Stochastic Modelling and Applied Probability 68

Carl Graham
Denis Talay

Stochastic Simulation and Monte Carlo Methods

Mathematical Foundations of Stochastic Simulation

Springer

Stochastic Mechanics
Random Media
Signal Processing
and Image Synthesis
Mathematical Economics and Finance
Stochastic Optimization
Stochastic Control
Stochastic Models in Life Sciences

Stochastic Modelling and Applied Probability

(Formerly:
Applications of Mathematics)

68

Edited by P.W. Glynn
Y. Le Jan

Advisory Board M. Hairer
I. Karatzas
F.P. Kelly
A. Kyprianou
B. Øksendal
G. Papanicolaou
E. Pardoux
E. Perkins
H.M. Soner

For further volumes:
www.springer.com/series/602

Carl Graham • Denis Talay

Stochastic Simulation and Monte Carlo Methods

Mathematical Foundations of Stochastic Simulation

Springer

Carl Graham
Centre de Mathématiques Appliquées
École Polytechnique, CNRS
Palaiseau, France

Denis Talay
INRIA
Sophia Antipolis, France

ISSN 0172-4568 Stochastic Modelling and Applied Probability
ISBN 978-3-642-39362-4 ISBN 978-3-642-39363-1 (eBook)
DOI 10.1007/978-3-642-39363-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013945076

Mathematics Subject Classification: 60H10, 65U05, 65C05, 60J30, 60E7, 65R20

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The extraordinary increase of computer capacity now encourages physicists, biologists, economists, and engineers to model and simulate numerically tremendously complex phenomena, in order to answer scientific questions, industrial needs, and societal requirements for risk evaluation and control. Therefore, experts in various fields aim to solve incredibly complex high dimensional systems of partial differential equations (PDE), couplings of local stochastic dynamics and deterministic macroscopic equations, etc.

Stochastic approaches appear to be useful, and sometimes mandatory, in the following two contexts. First, one cannot expect that very complex phenomena lead to perfectly calibrated mathematical models, or even to perfect mathematical models, so that uncertainties or stochastic components are involved in the equations. Second, stochastic numerical methods allow one to solve deterministic problems, of which the high dimension or singularities render classical deterministic methods of resolution intractable or inaccurate, provided that the solutions can be represented in terms of probability distributions of random variables or stochastic processes.

The combination of stochastic analysis and PDE theory are necessary to:

- obtain stochastic representations of solutions of deterministic PDE,
- construct effective stochastic numerical methods,
- obtain precise error estimates in terms of the numerical parameters of these methods, under the constraint that they take into account the critical situations where the stochastic numerical methods are used and the objectives of their users.

This monograph aims to introduce the reader to these difficult issues in a self-contained way. We particularly emphasize the essential role played by martingale theory in all the theoretical and numerical aspects of these issues. We also have devoted a substantial part of the book to the construction of simulation algorithms: the readers who are mainly interested in numerical issues may skip the mathematical proofs and concentrate on the algorithms and the convergence rate estimates.

Compared to other textbooks, this monograph presents the specificity of developing the mathematical tools which are necessary to construct effective stochastic

simulation methods and obtain accurate error estimates, and of reuniting in a non-classical way several advanced theoretical topics, some of which we now list.

We focus on non-asymptotic error estimates for Monte Carlo methods, on the backward martingale technique to prove the Strong Law of Large Numbers, and on elementary notions on logarithmic Sobolev inequalities to prove basic concentration inequalities. We place great emphasis on the pathwise construction and simulation of Poisson processes, discrete space Markov processes, and solutions of Itô stochastic differential equations.

We use the notions of infinitesimal generators and stochastic flows to establish stochastic representations of parabolic partial differential equations or integro-differential equations. We intensively use these stochastic representations of evolution equations to prove optimal error estimates for stochastic simulation methods. As explained in Chap. 1, stochastic numerical methods are used to compute quantities expressed in terms of the probability distribution of stochastic processes; we therefore essentially consider numerical errors in the weak sense rather than in an L^p-norm or in a pathwise sense which only provide crude information on the accuracy of practical simulations.

We present variance reduction techniques in nontrivial situations, which leads us to use optimized Girsanov transformations and to introduce the reader to stochastic optimization procedures.

For further information on the contents of the first two chapters, the reader is advised to consult the huge literature which concerns Central Limit Theorems, Edgeworth expansions, Large Deviations Principles, concentration inequalities, and simulation algorithms for finite-dimensional random variables. See, e.g., Devroye [10], Feller [15, 16], Petrov [43], Shiryayev [44], and references therein.

The first book on the discretization of stochastic differential equations is due to Milstein [37]. Several other books have been published on this topic with a rather different point of view to ours: most of them focus on particular applications or various discretization methods whereas, as already emphasized, we concentrate on the mathematical methodologies which allow one to get sharp convergence rates. Therefore we encourage the reader to consult the selected references below, references therein, and other useful references, to get further algorithmic or applied information on stochastic simulations.

For time dependent models, Monte Carlo methods are derived from the simulation of Markov processes, possibly discretized. In this context, e.g., Asmussen and Glynn [4] treat the mathematics of queueing theory and some related areas, with an emphasis on stationary regimes. Glasserman [20] focuses on numerical methods for financial models. Kloeden and Platen [28] present an extended catalog of variants of the Euler and Milstein discretization schemes for stochastic differential equations. Lapeyre, Pardoux and Sentis [32] present an overview of applications of Monte Carlo simulations of stochastic processes. Milstein and Tretyakov [38] notably study discretization methods for stochastic differential systems with symplectic structure, Hamiltonian systems, and small noise systems, layer simulation methods and random walk simulations for stochastic systems with boundary conditions. The CIME volume [21] developed the weak convergence of discretized pro-

cesses, with a strong emphasis on stochastic interacting particle systems related to non-linear partial differential equations.

We here have not tackled such important topics as Malliavin calculus techniques to get optimal convergence rates and develop variance reduction methods, long time simulations of ergodic Markov processes and the stochastic approximation of their invariant probability distributions, approximation methods for reflected and stopped diffusion processes, simulation of Lévy processes and discretization of Lévy driven stochastic differential equations, quantization simulation techniques, or exact simulation methods. These subjects need mathematical tools which are beyond the objectives of this first volume, and we will address them in a forthcoming volume.

Our other volumes will concern the stochastic simulation methods for Partial Differential Equations with non-smooth coefficients and the stochastic particle methods for the analysis and the numerical resolution of non-linear Partial Differential Equations.

We hope that Master or Ph.D. students with notions on stochastic calculus and researchers interested in the mathematical or numerical aspects of stochastic simulations will find this series of monographs useful, in order to be introduced into some advanced topics in Probability theory, to improve the accuracy and the confidence intervals of their simulations, or to acquire some fundamental knowledge before reading advanced research papers on numerical probability.

Palaiseau, France — Carl Graham
Sophia-Antipolis, France — Denis Talay
February 2012

Acknowledgements

We warmly thank Benjamin Jourdain for his very valuable comments on the first draft of this book, and he and Caroline Hillairet for having taught its contents with us at the École Polytechnique (Palaiseau, France).

Contents

Conventions and General Notation

A probability space $(\Omega, \mathscr{F}, \mathbb{P})$ will be given throughout the book. It supports all events, random variables, and stochastic processes in question. Often, $\mathbb{P}$-a.s. will be shortened to a.s. or omitted altogether.

Often a filtration $(\mathscr{F}_t) = (\mathscr{F}_t)_{t\in\mathbb{R}_+}$ or $(\mathscr{F}_n) = (\mathscr{F}_n)_{n\in\mathbb{N}}$ will be given on $(\Omega, \mathscr{F}, \mathbb{P})$. If not stated otherwise, processes will be assumed to be adapted to the filtration, stopping times, Brownian motions, (sub-, super-) martingales, Markov processes will be for the filtration, etc.

Similarly as for filtrations, sequences and processes will be denoted indifferently by (X_n), $(X_n, n \geq 1)$, (X_t), $(X_t)_{t\in\mathbb{R}_+}$, etc. For continuous time, sample paths will always be assumed to be right-continuous and have left limits.

In random variable or process constructions, draws are always assumed to be independent if not stated otherwise.

Paragraphs labeled "Algorithm" describe, in a compact fashion, either a precise mathematical construction, or with more practical details the basis for an actual simulation algorithm.

Abbreviations

a.s.	almost surely
c.d.f.	cumulative distribution function
i.i.d.	independent identically distributed
l.h.s.	left-hand side
r.h.s.	right-hand side
r.v.	random variable
resp.	respectively
s.t.	such that
w.r.t.	with respect to

Acronyms

CLT	Central Limit Theorem
LSI	logarithmic Sobolev inequality
ODE	ordinary differential equation

PDE partial differential equation
SDE stochastic differential equation
SLLN Strong Law of Large Numbers

General Mathematical Notation

$a \vee b$ $\max(a, b)$ for real a and b
$a \wedge b$ $\min(a, b)$ for real a and b
$\mathscr{B}(\mathscr{S})$ Borel σ-field generated by the topology of the space $\mathscr{S}$
$\sigma(\mathscr{S})$ smallest σ-field containing the set $\mathscr{S}$
$\sigma(X)$ smallest σ-field such that the random variable X is measurable
$\mathscr{V}$ state space; it is either discrete, or a closed subset of $\mathbb{R}^d$ with $\mathscr{B}(\mathscr{V})$

Number Sets, Euclidean Space, Matrix Space

$\mathbb{C}$ set of complex numbers, of the form $x + \mathrm{i}y$ for x and y in $\mathbb{R}$
$\mathbb{N}$ set of natural numbers, i.e., of non-negative integers $\{0, 1, \ldots\}$
$\mathbb{N}^*$ set of non-null natural numbers $\mathbb{N} - \{0\}$, i.e., of positive integers $\{1, 2, \ldots\}$
$\mathbb{Q}$, $\mathbb{Q}_+$ set of rational numbers, set of non-negative rational numbers
$\mathbb{R}$, $\mathbb{R}_+$ set of real numbers, set of non-negative real numbers
$\mathbb{R}^d$ Euclidean space of dimension d, with norm denoted by $|\cdot|$
$\mathbb{R}^{d\otimes r}$ space of $d \times r$ real matrices M, with norm $\|M\| = \mathrm{Trace}(MM^*)$

Sample Path Spaces

$\mathscr{C}(\mathbb{R}_+, \mathscr{V})$ space of continuous paths from $\mathbb{R}_+$ into $\mathscr{V}$, with product σ-field
$\mathscr{D}(\mathbb{R}_+, \mathscr{V})$ Skorohod space of right-continuous and left limited paths, with product σ-field

Part I
Principles of Monte Carlo Methods

The first part of this monograph starts by introducing and discussing the underlying principles of Monte Carlo methods.

The companion subject of the effective simulation on computers of random variables is then developed in several directions.

Discrete-time martingale theory is introduced, and the Strong Law of Large Numbers (SLLN) is given a proof using backward martingale convergence and the Kolmogorov zero-one law.

The random errors in Monte Carlo approximations are then discussed.

The Central Limit Theorem is stated and proved, and asymptotic confidence intervals are derived from it. Then, various rigorous results yielding non-asymptotic confidence intervals are given and partly proved: Berry–Esseen's and Bikelis' theorems, and confidence intervals based on concentration inequalities.

Lastly, the difficult topic of variance reduction is discussed, and several methods are explained.

Chapter 1
Introduction

Abstract This introduction starts by describing some motivations for using probabilistic models and stochastic simulations. Some examples are used to illustrate classical objectives of random simulations, and to derive relevant convergence criteria for which error estimates need to be developed. Then the goals and the organization of this monograph are presented.

1.1 Why Use Probabilistic Models and Simulations?

Probability theory was born from the study of games of chance and their remarkable properties. The issue was the construction of a mathematical model able to explain phenomena which seem universal, in the sense that they do not depend on a particular instance of a game, e.g., the convergence of empirical frequencies to deterministic values and the statistical properties of the random fluctuations around these limits, the occurrence of arbitrarily long sequences of events, etc.

In order to model games of chance and compute expected gains or loss risks, it was enough to define probability distributions on finite sets. However, to analyze the random fluctuations of the gains, as well as to model measurement uncertainties during physical experiments, the need to construct probability distributions on finite-dimensional spaces such as $\mathbb{R}^d$ arose at the end of the 18th century.

These constructions could not handle questions which appeared at the beginning of the 20th century. In order to model random time dependent quantities such as the history of a financial price or a temperature, one needs to construct probability distributions on suitable infinite-dimensional spaces. One can easily imagine that a rigorous mathematical definition of "sampling a continuous function" presents technical difficulties, and that other difficulties arise when one desires to design an algorithm to sample such a function on a computer.

Hence, the objective of this monograph is to present:

- examples of probability distributions on the space of continuous or piecewise continuous functions;
- examples of sampling methods for these probability distributions, and theoretical estimates describing sampling errors;
- a selection of applications which justify the examples and the need for theoretical sampling error estimates.

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_1, © Springer-Verlag Berlin Heidelberg 2013

The applications which we will study are typical of the situations where probabilistic models are used in practice. In the next section we emphasize two important issues which are not always well addressed in the literature, and then explain our selection of topics, results and techniques presented in the text.

1.1.1 What Are the Reasons for Probabilistic Models?

Probabilistic models are useful for *critical* situations only, that is, situations where deterministic models are inefficient or irrelevant. Such situations are becoming common in fundamental sciences and in engineering: it now is reasonable to use computers to simulate physical systems with large numbers of degrees of freedom, complex dynamics and short time scales; however, classical modeling procedures for such systems are inefficient. Let us give a few illustrative examples:

- A physical experiment which can be described by well-established physical laws, but for which certain model parameters are difficult to calibrate with accuracy. For example, the oil industry wishes to measure the geology and the cracks of the subsoil. For obvious reasons, the lack or inaccuracy of data is also inherent to the modeling of biological systems.
- A physical experiment which can be described by well-established physical laws, but which is submitted to a huge number of rapidly and unpredictably varying forces. For example, the shock absorbers of a car, whose mechanical fatigue studies are critical industrial issues, are submitted to the holes and bumps of the roads.
- A physical experiment which can be described by well-established physical laws, but has a number of degrees of freedom which is so large that one cannot list the reactions of each component of the system. For example, a huge set of colliding electrons and nuclei.
- A phenomenon driven by incompletely known physical laws. For example, exchanges of energy between particles and turbulent fluids, or interactions between solvents and polymers.
- A phenomenon which is not driven by physical laws and involves a huge number of heterogeneous sources of changes. For example, the price of a financial asset (one cannot establish a physical law which would describe the traders' psychological behavior when critical events occur) or the size of data flows on the Internet.

In addition, the complexity of certain phenomena necessitates to couple descriptions at micro and macro scales which combine random dynamics for the small scales and macroscopic equations for the large scales: such models are used to study the fatigue of mechanical structures with large numbers of degrees of freedom and submitted to wind and tide effects, the motion of huge assembles of atoms constituting a polymer in a turbulent fluid, etc.

So far we have not properly defined the probabilistic objects involved in the models we have in mind. The above examples illustrate the fact that probabilistic models

are often applied to time-varying phenomena. It is thus natural that *stochastic processes* are essential to probabilistic modeling.

To summarize, the imperfectly known parameters and geometries are considered as random, the poorly understood forces and the fast deformations of complex systems are described by random noises, the time evolution of systems which cannot be accurately measured is modeled by stochastic processes. Let us illustrate this summary by two particularly interesting examples.

Our first example concerns Molecular Dynamics. We quote the Ph.D. thesis of S. Park,[1] prepared under the supervision of K. Schulten who is a great specialist in the field:

> A protein is an unbranched chain of amino acids. (...) When placed in water (or, for some proteins, in a membrane), a protein folds into a particular structure which is also uniquely determined by the amino acid sequence. (...) In principle, the dynamics of biomolecules such as proteins must be described by solving Schrödinger's equation for every constituent particle (nuclei and electrons). However, this is impractical because of the large size of biomolecules. The Born–Oppenheimer approximation provides a more practical way to model large molecules. The basic idea is that since electrons are much lighter than nuclei, electrons can rapidly adjust to the motion of nuclei. (...) A model of molecular interactions in terms of nuclear coordinates (...) is called a *force field*. In principle, since proteins are supposed to fold by themselves, a good force field should be enough to obtain naive protein structures from simulations. But, whereas protein folding typically takes milliseconds or longer, today's molecular dynamics simulation is limited to the nanosecond time scale. This is because fast vibrational motions of atoms allow only a small time step for the integration of the equation of the motion. (...) Steered Molecular Dynamics is an efficient method which permits us to focus on important degrees of freedom while minimizing computational cost (...) The *reaction coordinate* is basically the most important degree of freedom, with the others being considered as fluctuations. With all the other degrees of freedom averaged out, the motion along the reaction coordinate is often well approximated by a diffusive Brownian motion on an effective potential.

Our second example concerns multi-scaled meteorological systems. The text below[2] has been written by a great specialist of stochastic models in Climatology, C. Penland of the NOAA Climate Diagnostics Center (Boulder, Colorado):

> The need to numerically model the interaction between geophysical processes having different timescales has led many modelers to represent rapidly varying components of the system as stochastic forcing. The methods these stochastic modelers use to do this are almost as numerous as the modelers themselves. Yet, there does exist a prescription for making the stochastic approximation in a systematic manner consistent with the multi-scale dynamics.
> Many of us are familiar with some of the Central Limit Theorem, which states how sums of weakly dependent quantities are approximately Gaussian distributed. There is another version that states the conditions under which a multi-scale dynamical system may be ap-

[1] Park, S.: Extracting equilibrium from nonequilibrium: free energy calculation from steered molecular dynamics simulations. Ph.D. thesis, University of Illinois at Urbana–Champaign (2004)

[2] Penland, C.: A stochastic approach to nonlinear dynamics: a review. Bull. Am. Meteorol. Soc. **84**, ES43–ES52 (2003)

> proximated as depending on the realizations of a white noise process, that is, as a stochastic differential equation.
> (...)
> The importance of stochastic differential equations to probabilistic forecasting is only now beginning to be appreciated in meteorology and climate research. The advances in the subject have already allowed us to make real-time predictions of climate systems much more complicated than the simple univariate, additive-noise case considered by Hassellman (1976) in his landmark paper. We hope, and expect, that climate researchers will increasingly value the great utility and intrinsic beauty of the theory.

1.1.2 What Are the Objectives of Random Simulations?

We have already emphasized that Probability theory allows to model complex phenomena whose states cannot be precisely deduced from accurate measurements. The study of such phenomena cannot be classical. Stochastic modeling consists in choosing the probability distributions of the data and aims to compute the probability distributions of important characteristics of the phenomena under consideration. Note that it would thus be useless and inconsistent to try to determine particular random states, since they would not give sufficient information on the desired probability distributions.

Let us revisit some of the examples we have already mentioned:

- In a (purely theoretical) perfect financial market (that is, complete and allowing no arbitrage), one is given a probability distribution to describe the time evolution of future stock prices, and one desires to compute mean values of contingent claims based on these stocks.
- In Neurosciences, one is given a probability distribution to describe the physical and chemical reactions along an axon, and one desires to determine the probability distribution of the passage times of the axon electric potential at the firing threshold.
- In network studies, one is given a probability distribution to describe the sizes and the emission dates of packets of bits sent to a network and one desires to estimate the probability that the network falls into congestion within a certain time interval.

To summarize, one desires *statistical* information on the system under consideration: the mean values of the velocity, the position, the energy, of the price; the probabilities of critical events such as long cracks, large financial losses, large number of individuals infected by an epidemic; the density of locations and speeds of particles, the probability distribution of neuronal firings. In all cases, one needs to compute a probability distribution, or an expectation, or probabilities of critical events.

The preceding quantities may often be obtained by solving deterministic equations. For example, contingent claims prices can often be expressed in terms of solutions of parabolic partial differential equations. It is actually the case of any

quantity of the type $\mathbb{E}\,\Psi(X_T)$, where Ψ is a given function and $(X_t, t \in [0,T])$ the solution of a stochastic differential equation.

The numerical resolution of these partial differential equations by deterministic methods is possible (and recommended) when the dimension of the state space is small enough. When the state space dimension increases, typically the numerical complexity of deterministic methods explodes and stochastic numerical methods are necessary to obtain quantitative information on probabilistic models and to solve fully deterministic equations, including equations describing purely deterministic phenomena.

The simplest situation of a deterministic quantity which requires probabilistic approximations is the following one: given an integrable function f in the hypercube $[0,1]^d$, one desires to approximate the integral

$$I := \int_{[0,1]^d} f(x_1, \ldots, x_d)\, dx_1 \cdots dx_d.$$

A classic deterministic method is the following.

Algorithm (Quadrature method) Discretize the hypercube to obtain a set of points $(\alpha_1^\ell, \ldots, \alpha_d^\ell)$ in $[0,1]^d$ and choose weights ρ_ℓ, appropriately for $1 \le \ell \le L$; approximate I by the weighted sum

$$\sum_{\ell=1}^{L} \rho_\ell f\left(\alpha_1^\ell, \ldots, \alpha_d^\ell\right).$$

How should one choose the number of points L?

The answer depends on the desired accuracy and on the dimension $d \ge 1$: typically, given the function f, if the desired accuracy $\varepsilon > 0$ requires that the quadrature method under consideration uses $M(\varepsilon)$ points when $d = 1$, then for the same accuracy $M(\varepsilon)^d$ points are necessary for dimension d.

Even for $d = 6$ (the dimension of the phase space in Mechanics) this number of points may be much too large to make the computation possible. Worse, in Statistical Physics the dimension d may well be of the order of hundreds or thousands.

In addition, the accuracies of classical quadrature methods are sensitive to the smoothness of the function f.

A probabilistic approach allows one to develop numerical methods whose complexity increases linearly only w.r.t. d and whose accuracy depends on the $L^2([0,1]^d)$ norm of f rather than the $L^\infty([0,1]^d)$ norm of its derivatives. Indeed, the integral I admits the following *probabilistic representation*:

$$I = \mathbb{E}\, f(U_1, \ldots, U_d),$$

where the U_i are independent random variables uniformly distributed on $[0,1]$. The Strong Law of Large Numbers yields that I can be approximated as follows.

Algorithm (Monte Carlo method) Simulate independent and uniformly distributed on [0, 1] random variables $U_i^{(\ell)}$ for $1 \le i \le d$ and $1 \le \ell \le N$. Approximate I by the random sum

$$\frac{1}{N} \sum_{\ell=1}^{N} f\left(U_1^{(\ell)}, \ldots, U_d^{(\ell)}\right).$$

Such a technique, consisting in approximating a deterministic quantity by an average of random values, is called a *Monte Carlo method*; N should be "large enough" to obtain a good accuracy, and typically increases linearly in d.

We cannot limit ourselves to consider this elementary situation. As we have seen in the preceding subsection, probabilistic models are often based on stochastic processes; therefore, we need to understand how to develop Monte Carlo methods based on sampled stochastic processes. In order to give a rigorous meaning to "sampling a stochastic process", we will define the *probability law of a stochastic process* as a probability measure on the space of the trajectories of the process.

We will limit ourselves to processes which are obtained by transformations of the Poisson process or of the Brownian motion, in particular the solutions of stochastic differential equations, because they are the core of most of the stochastic simulations used by practitioners.

We will carefully study the convergence rates of Monte Carlo methods in terms of the number of the simulations and other numerical parameters, particularly the discretization step of stochastic differential equations. We emphasize that our error analyses deeply reflect the objective of the numerical procedure, that is, the approximation of expectations of random variables or of functionals of stochastic processes.

To illustrate the relevance of our goal, let us quote the recommendations of the "Theory and Modeling in Nanoscience Report of the May 10–11 2002 Workshop" edited by the Basic Energy Sciences and Advanced Scientific Computing Advisory Committees to the Office of Science, U.S. Department of Energy:

> It was noted that upscaling (going to coarser length and time scales) typically results in the introduction of stochastic terms to reflect high-frequency motion at smaller scales; thus, more rigorous methods for including stochasm and more effective methods for characterizing and solving stochastic differential equations are required. New developments in nonequilibrium statistical mechanics, both of classical and quantum systems, are needed. For example, the recent discovery by nonequilibrium molecular dynamics, confirmed by nonlinear response theory, of violations of the second law of thermodynamics for nanoscale systems indicates the importance of new theoretical developments focused on nanoscale systems. Quantifying errors in calculated properties was identified as a significant problem that consists of two parts characterization of uncertainty due to inaccuracies inherent in the calculation (such as limitations of a force field in molecular dynamics or of a basis set in electronic structure calculations) and performing useful sensitivity analyses. The important aspect of nanoscale systems that makes them such a Theory Model Simulation (TMS) challenge is that the characterization of uncertainty will often have to be done in the absence of experimental data, since many of the property measurement experiments one would like to perform on nanoscale systems are impossible in many cases or unlikely to be done in cases where they are possible. Hence, the concept of self-validating TMS methods arises as a significant challenge in nanoscience. (...)

> Some degree of stochasticity is intrinsic at atomistic scales; hence, a Monte Carlo simulation will yield a fluctuation with statistics that must be carried to larger length scales. The numerical analysis of stochastic partial differential equations could enter here, the goal being a coarse solution with the right statistics. The theory of large deviation for stochastic PDEs allows the designing of sampling methods for so-called rare events. The long time integrations that are burdensome to the accuracy of deterministic models may actually be a benefit to the accuracy of statistical models. It is axiomatic to a mathematician that there are better ways than brute-force propagation of fine scales to attack a multiscale problem. A mathematician would argue that anyone requiring a billion degrees of freedom for a week of wall clock execution time is running the wrong algorithm. Indeed, the nanoscale community must learn to compute smarter and not just harder.

1.2 Organization of the Monograph

This monograph is organized so as to fulfill two main objectives:

- to describe the fundamental probabilistic numerical methods and analyze their convergence rates in terms of the numerical parameters of the simulations: e.g., the number of independent samples used in a Monte Carlo method and the time discretization step used in the discretization of a stochastic differential equation,
- to introduce in a self-contained way the advanced notions in Probability Theory and Stochastic Calculus (concentration inequalities, stochastic flows, probabilistic interpretation of partial differential equations, etc.) which are necessary to obtain sharp convergence rates.

In this perspective, the bibliography is voluntarily limited, and some fleeting references occur in footnotes. Exercises are proposed within the text, and usually bring complementary results. In addition, each chapter ends with problems which allow the reader to deepen the topics of the current chapter; the most difficult ones are flagged by the symbol ($\star$).

We now give a list of the various topics treated in this monograph after this introduction (Chap. 1).

Chapter 2 deals with the principles of Monte Carlo methods based on the Strong Law of Large Numbers (SLLN). A number of examples are described, some coming from concrete important applications. The description of algorithms of simulation of random variables follows; these are involved in the probabilistic numerical techniques which will be developed. The SLLN is proved using martingale techniques. These techniques play an essential role in modern Probability theory. Our presentation thus allows the reader to grow familiar with martingale theory, which is a core tool of the book.

Chapter 3 deals with the convergence rates for Monte Carlo methods. Its contents are different of most textbooks on Probability, which usually (and reasonably!) limit themselves to the classic Central Limit Theorem. Here, our goal is to provide theoretical results allowing to estimate the number of Monte Carlo simulations which are necessary to obtain a desired accuracy with a prescribed confidence interval. We

therefore develop *non-asymptotic* versions of the Central Limit Theorem (Berry–Esseen's and Bikelis' theorems, and concentration inequalities). We also tackle the difficult subject of variance reduction techniques for Monte Carlo methods.

Chapter 4 first introduces some practical and theoretical issues of modeling by means of Markov processes. Point processes are introduced in order to model jump instants. The Poisson process is then characterized as a point process without memory. The rest of the chapter consists in its rather detailed study, including various results concerning its simulation and approximation. This study is essential to understand the abstract constructions and the simulation methods for jump Markov processes developed in the following chapters.

Chapter 5 constitutes a rather detailed study of Markov processes with discrete state space. It focuses on sample path techniques in a perspective inspired by simulation needs. The relationship of these processes with Poisson processes and with discrete-time Markov chains is shown. Rigorous constructions and results are provided for Markov process with uniformly bounded jump rates. To this end, elements of the theory of bounded operators are introduced, which explain the relation between generator and semigroup, and provide a useful framework for the forward and backward Kolmogorov equations and the Feynman–Kac formula.

From Chap. 6 on, Markov processes with values in continuous space ($\mathbb{R}^d$ or one of its closed subsets) are considered. Their rigorous study requires advanced measure-theoretic tools, but we limit ourselves to developing the reader's intuition, notably by pathwise constructions leading to simulations.

In Chap. 6 we first emphasize the strong similarity between such Markov processes with constant trajectories between isolated jumps and discrete space ones. We then introduce Markov processes with sample paths following an ordinary differential equation between isolated jumps; these are often called piecewise deterministic Markov processes (PDMP). In both cases, the Kolmogorov equations and Feynman–Kac formula are established. This is applied to kinetic equations coming from statistical Mechanics. These describe the time evolution of the instantaneous distribution of particles in phase space (position-velocity), when the particle velocity jumps at random instants in function of the particle position and velocity.

Chapter 7 develops discretization schemes for stochastic differential equations and their applications to the probabilistic numerical resolution of deterministic parabolic partial differential equations. It starts with some important properties of Itô's Brownian stochastic calculus, and the existence and uniqueness theorem for stochastic differential equations with Lipschitz coefficients.

Then, using probabilistic techniques only, existence, uniqueness, and smoothness properties are proved for solutions of parabolic partial differential equations. To this end, we show that stochastic differential equations with smooth coefficients define stochastic flows, and we prove some properties of such flows. We are then in a position to prove an optimal convergence rate result for the discretization schemes.

Chapter 8 deepens the variance reduction subject and focuses on the Monte Carlo methods for deterministic parabolic partial differential equations. This topic requires advanced notions in stochastic calculus, particularly the Girsanov theorem, that we state and discuss at the beginning of the chapter.

Chapter 9 introduces a few theoretical and practical issues raised by stochastic optimization algorithms. These algorithms are efficient numerical tools in various applications and they are at the basis of the variance reduction techniques studied in the preceding chapter. We limit ourselves to a simple framework which allows to obtain convergence results by means of the dynamical system and martingale theories, without too many technical details.

Remark We do not tackle the difficult subject of testing, validating and comparing various simulation methods on full-scale problems encountered in Engineering, Physics, Biology, etc. Many academic papers report on numerical experiments for toy models. However we emphasize that each full-scale problem requires specific numerical approaches and strong expertise to control the discretization error and reduce the variance in an effective way. An entire volume written by practitioners would be necessary to present these topics and provide useful numerical recipes. The accuracy analysis of these numerical methods relies on the theoretical methodology which we develop here.

We also emphasize that full-scale stochastic problems concern complex models and, therefore, the question of accuracy should be completed by the question of robustness of the results with respect to modeling errors.

Chapter 2
Strong Law of Large Numbers and Monte Carlo Methods

Abstract The principles of Monte Carlo methods based on the Strong Law of Large Numbers (SLLN) are detailed. A number of examples are described, some of which correspond to concrete problems in important application fields. This is followed by the discussion and description of various algorithms of simulation, first for uniform random variables, then using these for general random variables. Eventually, the more advanced topic of martingale theory is introduced, and the SLLN is proved using a backward martingale technique and the Kolmogorov zero-one law.

2.1 Strong Law of Large Numbers, Examples of Monte Carlo Methods

The fundamental result for the numerical probability field is the Strong Law of Large Numbers, which will be proved at the end of the chapter.

2.1.1 Strong Law of Large Numbers, Almost Sure Convergence

A fundamental convergence result will now be stated.

Theorem 2.1 (Strong Law of Large Numbers) *Let* $(\xi^{(\ell)}, \ell \geq 1)$ *be a sequence of independent and identically distributed random variables with values in* $\mathbb{R}^d$. *Assume that*

$$\mathbb{E}\left|\xi^{(1)}\right| < \infty. \tag{2.1}$$

For $N \geq 1$, *denote the empirical mean of* $(\xi^{(1)}, \ldots, \xi^{(N)})$ *by*

$$\hat{S}_N := \frac{1}{N} \sum_{\ell=1}^{N} \xi^{(\ell)}.$$

Then, the Strong Law of Large Numbers holds true:

$$\lim_{N \to \infty} \hat{S}_N = \mathbb{E}\left(\xi^{(1)}\right), \quad \mathbb{P}\text{-}a.s. \tag{2.2}$$

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_2, © Springer-Verlag Berlin Heidelberg 2013

Remark 2.1 The Strong Law of Large Numbers admits a reciprocal statement, which we admit: if $\mathbb{E}|\xi^{(1)}| = \infty$ then the sequence $(\hat{S}_N, N \geq 1)$ diverges $\mathbb{P}$-a.s.

The Strong Law of Large Numbers can be stated as follows: the sequence of empirical means $(\hat{S}_N(\omega), N \geq 1)$ converges to $\mathbb{E}(\xi^{(1)})$ almost surely in ω. This is a particular case of almost sure convergence, the definition of which we now recall.

Definition 2.1 A property is said to hold almost surely (a.s.) if it holds except on an event of probability zero. The notation $\mathbb{P}$-a.s. is used to stress the underlying probability measure. In particular, a sequence $(\xi_N, N \geq 0)$ of random variables converges almost surely to a random variable ξ defined on the same probability space if

$$\mathbb{P}\Big(\omega \in \Omega : \lim_{N\to\infty} \xi_N(\omega) = \xi(\omega)\Big) = 1.$$

The Strong Law of Large Numbers is at the core of the following **Monte Carlo method**. Let γ be some quantity which must be approximated numerically. Assume that there exists a function f and a family $(X^{(1)}, \ldots, X^{(N)})$ of independent and identically distributed random variables, which are easy to simulate on computers[1] and satisfy

$$\mathbb{E} f\big(X^{(1)}\big) = \gamma. \tag{2.3}$$

Then, except on an event of probability zero, γ can be approximated as follows.

Algorithm (Monte Carlo method) Draw a sample $(X^{(1)}(\omega), \ldots, X^{(N)}(\omega))$, and approximate γ by the empirical mean:

$$\gamma \simeq \hat{S}_N(\omega) := \frac{1}{N} \sum_{\ell=1}^{N} f\big(X^{(\ell)}(\omega)\big).$$

This is a "good" approximation as soon as N is chosen "large enough". However the SLLN does not make precise the convergence rate of $\hat{S}_N$. Rigorously proving the SLLN and finding its precise convergence rate is one of our main goals in this chapter and the next one.

To summarize: the Monte Carlo methods in this book consist in:

- exhibiting a **probabilistic representation** of γ of the type (2.3) such that the probability distribution of $X^{(1)}$ can efficiently be simulated,
- and then applying the Strong Law of Large Numbers in order to approximate γ.

[1] This means that there exists a low complexity algorithm for generating sequences of independent samples from their common probability distribution.

Versions of the Strong Law of Large Numbers, under various sets of hypotheses can be proved in many ways. In particular, it is unnecessary to assume that the random variables $\xi^{(\ell)}$ are independent or identically distributed.

We will prove Theorem 2.1 in Sect. 2.3. We choose to use martingale techniques because this family of processes plays an important role in the sequel.

In the rest of this section we present some examples of Monte Carlo methods.

2.1.2 Buffon's Needle

Divide the two-dimensional space into vertical strips whose width is 1 cm. Throw at random a needle whose length is also 1 cm. What is the probability that the needle intersects one of the vertical lines?

To answer this question, one needs to make precise the probabilistic model. For instance, we define the random throwing of the needle as follows: the distance X of the center of the needle to the next line at its left side is a random variable with uniform distribution on $[0, 1]$, and the angle θ between the needle and the horizontal axis is a random variable with uniform distribution on $[-\frac{\pi}{2}, \frac{\pi}{2}]$ which is independent of X. The needle intersects a vertical line if

$$X(\omega) \in \left[0, \frac{1}{2}\cos\big(\theta(\omega)\big)\right] \cup \left[1 - \frac{1}{2}\cos\big(\theta(\omega)\big), 1\right].$$

An easy calculation then shows that the desired probability is $\frac{2}{\pi}$.

In 1850, an astronomer from Zürich, R. Wolf, approximated π by the Monte Carlo method using 5000 samples: he set $\xi^{(\ell)}(\omega) = 1$ when the needle intersected a vertical line at the sample ℓ, computed the average

$$\frac{2 \times 5000}{\xi^{(1)} + \cdots + \xi^{(5000)}},$$

and obtained 3.1596 as an approximation.

2.1.3 Neutron Transport Simulations

Consider a bounded continuous map λ from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}_+$ (in neutron transport theory this map is called a scattering diffusion cross-section). In addition, for any (x, y) in $\mathbb{R}^d \times \mathbb{R}^d$ a continuous probability density $\pi^{x,y}$ on $\mathbb{R}^d$ is given.

The random time evolution of the position of a neutron is described by the solution (X_t) of the differential equation

$$\frac{dX_t}{dt} = Y_t, \qquad (X_0, Y_0) = (x, y), \tag{2.4}$$

where the velocity (Y_t) is a pure jump process in the following sense: for any ω, the map $t \to Y_t(\omega)$ is piecewise constant and right continuous; the jump times of (Y_t) and the jump amplitudes are random. For any integer n and pair (x, y), the value of (Y_t) at the nth jump is a random variable with probability density $\pi^{x,y}$, where (x, y) is the state of (X_t, Y_t) at the time immediately preceding this jump; in other words, for any bounded continuous function f, if S_n is the nth jump time, the conditional expectation of $f(Y_{S_n})$ knowing that the state of (X_t, Y_t) immediately before S_n is (x, y), is equal to

$$\int_{R^d} f(z)\pi^{x,y}(z)\,dz.$$

Denote by T_n the time interval between S_n and S_{n+1}, that is, $T_n := S_{n+1} - S_n$. Knowing that the state of (X_t, Y_t) at time S_n is (x', y'), the distribution function of the random variable T_n is

$$F^{x',y'}(t) = 1 - \exp\Big(-\int_0^t \lambda\big(X_s^{x',y'}, y'\big)\,ds\Big),$$

where $X_s^{x',y'}$ solves (2.4) with $X_0^{x',y'} = x'$ and $Y_s = y'$ for any s. Note that the function $F^{x',y'}$ is independent of n. In addition, for any $i \geq 0$, the random variables T_i and $Y_{S_{i+1}}$ are independent.

The stochastic process (X_t, Y_t) is constructed by recursively solving (2.4) on each time interval $[S_n, S_{n+1}[$ with $Y_t = Y_{S_n}$. The pair (X_t, Y_t) is a homogeneous Markov process, and called a transport process. When λ and π do not depend on the space variable x, (Y_t) is called a pure jump process and describes the motion of particles in a homogeneous environment.

To simplify the notation we now limit ourselves to the case $d = 1$. Let g be a function from $\mathbb{R}^2$ to $\mathbb{R}$. Suppose that there exists a function $u(t, x, y)$ of class $\mathscr{C}^\infty(\mathbb{R}_+ \times \mathbb{R}^2)$, bounded with bounded derivatives of all orders, and such that

$$\begin{aligned}\frac{\partial u}{\partial t}(t,x,y) &= y\frac{\partial u}{\partial x}(t,x,y) - \lambda(x,y)u(t,x,y) + \lambda(x,y)\int_{\mathbb{R}} u(t,x,z)\pi^{x,y}(z)\,dz,\\ &\qquad t > 0,\ x \in \mathbb{R},\ y \in \mathbb{R},\\ u(0,x,y) &= g(x,y).\end{aligned} \tag{2.5}$$

One can show that

$$u(t,x,y) = \mathbb{E}_{x,y}\, g(X_t, Y_t), \tag{2.6}$$

where $\mathbb{E}_{x,y}$ denotes the conditional expectation knowing that the position and velocity at time 0 respectively are x and y.

The Monte Carlo method to approximate $u(t, x, y)$ consists in simulating large number of trajectories of the process (X_t, Y_t). The above construction of the process provides an algorithm of simulation of each trajectory.

This topic will be further developed in Sect. 6.3.4.

2.1.4 Stochastic Numerical Methods for Partial Differential Equations

The probabilistic representation (2.6) for the integro-differential equation (2.5) allowed us to construct a Monte Carlo method. This methodology can be extended to numerous linear and non-linear partial differential equations, provided that their solutions satisfy representations of the type

$$u(t,x) = \mathbb{E}\Psi\big(Z(t,x)\big),$$

where $(Z(t,x))$ is a family of suitable random variables.

Let us give an elementary example. Let ν be a strictly positive number and $u(t,x)$ be the solution of the heat equation

$$\frac{\partial u}{\partial t}(t,x) = \nu \Delta u(t,x), \quad \forall (t,x) \in]0,T] \times \mathbb{R}^d,$$

whose initial condition $u(0,\cdot) = u_0(\cdot)$ is assumed, say, to be continuous and bounded. By using the analytical expression of the Gaussian density of W_t one readily checks that

$$\forall (t,x) \in [0,T] \times \mathbb{R}^d, \quad u(t,x) = \mathbb{E}u_0(x + \sqrt{2\nu}W_t),$$

where (W_t) is an $\mathbb{R}^d$ valued standard Brownian motion (thus, for any t, the components of the random vector W_t are independent and Gaussian, have zero mean and variance equal to t). Therefore one can approximate $u(t,x)$ by

$$\frac{1}{N}\sum_{\ell=1}^{N} u_0\big(x + \sqrt{2\nu t}\xi^{(\ell)}(\omega)\big),$$

where the $\{\xi^{(\ell)}\}$ are $\mathbb{R}^d$ valued independent Gaussian vectors with zero mean and unit covariance matrix.

The linear parabolic partial differential equations related to European option prices in classical diffusion models are examples of equations whose solutions admit probabilistic representations. However these representations, which are called Feynman–Kac's formulas, involve processes which are much more complex than Brownian motions, that is, the solutions of stochastic differential equations (see Chap. 7).

Another example is the Poisson equation in $\mathbb{R}^d$

$$Lu(x) := \operatorname{div}\big(a(x)\nabla u(x)\big) = f(x),$$

where $a(x)$ is a real-valued function. This equation arises in various fields, e.g., in Geophysics and in Molecular Dynamics. When the function $a(x)$ is smooth, under

suitable other hypotheses, one can prove the following equality which is analogous to (2.6):

$$u(x) = \int_0^\infty \mathbb{E}_x\left(f(X_t) - \int f(\xi)\mu(d\xi) \right) dt, \tag{2.7}$$

where (X_t) is the solution to a certain stochastic differential equation, $\mathbb{E}_x$ denotes the conditional expectation knowing that X_0 is equal to x, and μ is the limit probability law, when t tends to infinity, of the law of X_t. The stochastic numerical method combines the standard Monte Carlo method and long time simulations of (X_t): this leads to important numerical difficulties which are current subjects of research.

In addition, one often needs to consider discontinuous functions $a(x)$. For example, in Geophysics, the discontinuities of $a(x)$ reflect the soil heterogeneity. In such cases, the formula (2.7) does not involve the solution of a classical stochastic differential equation and, when the state space is multi-dimensional, the construction of *easy-to-simulate* processes (X_t) satisfying (2.7) is being investigated by many authors.

Stochastic numerical methods are being developed for various Partial Differential Equations, including non-linear ones such as Boltzmann equations, Vlasov equations, Navier–Stokes equations, Burgers equation, variational inequalities, etc. This difficult subject is out of the scope of this monograph.

We conclude this subsection by emphasizing three advantages of the numerical resolution of partial differential equations by stochastic methods: it not only allows one to solve problems in large dimension, but also:

- Monte Carlo methods allow to compute solutions whose gradient is locally very large: whereas deterministic methods require thin grids in the areas where the gradient of the solution is large, stochastic particles methods are grid-free and concentrate the simulated particles, and therefore the numerical information, in these areas.
- Most often, Monte Carlo methods are simpler and faster to code than deterministic methods, and the computer programs for stochastic numerical methods are easier to modify and adapt.
- Monte Carlo methods are naturally propitious for parallel or grid computing.

2.2 Simulation Algorithms for Simple Probability Distributions

Before introducing the more advanced material necessary for the proof of the SLLN, we pursue the subject of stochastic simulation, which will play a key role in the actual implementation of the Monte Carlo methods developed in the sequel.

We describe various methods to simulate samples, first from the uniform distribution on $[0, 1]$, and then from classical probability distributions on $\mathbb{R}$ or $\mathbb{R}^d$.

For more insight in this topic, we recommend for instance the books of Devroye [10] and Asmussen and Glynn [4].

2.2.1 Uniform Distributions

The following theorem, which has many variants (see, e.g., Kuipers and Niederreiter [29]) allows one to produce sequences (u_n) on $[0,1]$ which are uniformly distributed in the following sense:

$$\forall 0 \le a \le b \le 1, \quad \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{(a,b)}(u_j) = b - a.$$

Theorem 2.2 *Let θ be a positive irrational number. The sequence*

$$u_n = n\theta \pmod 1$$

is dense everywhere in $[0,1]$ *and is uniformly distributed in* $[0,1]$.

The sequences (u_n) in the above theorem have poor statistical properties: in particular, two consecutive terms in the sequence are strongly correlated. In addition, irrational numbers cannot be represented exactly in computers.

Many algorithms have been designed to generate sequences of "pseudo-random" numbers with statistical properties close to those of sequences of samples of independent and uniformly distributed random variables. The most frequently used pseudo-random number generators are congruential methods.

Algorithm (Congruential method) Chose a triple (a, m, v_0) of integers, and compute inductively the successive samples u_k from the formula

$$v_k = a v_{k-1} = a^k v_0 \pmod m, \qquad u_k = \frac{v_k}{m}.$$

In practice, often $m = 2^\alpha$ is chosen, where α is the number of bits of the computer: then the congruence calculation reduces to truncating a bit sequence. The following simple statement indicates relevant choices of the parameters in order to maximize the periodicity of the method.[2]

Proposition 2.1 *If $m \ge 2^\alpha$ and $\alpha \ge 4$, then the period of the congruential method is less than $\frac{m}{4}$, and this upper bound is attained when v_0 is odd and $a = 3 \pmod 8$ or $a = 5 \pmod 8$.*

The preceding choices do not suffice to generate sequences with good statistical properties, that is, which statistically behave as sequences of independent samples from the uniform distribution. An example of a poor generator is

$$v_{n+1} = \left(2^{16} + 3\right) v_n \pmod{2^{32}}.$$

[2] This result is originally due to M. Greenberger, "Notes on a new pseudo-random number generator", J. Assoc. Comput. Mach. **8**, 163–167 (1961).

For a survey on random number generators and a discussion on statistical tests issues, see, e.g., L'Ecuyer [33] or Gentle [19] and references therein. For theoretical issues, we refer to Niederreiter [40].

Initialization of the Samples

Monte Carlo simulations require very long sampling sequences. The root v_0 of the generator must be chosen once only, before the very first trial.

Given a root, a good generator will produce a sampling sequence with good statistical properties; however, two different sequences issued from different roots may be correlated.

We do not recommend the use of automatic initializations, e.g., by means of the computer internal clock. Being able to choose the same root in several runs of a simulation program may be useful to correct programming errors.

A Natural Question

Under which conditions is a simulation method of the uniform distribution satisfying? This is a critical issue without a universal answer.

In practice, one tests the uniform distribution hypothesis and the independence hypothesis of sampling sequences by using classical statistical procedures such as the Kolmogorov–Smirnov test, the χ^2 test, etc.

For an extended discussion on this subject and for analyses of efficient statistical tests, see Asmussen and Glynn [4], L'Ecuyer and Hellekaleke [34] and L'Ecuyer and Simard [35], for example.

Modern generators are often non-linear: for a survey on this issue, see for instance Niederreiter and Shparlinski [41].

2.2.2 Discrete Distributions

A probability distribution on a discrete set $\{x_1, x_2, \ldots\}$ is given by the corresponding probability weights $p_1, p_2, \ldots$, and a random variable X has this distribution if

$$\mathbb{P}(X = x_1) = p_1, \quad \mathbb{P}(X = x_2) = p_2, \quad \ldots.$$

Such a random variable X can be simulated by the following procedure:

Algorithm (Discrete distribution) To obtain a sample x from a discrete distribution giving weight p_i to x_i for $i \geq 1$: draw a sample u from the uniform distribution on $[0, 1]$, and set $x = x_n$ for the $n \geq 1$ satisfying $\sum_{i=1}^{n-1} p_i < u \leq \sum_{i=1}^{n} p_i$.

2.2.3 Gaussian Distributions

There are various simulation methods for the standard $\mathcal{N}(0,1)$ Gaussian distribution, with density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ on $\mathbb{R}$.

An approximate simulation method uses the Central Limit Theorem, which will be recalled in Chap. 3:

Algorithm (Gaussian distribution, approximate) To obtain an approximate sample x of the $\mathcal{N}(0,1)$ distribution, draw N independent samples $u_1, \ldots, u_N$ of the uniform distribution on $[0,1]$, and compute

$$x = \sqrt{\frac{12}{N}}\left(\sum_{i=1}^{N} u_i - \frac{N}{2}\right).$$

In practice, the choice $N = 12$ is often made: empirical studies show that the corresponding sampling sequences have satisfying statistical properties, and the preceding formula simplifies into $\sum_{i=1}^{12} u_i - 6$.

The Box–Muller Method

This is an exact simulation technique, which moreover seems more efficient than the preceding algorithm in terms of computational time.

Two independent samples of the $\mathcal{N}(0,1)$ Gaussian distribution are obtained by using the following result (due to Wiener):

Proposition 2.2 *Let U and V be independent random variables uniformly distributed on $[0,1]$. Then the random variables X and Y defined by*

$$X = \sqrt{-2\log U}\,\sin(2\pi V), \qquad Y = \sqrt{-2\log U}\,\cos(2\pi V),$$

have a $\mathcal{N}(0,1)$ Gaussian distribution and are independent.

From this we deduce the following simulation algorithm.

Algorithm (Gaussian distribution, Box–Muller) To obtain two independent samples x and y from the $\mathcal{N}(0,1)$ Gaussian distribution: draw two independent samples u and v from the uniform distribution on $[0,1]$, and compute

$$x = \sqrt{-2\log u}\,\sin(2\pi v), \qquad y = \sqrt{-2\log u}\,\cos(2\pi v).$$

In order to simulate a random variable X with $\mathcal{N}(m,\sigma^2)$ Gaussian distribution, observe that if Y is a $\mathcal{N}(0,1)$ Gaussian variable then $X := \sigma Y + m$ has the appropriate distribution. More generally, the following allows to simulate a Gaussian random vector $X = (X^1, \ldots, X^d)$ with mean vector m and covariance matrix C.

Algorithm (Gaussian vector distribution) To obtain a sample x from the $\mathcal{N}(m, C)$ distribution, where m is in $\mathbb{R}^d$ and C is a symmetric non-negative $d \times d$ matrix: compute

$$\sigma := (\sigma_{ij})_{1\le i,j\le d} = \begin{cases} \sigma_{i1} := \frac{C_{i1}}{\sqrt{C_{11}}}, & 1 \le i \le d, \\ \sigma_{ii} := \sqrt{C_{ii}} - \sum_{j=1}^{i-1} |C_{ij}|^2, & 1 < i \le d, \\ \sigma_{ij} := \frac{C_{ij} - \sum_{k=1}^{j-1} \sigma_{ik}\sigma_{jk}}{\sigma_{jj}}, & 1 < j < i \le d, \\ \sigma_{ij} := 0, & i < j \le d, \end{cases}$$

draw a vector of independent samples $y = (y^1, \ldots, y^d)$ from the $\mathcal{N}(0, 1)$ Gaussian distribution, and compute $x = \sigma y + m$.

2.2.4 Cumulative Distribution Function Inversion, Exponential Distributions

Monte Carlo methods for neutron transport partial differential equations, network models, neurons firing train models, etc., require to sample the exponential distribution $\mathcal{E}(\lambda)$ with parameter $\lambda > 0$, with density $\lambda e^{-\lambda x} \mathbb{1}_{\{x\ge 0\}}$ on $\mathbb{R}$ (actually, $\mathbb{R}_+$).

This issue will be seen to be crucial in the sequel, and we will solve it using a general result for simulation of real random variables: the cumulative distribution function inversion method.

Definition 2.2 The cumulative distribution function (c.d.f.) of a probability distribution $\mathbb{P}$ on $\mathbb{R}$, or of a real random variable X with distribution P, is the function

$$F : x \in \mathbb{R} \mapsto F(x) = P\big((-\infty, x]\big) = \mathbb{P}(X \le x).$$

The (left-continuous) inverse of a c.d.f. F is the function

$$F^{\leftarrow} : u \in [0, 1] \mapsto F^{\leftarrow}(u) = \inf\big\{y \in \mathbb{R} : F(y) \ge u\big\}.$$

Note that if F is a bijection then $F^{\leftarrow} = F^{-1}$, and that the c.d.f. F characterizes the probability distribution P.

Theorem 2.3 *Let P be a probability distribution on $\mathbb{R}$ with c.d.f. F, and $F^{\leftarrow}$ be its inverse. If U is a uniformly random variable on $[0, 1]$, then the random variable $X := F^{\leftarrow}(U)$ has c.d.f. F and hence distribution P.*

Exercise 2.1 Prove this result.

An important application of the c.d.f. inversion method is the simulation of exponential $\mathcal{E}(\lambda)$ random variables.

Algorithm (Exponential distribution) To obtain a sample x from the $\mathscr{E}(\lambda)$ exponential distribution, $\lambda > 0$: draw a sample u from the uniform distribution on $[0, 1]$, and compute

$$x = -\frac{1}{\lambda}\log(u).$$

Another similar application is for Cauchy random variables.

Algorithm (Cauchy distribution) To obtain a sample x from the Cauchy distribution with density function $\frac{\sigma}{\pi(x^2+\sigma^2)}$ on $\mathbb{R}$, for $\sigma > 0$: draw a sample u from the uniform distribution on $[0, 1]$, and compute $x = \sigma\tan(\pi u)$.

The actual implementation of the c.d.f. inversion method requires an explicit representation of the function $F^{\leftarrow}$ (i.e., of F^{-1} when F is a bijection).

In practice, an alternative procedure consists in the numerical resolution of the equation $F(x) = u$ for any sampled value u of U, but the numerical cost may be high. The Newton–Raphson method is an example of this procedure.

Algorithm (Newton–Raphson method) To obtain a sample x from a strictly positive density f on $\mathbb{R}$, with c.d.f. $F(\cdot) = \int_0^{\cdot} f(y)\,dy$: draw a sample u from the uniform distribution on $[0, 1]$, set $x_0 = u$, compute

$$x_{k+1} = x_k - \frac{F(x_k) - u}{f(x_k)}, \quad k \geq 0$$

up to the step ℓ at which $|x_{\ell+1} - x_\ell|$ is less than a prescribed threshold, and set $x = x_{\ell+1}$.

2.2.5 Rejection Method

The rejection method is often used to simulate a random vector with density f on $\mathbb{R}^d$. It consists in choosing a density g on $\mathbb{R}^d$ such that:

- the random vectors with density g are easy to simulate (for instance, g is a Gaussian density),
- there exists $\varepsilon > 0$ such that $h(x) := \varepsilon\frac{f(x)}{g(x)} \leq 1$ for all x.

Then we proceed as follows.

Algorithm (Rejection method) To obtain a sample x from the density f:

- draw independent samples y from the density g and u from the uniform distribution on $[0, 1]$, then
- – if $u \leq h(y) := \varepsilon\frac{f(y)}{g(y)}$, accept the sample y for x (i.e., set $x = y$)
  – else, reject it and start over again,

(repeat until successful).

Note that a random number of rejections occur before a sample is accepted. This number of trials depends on the sampling sequence and on the chosen density g. This observation is important in practice: the rejection method is efficient if the acceptation rate is high.

The justification of the rejection method, and the control on the number of iterations, is based on what follows.

Proposition 2.3 *Let the random variables $Y_1, Y_2, \ldots$ have density g and $U_1, U_2, \ldots$ be uniform on $[0,1]$, and all be independent. Let the rank and value of the first accepted sample be given by the random variables*

$$M := \inf\left\{k \geq 1 : U_k \leq \varepsilon \frac{f(Y)}{g(Y)}\right\}, \qquad X := Y_M.$$

Then M is a.s. finite, $\mathbb{P}(M = k) = \varepsilon(1-\varepsilon)^{k-1}$ for $k \geq 1$ (geometric distribution), and X has density f and is independent of M. In particular $\mathbb{E}(M) = 1/\varepsilon$.

Proof By definition, for any $k \geq 2$ and open set A in $\mathbb{R}^d$,

$$\begin{aligned}\mathbb{P}(M = k, Y_M \in A) &= \mathbb{P}(M = k, Y_k \in A) \\ &= \prod_{\ell=1}^{k-1} \mathbb{P}\left(U_\ell > \varepsilon \frac{f(Y_\ell)}{g(Y_\ell)}\right) \mathbb{P}\left(U_k \leq \varepsilon \frac{f(Y_k)}{g(Y_k)}, Y_k \in A\right)\end{aligned}$$

and, since f is a probability density,

$$\mathbb{P}\left(U_1 \leq \varepsilon \frac{f(Y)}{g(Y)}\right) = \int \varepsilon \frac{f(y)}{g(y)} g(y)\, dy = \varepsilon \int f(y)\, dy = \varepsilon.$$

Therefore, for all $k \geq 1$,

$$\mathbb{P}(M = k, Y_M \in A) = (1-\varepsilon)^{k-1} \varepsilon \int_A \frac{f(y)}{g(y)} g(y)\, dy = \varepsilon(1-\varepsilon)^{k-1} \int_A f(y)\, dy.$$

Choosing $A = \Omega$ leads to the distribution of M and shows that M is a.s. finite; in addition, the product form of the right-hand side shows that M and Y_M are independent. Finally, we also deduce

$$\mathbb{P}(Y_M = A) = \int_A f(y)\, dy,$$

which shows that Y_M has density f. □

The preceding result and its proof can easily be extended, for instance as in the following proposition.

Proposition 2.4 *Let $(S, \mathscr{S})$ be a measurable space. Let ν and μ be two probability measures on this space. Suppose that μ is absolutely continuous w.r.t. ν and*

$$\exists \varepsilon > 0, \quad h(x) := \varepsilon \frac{d\mu}{d\nu}(x) \leq 1, \quad \nu\text{-a.s.}$$

Let (Y_n, I_n) be a sequence of independent and identically distributed random variables taking values in $S \times \{0, 1\}$. Suppose that the common probability distribution of the Y_i is ν and that

$$\mathbb{P}(I_1 = 1 \,|\, Y_1) = h(Y_1), \quad \text{a.s.}$$

Set

$$M := \inf\{k \geq 1 : I_k = 1\}, \qquad X := Y_M.$$

Then M is a.s. finite, $\mathbb{E}(M) < \infty$, and the probability distribution of X is μ.

2.3 Discrete-Time Martingales, Proof of the SLLN

In this section we introduce the important notion of martingale processes, which plays a key role in the sequel. We then prove the Strong Law of Large Numbers using a technique which relies on the convergence of a backward martingale.

Our presentation is inspired by Jacod and Protter [24]. For a rigorous construction of the abstract conditional expectation and a systematic study of discrete-time martingales, an excellent reference book is Williams [47].

2.3.1 Reminders on Conditional Expectation

A probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is given. Let X be an integrable $\mathbb{R}^n$-valued random variable:

$$X \in L^1\big(\Omega, \mathscr{F}, \mathbb{P}; \mathbb{R}^n\big).$$

For a sub-σ-field $\mathscr{G}$ of $\mathscr{F}$, the conditional expectation $\mathbb{E}(X \,|\, \mathscr{G})$ of X knowing $\mathscr{G}$ is defined as (a representative of) the class of random variables in $L^1(\Omega, \mathscr{G}, \mathbb{P}; \mathbb{R}^n)$ satisfying the characteristic property

$$\mathbb{E}\big(\mathbb{E}(X \,|\, \mathscr{G}) Z\big) = \mathbb{E}(XZ), \quad \forall Z \in L^\infty(\Omega, \mathscr{G}, \mathbb{P}; \mathbb{R}), \tag{2.8}$$

where it is in fact enough to consider all Z of the form $\mathbb{1}_A$ for A in $\mathscr{G}$.

It can be proved that any two random variables in this class are equal except on a null probability set. The fact that statements involving conditional expectations hold true $\mathbb{P}$-a.s. is often left implicit.

If Y is an $\mathbb{R}^k$-valued random variable, the conditional expectation of X knowing Y is defined by

$$\mathbb{E}(X \mid Y) := \mathbb{E}\big(X \mid \sigma(Y)\big) \in L^1\big(\Omega, \sigma(Y), \mathbb{P}; \mathbb{R}^n\big),$$

where $\sigma(Y)$ is the σ-field generated by Y, constituted of all events $\{Y \in A\}$ for sets A in $\mathscr{B}(\mathbb{R}^k)$. Doob's lemma shows that any $\sigma(Y)$-measurable random variable is of the form $f(Y)$ for some Borel function f. Therefore, $\mathbb{E}(X \mid Y)$ can be characterized as the $\mathbb{P}$-a.s. unique integrable random variable of the form $f(Y)$ for some Borel function f satisfying

$$\mathbb{E}\big(f(Y)g(Y)\big) = \mathbb{E}\big(Xg(Y)\big) \quad \text{for every real bounded Borel function } g.$$

One can then set $\mathbb{E}(X \mid Y = y) := f(y)$. Note that all suitable functions f are identical except on sets A such that $\mathbb{P}(Y \in A) = 0$. As expected, if $\mathbb{P}(Y = y) \neq 0$ then

$$\mathbb{E}(X \mid Y = y) = \frac{\mathbb{E}(X \mathbb{1}_{\{Y=y\}})}{\mathbb{P}(Y = y)}.$$

If $A \in \mathscr{F}$ then $\mathbb{P}(A \mid \mathscr{G})$ and $\mathbb{P}(A \mid Y)$ and $\mathbb{P}(A \mid Y = y)$ are respectively used as notations for $\mathbb{E}(\mathbb{1}_A \mid \mathscr{G})$ and $\mathbb{E}(\mathbb{1}_A \mid Y)$ and $\mathbb{E}(\mathbb{1}_A \mid Y = y)$.

We now recall some important properties of conditional expectation, which hold $\mathbb{P}$-a.s. Let $\mathscr{G}$ and $\mathscr{H}$ be two sub-σ-fields of $\mathscr{F}$.

- Most properties of expectation carry over to conditional expectation: linearity, positivity, Jensen's and Hölder inequalities, monotone and dominated convergence, Fatou's lemma, etc.
- One may "take out what is known": if Y is a $\mathscr{G}$-measurable random variable such that XY is integrable, then

$$\mathbb{E}(XY \mid \mathscr{G}) = Y\mathbb{E}(X \mid \mathscr{G}). \tag{2.9}$$

- The "tower property" holds:

$$\text{if } \mathscr{G} \subset \mathscr{H} \quad \text{then } \mathbb{E}\big(\mathbb{E}(X \mid \mathscr{H}) \mid \mathscr{G}\big) = \mathbb{E}(X \mid \mathscr{G}). \tag{2.10}$$

  In particular $\mathbb{E}(\mathbb{E}(X \mid \mathscr{H})) = \mathbb{E}(X)$ (take $\mathscr{G} = \{\emptyset, \Omega\}$).
- Knowing facts independent of our purpose does not help:

$$\text{if } \mathscr{H} \text{ is independent of } \sigma(X, \mathscr{G}), \quad \text{then } \mathbb{E}\big(X \mid \sigma(\mathscr{G}, \mathscr{H})\big) = \mathbb{E}(X \mid \mathscr{G}), \tag{2.11}$$

  where $\sigma(X, \mathscr{G})$ [resp. $\sigma(\mathscr{G}, \mathscr{H})$] denotes the smallest σ-field containing $\sigma(X)$ and $\mathscr{G}$ [resp. $\mathscr{G}$ and $\mathscr{H}$].

Exercise 2.2 Prove the statements in this subsection using the characteristic property of conditional expectation (2.8).

2.3.2 Martingales and Sub-martingales, Backward Martingales

Definition 2.3 Let $(\mathscr{F}_n, n \in \mathbb{N})$ be a filtration, i.e., a collection of sub-σ-fields of $\mathscr{F}$ such that $\mathscr{F}_n \subset \mathscr{F}_{n+1}$ for all n. Let $(M_n, n \in \mathbb{N})$ be a discrete-time process which is $(\mathscr{F}_n)$-adapted and integrable, i.e., such that $M_n \in L^1(\Omega, \mathscr{F}_n, \mathbb{P})$ for all n.

Then, the process $(M_n, n \in \mathbb{N})$ is called a martingale if

$$M_n = \mathbb{E}(M_{n+1} \,|\, \mathscr{F}_n), \quad \mathbb{P}\text{-a.s.}, \tag{2.12}$$

and a sub-martingale if

$$M_n \leq \mathbb{E}(M_{n+1} \,|\, \mathscr{F}_n), \quad \mathbb{P}\text{-a.s.} \tag{2.13}$$

The related notion of super-martingale is obtained by changing the sense of the inequality, so that $(M_n, n \in \mathbb{N})$ is a super-martingale if and only if $(-M_n, n \in \mathbb{N})$ is a sub-martingale. Note that $(M_n, n \in \mathbb{N})$ is a martingale if and only if it is both a sub-martingale and a super-martingale. Therefore, we concentrate on sub-martingales for most statements, except to make explicit some reinforcements for martingales.

Exercise 2.3 Let $(M_n, n \in \mathbb{N})$ be a sub-martingale [resp. a martingale], and ϕ be an increasing convex function [resp. a convex function] such that $\mathbb{E}|\phi(M_n)| < \infty$ for every n. Prove that $(\phi(M_n), n \in \mathbb{N})$ is again a sub-martingale.

One often needs to consider martingales and sub-martingales at random times, such as the first time at which the process hits a given threshold. Certain specific random times, called *stopping times*, have very useful properties.

Definition 2.4 A random variable T taking values in $\mathbb{N} \cup \{+\infty\}$ is a stopping time for the filtration $(\mathscr{F}_n)$ if $\{T \leq n\} \in \mathscr{F}_n$ for every n in $\mathbb{N}$.

If T is a stopping time, then $\mathscr{F}_T$ is defined as the σ-field constituted of all events A in $\mathscr{F}_\infty := \sigma(\cup_{n\in\mathbb{N}} \mathscr{F}_n)$ such that $A \cap \{T \leq n\} \in \mathscr{F}_n$ for every n in $\mathbb{N}$.

Our next result generalizes the sub-martingale property.

Theorem 2.4 (Optional sampling) *Let $(M_n, n \in \mathbb{N})$ be a sub-martingale. If S and T are two stopping times satisfying $S(\omega) \leq T(\omega) \leq K$, ω-a.s., for some deterministic integer K, then*

$$\mathbb{E}(M_S) \leq \mathbb{E}(M_T \,|\, \mathscr{F}_S), \qquad \mathbb{E}(M_S) \leq \mathbb{E}(M_T).$$

Proof Observe that

$$M_S = \sum_{j=0}^{K} M_j \mathbb{1}_{\{S=j\}} \mathbb{1}_{\{T \geq j\}}.$$

Since $\{T \geq K+1\} = \emptyset$, expressing $M_j \mathbb{1}_{\{T \geq j\}}$ as a telescopic sum yields

$$\begin{aligned} M_S &= \sum_{j=0}^{K} \sum_{k=j}^{K} (M_k \mathbb{1}_{\{T \geq k\}} - M_{k+1} \mathbb{1}_{\{T \geq k+1\}}) \mathbb{1}_{\{S=j\}} \\ &= \sum_{j=0}^{K} \sum_{k=j}^{K} M_k \mathbb{1}_{\{T=k\}} \mathbb{1}_{\{S=j\}} + \sum_{j=0}^{K} \sum_{k=j}^{K} (M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\}}. \end{aligned}$$

Moreover,

$$\sum_{j=0}^{K} \sum_{k=j}^{K} M_k \mathbb{1}_{\{T=k\}} \mathbb{1}_{\{S=j\}} = M_T \sum_{j=0}^{K} \sum_{k=j}^{K} \mathbb{1}_{\{T=k\}} \mathbb{1}_{\{S=j\}} = M_T$$

and hence

$$M_S = M_T + \sum_{j=0}^{K} \sum_{k=j}^{K} (M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\}}.$$

Now, if A is in $\mathscr{F}_S$ then

$$(M_S - M_T) \mathbb{1}_A = \sum_{j=0}^{K} \sum_{k=j}^{K} (M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\} \cap A}.$$

Definition 2.4 implies that $\{T \geq k+1\} = \{T \leq k\}^c \in \mathscr{F}_k$ and

$$\{S = j\} \cap A = \{S \leq j\} \cap A - \{S \leq j-1\} \cap A \in \mathscr{F}_j \subset \mathscr{F}_k$$

so that taking expectations, (2.13) yields

$$\mathbb{E}\big((M_S - M_T) \mathbb{1}_A\big) = \sum_{j=0}^{K} \sum_{k=j}^{K} \mathbb{E}\big((M_k - M_{k+1}) \mathbb{1}_{\{T \geq k+1\}} \mathbb{1}_{\{S=j\} \cap A}\big) \leq 0.$$

Since A is arbitrary, the characteristic property (2.8) allows to conclude the first inequality. The second follows by taking $A = \Omega$. □

Definition 2.5 Let real numbers $a < b$ be fixed. For $n \geq 1$, the number of upcrossings of $[a, b]$ between times 0 and n by a sequence of real numbers $(M_k, k \in \mathbb{N})$ is defined as

$$U_n := \max\{j \geq 0 : T_j \leq n\}$$

where the $(T_j)_{j \geq 0}$ are recursively defined as follows: $T_0 = 0$ and

$$\tau_{j+1} = \inf\{k \geq T_j : M_k \leq a\}, \qquad T_{j+1} = \inf\{k > \tau_{j+1} : M_k \geq b\}.$$

Theorem 2.5 *In the framework of Definition 2.5, if* $(M_k, k \in \mathbb{N})$ *is a sub-martingale, then*

$$\mathbb{E}(U_n) \leq \frac{1}{b-a}\mathbb{E}\big((M_n - a)^+\big).$$

Proof Set $Y_n := (M_n - a)^+$. As $\tau_{n+1} > n$, one has $Y_{\tau_{n+1}\wedge n} = Y_n$, which can be expressed as the telescopic sum

$$\begin{aligned} Y_n &= Y_{\tau_1\wedge n} + \sum_{i=1}^{n}(Y_{\tau_{i+1}\wedge n} - Y_{\tau_i\wedge n}) \\ &= Y_{\tau_1\wedge n} + \sum_{i=1}^{n}(Y_{\tau_{i+1}\wedge n} - Y_{T_i\wedge n}) + \sum_{i=1}^{n}(Y_{T_i\wedge n} - Y_{\tau_i\wedge n}). \end{aligned}$$

Observe that

$$Y_{\tau_1\wedge n} \geq 0, \qquad \sum_{i=1}^{n}(Y_{T_i\wedge n} - Y_{\tau_i\wedge n}) \geq (b-a)U_n,$$

and thus

$$Y_n \geq \sum_{i=1}^{n}(Y_{\tau_{i+1}\wedge n} - Y_{T_i\wedge n}) + (b-a)U_n.$$

The result of Question 2 of Exercise 2.3 applied to the increasing convex function $\phi(x) = (x-a)^+$ yields that $(Y_n, n \in \mathbb{N})$ is a sub-martingale. Taking expectations and applying Theorem 2.4 to the stopping times $T_i \wedge n \leq \tau_{i+1} \wedge n$ then yields that

$$\mathbb{E}(Y_n) \geq (b-a)\mathbb{E}(U_n),$$

from which the result follows. □

The notion of filtration expresses the practical fact that information is increasing with time: the σ-fields satisfy $\mathscr{F}_n \subset \mathscr{F}_{n+1}$. In the proof of Theorem 2.1, we will consider the σ-fields $\sigma(\hat{S}_N, \hat{S}_{N+1}, \ldots)$, which decrease w.r.t. time $N \geq 1$, but however define a filtration when time is run backwards. We thus are led to introduce the following definition.

Definition 2.6 Let $(\mathscr{G}_{-N}, N \geq 1)$ be a family of sub-σ-fields satisfying

$$\mathscr{G}_{-(N+1)} \subset \mathscr{G}_{-N}, \quad N \geq 1. \tag{2.14}$$

A process $(M_{-N}, N \geq 1)$ is a $(\mathscr{G}_{-N})$-backward martingale if the M_{-N} are integrable, $\mathscr{G}_{-N}$-measurable, and

$$M_{-(N+1)} = \mathbb{E}(M_{-N} \,|\, \mathscr{G}_{-(N+1)}), \quad \mathbb{P}\text{-a.s.}$$

Theorem 2.5 can then be easily adapted as follows. We leave the details as an exercise for the interested reader.

Theorem 2.6 *Let* $(\mathscr{G}_{-N}, N \geq 1)$ *be a family of sub-*σ*-fields satisfying* (2.14)*, and* $(M_{-N}, N \geq 1)$ *be a* $(\mathscr{G}_{-N})$*-backward martingale. Let* $a < b$ *be fixed, and for* $N \geq 1$ *let* U_{-N} *denote the number of upcrossings of the interval* $[a, b]$ *between times* 0 *and* $N - 1$ *by the sequence*

$$(\bar{M}_k, k \in \mathbb{N}) = (M_{-N}, \ldots, M_{-1}, M_{-1}, \ldots)$$

(*see Definition* 2.5)*. Then*

$$\mathbb{E}(U_{-N}) \leq \frac{1}{b-a}\mathbb{E}\big((M_{-1} - a)^+\big).$$

2.3.3 Proof of the Strong Law of Large Numbers

Now, Theorem 2.1 will be proved. Recall that $(\xi^{(\ell)}, \ell \geq 1)$ is a sequence of independent identically distributed random variables such that $\mathbb{E}|\xi^{(1)}| < \infty$, and that $\hat{S}_N := \frac{1}{N}\sum_{\ell=1}^{N} \xi^{(\ell)}$ for $N \geq 1$.

Step 1

The key idea consists in considering

$$\mathscr{G}_{-N} := \sigma(\hat{S}_N, \hat{S}_{N+1}, \ldots), \quad N \geq 1,$$

which clearly satisfies (2.14), and observing that $(\hat{S}_N, N \geq 1) = (M_{-N}, N \geq 1)$ for an appropriately defined $(\mathscr{G}_{-N})$-backward martingale $(M_{-N}, N \geq 1)$.

Indeed, from (2.10) it follows that

$$M_{-N} := \mathbb{E}\big(\xi^{(1)} \,|\, \mathscr{G}_{-N}\big) \tag{2.15}$$

defines a $(\mathscr{G}_{-N})$-backward martingale $(M_{-N}, N \geq 1)$. In addition, by symmetry,

$$\forall 1 \leq \ell \leq N, \quad M_{-N} = \mathbb{E}\big(\xi^{(\ell)} \,|\, \mathscr{G}_{-N}\big),$$

and therefore, since $\hat{S}_N$ is $\mathscr{G}_{-N}$ measurable,

$$M_{-N} = \frac{1}{N}\sum_{\ell=1}^{N} \mathbb{E}\big(\xi^{(\ell)} \,|\, \mathscr{G}_{-N}\big) = \mathbb{E}(\hat{S}_N \,|\, \mathscr{G}_{-N}) = \hat{S}_N.$$

Note that in particular $M_{-1} = \xi^{(1)}$.

Step 2

We prove that backward martingales such as $(M_{-N}, N \geq 1)$ converge a.s. to an integrable random variable $M_{-\infty}$.

Fix $a < b$ and use the notation of Theorem 2.6. For every ω in Ω, the increasing sequence $(U_{-N}(\omega), N \geq 1)$ has a limit $U_{-\infty}(\omega)$ in $\mathbb{R} \cup \{\infty\}$. The monotone convergence theorem and Theorem 2.6 yield

$$\mathbb{E}(U_{-\infty}) = \lim_{N\to\infty} \mathbb{E}(U_{-N}) \leq \frac{1}{b-a}\mathbb{E}\big((M_{-1}-a)^+\big) < \infty.$$

Thus the random variable $U_{-\infty}$ is finite a.s., and the sequence $(M_{-N}, N \geq 1)$ almost surely crosses $[a, b]$ a finite number of times, which implies that

$$\mathbb{P}\Big(\liminf_{N\to\infty} M_{-N} < a < b < \limsup_{N\to\infty} M_{-N}\Big) = 0.$$

This being true for all rational $a < b$, necessarily

$$\mathbb{P}\Big(\liminf_{N\to\infty} M_{-N} < \limsup_{N\to\infty} M_{-N}\Big) = 0$$

and thus, in $\mathbb{R} \cup \{-\infty, \infty\}$,

$$\lim_{N\to\infty} M_{-N} = M_{-\infty} := \liminf_{N\to\infty} M_{-N}, \quad \mathbb{P}\text{-a.s.}$$

Using (2.15) and (2.10) yields

$$\mathbb{E}\big(|M_{-N}|\big) = \mathbb{E}\big(\big|\mathbb{E}\big(\xi^{(1)} \,|\, \mathscr{G}_{-N}\big)\big|\big) \leq \mathbb{E}\big(\mathbb{E}\big(\big|\xi^{(1)}\big| \,|\, \mathscr{G}_{-N}\big)\big) = \mathbb{E}\big(\big|\xi^{(1)}\big|\big)$$

and Fatou's lemma yields

$$\mathbb{E}\big(|M_{-\infty}|\big) \leq \liminf_{N\to\infty} \mathbb{E}\big(|M_{-N}|\big) \leq \mathbb{E}\big(\big|\xi^{(1)}\big|\big) < \infty.$$

Step 3

We show that backward martingales such as $(M_{-N}, N \geq 1)$ converge in $L^1(\Omega)$.

This readily follows from a.s. convergence and (2.15) and classic uniform integrability results (see Problem 2.3), but we will prove it using less advanced notions.

For any $N \geq 1$ and $C \in \mathbb{R}_+$, it holds that

$$\begin{aligned}\mathbb{E}\big(|M_{-\infty} - M_{-N}|\big) \leq\ & \mathbb{E}\big(|M_{-\infty} - M_{-N}|\mathbb{1}_{\{|M_{-N}|\leq C\}}\big) \\ & + \mathbb{E}\big(\big(|M_{-\infty}| + |M_{-N}|\big)\mathbb{1}_{\{|M_{-N}|>C\}}\big).\end{aligned}$$

In view of (2.15), (2.9) and (2.10),

$$|M_{-N}|\mathbb{1}_{\{|M_{-N}|>C\}} = \left|\mathbb{E}\left(\xi^{(1)}\,|\,\mathscr{G}_{-N}\right)\mathbb{1}_{\{|M_{-N}|>C\}}\right| = \left|\mathbb{E}\left(\xi^{(1)}\mathbb{1}_{\{|M_{-N}|>C\}}\,|\,\mathscr{G}_{-N}\right)\right|$$
$$\leq \mathbb{E}\left(\left|\xi^{(1)}\right|\mathbb{1}_{\{|M_{-N}|>C\}}\,|\,\mathscr{G}_{-N}\right)$$

and hence

$$\mathbb{E}\left(|M_{-N}|\mathbb{1}_{\{|M_{-N}|>C\}}\right) \leq \mathbb{E}\left(\mathbb{E}\left(\left|\xi^{(1)}\right|\mathbb{1}_{\{|M_{-N}|>C\}}\,|\,\mathscr{G}_{-N}\right)\right) = \mathbb{E}\left(\left|\xi^{(1)}\right|\mathbb{1}_{\{|M_{-N}|>C\}}\right).$$

Thus, for $Y := |M_{-\infty}| + |\xi^{(1)}|$ it holds that

$$\mathbb{E}\left(|M_{-\infty} - M_{-N}|\right) \leq \mathbb{E}\left(|M_{-\infty} - M_{-N}|\mathbb{1}_{\{|M_{-N}|\leq C\}}\right) + \mathbb{E}(Y\mathbb{1}_{\{|M_{-N}|>C\}}). \quad (2.16)$$

Since $M_{-\infty}$ is integrable, by the Dominated Convergence Theorem,

$$\lim_{N\to\infty} \mathbb{E}\left(|M_{-\infty} - M_{-N}|\mathbb{1}_{\{|M_{-N}|\leq C\}}\right) = 0.$$

In addition, for any $B \in \mathbb{R}_+$,

$$\mathbb{E}(Y\mathbb{1}_{\{|M_{-N}|>C\}}) \leq \mathbb{E}(Y\mathbb{1}_{\{Y>B\}}) + B\mathbb{P}\left(|M_{-N}| > C\right)$$

and by dominated convergence, since Y is integrable, for any $\varepsilon > 0$ one can choose B and then C large enough to have

$$\mathbb{E}(Y\mathbb{1}_{\{Y>B\}}) < \varepsilon, \qquad B\mathbb{P}\left(|M_{-\infty}| > C\right) < \varepsilon,$$

and moreover

$$\lim_{N\to\infty} B\mathbb{P}\left(|M_{-N}| > C\right) = B\mathbb{P}\left(|M_{-\infty}| > C\right) < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrarily chosen, (2.16) and the results that follow it yield

$$\lim_{N\to\infty} \mathbb{E}\left(|M_{-\infty} - M_{-N}|\right) = 0.$$

Note that this $L^1(\Omega)$ convergence and $\mathbb{E}(M_{-N}) = \mathbb{E}(\mathbb{E}(\xi^{(1)}\,|\,\mathscr{G}_{-N})) = \mathbb{E}(\xi^{(1)})$, see (2.15) and (2.10), yield that

$$\mathbb{E}(M_{-\infty}) = \lim_{N\to\infty} \mathbb{E}(M_{-N}) = \mathbb{E}\left(\xi^{(1)}\right). \quad (2.17)$$

Step 4

For every $k \geq 1$,

$$M_{-\infty} = \lim_{N\to\infty} \hat{S}_N := \lim_{N\to\infty} \frac{\xi^{(1)} + \cdots + \xi^{(N)}}{N} = \lim_{N\to\infty} \frac{\xi^{(k)} + \cdots + \xi^{(N)}}{N}, \quad \mathbb{P}\text{-a.s.},$$

so that $M_{-\infty}$ is measurable w.r.t. the tail σ-field $\bigcap_{k\geq 1} \sigma(\xi^{(k)}, \xi^{(k+1)}, \ldots)$.

We admit the following classic result, which is an ingredient in most proofs of the Strong Law of Large Numbers (see, e.g., Williams [47] for a proof using martingales).

Theorem 2.7 (Kolmogorov zero-one law) *Let* $(\xi^{(\ell)}, \ell \geq 1)$ *be a sequence of independent random variables, and* $\mathscr{T} := \bigcap_{k\geq 1} \sigma(\xi^{(k)}, \xi^{(k+1)}, \ldots)$ *be its tail* σ*-field. Then, any* $\mathscr{T}$*-measurable random variable is a.s. constant.*

Applying this theorem to $M_{-\infty}$ yields that this random variable is a.s. constant. It must then be a.s. equal to its expectation $\mathbb{E}(\xi^{(1)})$, see (2.17).

This concludes the proof of the SLLN.

2.4 Problems

2.1 (A Poisson Distribution Simulation Method) Recall that the Laplace transform of a non-negative random variable X, or of its law $\mathbb{P}^X$, is given by $\mathbb{E}(\mathrm{e}^{-\theta X}) = \int \mathrm{e}^{-\theta x} \mathbb{P}^X(dx)$ for all $\theta \geq 0$ such that the expectation is finite.

1. Compute the Laplace transform of the distribution function of the exponential probability law with parameter $\lambda > 0$.
2. Let $(X_k, k \geq 1)$ be a family of independent random variables with the same exponential probability distribution with parameter $\lambda > 0$. Set

$$S_N := \sum_{i=1}^{N} X_k.$$

   Compute the Laplace transform of the distribution function of S_N. Deduce that the probability density function of S_N is

$$p_N(x) := \frac{\lambda^N}{(N-1)!} x^{N-1} \mathrm{e}^{-\lambda x} \mathbb{1}_{x\geq 0}.$$

3. Let M be the smallest integer N such that $S_{N+1} > \lambda$. Show that this random variable has a Poisson distribution.
4. Propose a simulation method of the Poisson distribution which requires samples only of the uniform distribution on [0, 1].

2.2 (Lyapunov Exponent of Linear Random Recursive Sequences) Let a and b be two real numbers. Let $1 > h > 0$ be a time discretization step. For any integer p set

$$\bar{X}^h_{p+1}(x) = \left(1 + b\sqrt{h} G_{p+1} + \left(a + \frac{b^2}{2}\right)h\right) \bar{X}^h_p(x),$$

where the G_p are mutually independent and centered Gaussian random variables with unit variance, and where $\bar{X}^h_0(x) = x$ a.s.

1. Check that the functions $|\log(|x|)|\exp(-x^2)$ and $(\log(|x|))^2\exp(-x^2)$ are integrable over $\mathbb{R}$.
   Hint: Use that $(\log(x))^2$ is the derivative of $x(\log(x))^2 - 2x\log(x) + 2x$.
2. Show that

$$\exists \bar{\lambda}^h \in \mathbb{R} \text{ for all } x \in \mathbb{R}^d - \{0\}, \quad \bar{\lambda}^h = \lim_{N \longrightarrow +\infty} \frac{1}{Nh} \log |\bar{X}_N^h(x)|, \quad \text{a.s.}$$

3. Show that, for all x in $\mathbb{R} - \{0\}$,

$$\exists C_h \in \mathbb{R}, \quad \frac{1}{N^2}\mathbb{E}\big[\log |\bar{X}_N^h(x)|\big]^2 < C_h \quad \text{for all } N \in \mathbb{N} - \{0\}.$$

   Problem 2.3 shows that the preceding inequality implies

$$\forall x \in \mathbb{R} - \{0\}, \quad \bar{\lambda}^h = \lim_{N \longrightarrow +\infty} \frac{1}{Nh}\mathbb{E}\log\big|\bar{X}_N^h(x)\big|.$$

   Deduce from this result that

$$\bar{\lambda}^h = \frac{1}{h}\mathbb{E}\log\left|1 + b\sqrt{h}G_1 + \left(a + \frac{b^2}{2}\right)h\right|$$

   for all h small enough and all N.
4. Let

$$Y := b\sqrt{h}G_1 + \left(a + \frac{b^2}{2}\right)h.$$

   Prove that

$$\mathbb{E}\log|1+Y| = \mathbb{E}\left[Y - \frac{Y^2}{2} + \frac{Y^3}{3}\right] + \mathbb{E}\left[\mathbb{1}_{|Y|<1}\left(\log(1+Y) - Y + \frac{Y^2}{2} - \frac{Y^3}{3}\right)\right]$$
$$+ \mathbb{E}\left[\mathbb{1}_{|Y|\geq 1}\left(\log|1+Y| - Y + \frac{Y^2}{2} - \frac{Y^3}{3}\right)\right].$$

   Deduce that $\bar{\lambda}^h = a + \mathscr{O}(h)$.

2.3 (Uniformly Integrable Random Variables ($\star$)) Consider a sequence (X_n) of random variables with finite expectations which are uniformly integrable, that is,

$$\lim_{C\to\infty} \sup_{n\geq 0} \mathbb{E}\big[|X_n|\mathbb{1}_{|X_n|\geq C}\big] = 0. \tag{2.18}$$

1. Show that (2.18) is equivalent to the conjunction of the two following ones:

$$\sup_n \big(\mathbb{E}|X_n|\big) < \infty, \tag{2.19}$$

$$\forall \varepsilon > 0,\ \exists \delta(\varepsilon),\ \forall A \in \mathscr{F}, \quad \mathbb{P}(A) \leq \delta(\varepsilon) \Rightarrow \sup_n \mathbb{E}\big(|X_n|\mathbb{1}_A\big) < \varepsilon. \tag{2.20}$$

2. Let (X_n) be a uniformly integrable sequence, and let X be an integrable random variable. Show that the sequence $(|X_n - X|)$ is uniformly integrable.
   Hint: Observe that $|X_n - X| \leq |X_n| + |X|$, and apply the preceding question.
3. Now assume in addition that (X_n) has an almost sure limit X. Show that

$$\mathbb{E}|X| < \infty.$$

   Hint: Start with observing that

$$\mathbb{E}\big(|X_n|\big) = \mathbb{E}\big(|X_n|\mathbb{1}_{|X_n|\leq C}\big) + \mathbb{E}\big(|X_n|\mathbb{1}_{|X_n|\geq C}\big); \tag{2.21}$$

   then use Fatou's lemma and Lebesgue's Dominated Convergence Theorem.
4. Deduce from Questions 1 and 2 that

$$\lim_{n\to\infty} \mathbb{E}|X_n - X| = 0. \tag{2.22}$$

   Hint: Observe that, for all positive ε,

$$\mathbb{E}|X_n - X| \leq \mathbb{E}\big[|X_n - X|\mathbb{1}_{|X_n - X|\geq\varepsilon}\big] + \varepsilon.$$

5. Write an alternative proof to Step 3 in Sect. 2.3.3 for

$$\lim_{N\to\infty} \mathbb{E}\big(|M_{-\infty} - M_{-N}|\big) = 0.$$

   Hint: Show that $\sup_{n\geq 0} \mathbb{E}(X_n)^2 < \infty$ implies uniform integrability.

Chapter 3
Non-asymptotic Error Estimates for Monte Carlo Methods

Abstract In order to effectively implement Monte Carlo methods, the random approximation errors must be controlled. For this purpose, theoretical results are provided for the estimation of the number of simulations necessary to obtain a desired accuracy with a prescribed confidence interval. Therefore *absolute*, i.e., *non-asymptotic*, versions of the Central Limit Theorem (CLT) are developed: Berry–Esseen's and Bikelis' theorems, as well as concentration inequalities obtained from logarithmic Sobolev inequalities. The difficult subject of variance reduction techniques for Monte Carlo methods arises naturally in this context, and is discussed at the end of this chapter.

Reference books for this chapter are Petrov [43] and Shiryayev [44], as well as Williams [47] up to the CLT result.

3.1 Convergence in Law and Characteristic Functions

In the preceding chapter we have proved the almost sure convergence of the Monte Carlo methods. Now our goal is to describe the approximation error. Almost sure bounds cannot be obtained; however, one can prove precise estimates on the probability distribution of this error.

Before treating this question, we start with a few reminders on the probability laws of random variables.

Definition 3.1 Let $X := (X_1, \ldots, X_d)$ be an $\mathbb{R}^d$-valued random variable. The *law* or *distribution* of X is the probability measure $\mathbb{P}^X$ on the measurable space $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ defined by

$$\forall A \in \mathscr{B}\big(\mathbb{R}^d\big), \quad \mathbb{P}^X(A) := \mathbb{P}\big(\omega;\, X(\omega) \in A\big).$$

In set-theoretic terms $\mathbb{P}^X = \mathbb{P} \circ X^{-1}$.

The following result states a classical property of laws of random variables.

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_3, © Springer-Verlag Berlin Heidelberg 2013

Theorem 3.1 *For any measurable non-negative function f from $\mathbb{R}^d$ to $\mathbb{R}$,*

$$\mathbb{E}f(X) := \int_\Omega f\big(X(\omega)\big)\mathbb{P}(d\omega) = \int_{\mathbb{R}^d} f(x)\mathbb{P}^X(dx).$$

In practice, it will often be seen that the exact Monte Carlo method cannot be implemented, due to the fact that a probabilistic representation $\gamma = \mathbb{E}f(X)$, see (2.3), can only be found for a random variable X which is difficult (or impossible) to simulate. It is then necessary to find some other easily simulable random variable $\bar{X}$ which satisfies $\mathbb{E}f(\bar{X}) \simeq \mathbb{E}f(X)$. For such approximations, we introduce the fundamental notion of convergence in law of random variables, i.e., of weak convergence of their laws. It will be seen that the limitations of continuity and of boundedness on the test functions can be sometimes removed.

Definition 3.2 Let $(\bar{X}^n, n \in \mathbb{N})$ be a sequence of $\mathbb{R}^d$-valued random variables. It is said to converge in law to a random variable X, or to its law $\mathbb{P}^X$, if the laws $(\mathbb{P}^{\bar{X}^n}, n \in \mathbb{N})$ converge weakly to $\mathbb{P}^X$, i.e., if for every bounded continuous function f from $\mathbb{R}^d$ to $\mathbb{R}$,

$$\lim_{n\to\infty} \int f(x)\mathbb{P}^{\bar{X}^n}(dx) = \int f(x)\mathbb{P}^X(dx).$$

In terms of the random variables themselves, this can be written as

$$\lim_{n\to\infty} \mathbb{E}f\big(\bar{X}^n\big) = \mathbb{E}f(X).$$

The approximation of $\gamma = \mathbb{E}f(X)$ by a Monte Carlo method often depends on two parameters:

- the index n of the term $\bar{X}^n$ actually chosen as an approximation, in the sequence $(\bar{X}^n)$ which converges in law to X,
- the number N of the simulated independent and identically distributed copies $\bar{X}^{n(1)}, \ldots, \bar{X}^{n(N)}$ of $\bar{X}^n$.

More precisely, the global approximation error can be decomposed as follows:

$$\gamma - \frac{1}{N}\sum_{\ell=1}^{N} f\big(\bar{X}^{n(\ell)}\big) = \underbrace{\mathbb{E}f(X) - \mathbb{E}f\big(\bar{X}^n\big)}_{\varepsilon_d(n)} + \underbrace{\mathbb{E}f\big(\bar{X}^n\big) - \frac{1}{N}\sum_{\ell=1}^{N} f\big(\bar{X}^{n(\ell)}\big)}_{\varepsilon_s(n,N)}, \quad (3.1)$$

- where the term $\varepsilon_d(n)$ measures the approximation error on the exact probability distribution, due to convergence in law,
- and the term $\varepsilon_s(n, N)$ measures the statistical error resulting from the application of the Strong Law of Large Numbers.

In the sequel to this monograph we study the behavior w.r.t. n and N of these two errors. The behavior w.r.t. N is described by the Central Limit Theorem and its variants, whose proofs rely on the notion of characteristic function.

Definition 3.3 The characteristic function of the $\mathbb{R}^d$-valued random variable X, or of its probability law $\mathbb{P}^X$, is given by

$$\phi_X : t \in \mathbb{R}^d \mapsto \phi_X(t) := \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}t\cdot X}\mathbb{P}^X(dx) = \mathbb{E}\mathrm{e}^{\mathrm{i}t\cdot X}.$$

Indeed, Definition 3.2 is inconvenient because the space of continuous bounded functions is huge. Fourier analysis of abstract measures allows one to limit oneself to the parametrized family of functions $x \mapsto \mathrm{e}^{\mathrm{i}t\cdot x}$, yielding the following important result (see Jacod and Protter [24, Thm 19.1 p. 167], e.g.).

Theorem 3.2 *A sequence $(\bar{X}^n)$ of $\mathbb{R}^d$-valued random variables converges in law to a random variable X if and only if*

$$\lim_{n\to\infty} \phi_{\bar{X}^n}(t) = \phi_X(t), \quad \forall t \in \mathbb{R}^d.$$

Notably, two random variables with the same characteristic function have the same law.

The next classic result will be useful.

Proposition 3.1 *Let X be a real-valued random variable. Suppose that there exists an integer K in $\mathbb{N} - \{0\}$ such that $\mathbb{E}|X|^K < \infty$. Then the function ϕ_X is everywhere differentiable up to the order K and*

$$\phi_X(t) = \sum_{k=1}^{K} \frac{(\mathrm{i}t)^k}{k!}\mathbb{E}\big(X^k\big) + \frac{t^K}{K!}\varepsilon_K(t), \quad \forall t \in \mathbb{R},$$

where $\sup_{t\in\mathbb{R}} |\varepsilon_K(t)| \le 3\mathbb{E}|X|^K$ and $\lim_{t\to 0}\varepsilon_K(t) = 0$.

Proof For any x in $\mathbb{R}$, Taylor's formula with integral remainder term yields

$$\begin{aligned}\mathrm{e}^{\mathrm{i}x} &= \cos(x) + \mathrm{i}\sin(x)\\ &= \sum_{k=0}^{K-1} \frac{(\mathrm{i}x)^k}{k!} + x^K \int_0^1 \frac{(1-\theta)^{K-1}}{(K-1)!}\big(\cos^{(K)}(\theta x) + \mathrm{i}\sin^{(K)}(\theta x)\big)\, d\theta\end{aligned}$$

and therefore the formula is valid for

$$\varepsilon_K(t) = K\mathbb{E}\left(X^K \int_0^1 (1-\theta)^{K-1}\big(\cos^{(K)}\big(\theta t X(\omega)\big) + \mathrm{i}\sin^{(K)}\big(\theta t X(\omega)\big) - \mathrm{i}^K\big)\, d\theta\right)$$

which satisfies $\sup_t |\varepsilon_K(t)| \le 3\mathbb{E}|X|^K$. One then proves that $\lim_{t\to 0}\varepsilon_K(t) = 0$ by successively examining the cases where K is odd or even and by using Lebesgue's dominated convergence theorem. □

A last result on weak convergence of real random variables involves the cumulative distribution function (c.d.f.), see Definition 2.2. It will be useful for the future notion of confidence intervals, or to consider histogram convergence. See Jacod and Protter [24, Thm 18.4 p. 153], e.g., also for the fact that a c.d.f. is right-continuous and has left limits. Let F_Y denote the c.d.f. of a random variable Y.

Theorem 3.3 *A sequence $(\bar{X}^n)$ of real-valued random variables converges in law to a random variable X if and only if*

$$\lim_{n\to\infty} F_{\bar{X}^n}(x) = F_X(x), \quad \forall x \in \mathbb{R} \text{ s.t. } F_X(x) = F_X(x-).$$

In particular, if F_X is continuous then there is convergence for all x.

3.2 Central Limit Theorem

The Strong Law of Large Numbers ensures the convergence of Monte Carlo methods, but does not allow one to describe the accuracy of the method or the number N of simulations in terms of the desired accuracy. We already emphasized that the approximation error is random. The Central Limit Theorem makes precise the limit behavior, when N tends to infinity, of the normalized error probability distribution.

Theorem 3.4 (Central Limit Theorem) *Let $(\xi^{(\ell)}, \ell \geq 1)$ be a sequence of real-valued independent and identically distributed random variables. Assume they are square integrable and set $\sigma^2 := \mathrm{var}(\xi^{(1)})$. Consider the random variables*

$$Y_N := \frac{\sqrt{N}}{\sigma}\left(\frac{1}{N}\sum_{\ell=1}^{N}\xi^{(\ell)} - \mathbb{E}\big(\xi^{(1)}\big)\right), \quad N \geq 1.$$

The sequence (Y_N) converges in law to the Gaussian law $\mathcal{N}(0,1)$.

Proof We wish to apply Theorem 3.2, and prove the convergence of the characteristic function of Y_N to $t \mapsto \exp(-\frac{t^2}{2})$, which classically is the characteristic function of the $\mathcal{N}(0,1)$ law. Set

$$\tilde{\xi}^{(\ell)} := \xi^{(\ell)} - \mathbb{E}\big(\xi^{(\ell)}\big)$$

and observe that, since the random variables $\xi^{(\ell)}$ are independent and identically distributed,

$$\phi_{Y_N}(t) = \mathbb{E}\left(\exp\left(\frac{\mathrm{i}t}{\sigma\sqrt{N}}\sum_{\ell=1}^{N}\tilde{\xi}^{(\ell)}\right)\right) = \left(\phi_{\tilde{\xi}^{(1)}}\left(\frac{t}{\sigma\sqrt{N}}\right)\right)^N. \tag{3.2}$$

To conclude, it then remains to use Proposition 3.1 with $K = 2$. □

In practice, one needs to estimate the probability that the approximation error of the Monte Carlo method for $\mathbb{E}(\xi^{(1)})$ is larger than a desired accuracy $\varepsilon > 0$. This error is given by

$$\varepsilon_s(N) := \mathbb{E}\big(\xi^{(1)}\big) - \frac{1}{N}\sum_{\ell=1}^{N} \xi^{(\ell)} = -\frac{\sigma}{\sqrt{N}} Y_N,$$

and thus the probability that $\mathbb{E}(\xi^{(1)})$ does not belong to the interval I, centered in $\frac{1}{N}\sum_{\ell=1}^{N} \xi^{(\ell)}$ with radius ε, should be evaluated. If this probability can be bounded above by δ, then I is called a confidence interval with risk level δ, or confidence level $1 - \delta$.

Our objective is to identify the dependency of each one of the parameters ε, δ, and N in terms of the others in such circumstances.

3.2.1 Asymptotic Confidence Intervals

The Central Limit Theorem is classically invoked to provide asymptotic confidence intervals, according to the following coarse analysis:

- For any "large enough" N, the probability law of $\varepsilon_s(N)$ is "close to" the Gaussian law with zero mean and variance equal to $\frac{\sigma}{\sqrt{N}}$, and in this "sense", if Y_∞ denotes an arbitrary $\mathcal{N}(0, 1)$ Gaussian variable,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{\ell=1}^{N} \xi^{(\ell)} - \mathbb{E}\big(\xi^{(1)}\big)\right| \geq \varepsilon\right) \simeq \mathbb{P}\left(|Y_\infty| \geq \frac{\sqrt{N}}{\sigma}\varepsilon\right).$$

- Classical statistical tables (coded in computer programs) provide precise evaluations of the r.h.s. as a function of $\frac{\sqrt{N}}{\sigma}\varepsilon$.
- This evaluation is used for an approximation of the actual risk level δ.

This procedure yields an approximate relation between the accuracy ε, number N of samples, risk level δ, and variance σ^2. For instance, this yields approximate values for the number N of samples necessary to achieve a prescribed risk level δ in terms of the rest; as an application, if $\delta = 5\ \% = 0.05$ then $\mathbb{P}(|Y_\infty| \geq 1.96) \simeq 0.05$ and thus

$$N \geq 1.96^2 \frac{\sigma^2}{\varepsilon^2}.$$

Remark 3.1 Note that:

- this bound is proportional to the variance $\sigma^2 = \mathbb{E}((\xi^{(1)})^2) - \mathbb{E}(\xi^{(1)})^2$ and inversely proportional to the square of the desired accuracy,
- the proportionality ratio depends only on the level of risk δ,

- in practice the variance σ^2 is unknown, but should be estimated by a Monte Carlo method simultaneously with $\mathbb{E}(\xi^{(1)})$, or bounded a priori.

Obviously, in order to validate the preceding coarse analysis, it is necessary to make precise the meaning of "large enough N" and "close to a Gaussian law".

The next sections concern extensions of the Central Limit Theorem. These allow to estimate the distance between the c.d.f. of Y_N and the c.d.f. of the Gaussian $\mathcal{N}(0, 1)$ law, and thus to obtain accurate non-asymptotic confidence intervals, valid for all N (see also Theorem 3.3). Motivations may be:

- Simulations or statistical measurements which are difficult or costly (in terms of time, money, availability, ...). Then N needs to be chosen as small as possible.
- Estimation of probabilities of rare events, say of order 10^{-8} for air control purposes. This requires to take ε and δ extremely small. It is crucial to be sure that the corresponding high N is compatible with realistic simulation program running times.
- Applications (including the preceding example) requiring to choose δ extremely small, but for which the tails of the law of $\varepsilon_s(N)$ may not be well approximated by the tails of the $\mathcal{N}(0, 1)$ law, even when N is quite large. Section 3.4 will show that this is the case when the law of $\xi^{(1)}$ has fat tails, for example when $\mathbb{E}(|\xi^{(1)}|^{2+\gamma})$ is very large for every $0 < \gamma < 1$.

In such cases, the blind use of asymptotic results valid only for N tending to infinity can lead to gross undervaluation of risk levels.

3.3 Berry–Esseen's Theorem

The Central Limit Theorem gives a precise asymptotic rate of convergence for the Strong Law of Large Numbers. The Berry–Esseen theorem provides non-asymptotic error estimates.

Its proof uses the following result, which is admitted. The very technical proof can be found in the textbooks by Petrov [43] or Shiryayev [44].

Theorem 3.5 (Esseen) *Let X and Y be two real-valued random variables. Let $T > 0$ be an arbitrary number. For all $b > \frac{1}{2\pi}$, there exists a positive number $C(b)$ (depending on b only) such that*

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(X \le x) - \mathbb{P}(Y \le x)\big| \le b \int_{-T}^{T} \left|\frac{\phi_X(t) - \phi_Y(t)}{t}\right| dt + \frac{b}{T} \sup_x \int_{|y|\le C(b)} \big|\mathbb{P}(Y \le x + y) - \mathbb{P}(Y \le x)\big| dy. \tag{3.3}$$

We now state and prove the Berry–Esseen theorem.

Theorem 3.6 (Berry–Esseen) *Let $(\xi^{(\ell)}, \ell \geq 1)$ be a sequence of real-valued independent and identically distributed random variables. Assume that $\mathbb{E}\xi^{(1)} = 0$ and $\mathbb{E}|\xi^{(1)}|^3 < +\infty$. Let $\sigma^2 := \mathrm{var}(\xi^{(1)})$ and Φ denote the c.d.f. of the $\mathcal{N}(0,1)$ Gaussian law. Then*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\xi^{(1)}+\cdots+\xi^{(N)}}{\sigma\sqrt{N}} \leq x\right) - \Phi(x)\right| \leq \frac{C\mathbb{E}|\xi^{(1)}|^3}{\sigma^3\sqrt{N}}.$$

It has been proved that $0.398 \leq C \leq 0.8$.

Remark 3.2 The hypothesis that $\mathbb{E}\xi^{(1)} = 0$ is not a restriction: the result can be applied to $\tilde{\xi}^{(\ell)} := \xi^{(\ell)} - \mathbb{E}\xi^{(1)}$, see the following section.

Proof Observe that $\Phi'(x) \leq \frac{1}{\sqrt{2\pi}}$ for all x. Set

$$\Phi_N(x) := \mathbb{P}\left(\frac{\xi^{(1)}+\cdots+\xi^{(N)}}{\sigma\sqrt{N}} \leq x\right), \qquad L_N := \frac{\mathbb{E}(|\xi^{(1)}|^3)}{\sigma^3\sqrt{N}}, \tag{3.4}$$

and apply Esseen's theorem 3.5 with $b = \frac{1}{\pi}$ and $T = \frac{1}{4L_N}$. Denote by $\phi_N(t)$ the characteristic function of $\frac{1}{\sigma\sqrt{N}}\sum_{\ell=1}^N \xi^{(\ell)}$. There exists $C > 0$ such that

$$\sup_x\left|\Phi_N(x) - \Phi(x)\right| \leq \frac{1}{\pi}\int_{|t|\leq\frac{1}{4L_N}}\left|\frac{\phi_n(t) - \mathrm{e}^{-t^2/2}}{t}\right|dt + CL_N.$$

To conclude it then remains to apply the following lemma. □

Lemma 3.1 *Under the assumptions of Theorem 3.6, let $\phi_N(t)$ be the characteristic function of $\frac{1}{\sigma\sqrt{N}}\sum_{\ell=1}^N \xi^{(\ell)}$. Then, for L_N defined in (3.4),*

$$\forall |t| \leq \frac{1}{4L_N}, \quad \left|\phi_N(t) - \mathrm{e}^{-t^2/2}\right| \leq 16L_N|t|^3\mathrm{e}^{-t^2/3}. \tag{3.5}$$

Proof We start by dealing with the case where, in addition to being less that $\frac{1}{4L_N}$, the number t satisfies $|t| \geq \frac{1}{2}L_N^{-1/3}$. Then $16L_N|t|^3 \geq 2$ and

$$\left|\phi_N(t) - \mathrm{e}^{-t^2/2}\right| \leq \left|\phi_N(t)\right| + \mathrm{e}^{-t^2/2}.$$

Therefore it suffices to show that $|\phi_N(t)| \leq \mathrm{e}^{-t^2/3}$. Let $\phi_\xi(t)$ be the common characteristic function of the $\xi^{(\ell)}$'s. The function $|\phi_\xi(t)|^2$ is the characteristic function of $\xi^{(\ell)} - \xi^{(j)}$ for any $j \neq \ell$ and, in addition,

$$\mathbb{E}\left(\left|\xi^{(\ell)} - \xi^{(j)}\right|^3\right) \leq 8\mathbb{E}\left(\left|\xi^{(1)}\right|^3\right).$$

Applying Proposition 3.1, we deduce

$$\left|\phi_\xi(t)\right|^2 \le 1 - \sigma^2 t^2 + \frac{4}{3}|t|^3 \mathbb{E}\left(\left|\xi^{(1)}\right|^3\right) \le \exp\left(-\sigma^2 t^2 + \frac{4}{3}|t|^3 \mathbb{E}\left(\left|\xi^{(1)}\right|^3\right)\right).$$

As $|t| \le \frac{1}{4L_N}$ by hypothesis, we have obtained $|\phi_N(t)| \le \mathrm{e}^{-t^2/3}$.

We now consider the case where $|t| \le \frac{1}{4L_N}$ and $|t| \le \frac{1}{2} L_N^{-1/3}$ and observe that

$$\frac{\sigma}{\sigma\sqrt{N}}|t| \le \frac{(\mathbb{E}(|\xi^{(1)}|^3))^{1/3}}{\sigma\sqrt{N}} \le L_N^{1/3}|t|.$$

Since

$$\begin{aligned}\mathrm{e}^{\mathrm{i}x} &= \cos(x) + \mathrm{i}\sin(x) \\ &= 1 + \sum_{k=1}^{2} \frac{(\mathrm{i}x)^k}{k!} + x^3 \int_0^1 \frac{(1-\theta)^2}{2}\left(\cos^{(3)}(\theta x) + \mathrm{i}\sin^{(3)}(\theta x)\right) d\theta,\end{aligned}$$

we also have

$$\phi_\xi\left(\frac{t}{\sigma\sqrt{N}}\right) = 1 - \frac{\sigma^2 t^2}{2\sigma^2 N} + \eta(N)$$

with

$$\left|\eta(N)\right| \le \frac{\mathbb{E}(|\xi^{(1)}|^3)}{6\sigma^3 N\sqrt{N}}|t|^3.$$

Therefore

$$\begin{aligned}\left(\frac{\sigma^2 t^2}{2\sigma^2 N} + \eta(N)\right)^2 &\le 2\left(\frac{\sigma^2 t^2}{2\sigma^2 N}\right)^2 + 2\eta(N)^2 \\ &\le \left(\frac{1}{4} + \frac{1}{126}\right)\frac{\mathbb{E}(|\xi^{(1)}|^3)}{\sigma^3 N\sqrt{N}}|t|^3 \le \frac{\mathbb{E}(|\xi^{(1)}|^3)}{3\sigma^3 N\sqrt{N}}|t|^3.\end{aligned}$$

One also has

$$\frac{\sigma^2 t^2}{2\sigma^2 N} + \left|\eta(N)\right| \le \frac{1}{6},$$

and thus

$$\log\left(\phi_\xi\left(\frac{t}{\sigma\sqrt{N}}\right)\right) \le -\frac{t^2}{2N} + \left|\eta(N)\right|,$$

from which

$$\log\left(\phi_N(t)\right) \le -\frac{t^2}{2} + \frac{L_N|t|^3}{2}.$$

From $L_N|t|^3 \le \frac{1}{8}$ it follows that $\exp(\frac{L_N|t|^3}{2}) \le 2$. Consequently we have obtained

$$\left|\phi_N(t) - \mathrm{e}^{-t^2/2}\right| \le \frac{L_N}{2}|t|^3 \exp\left(-\frac{t^2}{2} + \frac{L_N}{2}|t|^3\right) \le L_N|t|^3 \mathrm{e}^{-t^2/2},$$

which implies (3.5). □

3.4 Bikelis' Theorem

The bounds in Bikelis' theorem are more precise than those in Berry–Esseen's theorem, since now the error estimate is not uniform w.r.t. x. We refer to Petrov [43] for a proof and some extensions.

Theorem 3.7 (Bikelis) *Let $(\xi^{(\ell)}, \ell \ge 1)$ be a sequence of real-valued independent random variables (which are not necessarily identically distributed). Assume that $\mathbb{E}\xi^{(\ell)} = 0$ for all ℓ, and that there exists $0 < \gamma \le 1$ such that $\mathbb{E}|\xi^{(\ell)}|^{2+\gamma} < +\infty$ for all ℓ. Let*

$$B_N := \sum_{\ell=1}^{N} \mathrm{var}\big(\xi^{(\ell)}\big), \qquad \Phi_N(x) := \mathbb{P}\left(\frac{\xi^{(1)} + \cdots + \xi^{(N)}}{\sqrt{B_N}} \le x\right),$$

and Φ denote the c.d.f. of the $\mathcal{N}(0,1)$ Gaussian law. Then, there exists a universal constant A such that, for all x in $\mathbb{R}$,

$$\left|\Phi_N(x) - \Phi(x)\right| \le \frac{A}{B_N^{1+\gamma/2}(1+|x|)^{2+\gamma}} \sum_{\ell=1}^{N} \mathbb{E}\left|\xi^{(\ell)}\right|^{2+\gamma}. \tag{3.6}$$

It has been proved that $\frac{1}{\sqrt{2\pi}} \le A < 1$.

3.4.1 Absolute Confidence Intervals

Let $(\xi^{(\ell)}, \ell \ge 1)$ be a sequence of integrable real-valued independent and identically distributed random variables, such that $\mathbb{E}\xi^{(1)}$ is not necessarily null. In order to achieve an accuracy of order ε with a confidence level $1 - \delta$ for the Monte Carlo estimation of $\mathbb{E}\xi^{(1)}$, one has to choose a number N of simulations large enough that

$$\mathbb{P}\left(\left|\mathbb{E}\big(\xi^{(1)}\big) - \frac{1}{N}\sum_{\ell=1}^{N} \xi^{(\ell)}\right| \ge \varepsilon\right) \le \delta. \tag{3.7}$$

Let us show how the Bikelis theorem can be applied to establish absolute, or non-asymptotic, confidence intervals, under the further assumption $\mathbb{E}|\xi^{(1)}|^3 < \infty$.

Let $m := \mathbb{E}\xi^{(1)}$ and $\sigma^2 := \mathrm{var}(\xi^{(1)})$ and

$$\tilde{\xi}^{(\ell)} := \xi^{(\ell)} - m, \qquad Z_N := \sum_{\ell=1}^{N} \tilde{\xi}^{(\ell)} = \sum_{\ell=1}^{N} \xi^{(\ell)} - Nm.$$

Note that the $\tilde{\xi}^{(\ell)}$ are i.i.d., and that $\mathbb{E}\tilde{\xi}^{(1)} = 0$ and $\mathbb{E}|\tilde{\xi}^{(1)}|^3 < +\infty$ and

$$\mathbb{P}\left(\left|\mathbb{E}\big(\xi^{(1)}\big) - \frac{1}{N}\sum_{\ell=1}^{N} \xi^{(\ell)}\right| \geq \varepsilon\right) = \mathbb{P}\big(|Z_N| \geq \varepsilon N\big).$$

Setting

$$\Phi_N(x) := \mathbb{P}\left(\frac{\tilde{\xi}^{(1)} + \cdots + \tilde{\xi}^{(N)}}{\sigma\sqrt{N}} \leq x\right),$$

an application of Bikelis' theorem yields that

$$\begin{aligned}
\mathbb{P}\big(|Z_N| \geq \varepsilon N\big) &= 2\left(1 - \Phi_N\left(\frac{\varepsilon\sqrt{N}}{\sigma}\right)\right) \\
&\leq 2\left(1 - \Phi\left(\frac{\varepsilon\sqrt{N}}{\sigma}\right)\right) + \frac{2A\mathbb{E}|\tilde{\xi}^{(1)}|^3}{\sqrt{N}(\sigma + \varepsilon\sqrt{N})^3} \\
&\leq \frac{2}{\sqrt{2\pi}} \int_{\frac{\varepsilon\sqrt{N}}{\sigma}}^{+\infty} \mathrm{e}^{-y^2/2}\, dy + \frac{2A\mathbb{E}|\tilde{\xi}^{(1)}|^3}{\sqrt{N}(\sigma + \varepsilon\sqrt{N})^3}
\end{aligned}$$

so that, in order to achieve the bound (3.7), it suffices to choose N such that

$$\frac{2}{\sqrt{2\pi}} \int_{\frac{\varepsilon\sqrt{N}}{\sigma}}^{+\infty} \mathrm{e}^{-y^2/2}\, dy + \frac{2A\mathbb{E}|\tilde{\xi}^{(1)}|^3}{\sqrt{N}(\sigma + \varepsilon\sqrt{N})^3} \leq \delta.$$

For actual computation of numerical values, statistical tables can again be used, as well as bounds such as Komatsu's inequality:[1] for all $x > 0$, it holds that

$$\frac{2}{x + \sqrt{x^2 + 4}} \mathrm{e}^{-x^2/2} \leq \int_x^{+\infty} \mathrm{e}^{-y^2/2}\, dy \leq \frac{2}{x + \sqrt{x^2 + 2}} \mathrm{e}^{-x^2/2} \leq \sqrt{2}\mathrm{e}^{-x^2/2}.$$

In practice, σ and $\mathbb{E}|\tilde{\xi}^{(1)}|^3$ are unknown (otherwise, approximating $\mathbb{E}\tilde{\xi}^{(1)}$ would not be an issue). In order to effectively estimate the order of magnitude of a suitable N, two strategies are often followed:

- to estimate or bound σ and $\mathbb{E}|\tilde{\xi}^{(1)}|^3$ starting from the law of $\tilde{\xi}^{(1)}$,
- or, to proceed numerically in an iterative way, as in the following algorithm.

[1] Itô, K., McKean Jr., H.P.: Diffusion Processes and Their Sample Paths. Die Grundlehren der Mathematischen Wissenschaften, vol. 125, 2nd edn. Springer, Berlin (1974).

Algorithm (Estimating number of samples) Start from a first set of N_1 samples simulated from the law of $\xi^{(1)}$;

- compute an approximation for m and σ^2 and $\mathbb{E}|\xi^{(1)}|^3$,
- use these for Bikelis' theorem, which provides a value N_2 for the number of simulations that should be needed,
- – If $N_2 \leq N_1$ then all is well, terminate the simulation,
  – else, simulate a new set of $N_2 - N_1$ samples, and proceed from the start using the N_2 samples instead of N_1,

(repeat until successful).

Even though N_1 may be an educated guess, it may grossly underestimate the needed number of simulations.

In contrast, even though the estimation of m and σ^2 and $\mathbb{E}|\xi^{(1)}|^3$ from N_1 samples may be poor, Bikelis' theorem should provide an N_2 quite close to the mark. Using the N_2 samples yields a (statistically) better approximation for m and σ^2 and $\mathbb{E}|\xi^{(1)}|^3$, but seldom would these estimations differ enough that Bikelis' theorem would indicate a value N_3 for the desired number of simulations substantially larger than N_2, so that an additional $N_3 - N_2$ samples would be required to straighten things out. Thus, this algorithm should stop after two or at most three iterations.

Exercise 3.1 During an election between two candidates, the proportion p of voters who will vote for Candidate number 1 is estimated using an opinion poll. Suppose that the answers are independent identically distributed random variables $\xi^{(1)}, \ldots, \xi^{(N)}$ satisfying $\mathbb{P}(\xi^{(1)} = 1) = p = 1 - \mathbb{P}(\xi^{(1)} = 0)$.

1. Explain why p should be estimated by $\hat{p}_N := \frac{1}{N} \sum_{\ell=1}^{N} \xi^{(\ell)}$.
2. Fix $N = 1024$. Use various methods to estimate the probability that Candidate number 2 wins the election when $\hat{p}_N = 51.5\ \%$ or $\hat{p}_N = 53\ \%$.

3.5 Concentration Inequalities

The goal of this section is to estimate the probability that the Monte Carlo method error exceeds a given threshold, in the case where the simulated probability distribution satisfies an abstract property: the logarithmic Sobolev inequality. Although this property may be difficult to prove, it often is satisfied in practice.

The concentration inequalities in Theorem 3.9 have important numerical consequences: if the logarithmic Sobolev inequality is satisfied, then the probability that the Monte Carlo method errors take unacceptable values is small, in quantifiable terms.

Most of the material in this section comes from the collective pedagogical monograph Ané et al. [2] on the subject.

3.5.1 Logarithmic Sobolev Inequalities

The logarithmic Sobolev inequality relies on the notion of entropy which is classical in information theory. For a probability measure on $\mathbb{R}^d$ and a positive or integrable function g on $\mathbb{R}^d$, the notation $\mathbb{E}_\mu(g) := \int g\, d\mu$ is used.

Definition 3.4 Let μ be a probability measure on $\mathbb{R}^d$. For any positive measurable function f on $\mathbb{R}^d$, the entropy of f w.r.t. μ is defined by

$$\mathrm{Ent}_\mu(f) := \mathbb{E}_\mu\big(f\log(f)\big) - \mathbb{E}_\mu(f)\log\big(\mathbb{E}_\mu(f)\big).$$

Jensen's inequality implies that $\mathrm{Ent}_\mu(f)$ is non-negative, and vanishes if and only if the function f is μ-a.s. constant. Moreover, $\mathrm{Ent}_\mu(f)$ is finite if and only if the function $f(\log(f))^+$ is μ-integrable.

Exercise 3.2 Check that

$$\mathrm{Ent}_\mu(f) = \sup\big\{\mathbb{E}_\mu(fg);\ g \text{ s.t. } \mathbb{E}_\mu\big(\mathrm{e}^g\big) = 1\big\}. \tag{3.8}$$

Hint: Use the inequality $uv \le u\log(u) - u + \mathrm{e}^v$ for $u \ge 0$ and $v \in \mathbb{R}$.

The following proposition shows that a product measure enjoys the so-called *tensorization* property.

Proposition 3.2 *Let $\mu_1, \dots, \mu_N$ be probability measures on $\mathbb{R}$, and $\mu^N := \mu_1 \otimes \cdots \otimes \mu_N$ be their product measure on $\mathbb{R}^N$. Then, for any positive function f on $\mathbb{R}^N$,*

$$\mathrm{Ent}_{\mu^N}(f) \le \sum_{i=1}^N \mathbb{E}_{\mu^N}\big(\mathrm{Ent}_{\mu_i}(f)\big), \tag{3.9}$$

where $\mathrm{Ent}_{\mu_i}(f)$ is computed by integrating only w.r.t. the x_i variable, leaving the other fixed.

Proof Let g be a real-valued function defined on $\mathbb{R}^N$ such that $\mathbb{E}_{\mu^N}(\mathrm{e}^g) = 1$, and set

$$g_1 := g - \log \int \mathrm{e}^g \mu_1(dx_1),$$

$$g_i := \log \frac{\int \mathrm{e}^g \mu_1(dx_1)\cdots\mu_{i-1}(dx_{i-1})}{\int \mathrm{e}^g \mu_1(dx_1)\cdots\mu_i(dx_i)}, \quad i \ge 2.$$

As $\mathbb{E}_{\mu_i}\mathrm{e}^{g_i} = 1$, the formula (3.8) implies

$$\sum_{i=1}^N \mathbb{E}_{\mu_i}(fg_i) \le \sum_{i=1}^N \mathrm{Ent}_{\mu_i}(f).$$

It then suffices to observe that $g = \sum g_i$ and thus

$$\mathbb{E}_{\mu^N}(fg) = \sum_{i=1}^{N} \mathbb{E}_{\mu^N}\big(\mathbb{E}_{\mu_i}(fg_i)\big),$$

and then to again use the formula (3.8). □

We now are in a position to introduce the logarithmic Sobolev inequalities.

Definition 3.5 A measure ν on $\mathbb{R}^d$ satisfies the logarithmic Sobolev inequality (LSI) with constant $C > 0$ for the family of functions $\mathscr{A}$ if, for every function f in $\mathscr{A}$,

$$\mathrm{Ent}_\nu\big(f^2\big) \le C\mathbb{E}_\nu\big(|\nabla f|^2\big). \tag{3.10}$$

Many measures satisfy an LSI. The next theorem shows two elementary examples.

Theorem 3.8 *Let $\mathscr{A}$ be the family of the functions of class $\mathscr{C}^2(\mathbb{R}^d, \mathbb{R})$ with compact support.*

(a) *The Bernoulli law on $\{0, 1\}$ with parameter p satisfies the LSI for the family $\mathscr{A}$, with constant $C = 2$ if $p = \frac{1}{2}$ and $C = \frac{\log(1-p)-\log(p)}{1-2p}$ otherwise.*

(b) *The standard Gaussian law on $\mathbb{R}^d$ satisfies the LSI with constant $C = 2$ for any dimension d.*

Proof We observe that if the result holds true for the standard Gaussian law on $\mathbb{R}$, then the tensorization inequality (3.9) implies that the result also holds true for the standard Gaussian law on $\mathbb{R}^d$.

We now prove, using the Central Limit Theorem, that if the result holds for the Bernoulli laws, then it holds for the standard Gaussian law on $\mathbb{R}$. We consider the Bernoulli law of parameter $\frac{1}{2}$, with the notation

$$\nu_i := \frac{1}{2}(\delta_0 + \delta_1), \quad i \ge 1.$$

Let f be a function of class $\mathscr{C}^2(\mathbb{R}^d, \mathbb{R})$ with compact support. Apply the tensorization inequality (3.9) to the function $f \circ \Phi_N$, where

$$\Phi_N(x_1, \ldots, x_N) := \frac{2}{\sqrt{N}}\left(\sum_{\ell=1}^{N} x_i - \frac{N}{2}\right), \quad x_i \in \{0, 1\}.$$

Now observe that, for any function F defined on $\{0, 1\}^N$,

$$\mathrm{Ent}_{\nu_i}\big(F^2\big) = \frac{1}{2}\big(F_i(1)^2 - F_i(0)^2\big)\log\big[F_i(1) - \log\big(F_i(0)\big)\big],$$

where $F_i(a) := F_i(x_1, \dots, x_{i-1}, a, x_{i+1}, \dots, x_N)$, and that

$$\big|(f \circ \Phi_N)_i(1) - (f \circ \Phi_N)_i(0)\big| \le \frac{2}{\sqrt{N}} \big|f' \circ \Phi_{N,i}(x)\big| + \frac{C}{\sqrt{N}}$$

where

$$\Phi_{N,i}(x) := \frac{2}{\sqrt{N}} \left(x_1 + \cdots + x_{i-1} + x_{i+1} + \cdots + x_N - \frac{N}{2} \right),$$

since the second derivative of f is bounded.

In order to obtain an LSI for the standard Gaussian law, it then suffices to observe that the probability distribution $\nu^N \circ (\Phi_{N,i})^{-1}$ does not depend on i, and to apply the Central Limit Theorem to this probability distribution by letting N tend to infinity.

Let us now return to the proof for the Bernoulli law. By symmetry one can suppose $0 < p \le \frac{1}{2}$. One then use the formula (3.8) to compute the quantity

$$\sup \left\{ \frac{\mathbb{E}_\nu(f^2 g)}{\frac{1}{2}|f'(1)|^2 + \frac{1}{2}|f'(0)|^2};\ g \text{ s.t. } \mathrm{e}^{g(1)} + \mathrm{e}^{g(0)} = 2 \right\},$$

which is a simple but tedious calculation. □

3.5.2 Concentration Inequalities, Absolute Confidence Intervals

Theorem 3.9 below, which is interesting in its own, will be used to analyze the Monte Carlo method error. To prove it, we will twice need the following elementary result:

Lemma 3.2 *Let μ be a probability measure on $\mathbb{R}^d$, and let F be a measurable bounded function. Suppose that there exists $\gamma > 0$ such that*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}_\mu\big(\mathrm{e}^{\lambda(F - \mathbb{E}_\mu(F))}\big) \le \exp\left(\gamma \frac{\lambda^2}{2}\right). \tag{3.11}$$

Then

$$\forall r > 0, \quad \mu\big(\big|F - \mathbb{E}_\mu(F)\big| \ge r\big) \le 2 \exp\left(-\frac{r^2}{2\gamma}\right). \tag{3.12}$$

Proof Using the hypothesis and the Markov inequality,

$$\mu\big(F - \mathbb{E}\mu(F) \ge r\big) = \mu\big(\mathrm{e}^{\lambda(F - \mathbb{E}\mu(F))} \ge \mathrm{e}^{\lambda r}\big) \le \mathrm{e}^{-\lambda r} \mathbb{E}_\mu \mathrm{e}^{\lambda(F - \mathbb{E}\mu(F))} \le \mathrm{e}^{\gamma \frac{\lambda^2}{2} - \lambda r}.$$

Optimize the last term in the r.h.s. by choosing $\lambda = \frac{r}{\gamma}$, from which

$$\forall r > 0, \quad \mu\big(F - \mathbb{E}_\mu(F) \ge r\big) \le \mathrm{e}^{-\frac{r^2}{2\gamma}}.$$

It is then easy to obtain (3.12): one observes that

$$\mu\big(|F - \mathbb{E}_\mu(F)| \geq r\big) \leq \mu\big(F - \mathbb{E}_\mu(F) \geq r\big) + \mu\big(-F + \mathbb{E}_\mu(F) \geq r\big)$$

and applies the preceding inequality to the functions F and $\tilde{F} := -F$. □

Definition 3.6 In the rest of this subsection, we will write that a real function F on $\mathbb{R}^d$ has Lipschitz constant L if

$$\big|F(x_1, \ldots, x_d) - F(y_1, \ldots, y_d)\big| \leq L\sqrt{|x_1 - y_1|^2 + \cdots + |x_d - y_d|^2}.$$

Theorem 3.9 *Let μ be a measure on $\mathbb{R}^d$ satisfying the logarithmic Sobolev inequality with constant C for the family $\mathscr{A}$ of the bounded functions of class $\mathscr{C}^\infty(\mathbb{R}^d, \mathbb{R})$ with bounded partial derivatives of all orders. Then, for any function F with Lipschitz constant smaller than 1 and $r > 0$,*

$$\mu\big(F - \mathbb{E}\mu(F) \geq r\big) \leq \exp\left(-\frac{r^2}{C}\right), \tag{3.13}$$

$$\mu\big(|F - \mathbb{E}\mu(F)| \geq r\big) \leq 2\exp\left(-\frac{r^2}{C}\right). \tag{3.14}$$

Remark 3.3 For a function F with an arbitrary Lipschitz constant L, similar inequalities are obtained by considering the function F/L.

Proof Let us first prove (3.13). We limit ourselves to functions F with Lipschitz constant smaller than 1 which are bounded and of class $\mathscr{C}^\infty(\mathbb{R}^d)$. The general case will follow by applying Lemma 3.3 below to the approximating sequence

$$F_n := g_{1/n} * \max\big(-n, \min(n, F)\big),$$

where $g_{1/n}$ is the $\mathscr{N}(0, \frac{1}{n})$ Gaussian density and $*$ is the convolution operator.

Our objective is to apply Lemma 3.2. Set $H(\lambda) := \mathbb{E}_\mu(\mathrm{e}^{\lambda F})$. For $\lambda > 0$ we apply the LSI with constant C to $f := \mathrm{e}^{\frac{1}{2}\lambda F}$: using that $|\nabla F| \leq 1$ it follows that

$$\lambda H'(\lambda) - H(\lambda)\log H(\lambda) \leq C\frac{\lambda^2}{4}\int \mathrm{e}^{\lambda F}|\nabla F|^2 d\mu \leq C\frac{\lambda^2}{4}H(\lambda).$$

Therefore

$$\frac{H'(\lambda)}{\lambda H(\lambda)} - \frac{\log H(\lambda)}{\lambda^2} \leq \frac{C}{4}.$$

Set $\tilde{H}(\lambda) := \frac{1}{\lambda}\log H(\lambda)$. The preceding inequality shows that $\tilde{H}'(\lambda) \leq \frac{C}{4}$. In addition, in a neighborhood of 0, one has $\tilde{H}(\lambda) \simeq \frac{H(\lambda)-1}{\lambda} \simeq \mathbb{E}_\mu(F)$. Thus

$$\tilde{H}(\lambda) - \tilde{H}(0) = \int_0^\lambda \tilde{H}'(\alpha)\, d\alpha \leq \frac{C}{4}\lambda,$$

which is equivalent to

$$H(\lambda) \le \exp\left(\frac{C}{4}\lambda^2 + \lambda \mathbb{E}_\mu(F)\right).$$

It then suffices to apply Lemma 3.2 to get (3.13).

One finally obtains (3.14) by applying (3.13) to F and $-F$. □

The following lemma has been used in the preceding proof.

Lemma 3.3 *Let (F_n) be a sequence of μ-integrable and Lipschitz functions with Lipschitz constant smaller than 1 and tending to F μ-a.s. Assume that there exists $\gamma > 0$ such that*

$$\forall n \in \mathbb{N},\ \forall \lambda > 0, \quad \mathbb{E}_\mu\left(e^{\lambda(F_n - \mathbb{E}_\mu(F_n))}\right) \le e^{\gamma\lambda^2/2}.$$

Then F is integrable and

$$\forall \lambda > 0, \quad \mathbb{E}_\mu\left(e^{\lambda(F - \mathbb{E}_\mu(F))}\right) \le e^{\gamma\lambda^2/2}.$$

Proof Suppose first that the sequence (F_n) converges to F in $L^1(\Omega, \mathscr{F}, \mathbb{P})$. In view of the bound which is assumed, $\mathbb{E}_\mu(F_n)$ tends to $\mathbb{E}_\mu(F)$. Applying Fatou's lemma yields

$$\mathbb{E}_\mu\left(e^{\lambda F}\right) \le \liminf_n \mathbb{E}_\mu\left(e^{\lambda F_n}\right) \le \liminf_n e^{\lambda \mathbb{E}_\mu(F_n) + \gamma\lambda^2/2} \le e^{\lambda \mathbb{E}_\mu(F) + \gamma\lambda^2/2},$$

which provides the desired result.

To prove that (F_n), which converges to F a.s., also converges in $L^1(\Omega, \mathscr{F}, \mathbb{P})$, it classically suffices to show that $M := \sup_n \mathbb{E}_\mu((F_n)^2) < \infty$, for the following reason: since, for all $C \ge 0$,

$$\mathbb{E}\big(|F_n - F|\big) \le \mathbb{E}\big(|F_n - F|\mathbb{1}_{\{|F_n| \le C\}}\big) + \mathbb{E}\big(|F_n|\mathbb{1}_{\{|F_n| > C\}}\big) + \mathbb{E}\big(|F|\mathbb{1}_{\{|F_n| > C\}}\big),$$

by choosing $C > \frac{M}{\varepsilon}$ where $\varepsilon > 0$ is arbitrarily fixed,

$$\mathbb{E}\big(|F_n|\mathbb{1}_{\{|F_n| > C\}}\big) \le \frac{1}{C}\mathbb{E}\big(|F_n|^2\mathbb{1}_{\{|F_n| > C\}}\big) \le \frac{1}{C}\mathbb{E}\big(|F_n|^2\big) < \varepsilon,$$

and thus, by dominated convergence,

$$\limsup_n \mathbb{E}\big(|F_n - F|\big) < \varepsilon.$$

This observation leads us to seek bounds which are uniform w.r.t. n for $A_n := \mathbb{E}_\mu((F_n - \mathbb{E}_\mu(F_n))^2)$ and $B_n := \mathbb{E}_\mu((F_n)^2)$.

Lemma 3.2 and our hypotheses imply

$$\mu\big(|F_n - \mathbb{E}_\mu(F_n)| \ge r\big) \le 2e^{-\frac{r^2}{2\gamma}}.$$

Classically, the Fubini theorem yields that, for any probability measure ν,

$$\int_0^\infty x^2 \nu(dx) = 2\int_0^\infty x\,\nu(x,+\infty)\,dx. \tag{3.15}$$

Choose $\nu := \mu \circ |F_n - \mathbb{E}_\mu(F_n)|^{-1}$ and use Theorem 3.1. It then follows that

$$\mathbb{E}_\mu\big(\big(F_n - \mathbb{E}_\mu(F_n)\big)^2\big) = 2\int_0^\infty r\mu\big(\big|F_n - \mathbb{E}_\mu(F_n)\big| \geq r\big)\,dr \leq 4\int_0^\infty r\mathrm{e}^{-\frac{r^2}{2\gamma}}\,dr.$$

Now consider B_n. Let m be large enough to have $\mu(|F| \leq m) \geq \frac{3}{4}$, and let n_0 be such that $\mu(|F_n - F| \geq 1) \leq \frac{1}{4}$ for all $n \geq n_0$. As

$$\begin{aligned}\{F_n \geq m+1\} &\subset \big\{|F_n - F| + |F| \geq m+1\big\}\\ &\subset \big\{|F| \geq m\big\} \cup \big\{|F| \leq m \text{ and } |F_n - F| \geq 1\big\}\\ &\subset \big\{|F| \geq m\big\} \cup \big\{|F_n - F| \geq 1\big\},\end{aligned}$$

we have, for all $n \geq n_0$,

$$\mathbb{P}(F_n \geq m+1) \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Let r_1 be such that $\mathrm{e}^{-r_1^2/2\gamma} < \frac{1}{2}$. We also have

$$\mathbb{P}\big(\{F_n \geq m+1\} \cup \big\{|F_n - \mathbb{E}_\mu(F_n)| \leq r_1\big\}\big) > 1,$$

so that the events $\{F_n \geq m+1\}$ and $\{|F_n - \mathbb{E}_\mu(F_n)| \leq r_1\}$ have a non-empty intersection. For all ω in this intersection, we have $\mathbb{E}_\mu(F_n) \leq r_1 + F_n(\omega)$ and $F_n(\omega) \leq m+1$. It results that

$$\mathbb{E}_\mu(F_n) \leq r_1 + m + 1.$$

Now gather the preceding bounds from above for A_n and B_n: we have obtained

$$M := \sup_n \mathbb{E}_\mu\big((F_n)^2\big) \leq 4\int_0^\infty r\mathrm{e}^{-\frac{r^2}{2\gamma}}\,dr + r_1 + m + 1 < \infty,$$

which ends the proof. □

We are now in a position to state the following result. It allows to bound the probability that the Monte Carlo method approximation error takes large values, and thus to construct confidence intervals.

Theorem 3.10 (Concentration inequalities for the SLLN) *Let $(X^{(\ell)}, \ell \geq 1)$ be a sequence of $\mathbb{R}^d$-valued independent random variables with law μ. Assume that μ*

satisfies a logarithmic Sobolev inequality with constant C for the family $\mathscr{A}$ of functions with Lipschitz constant smaller than 1. *Then, for any f in $\mathscr{A}$ and $r > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{\ell=1}^{N} f\left(X^{(\ell)}\right) - \mathbb{E}_{\mu}(f)\right| \geq r\right) \leq 2\exp\left(-\frac{Nr^2}{C}\right). \tag{3.16}$$

Proof The tensorization property in Proposition 3.2 implies that $(X^{(\ell)}, 1 \leq \ell \leq N)$ satisfies a logarithmic Sobolev inequality with constant C. In addition, in view of Definition 3.6 and the following consequence of Cauchy–Schwarz's inequality:

$$\sum_{j=1}^{N} |a_j| = \sum_{j=1}^{N} |a_j| \times 1 \leq \sqrt{N}\sqrt{\sum_{j=1}^{N} |a_j|^2},$$

the function $F_N(x) := \frac{1}{N}\sum_{\ell=1}^{N} f(x_\ell)$ is Lipschitz on $\mathbb{R}^{dN}$ with Lipschitz constant $\frac{1}{\sqrt{N}}$. To conclude, it suffices to apply (3.14) to the function $\sqrt{N}F_N$. □

3.6 Elementary Variance Reduction Techniques

The Central Limit Theorem and its non-asymptotic versions show that, the number N of simulations being fixed, a Monte Carlo approximation is all the more effective when the variance σ^2 of $\xi^{(1)}$ is small. Equivalently, the desired accuracy being fixed, the number of simulations which is necessary to achieve this accuracy is all the weaker when the σ^2 is small. The quantity $\sigma^2 = \mathrm{var}(\xi^{(1)})$ will be called the "variance of the simulation".

Some variance reduction techniques are going to be described using some examples. We refer for instance to the book of Asmussen and Glynn [4] and the references therein for more insight into this delicate matter.

3.6.1 Control Variate

Consider a random variable X of which the probability distribution can be simulated, and a measurable function f such that $f(X)$ has finite expectation. The goal is to compute $\mathbb{E}f(X)$.

The control variate variance reduction method consists in using the Monte Carlo method to approximate $\mathbb{E}((f-g)(X))$, where the function g satisfies:

- the value of $\mathbb{E}g(X)$ is explicitly known;
- the variance of $f(X) - g(X)$ is substantially smaller than the variance of $f(X)$.

In practice, this double constraint is not easy to satisfy.

However the computation of European options provides a concrete application example. Consider the price at time 0 of a European call:

$$C_0 := \mathrm{e}^{-rT}\mathbb{E}^*\big((S_T - K)_+\big),$$

where T is the maturity of the option and K its exercise price, and where the expectation is computed under the risk neutral probability. The variance of the simulation is

$$\mathrm{e}^{-2rT}\mathbb{E}^*\big((S_T - K)_+\big)^2 - C_0^2.$$

Denote by P_0 the price of a European put option. We have

$$P_0 := \mathrm{e}^{-rT}\mathbb{E}^*\big((K - S_T)_+\big),$$

and thus, in view of the Put–Call parity,

$$C_0 = P_0 + S_0 - K\mathrm{e}^{-rT}.$$

Consequently the deterministic quantity $S_0 - K\mathrm{e}^{-rT}$ can be chosen as a control variate. The variance of the simulation for the Monte Carlo method which approximates P_0 is

$$\mathrm{e}^{-2rT}\mathbb{E}^*\big((K - S_T)_+\big)^2 - C_0^2$$

and is bounded from above by $\mathrm{e}^{-2rT}K^2 - C_0^2$. Therefore, when

$$K^2 \ll \mathbb{E}^*\big((S_T - K)_+\big)^2,$$

it is efficient to change the computation of a call price into the computation of a put price.

3.6.2 Importance Sampling

In this subsection, we start by presenting the importance sampling variance reduction technique in the following elementary situation. As above we wish to compute $\mathbb{E}f(X)$, but now we assume in addition that X is real-valued and that its law $\mathbb{P}^X$ has a density p_X.

For any real-valued random variable Y whose probability distribution has a strictly positive density p_Y it holds that

$$\mathbb{E}f(X) = \int f(x)p_X(x)\,dx = \int \frac{f(x)}{p_Y(x)}p_Y(x)p_X(x)\,dx = \mathbb{E}\left[\frac{f(Y)}{p_Y(Y)}p_X(Y)\right] = \mathbb{E}Z,$$

where the random variable

$$Z := \frac{f(Y)}{p_Y(Y)} p_X(Y)$$

has variance

$$\mathrm{var}(Z) = \int \frac{f^2(y)}{p_Y(y)} \big(p_X(y)\big)^2 \, dy - \big(\mathbb{E} f(X)\big)^2.$$

It may happen that one can suitably choose p_Y in such a way that $\mathrm{var}(Z)$ is much smaller than $\mathrm{var}(X)$, and such that Y can be easily simulated.

In such a case, it is worth computing $\mathbb{E}Z$ rather than $\mathbb{E}f(X)$ by the Monte Carlo method since, at fixed desired accuracy, the computation of $\mathbb{E}Z$ requires less simulations than the computation of $\mathbb{E}f(X)$.

In practice, it appears difficult to fulfil the preceding requirements. The following examples allow one to understand this limitation of the method.

Let $I := \int_0^1 \exp(x)\, dx$. Suppose that one has the curious objective to approximate I by a Monte Carlo method. The standard method leads one to write $I = \mathbb{E}f(X)$, where X is a uniformly distributed on $[0, 1]$ random variable; an easy calculation shows that $\mathrm{var}(f(U)) = 0.24$.

Now choose $p_Y(x) := \frac{2}{3}(1+x)\mathbb{1}_{[0,1]}(x)$ and apply the importance sampling variance reduction technique. One then has $\mathrm{var}(Z) = 0.027$.

One can improve the variance reduction by looking for a $p_Y(\cdot)$ of the (non-natural) type

$$p_Y(x) = \frac{1+\beta x}{1+\frac{1}{2}\beta} \mathbb{1}_{[0,1]}(x).$$

An easy calculation shows that the optimal choice of β is $\beta = 1.81$ for which $\mathrm{var}(Z) = 0.004$.

Of course, such choices are impossible when the density p_X and the function f are algebraically complex or are not explicitly known, which is the typical situation where Monte Carlo simulations are useful.

In particular, in Chap. 8 we will see the difficulties raised by the importance sampling variance reduction technique when applied to the computation of expectation of functionals of stochastic processes, and examples of numerical methodologies which, in favorable circumstances, allow one to overcome these difficulties.

Remark 3.4 The stratification variance reduction technique is a particular importance sampling technique for integrals of the type $I := \int_D f(x)\, dx$, where D is a domain in $\mathbb{R}^d$. The integration domain D is first partitioned into appropriate sub-domains. Then I is expressed as the sum of integrals on each sub-domain, and each of these new integrals is approximated by the Monte Carlo method, the respective numbers of simulations being adjusted to the variation of the function f on each sub-domain (few simulations for a sub-domain where f varies little, and many where f varies a lot).

Bermuda European Options and Gaussian Changes of Measure

We now consider the situation of a Bermuda European option: its risk neutral price has the form

$$V_0 = \mathrm{e}^{-rT_d}\mathbb{E}^* f(S_{T_1}, \ldots, S_{T_d})$$

where the dates T_i are deterministic, f is the pay-off function of the option, and (S_t) is the underlying asset price process. Here, the expectation is computed under the risk neutral probability $\mathbb{P}^*$.

The Monte Carlo method to approximate V_0 consists in simulating N independent simulations of the probability distribution under $\mathbb{P}^*$ of the random variables $S_{T_1}, \ldots, S_{T_d}$, and in computing

$$\bar{V}_0 := \mathrm{e}^{-rT_d} \sum_{\ell=1}^{N} f\big(S_{T_1}^{(\ell)}, \ldots, S_{T_d}^{(\ell)}\big).$$

The variance of the simulation is the variance under $\mathbb{P}^*$ of $f(S_{T_1}, \ldots, S_{T_d})$.

We will see in Chap. 8 that, by using Girsanov's theorem 8.1, one can, at least theoretically, construct an importance sampling variance reduction technique when (S_t) solves a stochastic differential equation, and that it sometimes is possible to optimize the variance reduction.

We describe this mechanism in the very particular case where (S_t) follows the log-normal Black and Scholes model. (This case presents no numerical interest but allows us to introduce a mathematically complex issue in a simplified framework.)

In this model, the asset price process under $\mathbb{P}^*$ is

$$S_t = S_0 \exp\left(r - \frac{\sigma^2}{2}t + \sigma W_t\right),$$

where (W_t) is a $\mathbb{P}^*$ standard Brownian motion (the definition is given in Chap. 7; here we only need to know that, for all $0 < s < t$, $W_t - W_s$ has a Gaussian distribution with zero mean and variance $t - s$ and is independent of W_s). Set

$$F(x_1, \ldots, x_d) = \mathrm{e}^{-rT_d} f\left(S_0 \exp\left(r - \frac{\sigma^2}{2}T_1 + \sigma x_1\right), \ldots, S_0 \exp\left(r - \frac{\sigma^2}{2}T_d + \sigma x_d\right)\right).$$

We can then rewrite the price V_0 as follows:

$$V_0 = \mathbb{E}^* F(W_{T_1}, \ldots, W_{T_d}).$$

Let θ be a real number and consider the new probability measure

$$\frac{d\mathbb{Q}}{d\mathbb{P}^*} = \exp\left(\theta W_{T_d} - \frac{\theta^2 T_d}{2}\right).$$

Changes of variables in integrals w.r.t. Gaussian densities (which avoids using Girsanov's theorem 8.1 in this simple setting) show that

$$\begin{aligned} V_0 &= \mathbb{E}^{\mathbb{Q}}\left[F(\tilde{W}_{T_1} + \theta T_1, \dots, \tilde{W}_{T_d} + \theta T_d) \exp\left(-\theta \tilde{W}_{T_d} - \frac{\theta^2 T_d}{2} \right) \right] \\ &= \mathbb{E}^*\left[F(W_{T_1} + \theta T_1, \dots, W_{T_d} + \theta T_d) \exp\left(-\theta W_{T_d} - \frac{\theta^2 T_d}{2} \right) \right] \\ &= \mathbb{E}^* F(W_{T_1}, \dots, W_{T_d}). \end{aligned} \tag{3.17}$$

One thus can approximate V_0 by the following Monte Carlo method:

$$\tilde{V}_0 := \mathrm{e}^{-rT_d} \frac{1}{N} \sum_{\ell=1}^{N} F\big(W_{T_1}^{(\ell)} + \theta T_1, \dots, W_{T_d}^{(\ell)} + \theta T_d\big) \exp\left(-\theta W_{T_d}^{(\ell)} - \frac{\theta^2 T_d}{2} \right).$$

The smaller variance of the simulation is obtained by minimizing the quantity

$$H(\theta) := \mathbb{E}^*\big[\big(F\big(W_{T_1} + \theta T_1, \dots, W_{T_d} + \theta T_d\big) \big)^2 \exp\big(-2\theta W_{T_d} - \theta^2 T_d\big) \big].$$

Proposition 3.3 *Suppose that the pay-off function f has an at most polynomial growth at infinity. Suppose also that there exists a ball $B(0, R)$ with radius R in $\mathbb{R}^d$ such that*

$$\int_{B(0,R)} \big(f(x_1, \dots, x_d) \big)^2 \, dx_1 \cdots dx_d > 0. \tag{3.18}$$

Then the function H is strictly convex and there exists a unique real number θ^* such that

$$H(\theta^*) = \min_{\theta \in \mathbb{R}} H(\theta).$$

Proof We start by proving that H is everywhere twice differentiable and by computing explicitly its second derivative. To this end, we observe that

$$\begin{aligned} H(\theta) &= \mathbb{E}^*\big[\big(F(W_{T_1} + \theta T_1, \dots, W_{T_d} + \theta T_d) \big)^2 \exp\big(-2\theta(W_{T_d} + \theta T_d) + \theta^2 T_d\big) \big] \\ &= \mathbb{E}^*\Bigg[\big(F(W_{T_1} + \theta T_1, \dots, W_{T_d} + \theta T_d) \big)^2 \\ &\qquad \times \exp\left(-\theta W_{T_d} - \frac{\theta^2 T_d}{2} - \theta(W_{T_d} + \theta_{T_d}) + \frac{\theta^2 T_d}{2} \right) \Bigg] \end{aligned}$$

so that (3.17) yields

$$H(\theta) = \mathbb{E}^*\left[\big(F(W_{T_1}, \dots, W_{T_d}) \big)^2 \exp\left(-\theta W_{T_d} + \frac{\theta^2 T_d}{2} \right) \right]. \tag{3.19}$$

As the pay-off function f has an at most polynomial growth at infinity, there exists $C > 0$ such that $|F(x_1, \dots, x_d)| \le C e^{C|x|_1 + \cdots C|x|_d}$ for all x, and one can differentiate under the integral sign in the r.h.s. of (3.19); it follows that

$$\frac{d}{d\theta} H(\theta) = \mathbb{E}^* \left[\big(F(W_{T_1}, \dots, W_{T_d})\big)^2 (\theta - W_{T_d}) \exp\left(-\theta W_{T_d} + \frac{\theta^2 T_d}{2} \right) \right].$$

Similarly,

$$\frac{d^2}{d\theta^2} H(\theta) = \mathbb{E}^* \left[\big(F(W_{T_1}, \dots, W_{T_d})\big)^2 \big(1 + (\theta - W_{T_d})^2\big) \exp\left(-\theta W_{T_d} + \frac{\theta^2 T_d}{2} \right) \right].$$

Therefore, denoting by g_d the standard Gaussian density on $\mathbb{R}^d$,

$$\frac{d^2}{d\theta^2} H(\theta) \ge \int_{B(0,R)} \exp\left(-|\theta| R + \theta^2 \frac{T_d}{2} \right) \times \big(F(\sqrt{T_1} x_1, \dots, \sqrt{T_d} x_d)\big)^2 g_d(x_1, \dots, x_d)\, dx_1 \cdots dx_d.$$

As the hypothesis (3.18) implies that

$$\int_{B(0,R)} \big(F(x_1, \dots, x_d)\big)^2 \, dx_1 \cdots dx_d > 0, \tag{3.20}$$

we are in a position to deduce that the second derivative of H is everywhere strictly positive.

To obtain the existence and uniqueness of the optimal value θ^* it therefore suffices to show that $H(\theta)$ tends to infinity when $|\theta|$ tends to infinity. We have:

$$H(\theta) \ge \exp\left(-|\theta| R + \theta^2 \frac{T_d}{2} \right) \times \int_{B(0,R)} \big(F(\sqrt{T_1} x_1, \dots, \sqrt{T_d} x_d)\big)^2 g_d(x_1, \dots, x_d)\, dx_1 \cdots dx_d.$$

We again use (3.20) to conclude. □

We have just shown that to reduce the variance of the Monte Carlo method for Bermuda European options prices, one can optimize the importance sampling technique by perturbing the original Brownian motion by the deterministic function $t \mapsto \theta t$. It however remains to compute the optimal parameter θ^*.

In Chap. 9 we will see that θ^* can be approximated by means of an optimization deterministic or stochastic algorithm before running the Monte Carlo procedure to approximate V_0. We will even develop an adaptive Monte Carlo method, due to B. Arouna [3], which approximates both θ^* and V_0 by using one single set of simulations.

Remark 3.5 In theory one can easily construct many other importance sampling techniques to approximate V_0. For instance, one may use other parametrized Girsanov transformations than the perturbation of (W_t) by a linear function of time, and optimize w.r.t. the new parameters. However, even for the log-normal model and a fortiori for less elementary models, it is not obvious to do so in a way which substantially decreases the variance of the simulation without substantially increasing the complexity of the simulations, and for which optimal values of the parameters can be approximated with a reasonable numerical cost. We refer to Chap. 8 for more on this topic.

3.7 Problems

3.1 (Convergence Rate in Donsker's Theorem) Bikelis' theorem allows to make precise the convergence rate in Donsker's invariance principle for appropriately normalized centered random walks, which converge to a standard Brownian motion.

Let $(X_n, n \geq 1)$ be a sequence of independent identically distributed random variables, with zero mean and standard deviation σ and satisfying $\mathbb{E}(|X_1|^3) < +\infty$. For $T > 0$ let

$$W_T^{(N)} = \frac{\sqrt{T}}{\sigma\sqrt{N}}(X_1 + \cdots + X_N).$$

Let f be a differentiable function with bounded derivative.

1. Prove that

$$\begin{aligned}\left|\int_{-a}^{a} f(x)\mathbb{P}\big(W_T^{(N)} \in dx\big) - \int_{-a}^{a} f(x)\mathbb{P}(W_T \in dx)\right| \\ \leq \int_{-a}^{a} \big|\mathbb{P}\big(W_T^{(N)} \leq x\big) - \mathbb{P}(W_T \leq x)\big|\big|f'(x)\big|\,dx \\ + \big|f(a)\big|\big|\mathbb{P}\big(W_T^{(N)} \leq a\big) - \mathbb{P}(W_T \leq a)\big| \\ + \big|f(-a)\big|\big|\mathbb{P}\big(W_T^{(N)} \leq -a\big) - \mathbb{P}(W_T \leq -a)\big|.\end{aligned}$$

2. Deduce that

$$\big|\mathbb{E}f\big(W_T^{(N)}\big) - \mathbb{E}f(W_T)\big| \leq \big\|f'\big\|_\infty \frac{C'\mathbb{E}|X_1|^3}{\sigma^3\sqrt{N}}.$$

3.2 (Pointwise Approximation of a Density) Let X be a real-valued random variable with a bounded continuous density p^X which is not explicitly known.

1. Let G_ε be a family of Gaussian random variables with zero mean and standard deviation ε. Show that this family converges in law to 0.

2. Let x_0 be an arbitrary real number. By considering $\mathbb{E}[p^X(x_0 - \varepsilon G)]$, construct a Monte Carlo method to approximate $p^X(x_0)$. Estimate the variance of the simulation.
3. Now suppose that the density p^X is everywhere strictly positive and differentiable. Identify classes of functions f such that

$$\mathbb{E}f'(X) = -\mathbb{E}\left(f(X)\frac{p'(X)}{p(X)}\right).$$

4. In which cases does this equality allow one to develop an efficient variance reduction technique for the preceding Monte Carlo approximation of $\mathbb{E}f'(X)$? Treat the particular case of the pointwise approximation of a standard Gaussian density.

3.3 (Approximation of Quantiles) Let X be a real-valued random variable with an everywhere continuous and strictly positive density. For any $0 < a < 1$ denote by $\rho(a)$ its quantile of order a: it is defined by $F(\rho(a)) = a$, where F is the cumulative distribution function of X.

Let N independent samplings of the law of X be ordered increasingly as

$$X_1 \leq X_2 \leq \cdots \leq X_N.$$

Let $0 < a < 1$ be such that Na has a non-integer value. Denote by μ its integer part and set $\rho_N(a) := X_{\mu+1}$.

1. For any x and h in $\mathbb{R}$, compute $\mathbb{P}(\rho_N(a) \in [x, x+h])$. By letting h tend to 0, deduce that the law of $\rho_N(a)$ has density

$$g_N(x) = \mathrm{C}_N^\mu (N-\mu)\big(F(x)\big)^\mu \big(1 - F(x)\big)^{N-\mu-1} F'(x).$$

2. Set

$$Y := \frac{\sqrt{N}}{\sqrt{a(1-a)}} F'\big(\rho(a)\big)\big(\rho_N(a) - \rho(a)\big).$$

Prove that the density q_N of Y is

$$q_N(x) := \frac{1}{F'(\rho(a))}\frac{\sqrt{a(1-a)}}{\sqrt{N}} g_N\left(\rho(a) + \frac{\sqrt{a(1-a)}}{\sqrt{N}}\frac{x}{F'(\rho(a))}\right).$$

3. By using the classic limit

$$\lim_{N\to\infty} \frac{\sqrt{a(1-a)}}{\sqrt{N}} \mathrm{C}_N^\mu a^\mu (1-a)^{N-\mu} \frac{N-\mu}{1-a} = \frac{1}{\sqrt{2\pi}},$$

show that, when N tends to infinity, the $\rho_N(a)$ converge in law to a Gaussian distribution with mean $\rho(a)$ and standard deviation

$$\frac{1}{F'(\rho(a))}\frac{\sqrt{a(1-a)}}{\sqrt{N}}.$$

Deduce an error estimate for the approximation of quantiles by a Monte Carlo method.

3.4 (Concentration Inequalities ($\star$)) Let U be a strictly convex function from $\mathbb{R}$ to $\mathbb{R}$ with bounded continuous second derivative. Assume there exists $\lambda > 0$ such that

$$U''(x) \geq \lambda \quad \text{for all } x \in \mathbb{R}. \tag{3.21}$$

1. Let (G_n) be a sequence of independent real-valued standard Gaussian random variables. For any $0 < h < \frac{1}{\lambda}$ and x_0 in $\mathbb{R}$, define a random sequence by $X_0 = x$ and

$$X_{n+1} = X_n - U'(X_n)h + \sqrt{h}G_n.$$

Let K be the operator which maps any bounded continuous function f to the function Kf defined as

$$Kf(x) := \mathbb{E}f(X_1).$$

Show that

$$\forall x \in \mathbb{R}, \quad \left|\frac{d}{dx}K(f)(x)\right| \leq (1 - \lambda h)K\big(|f'|\big)(x). \tag{3.22}$$

2. Denote by ν^n the law of the random variable X_n. Show that ν^1 satisfies a logarithmic Sobolev inequality with constant $2h$.
3. Set $K^0(f) := f$, $K^2(f) := K(K(f)), \ldots, K^p(f) := K(K^{p-1}(f))$. Using Definition 3.4 of the entropy, show that

$$\begin{aligned}&\mathrm{Ent}_{\nu^n}\big(f^2\big)\\ &= \sum_{i=1}^{n}\Big\{K^i\big(K^{n-i}\big(f^2\big)\log K^{n-i}\big(f^2\big)\big) - K^{i-1}\big(K^{n-i+1}\big(f^2\big)\log K^{n-i+1}\big(f^2\big)\big)\Big\}.\end{aligned}$$

4. Set $g_{n-i} := \sqrt{K^{n-i}(f^2)}$. Show

$$\mathrm{Ent}_{\nu^n}\big(f^2\big) \leq 2h\sum_{i=1}^{n} K^i\big(|g'_{n-i}|^2\big).$$

5. Using (3.22) show

$$|g'_{n-i}|^2 \leq (1 - \lambda h)^2 K\big(|g'_{n-i-1}|^2\big).$$

Deduce

$$|g'_{n-i}|^2 \leq (1 - \lambda h)^{2(n-i)} K^{n-i}\big(|f'|^2\big).$$

6. Show

$$\mathrm{Ent}_{\nu^n}\big(f^2\big) \leq \frac{2}{\lambda(2 - \lambda h)}\big(1 - (1 - \lambda h)^{2n}\big)K^n\big(|f'|^2\big).$$

7. Show

$$\left|g'_{n-i}\right|^2 \leq (1-\lambda h)^2 K\left(\left|g'_{n-i-1}\right|^2\right),$$

and deduce

$$\left|g'_{n-i}\right|^2 \leq (1-\lambda h)^{2(n-i)} K^{n-i}\left(\left|f'\right|^2\right).$$

8. Show

$$\mathrm{Ent}_{\nu^n}\left(f^2\right) \leq \frac{2}{\lambda(2-\lambda h)}\left(1-(1-\lambda h)^{2n}\right) K^n\left(\left|f'\right|^2\right).$$

Deduce a concentration inequality for the Monte Carlo method which approximates $\mathbb{E}f(X_n)$, where f is a Lipschitz function with Lipschitz constant smaller than 1, and then an estimate for

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{\ell=1}^{N} f\left(X_n^{(\ell)}\right) - \mathbb{E}_{\nu^n}(f)\right| \geq r\right)$$

in terms of n and the number N of simulations.

Part II
Exact and Approximate Simulation of Markov Processes

Markov processes of increasingly complex type will be introduced, from discrete to continuous space, and from pure jump to diffusion type. The study focuses on their concrete pathwise constructions and simulations.

This culminates with the solutions of Itô stochastic differential equations, for which implementable discretization schemes are derived, justified, and evaluated.

The generators of these Markov processes are successively computed, and the related equations (forward and backward Kolmogorov equations, Fokker–Planck equations, and the Feynman–Kac formula for solving terminal value equations) are given and discussed in a synthetic way. Several incursions are made into the realm of the profound relationship between Markov processes and martingales.

These Markov processes are used for Monte Carlo methods for the approximation of solutions of equations, ranging from systems of ordinary equations to integro-differential equations and parabolic partial differential equations, for which a number of results are proved using probabilistic techniques.

Chapter 4
Poisson Processes as Particular Markov Processes

Abstract We first introduce some practical and theoretical issues of modeling by means of Markov processes. Point processes are introduced in order to model jump instants. The Poisson process is then characterized as a point process without memory. The rest of the chapter consists in its rather detailed study, including various results concerning its simulation and approximation. This study is essential to understand the abstract constructions and the simulation methods for jump Markov processes developed in the following chapters.

4.1 Quick Introduction to Markov Processes

Markovian modeling of randomly evolving systems raises several issues. After giving some examples, we discuss a reasonably simple mathematical framework in which results can actually be proved.

4.1.1 Some Issues in Markovian Modeling

Many application fields feature random phenomena which evolve in continuous time, and in a continuous fashion between sudden state changes. The latter are called *jumps*, occur at *jump instants*, and may for instance be due to:

- arrivals, ends of service, and transfers of customers in communication networks,
- contagion, recovery, and deaths of individuals in epidemiology models,
- births, deaths, mutations, and transfers of individuals in ecological models,
- interactions or reactions between particles in physical or chemical models.

The natural state space may be *discrete*, as is the case when counting customers at the various network resources, or infected individuals in an epidemic. The evolution is then necessarily constant between the jumps.

The state space may also be *continuous*, as is the case when measuring phenotypes such as size or speed of individuals in ecology, or the energy in certain models of physics. The evolution may then be purely continuous, or mix jumps with continuous (possibly constant) stretches.

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_4, © Springer-Verlag Berlin Heidelberg 2013

The goal here will be to model random phenomena having a property of *lack of memory of the past* conditional on the *present*. A *Markov process* will be defined as a random evolution such that at any instant the law of the future evolution only depends on the past through the present state. The main case considered is when this law does not depend on the value of the present instant; else, the process will be said to be inhomogeneous (in time).

An important element of the study of Markov process with jumps is their set of discontinuities. In this chapter, *point processes* will be introduced as a representation for a sequence of jump instants without accumulation point. This is followed by detailed study of *Poisson processes*, which are point processes having a certain property of lack of memory. This study will be fundamental in the sequel, for abstract comprehension and construction as well as for effective simulation of Markov processes with jumps.

The next subsection is a quick introduction to some theoretical material.

4.1.2 Rudiments on Processes, Sample Paths, and Laws

Continuous time and state models for random phenomena necessitate a much more complex mathematical theory than *discrete* ones, and some abstract concepts must be introduced. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.

A (*stochastic*) *process* with values in a *state space* $\mathscr{V}$, which here will be either a discrete set or a closed subset of $\mathbb{R}^d$ with its Borel σ-field $\mathscr{B}(\mathscr{V})$, is a *random function* from $\mathbb{R}_+$, most often representing time, to $\mathscr{V}$:

$$(X_t)_{t\in\mathbb{R}_+} : \omega \in \Omega \mapsto \big(X_t(\omega)\big)_{t\in\mathbb{R}_+}, \quad X_t(\omega) \in \mathscr{V}.$$

It will always be assumed here that the *sample paths* $(X_t(\omega))_{t\in\mathbb{R}_+}$ belong to the Skorohod space $\mathscr{D}(\mathbb{R}_+, \mathscr{V})$ of functions from $\mathbb{R}_+$ to $\mathscr{V}$ which are *right-continuous* and have *left limits*.

This regularity assumption corresponds to mathematical constraints, as well as to the fact that the random phenomena on hand surprise us when they jump. For continuous evolutions, the sample paths belong to the subspace $\mathscr{C}(\mathbb{R}_+, \mathscr{V})$ of continuous functions.

The minimal measurability assumption is that each $X_t : \mathbb{R}_+ \to \mathscr{V}$ is a random variable (r.v.), i.e., is measurable. Hence $\mathscr{D}(\mathbb{R}_+, \mathscr{V})$ is equipped with the smallest such σ-field, called the *product* σ-field, engendered by the "cylindrical sets"

$$C = \big\{(x_t)_{t\in\mathbb{R}_+} \in \mathscr{D}(\mathbb{R}_+, \mathscr{V}) : x_{t_1} \in A_1, \ldots, x_{t_n} \in A_n\big\}$$

for $n \geq 1$ and $0 \leq t_1 < \cdots < t_n$ and $A_1, \ldots, A_n$ in $\mathscr{B}(\mathscr{V})$.

By a limit procedure, a function in $\mathscr{D}(\mathbb{R}_+, \mathscr{V})$ can be determined by its values on $\mathbb{Q}_+$, or on all finite subsets of $\mathbb{R}_+$, and this measurable space suits our purposes.

The law of a process $(X_t)_{t\in\mathbb{R}_+}$ is a probability measure on $\mathscr{D}(\mathbb{R}_+, \mathscr{V})$, and hence is characterized by the family of the laws of all random vectors

$$(X_{t_1}, \dots, X_{t_n}) = (X_{t_k})_{1\le k\le n}, \quad n \ge 1,\ 0 \le t_1 < \cdots < t_n,$$

called the *finite-dimensional marginals* of the process, or of its law.

Conversely, consider a family of laws

$$\mathscr{M} = \left\{\pi_{t_1,\dots,t_n} \in \mathscr{P}\left(\mathscr{V}^n\right) : n \ge 1,\ 0 \le t_1 < \cdots < t_n\right\}$$

satisfying the following natural *compatibility property*: if $0 \le s_1 < \cdots < s_k$ is a subsequence of $0 \le t_1 < \cdots < t_n$, then $\pi_{s_1,\dots,s_k}$ is the corresponding marginal of $\pi_{t_1,\dots,t_n}$. Then the Kolmogorov extension theorem (see, e.g., Theorem 2.2 in Karatzas and Shreve [26, Chap. 2]) ensures that there exists a (unique) law on $\mathscr{V}^{\mathbb{R}_+}$ with product σ-field having $\mathscr{M}$ for finite-dimensional marginals. Under further assumptions, it can be shown that this law corresponds to a process with sample paths in $\mathscr{D}(\mathbb{R}_+, \mathscr{V})$.

This book elects to construct the process sample paths from a simulation perspective, which enables to directly prove the existence of a law of a process with sample paths in $\mathscr{D}(\mathbb{R}_+, \mathscr{V})$ without invoking this theorem.

4.2 Poisson Processes: Characterization, Properties

Jump instants of stochastic processes will be modeled using point processes. Then, prefiguring the rigorous notion of Markov processes, the Poisson process will be characterized as a memoryless point process.

4.2.1 Point Processes and Poisson Processes

For pedagogical purposes, non-degenerate point processes are first introduced; the general case will be treated as an extension.

Definition 4.1 (Point process) A point process on $\mathbb{R}_+$ with an infinite number of (strictly) positive jump instants without accumulation point, is a process $(N_t)_{t\in\mathbb{R}_+}$ with values in $\mathbb{N} := \{0, 1, \dots\}$, vanishing at 0, non-decreasing, right continuous, with unit jumps, and with infinite limit: for $0 \le s \le t < \infty$,

$$0 = N_0 \le N_s \le N_t = N_{t+} \in \mathbb{N}, \quad N_t - N_{t-} \in \{0, 1\}, \quad \lim_{t\to\infty} N_t = \infty,$$

with the notation $N_{t+} = \lim_{u\to t^+} N_u$ and $N_{t-} = \lim_{u\to t^-} N_u$ and $N_{0-} = N_0$.

This kind of random object has two other equivalent descriptions, to which it makes implicit reference. All this is illustrated in Fig. 4.1.

A point process as in Definition 4.1 can be given equivalently by any of the two following random objects:

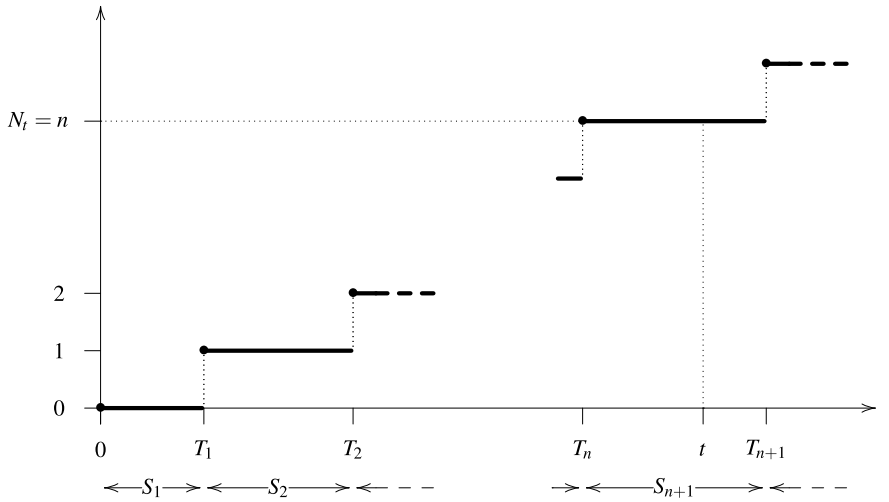

Fig. 4.1 A sample path of a point process, such as a Poisson process

(a) A sequence $(T_n)_{n\geq 1}$ of positive r.v.'s increasing to infinity:

$$0 < T_1 < T_2 < \cdots < \infty, \quad \lim_{n\to\infty} T_n = \infty.$$

The T_n are the *jump instants* of $(N_t)_{t\in\mathbb{R}_+}$, and are also called the *arrival instants* or the *points* of the point process (hence its name). Then $(N_t)_{t\in\mathbb{R}_+}$ is given by their counting process:

$$N_t = \sum_{n\geq 1} \mathbb{1}_{\{T_n\leq t\}} = \sup\{n \geq 1 : T_n \leq t\}, \quad \{N_t \geq n\} = \{T_n \leq t\}, \quad \text{etc.}$$

(b) A sequence $(S_n)_{n\geq 1}$ of positive r.v.'s such that $\sum_{n\geq 1} S_n = \infty$. The bijection between the *inter-arrivals* S_n and the jump instants T_n is given by

$$S_1 = T_1, \qquad S_n = T_n - T_{n-1}, \quad n \geq 2, \quad \text{or} \quad T_n = S_1 + \cdots + S_n, \quad n \geq 1.$$

Poisson Processes

An essential fact is that if a point process has the following property of *lack of memory* and of *stability* in time, then its law is determined up to time unit.

Definition 4.2 A point process $(N_t)_{t\in\mathbb{R}_+}$ has *independent and stationary increments* if for every $t, h \geq 0$ the r.v. $N_{t+h} - N_t$ is independent of $(N_s)_{0\leq s\leq t}$ and has the same law as N_h. To check this, it is enough to check that for every $n \geq 1$ and $0 = t_0 < t_1 < \cdots < t_n$ the $N_{t_k} - N_{t_{k-1}}$ for $1 \leq k \leq n$ are independent and have the same law as $N_{t_k - t_{k-1}}$.

Theorem 4.1 (Poisson process characterization)

(a) *Let $(N_t)_{t\in\mathbb{R}_+}$ be a point process with independent and stationary increments. Then there exists $\lambda > 0$ such that the law of N_t is Poisson $\mathscr{P}(\lambda t)$, i.e.,*

$$\mathbb{P}(N_t = k) = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k \in \mathbb{N},$$

and the inter-arrivals $(S_n)_{n\geq 1}$ are i.i.d. with $\mathscr{E}(\lambda)$ exponential law, i.e., have density $\mathbb{1}_{\{t>0\}}\lambda \mathrm{e}^{-\lambda t}$.

(b) *Conversely, the counting process $(N_t)_{t\in\mathbb{R}_+}$ for such inter-arrivals $(S_n)_{n\geq 1}$ is a point process with independent and stationary increments.*

Such a process $(N_t)_{t\in\mathbb{R}_+}$ is called a Poisson process *with* intensity *or* rate λ, *which can be interpreted as a* mean number of jumps per unit of time; *recall that* $\mathbb{E}(N_t) = \lambda t$ *and* $\mathbb{E}(S_1) = 1/\lambda$.

Proof (a) The proof introduces a time-scaling method on a grid with mesh $\varepsilon > 0$, which will be very useful in the sequel. This allows to exploit the links between binomial and Poisson random variables, and between geometric and exponential r.v.'s. Recall the notation

$$\lfloor s \rfloor = \sup\{n \in \mathbb{N} : n \leq s\}, \quad \lceil s \rceil = \inf\{n \in \mathbb{N} : s \leq n\}, \quad s \in \mathbb{R}_+. \tag{4.1}$$

Let

$$Y_k^\varepsilon = \mathbb{1}_{\{N_{k\varepsilon} > N_{(k-1)\varepsilon}\}}, \quad k \geq 1; \qquad N_t^\varepsilon = \sum_{k=1}^{\lfloor t/\varepsilon \rfloor} Y_k^\varepsilon, \quad t \geq 0.$$

Since the increments of $(N_t)_{t\in\mathbb{R}_+}$ are independent and stationary, the $(Y_k^\varepsilon)_{k\geq 1}$ are i.i.d. and

$$\mathbb{P}\big(Y_1^\varepsilon = 1\big) = p_\varepsilon := \mathbb{P}(N_\varepsilon \geq 1),$$

so that N_t^ε has binomial law $\mathscr{B}(\lfloor t/\varepsilon \rfloor, p_\varepsilon)$ with generating function

$$\mathbb{E}\big(z^{N_t^\varepsilon}\big) = \mathbb{E}\big(z^{Y_1^\varepsilon}\big)^{\lfloor t/\varepsilon \rfloor} = (1 - p_\varepsilon + p_\varepsilon z)^{\lfloor t/\varepsilon \rfloor}, \quad 0 \leq z \leq 1.$$

The fact that $(N_t)_{t\in\mathbb{R}_+}$ is right-continuous with isolated discontinuities implies that $\lim_{\varepsilon\to 0} N^\varepsilon_{\varepsilon\lceil t/\varepsilon \rceil} = N_t$ a.s., and thus also in law. In particular

$$\mathbb{P}(N_t = 0) = \lim_{\varepsilon\to 0} \mathbb{P}\big(N^\varepsilon_{\varepsilon\lceil t/\varepsilon \rceil} = 0\big) = \lim_{\varepsilon\to 0} (1 - p_\varepsilon)^{\lceil t/\varepsilon \rceil} = \lim_{\varepsilon\to 0} \mathrm{e}^{\frac{\log(1-p_\varepsilon)}{\varepsilon} t} \in [0, 1]$$

and there exists λ in $[0, \infty]$ such that

$$\lambda := -\lim_{\varepsilon\to 0}\frac{\log(1-p_\varepsilon)}{\varepsilon} = \lim_{\varepsilon\to 0}\frac{p_\varepsilon}{\varepsilon}, \qquad \mathbb{P}(N_t = 0) = \mathrm{e}^{-\lambda t},$$

and $\lim_{t\to 0} N_t = 0$ and $\lim_{t\to\infty} N_t = \infty$ imply that $0 < \lambda < \infty$. Then, the generating function of N_t is

$$\mathbb{E}\big(z^{N_t}\big) = \lim_{\varepsilon\to 0}\mathbb{E}\big(z^{N^\varepsilon_{\lceil t/\varepsilon\rceil}}\big) = \lim_{\varepsilon\to 0}\big(1+p_\varepsilon(z-1)\big)^{\lceil t/\varepsilon\rceil} = \mathrm{e}^{\lambda t(z-1)}, \quad 0\le z\le 1,$$

which corresponds to the Poisson law $\mathscr{P}(\lambda t)$.

Let us now prove the result on $(S_n)_{n\ge 1}$. For $n\ge 1$ we define approximate inter-arrivals G^ε_n in *mesh numbers* and S^ε_n in *natural time* by

$$G^\varepsilon_n = \inf\big\{j\ge 1 : Y^\varepsilon_{G^\varepsilon_1+\cdots+G^\varepsilon_{n-1}+j} = 1\big\}, \qquad S^\varepsilon_n = \varepsilon G^\varepsilon_n.$$

Clearly, a.s., for all n, $\lim_{\varepsilon\to 0}(S^\varepsilon_n)_{n\ge 1} = (S_n)_{n\ge 1}$. In addition, for all $j_k\ge 1$ for $k\ge 1$,

$$\begin{aligned}
&\mathbb{P}\big(G^\varepsilon_k = j_k : 1\le k\le n\big)\\
&\quad = \mathbb{P}\big(Y^\varepsilon_{j_1+\cdots+j_{k-1}+1} = \cdots = Y^\varepsilon_{j_1+\cdots+j_k-1} = 0, Y^\varepsilon_{j_1+\cdots+j_k} = 1 : 1\le k\le n\big)\\
&\quad = \prod_{k=1}^n (1-p_\varepsilon)^{j_k-1}p_\varepsilon,
\end{aligned}$$

thus the $(G^\varepsilon_n)_{n\ge 1}$ are i.i.d., hence so are the $(S^\varepsilon_n)_{n\ge 1}$. Taking limits, the $(S_n)_{n\ge 1}$ are i.i.d., and their common law is $\mathscr{E}(\lambda)$ since

$$\mathbb{P}(S_1 > t) = \mathbb{P}(N_t = 0) = \mathrm{e}^{-\lambda t}.$$

(This shows that the law of G^ε_1 is geometric on $\mathbb{N}^*$, and that the limit of the laws of the $\varepsilon G^\varepsilon_1$ is $\mathscr{E}(\lambda)$, which can be easily proved directly.)

(b) Let $(S_n)_{n\ge 1}$ be a sequence of i.i.d. inter-arrivals with law $\mathscr{E}(\lambda)$, and $(N_t)_{t\in\mathbb{R}_+}$ the corresponding point process. Let $n\ge 1$ and $0 = t_0\le t_1 < \cdots < t_n$ be arbitrary. For all $j_k\ge 0$ for $k\ge 1$, setting $J_k = j_1+\cdots+j_k$ it holds that

$$\{N_{t_k} - N_{t_{k-1}} = j_k : 1\le k\le n\} = \bigcap_{1\le k\le n}\{t_{k-1} < T_{J_{k-1}+1},\ldots,T_{J_k}\le t_k\}\cap\{t_n < T_{J_n+1}\}$$

and since $T_j = S_1+\cdots+S_j$ we obtain that

$$\begin{aligned}
&\mathbb{P}(N_{t_k} - N_{t_{k-1}} = j_k : 1\le k\le n) = \lambda^{J_n+1}\int_{\mathscr{D}}\mathrm{e}^{-\lambda(s_1+\cdots+s_{J_n+1})}\,ds_1\cdots ds_{J_n+1},\\
&\mathscr{D} = \Bigg[\prod_{\substack{1\le k\le n\\ j_k\ne 0}}\{t_{k-1} < s_1+\cdots+s_{J_{k-1}+1},\ldots,s_1+\cdots+s_{J_k}\le t_k\}\Bigg]\\
&\qquad\times\{t_n < s_1+\cdots+s_{J_n+1}\}.
\end{aligned}$$

The change of variables $s_1 + \cdots + s_j = u_j$, with Jacobian 1, yields

$$\mathbb{P}(N_{t_k} - N_{t_{k-1}} = j_k : 1 \le k \le n) = \lambda^{J_n+1} \int_{\mathscr{E}} \mathrm{e}^{-\lambda u_{J_n+1}}\, du_1 \cdots du_{J_n+1},$$

$$\mathscr{E} = \left[\prod_{\substack{1 \le k \le n \\ j_k \neq 0}} \{t_{k-1} < u_{J_{k-1}+1} \le \cdots \le u_{J_k} \le t_k\} \right] \times \{t_n < u_{J_n+1}\},$$

and the Fubini theorem allows to conclude that

$$\mathbb{P}(N_{t_k} - N_{t_{k-1}} = j_k : 1 \le k \le n)$$
$$= \left[\prod_{\substack{1 \le k \le n \\ j_k \neq 0}} \lambda^{j_k} \int_{t_{k-1} < z_1 < \cdots < z_{j_k} \le t_k} dz_1 \cdots dz_{j_k} \right] \lambda \int_{t_n < z} \mathrm{e}^{-\lambda z}\, dz.$$

This product form proves that the increments are independent, and the invariance of the Lebesgue measure under translations that they are stationary. □

Infinitesimal Point of View

The following infinitesimal point of view is made rigorous in the proof. For ε small,

$$\begin{cases} \mathbb{P}(N_{t+\varepsilon} - N_t = 0) = \mathrm{e}^{-\lambda\varepsilon} = 1 - \lambda\varepsilon + \mathscr{O}(\varepsilon^2), \\ \mathbb{P}(N_{t+\varepsilon} - N_t = 1) = \mathrm{e}^{-\lambda\varepsilon}\lambda\varepsilon = \lambda\varepsilon + \mathscr{O}(\varepsilon^2), \\ \mathbb{P}(N_{t+\varepsilon} - N_t \ge 2) = 1 - \mathrm{e}^{-\lambda\varepsilon} - \mathrm{e}^{-\lambda\varepsilon}\lambda\varepsilon \le \dfrac{1}{2}\lambda^2\varepsilon^2 = \mathscr{O}(\varepsilon^2). \end{cases} \tag{4.2}$$

Law of the Poisson Process

The law of $(N_t)_{t\in\mathbb{R}_+}$ is entirely determined by the fact that its inter-arrivals $(S_n)_{n\ge1}$ are i.i.d. $\mathscr{E}(\lambda)$ r.v.'s, which also allows to construct this process.

From a theoretical perspective, the law of $(N_t)_{t\in\mathbb{R}_+}$ is also characterized by its finite-dimensional marginals, which can be computed from the fact that the law of $(N_{t_k} - N_{t_{k-1}})_{1\le k\le n}$ is the product of the Poisson laws $\mathscr{P}(\lambda(t_k - t_{k-1}))$, for every $n \ge 1$ and $0 = t_0 < t_1 < \cdots < t_n$.

The end of the proof of Theorem 4.1 is an example of a computation on laws performed by expressing events in terms of $(S_n)_{n\ge1}$. It can be used to obtain results on the jump instants $(T_n)_{n\ge1}$ such as those in Problem 4.1.

Exercise 4.1 Use it to prove that the law of $N_{t_k} - N_{t_{k-1}}$ is Poisson $\mathscr{P}(\lambda(t_k - t_{k-1}))$.

Memoryless Laws

The property of the Poisson process is a generalization of the following classic result: the only memoryless $\mathbb{R}_+^*$-valued random variables are exponential.

Theorem 4.2 (Memoryless r.v.'s on $\mathbb{R}_+$) *A r.v. S with values in $\mathbb{R}_+ \cup \{\infty\}$ is* memoryless, *in the sense that*

$$\mathbb{P}(S > t + h \mid S > t) = \mathbb{P}(S > h), \quad t, h \geq 0,$$

if and only if either there exists $\lambda > 0$ such that S has $\mathscr{E}(\lambda)$ exponential law, or else S is a degenerate r.v. such that either $S = \infty$ a.s. or $S = 0$ a.s.

Proof With notation (4.1), for all $t > 0$ and $\varepsilon > 0$ the memoryless assumption yields that

$$\begin{aligned}\mathbb{P}\big(S > \lceil t/\varepsilon \rceil \varepsilon\big) &= \mathbb{P}(S > \varepsilon)\mathbb{P}\big(S > \lceil t/\varepsilon \rceil \varepsilon \mid S > \varepsilon\big) \\ &= \mathbb{P}(S > \varepsilon)\mathbb{P}\big(S > \big(\lceil t/\varepsilon \rceil - 1\big)\varepsilon\big)\end{aligned}$$

so that by induction

$$\mathbb{P}\big(S > \lceil t/\varepsilon \rceil \varepsilon\big) = \mathbb{P}(S > \varepsilon)^{\lceil t/\varepsilon \rceil}.$$

The right-continuity of $s \mapsto \mathbb{P}(S > s)$ implies that

$$\mathbb{P}(S > t) = \lim_{\varepsilon \to 0} \mathbb{P}\big(S > \lceil t/\varepsilon \rceil \varepsilon\big) = \lim_{\varepsilon \to 0} \mathbb{P}(S > \varepsilon)^{\lceil t/\varepsilon \rceil} = \lim_{\varepsilon \to 0} \mathrm{e}^{\lceil t/\varepsilon \rceil \log \mathbb{P}(S > \varepsilon)}$$

which yields the equality and the existence of the limit

$$\mathbb{P}(S > t) = \mathrm{e}^{-\lambda t}, \quad \lambda = -\lim_{\varepsilon \to 0} \frac{\log \mathbb{P}(S > \varepsilon)}{\varepsilon} \in [0, \infty].$$

Hence, according to whether $0 < \lambda < \infty$, $\lambda = 0$, or $\lambda = \infty$, the r.v. S has law $\mathscr{E}(\lambda)$, is infinite a.s, or vanishes a.s. Conversely, a simple computation shows that these r.v.'s are memoryless. □

Some Extensions

The proof of Theorem 4.1 easily extends to point processes which may have a *finite* number of jumps, or *not* vanish at zero. It yields two degenerate point processes:

- the first for $\lambda = 0$ with $N_t = 0$ and $T_n = \infty$ and $S_n = \infty$;
- the second for $\lambda = \infty$ with $N_t = \infty$ and $T_n = 0$ and $S_n = 0$.

A common *abuse of language* is to say that a Poisson process with intensity 0 corresponds to the first of these, a Poisson process with intensity ∞ to the second, and that an $\mathscr{E}(0)$ exponential r.v. is infinite and an $\mathscr{E}(\infty)$ exponential r.v. vanishes,

a.s. This allows to avoid being explicit about some trivial particular cases in some statements, but must be used with caution since it contradicts statements such as "exponential r.v.'s have a density". It is safer to be specific about the fact that these are *degenerate*.

The following exercise corresponds to i.i.d. jump sizes, instead of unit ones. Problem 4.3 will study a case with time-varying intensity.

Exercise 4.2 (Compound Poisson process) Take a Poisson process $(N_t)_{t\in\mathbb{R}_+}$ with intensity λ, and independently a sequence $(X_n)_{n\geq 1}$ of i.i.d. real r.v.'s with characteristic function ϕ. Let $Z_t = \sum_{n=1}^{N_t} X_n$. Prove for every $n \geq 1$ and $0 = t_0 < t_1 < \cdots < t_n$ that

$$\mathbb{E}\big(\mathrm{e}^{i\sum_{k=1}^n u_k(Z_{t_k}-Z_{t_{k-1}})}\big) = \prod_{k=1}^n \mathrm{e}^{\lambda(t_k-t_{k-1})(\phi(u_k)-1)}, \quad u_1,\ldots,u_n \in \mathbb{R}.$$

Deduce from this that $(Z_t)_{t\geq 0}$ has independent and stationary increments.

4.2.2 Simple and Strong Markov Property

Let $(N_t)_{t\in\mathbb{R}_+}$ be a Poisson process with intensity λ. For any $t \geq 0$, the process

$$(N_{t+h} - N_t)_{h\in\mathbb{R}_+}$$

counting all instants after time t is a point process with independent and stationary increments, and moreover $\mathbb{E}(N_{t+1} - N_t) = \lambda$. Theorem 4.1 yields that it is then *also* a Poisson process with intensity λ. Moreover, it is *independent* of $(N_s)_{0\leq s\leq t}$. See Fig. 4.2.

This fundamental property of lack of memory given the present and of temporal stability is a special case of the Markov property to be seen later.

It remains true if time t (which does not depend on chance) is replaced by certain *random* times T, for instance for T independent of $(N_t)_{t\in\mathbb{R}_+}$, or $T = T_n$ for $n \geq 1$. On the contrary, for example it is a simple matter to check that for

$$M = \inf\big\{n \geq 2 : |T_{n+1} - T_{n-1}| < 1\big\}, \quad T = T_M,$$

it holds that $\mathbb{P}(M < \infty) = 1$, but that the process $(N_{T+h} - N_T)_{h\in\mathbb{R}_+}$ is neither Poisson nor independent of $(N_s)_{0\leq s\leq T}$.

The following assumption on T prevents that it carry such a knowledge of the future behavior of the process, which transmits a memory.

Definition 4.3 A *stopping time* for $(N_t)_{t\in\mathbb{R}_+}$ is a r.v. T with values in $\mathbb{R}_+ \cup \{\infty\}$, such that for every $t \in \mathbb{R}_+$ the event $\{T \leq t\}$ belongs to the σ-field $\sigma((N_s)_{0\leq s\leq t})$.

Technically, this means that there exists a set of paths A_t which is measurable for the product σ-field and such that $\{T \leq t\} = \{(N_s)_{0\leq s\leq t} \in A_t\}$.

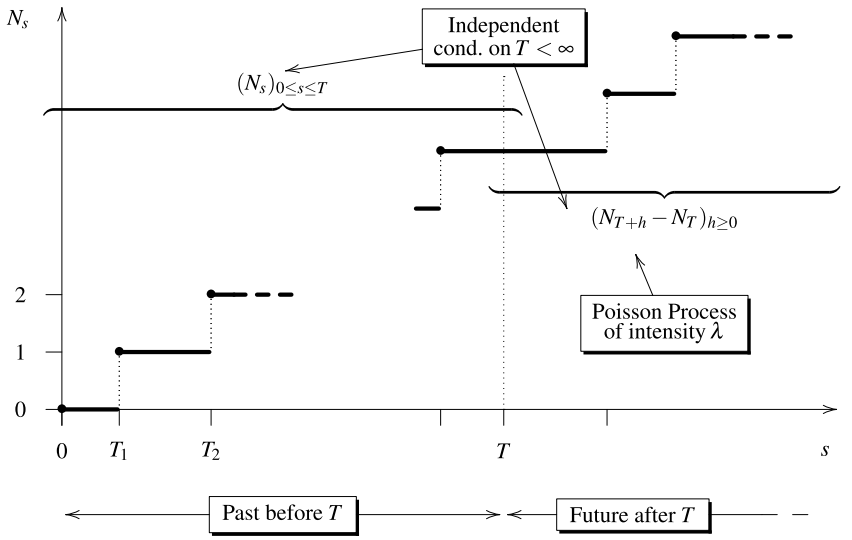

Fig. 4.2 The strong Markov property of a Poisson process of intensity λ, where T is a stopping time. The simple Markov property corresponds to the case when $T = t$, a.s.

Obviously, a deterministic r.v. T (such that $\mathbb{P}(T = t) = 1$ for some $t \in \mathbb{R}_+$) is a stopping time. The jump instants T_n are also stopping times.

Exercise 4.3 Prove that the T_ns are stopping times.

The Markov property for stopping times will be called the strong Markov property. The one for deterministic times is called the simple Markov property if such emphasis is needed.

Theorem 4.3 (Strong Markov property) *Let* $(N_t)_{t\in\mathbb{R}_+}$ *be a Poisson process with intensity* λ, *and* T *be a* stopping time. *Conditional on* $T < \infty$, *the process*

$$(N_{T+h} - N_T)_{h\in\mathbb{R}_+}$$

is a Poisson process with intensity λ *independent of* $(N_s)_{0\le s\le T}$, *i.e., independent of* T *and* $(N_{T\wedge s})_{s\ge 0}$. *See Fig.* 4.2.

Proof Conditioning allows to assume that $T < \infty$, a.s. First, it is enough to prove the result for stopping times with values in a discrete set: indeed, since T is the limit of the non-increasing sequence of stopping times

$$T^n = \sum_{k\in\mathbb{N}} \frac{k}{2^n} \mathbb{1}_{\{\frac{k-1}{2^n} < T \le \frac{k}{2^n}\}}, \quad n \in \mathbb{N}.$$

If the statement is true for all these T^n then the right-continuity of sample-paths allows to take limits and shows that the statement is also true for T.

Now, let T take values in an increasing sequence $(t_j)_{j\geq 0}$. Since

$$\{T = t_j\} = \{T \leq t_j\} - \{T \leq t_{j-1}\} \in \sigma\big((N_s)_{0\leq s\leq t_j}\big),$$

the simple Markov property at the times t_j yields, for $0 \leq s_1 < \cdots < s_n \leq t$ and $0 \leq h_1 < \cdots < h_n$ and $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{N}$, that

$$\begin{aligned}
&\mathbb{P}\big(T \leq t, (N_{T\wedge s_k})_{1\leq k\leq n} = (x_k)_{1\leq k\leq n}, (N_{T+h_k} - N_T)_{1\leq k\leq n} = (y_k)_{1\leq k\leq n}\big)\\
&\quad = \sum_{t_j\leq t} \mathbb{P}\big(T = t_j, (N_{t_j\wedge s_k})_{1\leq k\leq n} = (x_k)_{1\leq k\leq n}, (N_{t_j+h_k} - N_{t_j})_{1\leq k\leq n} = (y_k)_{1\leq k\leq n}\big)\\
&\quad = \sum_{t_j\leq t} \mathbb{P}\big(T = t_j, (N_{t_j\wedge s_k})_{1\leq k\leq n} = (x_k)_{1\leq k\leq n}\big)\mathbb{P}\big((N_{h_k})_{1\leq k\leq n} = (y_k)_{1\leq k\leq n}\big)\\
&\quad = \mathbb{P}\big(T \leq t, (N_{T\wedge s_k})_{1\leq k\leq n} = (x_k)_{1\leq k\leq n}\big)\mathbb{P}\big((N_{h_k})_{1\leq k\leq n} = (y_k)_{1\leq k\leq n}\big),
\end{aligned}$$

which suffices to conclude in this special case. □

Remark 4.1 This result is still true for T such that the $\{T \leq t\}$ can be expressed not only in terms of $(N_s)_{0\leq s\leq t}$ but also of events which are independent of $(N_{t+h} - N_t)_{h\in\mathbb{R}_+}$. For other generalizations, see below the proof of Theorem 4.4 or Exercise 5.2.

4.2.3 Superposition and Decomposition

The following are two natural operations on point processes:

- the *sum* or *superposition*, which *adds* the counting processes, and *reunites* the points (which must then be renumbered) into a single process;
- the *decomposition* or *marking*, which *marks* the points of a process, then reunites points accordingly to form a *family* of processes *indexed* by the marks.

In the following, Theorem 4.1 readily yields that the resulting processes are Poisson. The remarkable *independence* properties achieve the *symmetry* of the statement, and are related to the strong Markov property.

Theorem 4.4 (Superposition-decomposition) *Let $\mathscr{I}$ be a countable set.*

(a) *Let $(N_t^i)_{t\geq 0}$ be independent Poisson processes with intensities $\lambda_i > 0$ for i in $\mathscr{I}$ satisfying $\lambda := \sum_{i\in\mathscr{I}} \lambda_i < \infty$. Then $(N_t)_{t\geq 0}$ defined by*

$$N_t := \sum_{i\in\mathscr{I}} N_t^i$$

is a Poisson process with intensity λ, and is called the superposition, *or the* sum, *of the $(N_t^i)_{t\geq 0}$.*

Moreover, for $n \geq 1$, let T_n be the nth jump instant of $(N_t)_{t\geq 0}$, and Y_n the $\mathscr{I}$-valued r.v. given by

$$\{Y_n = i\} := \left\{N^i_{T_n} \neq N^i_{T_n-}\right\}, \quad i \in \mathscr{I}.$$

Then $(Y_n)_{n\geq 1}$ is a sequence of i.i.d. r.v.'s independent *of $(N_t)_{t\geq 0}$, and*

$$\mathbb{P}(Y_1 = i) = \frac{\lambda_i}{\lambda} = \frac{\lambda_i}{\sum_{j\in\mathscr{I}} \lambda_j}, \quad i \in \mathscr{I}.$$

(b) *Let $(N_t)_{t\geq 0}$ be a Poisson process with intensity $\lambda > 0$ and $(T_n)_{n\geq 1}$ its jump instants. Let $(Y_n)_{n\geq 1}$ be a sequence of i.i.d. r.v.'s with values in $\mathscr{I}$, independent of $(N_t)_{t\geq 0}$ and such that $\mathbb{P}(Y_1 = i) = p_i > 0$ for i in $\mathscr{I}$. Then the $(N^i_t)_{t\geq 0}$ defined by*

$$N^i_t := \sum_{n\geq 1} \mathbb{1}_{\{T_n\leq t, Y_n = i\}}, \quad i \in \mathscr{I},$$

are independent *Poisson processes with intensities $\lambda_i = \lambda p_i$, and are called the* decomposition, *or* marking, *of $(N_t)_{t\geq 0}$ according to $(Y_n)_{n\geq 1}$.*

Notably, the sampling *or* erasing *property holds: retaining among the instants of a Poisson process with intensity $\lambda > 0$ only those marked in i.i.d. fashion with probability $p > 0$ yields a Poisson process with intensity λp.*

Proof (a) Let us first assume that $\mathscr{I}$ is finite. Recall Definitions 4.1 and 4.2. The $(N^i_t)_{t\in\mathbb{R}_+}$ are point processes with independent and stationary increments (Theorem 4.1) and are independent, hence their finite sum $(N_t)_{t\in\mathbb{R}_+}$ is also such a process; the only fact which is not immediate is that it has unit jumps, i.e., that $N_t - N_{t-} \in \{0, 1\}$, which we now proceed to prove. For $\varepsilon > 0$ and $k \geq 1$ and $i \neq j$, independence implies that

$$\mathbb{P}\big(N^i_{k\varepsilon} - N^i_{(k-1)\varepsilon} \geq 1, N^j_{k\varepsilon} - N^j_{(k-1)\varepsilon} \geq 1\big) = \big(1 - \mathrm{e}^{\lambda_i\varepsilon}\big)\big(1 - \mathrm{e}^{\lambda_j\varepsilon}\big) \leq \lambda_i\lambda_j\varepsilon^2$$

which with (4.2) applied to the N^i yields

$$\mathbb{P}(N_{k\varepsilon} - N_{(k-1)\varepsilon} \geq 2) \leq \frac{1}{2}\sum_{i\in\mathscr{I}} \lambda_i^2\varepsilon^2 + \frac{1}{2}\sum_{i\neq j} \lambda_i\lambda_j\varepsilon^2 = \frac{1}{2}\lambda^2\varepsilon^2,$$

$$\mathbb{P}\big(\exists t \in [0, T] : N_t - N_{t-} \geq 2\big) \leq \frac{1}{2}\lceil T/\varepsilon\rceil\lambda^2\varepsilon^2 \xrightarrow[\varepsilon\to 0]{} 0, \quad \forall T > 0.$$

Then, Theorem 4.1 yields that $(N_t)_{t\in\mathbb{R}_+}$ is a Poisson process with intensity

$$\mathbb{E}(N_1) = \sum_{i\in\mathscr{I}} \mathbb{E}\big(N^i_1\big) = \sum_{i\in\mathscr{I}} \lambda_i = \lambda.$$

The case when $\mathscr{I}$ is countably infinite follows by taking increasing limit of sums on finite subsets and using the convergence of the series $\sum_{i\in\mathscr{I}} \lambda_i$. In particular N^i and N^j do not jump simultaneously for $i \neq j$, and the Y_n are well defined.

Now, for $n \geq 1$ and $t \geq 0$ it holds that

$$\{T_n \leq t\} = \left\{ \sum_{i \in \mathscr{I}} N_t^i \geq n \right\}$$

so that T_n is a stopping time for the *family* of processes $((N_t^i)_{t \geq 0} : i \in \mathscr{I})$. The proof of Theorem 4.3 (strong Markov property) can be easily extended to show that the family

$$\left(N^i_{T_n+h} - N^i_{T_n}\right)_{h \geq 0}, \quad i \in \mathscr{I},$$

is constituted of independent Poisson processes with intensities λ_i, and is independent of the family $(N_s^i)_{0 \leq s \leq T_n}$ for i in $\mathscr{I}$. This yields that (Y_{n+1}, T_{n+1}) is independent of the $(N_s^i)_{0 \leq s \leq T_n}$ and hence of $(Y_k, T_k)_{1 \leq k \leq n}$, and has the same law as (Y_1, T_1). Moreover, for i in $\mathscr{I}$ and $t \geq 0$,

$$\begin{aligned}
\mathbb{P}(Y_1 = i, T_1 > t) &= \mathbb{P}\big(S_1^j > S_1^i > t : j \neq i\big) \\
&= \lambda_i \int_t^\infty \mathrm{e}^{-\lambda_i s} \prod_{j \neq i} \mathbb{P}\big(S_1^j > s : j \neq i\big)\, ds \\
&= \lambda_i \int_t^\infty \mathrm{e}^{-\sum_{j \in \mathscr{I}} \lambda_j s}\, ds \\
&= \frac{\lambda_i}{\lambda} \mathrm{e}^{-\lambda t}
\end{aligned}$$

which characterizes the law of Y_1 and shows that Y_1 and T_1 are independent. Inductively, the sequence $(Y_n)_{n \geq 1}$ is constituted of i.i.d. r.v. with said law, and is independent of $(T_n)_{n \geq 1}$ and hence of $(N_t)_{t \in \mathbb{R}_+}$.

(b) For i in $\mathscr{I}$, clearly $(N_t^i)_{t \geq 0}$ is a point process with independent and stationary increments, and hence a Poisson process with intensity

$$\mathbb{E}\big(N_1^i\big) = \sum_{n \geq 1} \mathbb{E}(\mathbb{1}_{\{T_n \leq 1, Y_n = i\}}) = p_i \sum_{n \geq 1} \mathbb{E}(\mathbb{1}_{\{T_n \leq 1\}}) = p_i \mathbb{E}(N_1) = \lambda_i .$$

In order to prove that the $(N_t^i)_{t \geq 0}$ are independent, it is sufficient to prove that the $(N^i_{t_k} - N^i_{t_{k-1}})_{1 \leq k \leq n}$ are independent for every $0 = t_0 < t_1 < \cdots < t_n$, and since the increments of $(N_t^i)_{t \geq 0}$ are independent and stationary, it is sufficient to prove that the N_t^i are independent for every $t \geq 0$. Using the multinomial law, for all $n_i \in \mathbb{N}$ for i in $\mathscr{I}$ such that $n = \sum_{i \in \mathscr{I}} n_i < \infty$ it holds that

$$\begin{aligned}
\mathbb{P}\big(N_t^i = n_i, i \in \mathscr{I}\big) &= \mathbb{P}(N_t = n) \mathbb{P}\big(N_t^i = n_i, i \in \mathscr{I} \mid N_t = n\big) \\
&= \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} \frac{n!}{\prod_{i \in \mathscr{I}} n_i!} \prod_{i \in \mathscr{I}} p_i^{n_i} = \mathrm{e}^{-\lambda t} \prod_{i \in \mathscr{I}} \frac{(\lambda p_i t)^{n_i}}{n_i!},
\end{aligned}$$

and hence the N_t^i are independent. □

Application to Exponential Random Variables

The construction, and hence simulation, of Poisson processes relies on exponential random variables, and we state an important corollary.

Corollary 4.1 *Let $\mathscr{I}$ be a countable set, $\lambda_i > 0$ for i in $\mathscr{I}$ be such that $\lambda := \sum_{i\in\mathscr{I}} \lambda_i < \infty$, and S_i be independent r.v.'s with $\mathscr{E}(\lambda_i)$ exponential laws. Then the r.v.*

$$S := \inf_{i\in\mathscr{I}} S_i$$

has $\mathscr{E}(\lambda)$ exponential law, the infimum is attained at a unique random index Y independent of S and with law given by

$$\mathbb{P}(Y = i) = \frac{\lambda_i}{\lambda} = \frac{\lambda_i}{\sum_{j\in\mathscr{I}} \lambda_j}, \quad i \in \mathscr{I},$$

and conditionally on $Y = i$, the $S_j - S_i$ for $j \neq i$ form a family of independent r.v.'s with $\mathscr{E}(\lambda_j)$ exponential laws which is independent of S.

Exercise 4.4 Prove this corollary using Theorem 4.4.

4.3 Simulation and Approximation

Exact and approximate simulation of Poisson processes will be discussed, as well as some statistical and computational issues.

4.3.1 Simulation of Inter-arrivals

As seen in Theorem 4.1, the inter-arrivals $(S_n)_{n\geq 1}$ of a Poisson process of intensity $\lambda > 0$ are i.i.d. $\mathscr{E}(\lambda)$ exponential r.v.'s, and the jump instants $(T_n)_{n\geq 1}$ and counting process $(N_t)_{t\in\mathbb{R}_+}$ can be reconstituted from this sequence.

As seen in Sect. 2.2.4, $(S_n)_{n\geq 1}$ can be simulated using an i.i.d. sequence $(U_n)_{n\geq 1}$ of uniform r.v.'s on $[0, 1]$ by setting

$$S_n = -\frac{1}{\lambda} \log U_n, \quad n \geq 1.$$

From a programming perspective, such a sequence $(U_n)_{n\geq 1}$ is obtained by repeated calls of a function such as `rand`. Many programming languages have ready-made functions that yield draws according to the $\mathscr{E}(\lambda)$ law; if this is not the case, such a function must be carefully defined.

Algorithm (Poisson process simulation) To simulate a sample path of a Poisson process of intensity λ: set $t_0 = 0$, then iteratively for $n \geq 1$:

- draw a sample s_n from the $\mathscr{E}(\lambda)$ exponential law, and set $t_n = t_{n-1} + s_n$ for the value taken by T_n, the nth jump instant,
- for $t_n \leq t < t_{n+1}$, the counting function N_t takes the value n.

This terminates either when the simulated instant goes beyond some temporal horizon that was fixed beforehand, or due to some other criterion.

In order to simulate a Poisson $\mathscr{P}(\lambda)$ r.v. with a parameter λ which is not "too large", the most effective algorithm is usually to simulate a Poisson process $(N_t)_{t\geq 0}$ with intensity 1 for a duration λ and take its final value N_λ.

Approximations for large values of λ will be given in Sect. 4.3.3.

Exercise 4.5 Let $(U_n)_{n\geq 1}$ be i.i.d. uniform r.v.'s on $[0, 1]$. Give the law of

$$N = \min\left\{ n \in \mathbb{N} : \prod_{k=1}^{n+1} U_k < \mathrm{e}^{-\lambda} \right\}.$$

What is the use of such a formulation?

4.3.2 Simulation of Independent Poisson Processes

Theorem 4.4 and its Corollary 4.1 are *essential* for simulation. In order to simulate, for i in $\mathscr{I}$, independent Poisson processes $(N_t^i)_{t\geq 0}$ with intensities λ_i that have a finite sum, the by far most economical and practical method is to simulate a Poisson process $(N_t)_{t\geq 0}$ with intensity λ and i.i.d. r.v.'s Y_n with values in $\mathscr{I}$ for $n \geq 1$ such that

$$\lambda = \sum_{i\in\mathscr{I}} \lambda_i < \infty, \quad \mathbb{P}(Y_1 = i) = \frac{\lambda_i}{\lambda}, \quad i \in \mathscr{I},$$

and then to recover the $(N_t^i)_{t\geq 0}$ using Theorem 4.4(b). The corresponding algorithm can be summarized as follows.

Algorithm (Independent Poisson process simulation) To simulate the sample paths of independent Poisson processes of intensity λ_i, for i in $\mathscr{I}$: compute $\lambda = \sum_{i\in\mathscr{I}} \lambda_i$, set $t_0 = 0$ and $x^i = 0$ for i in $\mathscr{I}$, then iteratively for $n \geq 1$:

- draw a sample s_n from the $\mathscr{E}(\lambda)$ exponential law and set $t_n = t_{n-1} + s_n$ (the value of T_n, the nth jump instant of the superposition),
- draw a sample k in $\mathscr{I}$ from the discrete law $(\lambda_i/\lambda)_{i\in\mathscr{I}}$, set $x_n^k = x_n^{k-1} + 1$ and $x_n^i = x_n^{i-1}$ for $i \neq k$,
- for $t_n \leq t < t_{n+1}$, the N_t^i take the value x_n^i for all i in $\mathscr{I}$.

In contrast, the most direct and naive simulation algorithm, numerically much less efficient, would be as follows.

Algorithm (Simulation of renewal processes) To simulate the sample paths of independent Poisson processes of intensity λ_i, for i in $\mathscr{I}$: set $t_0 = 0$ and draw samples s_1^i from the $\mathscr{E}(\lambda_i)$ exponential law for i in $\mathscr{I}$, then iteratively for $n \geq 1$:

- determine the least inter-arrival and corresponding index

$$s_n := \min_{i \in \mathscr{I}} s_n^i, \qquad k_n := \operatorname*{argmin}_{i \in \mathscr{I}} s_n^i,$$

  and set $t_n^{k_n} = t_{n-1}^{k_n} + s_n$ for the next jump instant of $(N_t^{k_n})_{t \geq 0}$,
- set $s_{n+1}^i = s_n^i - s_n$ for $i \neq k_n$ and draw a new sample $s_{n+1}^{k_n}$ from the $\mathscr{E}(\lambda_{k_n})$ law.

This method can be used to simulate an independent family of point process with arbitrary i.i.d. inter-arrivals, which are called renewal processes. This exemplifies that it does not exploit at all the strong Markov property of the Poisson processes; for instance, another inefficient algorithm would be as follows.

Algorithm (Intermediate method) To simulate the sample paths of independent Poisson processes of intensity λ_i, for i in $\mathscr{I}$: set $t_0 = 0$, then iteratively for $n \geq 1$:

- draw samples s_n^i from the $\mathscr{E}(\lambda_i)$ exponential law for i in $\mathscr{I}$,
- determine the least inter-arrival and corresponding index

$$s_n := \min_{i \in \mathscr{I}} s_n^i, \qquad k_n := \operatorname*{argmin}_{i \in \mathscr{I}} s_n^i,$$

  and set $t_n^{k_n} = t_{n-1}^{k_n} + s_n$ for the next jump instant of $(N_t^{k_n})_{t \geq 0}$,

The initial, and most efficient, algorithm is obtained by putting to best use the superposition-decomposition property.

Exercise 4.6 Justify that the three algorithms allow to sample trajectories of the Poisson process.

4.3.3 Long Time or Large Intensity Limit, Applications

The asymptotic behavior of Poisson processes in long times, or with large intensities, is of interest. Recall that $\mathbb{E}(N_t) = \operatorname{var}(N_t) = \lambda t$.

Theorem 4.5 (SLLN and CLT)

(a) *A Poisson process* $(N_t)_{t\geq 0}$ *with intensity* λ *satisfies the* Strong Law of Large Numbers *and the* Central Limit Theorem

$$\frac{N_t}{t} \xrightarrow[t\to\infty]{\text{a.s.}} \lambda, \qquad \frac{N_t - \lambda t}{\sqrt{\lambda t}} \xrightarrow[t\to\infty]{\text{in law}} \mathcal{N}(0,1).$$

(b) *Poisson processes* $(N_t^\lambda)_{t\geq 0}$ *with intensities* $\lambda > 0$ *satisfy the* Weak Law of Large Numbers *and the* Central Limit Theorem *for every fixed* $t > 0$

$$\frac{N_t^\lambda}{\lambda} \xrightarrow[\lambda\to\infty]{\text{in probab}} t, \qquad \frac{N_t^\lambda - \lambda t}{\sqrt{\lambda t}} \xrightarrow[\lambda\to\infty]{\text{in law}} \mathcal{N}(0,1).$$

Proof (a) For $n \leq t < n+1$ it holds that

$$\frac{n}{n+1}\frac{1}{n}\sum_{k=1}^{n}(N_k - N_{k-1}) = \frac{N_n}{n+1} \leq \frac{N_t}{t} \leq \frac{N_{n+1}}{n} = \frac{n+1}{n}\frac{1}{n+1}\sum_{k=1}^{n+1}(N_k - N_{k-1})$$

and since the increments of $(N_t)_{t\geq 0}$ are independent and stationary, the strong law of large numbers for i.i.d. r.v., applied to the lower and upper bound, yields the a.s. result. To prove the Central Limit Theorem, we use that, for $n \geq 0$,

$$\frac{N_t - \lambda t}{\sqrt{\lambda t}} = \frac{N_n - \lambda n}{\sqrt{\lambda n}}\sqrt{\frac{n}{t}} + \frac{N_t - N_n}{\sqrt{\lambda t}} + \frac{n-t}{\sqrt{t}}\sqrt{\lambda}, \quad n \leq t < n+1,$$

where $\lim_{t\to\infty} \frac{n-t}{\sqrt{t}} = 0$ and $\lim_{t\to\infty} \sqrt{\frac{n}{t}} = 1$. Moreover, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{N_t - N_n}{\sqrt{\lambda t}}\right| > \varepsilon\right) = \mathbb{P}(N_{t-n} > \varepsilon\sqrt{\lambda t}) \leq \mathbb{P}(N_1 > \varepsilon\sqrt{\lambda t}) \xrightarrow[t\to\infty]{} 0.$$

Lastly, the Central Limit Theorem for i.i.d. r.v. yields

$$\frac{N_n - \lambda n}{\sqrt{\lambda n}} = \frac{1}{\sqrt{\lambda n}}\sum_{k=1}^{n}(N_k - N_{k-1} - \lambda) \xrightarrow[n\to\infty]{\text{in law}} \mathcal{N}(0,1).$$

From these facts, the characteristic function of $\frac{N_t^\lambda - \lambda t}{\sqrt{\lambda t}}$ converges pointwise to the characteristic function of the $\mathcal{N}(0,1)$ law, yielding convergence in law.

(b) These results in law follow immediately from (a), using the fact that N_t^λ and N_t are Poisson $\mathcal{P}(\theta)$-r.v.'s with $\theta = \lambda t$ going to infinity when one of λ or t goes to infinity and the other is fixed. □

Problem 4.2 will provide a direct proof for the Central Limit Theorem, as well as concentration inequalities, for Poisson r.v.'s.

Problem 4.4 will state and provide elements of proof for a stronger version of the Central Limit Theorem: the Brownian limit for an appropriately recentered and normalized Poisson process.

Estimation of the Intensity and Simulation Duration

It appears that the important parameter in these results is $\theta := \lambda t$. From a statistical perspective:

- the SLLN provides an unbiased consistent estimator for the intensity λ of an observed Poisson process of rate λ on a time interval $[0, t]$,
- the CLT provides asymptotic confidence intervals for this estimation, which are very good except for small values of θ.

Algorithm (Estimation of intensity) To estimate the intensity λ of a Poisson process $(N_s)_{s\geq 0}$ observed on a time interval $[0, t]$:

- estimate λ by $\hat{\lambda}_t = \frac{N_t}{t}$,
- derive asymptotic confidence intervals, etc., by pretending that $\frac{N_t}{t} - \lambda$ follows the $\mathcal{N}(0, \sqrt{\hat{\lambda}_t / t})$ Gaussian law.

Concentration inequalities are available if need be, see Problem 4.2.

After this incursion into statistics, let us come back to Monte Carlo simulation:

- the SLLN shows that the duration of the exact simulation of a Poisson process $(N_s)_{s\geq 0}$ of intensity λ on a time interval $[0, t]$ is of approximately θ simulation step durations,
- the CLT yields asymptotic confidence intervals for this estimation, which are very good except for small values of θ (for which the simulation is very short).

This allows to estimate the cost (in machine time) of such a simulation, which is proportional to θ.

Exercise 4.7 For not too small θ, show that the probability that this duration exceeds $\theta + 1.96\sqrt{\theta}$ simulation steps is approximately 2.5 %.

Approximate Simulation

Clearly, problems arise for large values of $\theta := \lambda t$.

These are even more apparent when one wishes to simulate a $\mathcal{P}(\theta)$ Poisson r.v., since then the numerous intermediate instants on $[0, \theta]$ of the Poisson process with intensity 1, computed by the classic simulation method, are not further used.

Approximate simulation methods based on Theorem 4.5 may then be used. For large θ, the lengthy exact simulation of a $\mathcal{P}(\theta)$ Poisson r.v. is then replaced:

- by the *deterministic approximation* given by its expectation θ with asymptotic confidence intervals,
- or, better still, by the more precise *random approximation* obtained by adding to θ a $\mathcal{N}(0, 1)$ Gaussian r.v. multiplied by $\sqrt{\theta}$,

rounding appropriately if one needs an integer value.

4.4 Problems

4.1 (Law of the Instants of a Poisson Process) Let $(N_t)_{t\in\mathbb{R}_+}$ be a Poisson process with intensity $\lambda > 0$ and instants $(T_n)_{n\geq 1}$.

1. Prove that the law of $(T_1, \dots, T_n)$ has density $\lambda^n \mathbb{1}_{0<t_1<\cdots<t_n} \mathrm{e}^{-\lambda t_n}$ on $\mathbb{R}^n$, and that the law of T_n is gamma $\Gamma(n, \lambda)$, with density $\frac{\lambda^n}{(n-1)!}\mathbb{1}_{t>0} t^{n-1}\mathrm{e}^{-\lambda t}$ on $\mathbb{R}$.
2. Prove that conditionally on $N_t = n$ the vector $(T_1, \dots, T_n)$ has uniform law on the simplex $\{(t_1, \dots, t_n) : 0 < t_1 < \cdots < t_n < t\}$ of $\mathbb{R}^n$. Prove that this law is also the law of the increasing reordering of n i.i.d. uniform r.v.'s on $[0, t]$.
3. Prove that conditionally on $T_{n+1} = t$ the vector $(T_1, \dots, T_n)$ again has uniform law on the simplex $\{(t_1, \dots, t_n) : 0 < t_1 < \cdots < t_n < t\}$ of $\mathbb{R}^n$.
4. Compute the law of $(\frac{T_1}{T_{n+1}}, \dots, \frac{T_n}{T_{n+1}})$.

4.2 (Central Limit Theorem and Concentration Inequalities for Poisson Laws) For each $\lambda > 0$, let X_λ be a Poisson $\mathscr{P}(\lambda)$ random variable and $W_\lambda = \frac{X_\lambda - \lambda}{\sqrt{\lambda}}$.

1. Compute the generalized Laplace transform $\psi_\lambda(z) := \mathbb{E}(\mathrm{e}^{zX_\lambda})$ for $z \in \mathbb{C}$.
2. Deduce from this that $\lim_{\lambda\to\infty} W_\lambda$ in law exists, and give the limit law.
3. Prove that, for $a \geq 0$ and $\theta > 0$,

$$\mathbb{P}(X_\lambda - \lambda \geq a) \leq \mathrm{e}^{\lambda \mathrm{e}^\theta - \lambda - \theta(\lambda + a)},$$

$$\mathbb{P}(X_\lambda - \lambda \geq a) \leq \mathrm{e}^{a - (\lambda+a)\log(1+\frac{a}{\lambda})} \leq \mathrm{e}^{-\frac{a^2}{2\lambda}(1-\frac{a}{\lambda})}.$$

   Deduce from this a bound for $\mathbb{P}(W_\lambda \geq c)$ for $c \geq 0$.
4. Prove that, for $b \geq 0$ and $\theta > 0$,

$$\mathbb{P}(X_\lambda - \lambda \leq -b) \leq \mathrm{e}^{\lambda \mathrm{e}^{-\theta} - \lambda + \theta(\lambda - b)}, \qquad \mathbb{P}(X_\lambda - \lambda \leq -b) \leq \mathrm{e}^{-b - (\lambda - b)\log(1-\frac{b}{\lambda})},$$

   and then (one must keep b far from λ, the inequality becoming weak)

$$\mathbb{P}(X_\lambda - \lambda \leq -b) \leq \mathrm{e}^{-\frac{b^2}{2\lambda}(1-\frac{b}{\lambda}) + (\lambda - b)\frac{8b^3}{3\lambda^3}}, \quad 0 \leq b \leq \frac{\lambda}{2}.$$

   Deduce from this a bound for $\mathbb{P}(W_\lambda \leq d)$ for $d \leq 0$.

4.3 (Inhomogeneous Poisson Process ($\star$)) Let $\theta : \mathbb{R}_+ \to \mathbb{R}_+$ be a locally integrable function such that $\int_0^\infty \theta(s)\, ds = \infty$. The objective is to construct a point process $(N_t)_{t\in\mathbb{R}_+}$ with independent increments and such that for $t, h \geq 0$ the law of $N_{t+h} - N_t$ is a function of the quantity $\int_t^{t+h} \theta(s)\, ds$ only (and not t).

Let $(T_n)_{n\geq 1}$ be the jump instants of $(N_t)_{t\in\mathbb{R}_+}$, and $(\hat{T}_n)_{n\geq 1}$ those of the point process $(\hat{N}_t)_{t\in\mathbb{R}_+}$ given by

$$\Theta(t) = \int_0^t \theta(s)\, ds, \qquad A(t) = \inf\big\{u \in \mathbb{R}_+ : \Theta(u) \geq t\big\}, \qquad \hat{N}_t = N_{A(t)}.$$

1. Let $t \in \mathbb{R}_+$. Prove that

$$\Theta\big(A(t)\big) = t, \qquad A\big(\Theta(t)\big) = \inf\big\{u \in \mathbb{R}_+ : \Theta(u) = \Theta(t)\big\} \le t, \qquad N_t = \hat{N}_{\Theta(t)}.$$

   Deduce from this that $T_n = A(\hat{T}_n)$ and $\hat{T}_n = \Theta(T_n)$ for $n \ge 1$.
2. Prove that $(\hat{N}_t)_{t\in\mathbb{R}_+}$ is a Poisson process.

This determines the law of $(N_t)_{t\in\mathbb{R}_+}$ up to the intensity $\lambda > 0$ of $(\hat{N}_t)_{t\in\mathbb{R}_+}$, i.e., up to time scale or a multiplicative factor for θ, which must be fixed. A natural choice is $\lambda = 1$, so that $\mathbb{E}(N_t) = \Theta(t)$.

3. Suggest a method for simulating $(N_t)_{t\in\mathbb{R}_+}$. What problem can arise in practice?
4. Describe the law of $(N_{t_1}, N_{t_2} - N_{t_1}, \dots, N_{t_n} - N_{t_{n-1}})$ for $0 < t_1 < \cdots < t_n$.
5. Prove that

$$\mathbb{P}(T_{n+1} - T_n \ge t \mid T_1, \dots, T_n) = \mathbb{P}(T_{n+1} - T_n \ge t \mid T_n) = \mathrm{e}^{\Theta(T_n+t)-\Theta(T_n)}.$$

6. Let $(U_k)_{k\ge1}$ be i.i.d. uniform r.v. on $[0, 1]$. Prove that $(T_n)_{n\ge1}$ can be simulated by setting

$$T_n = \inf\left\{t \ge 0 : \Theta(t) \ge -\sum_{k=1}^{n} \log U_k\right\} = \inf\left\{t \ge 0 : \prod_{k=1}^{n} U_k \ge \mathrm{e}^{-\Theta(t)}\right\}.$$

7. Assume that $\sup_{t>0} \theta(t) \le \lambda$. Let $(N_t^*)_{t\in\mathbb{R}_+}$ be a Poisson process with intensity λ and instants $(T_n^*)_{n\ge1}$, and independently let $(V_k)_{k\ge1}$ be i.i.d. uniform r.v. on $[0, 1]$. Prove that $(N_t)_{t\in\mathbb{R}_+}$ can be simulated by setting

$$N_t = \sum_{n\ge1} \mathbb{1}_{\{T_n^* \le t, \lambda V_n \le \theta(T_n^*)\}}.$$

4.4 (Brownian Limit for the Poisson Process ($\star$)) The goal of this problem is to establish the principal elements for the proof of the following result, which gives an approximation in certain scales for the Poisson process.

Theorem 4.6 *Let $(N_t)_{t\ge0}$ be a Poisson process with intensity 1. Then*

$$\big(W_t^\theta\big)_{t\ge0} := \left(\frac{N_{\theta t} - \theta t}{\sqrt{\theta}}\right)_{t\ge0} \xrightarrow[\theta\to\infty]{\text{in law}} (W_t)_{t\ge0}$$

where $(W_t)_{t\ge0}$ is a standard Brownian motion.

1. Let X_x be a Poisson $\mathscr{P}(x)$ random variable, with arbitrary parameter $x > 0$.
   (a) Compute the generalized Laplace transform $\psi_x(z) := \mathbb{E}(\mathrm{e}^{zX_x})$ for $z \in \mathbb{C}$.
   (b) Deduce from this that

$$\mathbb{E}(X_x) = x, \qquad \mathbb{E}\big(X_x^2\big) = x + x^2, \qquad \mathbb{E}\big(X_x^3\big) = x + 3x^2 + x^3,$$
$$\mathbb{E}\big(X_x^4\big) = x + 7x^2 + 6x^3 + x^4.$$

2. We are going to prove a convergence result for finite-dimensional marginals.
   (a) Prove for all $t \geq 0$ that
   $$W_t^\theta \xrightarrow[\theta\to\infty]{\text{in law}} W_t.$$
   (b) Prove for all $0 \leq t_1 < \cdots < t_n$ that (with $t_0 := 0$ by convention)
   $$\big(W_{t_1}^\theta - W_{t_0}^\theta, \ldots, W_{t_n}^\theta - W_{t_{n-1}}^\theta\big) \xrightarrow[\theta\to\infty]{\text{in law}} (W_{t_1} - W_{t_0}, \ldots, W_{t_n} - W_{t_{n-1}}).$$
   (c) Prove for all $0 \leq t_1 < \cdots < t_n$ that
   $$\big(W_{t_1}^\theta, \ldots, W_{t_n}^\theta\big) \xrightarrow[\theta\to\infty]{\text{in law}} (W_{t_1}, \ldots, W_{t_n}).$$
3. In the sequel, $T > 0$ and $\varepsilon > 0$ and $a > 0$ are arbitrary.
   (a) Prove for $K(T,\varepsilon) \in \mathbb{N}$ given by $K(T,\varepsilon)\varepsilon < T \leq K(T,\varepsilon)\varepsilon + \varepsilon$ that
   $$\sup_{0\leq s,t\leq T, |t-s|\leq\varepsilon} \big|W_t^\theta - W_s^\theta\big| \leq 3\sup\Big\{\sup_{k\varepsilon\leq t\leq k\varepsilon+\varepsilon} \big|W_t^\theta - W_{k\varepsilon}^\theta\big| : k = 0, \ldots, K(T,\varepsilon)\Big\}.$$
   (b) Prove that
   $$\mathbb{P}\Big(\sup_{0\leq s,t\leq T, |t-s|\leq\varepsilon} \big|W_t^\theta - W_s^\theta\big| \geq 3a\Big) \leq \frac{T+\varepsilon}{\varepsilon}\mathbb{P}\Big(\sup_{0\leq t\leq\varepsilon} \big|W_t^\theta\big| \geq a\Big).$$
4. We are going to prove a classic result. Let $(M_n)_{n\in\mathbb{N}}$ be a discrete-time non-negative sub-martingale. Let $m \geq 1$ be an integer, $c > 0$ a real number, and
   $$\tau = \inf\{n \in \mathbb{N} : M_n \geq c\} \wedge m \quad (\tau = m \text{ if } M_n < c \text{ for } 0 \leq n \leq m).$$
   Prove that
   $$\mathbb{P}\Big(\sup_{0\leq n\leq m} M_n \geq c\Big) = \mathbb{P}(M_\tau \geq c) \leq \frac{1}{c}\mathbb{E}(M_\tau) \leq \frac{1}{c}\mathbb{E}(M_m).$$
   We **admit** that this can be generalized to continuous time as follows:

Lemma 4.1 (Doob's maximal sub-martingale inequality) *Let $(M_t)_{t\in\mathbb{R}_+}$ be a non-negative sub-martingale, right-hand continuous with left-hand limits. Let $u > 0$ and $c > 0$ be real numbers. Then*
$$\mathbb{P}\Big(\sup_{0\leq t\leq u} M_t \geq c\Big) \leq \frac{1}{c}\mathbb{E}(M_u).$$

5. We are going to obtain a fundamental bound.

   (a) Prove that $(W_t^\theta)_{t\geq 0}$ is a martingale, for every $\theta > 0$.
   (b) Prove that $(|W_t^\theta|^p)_{t\geq 0}$ is a sub-martingale, for every $\theta > 0$ and $p \geq 1$.
   (c) Prove that

$$\mathbb{P}\Big(\sup_{0\leq s,t\leq T,|t-s|\leq\varepsilon} \big|W_t^\theta - W_s^\theta\big| \geq 3a\Big) \leq \frac{T+\varepsilon}{a^4}\left(\frac{1}{\theta} + 3\varepsilon\right).$$

Complementary Elements for the Interested Reader This theorem is very powerful, and concerns the approximation of functionals of the whole process, such as extrema or hitting times. A complete rigorous proof—and even an actual understanding of the statement—requires tools far beyond those introduced in this course, since it bears on convergence in law of processes, i.e., of weak convergence of their laws on the Skorokhod space $\mathscr{D}(\mathbb{R}_+, \mathbb{R})$.

The last bound we have established concerns the modulus of continuity of $(W_t^\theta)_{t\in[0,T]}$. For every $T > 0$ and $a > 0$ and $\delta > 0$, it allows to find $\varepsilon > 0$ small enough that, for θ large, δ is a bound for the probability that this modulus of continuity exceeds $2a$. Moreover, $W_0^\theta = 0$.

The Ascoli–Arzela theorem characterizes the subsets with compact closure of $C([0, T], R)$, for the uniform norm, as being uniformly bounded and uniformly continuous. An adaptation of this result shows that here there exists a compact subset $A(\delta)$ of $C([0, T], \mathbb{R})$, and a measurable subset $B(\delta)$ of $D([0, T], \mathbb{R})$ within uniform distance δ of $A(\delta)$, such that for θ large

$$\mathbb{P}\big(\big(W_t^\theta\big)_{t\in[0,T]} \notin B(\delta)\big) \leq \delta.$$

This and an advanced result in probability theory, the Prokhorov theorem, allow to show that any sub-sequence in θ of the laws of $(W_t^\theta)_{t\in[0,T]}$ contains a further sub-sequence weakly converging to a probability measure carried by $C([0, T], \mathbb{R})$.

The result on the convergence of finite-dimensional marginals, together with the fact that these characterize the law of a process, show that the limit of every such converging sub-sequence is the law of a standard Brownian motion.

This compactness-uniqueness method proves the convergence result. For a full treatment of this important subject, see Billingsley [6].

Chapter 5
Discrete-Space Markov Processes

Abstract A rather detailed study of Markov processes with discrete state space is provided. It focuses on sample path techniques in a perspective inspired by simulation needs. The relationship of these processes with Poisson processes and with discrete-time Markov chains is shown. Rigorous constructions and results are provided for Markov process with uniformly bounded jump rates. To this end, elements of the theory of bounded operators are introduced, which explain the relation between generator and semigroup, and provide a useful framework for the forward and backward Kolmogorov equations and the Feynman–Kac formula.

Anderson [1] is a reference book on the topic, with a much wider scope than here. Asmussen and Glynn [4] provides a short primer on the theory, and many queueing examples.

5.1 Characterization, Specification, Properties

We start with some general properties of a Markov process *assuming* that it exists. Measure theory on a discrete space reduces to series summation, and we give the definitions in this framework.

5.1.1 Measures, Functions, and Transition Matrices

Positive and Signed Measures, Integrals of Functions

For $\mu = (\mu(x))_{x\in\mathscr{V}}$ such that $\mu(x) \in \mathbb{R}_+$, let

$$\mu(A) := \sum_{x\in A} \mu(x) \in \mathbb{R}_+ \cup \{\infty\} = [0,\infty], \quad A \subset \mathscr{V},$$

be defined in the sense of positive series. Such a set-function is called a *measure* on $\mathscr{V}$, and sometimes a *positive* measure for precision. The measure μ is said to be *finite* if its *total mass* $\mu(\mathscr{V})$ is finite, and to be a *probability measure*, or a *law*, if $\mu(\mathscr{V}) = 1$. The space of probability measures is denoted by $\mathscr{P} = \mathscr{P}(\mathscr{V})$.

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_5, © Springer-Verlag Berlin Heidelberg 2013

For $\mu = (\mu(x))_{x\in\mathscr{V}}$ with $\mu(x) \in \mathbb{R}$ such that $\sum_{x\in\mathscr{V}} |\mu(x)| < \infty$, let

$$\mu(A) := \sum_{x\in A} \mu(x) \in \mathbb{R}, \quad A \subset \mathscr{V},$$

be defined in the sense of absolutely convergent series. Such a set-function is called a *signed* measure on $\mathscr{V}$. The vector space of signed measures is denoted by $\mathscr{M} = \mathscr{M}(\mathscr{V})$. Among positive measures, only finite ones belong to $\mathscr{M}$.

A duality bracket between measures μ and functions f is given by,

$$(\mu, f) \mapsto \langle \mu, f \rangle := \sum_{x\in\mathscr{V}} \mu(x) f(x), \tag{5.1}$$

in the sense of positive series if μ and f are positive (wide sense), and in the sense of absolutely convergent series if μ belongs to $\mathscr{M}$ and f to the Banach space of bounded functions $L^\infty = L^\infty(\mathscr{V})$.

Markovian Matrices and Their Actions, Line and Column Vectors

In the following, classic matrix multiplication conventions are used. A positive or signed measure μ on $\mathscr{V}$ is considered as a *line* vector $(\mu(x))_{x\in\mathscr{V}}$, a function f on $\mathscr{V}$ as a *column* vector $(f(x))_{x\in\mathscr{V}}$, both of infinite length if $\mathrm{Card}(\mathscr{V}) = \infty$, and their matrix product corresponds to the above duality bracket

$$\mu f = \begin{pmatrix} \cdots & \mu(x) & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ f(x) \\ \vdots \end{pmatrix} = \sum_{x\in\mathscr{V}} \mu(x) f(x) = \langle \mu, f \rangle. \tag{5.2}$$

A matrix P is said to be a *transition* or *Markovian* matrix if

$$P = \big(P(x, y)\big)_{x,y\in\mathscr{V}}, \qquad P(x, y) \geq 0, \qquad \sum_{y\in\mathscr{V}} P(x, y) = 1,$$

i.e., if its line vectors $P(x, \cdot) = (P(x, y))_{y\in\mathscr{V}}$ constitute probability measures on $\mathscr{V}$ for all x. Multiplying a Markovian matrix P on its right by a non-negative [resp. bounded] function f yields a non-negative [resp. bounded] function Pf, on its left by a positive [resp. signed] measure μ yields a positive [resp. signed] measure μP, specifically

$$Pf : x \in \mathscr{V} \mapsto Pf(x) = P(x, \cdot) f = \sum_{y\in\mathscr{V}} P(x, y) f(y),$$

$$\mu P = \big(\mu P(y)\big)_{y\in\mathscr{V}}, \qquad \mu P(y) = \sum_{x\in\mathscr{V}} \mu(x) P(x, y),$$

and in matrix notation

$$Pf = \begin{pmatrix} \ddots & \vdots & \ddots \\ \cdots & P(x,y) & \cdots \\ \ddots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ f(y) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ Pf(x) = \sum_{y\in\mathscr{Y}} P(x,y) f(y) \\ \vdots \end{pmatrix},$$

$$\mu P = \begin{pmatrix} \cdots & \mu(x) & \cdots \end{pmatrix} \begin{pmatrix} \ddots & \vdots & \ddots \\ \cdots & P(x,y) & \cdots \\ \ddots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \cdots & \mu P(y) = \sum_{x\in\mathscr{Y}} \mu(x) P(x,y) & \cdots \end{pmatrix}.$$

Moreover, in $[0,\infty]$ [resp. $\mathbb{R}$],

$$\mu P f = \sum_{x,y\in\mathscr{Y}} \mu(x) P(x,y) f(y) = \begin{pmatrix} \cdots & \mu(x) & \cdots \end{pmatrix} \begin{pmatrix} \ddots & \vdots & \ddots \\ \cdots & P(x,y) & \cdots \\ \ddots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ f(y) \\ \vdots \end{pmatrix}.$$

Remark 5.1 A Markovian matrix $P = (P(x,y))_{x,y\in\mathscr{Y}}$ thus corresponds to an operator on L^∞ and to another on $\mathscr{M}$. The first operator has matrix P in the base in which $g \in L^\infty$ has for coordinates its values $(g(x))_{x\in\mathscr{Y}}$, the second has matrix P^*, the transpose of P, in the base in which $\nu \in \mathscr{M}$ has for coordinates its atoms $(\nu(x))_{x\in\mathscr{Y}}$. Identifying these operators to their matrices,

$$P : f \in L^\infty \mapsto Pf \in L^\infty, \qquad P^* : \mu \in \mathscr{M} \mapsto \mu P \in \mathscr{M}.$$

These operators are adjoint for the duality bracket (5.1) since

$$\langle \mu, Pf \rangle = \mu P f = \langle P^* \mu, f \rangle.$$

5.1.2 Simple and Strong Markov Property

Markov Property, Transition Matrices, and Instantaneous Laws

This book studies mainly *homogeneous* (in time) Markov processes, and not more general *inhomogeneous* Markov processes, for which in Definition 5.1 below $\mathbb{P}(X_h = y \mid X_0 = x)$ would be replaced by $\mathbb{P}(X_{t+h} = y \mid X_t = x)$.

In the sequel, the term "homogeneous" will be implicitly understood, as well as the fact that the conditional probabilities are limited to the cases when they are well defined, i.e., when the conditioning event has *non-zero* probability.

A Markov process with values in a discrete space is sometimes also called a *continuous-time Markov chain*.

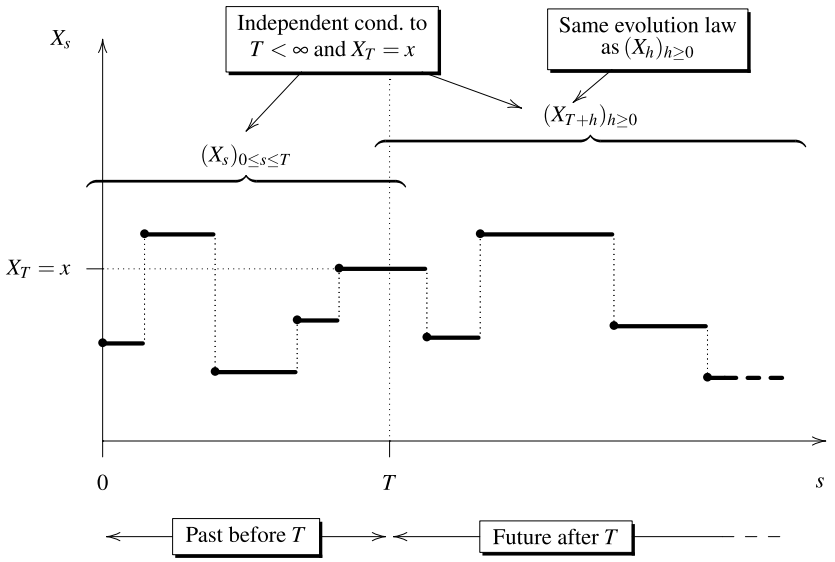

Fig. 5.1 The strong Markov property (b), where T is a stopping time. The simple Markov property corresponds to taking $T = t$, a.s.

Definition 5.1 A process $(X_t)_{t\in\mathbb{R}_+}$ with values in a discrete space $\mathscr{V}$ is a *Markov process* if, for all $n \geq 0$ and $0 \leq s_1 < \cdots < s_n < t$ and $h \geq 0$ and $x_1, \ldots, x_n$, x, y in $\mathscr{V}$,

$$\mathbb{P}(X_{t+h} = y \,|\, X_{s_1} = x_1, \ldots, X_{s_n} = x_n, X_t = x) = \mathbb{P}(X_h = y \,|\, X_0 = x).$$

Its *transition matrices* are defined for $t \in \mathbb{R}_+$ by

$$P_t = \big(P_t(x, y)\big)_{x,y\in\mathscr{V}}, \quad P_t(x, y) = \mathbb{P}(X_t = y \,|\, X_0 = x).$$

The line vector $P_t(x, \cdot)$ is the conditional law of X_t given that $X_0 = x$, as well as of X_{s+t} given that $X_s = x$ for any $s \geq 0$.

A fundamental result is that this property is equivalent to properties which seem at first stronger, and which are of more immediate use for proofs.

Theorem 5.1 (Markov property) *Let $(X_t)_{t\in\mathbb{R}_+}$ be a Markov process with values in a discrete space $\mathscr{V}$, and $(P_t)_{t\in\mathbb{R}_+}$ its transition matrices. It satisfies the following properties, which are equivalent.*

(a) *The property of Definition 5.1.*
(b) *Conditional on $X_t = x$, the process $(X_{t+h})_{h\geq 0}$ is independent of $(X_s)_{0\leq s\leq t}$ and has the same law as $(X_h)_{h\geq 0}$ for $X_0 = x$. See Fig. 5.1.*

(c) *For all* $n \geq 1$ *and* $0 = t_0 \leq t_1 < \cdots < t_n$ *and* $x_1, \ldots, x_n$ *in* $\mathscr{V}$,

$$\mathbb{P}(X_{t_1} = x_1, \ldots, X_{t_n} = x_n) = \sum_{x_0 \in \mathscr{V}} \mathbb{P}(X_0 = x_0) \prod_{k=1}^{n} P_{t_k - t_{k-1}}(x_{k-1}, x_k).$$

Proof Property (b) implies (a) as a special case. Property (a) implies that

$$\begin{aligned}
&\mathbb{P}(X_{t_1} = x_1, \ldots, X_{t_n} = x_n) \\
&\quad = \mathbb{P}(X_{t_1} = x_1, \ldots, X_{t_{n-1}} = x_{n-1}) \mathbb{P}(X_{t_n} = x_n \mid X_{t_1} = x_1, \ldots, X_{t_{n-1}} = x_{n-1}) \\
&\quad = \mathbb{P}(X_{t_1} = x_1, \ldots, X_{t_{n-1}} = x_{n-1}) P_{t_n - t_{n-1}}(x_{n-1}, x_n)
\end{aligned}$$

which implies (c) by induction. Property (c) implies that, for arbitrary $n \geq 1$ and $0 \leq s_1 < \cdots < s_n < t$ and $0 < h_1 < \cdots < h_n$ and $x_1, \ldots, x_n, x, y_1, \ldots, y_n$ in $\mathscr{V}$,

$$\begin{aligned}
&\mathbb{P}\big((X_{s_k})_{1 \leq k \leq n} = (x_k)_{1 \leq k \leq n}, X_t = x, (X_{t+h_k})_{1 \leq k \leq n} = (y_k)_{1 \leq k \leq n}\big) \\
&\quad = \sum_{x_0 \in \mathscr{V}} \mathbb{P}(X_0 = x_0) \prod_{k=1}^{n} P_{s_k - s_{k-1}}(x_{k-1}, x_k) P_{t - s_n}(x_n, x) \\
&\qquad \times P_{h_1}(x, y_1) \prod_{k=2}^{n} P_{h_k - h_{k-1}}(y_{k-1}, y_k) \\
&\quad = \mathbb{P}\big((X_{s_k})_{1 \leq k \leq n} = (x_k)_{1 \leq k \leq n}, X_t = x\big) \mathbb{P}\big((X_{h_k})_{1 \leq k \leq n} = (y_k)_{1 \leq k \leq n} \mid X_0 = x\big),
\end{aligned}$$

which implies (b). □

Remark 5.2 The law of a Markov process $(X_t)_{t \geq 0}$ is characterized by its finite-dimensional marginals, which are determined in terms of the law of X_0 and its transition matrices $(P_t)_{t \geq 0}$ by Property (c). One remaining problem is that this cannot yield existence of the process, since the $(P_t)_{t \geq 0}$ themselves are *assumed* to exist.

The notations $\mathbb{P}_x$ and $\mathbb{E}_x$ are used to indicate that $X_0 = x$, and $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ that X_0 is of law $\mu \in \mathscr{P}$. Thus

$$\mathbb{P}_\mu := \sum_{x \in \mathscr{V}} \mu(x) \mathbb{P}_x, \qquad \mathbb{E}_\mu := \sum_{x \in \mathscr{V}} \mu(x) \mathbb{E}_x.$$

For y in $\mathscr{V}$, $f \in L^\infty$ and $\mu \in \mathscr{P}$,

$$\begin{aligned}
&\mathbb{P}_x(X_t = y) := P_t(x, y), && \mathbb{P}_\mu(X_t = y) := \mu P_t(y) = P_t^* \mu(y), \\
&\mathbb{E}_x\big(f(X_t)\big) := P_t f(x) = P_t(x, \cdot) f, && \mathbb{E}_\mu\big(f(X_t)\big) := \mu P_t f.
\end{aligned}$$

The notation $\mathbb{P}$ and $\mathbb{E}$ avoids stating the law of X_0. A frequent abuse of notation is

$$\mathbb{P}(\cdot \mid X_0 = x) = \mathbb{P}_x(\cdot), \qquad \mathbb{E}(\cdot \mid X_0 = x) = \mathbb{E}_x(\cdot).$$

The laws of the X_t are called the *instantaneous laws* of the process and denoted by π_t, and π_0 is called its *initial law*. It holds that

$$\pi_t(y) := \mathbb{P}(X_t = y) \quad \text{for } y \in \mathscr{V}, \qquad \pi_t := \pi_0 P_t, \qquad \pi_t := P_t(x, \cdot) \quad \text{when } \pi_0 = \delta_x.$$

Stopping Times, Strong Markov Property

The Markov property can again be extended to stopping times, and is then called the *strong* Markov property. If it must be stressed, the property for deterministic times is called the *simple* Markov property.

Definition 5.2 A *stopping time* for the process $X = (X_t)_{t\in\mathbb{R}_+}$ is a r.v. T with values in $\mathbb{R}_+ \cup \{\infty\}$ such that for every $t \in \mathbb{R}_+$ the event $\{T \le t\}$ belongs to the σ-field $\mathscr{F}_t^X := \sigma((X_s)_{0\le s\le t})$.

Technically, this means that there exists a set of paths A_t which is measurable for the product σ-field and such that $\{T \le t\} = \{(X_t)_{0\le s\le t} \in A_t\}$.

Deterministic (non-random) times are obviously stopping times. More interesting examples are the *hitting time* and the *strict future hitting time* of $E \subset \mathscr{V}$, given by

$$T_E = \inf\{t \ge 0 : X_t \in E\}, \qquad T_E^+ = \inf\{t > 0 : X_{t-} \notin E, X_t \in E\}.$$

Exercise 5.1 Prove this; recall that $A_t \in \sigma((X_s)_{0\le s\le t})$ must be expressed in terms of a *countable* set of X_s due to the definition of the product σ-algebra.

Recall that on $\{T < \infty\}$ the r.v. X_T is defined by $X_T(\omega) = (X_{T(\omega)})(\omega)$.

Theorem 5.2 (Strong Markov property) *Let $(X_t)_{t\in\mathbb{R}_+}$ be a Markov process with values in a discrete space $\mathscr{V}$, and T be a* stopping time. *Conditional on $T < \infty$ and $X_T = x$, the process $(X_{T+h})_{h\ge 0}$ is independent of $(X_s)_{0\le s\le T}$, i.e., of T and $(X_{T\wedge s})_{s\ge 0}$, and has the same law as $(X_h)_{h\ge 0}$ for $X_0 = x$. See Fig. 5.1.*

Proof As for Theorem 4.3, we may consider that T takes values in an increasing sequence $(t_j)_{j\ge 0}$. Since

$$\{T = t_j\} = \{T \le t_j\} - \{T \le t_{j-1}\} \in \sigma\big((X_s)_{0\le s\le t_j}\big),$$

the Markov property of Theorem 5.1 5.1) applied to the $t = t_j$ implies, for all $n \in \mathbb{N}$ and $0 \le s_1 < \cdots < s_n \le t$ and $0 \le h_1 < \cdots < h_n$ and $x_1, \ldots, x_n$ and x and $y_1, \ldots, y_n$ in $\mathscr{V}$, that

$$\begin{aligned}&\mathbb{P}\big(T \le t, (X_{T\wedge s_k})_{1\le k\le n} = (x_k)_{1\le k\le n}, X_T = x, (X_{T+h_k})_{1\le k\le n} = (y_k)_{1\le k\le n}\big)\\ &= \sum_{t_j\le t} \mathbb{P}\big(T = t_j, (X_{t_j\wedge s_k})_{1\le k\le n} = (x_k)_{1\le k\le n}, X_{t_j} = x, (X_{t_j+h_k})_{1\le k\le n} = (y_k)_{1\le k\le n}\big)\end{aligned}$$

$$= \sum_{t_j \le t} \mathbb{P}\big(T = t_j, (X_{t_j \wedge s_k})_{1\le k\le n} = (x_k)_{1\le k\le n}, X_{t_j} = x\big)\mathbb{P}_x\big((X_{h_k})_{1\le k\le n} = (y_k)_{1\le k\le n}\big)$$

$$= \mathbb{P}\big(T \le t, (X_{T\wedge s_k})_{1\le k\le n} = (x_k)_{1\le k\le n}, X_T = x\big)\mathbb{P}_x\big((X_{h_k})_{1\le k\le n} = (y_k)_{1\le k\le n}\big),$$

which allows to conclude. □

Exercise 5.2 Let $(\mathscr{F}_t)_{t\ge 0}$ be a filtration, i.e., a non-decreasing family of σ-fields. It is said that $(X_t)_{t\ge 0}$ is a $(\mathscr{F}_t)_{t\ge 0}$-Markov process if

$$\mathbb{P}(X_{t+h} = y \,|\, \mathscr{F}_t) = P_h(X_t, y), \quad t, h \ge 0,\ y \in \mathscr{V},$$

and that T is a $(\mathscr{F}_t)_{t\ge 0}$ stopping time if $\{T \le t\} \in \mathscr{F}_t$ for all t. Moreover, $\mathscr{F}_T$ is defined as the σ-field constituted of the events A in $\vee_{t\ge 0}\mathscr{F}_t$ such that $A \cap \{T \le t\} \in \mathscr{F}_t$. *Prove that*: Conditional on $T < \infty$ and $X_T = x$, the process $(X_{T+h})_{h\ge 0}$ is independent of $\mathscr{F}_T$, and that

$$\mathbb{P}(X_{T+t+h} = y \,|\, T < \infty, \mathscr{F}_{T+t}) = P_h(X_{T+t}, y), \quad t, h \ge 0,\ y \in \mathscr{V}.$$

5.1.3 Semigroup, Infinitesimal Generator, and Evolution Law

Semigroup Property, Generator and Jump Instants

Using $\mathbb{P}(X_0 = x, X_t = y) = \sum_{z\in\mathscr{V}} \mathbb{P}(X_0 = x, X_t = z, X_{t+s} = y)$ and Theorem 5.1, the transition matrices P_t of a Markov process satisfy the *Chapman–Kolmogorov* equations

$$P_{t+s}(x, y) = \sum_{z\in\mathscr{V}} P_t(x, z)P_s(z, y), \quad x, y \in \mathscr{V},\ t, s \ge 0,$$

in matrix notation

$$P_{t+s} = P_t P_s, \quad t, s \ge 0.$$

Thus, $(P_t)_{t\ge 0}$ constitutes a *semigroup* for matrix multiplication, which is called the semigroup (of transition matrices) of the Markov process.

Our next theorem gathers other important structure properties of Markov processes: the time spent in each state is exponentially distributed; the pathwise construction of (X_t) is determined by the generator of its semigroup.

Theorem 5.3 *Let $(X_t)_{t\in\mathbb{R}_+}$ be a Markov process with values on a discrete space $\mathscr{V}$, and $(P_t)_{t\in\mathbb{R}_+}$ its semigroup. Let the jump instants $(T_n)_{n\ge 1}$ of the process be defined by*

$$T_n = \inf\{t > T_{n-1} : X_t \ne X_{t-}\}$$

with the conventions $T_0 = 0$ and $X_{T_n} = X_{T_{n-1}}$ if $T_n = \infty$.

(a) *Then*

$$\begin{cases} 0 = T_0 < T_1 \le T_2 \le \cdots \le \infty \quad \text{and} \lim_{n\to\infty} T_n = \infty, \\ X_t = \sum_{n\ge 0} X_{T_n} \mathbb{1}_{\{T_n \le t < T_{n+1}\}} \quad \text{for } t \ge 0. \end{cases}$$

(b) *There exists a matrix* $Q = (Q(x,y))_{x,y\in\mathscr{V}}$ *such that* $Q = \lim_{\varepsilon\to 0^+} \frac{P_\varepsilon - I}{\varepsilon}$ *in the sense that, for every* f *in* L^∞ *and* $x \in \mathscr{V}$,

$$\lim_{\varepsilon\to 0^+} \frac{P_\varepsilon f(x) - f(x)}{\varepsilon} = Qf(x) = \sum_{y\ne x} Q(x,y)\big(f(y) - f(x)\big).$$

It satisfies

$$Q(x,y) \ge 0 \quad \text{for } x \ne y, \qquad q(x) := -Q(x,x) = \sum_{y\ne x} Q(x,y) \ge 0, \tag{5.3}$$

the latter being equivalent to $\sum_{y\in\mathscr{V}} Q(x,y) = 0$, *in matrix notation* $Q1 = 0$.

(c) *Conditional on* $T_{n-1} < \infty$ *and* $X_{T_{n-1}} = x$, *the r.v.'s*

$$T_n - T_{n-1}, \quad X_{T_n}, \quad (T_k, X_{T_k})_{0\le k\le n-1},$$

are independent. If $q(x) = 0$ *then* $T_n - T_{n-1} = \infty$ *and by convention* $X_{T_n} = x$ *and* $\Pi(x,x) = 1$ *and* $\Pi(x,y) = 0$ *for* $y \ne x$. *If* $q(x) > 0$ *then*

$$\begin{cases} T_n - T_{n-1} \text{ has } \mathscr{E}\big(q(x)\big) \text{ exponential law}, \\ X_{T_n} \text{ has law } \Pi(x,\cdot) = \big(\Pi(x,y)\big)_{y\in\mathscr{V}} \quad \text{for } \Pi(x,y) = \dfrac{Q(x,y)}{q(x)} \mathbb{1}_{\{y\ne x\}}, \end{cases}$$

see Fig. 5.2. These quantities satisfy

$$q(x) \ge 0, \qquad \Pi(x,x) = \mathbb{1}_{\{q(x)=0\}}, \qquad \Pi(x,y) \ge 0, \qquad \sum_{y\in\mathscr{V}} \Pi(x,y) = 1, \tag{5.4}$$

and are related to those of (5.3) *by*

$$Q(x,y) = q(x)\Pi(x,y), \quad x \ne y \in \mathscr{V}. \tag{5.5}$$

Remark 5.3 In particular, for all $x \ne y$ in $\mathscr{V}$,

$$P_\varepsilon(x,y) = Q(x,y)\varepsilon + o(\varepsilon), \qquad P_\varepsilon(x,x) = 1 + Q(x,x)\varepsilon + o(\varepsilon),$$

and the process has probability $Q(x,y)\varepsilon + o(\varepsilon)$ to jump from x to $y \ne x$ within ε.

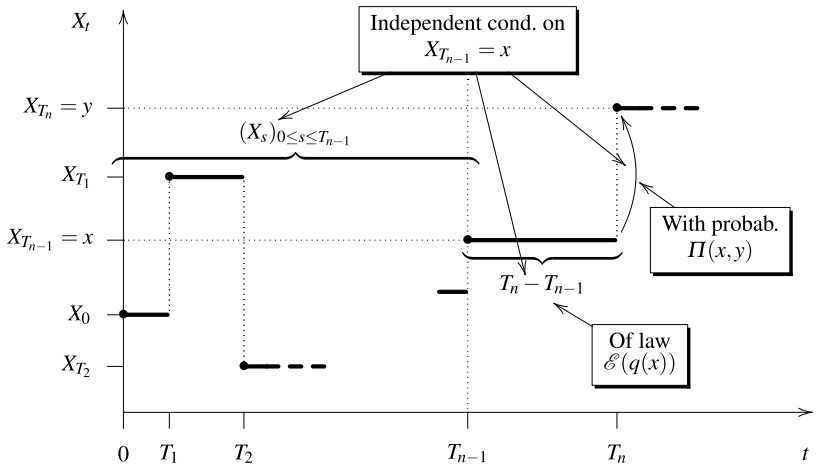

Fig. 5.2 Structure and construction of a Markov process with generator Q, with $x \neq y$ and $q(x) := \sum_{z \neq x} Q(x,z) > 0$ and $\Pi(x,y) = Q(x,y)/q(x)$

The remarkable independence properties recall Theorem 4.4. One calls:

- the matrix Q: the (*infinitesimal*) *generator* of the Markov process or of its semigroup,
- the non-negative real number $Q(x,y)$ for $y \neq x$: the *intensity* or *rate* of jumps from x to y,
- the non-negative real number $q(x)$: the *intensity* or *rate* of jumps from x; if $q(x) = 0$ then x is said to be an absorbing state,
- the matrix $\Pi = (\Pi(x,y))_{x,y \in \mathscr{V}}$: the *induced* (*transition*) *matrix*,
- the sequence $(X_{T_n})_{n \in \mathbb{N}}$ visiting the states in the same order as $(X_t)_{t \in \mathbb{R}_+}$: the *induced Markov chain*, which has transition matrix Π.

The notation Q is so common that the expression "Q-matrix for a process" or "q-matrix for a process" is often used, as in Anderson [1].

Proof (a) Since $\mathscr{V}$ is discrete, $\lim_{n \to \infty} T_n = \infty$ a.s., since else there could not be a left limit for the process sample-path at that instant.

(b) Time will be discretized by a grid of mesh $\varepsilon > 0$. With notation (4.1), for $t > 0$,

$$\left\{T_1 > \lceil t/\varepsilon \rceil \varepsilon\right\} \subset \{X_0 = X_\varepsilon = \cdots = X_{\lceil t/\varepsilon \rceil \varepsilon}\} \subset \left\{T_1 > \lceil t/\varepsilon \rceil \varepsilon\right\} \cup \{T_2 - T_1 < \varepsilon\}$$

and $\lim_{\varepsilon \to 0^+} \mathbb{P}(T_2 - T_1 < \varepsilon) = 0$ by right continuity, and hence for $x \in \mathscr{V}$, by the squeeze rule,

$$\begin{aligned} \mathbb{P}_x(T_1 > t) &= \lim_{\varepsilon \to 0^+} \mathbb{P}_x(X_0 = X_\varepsilon = \cdots = X_{\lceil t/\varepsilon \rceil \varepsilon}) \\ &= \lim_{\varepsilon \to 0^+} P_\varepsilon(x,x)^{\lceil t/\varepsilon \rceil} = \lim_{\varepsilon \to 0^+} \mathrm{e}^{\lceil t/\varepsilon \rceil \log P_\varepsilon(x,x)} \end{aligned}$$

which implies the existence of $q(x) := -Q(x,x) \in \mathbb{R}_+$ and the equalities in

$$q(x) := -\lim_{\varepsilon\to 0^+} \frac{\log P_\varepsilon(x,x)}{\varepsilon} = \lim_{\varepsilon\to 0^+} \frac{1-P_\varepsilon(x,x)}{\varepsilon}, \qquad \mathbb{P}_x(T_1 > t) = \mathrm{e}^{-q(x)t}.$$

If $q(x) = 0$ then $\mathbb{P}_x(T_1 = \infty) = 1$, and if $q(x) > 0$ then T_1 has law $\mathscr{E}(q(x))$. In the latter case, if $y \neq x$ then likewise

$$\begin{aligned}
&\bigcup_{1\le k\le \lceil t/\varepsilon\rceil} \{X_0 = X_\varepsilon = \cdots = X_{(k-1)\varepsilon} = x, X_{k\varepsilon} = y\} \\
&\quad \subset \{T_2 - T_1 < \varepsilon\} \cup \big\{T_1 \le \lceil t/\varepsilon\rceil\varepsilon, X_0 = x, X_{T_1} = y\big\} \\
&\quad \subset \{T_2 - T_1 < \varepsilon\} \cup \bigcup_{1\le k\le \lceil t/\varepsilon\rceil} \{X_0 = X_\varepsilon = \cdots = X_{(k-1)\varepsilon} = x, X_{k\varepsilon} = y\}
\end{aligned}$$

and there is existence of the limits

$$\begin{aligned}
\mathbb{P}_x(T_1 \le t, X_{T_1} = y) &= \lim_{\varepsilon\to 0^+} \sum_{1\le k\le \lceil t/\varepsilon\rceil} P_\varepsilon(x,x)^{k-1} P_\varepsilon(x,y) \\
&= \lim_{\varepsilon\to 0^+} \frac{1-P_\varepsilon(x,x)^{\lceil t/\varepsilon\rceil}}{1-P_\varepsilon(x,x)} P_\varepsilon(x,y)
\end{aligned}$$

and hence, using the previous limits, of

$$Q(x,y) := \lim_{\varepsilon\to 0^+} \frac{P_\varepsilon(x,y)}{\varepsilon}, \qquad \mathbb{P}_x(T_1 \le t, X_{T_1} = y) = \frac{\mathbb{P}_x(T_1 \le t)}{q(x)} Q(x,y).$$

This product form shows that T_1 and X_{T_1} are independent, and that, starting at x, the law $(\Pi(x,y))_{y\in\mathscr{V}}$ of X_{T_1} is given by

$$\Pi(x,y) = \frac{Q(x,y)}{q(x)}, \quad y \in \mathscr{V} - \{x\}, \quad \Pi(x,x) = 0,$$

which implies that $q(x) = \sum_{y\neq x} Q(x,y)$.

More generally, one may likewise prove for $f \in L^\infty$ that

$$\mathbb{E}_x\big(\mathbb{1}_{\{T_1\le t\}}\big(f(X_{T_1}) - f(x)\big)\big) = \mathbb{P}_x(T_1 \le t)\frac{1}{q(x)} \lim_{\varepsilon\to 0^+} \frac{P_\varepsilon f(x) - f(x)}{\varepsilon}$$

and that this limit is equal to

$$q(x)\mathbb{E}_x\big(f(X_{T_1}) - f(x)\big) = \sum_{y\neq x} Q(x,y)\big(f(y) - f(x)\big) = Qf(x).$$

(c) The strong Markov property (Theorem 5.2) allows to conclude. □

The strong Markov property applied at the first jump instant T_1 of the Markov process allows to establish certain equations on its law, in terms of its generator.

The next exercise uses this to obtain a set of equations for the semigroup called the backward Kolmogorov equations. The following one provides a probabilistic solution for some Dirichlet problems, a theme that will not be further developed in this book.

Exercise 5.3 Let $(X_t)_{t\geq 0}$ be a Markov process with values in a discrete space $\mathscr{V}$ and generator Q. Prove that the semigroup $(P_t)_{t\geq 0}$ satisfies

$$P_t(x,y) = \mathbb{1}_{\{x=y\}} \mathrm{e}^{-q(x)t} + \int_0^t \mathrm{e}^{-q(x)s} \sum_{z\neq x} Q(x,z) P_{t-s}(z,y)\, ds, \quad x,y \in \mathscr{V}.$$

Prove that this system of equations is equivalent to the system

$$\frac{d}{dt} P_t(x,y) = \sum_{z\in\mathscr{V}} Q(x,z) P_t(z,y), \quad x,y \in \mathscr{V}.$$

Exercise 5.4 Let $(X_t)_{t\geq 0}$ be a Markov process with values in a discrete space $\mathscr{V}$ and generator Q. Let $A \subset \mathscr{V}$ and $T_A = \inf\{t \geq 0 : X_t \in A\}$. Prove that, for any $\lambda \geq 0$ and non-negative or bounded function $f : A \to \mathbb{R}$, the function

$$u_\lambda^f : x \in \mathscr{V} \mapsto u_\lambda^f(x) = \mathbb{E}_x\left(\mathrm{e}^{-\lambda T_A} f(X_{T_A}) \mathbb{1}_{\{T_A<\infty\}}\right)$$

is a solution to the Dirichlet problem, in which $u : \mathscr{V} \to \mathbb{R}$ is a function,

$$u = f \quad \text{on } A, \qquad (\lambda I - Q)u = 0 \quad \text{on } \mathscr{V} - A.$$

Assuming you knew how to simulate $(X_t)_{t\geq 0}$, propose a Monte Carlo method yielding approximate solutions to this equation, and discuss the natural issues involved.

5.2 Constructions, Existence, Simulation, Equations

We now provide effective construction and simulation techniques for a Markov process, starting from its generator.

5.2.1 Fundamental Constructions

According to Theorem 5.3, the random evolution of a Markov process $(X_t)_{t\geq 0}$ can be characterized and described in terms of

- its generator $Q = (Q(x,y))_{x,y\in\mathscr{V}}$ satisfying (5.3),
- or its rates $(q(x))_{x\in\mathscr{V}}$ and induced transition matrix $\Pi = (\Pi(x,y))_{x,y\in\mathscr{V}}$ satisfying (5.4),

see also Fig. 5.2. These quantities are related by formula (5.5).

Conversely, Theorem 5.3 is key to try to *construct* and *simulate* a Markov process $(X_t)_{t\geq 0}$ starting from such quantities. In the following constructions, draws are always assumed to be independent.

Algorithm (First construction) Draw a sample x_0 for the initial value; then iteratively, if the process is in state x:

- draw a sample from an $\mathscr{E}(q(x))$ exponential law for the sojourn duration at x,
- draw the next state $y \neq x$ from the law $\Pi(x, \cdot)$.

(If $q(x) = 0$ then the process is absorbed at x.)

This construction can be used for simulation. The results on superposition-decomposition (Theorem 4.4 and its Corollary 4.1) provide constructions which are equivalent in law, and are useful for theoretical comprehension and results, but which must be avoided for simulation.

Algorithm (Second construction) Draw a sample x_0 for the initial value; then iteratively, if the process is in state x:

- draw samples from the $\mathscr{E}(Q(x, y))$ exponential laws for $y \neq x$, and determine their minimum, which is the sojourn duration of the process at x,
- the next state $y' \neq x$ is determined by the fact that the sample from $\mathscr{E}(Q(x, y'))$ achieves this infinimum.

(If $Q(x, y) = 0$ for all $y \neq x$, then the process is absorbed at x.)

The following interpretations of these two constructions define globally the process $(X_t)_{t\in\mathbb{R}_+}$ in terms of Poisson processes and sequences of i.i.d. r.v.'s.

Algorithm (First construction, bis) For every x in $\mathscr{V}$, consider a Poisson process N^x with intensity $q(x)$ and jump instants $(T_n^x)_{n\geq 1}$, and a sequence $(Y_n^x)_{n\geq 1}$ of i.i.d. r.v.'s with laws $\Pi(x, \cdot)$, independently. Draw a sample x_0 for the initial value; then iteratively, if the process reaches state x, let it stay there until the next jump instant of N^x, and if this is T_m^x then let the process jump to Y_m^x.

Algorithm (Second construction, bis) For every $x \neq y$ in $\mathscr{V}$, consider a Poisson process $N^{x,y}$ with intensity $Q(x, y)$, independently. Draw a sample x_0 for the initial value; then iteratively, if the process reaches state x, let it stay there until the first (least) of the next jump instants of the $N^{x,y}$ for $y \neq x$, and if this is a jump of $N^{x,y'}$ then let the process jump to y'.

Interpreting a random evolution in terms of Theorem 4.4 and of one of these constructions often allows to show that the underlying process is Markov and to identify its generator Q, or equivalently its jump intensities $q(x)$ and its induced transition matrix Π.

The above exponential random variables, or Poisson processes, are often referred to as exponential, or Poisson, clocks.

5.2.2 Explosion or Existence for a Markov Process

These constructions are all equivalent in law. They define an increasing sequence constituted of $T_0 = 0$ and the jump instants $(T_n)_{n\geq 1}$, between which the process is constant, and a Markov chain $(X_{T_n})_{n\in\mathbb{N}}$ with matrix Π giving the successive distinct values taken by the process. This allows to define X_t for $0 \leq t < \lim_{n\to\infty} T_n$.

Theorem 5.4 *Let $Q = (Q(x, y))_{x,y\in\mathscr{Y}}$ be a matrix satisfying (5.3). If any of the above constructions (equivalent in law) is such that*

$$\lim_{n\to\infty} T_n = \infty, \quad a.s.,$$

then the resulting process $(X_t)_{t\in\mathbb{R}_+}$ is Markov and has generator Q.

Proof The "Second construction, bis" is used. The resulting process $(X_t)_{t\in\mathbb{R}_+}$ can readily be shown to be Markov using the Markov property of the Poisson processes $N^{x,y}$ (Sect. 4.2.2). The evolution description in Theorem 5.3 allows to identify Q as its generator. □

The following result provides a criterion for this abstract existence result.

Theorem 5.5 *Let $Q = (Q(x, y))_{x,y\in\mathscr{Y}}$ be a matrix satisfying (5.3). For any of the above constructions, a.s.,*

$$\lim_{n\to\infty} T_n = \infty \quad \textit{if and only if} \quad \sum_{n\in\mathbb{N}} \frac{1}{q(X_{T_n})} = \infty.$$

Proof Conditionally on $(X_{T_k})_{k\in\mathbb{N}}$ the $T_{n+1} - T_n$ for $n \geq 0$ are independent exponential r.v.'s of parameters $q(X_{T_n})$ (Theorem 5.3). Moreover

$$\lim_{n\to\infty} T_n = \sum_{n\in\mathbb{N}} (T_{n+1} - T_n) \in [0, \infty], \qquad A := \left\{ \sum_{n\in\mathbb{N}} \frac{1}{q(X_{T_n})} = \infty \right\} \in \sigma\big((X_{T_k})_{k\in\mathbb{N}}\big).$$

On the one hand, using the Lebesgue Monotone Convergence Theorem or the Fubini theorem,

$$\mathbb{E}\Big(\lim_{n\to\infty} T_n \,|\, (X_{T_k})_{k\in\mathbb{N}}\Big) = \sum_{n\in\mathbb{N}} \mathbb{E}\big(T_{n+1} - T_n \,|\, (X_{T_k})_{k\in\mathbb{N}}\big) = \sum_{n\in\mathbb{N}} \frac{1}{q(X_{T_n})}.$$

This is finite on $A^c := \Omega - A$, hence

$$\mathbb{E}\Big(\mathbb{1}_{A^c} \lim_{n\to\infty} T_n \,|\, (X_{T_k})_{k\in\mathbb{N}}\Big) = \mathbb{1}_{A^c} \mathbb{E}\Big(\lim_{n\to\infty} T_n \,|\, (X_{T_k})_{k\in\mathbb{N}}\Big) < \infty$$

which implies

$$\mathbb{P}\Big(\mathbb{1}_{A^c} \lim_{n\to\infty} T_n < \infty \,|\, (X_{T_k})_{k\in\mathbb{N}}\Big) = 1.$$

The fact that

$$\lim_{n\to\infty} T_n < \infty \quad \text{on } A^c, \quad \text{a.s.},$$

follows by taking expectations.

On the other hand, using the Lebesgue Dominated Convergence Theorem,

$$\mathbb{E}\big(\mathrm{e}^{-\lim_{n\to\infty} T_n} \,|\, (X_{T_k})_{k\in\mathbb{N}}\big) = \mathbb{E}\bigg(\prod_{n\in\mathbb{N}} \mathrm{e}^{-(T_{n+1}-T_n)} \,|\, (X_{T_k})_{k\in\mathbb{N}}\bigg) = \prod_{n\in\mathbb{N}} \frac{q(X_{T_n})}{1+q(X_{T_n})}.$$

This vanishes on A since

$$\prod_{n\in\mathbb{N}} \frac{q(X_{T_n})}{1+q(X_{T_n})} = \prod_{n\in\mathbb{N}} \bigg(1+\frac{1}{q(X_{T_n})}\bigg)^{-1} \le \bigg(1+\sum_{n\in\mathbb{N}} \frac{1}{q(X_{T_n})}\bigg)^{-1},$$

hence

$$\mathbb{E}\big(\mathbb{1}_A \mathrm{e}^{-\lim_{n\to\infty} T_n} \,|\, (X_{T_k})_{k\in\mathbb{N}}\big) = \mathbb{1}_A \mathbb{E}\big(\mathrm{e}^{-\lim_{n\to\infty} T_n} \,|\, (X_{T_k})_{k\in\mathbb{N}}\big) = 0.$$

The fact that

$$\lim_{n\to\infty} T_n < \infty \quad \text{on } A, \quad \text{a.s.},$$

follows by taking expectations. □

This yields the two following existence results, under verifiable assumptions which are *always* true if $\mathscr{V}$ is *finite*. The first one uses only the jump intensities.

Theorem 5.6 *Let $Q = (Q(x,y))_{x,y\in\mathscr{V}}$ be a matrix satisfying (5.3). If*

$$\sup_{x\in\mathscr{V}} q(x) < \infty$$

then any of the above constructions is such that $\lim_{n\to\infty} T_n = \infty$ a.s., and yields a Markov process $(X_t)_{t\in\mathbb{R}_+}$ with generator Q.

Note that this is always true if $\mathscr{V}$ is finite.

Proof This follows from

$$\sum_{n\in\mathbb{N}} \frac{1}{q(X_{T_n})} \ge \sum_{n\in\mathbb{N}} \frac{1}{\sup_{x\in\mathscr{V}} q(x)} = \infty$$

and Theorems 5.5 and 5.4. □

The second one uses only the induced transition matrix Π, through the induced Markov chain $(X_{T_n})_{n\in\mathbb{N}}$ it generates. It uses a recurrence assumption, for which the theory of discrete time Markov chains gives several verifiable criteria.

Theorem 5.7 *Let $Q = (Q(x, y))_{x,y\in\mathscr{V}}$ be a matrix satisfying (5.3), and consider one of the above constructions. If there exists a random finite subset F of $\mathscr{V}$ such that*

$$\sum_{n\in\mathbb{N}} \mathbb{1}_{\{X_{T_n}\in F\}} = \infty \quad a.s.,$$

then $\lim_{n\to\infty} T_n = \infty$ a.s., and the construction yields a Markov process $(X_t)_{t\in\mathbb{R}_+}$ with generator Q.

Proof If $\lim_{n\to\infty} T_n < \infty$ then Theorem 5.5 implies that $\sum_{n\in\mathbb{N}} \frac{1}{q(X_{T_n})} < \infty$ and hence $\lim_{n\to\infty} q(X_{T_n}) = \infty$, and thus any finite subset F may only contain X_{T_n} for a finite number of n since $q(F)$ being finite is upper-bounded. □

As a consequence, if $T_\infty := \lim_n T_n < \infty$ then X_t remains outside of any finite set F for large enough $t < T_\infty$ (random, depending on F), which explains why it is said that there is *explosion* at time T_∞. The existence condition "$\lim_n T_n = \infty$" a.s. is often referred to as "non-explosion".

5.2.3 Fundamental Simulation, Fictitious Jump Method

Section 4.3 contains many useful facts for the sequel.

Fundamental Simulation

The "First construction" in Sect. 5.2.1 can be used to simulate a Markov process taking values in a discrete space. This natural method relies on the fundamental description in Theorem 5.3, see Fig. 5.2, and uses the jump rates $q(x)$ and induced transition matrix Π, which can be obtained from the generator Q by

$$q(x) = \sum_{y\neq x} Q(x, y), \quad Q(x, y) = q(x)\Pi(x, y), \quad x \neq y \in \mathscr{V}.$$

This method simulates recursively the jumps of the process. In an actual code, this is obtained by a loop.

Algorithm (Fundamental simulation) Draw a sample x_0 for the initial value; then iteratively, if the process is in state x:

- compute the intensity $q(x)$, and draw a sample from the $\mathscr{E}(q(x))$ exponential law for the duration that the process stays at x,
- compute $\Pi(x, y) = Q(x, y)/q(x)$ for $y \neq x$, and a draw a sample from this discrete law for the next state of the process.

(If $q(x) = 0$ then the process is absorbed at x.)

It is sometimes possible to perform the latter draw in steps without having to compute the $\Pi(x, y)$ for all $y \neq x$ beforehand, by regrouping the y into appropriate subsets.

To avoid repetitions, in a preliminary phase one can decide to compute systematically the $q(x)$, and perhaps even the $\Pi(x, \cdot)$, for all $x \in \mathscr{V}$, and store them in memory to be accessed as needed at each iteration. This may be much more efficient when computations are lengthy, but is possible only for small sample spaces.

Fictitious Jump Method

Here is a family of variant methods of construction, which may be useful for simulation. If $\hat{q}(x) \geq q(x)$ for all $x \in \mathscr{V}$ then the matrix $\hat{\Pi}$ defined by

$$\hat{\Pi}(x, y) = \frac{q(x)}{\hat{q}(x)} \Pi(x, y) \quad \text{for } y \neq x \in \mathscr{V}, \qquad \hat{\Pi}(x, x) = 1 - \frac{q(x)}{\hat{q}(x)}, \tag{5.6}$$

is Markovian. Note that (5.5) can be generalized to

$$Q(x, y) = q(x)\Pi(x, y) = \hat{q}(x)\hat{\Pi}(x, y), \quad x \neq y \in \mathscr{V}.$$

A special interesting case is when

$$\hat{q}(x) = \lambda \geq \sup_{y \in \mathscr{V}} q(y), \quad x \in \mathscr{V},$$

particularly so when the inequality is an equality.

It is then possible to construct $(X_t)_{t \in \mathbb{R}_+}$ similarly to the "First construction" in Sect. 5.2.1, but using $\hat{q}$ and $\hat{\Pi}$ instead of q and Π. In the second step, the draw according to $\hat{\Pi}$ is decomposed into two separate steps, as would be the case in actual code. All draws are independent.

Algorithm (Fictitious jump method) Draw a sample x_0 for the initial value; then iteratively, if the process is in state x:

- compute the intensity $\hat{q}(x)$, and draw a sample from the $\mathscr{E}(\hat{q}(x))$ exponential law for a duration after which:
- – either, with probability $1 - q(x)/\hat{q}(x)$, the process remains at x,
  – or else (with complementary probability $q(x)/\hat{q}(x)$), the process leaves x; compute $\Pi(x, y) = Q(x, y)/q(x)$ for $y \neq x$, and draw a sample from this discrete law for the next state.

(If $\hat{q}(x) = 0$ then the process is absorbed at x.)

This method can be advantageous for simulation, when there are good upper bounds $\hat{q}(x)$ for the $q(x)$ and the former are much easier to compute than the

latter; a natural choice is to take $\hat{q}(x)$ constant in x, so as not to have to compute it at each actual jump of the process. We shall see an example at the end of Sect. 5.2.6.

The sampling result in Theorem 4.4(b), and a simple conditioning show that the fictitious jump method is equivalent in law to the "First construction". Its name comes from the fact that it provides too many "potential" jump instants for the process, some of which are then rejected (and are thus "fictitious"). This is reminiscent of the rejection method in Sect. 2.2.5, but the justification goes much deeper.

The results on superposition-decomposition (Theorem 4.4 and its Corollary 4.1) provide constructions which are equivalent in law. These must be avoided for simulation, but may be useful for theoretical comprehension and results.

For instance, a proof of Theorem 5.6 can be thus obtained, by embedding the jump instants among those of a Poisson process with intensity $\lambda = \sup_{x\in\mathscr{V}} q(x)$ (which are known not to accumulate).

5.2.4 Kolmogorov Equations, Feynman–Kac Formula

Theorem 5.3 has extensions in terms of linear differential equations, in infinite dimensions for infinite $\mathscr{V}$. Under the assumptions of Theorem 5.6, which ensures the existence of a Markov process of generator Q and of its semigroup, results can be stated and proved without too many complications.

Theorem 5.8 *Let $Q = (Q(x,y))_{x,y\in\mathscr{V}}$ be a matrix satisfying (5.3) and*

$$\sup_{x\in\mathscr{V}} q(x) < \infty.$$

Let $(P_t)_{t\geq 0}$ be the semigroup of the corresponding Markov process $(X_t)_{t\in\mathbb{R}_+}$.

(a) *The semigroup $(P_t)_{t\geq 0}$ is the unique solution of the* backward Kolmogorov *equation*

$$\begin{cases} \dfrac{d}{dt}P_t = QP_t, \quad i.e., \quad \dfrac{d}{dt}P_t(x,y) = \displaystyle\sum_{z\in\mathscr{V}} Q(x,z)P_t(z,y), \quad x,y\in\mathscr{V}, \\ P_0 = I. \end{cases}$$

In L^∞, for $f\in L^\infty$, the $v_t : x \mapsto P_t f(x) = \mathbb{E}_x(f(X_t))$ for $t\geq 0$ constitute the unique solution of the backward Kolmogorov *equation*

$$\begin{cases} \dfrac{d}{dt}v_t = Qv_t, \quad i.e., \quad \dfrac{d}{dt}v_t(x) = \displaystyle\sum_{z\in\mathscr{V}} Q(x,y)v_t(y), \quad x\in\mathscr{V}, \\ v_0 = f. \end{cases}$$

(b) *The semigroup $(P_t)_{t\geq 0}$ is the unique solution of the* forward Kolmogorov *equation*

$$\begin{cases} \dfrac{d}{dt}P_t = P_t Q, \quad \textit{i.e.}, \quad \dfrac{d}{dt}P_t(x,y) = \displaystyle\sum_{z\in\mathscr{V}} P_t(x,z)Q(z,y), \quad x,y\in\mathscr{V}, \\ P_0 = I. \end{cases}$$

In $\mathscr{M}$, for any initial law π, the instantaneous laws $\pi_t = \pi P_t$ for $t \geq 0$ constitute the unique solution of the forward Kolmogorov *equation* (*or* Fokker–Planck *equation*)

$$\begin{cases} \dfrac{d}{dt}\pi_t = \pi_t Q, \quad \textit{i.e.}, \quad \dfrac{d}{dt}\pi_t(y) = \displaystyle\sum_{x\in\mathscr{V}} \pi_t(x)Q(x,y), \quad y\in\mathscr{V}, \\ \pi_0 = \pi. \end{cases}$$

This equation with initial condition $\mu \in \mathscr{M}$ has unique solution $(\mu P_t)_{t\geq 0}$.

The proof of a more precise statement will be given at the end of the following subsection. When $\sup_{x\in\mathscr{V}} q(x) = \infty$, this theorem remains valid in a weaker sense, statements and proofs are more delicate, and the equations must usually be interpreted termwise. The backward equation is always true, see Exercise 5.3, but not necessarily the forward equation. For all this we refer to Anderson [1].

The terms "backward" and "forward" come from the fact that Q acts on $P_t = (P_t(x,y))_{x,y\in\mathscr{V}}$ respectively on the "departure" state x and the "arrival" state y. In the proof, this will correspond respectively to a differentiation performed respectively "at the beginning" and "at the end" of the path.

Hence, the backward equation is studied and interpreted often in reversed time, as in the following result for a terminal-value equation.

Corollary 5.1 (Feynman–Kac formula) *Let the assumptions and notation of Theorem 5.8 hold. For $T > 0$ and $f \in L^\infty$, the backward equation*

$$\begin{cases} \dfrac{d}{dt}u_t + Qu_t = 0, \quad t\in[0,T], \\ u_T = f, \end{cases}$$

has a unique differentiable solution $t \in [0,T] \mapsto u_t \in L^\infty$, given by

$$x \mapsto u_t(x) = \mathbb{E}\big(f(X_T) \,|\, X_t = x\big) = P_{T-t}f(x).$$

Proof The time change $t \mapsto T - t$ on $[0,T]$ yields a bijection between the solutions of this equation and those of the backward Kolmogorov equation in L^∞. □

A differential equation (possibly of infinite dimension) arising in applications may have the requisite structure to be interpreted as a forward or backward equation

for a discrete-space Markov process. It can thus be given a *probabilistic representation*. This allows its approximate solution by Monte Carlo methods.

Such a probabilistic representation can also be useful for theoretical results, such as those in the corollary.

Forward Kolmogorov Equation as Balance Equation

The forward equation on $\mathscr{M}$ can be written

$$\frac{d}{dt}\pi_t = \sum_{x \neq y} \pi_t(x) Q(x, y) - \pi_t(y) q(y), \quad y \in \mathscr{V}$$

and has the following interpretation as a balance equation. The instantaneous variation of π_t is due to mass exchanges, and for any state y:

- mass flows into y from each state $x \neq y$ at rate $\pi_t(x)$ (probability for the process to be at x) times $Q(x, y)$ (jump rate from x to y);
- mass flows out of state y at rate $\pi_t(y)$ (probability for the process to be at y) times $q(y) = -Q(x, x) = \sum_{x \neq y} Q(x, y)$ (jump rate from y).

Generators and Martingales

Modern Markov process theory goes much further than these equations, and exploits strong links between Markov processes and martingales, established using the generator, which extend the Kolmogorov equations. Problem 5.3 provides a short initiation to this topic.

5.2.5 Generators and Semigroups in Bounded Operator Algebras

This section, which can be skipped in a first reading, introduces some rudiments about algebras of bounded operators on a Banach space, which will allow to better understand the relation between semigroup and generator, and prove Theorem 5.8 in a more precise form.

This relies heavily on the boundedness assumption on the rates in the theorem, but allows for straightforward proofs once the set-up is perfected. We do not insist further than necessary for our purposes on this matter.

Anderson [1] develops the general theory of Markovian semigroups and their generators without this boundedness assumption.

Measure-Function Duality, Total Variation Norm

The integral by a measure μ of a function f given in (5.1), see also (5.2), provides a natural duality bracket between $\mathscr{M}$ and the Banach space L^∞ of bounded functions with the uniform norm, and $\mathscr{M}$ with the dual norm

$$\|\mu\|_{\mathrm{VT}} := \sup_{\|f\|_\infty \le 1} \langle \mu, f \rangle = \sup_{\|f\|_\infty \le 1} \sum_{x \in \mathscr{V}} \mu(x) f(x), \quad \mu \in \mathscr{M}, \tag{5.7}$$

called the *total variation norm*, is a Banach space. It is straightforward to check that

$$\|\mu\|_{\mathrm{VT}} = \sum_{x \in \mathscr{V}} \big|\mu(x)\big|, \quad \mu \in \mathscr{M}. \tag{5.8}$$

We could identify $\mathscr{M}$ to the space $\ell^1(\mathscr{V})$ of summable sequences, and L^∞ to the space $\ell^\infty(\mathscr{V})$ of bounded sequences, but avoid this for the sake of clarity.

The space $\mathscr{P}$ of probability measures is the intersection of the unit sphere of $\mathscr{M}$ for the total variation norm and of the cone of positive measures, and hence is a closed subset of $\mathscr{M}$. The norm induces a complete metric, and it is straightforward to check that

$$\|\mu - \nu\|_{\mathrm{VT}} = 2 \sup_{A \subset \mathscr{V}} \big(\mu(A) - \nu(A)\big), \quad \mu, \nu \in \mathscr{P}.$$

Due to this and other reasons, some authors introduce a factor $1/2$ in the definition.

Remark 5.4 The duality bracket (5.1) allows to identify the topological dual of $\mathscr{M}$ with L^∞. The topological dual of L^∞ is constituted of the *finitely* additive set functions, and contains strictly $\mathscr{M}$ when $\mathscr{V}$ is *infinite*. Nevertheless, it is straightforward to prove that

$$\|f\|_\infty = \sup_{\mu \in \mathscr{M}, \|\mu\|_{\mathrm{VT}} \le 1} \langle \mu, f \rangle. \tag{5.9}$$

Exercise 5.5 Prove the various formulas given for $\|\cdot\|_{\mathrm{VT}}$ and $\|\cdot\|_\infty$.

Bounded Operator Algebras, Generators with Bounded Jump Rates

The operator norm of an operator A on a Banach space V with norm $\|\cdot\|$ is defined by

$$\|A\|_{\mathrm{op}} := \sup_{\|v\| \le 1} \|Av\|$$

(the supremum is over all $v \in V$ satisfying $\|v\| \le 1$). If A and B are two operators such that the composed operator AB is defined, then clearly

$$\|AB\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}}. \tag{5.10}$$

An operator is said to be *bounded* if its operator norm is finite, and the set of all bounded operators on a given Banach space constitutes a *Banach algebra*, i.e., a Banach space which is an associative algebra for a multiplication (here operator composition) satisfying the bound (5.10).

In the sequel, the notation $\|\cdot\|_{\mathrm{op}}$ will be used for the operator norm on various Banach spaces, and care will be taken to specify the space under interest.

In an extension of Remark 5.1, any real matrix $A = (A(x,y))_{x,y\in\mathscr{V}}$ satisfying

$$\sup_{x\in\mathscr{V}} \sum_{y\in\mathscr{V}} \big|A(x,y)\big| < \infty \tag{5.11}$$

corresponds to operators (identified with their matrices, see the remark)

$$A : f \in L^\infty \mapsto Af \in L^\infty, \qquad A^* : \mu \in \mathscr{M} \mapsto \mu A \in \mathscr{M}, \tag{5.12}$$

which are adjoint for the duality bracket (5.1) since

$$\langle \mu, Af\rangle = \mu Af = \big\langle A^*\mu, f\big\rangle, \quad \mu \in \mathscr{M},\ f \in L^\infty. \tag{5.13}$$

(Some bounded operators on L^∞ may not be of this form, see Remark 5.4.)

For any $x \in \mathscr{V}$ it holds that $Af(x) = \langle A(x,\cdot), f\rangle$ for $A(x,\cdot) = (A(x,y))_{y\in\mathscr{V}}$ in $\mathscr{M}$, and using $|Af(x)| = \max\{Af(x), A(-f)(x)\}$ and (5.7), (5.8) and (5.11),

$$\begin{aligned}\|A\|_{\mathrm{op}} &:= \sup_{\|f\|_\infty\le 1} \|Af\|_\infty = \sup_{x\in\mathscr{V}} \sup_{\|f\|_\infty\le 1} Af(x)\\ &= \sup_{x\in\mathscr{V}} \big\|A(x,\cdot)\big\|_{\mathrm{VT}} = \sup_{x\in\mathscr{V}} \sum_{y\in\mathscr{V}} \big|A(x,y)\big| < \infty\end{aligned}$$

and (5.7), (5.13) and (5.9) yield

$$\big\|A^*\big\|_{\mathrm{op}} := \sup_{\|\mu\|_{\mathrm{VT}}\le 1} \big\|A^*\mu\big\|_{\mathrm{VT}} = \sup_{\|\mu\|_{\mathrm{VT}}\le 1} \sup_{\|f\|_\infty\le 1} \mu Af = \sup_{\|f\|_\infty\le 1} \|Af\|_\infty := \|A\|_{\mathrm{op}} \tag{5.14}$$

so that A and A^* have the same bounded operator norm.

The transition semigroup $(P_t)_{t\ge 0}$ of a Markov process is constituted of Markovian matrices which enter in this framework. A matrix Q satisfying (5.3) enters this framework if and only if it satisfies

$$\|Q\|_{\mathrm{op}} := 2 \sup_{x\in\mathscr{V}} q(x) < \infty, \tag{5.15}$$

i.e., the assumption in Theorems 5.6 and 5.8.

Proof of a Generalization of Theorem 5.8

Lemma 5.1 *Let the assumptions and notation of Theorem 5.8 hold. In the Banach algebra of operators on L^∞ with the operator norm,*

$$\lim_{\varepsilon\to 0^+} \frac{P_\varepsilon - I}{\varepsilon} = Q, \quad \textit{and more generally} \quad \frac{d}{dt} P_t = Q P_t = P_t Q, \quad t \ge 0.$$

In the Banach algebra of operators on $\mathscr{M}$, the analogous results for the adjoint operators hold.

Proof Let us construct a Markov process $(X_t)_{t\geq 0}$ with generator Q using the "Fictitious jump method" of Sect. 5.2.3, with jump instants $(T_n)_{n\geq 1}$ chosen among those of a Poisson process $(N_t)_{t\geq 0}$ of intensity

$$\lambda := \sup_{x\in\mathscr{V}} q(x) = \frac{1}{2}\|Q\|_{\mathrm{op}} < \infty.$$

Let us first prove that its semigroup $(P_t)_{t\geq 0}$ is continuous. For $x \in \mathscr{V}$ and $0 \leq s \leq t$ and $f \in L^\infty$ such that $\|f\|_\infty \leq 1$,

$$(P_t - P_s)f(x) = \mathbb{E}_x\big(f(X_t) - f(X_s)\big) \leq 2\mathbb{P}(N_t - N_s \geq 1) = 2\big(1 - \mathrm{e}^{-\lambda(t-s)}\big)$$

so that, in the bounded operator algebra of L^∞,

$$\|P_t - P_s\|_{\mathrm{op}} = \sup_{\|f\|_\infty\leq 1} \sup_{x\in\mathscr{V}} (P_t - P_s)f(x) \leq 2\big(1 - \mathrm{e}^{-\lambda(t-s)}\big) \leq 2\lambda(t-s).$$

Let us now prove differentiability at $t = 0$. For $x \in \mathscr{V}$ and $\varepsilon > 0$ and $f \in L^\infty$ such that $\|f\|_\infty \leq 1$,

$$\begin{aligned}(P_\varepsilon - I)f(x) &= \mathbb{E}_x\big(f(X_\varepsilon) - f(x)\big)\\ &= \mathbb{E}_x\big(\big(f(X_{T_1}) - f(x)\big)\mathbb{1}_{\{T_1\leq\varepsilon\}}\big) + \mathbb{E}_x\big(\big(f(X_\varepsilon) - f(X_{T_1})\big)\mathbb{1}_{\{T_2\leq\varepsilon\}}\big)\end{aligned}$$

in which the independence properties and $Q(x, y) = q(x)\Pi(x, y)$ imply that

$$\begin{aligned}\mathbb{E}_x\big(\big(f(X_{T_1}) - f(x)\big)\mathbb{1}_{\{T_1\leq\varepsilon\}}\big) &= \big(1 - \mathrm{e}^{-q(x)\varepsilon}\big)\sum_{y\neq x}\Pi(x, y)\big(f(y) - f(x)\big)\\ &= \varepsilon Qf(x) + \mathscr{O}\big(\lambda^2\varepsilon^2\big)\end{aligned}$$

and moreover (4.2) implies that

$$\mathbb{E}_x\big(\big(f(X_\varepsilon) - f(X_{T_1})\big)\mathbb{1}_{\{T_2\leq\varepsilon\}}\big) \leq 2\mathbb{P}(N_\varepsilon \geq 2) = 2\big(1 - \mathrm{e}^{-\lambda\varepsilon} - \lambda\varepsilon\mathrm{e}^{-\lambda\varepsilon}\big) \leq \lambda^2\varepsilon^2.$$

Hence,

$$\|P_\varepsilon - I - \varepsilon Q\|_{\mathrm{op}} = \sup_{\|f\|_\infty\leq 1} \sup_{x\in\mathscr{V}} (P_\varepsilon - I - \varepsilon Q)f(x) = \mathscr{O}\big(\lambda^2\varepsilon^2\big)$$

and thus $\lim_{\varepsilon\to 0^+} \frac{P_\varepsilon - I}{\varepsilon} = Q$ for the operator norm for L^∞.

The semigroup property and the preceding results yield

$$\frac{P_{t+\varepsilon} - P_t}{\varepsilon} = \begin{cases}\frac{P_\varepsilon - I}{\varepsilon}P_t \xrightarrow[\varepsilon\to 0^+]{} QP_t,\\ P_t\frac{P_\varepsilon - I}{\varepsilon} \xrightarrow[\varepsilon\to 0^+]{} P_tQ,\end{cases} \qquad \frac{P_{t-\varepsilon} - P_t}{-\varepsilon} = \begin{cases}\frac{I - P_\varepsilon}{-\varepsilon}P_{t-\varepsilon} \xrightarrow[\varepsilon\to 0^+]{} QP_t,\\ P_{t-\varepsilon}\frac{I - P_\varepsilon}{-\varepsilon} \xrightarrow[\varepsilon\to 0^+]{} P_tQ,\end{cases}$$

in the bounded operator algebra for L^∞, (where (5.10) allows to take limits of products), and this concludes the proof in this setting.

The corresponding results for the adjoint operators follow using (5.14). □

Completion of the Proof of Theorem 5.8

Lemma 5.1 shows that $(P_t)_{t\geq 0}$ solves the differential equations $\frac{d}{dt}P_t = QP_t$ and $\frac{d}{dt}P_t = P_tQ$ given in Theorem 5.8, each of these both in the Banach algebra of bounded operators on L^∞ and in the Banach algebra of bounded operators on $\mathscr{M}$.

Using (5.10), the operators $P \mapsto QP$ and $P \mapsto PQ$ on these algebras are bounded with operator norm $\|Q\|_{\mathrm{op}}$, and the Gronwall lemma yields uniqueness of solution for these equations.

Multiplication on the right by $f \in L^\infty$ of the backward equation $\frac{d}{dt}P_t = QP_t$ for operators on L^∞ yields that $(P_t f)_{t\geq 0}$ solves the backward equation on L^∞. Multiplication on the left by $\mu \in \mathscr{M}$ of the forward equation $\frac{d}{dt}P_t = P_tQ$ for operators on $\mathscr{M}$ yields that $(\mu P_t)_{t\geq 0}$ solves the forward equation on $\mathscr{M}$.

Uniqueness for these equations follows from the fact that Q is a bounded operator and the Gronwall lemma.

Operator Exponential

Under the assumptions of Theorem 5.8, one can define

$$\mathrm{e}^{tQ} = \sum_{k\geq 0} \frac{t^k Q^k}{k!}, \quad t \geq 0, \tag{5.16}$$

in the Banach algebra of bounded operators on L^∞, in which then

$$\frac{d}{dt}\mathrm{e}^{tQ} = Q\mathrm{e}^{tQ} = \mathrm{e}^{tQ}Q, \quad \mathrm{e}^{(s+t)Q} = \mathrm{e}^{sQ}\mathrm{e}^{tQ}, \quad s,t \geq 0.$$

These results can be proved using the bounds obtained using (5.10), which are uniform on bounded time intervals,

$$\sum_{k\geq 0} \left\| \frac{t^k Q^k}{k!} \right\|_{\mathrm{op}} \leq \sum_{k\geq 0} \frac{t^k \|Q\|_{\mathrm{op}}^k}{k!} = \mathrm{e}^{t\|Q\|_{\mathrm{op}}}.$$

Recall that $(P_t)_{t\geq 0}$ is defined as the semigroup of the Markov process with generator Q, which exists as the result of a probabilistic pathwise construction. Hence, P_t is automatically given by a Markovian matrix.

From a theoretical point of view, it is not a priori clear that the operator e^{tQ} on L^∞ corresponds to a Markovian matrix, see Remark 5.4. Nevertheless, the uniqueness result in Theorem 5.8 implies that

$$P_t = \mathrm{e}^{tQ}$$

as an operator on L^∞, and also as an operator on $\mathscr{M}$ defined by duality or by right-multiplication by measures.

From a practical perspective, this equality is rarely of use, since the *effective* computation of the series defining e^{tQ} requires finding a tractable *spectral decomposition* for Q, and is most often impossible.

5.2.6 A Few Case Studies

In all the following examples, the description of a random evolution phenomena will allow to establish the generator for a Markov process which could represent it. It is then left to prove the actual existence of this process.

The $M/M/\infty$ Queue

A new customer arrives at each instant of a Poisson process with intensity $\lambda > 0$, is served starting at their arrival time for a duration with $\mathscr{E}(\mu)$ exponential law, and then leaves. This corresponds to the $M/M/\infty$ queue in Kendall's nomenclature: Markovian (or Memoryless) arrivals, Markovian (or Memoryless) services, ∞ number of servers.

Let X_t be the number of customers in the system at time $t \in \mathbb{R}_+$. When the process $(X_t)_{t\in\mathbb{R}_+}$ is in state $x \in \mathbb{N}$, it waits there until the first among the next jump instants of the arrival Poisson process and of x Poisson processes of intensity μ corresponding to the service times of the x customers. Then it jumps to either $x+1$ if this first instant belonged to the arrival process, or else to $x-1$ (in which case $x \geq 1$). The generator thus satisfies

$$Q(x, x+1) = \lambda, \quad Q(x, x-1) = x\mu, \quad x \in \mathbb{N},$$

and all other non-diagonal terms vanish. Hence

$$q(x) = \lambda + x\mu, \quad \Pi(x, x+1) = \frac{\lambda}{\lambda + x\mu}, \quad \Pi(x, x-1) = \frac{x\mu}{\lambda + x\mu}, \quad x \in \mathbb{N}.$$

The problem is now to prove existence of the corresponding Markov process. Since $\sup_{x\in\mathbb{N}} q(x) = \infty$, Theorem 5.6 does not apply.

A possibility would be to use Theorem 5.7 and a recurrence criterion for Markov chains; for instance, the detailed balance equations for the induced transition matrix Π yield an invariant measure, which is readily seen to be summable, and this proves that Π is positive recurrent.

A more direct proof follows.

The process $(X_t)_{t\in\mathbb{R}_+}$ can be constructed using independent Poisson processes, $(A_t)_{t\in\mathbb{R}_+}$ of intensity λ for the arrivals, and $(D_t^k)_{t\in\mathbb{R}_+}$ of intensity μ for $k \geq 1$ for the departures, so that $(D_t^1 + \cdots + D_t^k)_{t\in\mathbb{R}_+}$ is used for the next end of service when there are k customers in the system.

The issue is to prove that the jump instants of $(X_t)_{t\in\mathbb{R}_+}$ do not accumulate, a.s., for any $X_0 = x$. For any $n \geq 1$ there is $y_n \in \mathbb{N}$ large enough that

$$\mathbb{P}(A_n \leq y_n) \geq 1 - 1/n.$$

On the event $\{A_n \leq y\}$,

$$\sup_{0\leq s\leq n} X_s \leq x + y_n$$

and hence the jump instants $(T_k)_{k\geq 1}$ of $(X_t)_{t\in\mathbb{R}_+}$ on $[0,n]$ are among those of

$$\left(A_t + D_t^1 + \cdots + D_t^{x+y_n}\right)_{t\in\mathbb{R}_+}.$$

This is a Poisson process with intensity $\lambda + (x+y_n)\mu < \infty$, and its jump instants do not accumulate. Hence, by the monotone limit theorem,

$$\mathbb{P}\Big(\lim_{k\to\infty} T_k = \infty\Big) = \lim_{n\to\infty} \downarrow \mathbb{P}\Big(\lim_{k\to\infty} T_k > n\Big) \geq \lim_{n\to\infty} \downarrow \mathbb{P}(A_n \leq y_n) = 1.$$

Pure Birth and Birth and Death Processes

A process with values in $\mathbb{N}$ with generator Q such that

$$Q(x, x+1) = q(x) \geq 0, \quad x \in \mathbb{N},$$

and all other non-diagonal terms vanish is called a *pure birth* process. The only non-vanishing terms of the induced transition matrix Π are $\Pi(x, x+1) = 1$ for $x \geq 0$. Theorems 5.5 and 5.4 show that the corresponding Markov process exists if and only if

$$\sum_{x\in\mathscr{V}} \frac{1}{q(x)} = \infty.$$

More generally, a process with values in $\mathbb{N}$ with generator Q such that

$$Q(x, x+1) = \lambda_x \geq 0, \quad x \in \mathbb{N}, \qquad Q(x, x-1) = \mu_x \geq 0, \quad x \geq 1,$$

and its other non-diagonal terms vanish is called a *birth and death* process. With the convention $\mu_0 = 0$ and $\frac{0}{0} = 0$,

$$q(x) = \lambda_x + \mu_x, \quad \Pi(x, x+1) = \frac{\lambda_x}{\lambda_x + \mu_x}, \quad \Pi(x, x-1) = \frac{\mu_x}{\lambda_x + \mu_x}, \quad x \in \mathbb{N}.$$

Necessary and sufficient conditions of non-explosion in terms of these coefficients can be found in Anderson [1]. Here, we are going to give a simple proof that a sufficient condition is

$$\sum_{x\in\mathscr{V}} \frac{1}{\lambda_x} = \infty. \tag{5.17}$$

It is satisfied by the $M/M/\infty$ queue process, which is an example of birth and death process, which gives another existence proof for it.

Let $\hat{Q}$ be the pure jump generator such that

$$\hat{Q}(x, x+1) = Q(x, x+1) := \lambda_x, \quad x \in \mathbb{N}.$$

The "Second construction, bis" in Sect. 5.2.1 can be used to construct *simultaneously* the processes $(X_t)_{t\in\mathbb{R}_+}$ and $(\hat{X}_t)_{t\in\mathbb{R}_+}$ with respective generators Q and $\hat{Q}$,

with $X_0 = \hat{X}_0$ and the *same* Poisson processes $N^{x,x+1}$ of intensity λ_x for determining the jumps from x to $x+1$ for *both* processes. Let $(T_n)_{n\geq 1}$ and $(\hat{T}_n)_{n\geq 1}$ denote the respective sequences of jump instants.

Assume that (5.17) holds. Theorem 5.5 yields that $\lim_{n\to\infty} \hat{T}_n = \infty$, a.s. The processes have unit jumps, and if they are both at x then they can jump from x to $x+1$ only simultaneously by construction, hence $X_t \leq \hat{X}_t$ for $t < \lim_{n\to\infty} T_n$. Theorem 5.7 then implies that $\lim_{n\to\infty} T_n = \infty$, a.s.

A Network in Which Customers Join the Shortest Available Queue

There are $K \geq 1$ queues, numbered by k from 1 to K, and queue number k has a unique server at rate μ_k and a waiting queue with infinite capacity. By convention, queue $K+1$ denotes queue 1, and queue 0 denotes queue K.

For $1 \leq k \leq K$, customers arrive at the instants of a Poisson process with intensity λ_k, and join either queue k or $k+1$ according to their sizes x_k and x_{k+1}: queue k with probability $p_k(x_k, x_{k+1}) \in [0,1]$, or queue $k+1$ with complementary probability. The case $p_k(x_k, x_{k+1}) \in \{0,1\}$ constitutes a deterministic routing rule.

One may wish to use simulation in order to determine the functions p_k in terms of the other parameters, so as to ensure an approximately optimal operation for this network.

Let $X_t = (X_t^k)_{1\leq k\leq K}$ for $t \in \mathbb{R}_+$, where X_t^k is the size of queue k (including the customer in service) at time t. The state space is $\mathbb{N}^K$, and $(e_k)_{1\leq k\leq K}$ denotes the canonical basis. If the process is at $x = (x_k)_{1\leq k\leq K}$ then:

- it will either jump to a state $x + e_j$ when a customer joins queue j for $1 \leq j \leq K$, at rates

$$Q(x, x+e_j) = \lambda_j p_j(x_j, x_{j+1}) + \lambda_{j-1}\big(1 - p_{j-1}(x_{j-1}, x_j)\big);$$

- or else to a state $x - e_j$ when a customer service ends at queue j for $1 \leq j \leq K$, at rates μ_j if $x_j > 0$, so that

$$Q(x, x-e_j) = \mu_j \mathbb{1}_{\{x_j>0\}}.$$

Thus

$$q(x) = \sum_{j=1}^{K} (\lambda_j + \mu_j \mathbb{1}_{\{x_j>0\}}), \qquad \sup_{x\in\mathbb{N}^K} q(x) = \sum_{j=1}^{K} (\lambda_j + \mu_j) < \infty,$$

and Theorem 5.6 ensures existence of the corresponding Markov process.

For the practical simulation of the network, care should be made to avoid useless computations, and some draws ought to be performed in two steps, as follows.

Algorithm (Fundamental simulation) Draw a sample x_0 for the initial value; then iteratively, if the process is in state $x = (x_k)_{1\leq k\leq K}$:

- compute the intensity $q(x)$, and draw a sample from the $\mathscr{E}(q(x))$ exponential law for the duration that the process stays at x,
- the jump corresponds, for some $1 \leq j \leq K$, either to an arrival concerning both queues j and $j+1$, or to a departure from queue j, with respective probabilities $\lambda_j/q(x)$ and $\mu_j \mathbb{1}_{\{x_j>0\}}/q(x)$; determine which is the case, then:
  - for an arrival as above, either with probability $p_j(x_j, x_{j+1})$ increment the size of queue j by 1, or else increment the size of queue $j+1$ by 1,
  - for a departure from queue j, decrement its size by 1.

Note that only one probability $p_j(x_j, x_{j+1})$ has to be computed, and only in case of an arrival. Obviously, the value of $q(x)$ should be updated at each iteration, by incrementing or decrementing it by 1 whenever there is one more or one less non-empty queue. The following method avoids doing so, but has more iterations.

Algorithm (Fictitious jump method) Compute $\lambda = \sum_{j=1}^{K}(\lambda_j + \mu_j)$, and draw a sample x_0 for the initial value; then iteratively, if the process is in state $x = (x_k)_{1\leq k\leq K}$:

- draw a sample from the $\mathscr{E}(\lambda)$ exponential law,
- the jump corresponds, for some $1 \leq j \leq K$, either to an arrival concerning both queues j and $j+1$, or to a *potential* departure from queue j, with respective probabilities λ_j/λ and μ_j/λ; determine which is the case, then:
  - for an arrival as above, either with probability $p_j(x_j, x_{j+1})$ increment the size of queue j by 1, or else increment the size of queue $j+1$ by 1,
  - for a potential departure from queue j, decrement its size by 1 if it is non-zero, else leave it untouched (fictitious jump).

Conclusion to Be Drawn

The previous examples show that, for obtaining existence results for a Markov process as well as for its efficient simulation, one should think ahead and find inspiration in the phenomenon of interest. Theorem 4.4 (superposition-decomposition of Poisson processes) is fundamental for finding efficient simulations for processes, respecting the adequate Markovian evolution.

5.3 Problems

5.1 (Assymetric Ehrenfest Urn) Molecules labeled by $1, \ldots, N$ are put in a box, constituted of two compartments A and B separated by an asymmetric porous barrier. Independently, any molecule in B passes into A at rate $\lambda > 0$, any molecule in A passes into B at rate $\mu > 0$. For $1 \leq i \leq N$ and $t \in \mathbb{R}_+$, let

$$Y_i(t) = 1 \quad \text{if molecule } i \text{ is in } A \text{ at time } t, \qquad Y_i(t) = 0 \quad \text{if not;}$$

$$X_t := \sum_{1 \le i \le N} Y_i(t).$$

The processes $(Y_i(t), 1 \le i \le N)_{t \ge 0}$ and $(X_t)_{t \ge 0}$ provide microscopic and macroscopic representations of the system.

1. Show that $(Y_i(t), 1 \le i \le N)_{t \ge 0}$ is a Markov process on $\{0, 1\}^N$ such that, starting from a state $(y_i, 1 \le i \le N)$ in $\{0, 1\}^N$, for $1 \le k \le N$,
   - the rate of jumps to $(y_i + \mathbb{1}_{\{i=k\}}, 1 \le i \le N)$ is $\lambda(1 - y_k)$;
   - the rate of jumps to $(y_i - \mathbb{1}_{\{i=k\}}, 1 \le i \le N)$ is μy_k.
2. Prove that $(X_t)_{t \ge 0}$ is a Markov process on $\{0, \dots, N\}$ with infinitesimal generator $Q = (Q(x, y))_{x,y=0,\dots,N}$ specified by its non-null off-diagonal terms

$$Q(x, x+1) = \lambda(N - x), \quad Q(x, x-1) = \mu x, \quad x = 0, \dots, N.$$

3. Let $\pi_t = (\pi_t(x))_{x=0,\dots,N}$ denote the law of X_t, and $f_t = (f_t(x))_{x=0,\dots,N}$ be given by $f_t(x) := \pi_t(0) + \cdots + \pi_t(x)$.
   (a) Write an ordinary differential equation satisfied by $(\pi_t)_{t \ge 0}$.
   (b) Prove that $\frac{d}{dt}\pi_t = 0 \Leftrightarrow \frac{d}{dt} f_t = 0$.
   (c) Write $\frac{d}{dt} f_t(x)$ in terms of π_t.
   (d) Find a simple recurrence equation (in x) satisfied by any probability measure $\tilde{\pi} = (\tilde{\pi}(x))_{x=0,\dots,N}$ such that if $\pi_0 = \tilde{\pi}$ then $\pi_t = \tilde{\pi}$ for every t.
   (e) Compute $\tilde{\pi}$ from this; do not forget that it is a probability measure.

5.2 (Chain Reaction) Let be given a probability space, on which all random variables are constructed, a constant $\lambda > 0$, and a law $\eta = (\eta(k))_{k \in \mathbb{N}}$ on $\mathbb{N}$ satisfying

$$\eta(1) \ne 1, \quad m := \sum_{k \in \mathbb{N}} k\eta(k) < \infty. \tag{5.18}$$

In a model for a nuclear reaction, every particle **exists** for a duration with exponential law of parameter $\lambda > 0$, and then **disappears** in a collision with the medium resulting in the emission of $k \ge 0$ **new** particles with probability $\eta(k)$. The initial number of particles, the existence durations, and the numbers of particles emitted during the collisions are **independent**.

A Markov process $(X_t)_{t \in \mathbb{R}_+}$ with values in $\mathbb{N}$, where X_t is the number of particles at time t, is to be constructed. Let T_n for $n \ge 1$ be its jump instants, $T_0 = 0$, and $(\hat{X}_n)_{n \in \mathbb{N}}$ be the induced Markov chain given by $\hat{X}_n = X_{T_n}$.

Let $(Z_k)_{k \ge 1}$ be a sequence of i.i.d. r.v.'s of law η, independent of X_0 and of the existence durations, giving the numbers of **new** particles emitted during the successive collisions; we recall that an existing particle **disappears** at every collision.

1. Prove that there is $\lambda' > 0$ and a law $\eta' = (\eta'(k))_{k\in\mathbb{N}}$ on $\mathbb{N}$ satisfying $\eta'(1) = 0$, to be computed in terms of λ and $\eta = (\eta(k))_{k\in\mathbb{N}}$, such that the law of the evolution of $(X_t)_{t\in\mathbb{R}_+}$ is the same as that in which λ and η are replaced by λ' and η'.

   In the sequel it is **always assumed** that $\eta(1) = 0$.

2. (a) Explain in detail how to construct and simulate $(X_t)_{t\in\mathbb{R}_+}$, and compute its infinitesimal generator $Q = (Q(x, y))_{x,y\in\mathbb{N}}$. What problem can arise?
   (b) Give a simple expression for $\hat{X}_n$ in terms of X_0, the Z_k for $k \geq 1$, and
   $$M = \inf\{k \geq 0 : X_0 + Z_1 + \cdots + Z_k - k = 0\}.$$
   Deduce from it that $\hat{X}_n = \mathcal{O}(n)$, i.e., $(\frac{1}{n}\hat{X}_n, n \geq 1)$ is uniformly bounded, a.s.
   (c) Compute $\mathbb{E}(\mathrm{e}^{-T_n} \mid (\hat{X}_k)_{0\leq k\leq n-1})$, then prove that $\lim_{n\to\infty} \mathbb{E}(\mathrm{e}^{-T_n}) = 0$.
   Deduce from this that the Markov process $(X_t)_{t\in\mathbb{R}_+}$ is well defined.

The generating functions of the emission law and of the instantaneous laws are

$$g : z \in [0, 1] \mapsto g(z) = \sum_{k\in\mathbb{N}} z^k \eta(k) \in [0, 1],$$

$$f_t : z \in [0, 1] \mapsto f_t(z) = \mathbb{E}_1\big(z^{X_t}\big) \in [0, 1].$$

Let $h : z \in [0, 1] \mapsto h(z) = \lambda(g(z) - z)$.

3. (a) Explain how to obtain the case when X_0 is arbitrary from the case $X_0 = 1$.
   Deduce from this $\mathbb{E}_x(z^{X_t})$ in terms of $f_t(z)$ and $x \in \mathbb{N}$.
   (b) Prove that
   $$f_{t+\varepsilon}(z) = f_\varepsilon\big(f_t(z)\big), \quad t, \varepsilon \geq 0,\ z \in [0, 1],$$
   (c) Let $X_0 = 1$. Prove that the probability that $(X_t)_{t\in\mathbb{R}_+}$ jumps one or more times on $[0, \varepsilon]$ is $\lambda\varepsilon + \mathcal{O}(\lambda^2\varepsilon^2)$, then that the probability that it jumps two or more times on $[0, \varepsilon]$ is of order $\mathcal{O}(m\lambda^2\varepsilon^2)$ (where m is given in (5.18)).
   (d) Prove that
   $$f_{t+\varepsilon}(z) - f_t(z) = h\big(f_t(z)\big)\varepsilon + \mathcal{O}\big((m + 1)\lambda^2\varepsilon^2\big), \quad t, \varepsilon \geq 0,\ z \in [0, 1],$$
   then that
   $$\frac{\partial}{\partial t} f_t(z) = h\big(f_t(z)\big), \quad t \geq 0,\ z \in [0, 1].$$
   To which result in the book does this correspond?

4. Let $\mathscr{E} = \{\exists t \geq 0 : X_t = 0\}$ and $\theta = \mathbb{P}_1(\mathscr{E})$ denote the event of extinction of the reaction and its probability.
   (a) Prove that $\theta = \lim_{t\to\infty} f_t(0)$, then that $h(\theta) = 0$ (i.e., θ is a root for h).
   (b) Prove that if $m \leq 1$ then $\theta = 1$.
   (c) Prove that if $m > 1$ then θ is the unique root for h on $[0, 1[$.
5. (a) Prove that
$$\frac{\partial}{\partial z} f_t(z) = \exp\left(\int_0^t h'\big(f_s(z)\big)\, ds\right), \quad t \geq 0,\ z \in [0, 1].$$
   (b) Deduce from this that $\mathbb{E}_1(X_t) = \mathrm{e}^{\lambda(m-1)t}$.
   (c) Prove that if $m = 1$ then $\mathbb{E}_1(\sup_{t\geq 0} X_t) = \infty$.
6. Consider the special case when $\eta(0) = 1 - p$ and $\eta(2) = p$, for $0 < p < 1$.
   (a) Compute f_t for $t \geq 0$.
   (b) Deduce from this the value of θ.

5.3 (Markov process, generator, and martingales ($\star$)) Let $(X_t)_{t\in\mathbb{R}_+}$ be a Markov process with values in a discrete space $\mathscr{V}$, with bounded generator Q. For every $f \in L^\infty$, define

$$\big(M_t^f\big)_{t\geq 0}, \quad M_t^f = f(X_t) - f(X_0) - \int_0^t Qf(X_s)\, ds.$$

1. Prove that $(M_t^f)_{t\geq 0}$ is a martingale w.r.t. $(X_t)_{t\geq 0}$, i.e., that $\mathbb{E}(M_{t+h}^f \mid (X_s)_{0\leq s\leq t}) = M_t^f$ for $t, h \geq 0$.

This expresses the *deterministic tendency* of the evolution of $(X_t)_{t\geq 0}$ in terms of Q, with *random fluctuations* given by the martingales $(M_t^f)_{t\geq 0}$. Let

$$(A_t)_{t\geq 0}, \quad A_t = \int_0^t \big(Q\big(f^2\big)(X_s) - 2f(X_s)Qf(X_s)\big)\, ds.$$

2. Assume first that $X_0 = x$ and $f(x) = 0$ (to simplify computations).
   (a) Prove that
$$M_t^{f^2} = \big(M_t^f\big)^2 + 2M_t^f \int_0^t Qf(X_s)\, ds + \left(\int_0^t Qf(X_s)\, ds\right)^2 - \int_0^t Q\big(f^2\big)(X_s)\, ds.$$
   (b) Prove that
$$\left(\int_0^t Qf(X_s)\, ds\right)^2 = 2\int_0^t Qf(X_s) \int_0^s Qf(X_u)\, du\, ds.$$

(c) Prove that a martingale is obtained by setting

$$M_t^f \int_0^t Qf(X_s)\,ds - \int_0^t M_s^f\, Qf(X_s)\,ds, \quad t \geq 0.$$

(d) Prove that $((M_t^f)^2 - A_t)_{t\geq 0}$ is a martingale vanishing at time zero, and then that $\mathrm{Var}(M_t^f) = \mathbb{E}(A_t)$.

3. Prove that all this still holds for $f(x) = a \neq 0$, then when X_0 is arbitrary.

With the notation $A_t = \langle M^f \rangle_t$, the process $(\langle M^f \rangle_t)_{t\geq 0}$ can be written in terms of Q, and allows to further evaluate the martingale $(M_t^f)_{t\geq 0}$ by the fact that $(M_t^f)^2 - \langle M^f \rangle_t$ is a martingale. The following computation yields a more explicit expression.

4. Prove that

$$\begin{aligned}\big\langle M^f \big\rangle_t &= \int_0^t \sum_{y\neq X_s} Q(X_s, y)\big(f(y) - f(X_s)\big)^2\, ds \\ &= \int_0^t \sum_{y\in\mathscr{Y}} Q(X_s, y)\big(f(y) - f(X_s)\big)^2\, ds.\end{aligned}$$

All this can be appropriately generalized to Markov processes taking values in a continuous state space.

Chapter 6
Continuous-Space Markov Processes with Jumps

Abstract From now on, Markov processes with continuous state space ($\mathbb{R}^d$ for some or one of its closed subsets) are considered. Their rigorous study requires advanced measure-theoretic tools, but we limit ourselves to developing the reader's intuition, notably by pathwise constructions leading to simulations. We first emphasize the strong similarity between such Markov processes with constant trajectories between isolated jumps and discrete space ones. We then introduce Markov processes with sample paths following an ordinary differential equation between isolated jumps. In both cases, the Kolmogorov equations and Feynman–Kac formula are established. This is applied to kinetic equations coming from statistical Mechanics. These describe the time evolution of the instantaneous distribution of particles in phase space (position-velocity), when the particle velocity jumps at random instants in function of the particle position and velocity.

6.1 Preliminaries

The construction of continuous-space Markov processes requires the use of measure and integration theory, and it is always assumed implicitly that functions and subsets are measurable w.r.t. the Borel σ-field $\mathscr{B}(\mathscr{V})$ of $\mathscr{V}$. Intrinsic notations will be used for integrals and adjoint operators, as well as more intuitive notations which establish links with the discrete space theory and allow to quickly write formulas; there will be some redundancy.

6.1.1 Measures, Functions, and Transition Kernels

Positive and Signed Measures, Integrals of Functions

A positive measure μ on $\mathscr{V}$ is a σ-additive function from $\mathscr{B}(\mathscr{V})$ to $[0, \infty]$. It is *finite* if its *total mass* $\mu(\mathscr{V})$ is finite, and a *probability measure*, or a *law*, if $\mu(\mathscr{V}) = 1$. The space of probability measures is denoted by $\mathscr{P} = \mathscr{P}(\mathscr{V})$.

A signed measure μ is a σ-additive function from $\mathscr{B}(\mathscr{V})$ to $\mathbb{R}$. It can be written as the difference $\mu = \mu^+ - \mu^-$ of two positive finite measures which are mutually

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*, Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_6, © Springer-Verlag Berlin Heidelberg 2013

singular, i.e., such that $\mu^+(A)\mu^-(A) = 0$ for every $A \in \mathscr{B}(\mathscr{V})$. The vector space of signed measures is denoted by $\mathscr{M} = \mathscr{M}(\mathscr{V})$. A positive measure belongs to $\mathscr{M}$ if and only if it is finite.

The Lebesgue integral by a measure μ of a function f is defined in $[0, \infty]$ if μ and f are positive (wide-sense), and in $\mathbb{R}$ if μ belongs to $\mathscr{M}$ and f to the space of bounded functions $L^\infty = L^\infty(\mathscr{V})$. It provides a natural duality bracket

$$(\mu, f) \mapsto \langle \mu, f \rangle := \int f \, d\mu = \int f(x)\mu(dx) = \int \mu(dx) f(x) \tag{6.1}$$

and all these notations will be used. The notations $\int \mu(dx) f(x)$ and $\langle \mu, f \rangle$ are close to the discrete state space matrix notation, and allow to recover some formulas in an easier way.

Transition Kernels and Their Actions

A transition or Markovian *kernel* P is a measurable mapping

$$x \in \mathscr{V} \mapsto P(x, \cdot) = P(x, dy) \in \mathscr{P}(\mathscr{V}).$$

A kernel of positive measures or of signed measures is defined similarly. The abuse of notation $P(x, y) = P(x, \{y\})$ is common, as it is for any measure.

A transition kernel $P = (P(x, dy))_{x \in \mathscr{V}}$ can be identified with the operator $f \mapsto Pf$ on L^∞ (or on $[0, \infty]$-valued functions) defined by the formula

$$Pf : x \in \mathscr{V} \mapsto Pf(x) = \big\langle P(x, \cdot), f \big\rangle := \int f(y) P(x, dy) := \int P(x, dy) f(y).$$

The adjoint operator $P^* : \mu \mapsto P^*\mu$ on $\mathscr{M}$ of P for the duality bracket (6.1) satisfies $\langle P^*\mu, f \rangle = \langle \mu, Pf \rangle$. Developing Pf, by the Fubini theorem

$$\langle \mu, Pf \rangle = \iint f(y) P(x, dy) \mu(dx) := \iint \mu(dx) P(x, dy) f(y).$$

Thus P^* can be expressed in terms of the kernel $(P(x, dy))_{x \in \mathscr{V}}$ by the formula

$$P^*\mu(dy) = \int_x P(x, dy)\mu(dx) := \int_x \mu(dx) P(x, dy) := \mu P(dy).$$

The last notation can be interpreted as defining the operation of P on its *left* on a measure μ, which allows to recover quickly this and future formulas.

Remark 6.1 If P and R are both Markovian kernels, then the composed operator $PR : f \mapsto PRf$ on L^∞ is given for $x \in \mathscr{V}$ by

$$PRf(x) = \int PR(x, dy) f(y), \qquad PR(x, dy) = \int P(x, dz) R(z, dy).$$

This defines a natural product for kernels, analogous to the one for matrices.

6.1.2 Markov Property, Finite-Dimensional Marginals

Transition Kernels, Instantaneous Laws, Markov Property

Abstract conditional expectation must now be used, see Sect. 2.3.1. The r.v.'s thus defined are defined only a.s., but this is mainly left implicit.

Definition 6.1 A process $(X_t)_{t\in\mathbb{R}_+}$ with values in $\mathscr{V}$ is a *Markov process* if there exists a family $(P_t)_{t\geq 0}$ of Markovian kernels $P_t = (P_t(x,dy))_{x\in\mathscr{V}}$ such that, for all $s,t\geq 0$ and $A\in\mathscr{B}(\mathscr{V})$,

$$\mathbb{P}\big(X_{s+t}\in A \,|\, (X_\theta)_{0\leq\theta\leq s}\big) = P_t(X_s, A).$$

The P_t are called the *transition kernels* of the process.

Remark 6.2 Then

$$\mathbb{P}(X_{s+t}\in A \,|\, X_s) = P_t(X_s, A).$$

Moreover, if f is non-negative or bounded,

$$\mathbb{E}\big(f(X_{s+t}) \,|\, (X_\theta)_{0\leq\theta\leq s}\big) = \mathbb{E}\big(f(X_{s+t}) \,|\, X_s\big) = P_t f(X_s) := \int P_t(X_s, dy) f(y),$$

and $P_t(x,dy)$ is the conditional law of X_{s+t} given $X_s = x$, and in particular of X_t given $X_0 = x$.

The notation for discrete spaces (see Sect. 5.1.2) is adapted to this framework: $\mathbb{P}_x$ and $\mathbb{E}_x$ are used to indicate that $X_0 = x$, and $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ that X_0 is of law $\mu\in\mathscr{P}$. For $f\in L^\infty$ and $\mu\in\mathscr{P}$ and $A\in\mathscr{B}(\mathscr{V})$,

$$\mathbb{P}_x(X_t\in A) := P_t(x,A), \qquad \mathbb{P}_\mu(X_t\in A) := P_t^*\mu(A) = \mu P_t(A),$$
$$\mathbb{E}_x\big(f(X_t)\big) := P_t f(x) = \big\langle P(x,\cdot), f\big\rangle, \qquad \mathbb{E}_\mu\big(f(X_t)\big) := \langle \mu, Pf\rangle,$$

and denoting by π_t the law of X_t,

$$\pi_t(A) := \mathbb{P}(X_t\in A), \qquad \pi_t := P_t^*\pi_0 = \pi_0 P_t, \qquad \pi_t := P_t(x,\cdot) \quad \text{for } \pi_0 = \delta_x.$$

A fundamental result is that Definition 6.1 is equivalent to apparently stronger properties, quite useful for proofs.

Theorem 6.1 (Markov property) *Let $(X_t)_{t\in\mathbb{R}_+}$ be a Markov process with values in $\mathscr{V}$, and $(P_t)_{t\in\mathbb{R}_+}$ denote its transition kernels. Then it satisfies the following properties, which are equivalent.*

(a) *The property in Definition 6.1.*

(b) *For any measurable* $C \subset \mathscr{D}(\mathbb{R}_+, \mathscr{V})$, *let*

$$F^C : x \in \mathscr{V} \mapsto F^C(x) = \mathbb{P}_x\big((X_h)_{h\geq 0} \in C\big).$$

Then

$$\mathbb{P}\big((X_{t+h})_{h\geq 0} \in C \mid (X_s)_{0\leq s\leq t}\big) = F^C(X_t) : \omega \mapsto \mathbb{P}_{X_t(\omega)}\big((X_h)_{h\geq 0} \in C\big).$$

This can be interpreted as follows: *at time* t, *chance provides a value* $x = X_t(\omega)$, *and given that*, *the law of the future evolution* $(X_{t+h})_{h\geq 0}$ *is that of a* copy *of the Markov process* $(X_h)_{h\geq 0}$ *started at* x *and independent from the past* $(X_s)_{0\leq s\leq t}$. *In this specific sense*, *conditional on* X_t, *the process* $(X_{t+h})_{h\geq 0}$ *is independent of* $(X_s)_{0\leq s\leq t}$ *and has the same law as the Markov process starting at* X_t.

(c) *For all* $0 = t_0 \leq t_1 < \cdots < t_n$ *and* $A_0, A_1, \ldots, A_n \in \mathscr{B}(\mathscr{V})$ *and law* π_0 *of* X_0,

$$\begin{aligned}
&\mathbb{P}(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) \\
&\quad = \int_{\mathscr{V}} \pi_0(dx_0) \int_{A_1} P_{t_1-t_0}(x_0, dx_1) \cdots \int_{A_n} P_{t_n-t_{n-1}}(x_{n-1}, dx_n) \\
&\quad = \int_{\mathscr{V}} \pi_0(dx_0) \prod_{k=1}^{n} \int_{A_k} P_{t_k-t_{k-1}}(x_{k-1}, dx_k).
\end{aligned}$$

Proof Property (b) implies (a) as a special case. Since

$$\begin{aligned}
\mathbb{P}(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) &= \mathbb{E}\big(\mathbb{E}\big(\mathbb{1}_{\{X_{t_1}\in A_1, \ldots, X_{t_n}\in A_n\}} \mid (X_s)_{0\leq s\leq t_{n-1}}\big)\big) \\
&= \mathbb{E}\big(\mathbb{1}_{\{X_{t_1}\in A_1, \ldots, X_{t_{n-1}}\in A_{n-1}\}} \mathbb{E}\big(\mathbb{1}_{\{X_{t_n}\in A_n\}} \mid (X_s)_{0\leq s\leq t_{n-1}}\big)\big).
\end{aligned}$$

Property (a) implies that

$$\mathbb{P}(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = \mathbb{E}\big(\mathbb{1}_{\{X_{t_1}\in A_1, \ldots, X_{t_{n-1}}\in A_{n-1}\}} P_{t_n-t_{n-1}}(X_{t_{n-1}}, A_n)\big)$$

which implies (c) by induction. Now, Property (b) will be obtained from (c).

Since the product σ-field on $\mathscr{D}(\mathbb{R}_+, \mathscr{V})$ is generated by the cylindrical sets, it is enough to prove the result when C is such a set, i.e., of the form

$$C = \big\{(x_h)_{h\geq 0} \in \mathscr{D}(\mathbb{R}_+, \mathscr{V}) : x_{h_1} \in A_1, \ldots, x_{h_n} \in A_n\big\}$$

for $0 \leq h_1 < \cdots < h_n$ and $A_1, \ldots, A_n \in \mathscr{B}(\mathscr{V})$. Then, applying three times Property (c), for all $0 = s_0 \leq s_1 < \cdots < s_n < t$ and $B_1, \ldots, B_n$ and B in $\mathscr{B}(\mathscr{V})$,

$$\begin{aligned}
&\mathbb{P}\big(X_{s_1} \in B_1, \ldots, X_{s_n} \in B_n, X_t \in B, (X_{t+h})_{h\geq 0} \in C\big) \\
&\quad = \int_{\mathscr{V}} \pi_0(dx_0) \prod_{k=1}^{n} \int_{B_k} P_{s_k-s_{k-1}}(x_{k-1}, dx_k)
\end{aligned}$$

$$\times \int_B P_{t-s_n}(x_n, dx_{n+1}) \prod_{k=1}^{n} \int_{A_k} P_{h_k}(x_{n+k}, dx_{n+k+1})$$

$$= \int_{\mathscr{Y}} \pi_0(dx_0) \prod_{k=1}^{n} \int_{B_k} P_{s_k - s_{k-1}}(x_{k-1}, dx_k) \int_B P_{t-s_n}(x_n, dx_{n+1}) F^C(x_{n+1})$$

$$= \mathbb{E}\big(\mathbb{1}_{\{X_{s_1} \in B_1, \ldots, X_{s_n} \in B_n, X_t \in B\}} F^C(X_t)\big)$$

and hence $F^C(X_t)$ satisfies the characteristic property (2.8) of conditional expectations. □

Remark 6.3 The law of a Markov process $(X_t)_{t\geq 0}$ is characterized by its finite-dimensional marginals, which are determined in terms of the law of X_0 and its transition kernels $(P_t)_{t\geq 0}$ by Property (c).

Strong Markov Property

It is natural to try to extend the Markov property by replacing t by a stopping time, but the situation for a continuous state space is much more delicate than for a discrete space. This would exceed the scope of this book, and our true needs, since the construction and simulation of the processes we consider will necessitate only to use the property for very special stopping times: the *jump instants* of the process.

6.1.3 Semigroup, Infinitesimal Generator

Theorem 6.1 implies that the transition kernels P_t of a Markov process satisfy the *Chapman–Kolmogorov* equation

$$P_{t+s}(x, dy) = \int_{z\in\mathscr{Y}} P_t(x, dz) P_s(z, dy), \quad x, y \in \mathscr{Y},\ t, s \geq 0,$$

which translates for the kernel product of Remark 6.1, or for operator composition, into

$$P_{t+s} = P_t P_s, \quad t, s \geq 0.$$

Hence, $(P_t)_{t\geq 0}$ forms a *semigroup* of Markovian kernels, called the *(transition) semigroup* of the Markov process.

This semigroup induces an operator semigroup $f \mapsto P_t f$ on L^∞ and another $\mu \mapsto P_t^* \mu = \mu P_t$ on $\mathscr{M}$ by duality, or letting it operate on its left on measures. The theory of such semigroups shows that they are characterized by their left derivative at 0.

Specifically, the (*infinitesimal*) *generator* of the Markov process, or of its semigroup, is the operator defined by

$$\begin{cases} \mathscr{A} : f \in \mathscr{D}(\mathscr{A}) \subset L^\infty \mapsto \mathscr{A} f \in L^\infty, \\ \mathscr{A} f : x \in \mathscr{V} \mapsto \mathscr{A} f(x) = \lim\limits_{\varepsilon \to 0^+} \dfrac{P_\varepsilon f(x) - f(x)}{\varepsilon} = \lim\limits_{\varepsilon \to 0^+} \dfrac{\mathbb{E}_x(f(X_\varepsilon) - f(x))}{\varepsilon}, \end{cases} \tag{6.2}$$

where its domain $\mathscr{D}(\mathscr{A})$ is constituted of the functions such that these limits exist and define a function $\mathscr{A} f$ in L^∞.

It characterizes the law of the evolution of the Markov process, and we can hope to obtain it in *explicit* form from the model under study.

6.2 Markov Processes Evolving Only by Isolated Jumps

The simplest class of Markov processes on $\mathscr{V}$ is now studied, and the relation with Markov processes on discrete space made clear.

6.2.1 Semigroup, Infinitesimal Generator, and Evolution Law

Let us first consider Markov processes $(X_t)_{t \in \mathbb{R}_+}$ which evolve *only* by *isolated* jumps. These can be written in the form

$$\begin{cases} X_t = \sum\limits_{n \geq 0} X_{T_n} \mathbb{1}_{\{T_n \leq t < T_{n+1}\}} \quad \text{for } t \geq 0, \\ 0 = T_0 < T_1 \leq T_2 \leq \cdots \leq \infty, \quad \lim\limits_{n \to \infty} T_n = \infty, \quad T_n < \infty \Leftrightarrow X_{T_n} \neq X_{T_{n-1}}. \end{cases} \tag{6.3}$$

For $n \geq 1$, if $T_n < \infty$ then it is a *true* jump instant (as opposed to a *fictitious* one), and necessarily $T_n > T_{n-1}$. All this is analogous to Theorem 5.3.

The theory of these processes is quite similar to that of Markov processes with a discrete state space. The hypothesis of "isolated jumps" is a serious limitation: a process evolving by jumps in continuous space can have converging jump instants as well as states at those instants, and then "start over" and be defined for all times.

Theorem 6.2 *Let $(X_t)_{t \in \mathbb{R}_+}$ be a Markov process with values in $\mathscr{V}$ of the form (6.3), and $(P_t)_{t \in \mathbb{R}_+}$ denote its semigroup.*

(a) *The generator $\mathscr{A}$ of the Markov process, defined in (6.2), is given by a kernel $(\alpha(x, dy))_{x \in \mathscr{V}}$ of finite positive measures on $\mathscr{V}$ as follows: for $x \in \mathscr{V}$ and $f \in L^\infty$,*

$$\begin{cases} \alpha(x,dy) \geq 0, \quad \alpha\big(x,\{x\}\big) = 0, \quad q(x) := \alpha(x,\mathscr{V}) < \infty, \\ \mathscr{A}f(x) = \displaystyle\int \big(f(y) - f(x)\big)\alpha(x,dy) = \int f(y)\alpha(x,dy) - q(x)f(x). \end{cases} \tag{6.4}$$

In particular

$$\begin{aligned} P_\varepsilon(x,A) &= \alpha(x,A)\varepsilon + o(\varepsilon) \quad \textit{for } A \in \mathscr{B}\big(\mathscr{V} - \{x\}\big), \\ P_\varepsilon\big(x,\{x\}\big) &= 1 - q(x)\varepsilon + o(\varepsilon). \end{aligned}$$

(b) *When $X_0 = x$ the r.v.'s T_1 and X_{T_1} are independent. If $q(x) = 0$ then $T_1 = \infty$ and we set $X_{T_1} = x$ and $\Pi(x,dy) = \delta_x(dy)$. If $q(x) > 0$ then*

$$\begin{cases} T_1 \textit{ has } \mathscr{E}\big(q(x)\big) \textit{ exponential law}, \\ X_{T_1} \textit{ has law } \Pi(x,dy) = \dfrac{\alpha(x,dy)}{q(x)}. \end{cases}$$

These quantities satisfy

$$q(x) \geq 0, \qquad \Pi\big(x,\{x\}\big) = \mathbb{1}_{\{q(x)=0\}}, \qquad \Pi(x,dy) \geq 0, \qquad \Pi(x,\mathscr{V}) = 1, \tag{6.5}$$

and are related to those of (6.4) by

$$\alpha(x,dy) = q(x)\Pi(x,dy). \tag{6.6}$$

Moreover, for $A \in \mathscr{B}(\mathbb{R}_+)$ and $B \in \mathscr{B}(\mathscr{V})$,

$$\begin{aligned} &\mathbb{P}\big(T_n - T_{n-1} \in A, X_{T_n} \in B \mid (T_k, X_{T_k})_{0\leq k\leq n-1}\big) \\ &\quad = \mathbb{P}(T_n - T_{n-1} \in A, X_{T_n} \in B \mid X_{T_{n-1}}) \\ &\quad = q(X_{T_{n-1}}) \int_A \mathrm{e}^{-q(X_{T_{n-1}})s}\, ds \times \Pi(X_{T_{n-1}}, B). \end{aligned}$$

Thus, in order to construct T_n and X_{T_n} starting from $(T_k, X_{T_k})_{0\leq k\leq n-1}$, it suffices to use the observed value $X_{T_{n-1}} = x$ and draw $T_n - T_{n-1}$ and X_{T_n} independently according to the laws $\mathscr{E}(q(x))$ and $\Pi(x,dy)$. Figure 5.2 can be readily adapted.

Remark 6.4 In particular, for all A in $\mathscr{B}(\mathscr{V})$ not containing x,

$$P_\varepsilon(x,A) = Q(x,A)\varepsilon + o(\varepsilon), \qquad P_\varepsilon(x,x) = 1 + Q(x,x)\varepsilon + o(\varepsilon),$$

and the process has probability $Q(x,A)\varepsilon + o(\varepsilon)$ to jump from x to A within ε.

One calls:

- the operator $\mathscr{A}$: the *generator* of the Markov process or of its semigroup,
- its kernel $\alpha = (\alpha(x, dy))_{x\in\mathscr{Y}}$: the *jump measure kernel*,
- the non-negative real number $q(x)$: the jump *intensity* or *rate* from x; if $q(x) = 0$ then x is an absorbing state,
- the kernel $\Pi = (\Pi(x, dy))_{x\in\mathscr{Y}}$: the *jump law kernel*,
- the sequence $(X_{T_n})_{n\in\mathbb{N}}$ visiting the states in the same order as $(X_t)_{t\in\mathbb{R}_+}$: the *induced Markov chain* of transition kernel Π (*discrete* time, *continuous* space).

Proof (a) As in the proof for Theorem 5.3, a discretization of time by a grid with mesh $\varepsilon > 0$ is used to yield existence of

$$q(x) := \lim_{\varepsilon\to 0^+} \frac{1 - P_\varepsilon(x, \{x\})}{\varepsilon} \in \mathbb{R}_+, \qquad \mathbb{P}_x(T_1 > t) = \mathrm{e}^{-q(x)t}.$$

The case $q(x) = 0$ corresponds to $\mathbb{P}_x(T_1 = \infty) = 1$ and the case $q(x) = \infty$ is eliminated since $T_1 > T_0 = 0$ by hypothesis. Likewise, for $q(x) > 0$ and $f \in L^\infty$ this discretization yields existence of the limit

$$\mathbb{E}_x\big(\mathbb{1}_{\{T_1\le t\}}\big(f(X_{T_1}) - f(x)\big)\big) = \frac{\mathbb{P}_x(T_1 \le t)}{q(x)} \lim_{\varepsilon\to 0^+} \frac{P_\varepsilon f(x) - f(x)}{\varepsilon}$$

which is thus equal to

$$q(x)\mathbb{E}_x\big(f(X_{T_1}) - f(x)\big) = \int \big(f(y) - f(x)\big)\alpha(x, dy).$$

The result for $P_\varepsilon(x, A)$ corresponds to $f = \mathbb{1}_A$, for $P_\varepsilon(x, \{x\})$ to $f = \mathbb{1}_{\{x\}}$.

(b) We give a proof which avoids using the strong Markov property. Since the T_n are isolated, using right continuity and setting $J_k = j_1 + \cdots + j_k$ results in

$$\begin{aligned}
&\mathbb{P}(T_k - T_{k-1} \le s_k, X_{T_k} \in A_k : 1 \le k \le n)\\
&\quad = \lim_{\varepsilon\to 0^+} \sum_{\substack{0\le j_k\le \lceil s_k/\varepsilon\rceil\\ 1\le k\le n}} \mathbb{P}(X_{J_{k-1}\varepsilon} = \cdots = X_{(J_k-1)\varepsilon} \ne X_{J_k\varepsilon}, X_{J_k\varepsilon} \in A_k : 1 \le k \le n)
\end{aligned}$$

and the Markov property in Theorem 6.1(c)yields

$$\begin{aligned}
&\mathbb{P}(X_{J_{k-1}\varepsilon} = \cdots = X_{(J_k-1)\varepsilon} \ne X_{J_k\varepsilon}, X_{J_k\varepsilon} \in A_k : 1 \le k \le n)\\
&\quad = \int_{\mathscr{Y}} \pi_0(dx_0) P_\varepsilon\big(x_0, \{x_0\}\big)^{j_1-1} \int_{A_1-\{x_0\}} P_\varepsilon(x_0, dx_1) P_\varepsilon\big(x_1, \{x_1\}\big)^{j_2-1}\\
&\qquad \cdots \int_{A_{n-1}-\{x_{n-2}\}} P_\varepsilon(x_{n-2}, dx_{n-1}) P_\varepsilon\big(x_{n-1}, \{x_{n-1}\}\big)^{j_n-1} \int_{A_n-\{x_{n-1}\}} P_\varepsilon(x_{n-1}, dx_n)
\end{aligned}$$

and hence, similarly to (a),

$$\begin{aligned}&\mathbb{P}(T_k - T_{k-1} \le s_k, X_{T_k} \in A_k : 1 \le k \le n)\\ &\quad = \int_{\mathscr{V}} \pi_0(dx_0)\big(1 - \mathrm{e}^{-q(x_0)s_1}\big) \int_{A_1} \Pi(x_0, dx_1)\\ &\qquad \cdots \int_{A_{n-1}} \Pi(x_{n-2}, dx_{n-1})\big(1 - \mathrm{e}^{-q(x_{n-1})s_n}\big) \int_{A_n} \Pi(x_{n-1}, dx_n).\end{aligned}$$

The conclusion follows from the conditional expectation characteristic property (2.8). □

Operator Kernels

The operator $\mathscr{A}$ in (6.4) is an integral operator, and as such is said to have a kernel $(Q(x, dy))_{x\in\mathscr{V}}$ of signed measures, satisfying $\mathscr{A}f(x) = \langle Q(x, \cdot), f\rangle := \int f(y)Q(x, dy)$ and hence

$$Q(x, dy) = \alpha(x, dy) - \alpha(x, \mathscr{V})\delta_x(dy) = \alpha(x, dy) - q(x)\delta_x(dy).$$

Such an operator is often identified with its kernel, but Markov process theory stresses instead the jump measure kernel $(\alpha(x, dy))_{x\in\mathscr{V}}$, constituted of positive measures describing more directly the instantaneous evolution by jumps. One of the reasons is that the case $\alpha(x, \mathscr{V}) = \infty$ is also of interest.

Back to a Discrete State Space

When $\mathscr{V}$ is discrete, identifying a function f to the sequence $(f(x))_{x\in\mathscr{V}}$ of its values specifies a basis, in which an operator $\mathscr{A}$ satisfying (6.4) has matrix $Q = (Q(x, y))_{x,y\in\mathscr{V}}$ satisfying

$$Q(x, y) = \alpha\big(x, \{y\}\big) \ge 0, \quad x \ne y; \qquad Q(x, x) = -\alpha(x, \mathscr{V}) = -\sum_{y\ne x} Q(x, y),$$

and thus (5.3). Conversely, given a matrix $Q = (Q(x, y))_{x,y\in\mathscr{V}}$ satisfying (5.3), an operator $\mathscr{A}$ satisfying (6.4) is obtained using the jump kernel

$$\alpha(x, dy) = \sum_{z\ne x} Q(x, z)\delta_z(dy), \quad x \in \mathscr{V},$$

and $\mathscr{A}$ will have matrix Q in the aforementioned basis.

6.2.2 Construction, Simulation, Existence

Fundamental Constructions

According to Theorem 6.2, the random evolution of a Markov process $(X_t)_{t\geq 0}$ can be characterized and described in terms of

- its generator $\mathscr{A}$, or corresponding jump kernel $(\alpha(x,dy))_{x\in\mathscr{V}}$ of positive measures, satisfying (6.4),
- or the jump rates $(q(x))_{x\in\mathscr{V}}$ and the induced transition kernel $(\Pi(x,dy))_{x\in\mathscr{V}}$ satisfying (6.5),

and Fig. 5.2 can readily be adapted. These quantities are related by (6.6).

Conversely, Theorem 6.2 is key to try to *construct* and *simulate* a Markov process $(X_t)_{t\geq 0}$ starting from such quantities. In the following two constructions, which are direct transcriptions of the relevant ones in Sect. 5.2.1, draws are always assumed to be *independent*.

Algorithm (First construction) Draw a sample for the initial value x_0; then iteratively, if the process is in state x:

- draw a sample from an $\mathscr{E}(q(x))$ exponential law for the sojourn duration at x,
- draw the next state $y\neq x$ from the law $\Pi(x,dy)=\frac{\alpha(x,dy)}{q(x)}$.

(If $q(x)=0$ then the process is absorbed at x.)

This construction constitutes a practical *simulation* method for the process. It has the following interpretation in terms of Poisson processes, which is useful for comprehension as well as for some theoretical results.

Algorithm (First construction, bis) For every x in $\mathscr{V}$, consider a Poisson process N^x with intensity $q(x)$ and jump instants $(T_n^x)_{n\geq 1}$, and a sequence $(Y_n^x)_{n\geq 1}$ of i.i.d. r.v.'s with laws $\Pi(x,dy)$, independently. Draw a sample for the initial value x_0; then iteratively, if the process reaches state x, let it stay there until the next jump instant of N^x, and if this is T_m^x then let the process jump to Y_m^x.

From a theoretical perspective, in order to construct the $(Y_n^x)_{n\geq 1}$, a non-countable product of spaces indexed by $x\in\mathscr{V}$ and equipped with the product σ-field can be used. Any subset of the product space which is a cylinder set based only on a countable set of coordinates is measurable for the product σ-field, and these are the only kind of sets used in the construction.

The "Second construction" and the "Second construction, bis" in Sect. 5.2.1 are more difficult to adapt. This can be done using the more advanced tool of Poisson point processes, also called Poisson random measures.

Interpreting a random jump evolution in terms of one of these constructions often allows to prove that it corresponds to a Markov process and to identify its generator $\mathscr{A}$, through its jump kernel α or its jump rates $q(x)$ and its jump law kernel Π.

Non-explosion and Existence of the Markov Process

Section 5.2.2 can be readily adapted, and we refer to it for explanations. Let $(T_n)_{n\geq 1}$ be the sequence of jump instants obtained in one of the above construction attempts.

Theorem 6.3 *Let $\mathscr{A}$ be an operator satisfying (6.4). If any of the above (equivalent in law) constructions is such that*

$$\lim_{n\to\infty} T_n = \infty, \quad a.s.,$$

then the resulting process $(X_t)_{t\in\mathbb{R}_+}$ is Markov and has generator $\mathscr{A}$.

Proof The above "First construction, bis" will be used. The conclusion follows from the Markov property of the Poisson processes N^x, the independence of these and the $(T_n^x)_{n\geq 1}$, and the fact that the latter are all i.i.d. of appropriate law. Details are left to the interested reader. □

The following result provides a criterion for this abstract existence result.

Theorem 6.4 *Let $\mathscr{A}$ be an operator satisfying (6.4). For any of the above constructions, a.s.,*

$$\lim_{n\to\infty} T_n = \infty \quad \textit{if and only if} \quad \sum_{n\in\mathbb{N}} \frac{1}{q(X_{T_n})} = \infty.$$

Proof See the proof for Theorem 5.5. □

This yields the two following existence results, under verifiable assumptions. The first one uses the jump intensities, the second the induced kernel.

Theorem 6.5 *Let $\mathscr{A}$ be an operator satisfying (6.4). If it is bounded in the sense that*

$$\sup_{x\in\mathscr{V}} q(x) < \infty,$$

then any of the above constructions is such that $\lim_{n\to\infty} T_n = \infty$ a.s., and yields a Markov process $(X_t)_{t\in\mathbb{R}_+}$ with generator $\mathscr{A}$.

Proof See the proof for Theorem 5.6. □

Theorem 6.6 *Let $\mathscr{A}$ be an operator satisfying (6.4), and consider one of the above constructions. If the jump rates are locally bounded, i.e., if*

$$\sup_{x\in B} q(x) < \infty \quad \textit{for every bounded set } B \subset \mathscr{V},$$

and if there exists a random bounded subset B of $\mathscr{V}$ such that

$$\sum_{n\in\mathbb{N}} \mathbb{1}_{\{X_{T_n}\in B\}} = \infty, \quad a.s.,$$

then $\lim_{n\to\infty} T_n = \infty$ *a.s., and the construction yields a Markov process* $(X_t)_{t\in\mathbb{R}_+}$ *of generator* $\mathscr{A}$.

Proof The proof for Theorem 5.7 is readily adapted. □

This again allows to understand the terminology "explosion".

Fundamental Simulation

The "First construction" at the beginning of this Sect. 6.2.2 allows to simulate efficiently a Markov process evolving by isolated jumps on a continuous state space, with generator $\mathscr{A}$ satisfying (6.3). This natural method relies on the fundamental construction in Theorem 6.2, see also Fig. 5.2, and uses the jump rates $q(x)$ and the induced transition kernel $\Pi(x, dy)$, which can be deduced from the jump measure kernel α of the generator $\mathscr{A}$ by

$$q(x) = \alpha(x, \mathscr{V}), \qquad \alpha(x, dy) = q(x)\Pi(x, dy).$$

This method simulates recursively the jumps of the process. In an actual code, this is obtained by a loop.

Algorithm (Fundamental simulation) Draw a sample for the initial value x_0; then iteratively, if the process is in state x:

- compute the intensity $q(x)$, and draw a sample from the $\mathscr{E}(q(x))$ exponential law for the duration that the process stays at x,
- compute $\Pi(x, dy) = \frac{\alpha(x,dy)}{q(x)}$, and a draw a sample from this law for the next state.

(If $q(x) = 0$ then the process is absorbed at x.)

It is sometimes possible to perform the latter draw more economically, by steps, by regrouping the y into adequate subsets.

Fictitious Jump Method

We give another construction method which can be useful in simulations. If $\hat{q}(x) \geq q(x)$ for $x \in \mathscr{V}$ then a Markovian kernel $\hat{\Pi}(x, dy)$ is given by

$$\hat{\Pi}(x, dy) = \frac{q(x)}{\hat{q}(x)}\Pi(x, dy), \qquad \hat{\Pi}\big(x, \{x\}\big) = 1 - \frac{q(x)}{\hat{q}(x)}; \tag{6.7}$$

note that (6.6) can be extended to

$$\alpha(x,dy)\mathbb{1}_{\{y\neq x\}} = q(x)\Pi(x,dy)\mathbb{1}_{\{y\neq x\}} = \hat{q}(x)\hat{\Pi}(x,dy)\mathbb{1}_{\{y\neq x\}}$$

and that

$$\mathscr{A}f(x) = q(x)\int \big(f(y)-f(x)\big)\Pi(x,dy) = \hat{q}(x)\int \big(f(y)-f(x)\big)\hat{\Pi}(x,dy).$$

An important special case is when

$$\hat{q}(x) = \lambda \geq \sup_{y\in\mathscr{V}} q(y), \quad x\in\mathscr{V}.$$

Here is a method constructing $(X_t)_{t\in\mathbb{R}_+}$ much like in the above "First construction", but using $\hat{q}$ and $\hat{\Pi}$ instead of q and Π; the second step (the draw according to $\hat{\Pi}$) is decomposed into two separate steps, as would be done in an actual algorithm.

Algorithm (Fictitious jump method) Draw a sample for the initial value x_0; then iteratively, if the process is in state x:

- compute the intensity $\hat{q}(x)$, and draw a sample from the $\mathscr{E}(\hat{q}(x))$ exponential law for a duration after which:
- – either, with probability $1-\frac{q(x)}{\hat{q}(x)}$, the process remains at x,
  – or else (with complementary probability $\frac{q(x)}{\hat{q}(x)}$), the process leaves x; compute $\Pi(x,dy)=\frac{\alpha(x,dy)}{q(x)}$, and draw a sample from this law for the next state.

(If $\hat{q}(x)=0$ then the process is absorbed at x.)

This method can be advantageous for simulation, when there are good upper bounds $\hat{q}(x)$ for the $q(x)$ and the former are much easier to compute than the latter; a natural choice is to take $\hat{q}(x)$ constant in x, so as not to have to compute it at each actual jump of the process.

The sampling result in Theorem 4.4(b), and a simple conditioning show that the fictitious jump method is equivalent in law to the "First construction". Its name comes from the fact that it provides too many "potential" jump instants for the process, some of which are then rejected (and are thus "fictitious").

For instance, a proof of Theorem 6.5 can be thus obtained, by embedding the jump instants among those of a Poisson process with intensity $\lambda = \sup_{x\in\mathscr{V}} q(x)$ (which are known not to accumulate).

6.2.3 *Kolmogorov Equations, Feynman–Kac Formula, Bounded Generator Case*

The computations in Sect. 6.1.1 are readily generalized to integral operators with signed measure kernels. If $\mathscr{A}$ is an operator satisfying (6.4) with jump kernel α,

then its adjoint operator $\mathscr{A}^*$ acting on measures μ is given by

$$\mathscr{A}^*\mu(dy) = \int_x \alpha(x,dy)\mu(dx) - q(y)\mu(dy) = \int_x \big[\alpha(x,dy)\mu(dx) - \alpha(y,dx)\mu(dy)\big] \tag{6.8}$$

(using $q(y) = \alpha(y,\mathscr{V}) = \int_x \alpha(y,dx)$). This can be readily recovered, in the form where μ is written before q or α, from the formal expression $\mathscr{A}^*\mu = \mu\mathscr{A}$.

For the interested reader, we leave as an exercise to obtain these expressions rigorously using the duality formula.

Theorem 6.2 again has extensions in terms of linear differential equations. Sections 5.2.4 and 5.2.5 are readily adapted, notably Lemma 5.1. We use the duality bracket (6.1), and

$$\|\mu\|_{\mathrm{VT}} = \sup_{\|f\|_\infty \le 1} \langle \mu, f\rangle = \sup_{\|f\|_\infty \le 1} \int f\, d\mu, \quad \mu \in \mathscr{M},$$

an operator $\mathscr{A}$ satisfying (6.4) has operator norm

$$\|\mathscr{A}\|_{\mathrm{op}} = \big\|\mathscr{A}^*\big\|_{\mathrm{op}} = 2\sup_{x\in\mathscr{V}} \big\|\alpha(x,dy)\big\|_{\mathrm{VT}} = 2\sup_{x\in\mathscr{V}} \alpha(x,\mathscr{V}) = 2\sup_{x\in\mathscr{V}} q(x), \tag{6.9}$$

and hence $\mathscr{A}$ is bounded if and only if $\sup_{x\in\mathscr{V}} q(x) < \infty$. In this case, Theorem 6.4 proves existence of the Markov process with generator $\mathscr{A}$.

Theorem 6.7 *Let $\mathscr{A}$ be an operator satisfying* (6.4)*, with jump measure kernel α, and such that*

$$\sup_{x\in\mathscr{V}} q(x) < \infty.$$

Let $(P_t)_{t\ge 0}$ be the semigroup of the Markov process $(X_t)_{t\ge 0}$ with generator $\mathscr{A}$.

(a) *In the Banach algebra of the operators on L^∞, the semigroup $(P_t)_{t\ge 0}$ is the unique solution of the* backward Kolmogorov *equation*

$$\begin{cases} \dfrac{d}{dt}P_t = \mathscr{A}P_t, \quad i.e., \quad \dfrac{d}{dt}P_t(x,dy) = \displaystyle\int_z \alpha(x,dz)P_t(z,dy) - q(x)P_t(x,dy), \\ \qquad\qquad\qquad\qquad x \in \mathscr{V}, \\ P_0 = I. \end{cases}$$

In the Banach space L^∞, for $f \in L^\infty$, the collection $(v_t)_{t\ge 0}$ of functions $v_t : x \mapsto P_t f(x) = \mathbb{E}_x(f(X_t))$ is the unique solution of the backward Kolmogorov *equation*

$$\begin{cases} \dfrac{d}{dt}v_t = \mathscr{A}v_t, \quad i.e., \quad \dfrac{d}{dt}v_t(x) = \displaystyle\int v_t(y)\alpha(x,dy) - q(x)v_t(x), \quad x \in \mathscr{V}, \\ v_0 = f. \end{cases}$$

(b) *In the Banach algebra of the operators on* $\mathscr{M}$, *the adjoint semigroup* $(P_t^*)_{t\geq 0}$ *is the unique solution of the* forward Kolmogorov *equation*

$$\begin{cases} \dfrac{d}{dt}P_t^* = \mathscr{A}^* P_t^*, \quad i.e., \quad \dfrac{d}{dt}P_t(x,dy) \\ \qquad\qquad\qquad\qquad = \displaystyle\int_z P_t(x,dz)\alpha(z,dy) - P_t(x,dy)q(y), \quad x \in \mathscr{V}, \\ P_0^* = I. \end{cases}$$

In the Banach space $\mathscr{M}$, *for any initial law* π, *the collection* $(\pi_t)_{t\geq 0}$ *of instantaneous laws, given by* $\pi_t = P_t^*\pi = \pi P_t$ *where* $\pi_t(A) = \mathbb{P}_\pi(X_t \in A)$ *for* $A \in \mathscr{B}(\mathscr{V})$, *is the unique solution of the* forward Kolmogorov *equation* (*or* Fokker–Planck *equation*)

$$\begin{cases} \dfrac{d}{dt}\pi_t = \mathscr{A}^*\pi_t, \quad i.e., \quad \dfrac{d}{dt}\pi_t(dy) = \displaystyle\int_z \pi_t(dx)\alpha(x,dy) - \pi_t(dy)q(y), \\ \pi_0 = \pi. \end{cases}$$

This equation with initial condition $\mu \in \mathscr{M}$ *has unique solution* $(\mu P_t)_{t\geq 0}$.

Proof The proof of Theorem 5.8 is readily adapted. □

The comments following Theorem 5.8 can be readily adapted, notably that weakened results are true when $\sup_{x\in\mathscr{V}} q(x) = \infty$, and the following result for some terminal-value equations.

Corollary 6.1 (Feynman–Kac formula) *Under the assumptions and with the notation of Theorem* 6.7, *for* $T > 0$ *and* $f \in L^\infty$ *the backward equation*

$$\begin{cases} \dfrac{d}{dt}u_t + \mathscr{A}u_t = 0, \quad t \in [0,T], \\ u_T = f, \end{cases}$$

has a unique differentiable solution $t \in [0,T] \mapsto u_t \in L^\infty$, *which is given by*

$$x \mapsto u_t(x) = \mathbb{E}\big(f(X_T) \,|\, X_t = x\big) = P_{T-t}f(x).$$

Proof The time change $t \mapsto T - t$ on $[0,T]$ yields a bijection between the solutions of this equation and those of the backward Kolmogorov equation in L^∞. □

An integral equation arising in applications, usually through balance considerations, may have the requisite structure to be interpreted as a forward or backward equation for a pure-jump Markov process. This *probabilistic representation* allows its approximate solution by Monte Carlo methods, and can also be useful for theoretical results, such as those in the corollary.

Forward Kolmogorov Equation as Balance Equation

The forward equation in $\mathscr{M}$ can again be interpreted as a balance equation. When all measures have densities, i.e., when

$$\alpha(x, dy) = a(x, y)\, dy, \qquad \pi_t(dx) = p_t(x)\, dx,$$

this can be written as a functional integral equation: for $y \in \mathscr{V}$,

$$\frac{d}{dt} p_t(y) = \int_x a(x, y) p_t(x)\, dx - q(y) p_t(y) = \int_x \big(a(x, y) p_t(x) - a(y, x) p_t(y)\big)\, dx$$

(using $q(y) = \alpha(y, \mathscr{V}) = \int a(y, x)\, dx$). It is often obtained in this form in physics by balance arguments such as the following: the (massic, numerical, etc.) density $p_t(y)$ of particles at y increases due to arrivals from x at rate $p_t(x)$ (density at x) times $a(x, y)$ (transition rate from x to y) and decreases at rate $p_t(y)$ times $q(y) = \int_x a(y, x)\, dx$ (departure rate from y).

Generators and Martingales

Modern Markov process theory goes much further than these equations, and exploits strong links between Markov processes and martingales, established using the generator, which extend the Kolmogorov equations. Problem 5.3 provided a short initiation to this topic, which can readily be extended here.

6.3 Markov Processes Following an Ordinary Differential Equation Between Jumps: PDMP

A more general class of Markov processes is introduced. The processes in this class are often called piecewise deterministic Markov processes (PDMP).

One of our goals will be to obtain probabilistic representations for kinetic equations arising in statistical Mechanics, which thus can be given approximate solutions by Monte Carlo methods. These equations describe the time evolution of the instantaneous distribution of particles in phase space (position-velocity), when the position evolves according to the velocity, and the velocity jumps at random instants in function of the particle position and velocity.

6.3.1 Sample Paths, Evolution, Integro-Differential Generator

A family of deterministic evolutions between jumps will now be specified.

Definition 6.2 (Ordinary differential equation) Let $b : \mathbb{R}^d \to \mathbb{R}^d$. It is said that

$$(x_t)_{t \in I} : t \in I \mapsto x_t \in \mathbb{R}^d$$

is a *local solution, started at* x, on the interval of definition I, of the ordinary differential equation (ODE) with *vector field* b, if the interval of definition I contains the origin 0 and if, for all t in I,

$$\frac{d}{dt}x_t = b(x_t) \quad \text{and} \quad x_0 = x, \quad \text{and more precisely} \quad x_t = x + \int_0^t b(x_s)\,ds. \tag{6.10}$$

It is said that there is *uniqueness* if any two local solutions started at x *coincide* on the intersection of their intervals of definition, and then a *maximal* solution is defined on the reunion of all such intervals.

The function $b : \mathbb{R}^d \to \mathbb{R}^d$ is *Lipschitz* with constant $L < \infty$ if

$$\big|b(x) - b(y)\big| \le L|x - y|, \quad \forall x, y \in \mathbb{R}^d.$$

It is *locally Lipschitz* if for every $z \in \mathbb{R}^d$ there exists a neighborhood V_z of z such that the restriction of b on V_z is Lipschitz, or equivalently if the restriction of b on any compact set C is Lipschitz. If b is *differentiable* then b is locally Lipschitz, and if the differential is *bounded* then b is Lipschitz.

The function b has *at most affine growth* if there exist α, β in $\mathbb{R}$ such that

$$\big|b(x)\big| \le \alpha|x| + \beta, \quad \forall x \in \mathbb{R}^d,$$

which is the case if b is Lipschitz.

Theorem 6.8 (Cauchy–Lipschitz) *Consider the ODE of Definition* 6.2, *and assume that the vector field* b *is* locally Lipschitz. *Then*:

(a) *For every starting point* x, *there is* uniqueness *and* existence of a maximal solution *on a neighborhood of* 0.

(b) *If these maximal solutions are defined on* $\mathbb{R}_+$, *let* $\phi_t : x \in \mathbb{R}^d \mapsto \phi_t(x) \in \mathbb{R}^d$, *where* $(\phi_t(x))_{t\in\mathbb{R}_+}$ *is the solution starting at* x, *which satisfies*

$$\phi_t(x) = x + \int_0^t b\big(\phi_s(x)\big)\,ds.$$

This defines the flow of homeomorphisms $(\phi_t)_{t\in\mathbb{R}_+}$ *of the ODE*: *the* ϕ_t *are bijective bicontinuous mappings from* $\mathbb{R}^d$ *into* $\phi_t(\mathbb{R}^d)$ *and form a* semigroup, *i.e.*, $\phi_t \circ \phi_s = \phi_{t+s}$ *for* $t, s \ge 0$.

Proof (Main elements) Uniqueness and continuity use the Gronwall inequality. Existence uses the Picard iteration method. We refer to classic textbooks on ordinary differential equations. □

Definition of a Class of Markov Processes

It will always be assumed that the ODE in Definition 6.2 has a flow of homeomorphisms $(\phi_t)_{t\ge0}$, see Theorem 6.8 for some verifiable sufficient conditions.

We consider Markov processes with values in $\mathbb{R}^d$ which evolve solely under this flow and by *isolated jumps*, and which can specifically be written in the form, extending (6.3),

$$\begin{cases} X_t = \displaystyle\sum_{n\geq 0} \phi_{t-T_n}(X_{T_n})\mathbb{1}_{\{T_n\leq t<T_{n+1}\}} \quad \text{for } t\geq 0, \\ 0 = T_0 < T_1 \leq T_2 \leq \cdots \leq \infty, \quad \lim T_n = \infty, \\ T_n < \infty \Rightarrow T_n > T_{n-1} \quad \text{and} \quad X_{T_n} \neq X_{T_n-}. \end{cases} \tag{6.11}$$

For $n \geq 1$, if $T_n < \infty$ then it is an *actual* jump instant. The fact that these have no accumulation point is again a limitation. However, generally there is no natural definition for X_{T_n} if $T_n = \infty$.

In order to study these processes, we shall not try to be wholly rigorous, but will make rather general ad hoc assumptions so as to introduce the essential parameters. We shall later give *verifiable* sufficient hypotheses, which will allow to construct rigorously a sub-class of these processes, in a way appropriate for simulations.

Law of the First Jump Instant

Since T_1 is a stopping time, the Markov property of Theorem 6.1(b), implies for any u and v in $\mathbb{R}_+$ that

$$\begin{aligned} \mathbb{P}_x(T_1 > u+v) &= \mathbb{P}_x\big(T_1 > u, X_u = \phi_u(x), T_1 > u+v\big) \\ &= \mathbb{P}_x\big(T_1 > u, X_u = \phi_u(x)\big)\mathbb{P}_{\phi_u(x)}(T_1 > v) \\ &= \mathbb{P}_x(T_1 > u)\mathbb{P}_{\phi_u(x)}(T_1 > v) \end{aligned}$$

and the properties of the flow $(\phi_t)_{t\geq 0}$ and induction yield, for $n \geq 1$ and $\varepsilon > 0$,

$$\mathbb{P}_x(T_1 > n\varepsilon) = \prod_{k=1}^{n} \mathbb{P}_{\phi_{(k-1)\varepsilon}(x)}(T_1 > \varepsilon) = \exp\left(\sum_{k=1}^{n} \log \mathbb{P}_{\phi_{(k-1)\varepsilon}(x)}(T_1 > \varepsilon)\right).$$

As the right continuity of sample paths and the Monotone Convergence Theorem imply that

$$\forall y, \quad \lim_{\varepsilon\to 0^+} \mathbb{P}_y(T_1 > \varepsilon) = 1, \quad \mathbb{P}_y(T_1 > t) = \lim_{\varepsilon\to 0^+} \mathbb{P}_y\big(T_1 > \lceil t/\varepsilon\rceil\varepsilon\big),$$

we deduce

$$\log \mathbb{P}_x(T_1 > t) = \lim_{\varepsilon\to 0^+} \varepsilon \sum_{k=1}^{\lceil t/\varepsilon\rceil} \frac{\log \mathbb{P}_{\phi_{(k-1)\varepsilon}(x)}(T_1 > \varepsilon)}{\varepsilon}.$$

Heuristically, this leads us to *expect* that there exists $q : \mathbb{R}^d \to \mathbb{R}_+$ satisfying

$$\lim_{\varepsilon\to 0^+} \frac{\log \mathbb{P}_x(T_1 > \varepsilon)}{\varepsilon} = q(x), \quad \text{or equivalently} \quad \lim_{\varepsilon\to 0^+} \frac{\mathbb{P}_x(T_1 \le \varepsilon)}{\varepsilon} = q(x),$$

and further that

$$\mathbb{P}_x(T_1 > t) = \exp\Big(-\int_0^t q\big(\phi_s(x)\big)\,ds\Big). \tag{6.12}$$

Then, the law of T_1 on $\mathbb{R}_+ \cup \{\infty\}$ is given by

$$P_{T_1}(dt) = q\big(\phi_t(x)\big)\exp\Big(-\int_0^t q\big(\phi_s(x)\big)\,ds\Big)\,dt + \exp\Big(-\int_0^\infty q\big(\phi_s(x)\big)\,ds\Big)\delta_\infty(dt)$$

and

$$\begin{aligned}
\mathbb{P}_x(T_1 \le \varepsilon) &= q(x)\varepsilon + o(\varepsilon),\\
\mathbb{P}_x(T_2 \le \varepsilon) &= \int \mathbb{P}_x(t < T_2 \le \varepsilon \mid T_1 = t) P_{T_1}(dt)\\
&= \int_0^\varepsilon \mathbb{P}_{\phi_t(x)}(T_1 \le \varepsilon - t) P_{T_1}(dt) = o(\varepsilon),
\end{aligned}$$

so that

$$P_\varepsilon\big(x, \{\phi_\varepsilon(x)\}\big) = 1 - q(x)\varepsilon + o(\varepsilon).$$

Heuristic Computation of the Generator, Infinitesimal Evolution

The definition of the generator $\mathscr{A}$ and of its domain $\mathscr{D}(\mathscr{A})$ are given in (6.2), and

$$\begin{aligned}
P_\varepsilon f(x) - f(x) &= \int \big(f(y) - f(x)\big) P_\varepsilon(x, dy)\\
&= \big(f\big(\phi_\varepsilon(x)\big) - f(x)\big) P_\varepsilon\big(x, \{\phi_\varepsilon(x)\}\big)\\
&\quad + \int \big(f(y) - f(x)\big)\mathbb{1}_{\{y \ne \phi_\varepsilon(x)\}} P_\varepsilon(x, dy).
\end{aligned}$$

Note that $\lim_{\varepsilon\to 0^+} P_\varepsilon(x, \{\phi_\varepsilon(x)\}) = 1$. In addition, the chain rule yields that, for all bounded functions f with continuous bounded differentials,

$$\lim_{\varepsilon\to 0^+} \frac{f(\phi_\varepsilon(x)) - f(x)}{\varepsilon} = \sum_{i=1}^d b_i(x)\partial_i f(x).$$

Therefore, for all smooth enough functions f in $\mathscr{D}(\mathscr{A})$ and all x, the limit

$$\frac{1}{\varepsilon}\int \big(f(y) - f(x)\big)\mathbb{1}_{\{y \ne \phi_\varepsilon(x)\}} P_\varepsilon(x, dy)$$

exists. Heuristically, we *expect* that there exists a positive measure kernel $(\alpha(x,dy))_{x\in\mathbb{R}^d}$ such that

$$\lim_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\int \big(f(y)-f(x)\big)\mathbb{1}_{\{y\neq\phi_\varepsilon(x)\}}P_\varepsilon(x,dy)=\int_y \big(f(y)-f(x)\big)\alpha(x,dy).$$

Assuming that

$$\alpha\big(x,\{x\}\big)=0 \quad\text{and}\quad \alpha\big(x,\mathbb{R}^d\big)<\infty,$$

we obtain

$$q(x)=\alpha\big(x,\mathbb{R}^d\big).$$

Then the generator acts on f in $C_b^1(\mathbb{R}^d)$ as an integro-differential operator which can be given under various forms such as, with the classical notation ∇ for the gradient,

$$\begin{aligned}\mathscr{A}f(x)&=\sum_{i=1}^d b_i(x)\partial_i f(x)+\int_y \big(f(y)-f(x)\big)\alpha(x,dy)\\&=b(x)\cdot\nabla f(x)+\int_y \alpha(x,dy)f(y)-q(x)f(x).\end{aligned}\tag{6.13}$$

Note that the singularity in $\phi_\varepsilon(x)$, which converges to x, shows that this convergence can be problematic if there is an accumulation of "small jumps".

For a rigorous characterization of the generator, we refer to Davis [9, Chap. 2].

Remark 6.5 The probabilistic interpretation is that the Markov process is subject to two infinitesimal tendencies:

- a *continuous evolution* according to the flow $(\phi_t)_{t\geq 0}$ of the ODE with vector field b, called the *drift* of the process, which during an infinitesimal time dt tends to make it go from a state x to the state $x+b(x)\,dt$,
- a *jump evolution* determined equivalently by the positive measure jump kernel α, or by the jump intensities q and the jump law kernel Π, these quantities being related by

$$\alpha(x,dy)=q(x)\Pi(x,dy),\quad x\in\mathbb{R}^d,\tag{6.14}$$

which during an infinitesimal time dt tends to make it jump from a state x with probability $q(x)dt$ to go to a state chosen according to the law $\Pi(x,dy)$.

There is a kind of separation of scales: the evolution due to the ODE and the probability of undertaking a jump are of infinitesimal order dt, but the sizes of the jumps are macroscopic.

6.3.2 Construction, Simulation, Existence

Let be given b and α, or equivalently b and q and Π, see (6.14). These specify the generator $\mathscr{A}$ of the Markov process $(X_t)_{t\in\mathbb{R}_+}$ using (6.13), and describe its infinitesimal evolution tendencies.

We wish to construct the process, mathematically as well as in an effective way for simulations, under *verifiable* assumptions. We must hence concatenate two kinds of evolutions, which separately are well understood:

- The *continuous evolution* corresponding to the differential part of the generator. We shall assume that the ODE in Definition 6.2 has a flow of homeomorphisms $(\phi_t)_{t\geq 0}$, see Theorem 6.8 for some verifiable sufficient conditions.
- The *jump evolution* corresponding to the integral part of the generator. For effective simulation of jumps in the following mathematical constructions, it is advised to carefully reread Sects. 5.2.3 and 6.2.2, and even Sect. 4.3.

An actual simulation requires either to have an explicit expression for the flow, or to approximate it in a well-controlled fashion, for instant using an Euler discretization scheme, or higher order generalizations such as Runge–Kutta methods.

Several construction methods for the Markov process will now be described.

Fictitious Jump Method

Assume that

$$\sup_{x\in\mathbb{R}^d} q(x) := \sup_{x\in\mathbb{R}^d} \alpha\big(x, \mathbb{R}^d\big) < \infty. \tag{6.15}$$

The jump instants $(T_n)_{n\geq 1}$ of the Markov process $(X_t)_{t\in\mathbb{R}_+}$ will be chosen among those $(\tau_n)_{n\geq 1}$ of a Poisson process $(N_t)_{t\in\mathbb{R}_+}$ with finite intensity $\lambda \geq \sup_{x\in\mathbb{R}^d} q(x)$ by a sampling or erasing principle (a "fictitious jump method").

The sample path is iteratively constructed from one jump instant to the next jump instant of $(N_t)_{t\in\mathbb{R}_+}$, and if $(X_t)_{t\in\mathbb{R}_+}$ has value x just before one of these instants, then its value at that instant will be chosen according to the Markovian kernel

$$\hat{\Pi}(x, dy) = \frac{q(x)}{\lambda}\Pi(x, dy) + \left(1 - \frac{q(x)}{\lambda}\right)\delta_x(dy).$$

Here is a concrete implementation of this method.

Algorithm (Fictitious jump method) Draw the initial value X_0; then iteratively for $n \geq 1$, when the process is defined for $0 \leq t \leq \tau_{n-1}$:

- draw S_n according to an $\mathscr{E}(\lambda)$ exponential law and set

$$\begin{cases} \tau_n = \tau_{n-1} + S_n, \\ X_t = \phi_{t-\tau_{n-1}}(X_{\tau_{n-1}}) \quad \text{for } \tau_{n-1} \leq t < \tau_n, \text{ and thus } X_{\tau_n -} = \phi_{S_n}(X_{\tau_{n-1}}), \end{cases}$$

- – either, with probability $1 - \frac{q(X_{\tau_n-})}{\lambda}$, set $X_{\tau_n} = X_{\tau_n-}$,
  – or else (with complementary probability $\frac{q(X_{\tau_n-})}{\lambda}$) draw X_{τ_n} according to the law $\Pi(X_{\tau_n-}, dy)$,
  which determines the process for $0 \le t \le \tau_n$.

The τ_n are the jump instants of a Poisson process of intensity λ, and thus $\lim_{n\to\infty} \tau_n = \infty$ and hence $\lim_{n\to\infty} T_n = \infty$, so that this construction indeed defines a process $(X_t)_{t\in\mathbb{R}_+}$. This process is Markov and has generator (6.13). Indeed, $T_1 := \inf\{t > 0 : X_t \neq X_{t-}\} > t$ if and only if there is no true jump at the instants τ_n preceding t, and classically (see Problem 4.1)

$$\mathbb{P}(T_1 > t) = \sum_{n\ge 1} \mathrm{e}^{-\lambda t} \frac{\lambda^n t^n}{n!} \left(\frac{1}{t} \int_0^t \left(1 - \frac{q(\phi_s(x))}{\lambda} \right) ds \right)^n = \exp\left(- \int_0^t q\big(\phi_s(x)\big)\, ds \right)$$

so that (6.12) holds; moreover, the evolution on the time interval $[0, T_1[$ and the law of the first jump $X_{T_1} - X_{T_1-}$ are indeed as they should be; then, the strong Markov property of the Poisson process allows to conclude. Details are left as an exercise.

For an efficient simulation, a good upper bound λ for the jump intensities must be found, but there will always be a problem if the product of $\sup_{x\in\mathbb{R}^d} q(x)$ and the temporal horizon of the simulation is large.

In the latter case, if the function q varies greatly then the following methods might result in faster simulations.

Sub-domain Method

Assume that for a sub-domain $\mathscr{S}$ of $\mathbb{R}^d$ it holds that

$$\sup_{x\in\mathscr{S}} q(x) \le \sigma \ll \sup_{x\in\mathbb{R}^d} q(x) \le \lambda.$$

A Poisson process of intensity σ can then be used as long as the process evolves *within* $\mathscr{S}$; in between its jump instants, one must test whether the flow of homeomorphisms has remained inside $\mathscr{S}$, and else start the simulation anew from the instant and the state of exit (strong Markov property of the Poisson process).

A Poisson process of intensity λ must be used when the process is outside $\mathscr{S}$.

Here is a concrete implementation of this method when $\mathscr{S}$ is an open set.

Algorithm (Sub-domain method) Draw the initial value X_0; then iteratively for $n \ge 1$, when the process is defined for $0 \le t \le \tau_{n-1}$:

- if $X_{\tau_{n-1}}$ is in $\mathscr{S}$, draw S_n according to an $\mathscr{E}(\sigma)$ exponential law, set

$$\begin{cases} \tau_n' = \tau_{n-1} + S_n, \\ X_t = \phi_{t-\tau_{n-1}}(X_{\tau_{n-1}}) \quad \text{for } \tau_{n-1} \le t < \tau_n', \text{ and thus } X_{\tau_n'-} = \phi_{S_n}(X_{\tau_{n-1}}), \end{cases}$$

- check whether $(X_t)_{\tau_{n-1}<t\le\tau_n'}$ stays inside $\mathscr{S}$, and
  - if so, set $\tau_n = \tau_n'$ and:
    - either, with probability $1 - \frac{q(X_{\tau_n-})}{\sigma}$, set $X_{\tau_n} = X_{\tau_n-}$,
    - or else (with complementary probability $\frac{q(X_{\tau_n-})}{\sigma}$) draw X_{τ_n} according to the law $\Pi(X_{\tau_n-}, dy)$,
  - if not, define τ_n to be the first exit time of $\mathscr{S}$,
- if $X_{\tau_{n-1}}$ is not in $\mathscr{S}$ then proceed as in the previous algorithm (with $\mathscr{E}(\lambda)$),

which determines the process for $0 \le t \le \tau_n$.

In order to try to do even better, one can proceed to decompose the state space in sub-domains within which the function q does not vary much. For instance, as long as the process evolves within $\{x \in \mathbb{R}^d : n-1 \le q(x) \le n\}$, a Poisson process with intensity n can be used; or else, as long as the process evolves within $\{x \in \mathbb{R}^d : n-1 \le \|x\| \le n\}$, a Poisson process with intensity $\sup_{n-1\le\|x\|\le n} q(x)$.

Starting from a sub-domain, one must test at each draw of a new jump instant whether the flow remains inside, and else start the simulation anew from the instant and the state of exit into a sub-domain in which q can take higher values, using the *higher* bound for q there for the next jump instant. This last point is important: if $\mathscr{S}$ were closed, then the previous algorithm would remain trapped at the boundary.

This method can be implemented even if $\sup_{x\in\mathbb{R}^d} q(x) = \infty$ if q is locally bounded, and results in an actual process $(X_t)_{t\in\mathbb{R}_+}$ if there is no explosion, i.e.,

$$\lim_{n\to\infty} T_n = \infty \quad \text{a.s.},$$

which is automatic if $\sup_{x\in\mathbb{R}^d} q(x) < \infty$.

It is a simple matter then to prove that $(X_t)_{t\in\mathbb{R}_+}$ is a Markov process of generator given by (6.13), and we leave this as an exercise.

From a practical point of view in actual simulations, problems arise from the multiplication of tests needed to check when the process exits a given sub-domain, and the restarts at the instants of exit into a domain with larger bound on q.

True Jump Method

Let us push this idea to its limit, or take inspiration in pure jump processes and in (6.12), or in Problem 4.3. The jump instants $(T_n)_{n\ge1}$ of the Markov process $(X_t)_{t\in\mathbb{R}_+}$ can be obtained by a *time change* from the jump instants $(\tau_n)_{n\ge1}$ of a Poisson process with intensity 1; all these will be used, hence the name "true jump method". The sample path is iteratively constructed from one jump instant to the next jump instant.

Algorithm (True jump method) Draw the initial value X_0; then iteratively for $n \ge 1$, when the process is defined for $0 \le t \le T_{n-1}$:

- draw S_n according to an $\mathscr{E}(1)$ exponential law and set

$$\begin{cases} T_n = \inf\left\{u > T_{n-1} : \displaystyle\int_{T_{n-1}}^{u} q\left(\phi_s(X_{T_{n-1}})\right) ds = S_n\right\}, \\ X_t = \phi_{t-T_{n-1}}(X_{T_{n-1}}) \quad \text{for } T_{n-1} \le t < T_n, \text{ and thus } X_{T_n-} = \phi_{T_n-T_{n-1}}(X_{T_{n-1}}), \end{cases}$$

- draw X_{T_n} according to the law $\Pi(X_{T_n-}, dy)$,

which determines the process for $0 \le t \le T_n$.

This method can be implemented even if $\sup_{x\in\mathbb{R}^d} q(x) = \infty$, when q is locally bounded, and results in an actual process $(X_t)_{t\in\mathbb{R}_+}$ if there is no explosion, i.e.,

$$\lim_{n\to\infty} T_n = \infty \quad \text{a.s.},$$

which is automatic if $\sup_{x\in\mathbb{R}^d} q(x) < \infty$.

It is a simple matter then to prove that $(X_t)_{t\in\mathbb{R}_+}$ is a Markov process of generator given by (6.13), and we leave this as an exercise.

From a practical point of view in actual simulations, problems arise from the necessity to compute integrals along the sample paths in order to determine the T_n by time-change.

Remark 6.6 The fictitious jump method is reminiscent of the rejection method (Sect. 2.2.5), the sub-domain method of the stratification variance reduction technique (Remark 3.4), and the true jump method of the c.d.f. inversion method (Theorem 2.3), but the justification goes much deeper.

6.3.3 Kolmogorov Equations, Feynman–Kac Formula

All statements in Sect. 6.2.3 can be adapted to this framework.

Note that if b does not constantly vanish, then the integro-differential operator cannot be a bounded operator, and hence does not belong to the corresponding Banach algebra; nevertheless, the results are true in a weakened form. No rigorous statement or proof will be given, even though these can be obtained rather simply if the flow of homeomorphisms $(\phi_t)_{t\ge0}$ of the ODE is well defined (see Theorem 6.8, e.g.), and the integral part of the operator is bounded (see (6.15)).

An integro-differential arising in applications may have the requisite structure to be interpreted as a forward or backward equation for a Markov process in the class we have just studied. This *probabilistic representation* allows its approximate solution by Monte Carlo methods, and can also be useful for theoretical results.

Such equations, notably those of the forward Kolmogorov type, are often obtained by balance considerations. An example will be detailed for a kinetic equation in the next subsection.

Recall that the generator $\mathscr{A}$ of the Markov process acts on f in $C_b^1(\mathbb{R}^d)$ as the integro-differential operator, given in (6.13),

$$\mathscr{A} f(x) = \sum_{i=1}^{d} b_i(x)\partial_i f(x) + \int_y \big(f(y) - f(x)\big)\alpha(x, dy).$$

Its adjoint $\mathscr{A}^*$ acting on measures is obtained by duality for the bracket (6.1).

The adjoint of the integral part of $\mathscr{A}$ has already been established in (6.8). The adjoint of its differential part can be written, in distributional sense as

$$(b_i\partial_i)^*\mu = -\partial_i(b_i\mu),$$

this derivative of a measure being interpreted by its action on appropriate test-functions f as

$$\big\langle \partial_i(b_i\mu), f\big\rangle = -\big\langle (b_i\partial_i)^*\mu, f\big\rangle = -\langle \mu, b_i\partial_i f\rangle$$

(extending integration by parts when μ has a differentiable density and b is differentiable and f has bounded support). Thus, $\mathscr{A}^*$ is the *integro-differential* operator

$$\begin{aligned}\mathscr{A}^*\mu(dx) &= -\sum_{i=1}^{d} \partial_i\big(b_i(x)\mu(dx)\big) + \int_y \big(\alpha(y, dx)\mu(dy) - \alpha(x, dy)\mu(dx)\big)\\ &= -\sum_{i=1}^{d} \partial_i\big(b_i(x)\mu(dx)\big) + \int_y \alpha(y, dx)\mu(dy) - q(x)\mu(dx). \qquad (6.16)\end{aligned}$$

If $\mu(dx) = m(x)\,dx$ and $\alpha(y, dx) = a(y, x)\,dx$ then $\mathscr{A}^*$ is made to operate on the densities by setting (with a classic abuse of notation)

$$\mathscr{A}^*\mu(dx) = \mathscr{A}^*\big(m(x)\,dx\big) = \big(\mathscr{A}^*m(x)\big)\,dx$$

and, since $\partial_i(b_i(x)m(x)) = b_i(x)\partial_i m(x) + \partial_i b_i(x)m(x)$,

$$\begin{aligned}\mathscr{A}^*m(x) &= -\sum_{i=1}^{d} b_i(x)\partial_i m(x) - \sum_{i=1}^{d} \partial_i b_i(x)m(x)\\ &\quad + \int_y \big(a(y, x)m(y) - a(x, y)m(x)\big)\,dy\\ &= -\sum_{i=1}^{d} b_i(x)\partial_i m(x) + \int_y a(y, x)m(y)\,dy - \left(\sum_{i=1}^{d} \partial_i b_i(x) + q(x)\right)m(x)\\ &= -b(x)\cdot\nabla m(x) + \int_y a(y, x)m(y)\,dy - \big(\nabla\cdot b(x) + q(x)\big)m(x)\end{aligned} \qquad (6.17)$$

which features the divergence $\nabla\cdot b$ of the vector field b.

6.3.4 Application to Kinetic Equations

Section 2.1.3 will now be revisited and completed, with some change of notation. A reference for what follows is the book by Lapeyre et al. [32].

Kinetic equations (also called transport equations in some situations) are statistical physics equations, which notably appear in nuclear physics models. Indistinguishable particles (neutrons, photons, charged particles, ...) evolve, according to classical mechanics and its fundamental principle, in a medium. The medium interacts with the particles by submitting them to accelerations which change their speeds. The effect of the particles on the medium is neglected.

Any particle is represented in phase space by its position $r \in \mathbb{R}^3$ and its velocity $v \in \mathbb{R}^3$ at time $t \in \mathbb{R}_+$. This allows to obtain a first order differential equation for the motion, starting from the fundamental principle of dynamics. The point in phase space can be denoted as

$$x = (r, v) = (r_1, r_2, r_3, v_1, v_2, v_3) \in \mathbb{R}^6$$

and adequate notation will be used, such as for the ODE vector field

$$b(x) = b(r, v) = \big(b_r(r, v), b_v(r, v)\big) = \big(v, b_v(r, v)\big) \in \mathbb{R}^6. \tag{6.18}$$

Here $b_r(r, v) = v$ since the position of the particle evolves according to its velocity, and $b_v(r, v)$ is the acceleration of the particle caused by a force field, due to the medium, which may depend both on velocity and on position.

Given the huge number of particles (typically of the order of the Avogadro number), it is impossible and useless to keep track of all their positions and velocities, and a statistical mechanics perspective is taken.

Specifically, a limit procedure is performed in an adequate scale (determined by the physics of the problem), in which the number of particles goes to *infinity* and some significant quantities remain *fixed*.

After this limit is taken, from a *probabilistic* perspective the study bears on the probability $\pi_t(dr, dv)$ of finding in position r a particle of velocity v at time t, and the corresponding density $p_t(r, v)$ if it exists. Physicists and numerical analysts usually consider the density (e.g., of mass or of number) $f_t(r, v)$ of particles in position r of velocity v at time t. An equation derived in this limit for the evolution of these laws or densities is called a *kinetic equation*.

Some physical phenomena are such that the particle velocities vary progressively for macroscopic durations, at the end of which they undergo *brusque* deviations due to *rare* but *brutal* interactions with the medium, which are called *collisions*. In such media, which are called collisional, there is a separation of scale between the smooth and the brusque variations of velocities.

Then, the *mean free path* is defined as the typical distance a particle travels before its velocity is significantly deviated (e.g., above some threshold). A limit is taken in which this important quantity remains fixed, after which the particle will travel a random distance of order this quantity before having its velocity undergo a *jump*.

The jump measure and jump law kernels act only on the velocity variable, and

$$\alpha\big(r, v, dr'\, dv'\big) = \delta_r\big(dr'\big)\alpha\big(r, v, dv'\big), \qquad q(r, v) = \int \alpha\big(r, v, dv'\big),$$

$$\Pi\big(r, v, dr'\, dv'\big) = \delta_r\big(dr'\big)\Pi\big(r, v, dv'\big) = \delta_r\big(dr'\big)\frac{\alpha(r, v, dv')}{q(r, v)},$$

with the abuse of notation $\alpha(r, v, dv') = \int_{r'} \alpha(r, v, dr'\, dv')$. The jump measure kernels are obtained by physical considerations, and often have densities:

$$\alpha\big(r, v, dv'\big) = a\big(r, v, v'\big)\, dv'.$$

The medium is said to be purely collisional if in the limit the velocity evolves only by jumps, i.e., if

$$b(x) = b(r, v) = (v, 0)$$

so that the particle position evolves with constant speed between the velocity jumps (this is called free transport, and the equation is called a transport equation). The general form (6.18) allows to take into account continuous velocity variations due to force fields, for instance an external gravitation field, and for a charged particle an electrical and a magnetic field.

The *kinetic* or *transport equations* in statistical physics are given for laws $\pi_t(dr, dv)$, and for their densities $p_t(r, v)$ if these exist: $\pi_t(dr, dv) = p_t(r, v)\, dr\, dv$.

These *integro-differential* equations correspond to the forward Kolmogorov (or Fokker–Planck) equations for the integro-differential operator $\mathscr{A}$ in (6.13) corresponding to the above situation, and feature the adjoint operator $\mathscr{A}^*$, which also is an integro-differential operator given in (6.16) and (6.17). For densities, these equations are often written as

$$\begin{aligned}\frac{\partial}{\partial t} p_t(r, v) + v \cdot \nabla_r p_t(r, v) + b_v(r, v) \cdot \nabla_v p_t(r, v) + \big(\nabla_v \cdot b_v(r, v)\big) p_t(r, v) \\ = \int_{v'} a\big(r, v', v\big) p_t\big(r, v'\big)\, dv' - q(r, v)\, p_t(r, v)\end{aligned} \tag{6.19}$$

where the r.h.s. contains the collision terms. In the l.h.s., the first two terms are the only ones in the purely collisional case $b_v(r, v) = 0$ (free transport).

Kinetic Equation as a Balance Equation

These equations are often derived directly by balance considerations from the models in statistical mechanics.

In this specific situation, the collisional term derivation (and interpretation) is sometimes obscured by the symmetries in physical collisions w.r.t. inversion of time. The particle density $p_t(r, v)$ in position r and with velocity v has rates of

- increase, due to particles in the same position r and with velocity v' before a collision which sends them into velocity v, which happens at rate $p_t(r, v')a(r, v', v)$, and this should be integrated over all possible v':

$$\int_{v'} a\big(r, v', v\big)\, dt\; p_t\big(r, v'\big)\, dv',$$

- decrease, due to particles with the same position r and velocity v before a collision which sends them into some other velocity v':

$$p_t(r, v) \int_{v'} a\big(r, v, v'\big)\, dv' = p_t(r, v) q(r, v).$$

The differential terms can be interpreted analogously.

A heuristic derivation from balance considerations of the equation, using the infinitesimal point of view on Markov processes, goes as follows: up to $o(dt)$ terms,

$$\begin{aligned} p_{t+dt}(r, v) - p_t(r, v) &= p_t(r - v\, dt, v) - p_t(r, v) + p_t\big(r, v - b_v(r, v)\, dt\big) - p_t(r, v) \\ &\quad + \int_{v'} a\big(r, v', v\big)\, dt\; p_t\big(r, v'\big)\, dv' - q(r, v)\, dt\; p_t(r, v) \end{aligned}$$

where the first two terms on the r.h.s. come respectively from transport of the position according to velocity, and to velocity change due to its drift (due to a force field); the last term (already computed above) comes from collisional change of velocity. By a first order Taylor expansion and the chain rule, up to $o(dt)$ terms,

$$p_t(r - v\, dt, v) = p_t(r, v) - v \cdot \nabla_r p_t(r, v)\, dt,$$

$$p_t\big(r, v - b_v(r, v)\, dt\big) = p_t(r, v) - \big(b_v(r, v) \cdot \nabla_v p_t(r, v) + \big(\nabla_v \cdot b_v(r, v)\big) p_t(r, v)\big)\, dt,$$

so that dividing through by dt tending to 0 yields (6.19) after some rearrangement.

Remark 6.7 Such a heuristic derivation requires little probabilistic knowledge. However, the Markovian interpretation captures much more of the underlying statistical physics models, and is quite appropriate for these phenomena. It allows a much richer description than the kinetic equation, which is solely for instantaneous laws, whereas the law of the process carries incomparably more information, and notably allows to consider first hitting times, maxima, etc., in scenarios of interest.

Monte Carlo Methods vs. Curse of Dimensionality

This class of integro-differential equations arises in important physical and industrial applications. These equations have the requisite structure so as to be interpreted as forward or backward equations for a Markov process, which we have described in detail. This probabilistic representation is of no great surprise for a probabilist, considering the underlying statistical mechanics models.

Such an equation can then thus be given a *probabilistic representation*, which allows its approximate solution by Monte Carlo methods.

The efficient computation of approximations for solutions has *colossal* implications for scientific as well as technological and industrial issues. Such Monte Carlo methods are intensively (and almost exclusively) used, since the high dimension renders deterministic methods prohibitive.

Indeed, these integro-differential equations were introduced here in dimension 6 for simplicity. More elaborate physical models may be in a much higher dimension d, since they may take into account degrees of freedom or parameters such as internal energies or relative positions of atoms in molecules, or (even worse but quite natural) mixtures of multi-type particles.

Deterministic methods are most often computationally untractable due to their prohibitive cost in computing time and memory, since these typically scale like $1/\varepsilon^d$ if ε is the space step-size.

This is an example of what is called the "curse of dimensionality".

Probabilistic representations allow to develop and prove convergence of Monte Carlo methods for approximating the solutions of these equations, and give rates of convergence. These methods scale well with dimension, since typically only the variance is affected. They are much used in practice since they are virtually the only efficient methods.

Physicists have intuitively introduced what they called "Direct Simulation Monte Carlo" (DSMC) methods, by arguing that they are similar to the physical phenomena. Numerical analysts have sometimes tried to justify these methods by reducing them to a mere probabilistic approximate quadrature method for the integral part of the operator.

All this does not capture the subtlety of the stochastic Markovian approach, which sets up an appropriate framework and allows to speak of important pathwise quantities, outside the scope of a deterministic approach, such as hitting times, extrema, etc.

6.3.5 Further Extensions

Boltzmann Equations

The Boltzmann equation has been derived for the study of similar phenomena, in which particles evolve in a medium composed of like particles, resulting in feedback. The interaction between similar particles must be studied.

It is a nonlinear equation, in which the collision operator is quadratic in the law or its density, due to the fact that the evolution of the density in a point depends on collisions with similar particles having the same density.

Specifically, the term $a(r, v', v)$ in (6.19), giving the effect of the medium on a particle in r and with velocity v', is replaced by a term of the form

$$\int_{v^*} a\big(r, v^*, v', v\big)\, p_t\big(r, v^*\big)\, dv^*$$

since in position r, the transition of the velocity of a particle from v' to v is due to a collision with a particle of velocity v^*, which is there with density $p_t(r, v^*)$. Its rigorous study is very arduous due to the local behavior in r of the collision term, all the more so that the coefficients obtained in the limit from physical interactions have singularities of varied types.

Nevertheless Monte Carlo methods can be adapted to this equation. They cannot any longer be based on independence of particles and the Strong Law of Large Numbers, but on a related asymptotic property called propagation of chaos; they are often called mean-field Monte Carlo methods.

An important reference book on the Boltzmann equation is Cercignani, Illner and Pulvirenti [7]. Remarkable probabilistic studies on mean-field models, including those related to this equation, can be found in Sznitman [45] and the references therein. See also, e.g., Graham and Méléard [22] and the references therein.

Again, the efficient computation of approximations for solutions has *colossal* scientific and technological implications, and Monte Carlo methods are intensively (and almost exclusively) used.

Linearized versions of the Boltzmann equation (for example around a stationary solution) are very similar to transport equations. This is why the latter are sometimes called *linear Boltzmann equations*, but this oxymoron can be confusing.

Accumulation of Jumps, Compensated Jumps

The most simple and natural assumption on the jump measure kernel α is that it is constituted of *finite* measures, and its mathematical construction and simulation are greatly simplified under the uniform bound assumption (6.15). The latter assumption ensures that the jump instants are isolated, and enables to recursively construct the process from jump to jump.

However, when the integral operator corresponding to the generator given in Theorem 6.2, which is also the integral part of the operator (6.13), is written in the form

$$\int \big(f(y) - f(x)\big)\alpha(x, dy)$$

it acts on functions f in C_b^1 if the kernel α of measures is not necessarily finite but satisfies the weaker integrability assumption

$$\int \min\big(1, |y - x|\big)\alpha(x, dy) < \infty.$$

Inspired by the central limit theorem, martingales, Itô's stochastic calculus, and the differential terms in the generator, one can even try to go further by "compensating" the small jumps by their mean. One must then arbitrarily separate small jumps from large ones, and introduce an operator of the form

$$\int \big(f(y) - f(x) - \mathbb{1}_{|y-x|<1}(y - x) \cdot \nabla f(x)\big)\alpha(x, dy),$$

which acts on f in C_b^2 under the still weaker integrability assumption

$$\int \min\bigl(1, |y-x|^2\bigr)\alpha(x, dy) < \infty.$$

The probabilistic interpretation of such kernels is that they describe accumulations of small jumps, of which there may be a *summable infinity* in a bounded time interval. The jump instants are no longer isolated, but may accumulate in finite time; nevertheless, as in Zeno's paradox, the Markov process may still proceed beyond that time. The mathematical construction and the effective simulation of Markov processes having such a generator is extremely delicate, and there are few general assumptions guaranteeing its existence.

Problems 6.2 and 6.3 will give some indications about this problem.

6.4 Problems

6.1 (Binary Energy Exchanges) A system of $N \geq 2$ particles is modeled by a Markov process $(X_t)_{t\in\mathbb{R}_+}$ with values in $\mathbb{R}_+^N$, in which $X_t = (X_t^i)_{1\leq i\leq N} = (X_t^1, \ldots, X_t^N)$ and $X_t^i \in \mathbb{R}_+$ represents the state (energy) of particle number i at time t.

It evolves under the superposition of gains, losses, and binary exchanges of energy, which are described as follows, where $\alpha \geq 0$, $\beta \geq 0$, and $\gamma \geq 0$ are parameters, and all draws are independent:

(a) Each particle has a dedicated Poisson process with intensity α, at the jump instants of which its energy is incremented by 1.
(b) Each particle of energy x loses a fraction θ chosen uniformly in $[0, 1]$ of it to the medium after an exponential duration of parameter βx.
(c) For each $1 \leq i < j \leq N$, the two particles of indices i and j, while in states x and y, interact by exchanging energy after an exponential duration of intensity γxy, and then particle i takes a state z uniformly chosen in $[0, x + y]$ and particle j takes state $x + y - z$.

1. Compute the generator of $(X_t)_{t\in\mathbb{R}_+}$. Is it bounded?
2. Prove rapidly that $(X_t)_{t\in\mathbb{R}_+}$ is well defined.
   Hint: Introduce $B_k := \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 + \cdots + x_n \leq k\}$ for $k \in \mathbb{N}$.
3. Describe a natural simulation method for this process, and show that it may be implemented using summations over N terms at each iteration.
4. Suggest a simple fictitious jump method using fewer computations at each iteration. What could you try to possibly further accelerate the method?

6.2 (Jump Accumulation) Let $0 = a_0 < a_1 < a_2 < \cdots$ be a sequence converging to 1. Let $(N_t^n)_{t\geq 0}$ and $(N_t)_{t\geq 0}$ be Poisson processes of intensities λ_n and λ satisfying $\sum_{n\geq 1} \frac{1}{\lambda_n} < \infty$, and U_k for $k \geq 1$ be uniform on $[0, 1]$ for $k \geq 1$. Let X_0 take values

in $[0, 1]$, all these processes and r.v. be independent, and $(X_t)_{t\in\mathbb{R}_+}$ be given by

$$X_t = X_0 + \sum_{0\le s\le t} (X_s - X_{s-}),$$

$$X_s - X_{s-} = \sum_{n\ge 1} \mathbb{1}_{\{a_{n-1}\le X_{s-}<a_n\}} \left(\frac{a_n - X_{s-}}{a_n - a_{n-1}} a_n + \frac{X_{s-} - a_{n-1}}{a_n - a_{n-1}} a_{n+1} - X_{s-} \right)$$
$$\times \left(N^n_s - N^n_{s-}\right) - \mathbb{1}_{\{X_{s-}=1\}} U_{N_s} (N_s - N_{s-}).$$

The justifications may be given in a not wholly rigorous fashion.

1. Describe the process evolution.
2. Show that it is a well-defined Markov process on $[0, 1]$. Compute its generator and discuss its domain.
3. Say whether the process can be simulated exactly, or approximately, and explain how.

6.3 (Generalized Kac Equation ($\star$)) The velocities of an interacting system of $N \ge 2$ particles evolve as follows.

A Poisson process with intensity 1 is dedicated to *each* particle, and at each jump instant the particle collides with a particle chosen uniformly among the $N - 1$ *other* particles.

At each such collision, if the (real-valued) velocities of the particles just before collision were $v_1 \in \mathbb{R}$ and $v_2 \in \mathbb{R}$, then the collision transforms them respectively into

$$v_1' = v_1 \cos\theta - v_2 \sin\theta, \qquad v_2' = v_1 \sin\theta + v_2 \cos\theta,$$

for an angle θ drawn in $[-\pi, \pi]$ according to a probability measure $\beta(\theta)\, d\theta$ with even density:

$$\beta : \theta \in [-\pi, \pi] \to \beta(\theta) \in \mathbb{R}_+, \qquad \beta(\theta) = \beta(-\theta), \qquad \|\beta\|_1 := \int_{-\pi}^{\pi} \beta(\theta)\, d\theta = 1.$$

Let $(V_t)_{t\in\mathbb{R}_+}$ denote the $\mathbb{R}^N$-valued process given by $V_t = (V_t^k)_{1\le i\le N}$, where V_t^k denotes the velocity of particle number k at time t.

Let $v = (v_k)_{1\le k\le N}$ be the generic point in $\mathbb{R}^N$, and for $1 \le i \le N$ let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^N$, with a 1 in ith position, denote the ith vector of the canonical basis.

1. Prove that the kinetic energy $K_t = \sum_{k=1}^N (V_t^k)^2$ of the system is constant.
2. (a) Describe how to efficiently simulate $(V_t)_{t\in\mathbb{R}_+}$. Which problems may arise?
   (b) Prove that $(V_t)_{t\in\mathbb{R}_+}$ is a pure jump Markov process, with generator $\mathscr{A}$ acting on $f \in L^\infty(\mathbb{R}^N)$ as follows: for $v \in \mathbb{R}^N$,

$$\mathscr{A} f(v) = \frac{2}{N-1} \sum_{1 \le i < j \le N} \int_{-\pi}^{\pi} \big\{ f\big(v + \big[v_i(\cos\theta - 1) - v_j \sin\theta\big] e_i + \big[v_i \sin\theta + v_j(\cos\theta - 1)\big] e_j\big) - f(v) \big\} \beta(\theta)\, d\theta.$$

Is this Markov process well defined?

3. Let $h : v = (v_k)_{1 \le k \le N} \mapsto h(v) = \sum_{k=1}^{N} v_k$. Let $M_t = h(V_t)$ be the impetus of the system, and $m_t = \mathbb{E}(M_t)$ its expectation, $t \ge 0$.
   (a) Compute $\mathscr{A} h$ in terms of the parameter $c = \int_{-\pi}^{\pi} \cos\theta \beta(\theta)\, d\theta$.
   (b) Prove that
   $$\frac{d}{dt} m_t = 2(c-1) m_t, \quad t \ge 0,$$
   then that $m_t = m_0 \mathrm{e}^{2(c-1)t}$.
   (c) Prove that $c < 1$.

In the sequel, β does not necessarily satisfy $\|\beta\|_1 := \int_{-\pi}^{\pi} \beta(\theta)\, d\theta = 1$.

4. Assume here that $0 < \|\beta\|_1 < \infty$.
   (a) Is $(V_t)_{t \in \mathbb{R}_+}$ a well-defined pure jump Markov process with isolated jumps?
   (b) Describe how to simulate it efficiently.
   (c) What practical problem arises when $\|\beta\|_1$ is large?

5. It is now assumed that $\|\beta\|_1 = \infty$ but $\int_{-\pi}^{\pi} \theta \beta(\theta)\, d\eta < \infty$.

   The operator $\mathscr{A}$ is made to act on the space $C_b^1(\mathbb{R}^N)$ of real bounded functions with continuous and bounded partial derivatives.

   For $0 < \varepsilon \le \pi$ let $b_\varepsilon = 2 \int_0^\varepsilon \theta^2 \beta(\theta)\, d\theta < \infty$, and define $\mathscr{A}_\varepsilon$ acting on $f \in C_b^1(\mathbb{R}^N)$ as follows: for $v \in \mathbb{R}^N$,

$$\mathscr{A}_\varepsilon f(v) = \frac{2}{N-1} \sum_{1 \le i < j \le N} \int_{-\pi}^{\pi} \big\{ f\big(v + \big[v_i(\cos\theta - 1) - v_j \sin\theta\big] e_i + \big[v_i \sin\theta + v_j(\cos\theta - 1)\big] e_j\big) - f(v) \big\} \beta(\theta) \mathbb{1}_{|\theta| \ge \varepsilon}\, d\theta - b_\varepsilon \sum_{i=1}^{N} v_i \frac{\partial}{\partial v_i} f(v).$$

   (a) Prove that $\mathscr{A} f \in L^\infty(\mathbb{R}^N)$ for $f \in C_b^1(\mathbb{R}^N)$.
   (b) Prove that $\lim_{\varepsilon \to 0} \mathscr{A}_\varepsilon f(v) = \mathscr{A} f(v)$ for $f \in C_b^1(\mathbb{R}^N)$ and $v \in \mathbb{R}^N$.
   (c) To what kind of Markov process does the generator $\mathscr{A}_\varepsilon$ correspond, and is it well defined?
   (d) How can one simulate it efficiently? What practical problem arises when ε is small?

Chapter 7
Discretization of Stochastic Differential Equations

Abstract This chapter develops discretization schemes for stochastic differential equations and their applications to the probabilistic numerical resolution of deterministic parabolic partial differential equations. It starts with some important properties of Itô's Brownian stochastic calculus, and the existence and uniqueness theorem for stochastic differential equations with Lipschitz coefficients. Then, using probabilistic techniques only, existence, uniqueness, and smoothness properties are proved for solutions of parabolic partial differential equations. To this end, we show that stochastic differential equations with smooth coefficients define stochastic flows, and we prove some properties of such flows. We are then in a position to prove an optimal convergence rate result for the discretization schemes.

7.1 Reminders on Itô's Stochastic Calculus

Notable reference books on Brownian motion and on Itô's stochastic integral calculus are Ikeda and Watanabe [23] and Karatzas and Shreve [26].

7.1.1 Stochastic Integrals and Itô Processes

In this chapter and the next, it is always assumed that a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is given, as well as a filtration on it, i.e., a family $(\mathscr{F}_t) := (\mathscr{F}_t)_{t\in\mathbb{R}_+}$ of time-indexed sub-σ-fields of $\mathscr{F}$ satisfying

$$0 \le s \le t \Rightarrow \mathscr{F}_s \subset \mathscr{F}_t.$$

For technical reasons and without loss of generality, it is further assumed that the filtration satisfies the so-called *usual conditions*: for every t in $\mathbb{R}_+$, $\mathscr{F}_t$ contains all the $\mathbb{P}$-negligible sets in $\mathscr{F}$, and $\mathscr{F}_t = \mathscr{F}_{t+} := \bigcap_{\theta>t} \mathscr{F}_\theta$. The collection $(\Omega, \mathscr{F}, \mathbb{P}, (\mathscr{F}_t))$ is called a filtered probability space.

A stochastic process $(\phi_t) := (\phi_t)_{t\in\mathbb{R}_+}$ taking values in $\mathbb{R}^d$ equipped with its Borel sigma-field $\mathscr{B}(\mathbb{R}^d)$ is said to be $(\mathscr{F}_t)$-adapted if the random variable ϕ_t is $\mathscr{F}_t$-measurable for every t. To simplify, most often we write "adapted" without

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*, Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_7, © Springer-Verlag Berlin Heidelberg 2013

mentioning the underlying filtration. The maps

$$t \in \mathbb{R}_+ \mapsto \phi_t(\omega)$$

are called the *trajectories* or the *sample paths* of (ϕ_t), and play an important role in its analysis. A process is said to be continuous [resp. right-continuous] if, ω-a.s., its trajectories are continuous [resp. right-continuous]. To avoid heavy technicalities, we *implicitly* limit ourselves to processes (ϕ_t) which are *measurable*, in the sense that the mapping

$$(t, \omega) \in \mathbb{R}_+ \times \Omega \mapsto \phi_t(\omega)$$

is $\mathscr{B}(\mathbb{R}_+) \otimes \mathscr{F}$-measurable. This is not a severe restriction: in particular, all the processes which we need to consider in practice satisfy it.

A very important process in all the sequel is the following.

Definition 7.1 (Standard Brownian motion) A real process $(W_t) := (W_t)_{t \geq 0}$ is a standard Brownian motion, or a Wiener process, if it is $(\mathscr{F}_t)$-adapted and continuous and

(i) $W_0 = 0$, $\mathbb{P}$-a.s.,
(ii) for all $0 \leq s \leq t$, the random variable $W_t - W_s$ is independent of $\mathscr{F}_s$,
(iii) for all $0 \leq s \leq t$, the random variable $W_t - W_s$ has a $\mathscr{N}(0, t-s)$ Gaussian distribution (zero mean, variance $t - s$).

Remark 7.1 If $(W_t)_{t \geq 0}$ is a standard Brownian motion, then $(W_{t+h} - W_t)_{h \geq 0}$ is a standard Brownian motion independent of $\mathscr{F}_t$ for every $t \geq 0$. This remains true if t is replaced by an a.s. finite stopping time.

Itô's Stochastic Integral

Now consider a finite time horizon $T > 0$. Our goal is to construct stochastic integrals w.r.t. a standard Brownian motion (W_t) for an appropriate class of real processes. This will be then be straightforwardly extended to vector-valued Brownian motions and processes at the end of this section.

Consider the class $\mathscr{L}^2_{\mathscr{F}}(0, T)$ of the adapted processes (ϕ_t) such that

$$\int_0^T (\phi_\theta)^2 \, d\theta < +\infty, \quad \mathbb{P}\text{-a.s.}$$

An example of such a process is an *elementary* process, which is a piecewise constant process of the type

$$\bar{\phi}_t(\omega) = \bar{\bar{\phi}}_0(\omega) \mathbb{1}_{\{0\}}(t) + \sum_{k=0}^{n-1} \bar{\bar{\phi}}_k(\omega) \mathbb{1}_{(t_k, t_{k+1}]}(t),$$

where $0 = t_0 < t_1 < \cdots < t_n = T$ and for all $k = 0, \ldots, n-1$ the random variable $\bar{\bar{\phi}}_k$ is $\mathscr{F}_{t_k}$-measurable. The *Itô stochastic integral process* of an elementary process $(\bar{\phi}_t)$ w.r.t. (W_t) is defined as

$$\forall 0 \le t \le T, \quad I_t(\bar{\phi}) := \sum_{k=0}^{n-1} \bar{\bar{\phi}}_k(\omega)(W_{t \wedge t_{k+1}} - W_{t_k \wedge t}).$$

For any process (ϕ_t) in $\mathscr{L}^2_{\mathscr{F}}(0,T)$, it is possible to construct a sequence of elementary processes $(\bar{\phi}^n_t)$ such that $\int_0^T |\phi_s - \bar{\phi}^n_s|^2 \, ds$ converges to 0 in probability (for $\mathbb{P}$), and that there exists an adapted continuous process $(I_t(\phi))$, which does not depend on the particular such sequence, such that

$$\lim_{n \to \infty} \sup_{0 \le t \le T} \left| I_t(\phi) - I_t\left(\bar{\phi}^n\right) \right| = 0, \quad \text{in probability for } \mathbb{P}.$$

For all t in the time interval $[0, T]$, by definition

$$\int_0^t \phi_\theta \, dW_\theta := I_t(\phi)$$

is called the *Itô stochastic integral* of (ϕ_t) w.r.t. (W_t), which defines a continuous adapted process. By construction, this integral is linear w.r.t. the integrated process: for any (ϕ_t) and (ψ_t) in $\mathscr{L}^2_{\mathscr{F}}(0,T)$ and real a and b,

$$\forall 0 \le t \le T, \quad \int_0^t (a\phi_\theta + b\psi_\theta) \, dW_\theta = a \int_0^t \phi_\theta \, dW_\theta + b \int_0^t \psi_\theta \, dW_\theta.$$

A more sophisticated property concerns stopped stochastic integrals. If τ is a stopping time for the filtration $(\mathscr{F}_t)$, then

$$\forall 0 \le t \le T, \quad \int_0^{t \wedge \tau} \phi_\theta \, dW_\theta = \int_0^t \phi_{\theta \wedge \tau} \, dW_\theta \tag{7.1}$$

(it is straightforward to extend Definition 2.4 in continuous time).

The Itô stochastic integral on $\mathscr{L}^2_{\mathscr{F}}(0,T)$ does not enjoy many other properties which allow one to develop exact computations, notably because its expectation may not be defined.

However, in practice the process which must be integrated often belongs to a large subclass of $\mathscr{L}^2_{\mathscr{F}}(0,T)$, namely the class $\mathscr{M}^2_{\mathscr{F}}(0,T)$ constituted of the adapted processes (ϕ_t) such that

$$\mathbb{E} \int_0^T (\phi_\theta)^2 \, d\theta < +\infty.$$

The corresponding stochastic integrals have remarkable properties. For any process (ϕ_t) in $\mathscr{M}^2_{\mathscr{F}}(0,T)$, it is possible to construct a sequence of elementary pro-

cesses $(\bar{\phi}_t^n)$ which are bounded in (t, ω), and moreover such that the convergence is stronger than the previous one:

$$\lim_{n\to\infty} \mathbb{E}\int_0^T \left|\phi_\theta - \bar{\phi}_\theta^n\right|^2 d\theta = 0.$$

Martingale Properties

The previous approximation result allows to prove the following fundamental theorem on Itô stochastic integrals, which is at the basis of stochastic calculus.

Theorem 7.1 *Let (W_t) and $(\tilde{W}_t)$ be two independent standard Brownian motions. Let (ϕ_t) and (ψ_t) be two processes in $\mathscr{M}^2_{\mathscr{F}}(0,T)$. Then, $\mathbb{P}$-a.s.,*

$$\forall 0 \le s \le t \le T, \quad \mathbb{E}\left[\int_s^t \phi_\theta \, dW_\theta \mid \mathscr{F}_s\right] = 0, \tag{7.2a}$$

$$\forall 0 \le s \le t \le T, \quad \mathbb{E}\left[\int_s^t \phi_\theta \, dW_\theta \int_s^t \psi_\theta \, dW_\theta \mid \mathscr{F}_s\right] = \mathbb{E}\left[\int_s^t \phi_\theta \psi_\theta \, d\theta \mid \mathscr{F}_s\right], \tag{7.2b}$$

and

$$\forall 0 \le s \le t \le T, \quad \mathbb{E}\left[\int_s^t \phi_\theta \, dW_\theta \int_s^t \psi_\theta d\tilde{W}_\theta \mid \mathscr{F}_s\right] = 0. \tag{7.2c}$$

Remark 7.2 Under the assumptions of Theorem 7.1, let

$$M_t^\phi := \int_0^t \phi_\theta \, dW_\theta, \qquad M_t^\psi := \int_0^t \psi_\theta \, dW_\theta, \qquad \tilde{M}_t^\psi := \int_0^t \psi_\theta \, d\tilde{W}_\theta.$$

Equation (7.2a) indicates that (M_t^ϕ) is a continuous $(\mathscr{F}_t)$-martingale; so are (M_t^ψ) and $(\tilde{M}_t^\psi)$. By definition, the Doob–Meyer brackets $(\langle M^\phi, M^\psi\rangle_t)$ and $(\langle M^\phi, \tilde{M}^\psi\rangle_t)$ are the continuous processes with bounded variation and vanishing at time 0 such that $(M_t^\phi M_t^\psi - \langle M^\phi, M^\psi\rangle_t)$ and $(M_t^\phi \tilde{M}_t^\psi - \langle M^\phi, \tilde{M}^\psi\rangle_t)$ are martingales. Equalities (7.2b) and (7.2c) show that

$$\left\langle M^\phi, M^\psi\right\rangle_t = \int_0^t \phi_\theta \psi_\theta \, d\theta, \qquad \left\langle M^\phi, \tilde{M}^\psi\right\rangle_t = 0.$$

Notably, $(M_t^\phi)_{t\in[0,T]}$ is square integrable and $\langle M^\phi\rangle_t = \int_0^t (\phi_\theta)^2 \, d\theta$.

The toolbox of martingale theory now opens for us. Let us recall a few important results which will be used, some of them directly in terms of the Itô integrals involved. The first one is the direct consequence of (7.1) and Theorem 7.1, or of classic results on stopped martingales.

Theorem 7.2 (Optional sampling) *Under the assumption of Theorem 7.1, if τ is a $(\mathscr{F}_t)$-stopping time, then*

$$\forall 0 \le s \le t \le T, \quad \mathbb{E}\left[\int_{s\wedge\tau}^{t\wedge\tau} \phi_\theta \, dW_\theta \mid \mathscr{F}_s\right] = 0,$$

$$\forall 0 \le s \le t \le T, \quad \mathbb{E}\left[\left(\int_{s\wedge\tau}^{t\wedge\tau} \phi_\theta \, dW_\theta\right)^2 \mid \mathscr{F}_s\right] = \mathbb{E}\left[\int_{s\wedge\tau}^{t\wedge\tau} (\phi_\theta)^2 \, d\theta \mid \mathscr{F}_s\right].$$

The following one can be applied to the Itô integrals in Remark 7.2; see, e.g., Karatzas and Shreve [26, Thm 1.3.8].

Theorem 7.3 (Doob's L^p maximal inequalities) *Let $(M_t)_{t\ge 0}$ be a martingale with right-continuous paths. For any real number $p > 1$, it holds that*

$$\mathbb{E}\Big(\sup_{0\le s\le t} |M_s|^p\Big) \le \left(\frac{p}{p-1}\right)^p \mathbb{E}\big(|M_t|^p\big).$$

Some more advanced martingale inequalities come last; see, e.g., Karatzas and Shreve [26, Thm 3.3.28].

Theorem 7.4 (Burkholder–Davis–Gundy (BDG) inequalities) *Let $m \ge 1$ be an arbitrary real number. There exist universal constants k_m and K_m such that, under the assumption of Theorem 7.1,*

$$k_m \mathbb{E}\left[\left(\int_0^t (\phi_\theta)^2 \, d\theta\right)^m\right] \le \mathbb{E}\left[\sup_{0\le s\le t}\left|\int_0^s \phi_\theta \, dW_\theta\right|^{2m}\right] \le K_m \mathbb{E}\left[\left(\int_0^t (\phi_\theta)^2 \, d\theta\right)^m\right].$$

Exercise 7.1 Deduce the BDG inequality for $m = 1$ from the Doob inequalities, giving the best constants k_1 and K_1 you can.

Multidimensional Extension

Vector-valued extensions are straightforward. Let $d, r \ge 1$ be two integers.

An r-dimensional, or $\mathbb{R}^r$-valued, standard Brownian motion $(W_t)_{t\ge 0}$ is of the form $W_t = (W_t^1, \ldots, W_t^r)$ where $(W_t^1)_{t\ge 0}, \ldots, (W_t^r)_{t\ge 0}$ are r independent standard Brownian motions. This is preserved by change of orthonormal basis.

A process $(\sigma_t) = (\sigma_t^{ij})$ with values in the space $\mathbb{R}^{d\otimes r}$ of matrices with d rows and r columns is in $\mathscr{L}^2_{\mathscr{F}}(0,T)$ or in $\mathscr{M}^2_{\mathscr{F}}(0,T)$ if all its terms (or coordinates) are, and then the d-dimensional Itô integral of (σ_t) w.r.t. (W_t) is defined by

$$\int_0^t \sigma_\theta \, dW_\theta := \left(\sum_{j=1}^r \int_0^t \sigma_\theta^{ij} \, dW_\theta^j\right)_{1\le i\le d}.$$

7.1.2 Itô's Formula, Existence and Uniqueness of Solutions of Stochastic Differential Equations

Let (b_t) be an $\mathbb{R}^d$-valued adapted process such that

$$\int_0^T |b_\theta|\, d\theta < \infty, \quad \mathbb{P}\text{-a.s.},$$

and $(\sigma_t) = (\sigma_t^{ij})$ be an $\mathbb{R}^{d\otimes r}$-valued process such that each coordinate process (σ_t^{ij}) belongs to $\mathscr{L}^2_{\mathscr{F}}(0,T)$. The $\mathbb{R}^{d\otimes d}$-valued process $a_t := \sigma_t \cdot \sigma_t^*$ has symmetric non-negative values; σ_t^* denotes the transposed matrix of σ_t, and we use matrix multiplication.

Consider an $\mathbb{R}^r$-valued standard Brownian motion $(W_t)_{t\geq 0}$ and an $\mathbb{R}^d$-valued and $\mathscr{F}_0$-measurable random vector X_0. Let (X_t) be the $\mathbb{R}^d$-valued process defined by

$$\forall 1 \leq i \leq d, \quad X_t^i = X_0^i + \int_0^t b_\theta^i\, d\theta + \sum_{j=1}^r \int_0^t \sigma_\theta^{ij}\, dW_\theta^j \tag{7.3a}$$

which can be written in vector form as

$$X_t = X_0 + \int_0^t b_\theta\, d\theta + \int_0^t \sigma_\theta\, dW_\theta. \tag{7.3b}$$

This is called an Itô process.

Itô's Formula for Change of Variable

Given a smooth function $u(t,x)$, the following fundamental theorem allows to express $u(t,X_t)$ by means of stochastic integrals; see, e.g., Karatzas and Shreve [26, Thm 3.3.3]. For a real function $u : (t,x) \in [0,T]\times\mathbb{R}^d \mapsto u(t,x)$ of class $\mathscr{C}^2$ in the space coordinate $x = (x_i)_{1\leq i\leq d}$, the corresponding partial derivatives are denoted by

$$\partial_i u := \frac{\partial u}{\partial x_i}, \quad \partial^2_{ij} u := \frac{\partial^2 u}{\partial x_i \partial x_j}, \quad 1 \leq i,j \leq d.$$

Theorem 7.5 (Itô's formula) *Let $u(t,x)$ be a real function in $\mathscr{C}^{1,2}([0,T]\times\mathbb{R}^d)$. Let $(X_t)_{t\geq 0}$ be given as in (7.3a). Then, $\mathbb{P}$-a.s., for all t in $[0,T]$,*

$$\begin{aligned} u(t,X_t) = u(0,X_0) &+ \int_0^t \frac{\partial u}{\partial\theta}(\theta,X_\theta)\, d\theta + \sum_{i=1}^d \int_0^t \partial_i u(\theta,X_\theta) b_\theta^i\, d\theta \\ &+ \sum_{i=1}^d \sum_{j=1}^r \int_0^t \partial_i u(\theta,X_\theta)\sigma_\theta^{ij}\, dW_\theta^j + \frac{1}{2}\sum_{i,k=1}^d \int_0^t \partial^2_{ik} u(\theta,X_\theta) a_\theta^{ik}\, d\theta. \end{aligned} \tag{7.4}$$

The equality (7.4) is the celebrated Itô formula, often written compactly as

$$du(\theta, X_\theta) = \frac{\partial u}{\partial \theta}(\theta, X_\theta)\, d\theta + \nabla u(\theta, X_\theta) b_\theta\, dX_\theta + \frac{1}{2} \operatorname{Trace}\big(\nabla\nabla^* u(\theta, X_\theta) a_\theta\big)\, d\theta$$

where ∇u denotes the (column) vector in $\mathbb{R}^d$ with coordinates ∂_i, accordingly $\nabla\nabla^* u$ denotes the $d \times d$ matrix with entries $\partial^2_{ij} u$. This can be seen as an extension of the chain rule for differentiation to stochastic integrals.

Stochastic Differential Equations

The next theorem provides useful existence and uniqueness results.

Theorem 7.6 *Assume that a filtered probability space* $(\Omega, \mathscr{F}, \mathbb{P}, (\mathscr{F}_t))$ *and an* $\mathbb{R}^r$*-valued standard Brownian motion* (W_t) *are given. Let* b *be a Lipschitz function from* $\mathbb{R}^d$ *to* $\mathbb{R}^d$*, and* σ *a Lipschitz function from* $\mathbb{R}^d$ *to* $\mathbb{R}^{d\otimes r}$*.*

Then, there is pathwise existence and uniqueness *for the following* stochastic differential equation:

For any $\mathbb{R}^d$*-valued random variable* X_0 *which is* $\mathscr{F}_0$*-measurable* (*thus independent of* (W_t)) *and has finite second moment* $\mathbb{E}|X_0|^2 < \infty$*, and any arbitrary deterministic final time* $T > 0$*, there exists a unique* $\mathbb{R}^d$*-valued process* (X_t) *in* $\mathscr{M}^2_{\mathscr{F}}(0, T)$ *such that,* $\mathbb{P}$*-a.s., for all* t *in* $[0, T]$*,*

$$\forall 1 \le i \le d, \quad X^i_t = X^i_0 + \int_0^t b^i(X_\theta)\, d\theta + \sum_{j=1}^{r} \int_0^t \sigma^{ij}(X_\theta)\, dW^j_\theta, \tag{7.5a}$$

or in vector notation

$$X_t = X_0 + \int_0^t b(X_\theta)\, d\theta + \int_0^t \sigma(X_\theta)\, dW_\theta, \tag{7.5b}$$

in short $dX_t = b(X_t)\, dt + \sigma(X_t)\, dW_t$*.*

Moreover, there is uniqueness in law: *the law of* (X_t) *depends only on* b*,* σ*, and the law of* X_0*, but not on the specific choice of* $(\Omega, \mathscr{F}, \mathbb{P}, (\mathscr{F}_t), (W_t))$*.*

The process (X_t) is called the unique strong or pathwise solution of the stochastic differential equation (7.5a)–(7.5b) starting at X_0.

The proof of the preceding theorem uses the Itô formula, martingale inequalities, and (as for Theorem 6.8) the Gronwall inequality for uniqueness and a Picard approximation sequence for existence: see, e.g., Sect. 5.2 in Karatzas and Shreve [26]. This "constructive" procedure yields that, ω-a.s., $(X_t(\omega))$ can be expressed as a measurable functional of $X_0(\omega)$ and the trajectory $(W_t(\omega))$. This functional is not explicit, but allows to express the law of (X_t) as the image measure of the law of $(X_0, (W_t))$, yielding uniqueness in law.

The fact that pathwise uniqueness implies uniqueness in law is actually a special case of a very general result, the Yamada–Watanabe theorem, see Karatzas and Shreve [26, Prop 5.3.20].

Remark 7.3 Theorem 7.6 provides an existence and uniqueness result for strong solutions, that is, when the filtered probability space and the Brownian motion are fixed. A related question can be asked: given b, σ, and the *law* of X_0, can one *find* $(\Omega, \mathscr{F}, \mathbb{P}, (\mathscr{F}_t), (W_t))$ and an adapted process (X_t) such that (7.5a)–(7.5b) holds? The quantity of interest is then the *law* of the process (X_t). Such solutions in law are often referred to as *weak* solutions. For a discussion of these issues, see, e.g., Karatzas and Shreve [26, Sects. 5.2 and 5.3].

Definition 7.2 Given a time origin $0 \le s$, we define the flow $(X_t(x))$ from time s to time $T > s$ as the family of solutions (when the deterministic initial condition x varies) to the equations

$$X_t = x + \int_s^t b(X_\theta)\, d\theta + \int_s^t \sigma(X_\theta)\, dW_\theta, \quad s \le t \le T. \tag{7.6}$$

7.1.3 Markov Properties, Martingale Problems and Fokker–Planck Equations

Under the assumptions of Theorem 7.6, for all $t, h \ge 0$ it holds that

$$X_{t+h} = X_t + \int_t^{t+h} b(X_\theta)\, d\theta + \int_t^{t+h} \sigma(X_\theta)\, dW_\theta.$$

Let us admit for the time being that X_t has a second moment (which will be proved soon), and recall Remarks 7.1 and 7.3.

Then, $(X_{t+h})_{h\ge 0}$ is conditionally independent of $\mathscr{F}_t$ knowing X_t, and has the same law as the solution of the stochastic differential equation starting at X_t.

Hence, $(X_t)_{t\ge 0}$ is a $(\mathscr{F}_t)$-Markov process, see Definition 6.1 and Theorem 6.1. It is even strong Markov, see Remark 7.1.

Moreover, Itô's formula yields for every real f in $C_b^2(\mathbb{R}^d)$ and $t \ge 0$ that

$$f(X_t) = f(X_0) + \sum_{i=1}^d \int_0^t b^i(X_\theta)\partial_i f(X_\theta)\, d\theta + \sum_{i=1}^d \sum_{j=1}^r \int_0^t \sigma^{ij}(X_\theta)\partial_i f(X_\theta)\, dW_\theta^j$$
$$+ \frac{1}{2} \sum_{i,k=1}^d \int_0^t a^{ik}(X_\theta)\partial_{ik}^2 f(X_\theta)\, d\theta$$

and, defining the second order differential operator

$$\mathscr{L} = \sum_{i=1}^{d} b^i(\cdot)\partial_i + \frac{1}{2}\sum_{i,k=1}^{d} a^{ik}(\cdot)\partial^2_{ik}$$

which acts on f in $C^2_b(\mathbb{R}^d)$ by

$$\mathscr{L} f : x \in \mathbb{R}^d \mapsto \sum_{i=1}^{d} b^i(x)\partial_i f(x) + \frac{1}{2}\sum_{i,k=1}^{d} a^{ik}(x)\partial^2_{ik} f(x), \tag{7.7}$$

this can be written as

$$f(X_t) = f(X_0) + \int_0^t \mathscr{L} f(X_\theta)\, d\theta + \sum_{i=1}^{d}\sum_{j=1}^{r} \int_0^t \sigma^{ij}(X_\theta)\partial_i f(X_\theta)\, dW^j_\theta. \tag{7.8}$$

Note that the sum of stochastic integrals may seem messy, but it defines a *martingale*. We have just proved that (X_t) (more precisely its law) satisfies a martingale problem.

Martingale problems have become a central topic in modern probability theory for the study of Markov processes, and are profoundly related to weak solutions such as in Remark 7.3.

Generator and Fokker–Planck Equation

Taking expectation of (7.8) yields that, for all f in $C^2_b(\mathbb{R}^d)$,

$$\mathbb{E} f(X_t) = \mathbb{E} f(X_0) + \int_0^t \mathbb{E}\mathscr{L} f(X_\theta)\, d\theta \tag{7.9}$$

and hence, for $X_0 = x$ and $t = \varepsilon > 0$,

$$\frac{\mathbb{E} f(X_\varepsilon) - f(x)}{\varepsilon} = \frac{1}{\varepsilon}\int_0^\varepsilon \mathbb{E}\mathscr{L} f(X_\theta)\, d\theta \mapsto [\varepsilon \to 0]\mathscr{L} f(x)$$

so that, considering (6.2), the generator of the Markov process $(X_t)_{t\geq 0}$ is (an extension of) $\mathscr{L}$, acting on a domain containing $C^2_b(\mathbb{R}^d)$.

Such a Markov process is often called a diffusion process.

More can be said. Let π_t denote the law of X_t. Using the duality bracket (6.1) given by integration, we can rewrite (7.9) as

$$\langle \pi_t, f\rangle = \langle \pi_0, f\rangle + \int_0^t \langle \pi_\theta, \mathscr{L} f\rangle\, d\theta.$$

This is a weak formulation, with C_b^2 test functions, for the Fokker–Planck equation for measures

$$\frac{\partial \pi_t}{\partial t} = \mathscr{L}^* \pi_t$$

where the adjoint $\mathscr{L}^*$ is taken w.r.t. the duality bracket with $C_b^2(\mathbb{R}^d)$ and can be written in the sense of the distributions, as

$$\mathscr{L}^* \pi_t(dx) = -\sum_{i=1}^{d} \partial_i \big[b^i(x)\pi_t(dx)\big] + \frac{1}{2}\sum_{i,k=1}^{d} \partial_{ik}^2 \big[a^{ik}(x)\pi_t(dx)\big].$$

This is a forward Kolmogorov equation, akin to those we have already encountered in Theorems 5.8 and 6.7.

If b is C^1 and a is C^2 and $\pi_t(dx) = p_t(x)dx$ for a C^2 density $p(\cdot)$ then (p_t) satisfies the Fokker–Planck partial differential equation

$$\frac{\partial p_t}{\partial t} = \mathscr{L}^* p_t(x) = -\sum_{i=1}^{d} \partial_i \big[b^i(x)\, p_t(x)\big] + \frac{1}{2}\sum_{i,k=1}^{d} \partial_{ik}^2 \big[a^{ik}(x)\, p_t(x)\big].$$

Reciprocally, given a parabolic partial differential equation of one of the above forms (for instance derived by physical balance considerations), we can give it a probabilistic representation using the solution $(X_t)_{t\geq 0}$ of the corresponding stochastic differential equation. This allows us to approximate the solution of the parabolic equation by a Monte Carlo method: for any f in C^b,

$$\langle \pi_t, f\rangle = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f\big(X_t^{(i)}\big), \quad \text{a.s.},$$

where the $(X_t^{(i)})_{t\geq 0}$ are independent copies of $(X_t)_{t\geq 0}$.

Since most stochastic differential equations cannot be solved explicitly, the issue is now to be able to effectively simulate good approximations $(X_t^n)_{t\geq 0}$ for $(X_t)_{t\geq 0}$, where n denotes a precision parameter destined to become large.

The next section introduces such approximations. Their convergence properties will then be studied in detail. This will be applied to the analysis of Monte Carlo methods using the probabilistic representation given by the Feynman–Kac formula, the transposition for diffusion processes of Corollaries 5.1 and 6.1 related to the backward Kolmogorov equations in Theorems 5.8 and 6.7.

Generalizations

The study in Sect. 6.3 can be adapted to obtain a Markov process $(X_t)_{t\in\mathbb{R}_+}$ which evolves according to a stochastic differential equation between isolated jumps. The

generator $\mathscr{A}$ of the resulting process can be seen to act on f in $C_b^2(\mathbb{R}^d)$, generalizing (7.7) into

$$\mathscr{A} f(x) = \sum_{i=1}^{d} b^i(x)\partial_i f(x) + \frac{1}{2}\sum_{i,k=1}^{d} a^{ik}(x)\partial_{ik}^2 f(x) + \int \big(f(y) - f(x)\big)\alpha(x, dy).$$

Many operators issuing from physics exhibit such a combination of an integral term and a second order differential term. The latter is usually interpreted as a diffusive term, due for instance to an accumulation of small jumps on velocity, or to analogies with the heat equation.

This can happen, for various models, in the kinetic equations we have studied in Sect. 6.3.4, for the velocity as well as for the position variable, yielding for instance so-called *transport-diffusion* equations.

7.2 Euler and Milstein Schemes

When the function σ is null, (7.5a)–(7.5b) is an ordinary differential equation; even then, for most functions b an exact solution (X_t) cannot be exhibited under the form $X_t = \phi(X_0, t)$ with a function ϕ explicitly defined using usual functions only. The situation is even worse when σ is not null, and seldom can one write $X_t = \phi(X_0, t, (W_s, s \le t))$ for some function ϕ satisfying the same constraint as above.

It is thus necessary to approximate (X_t) numerically, in a way for which simulations can be implemented simply and quickly on a computer.

In our introduction we have insisted on the fact that, in practice, one simulates stochastic processes with the aim of computing information relating to their law, that is, computing by Monte Carlo methods expectations of the type $\mathbb{E} f(X_T)$ or, more generally, $\mathbb{E} f(X_{T_1}, \dots, X_{T_k})$ or $\mathbb{E} f(X_s, 0 \le s \le T)$, where f is a function respectively defined on $\mathbb{R}^d$ or $(\mathbb{R}^d)^k$ or $\mathscr{C}([0, T], \mathbb{R}^d)$.

Our first objective is to construct discrete time Markov chains X^n which are easy to simulate[1] and are such that $\mathbb{E} f(X_T^n) \simeq \mathbb{E} f(X_T)$ for all f in a large class of functions. The processes X^n which we will construct depend on a step-size T/n, where n is destined to become large.

Our second objective is to estimate the approximation error $\mathbb{E} f(X_T^n) - \mathbb{E} f(X_T)$ in terms of f and n and T.

The convergence rate results which will be proved below sustain the following good news: without particular reason due to local singularities or non-linear growth of the coefficients b and σ, one can obtain satisfying convergence rates by using the simpler possible approximation method, the *Euler scheme*.

[1] This means that there exists a low complexity algorithm for generating sequences of independent draws of sample paths of X^n.

Let us intuitively derive this scheme. Observing that

$$X_{(p+1)T/n} = X_{pT/n} + \int_{pT/n}^{(p+1)T/n} b(X_s)\,ds + \int_{pT/n}^{(p+1)T/n} \sigma(X_s)\,dW_s$$
$$\simeq X_{pT/n} + b(X_{pT/n})\frac{T}{n} + \sigma(X_{pT/n})(W_{(p+1)T/n} - W_{pT/n}) \quad (7.10)$$

justifies the following discretization scheme.

Algorithm (Euler's scheme) Select $T > 0$ and $n \geq 1$, and construct iteratively $(X^n_{pT/n}, 0 \leq p \leq n)$ by setting

$$\begin{cases} X^n_0 = X_0, \\ X^n_{(p+1)T/n} = X^n_{pT/n} + b\big(X^n_{pT/n}\big)\dfrac{T}{n} + \sigma\big(X^n_{pT/n}\big)\big(W_{(p+1)T/n} - W_{pT/n}\big). \end{cases} \quad (7.11)$$

For actual simulation of one realization, draw a sample for X_0 from the initial law, then independent samples of $W_{(p+1)T/n} - W_{pT/n}$ for $p = 0, \ldots, n-1$ from the $\mathcal{N}(0, T/n)$ Gaussian law (zero mean, variance T/n).

Note that no partial derivative of the coefficients b and σ appears in the scheme.

We now construct another discretization scheme, the *Milstein scheme*, by adding terms to the expansion in (7.10). A naive analysis shows that the drift increment is of order T/n and the Brownian increment is of order $\sqrt{T/n}$, and we complement only the latter.

Denote by σ_j the jth column vector of the matrix σ and by $\partial\sigma_j$ its Jacobian matrix $(\frac{\partial\sigma^i_j}{\partial x_n})_{1\leq i,n\leq d}$. Applying Itô's formula to $\sigma(X_s) - \sigma(X_{pT/n})$ yields

$$X_{(p+1)T/n} = X_{pT/n} + \int_{pT/n}^{(p+1)T/n} b(X_s)\,ds + \int_{pT/n}^{(p+1)T/n} \sigma(X_s)\,dW_s$$
$$\simeq X_{pT/n} + b(X_{pT/n})\frac{T}{n} + \sigma(X_{pT/n})(W_{(p+1)T/n} - W_{pT/n})$$
$$+ \sum_{j,k=1}^{r} \partial\sigma_j(X_{pT/n})\sigma_k(X_{pT/n}) \int_{pT/n}^{(p+1)T/n} \big(W^k_s - W^k_{pT/n}\big)\,dW^j_s \quad (7.12)$$

and this leads to the following new scheme.

Algorithm (Milstein's scheme) Select $T > 0$ and $n \geq 1$, and construct iteratively $(\bar{X}^n_{pT/n}, 0 \leq p \leq n)$ by setting

$$\begin{cases} \bar{X}_0^n = X_0, \\ \bar{X}_{(p+1)T/n}^n = \bar{X}_{pT/n}^n + b\big(\bar{X}_{pT/n}^n\big)\dfrac{T}{n} + \displaystyle\sum_{j=1}^{r} \sigma_j\big(\bar{X}_{pT/n}^n\big)\big(W_{(p+1)T/n}^j - W_{pT/n}^j\big) \\ \qquad + \displaystyle\sum_{j,k=1}^{r} \partial\sigma_j\big(\bar{X}_{pT/n}^n\big)\sigma_k\big(\bar{X}_{pT/n}^n\big) \int_{pT/n}^{(p+1)T/n} \big(W_s^k - W_{pT/n}^k\big)\, dW_s^j. \end{cases} \tag{7.13}$$

For actual simulation of one realization, it is necessary to be able to efficiently draw samples for the last stochastic integrals.

This scheme has been introduced by Milstein to improve the Euler scheme convergence rate in $L^p(\Omega)$ norm, see below. Unfortunately, its interest is more theoretical than practical since, when $r > 1$, its algorithmic complexity is substantially larger than the complexity of the Euler scheme.

Indeed, in addition to the fact that it requires the explicit knowledge or the numerical approximation of partial derivatives of the functions σ_j (which may have a prohibitive cost in the simulation of large dimensional processes), the Milstein scheme requires to simulate the many stochastic integrals

$$\int_{pT/n}^{(p+1)T/n} \big(W_s^k - W_{pT/n}^k\big)\, dW_s^j$$

and major difficulties arise for $k \neq j$: the joint probability distribution of these integrals and of the increments $W_{(p+1)T/n}^k - W_{pT/n}^k$ seems impossible to simulate exactly.

In addition, even if we knew how to simulate this joint probability distribution, we would not improve the Euler scheme convergence rate for the approximation of $\mathbb{E}f(X_T)$: we will see that, for the family of problems of interest, the Milstein and Euler schemes generically are of identical accuracy.

When $r = 1$, the Milstein scheme simplifies notably and can easily be simulated. Indeed, the Itô formula yields that

$$\int_{pT/n}^{(p+1)T/n} (W_s - W_{pT/n})\, dW_s = \frac{1}{2}\left[(W_{(p+1)T/n} - W_{pT/n})^2 - \frac{T}{n} \right]$$

and the Milstein scheme simplifies into

$$\begin{aligned} \bar{X}_{(p+1)T/n}^n = \bar{X}_{pT/n}^n + b\big(\bar{X}_{pT/n}^n\big)\frac{T}{n} + \sigma\big(\bar{X}_{pT/n}^n\big)(W_{(p+1)T/n} - W_{pT/n}) \\ + \frac{1}{2}\sigma\big(\bar{X}_{pT/n}^n\big)\sigma'\big(\bar{X}_{pT/n}^n\big)\left((W_{(p+1)T/n} - W_{pT/n})^2 - \frac{T}{n} \right) \end{aligned} \tag{7.14}$$

which requires only the simulation of the Brownian increments.

A similar situation arises for $r > 1$ when the column vectors of the matrix σ satisfy the restrictive algebraic condition

$$\forall i,\ \forall j, \quad \partial\sigma_j(\cdot)\sigma_i(\cdot) = \partial\sigma_i(\cdot)\sigma_j(\cdot). \tag{7.15}$$

This condition is obviously satisfied when $r = 1$ and when the function σ is constant.

Indeed, Itô's formula implies for $j \neq k$ that

$$\int_{pT/n}^{(p+1)T/n} \left(W_s^k - W_{pT/n}^k\right) dW_s^j + \int_{pT/n}^{(p+1)T/n} \left(W_s^j - W_{pT/n}^j\right) dW_s^k = \left(W_{(i+1)T/n}^k - W_{pT/n}^k\right)\left(W_{(p+1)T/n}^j - W_{pT/n}^j\right)$$

and if (7.15) holds then the Milstein scheme simplifies into

$$\begin{aligned}\bar{X}_{(p+1)T/n}^n &= \bar{X}_{pT/n}^n + b\left(\bar{X}_{pT/n}^n\right)\frac{T}{n} + \sigma\left(\bar{X}_{pT/n}^n\right)\left(W_{(p+1)T/n} - W_{pT/n}\right) \\ &\quad + \sum_{k=2}^{r}\sum_{j<k} \partial\sigma_j\left(\bar{X}_{pT/n}^n\right)\sigma_k\left(\bar{X}_{pT/n}^n\right)\left(W_{(p+1)T/n}^k - W_{pT/n}^k\right)\left(W_{(p+1)T/n}^j - W_{pT/n}^j\right) \\ &\quad + \frac{1}{2}\sum_{j=1}^{r} \partial\sigma_j\left(\bar{X}_{pT/n}^n\right)\sigma_j\left(\bar{X}_{pT/n}^n\right)\left[\left(W_{(p+1)T/n}^j - W_{pT/n}^j\right)^2 - \frac{T}{n}\right] \qquad (7.16)\end{aligned}$$

which requires only the simulation of the Brownian increments.

A beautiful result by Clark and Cameron [8] shows that, under the commutativity hypothesis (7.15), the Milstein scheme leads to the optimal convergence rate in $L^2(\Omega)$ norm among all the discretization schemes which use only random variables equal to the increments of the process $(W(t))$ at times $\{pT/n, 0 \leq p \leq n\}$.

7.3 Moments of the Solution and of Its Approximations

We now show that the solution of a stochastic differential equation with Lipschitz coefficients has the same moments as its initial condition, and then that the Euler scheme shares the same property uniformly w.r.t. n.

Notation 7.7 *From now on, the Euclidean norm $|\cdot|$ on $\mathbb{R}^d$ and the norm $\|M\|^2 := \mathrm{Trace}(MM^*) = \mathrm{Trace}(M^*M)$ for matrices M will be used. In addition, C will denote any constant depending only on the discretization parameter n. The value of C may change from line to line, but in every instance it remains uniform w.r.t. n.*

Let us now recall the classic Gronwall inequality, which will be of much use.

Lemma 7.1 (Gronwall's inequality) *Let g be a continuous function on $[0, T]$. Suppose that there exists a constant $\beta > 0$ and a positive integrable function α on $[0, T]$*

such that

$$g(t) \le \alpha(t) + \beta \int_0^t g(s)\, ds, \quad \forall 0 \le t \le T.$$

Then

$$g(t) \le \alpha(t) + \beta \int_0^t \alpha(s) \mathrm{e}^{\beta(t-s)s}\, ds, \quad \forall 0 \le t \le T.$$

If in addition α is an increasing function,

$$g(t) \le \alpha(t) \mathrm{e}^{\beta t}, \quad \forall 0 \le t \le T.$$

Proof Then

$$\frac{d}{dt}\left(\mathrm{e}^{-\beta t} \int_0^t g(s)\, ds\right) = \mathrm{e}^{-\beta t} g(t) - \beta \mathrm{e}^{-\beta t} \int_0^t g(s)\, ds \le \alpha(t) \mathrm{e}^{-\beta t}$$

and hence

$$\mathrm{e}^{-\beta t} \int_0^t g(s)\, ds \le \int_0^t \alpha(s) \mathrm{e}^{-\beta s}\, ds,$$

which we use on the r.h.s. of the inequality in the assumption to conclude. □

Theorem 7.8 *Let Y_t be a d-dimensional Itô process as in* (7.3a), *of the type*

$$Y_t = Y_0 + \int_0^t b_s\, ds + \int_0^t \sigma_s\, dW_s,$$

where (W_s) is an $\mathbb{R}^r$-valued standard Brownian motion, and (b_s) and (σ_s) are adapted processes taking values respectively in $\mathbb{R}^d$ and $\mathbb{R}^{d\otimes r}$. Let $m \ge 1$ be an integer such that $\mathbb{E}|Y_0|^{2m} < \infty$ and $s \mapsto \mathbb{E}|b_s|^{2m} + \mathbb{E}\|\sigma_s\|^{2m}$ is bounded on $[0, T]$.

Then, there exists a constant $C > 0$ such that

$$\mathbb{E}|Y_t|^{2m} \le \mathrm{e}^{Ct}\mathbb{E}|Y_0|^{2m} + C\mathrm{e}^{Ct} \int_0^t \left(\mathbb{E}|b_s|^{2m} + \mathbb{E}\|\sigma_s\|^{2m}\right) ds, \quad \forall 0 \le t \le T. \quad (7.17)$$

Proof The function $|x|^{2m} = (\sum_{i=1}^d x_i^2)^m$ is in $C^\infty(\mathbb{R}^d)$ and has first and second order partial derivatives given by

$$\begin{aligned} \partial_i |x|^{2m} &= 2m x_i |x|^{2(m-1)}, \\ \partial_{ij}^2 |x|^{2m} &= 4m(m-1) x_i x_j |x|^{2(m-2)} + 2m|x|^{2(m-1)} \mathbb{1}_{\{i=j\}}. \end{aligned} \quad (7.18)$$

Note that $x \mapsto x^{2m}$ is C^2 but not C_b^2, and that $x \mapsto x^{2m}$ is of class $\mathscr{C}^2$ but not $\mathscr{C}_b^2$. We are going to apply Itô's formula. To avoid integrability issues in each of its terms, a classic localization procedure using stopping times allows to deal with bounded processes.

Define $\tau_K := \inf\{s > 0, |Y_s| \geq K\}$ for $K \geq 1$. This localizing sequence of stopping times is such that $|Y_{t\wedge\tau_K}| \leq K$ for all t and $\lim_{K\to\infty} Y_{t\wedge\tau_K} = Y_t$. Using the localization property (7.1) of stochastic integrals, Itô's formula and equalities (7.18) yield

$$\begin{aligned}|Y_{t\wedge\tau_K}|^{2m} = |Y_0|^{2m} &+ 2m\int_0^{t\wedge\tau_K} |Y_{s\wedge\tau_K}|^{2m-2} Y_{s\wedge\tau_K}\cdot b_s\,ds \\ &+ 2m\int_0^{t\wedge\tau_K} |Y_{s\wedge\tau_K}|^{2m-2} Y_{s\wedge\tau_K}\cdot \sigma_s\,dW_s \\ &+ m\int_0^{t\wedge\tau_K} |Y_{s\wedge\tau_K}|^{2m-2}\|\sigma_s\|^2\,ds \\ &+ 2m(m-1)\int_0^{t\wedge\tau_K} |Y_{s\wedge\tau_K}|^{2m-4}|\sigma_s^* Y_{s\wedge\tau_K}|^2\,ds\end{aligned}$$

(the last term vanishes for $m = 1$). The stochastic integral in the r.h.s. is a martingale since $|Y_{s\wedge\tau_K}| \leq K$ for all s, and since the stopping time $t \wedge \tau_K$ is bounded from above by T, optional sampling (Theorem 7.2) yields

$$\mathbb{E}\int_0^{t\wedge\tau_K} |Y_{s\wedge\tau_K}|^{2m-2} Y_{s\wedge\tau_K}\cdot\sigma_s\,dW_s = 0.$$

In addition, there exists a constant C_m such that, for all non-negative x and y, $x^{2m-2}y^2 \leq C_m(x^{2m} + y^{2m})$.[2] One deduces that, possibly changing C_m,

$$\begin{aligned}\mathbb{E}|Y_{t\wedge\tau_K}|^{2m} \leq \mathbb{E}|Y_0|^{2m} &+ C_m\int_0^t \mathbb{E}|b_s|^{2m}\,ds + C_m\int_0^t \mathbb{E}\|\sigma_s\|^{2m}\,ds \\ &+ C_m\int_0^t \mathbb{E}|Y_{s\wedge\tau_K}|^{2m}\,ds. \end{aligned} \tag{7.19}$$

The second version of Gronwall's inequality (Lemma 7.1) yields

$$\mathbb{E}|Y_{t\wedge\tau_K}|^{2m} \leq \mathrm{e}^{C_m t}\mathbb{E}|Y_0|^{2m} + C_m\mathrm{e}^{C_m t}\int_0^t \left(\mathbb{E}|b_s|^{2m} + \mathbb{E}\|\sigma_s\|^{2m}\right) ds. \tag{7.20}$$

It then remains to let K tend to infinity, observe that

$$\liminf_{K\to\infty} |Y_{t\wedge\tau_K}|^{2m} = \lim_{K\to\infty} |Y_{t\wedge\tau_K}|^{2m} = |Y_t|^{2m},$$

and apply Fatou's lemma to conclude. □

This result on Itô processes will be now used to prove a similar result for the solution $(X_t)_{t\geq 0}$ of a stochastic differential equation, which is our central interest.

[2] Actually $x^{2m-2}y^2 \leq \max(x^{2m}, y^{2m})$.

Theorem 7.9 *Under the assumptions of Theorem 7.6, let $m \geq 1$ be an integer such that $\mathbb{E}|X_0|^{2m} < \infty$. Then, there exists $C < \infty$ only depending on m, T, and the Lipschitz constants of b and σ, such that*

$$\mathbb{E}|X_t|^{2m} \leq \big(1 + \mathbb{E}|X_0|^{2m}\big)\mathrm{e}^{Ct}, \quad \forall 0 \leq t \leq T, \tag{7.21a}$$

$$\mathbb{E} \sup_{0 \leq t \leq T} |X_t|^{2m} \leq C\big(1 + \mathbb{E}|X_0|^{2m}\big)\mathrm{e}^{CT}, \tag{7.21b}$$

$$\mathbb{E}|X_t - X_s|^{2m} \leq C\big(1 + \mathbb{E}|X_0|^{2m}\big)(t-s)^m, \quad \forall 0 \leq s \leq t \leq T. \tag{7.21c}$$

Proof Theorem 7.8 cannot be directly applied, since so far we do not know anything about the moments of $b_s := b(X_s)$ and $\sigma_s := \sigma(X_s)$, and we instead use (7.19):

$$\mathbb{E}|X_{t\wedge\tau_K}|^{2m} \leq \mathbb{E}|X_0|^{2m} + C\int_0^t \mathbb{E}|b_s|^{2m}\,ds + C\int_0^t \mathbb{E}\|\sigma_s\|^{2m}\,ds + C\int_0^t \mathbb{E}|X_{s\wedge\tau_K}|^{2m}\,ds.$$

The functions b and σ are Lipschitz, thus there exists $L > 0$ such that

$$\big|b(x)\big| + \big\|\sigma(x)\big\| \leq L\big(1 + |x|\big), \quad \forall x \in \mathbb{R}^d, \tag{7.22}$$

from which

$$\big|b(x)\big|^{2m} + \big\|\sigma(x)\big\|^{2m} \leq C_m\big(1 + |x|^{2m}\big), \quad \forall x \in \mathbb{R}^d. \tag{7.23}$$

Consequently, there exists $C > 0$ depending on m and L only such that

$$\mathbb{E}|X_{t\wedge\tau_K}|^{2m} \leq \mathbb{E}|X_0|^{2m} + C\int_0^t \big(1 + \mathbb{E}|X_{s\wedge\tau_K}|^{2m}\big)\,ds.$$

Then (7.21a) follows by applying Gronwall's inequality to the function $g(t) := 1 + \mathbb{E}|X_{t\wedge\tau_K}|^{2m}$ and using Fatou's lemma to let K go to infinity. Moreover (7.21b) follows from (7.5a)–(7.5b) by noting that

$$\left(\sum_{i=1}^k |x_i|\right)^{2m} \leq k^{2m-1}\sum_{i=1}^k |x_i|^{2m}, \quad \forall x_i \in \mathbb{R}, \tag{7.24}$$

and by applying the Hölder inequality to the integral $\int_0^t b(X_s)\,ds$ and the Burkholder–Davis–Gundy inequality (Theorem 7.4) to the stochastic integral $\int_0^t \sigma(X_s)\,dW_s$.

Similar arguments allow to prove (7.21c). □

We now extend the inequalities of this last theorem to the Euler scheme, taking care to have constants C which are uniform w.r.t. the discretization parameter n.

In order to simplify computations, we introduce a continuous-time interpolation process: for all n, we define a continuous process (X_t^n) which takes the values $X_{pT/n}^n$ at times pT/n; specifically,

$$X_t^n = X_{pT/n}^n + b\big(X_{pT/n}^n\big)\left(t - \frac{pT}{n}\right) + \sigma\big(X_{pT/n}^n\big)(W_t - W_{pT/n}),$$
$$\frac{pT}{n} \le t < \frac{(p+1)T}{n}. \tag{7.25a}$$

Observe that

$$X_t^n = X_0 + \int_0^t b\big(X_{\eta_n(s)}^n\big)\,ds + \int_0^t \sigma\big(X_{\eta_n(s)}^n\big)\,dW_s \tag{7.25b}$$

for

$$\eta_n(s) := \frac{pT}{n}, \quad \frac{pT}{n} \le s < \frac{(p+1)T}{n}. \tag{7.25c}$$

Remark 7.4 The above interpolation process will allow us to apply Itô's formula to the error process $X_t - X_t^n$ which, by construction, is an Itô process. In order to apply Gronwall's inequality, we will make appear the local error term which involves the dynamics of the interpolation scheme between times s and $\eta_n(s)$, which governs the global convergence rate.

This technique allows us to present in an elegant and concise way the error propagation of the discretization error from step to step.

The computation can also be done using only the Euler scheme dynamics: it consists in establishing a precise induction formula relying on the approximation error at time $(p+1)T/n$ in terms of the approximation error at time pT/n and a local approximation error, and then using a discrete-time version of Gronwall's inequality.

We are now in a position to estimate the moments of the Euler scheme.

Theorem 7.10 *Under the assumptions of Theorem 7.9, there exists $C < \infty$ which is uniform w.r.t. all $n \le T$ and satisfies*

$$\mathbb{E}\big|X_t^n\big|^{2m} \le \big(1 + \mathbb{E}|X_0|^{2m}\big)\mathrm{e}^{Ct}, \quad 0 \le t \le T. \tag{7.26}$$

Proof Similarly to the beginning of the proof of Theorem 7.9 but setting $b_s := b(X_{\eta_n(s)})$ and $\sigma_s := \sigma(X_{\eta_n(s)})$, there exists $C > 0$ depending on m and L only (and not on n) such that

$$\mathbb{E}\big|X_{t\wedge\tau_K}^n\big|^{2m} \le \mathbb{E}|X_0|^{2m} + C\int_0^t \big(1 + \mathbb{E}\big|X_{\eta_n(s)\wedge\tau_K}^n\big|^{2m}\big)\,ds. \tag{7.27}$$

Since $X^n_{\eta_n(s)}$ is independent of $W_s - W_{\eta_n(s)}$, using (7.23),

$$\begin{aligned}\mathbb{E}\left|X^n_s - X^n_{\eta_n(s)}\right|^{2m} &\le C\mathbb{E}\left|b\left(X^n_{\eta_n(s)}\right)\left(s-\eta_n(s)\right) + \sigma\left(X^n_{\eta_n(s)}\right)(W_s - W_{\eta_n(s)})\right|^{2m} \\ &\le C\left(1+\mathbb{E}\left|X^n_{\eta_n(s)}\right|^{2m}\right)\left(\left(s-\eta_n(s)\right)^{2m} + \mathbb{E}(W_s - W_{\eta_n(s)})^{2m}\right) \\ &\le C\left(\frac{T}{n}\right)^m. \end{aligned} \tag{7.28}$$

Coming back to the inequality (7.27) we deduce

$$\mathbb{E}\left|X^n_{t\wedge\tau_K}\right|^{2m} \le \mathbb{E}|X_0|^{2m} + C\int_0^t \left(1+\mathbb{E}\left|X^n_{s\wedge\tau_K}\right|^{2m}\right) ds + Ct\left(\frac{T}{n}\right)^m. \tag{7.29}$$

Recall that $T/n \le 1$; applying Gronwall's inequality (Lemma 7.1) for $g(t) := 1 + \mathbb{E}|X^n_t|^{2m}$, letting K go to infinity, and finally using Fatou's lemma allow to conclude. □

Similar arguments to those which allowed us to deduce (7.21b) from (7.21a) easily allow to deduce from (7.26) the following result:

Corollary 7.1 *Under the hypotheses of Theorem 7.9, there exists a positive real number C which is uniform w.r.t. all $n \ge T$ such that*

$$\mathbb{E}\sup_{0\le t\le T}|X^n_t|^{2m} \le C\left(1+\mathbb{E}|X_0|^{2m}\right)\mathrm{e}^{CT}. \tag{7.30}$$

Exercise 7.2 Consider the one-dimensional stochastic differential equation

$$dX_t = \mu X_t\,dt + \sigma X_t\,dW_t \tag{7.31}$$

where μ and σ are non-negative real numbers and (W_t) is a standard Brownian motion. Let X^n be the Euler scheme with step-size T/n for this equation. Show that X^n_T can be written as the product of independent random variables. Deduce a bound from above of $\mathbb{E}|X^n_T|^{2m}$ uniform w.r.t. n. Solve the same questions for the Milstein scheme.

7.4 Convergence Rates in $L^p(\Omega)$ Norm and Almost Surely

The objective of this section is to estimate the error

$$\mathbb{E}\max_{0\le p\le n}\left|X_{pT/n} - X^n_{pT/n}\right|^{2m}$$

where X^n is the result of the Euler or of the Milstein scheme. We even obtain convergence rates for the interpolated processes which imply the same rates for the discrete time schemes.

We again emphasize that these rates do not optimally describe the actual accuracy of stochastic simulations. On the one hand, such simulations aim to provide information on the laws of processes. On the other hand, pathwise error estimates assume that the exact process is driven by the specific Brownian motion which is constructed along the simulations, which is unreasonable (see also our discussion at the beginning of Sect. 7.5.2).

The following theorem originally comes from Kanagawa [25] and the Ph.D. thesis of Faure [14].

Theorem 7.11 *Suppose that the functions b and σ are Lipschitz and there exists an integer $m \geq 1$ such that $\mathbb{E}|X_0|^{2m} < \infty$. Then there exists a function K such that, for any integer $n \geq T$, the Euler scheme satisfies*

$$\mathbb{E} \sup_{0 \leq t \leq T} \left| X_t - X_t^n \right|^{2m} \leq \frac{K(T)}{n^m}. \tag{7.32}$$

The function K only depends on the Lipschitz constants of the functions b and σ, on m, and on $\mathbb{E}|X_0|^{2m}$.

Proof As we observed in the proof of (7.21b) and (7.30), it suffices to prove

$$\sup_{0 \leq t \leq T} \mathbb{E}\left[\left| X_t - X_t^n \right|^{2m} \right] \leq \frac{K(T)}{n^m}. \tag{7.33}$$

To this end, let us set $\varepsilon_t^n := X_t - X_t^n$ and note that

$$\begin{aligned}
\varepsilon_t^n &= \int_0^t \left(b(X_s) - b\left(X_{\eta_n(s)}^n\right) \right) ds + \int_0^t \left(\sigma(X_s) - \sigma\left(X_{\eta_n(s)}^n\right) \right) dW_s \\
&= \int_0^t \left(b(X_s) - b\left(X_s^n\right) \right) ds + \int_0^t \left(\sigma(X_s) - \sigma\left(X_s^n\right) \right) dW_s \\
&\quad + \int_0^t \left(b\left(X_s^n\right) - b\left(X_{\eta_n(s)}^n\right) \right) ds + \int_0^t \left(\sigma\left(X_s^n\right) - \sigma\left(X_{\eta_n(s)}^n\right) \right) dW_s.
\end{aligned} \tag{7.34}$$

We now proceed as in the proof of Theorem 7.10, in particular concerning (7.28) and (7.29), and obtain:

$$\mathbb{E}\left|\varepsilon_t^n\right|^{2m} \leq C \int_0^t \mathbb{E}\left|\varepsilon_s^n\right|^{2m} ds + C \frac{T^m}{n^m}.$$

Gronwall's inequality (Lemma 7.1) allows us to conclude. □

Borel–Cantelli's lemma leads to the following result.

Theorem 7.12 *Suppose that the hypotheses of Theorem 7.11 are satisfied. For any $m > 1$ it holds that*

$$\forall 0 \le \alpha < \frac{1}{2} - \frac{1}{2m}, \quad n^\alpha \sup_{0\le t\le T} \left|X_t - X_t^n\right| \underset{n\to\infty}{\longrightarrow} 0, \quad a.s. \tag{7.35}$$

In particular, if X_0 has finite moments of all orders, then

$$\forall 0 \le \alpha < \frac{1}{2}, \quad n^\alpha \sup_{0\le t\le T} \left|X_t - X_t^n\right| \underset{n\to\infty}{\longrightarrow} 0, \quad a.s. \tag{7.36}$$

Proof Set $Z_n := n^\alpha \sup_{0\le t\le T} |X_t - X_t^n|$ and observe that (7.32) and the Markov inequality imply for any $\varepsilon > 0$ that

$$\sum_{n\ge T} \mathbb{P}(Z_n \ge \varepsilon) \le \frac{K(T)}{\varepsilon^{2m}} \sum_{n\ge T} \frac{1}{n^{m(1-2\alpha)}} < \infty.$$

Borel–Cantelli's lemma implies that $\mathbb{P}$ almost all ω belong to a finite number of events $\{Z_n \ge \varepsilon\}$. This implies that the sequence (Z_n) almost surely converges to 0, from which we deduce (7.35). The result (7.36) is then obvious. □

Remark 7.5 All the convergence results in this section concern the interpolation scheme. Therefore, they bring more information than necessary from a numerical point of view, since it is impossible to simulate continuous time trajectories on computers. The essential practical information concerns the errors

$$\max_{0\le p\le n} \left|X_{pT/n} - X^n_{pT/n}\right| \le \sup_{0\le t\le T} \left|X_t - X_t^n\right|.$$

The Milstein scheme leads to convergence rates in $L^p(\Omega)$ norm and in the a.s. sense which are better than those of the Euler scheme. We admit the next result, of which the proof mainly consists in improving the estimate (7.28); it originally was obtained in Milstein [36].

Theorem 7.13 *Suppose that the functions b and σ are twice continuously differentiable and that their partial derivatives up to order 2 are bounded. Let $m \ge 1$ be an integer such that $\mathbb{E}|X_0|^{4m} < \infty$. The Milstein scheme satisfies: There exists a non-decreasing function K such that*

$$\sup_{0\le t\le T} \mathbb{E}\left|X_t - X_t^n\right|^{2m} \le \frac{K(T)}{n^{2m}}. \tag{7.37}$$

Moreover,

$$\forall 0 \le \alpha < \frac{1}{2} - \frac{1}{m}, \quad n^\alpha \sup_{0\le t\le T} \mathbb{E}\left|X_t - X_t^n\right| \underset{n\to\infty}{\longrightarrow} 0, \quad a.s. \tag{7.38}$$

Exercise 7.3 Consider the situation and the hypotheses of Exercise 7.2. Show that the convergence rates in Theorems 7.11, 7.12, and 7.13 are optimal.

7.5 Monte Carlo Methods for Parabolic Partial Differential Equations

We are now in a position to develop a Monte Carlo method for the solutions of a class of parabolic partial differential equations.

7.5.1 The Principle of the Method

Let $a = (a^i_j)$ be the function on $\mathbb{R}^d$ taking values in the space of $d \times d$ matrices defined as

$$a := \sigma\sigma^*.$$

Consider the second order differential operator

$$\mathscr{L} := \sum_{i=1}^{d} b_i(\cdot)\partial_i + \frac{1}{2}\sum_{i,j=1}^{d} a^i_j(\cdot)\partial^2_{ij} \tag{7.39}$$

(the generator of the solution of the stochastic differential equation (7.5a)–(7.5b), see Sect. 7.1.3) and the parabolic partial differential equation with terminal condition

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + \mathscr{L}u(t,x) = 0, & (t,x) \in [0,T) \times \mathbb{R}^d, \\ u(T,x) = f(x), & x \in \mathbb{R}^d. \end{cases} \tag{7.40}$$

Under the hypotheses that we will formulate below on the functions b, σ and f, the problem (7.40) has a unique solution which is continuous on $[0,T] \times \mathbb{R}^d$ and belongs to the space $\mathscr{C}^{1,2}([0,T) \times \mathbb{R}^d)$. We will see that this unique solution satisfies

$$u(t,x) = \mathbb{E} f\big(X_{T-t}(x)\big), \quad \forall (t,x) \in [0,T] \times \mathbb{R}^d,$$

where $(X_\theta(x))$ is the solution to the (7.5a)–(7.5b) with initial condition $X_0 = x$, i.e.,

$$X_\theta(x) = x + \int_0^\theta b\big(X_s(x)\big)\,ds + \int_0^\theta \sigma\big(X_s(x)\big)\,dW_s, \quad 0 \le \theta \le T. \tag{7.41}$$

The preceding consideration suggests the following Monte Carlo method. If $\{X^{(i)}, i \in \mathbb{N}\}$ is a family of independent copies of the process X, the Strong Law of Large Numbers implies that

$$u(t,x) = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f\big(X^{(i)}_{T-t}(x)\big), \quad \text{a.s.}$$

As, in general, it is impossible to obtain the exact values of the $X^{(i)}_{pT/n}(x)$, we replace them with the approximations resulting from the Euler scheme.

Algorithm (Monte Carlo for terminal value PDE) To obtain an approximation for the solution $(u(t,x))$ for (7.40): simulate N independent copies $(X^{n(i)}, 1 \le i \le N)$ of the Markov chain X^n given by the Euler scheme, and take

$$u^{n,N}(t,x) := \frac{1}{N}\sum_{i=1}^{N} f\left(X_{T-t}^{n(i)}(x)\right). \tag{7.42}$$

Why is this Monte Carlo method interesting? Does it compete with deterministic numerical methods such as finite difference or finite elements methods?

The answers to these questions depend on the dimension d and the functions b and σ. As emphasized in Chap. 1, numerical experiments show that a Monte Carlo method cannot have better performances than an effective deterministic numerical method. Fortunately for the specialists in probabilistic numerical methods, situations do exist where deterministic methods are ineffective and probabilistic approaches are satisfying.

An example of such a case concerns high dimensional partial differential equations. The numerical cost of the deterministic algorithms exponentially increases with the dimension d of the state space as the number of points in the discretization grids which are involved in these algorithms. On the contrary, the representation (7.42) illustrates that the complexity of a probabilistic method often increases linearly only with d.

This being so, the method (7.42) may be interesting even in small dimensions when one desires to approximate the solution $u(t,x)$ in one single point or in a small number of points: this typically occurs in financial option prices (these prices are functions of the asset spot prices which underlie the optional contracts), and in the numerical resolution of certain inverse problems (when one desires to calibrate unknown parameters of the differential operator $\mathscr{L}$ by using measurements at points x_i which provide values $u(t,x_i)$).

Similarly, Monte Carlo methods may be effective to compute artificial Dirichlet conditions at a small number of points located at the boundary of a bounded domain where one limits the numerical resolution by a deterministic method.

There exist other situations, which we do not detail here, where Monte Carlo methods are effective: for example, when operators $\mathscr{L}$ are locally degenerate, deterministic methods need thin grids around the degeneracy manifolds in order to preserve stability conditions, which may dramatically increase the number of discretization points; for certain McKean–Vlasov non-linear partial differential equations, interacting stochastic particles methods, which generalize Monte Carlo methods and consist in approximating the solutions by particles empirical distributions, are effective because the particle dynamics make them naturally concentrate where the gradients of the exact solutions are large.

7.5.2 Introduction of the Error Analysis

Define $u^{n,N}(t,x)$ as in (7.42). Our objective is to estimate the approximation error

$$\left|u(T,x) - u^{n,N}(T,x)\right|$$

in terms of T and the numerical parameters N and n. This error is decomposed into

$$\left|u(T,x)-u^{n,N}(T,x)\right| \le \underbrace{\left|u(T,x)-\mathbb{E}f\left(X_T^n(x)\right)\right|}_{\alpha(n)} + \underbrace{\left|\mathbb{E}f\left(X_T^n(x)\right)-u^{n,N}(T,x)\right|}_{\beta(n,N)}. \tag{7.43}$$

The analysis of the *statistical error* $\beta(n,N)$ consists in proving that the constants in the Berry–Esseen and Bikelis theorems and concentration inequalities (see Chap. 3) are bounded from above uniformly w.r.t. n: this typically results from the estimates on the *discretization error* $\alpha(n)$ which we get below.

Let us start with a coarse grained analysis of $\alpha(n)$. When the function f is Lipschitz, the error estimate (7.32) with $2m=1$ leads to

$$\left|\alpha(n)\right| \le \frac{C}{\sqrt{n}}.$$

However this rate is a poor estimate, notably because $\mathbb{E}|X-Y|$ generically is a bad approximation of $|\mathbb{E}(X)-\mathbb{E}(Y)|$. Note also that the error analysis of $\mathbb{E}f(X_T(x))-\mathbb{E}f(X_T^n(x))$ should not suppose that $X_T(x)$ and $X_T^n(x)$ belong to the same probability space, which is implicitly required in (7.32).

Exercise 7.4 Consider the situation of Exercise 7.2. Let X^n be the Euler scheme with step-size T/n. Show that $\mathbb{E}(X_T)-\mathbb{E}(X_T^n)$ and $\mathbb{E}((X_T)^2)-\mathbb{E}((X_T^n)^2)$ are of order $1/n$. Show that the Milstein scheme satisfies the same property.

We now show that the convergence rate $1/n$ is very general. We will even get a much better result: the discretization error $\alpha(n)$ admits a Taylor expansion w.r.t. $1/n$. We will see that this theoretical result has important numerical consequences.

A reasonable methodology to precisely estimate the error $\alpha(n)$ seems to be to study the time evolution of $\mathbb{E}f(X_{pT/n})-\mathbb{E}f(X^n_{pT/n})$. Unfortunately this technique is not successful, so that we need to develop another approach which reflects that, at each time step, the transition operator of the Markov process X is approximated by the transition operator of the Markov chain X^n.

The proof deeply uses the following classical result yielding a probabilistic representation for the solution.

Theorem 7.14 (Feynman–Kac formula) *Suppose that the functions b and σ are Lipschitz and that the function f has an at most polynomial growth at infinity. Suppose also that the partial differential equation (7.40) admits a solution of class $\mathscr{C}^{1,2}([0,T]\times\mathbb{R}^d)$ whose partial derivatives of all orders have an at most polynomial growth at infinity. Then*

$$u(t,x)=\mathbb{E}f\left(X_T(t,x)\right), \quad \forall (t,x)\in[0,T]\times\mathbb{R}^d, \tag{7.44}$$

where

$$X_\theta(t,x)=x+\int_t^\theta b\left(X_s(t,x)\right)ds+\int_t^\theta \sigma\left(X_s(t,x)\right)dW_s, \quad 0\le t\le\theta\le T. \tag{7.45}$$

Similarly, given the solution $(X_\theta(x))$ *to the stochastic differential equation* (7.41),

$$u(t,x) = \mathbb{E} f\big(X_{T-t}(x)\big), \quad \forall (t,x) \in [0,T] \times \mathbb{R}^d, \tag{7.46}$$

where $(X_\theta(x))$ *is the solution of* (7.41).

Proof Apply Itô's formula (7.4) to $u(s, X_s(t,x))$ from time t to T. The function u solves the partial differential equation (7.40), hence the integral w.r.t. Lebesgue measure vanishes, and we obtain

$$u\big(T, X_T(t,x)\big) = u(t,x) + \int_t^T \nabla u\big(s, X_s(t,x)\big) \cdot \sigma\big(X_s(t,x)\big)\, dW_s.$$

In view of Theorem 7.9 and inequality (7.22) the preceding stochastic integral is a martingale, and taking expectations yields

$$u(t,x) = \mathbb{E} u\big(T, X_T(t,x)\big).$$

Equality (7.44) follows by noting that, by hypothesis, $u(T,y) = f(y)$.

Then, (7.46) follows from the fact that the laws of $(X_\theta(x), 0 \le \theta \le T-t)$ and $(X_\theta(t,x), t \le \theta \le T)$ are identical, since $(W_{t+s} - W_t, 0 \le s \le T-t)$ is a Brownian motion, and there is uniqueness in law of the solutions to the stochastic differential equations (7.41) and (7.45), see Theorem 7.6. □

Remark 7.6 Various sets of hypotheses on the functions b, σ and f imply the conditions imposed on $u(t,x)$ in Theorem 7.14. Classical results, obtained by PDE analysis techniques, require ellipticity or hypo-ellipticity conditions on the differential operator $\mathscr{L}$. By using probabilistic techniques only, we will see below that smoothness hypotheses on the functions b, σ and f also allow one to satisfy the desired conditions on $u(t,x)$ without any further condition on $\mathscr{L}$.

Before proving the optimal convergence rate for the Euler scheme, we present a short computation which introduces the main ingredient of the technical proof and shows that the discretization error is of order $1/n$. It relies on the following property: the two first moments of the increments of the Euler scheme satisfy

$$\mathbb{E}\big[X^n_{(p+1)T/n}(x) - X^n_{pT/n}(x)\big] = \mathbb{E} b\big(X^n_{pT/n}(x)\big)\frac{T}{n}, \tag{7.47a}$$

$$\mathbb{E}\big[\big(X^n_{(p+1)T/n}(x) - X^n_{pT/n}(x)\big) \cdot \big(X^n_{(p+1)T/n}(x) - X^n_{pT/n}(x)\big)^t\big]$$
$$= \mathbb{E} a\big(X^n_{pT/n}(x)\big)\frac{T}{n} + \mathscr{O}\left(\frac{1}{n^2}\right), \tag{7.47b}$$

and moreover it is simple to similarly check that the product of three coordinates with indices $1 \le \ell_1, \ell_2, \ell_3 \le d$ of an X^n increment

$$\big(X^n_{(p+1)T/n}(x) - X^n_{pT/n}(x)\big)_{\ell_1}\big(X^n_{(p+1)T/n}(x) - X^n_{pT/n}(x)\big)_{\ell_2} \times \big(X^n_{(p+1)T/n}(x) - X^n_{pT/n}(x)\big)_{\ell_3}$$

has an expectation of order $\frac{1}{n^2}$, and that all products of four coordinates or more have expectations with the same order or a better one.

We now observe that (7.46) and a telescopic sum yield

$$\begin{cases} \mathbb{E}f\big(X^n_T(x)\big) - \mathbb{E}f\big(X_T(x)\big) = \mathbb{E}u\big(T, X^n_T(x)\big) - u(0,x) = \displaystyle\sum_{p=0}^{n-1} \delta^n_p, \\ \delta^n_p := \mathbb{E}\big[u\big((p+1)T/n, X^n_{(p+1)T/n}(x)\big) - u\big(pT/n, X^n_{pT/n}(x)\big)\big]. \end{cases} \tag{7.48}$$

A Taylor expansion of u around $(pT/n, X^n_{pT/n}(x))$ and the equality $\frac{\partial u}{\partial t} + \mathscr{L}u = 0$ lead to

$$\begin{aligned} &\mathbb{E}f\big(X^n_T(x)\big) - \mathbb{E}f\big(X_T(x)\big) \\ &\quad = \sum_{p=0}^{n-1} \mathbb{E}\big[u\big((p+1)T/n, X^n_{pT/n}(x)\big) - u\big(pT/n, X^n_{pT/n}(x)\big)\big] \\ &\qquad + \frac{T}{n}\sum_{p=0}^{n-1} \mathbb{E}\big[\mathscr{L}u\big((p+1)T/n, X^n_{pT/n}(x)\big)\big] + \sum_{p=0}^{n-1} \mathscr{O}\left(\frac{1}{n^2}\right) \\ &\quad = \frac{T}{n}\sum_{p=0}^{n-1} \mathbb{E}\big[\mathscr{L}u\big(pT/n, X^n_{pT/n}(x)\big) + \frac{\partial u}{\partial t}\big(pT/n, X^n_{pT/n}(x)\big)\big] + \sum_{p=0}^{n-1} \mathscr{O}\left(\frac{1}{n^2}\right) \\ &\quad = \mathscr{O}\left(\frac{1}{n}\right). \end{aligned}$$

Remark 7.7 To render the preceding expansion rigorous, we need to prove that the function u is smooth enough, all the expectations in the r.h.s. terms are finite and bounded from above uniformly w.r.t. n, etc. This is done in the next section.

7.6 Optimal Convergence Rate: The Talay–Tubaro Expansion

The goal of this section is to prove the following theorem in Talay and Tubaro [46].

Theorem 7.15 *Suppose that*

(H) *The functions b and σ are of class $\mathscr{C}^\infty(\mathbb{R}^d)$, and their partial derivatives of all orders are bounded (b and σ are not supposed bounded themselves),*

(H') *The function f is of class $\mathscr{C}^\infty(\mathbb{R}^d)$ and, for any multi-index α, there exists an integer s and a positive real number K such that*

$$\forall x \in \mathbb{R}^d, \quad \left|\partial_\alpha^{|\alpha|} f(x)\right| \le K\left(1+|x|^s\right). \tag{7.49}$$

Then, the discretization error of the Euler scheme satisfies

$$\mathbb{E} f\left(X_T(x)\right) - \mathbb{E} f\left(X_T^n(x)\right) = -\frac{T}{n}\int_0^T \mathbb{E}\Psi\left(s, X_s(x)\right) ds + \frac{Q_T^n(x)}{n^2} \tag{7.50}$$

where

$$\Psi(t,\cdot) = \frac{1}{2}\sum_{i,j=1}^d b^i b^j \partial_{ij}^2 u(t,\cdot) + \frac{1}{2}\sum_{i,j,k=1}^d b^i a_k^j \partial_{ijk}^3 u(t,\cdot) + \frac{1}{8}\sum_{i,j,k,l=1}^d a_j^i a_l^k \partial_{ijkl}^4 u(t,\cdot)$$
$$+ \frac{1}{2}\frac{\partial^2}{\partial t^2} u(t,\cdot) + \sum_{i=1}^d b^i \frac{\partial}{\partial t}\partial_i u(t,\cdot) + \frac{1}{2}\sum_{i,j=1}^d a_j^i \frac{\partial}{\partial t}\partial_{ij}^2 u(t,\cdot). \tag{7.51}$$

In addition, there exists an increasing function K and an integer m such that

$$\left|Q_T^n(x)\right| \le K(T)\left(1+|x|^m\right). \tag{7.52}$$

Remark 7.8 No ellipticity condition is made on the differential operator $\mathscr{L}$, which therefore can be degenerate.

The proof of the preceding theorem is technical. To help the reader we now use the notation already used in Sects. 5.1.2 and 6.1.1.

Notation 7.16 *For any measurable function ϕ from $\mathbb{R}^{2d}$ to $\mathbb{R}$ with an at most polynomial growth at infinity, let*

$$\mathbb{E}_x \phi\left(X_t, X_t^n\right) := \mathbb{E}\phi\left(X_t(x), X_t^n(x)\right).$$

Note that Theorems 7.9 and 7.10 imply that these expectations are well defined.

Notation 7.17 *Denote by (P_θ) the transition operators of the Markov process (X_t): for any measurable function ϕ from $\mathbb{R}^d$ to $\mathbb{R}$ with an at most polynomial growth at infinity and $\theta \ge 0$,*

$$P_\theta \phi(x) := \mathbb{E}_x \phi(X_\theta), \quad x \in \mathbb{R}^d.$$

The proof of Theorem 7.15 uses the following lemma.

Lemma 7.2 *The discretization error satisfies*:

$$\mathbb{E}_x f\left(X_T^n\right) - \mathbb{E}_x f(X_T) = \frac{T^2}{n^2}\sum_{p=0}^{n-2} \mathbb{E}_x \Psi\left(pT/n, X_{pT/n}^n\right) + \sum_{p=0}^{n-1} R_p^n(x), \tag{7.53}$$

where

$$R_{n-1}^n(x) := \mathbb{E}_x f\big(X_T^n\big) - \mathbb{E}_x\big[(P_{T/n} f)\big(X_{T-T/n}^n\big)\big]$$

and $R_k^n(x)$ *for* $k < n-1$ *is a sum of terms of the form*

$$\mathbb{E}_x\Bigg[\varphi_\alpha^\natural\big(X_{pT/n}^n\big)\int_{pT/n}^{(p+1)T/n}\int_{pT/n}^{s_1}\int_{pT/n}^{s_2}\big(\varphi_\alpha^\sharp\big(X_{s_3}^n\big)\partial_\alpha^{|\alpha|}u\big(s_3, X_{s_3}^n\big) + \varphi_\alpha^\flat(X_{s_3})\partial_\alpha^{|\alpha|}u(s_3, X_{s_3})\big)\,ds_3\,ds_2\,ds_1\Bigg] \tag{7.54}$$

in which the multi-indices satisfy $|\alpha| \le 6$, *and the functions* $\varphi_\alpha^\natural$, $\varphi_\alpha^\sharp$, $\varphi_\alpha^\flat$ *are products of partial derivatives of order less than* 3 *of the functions* a^{ij} *and* b^i.

Proof For all $z \in \mathbb{R}^d$ define the differential operator $\mathscr{L}_z$ by

$$\mathscr{L}_z := \sum_{i=1}^d b^i(z)\partial_i + \frac{1}{2}\sum_{i,j=1}^d a_j^i(z)\partial_{ij}^2$$

which are a version of (7.39) in which the coefficients are "frozen".

Let us go back to (7.48). Itô's formula (7.4) and (7.40) yield

$$\begin{aligned}\delta_p^n &= \mathbb{E}_x\int_{pT/n}^{(p+1)T/n}\big(\partial_t u\big(t, X_t^n\big) + \mathscr{L}_z u\big(t, X_t^n\big)\big|_{z=X_{pT/n}^n}\big)\,dt\\ &= \mathbb{E}_x\int_{pT/n}^{(p+1)T/n}\big(-\mathscr{L}u\big(t, X_t^n\big) + \mathscr{L}_z u\big(t, X_t^n\big)\big|_{z=X_{pT/n}^n}\big)\,dt.\end{aligned}$$

Defining

$$\begin{aligned}I_p^n(t) &:= \mathscr{L}_z u\big(t, X_t^n\big)\big|_{z=X_{pT/n}^n} - \mathscr{L}_z u\big(pT/n, X_{pT/n}^n\big)\big|_{z=X_{pT/n}^n},\\ J_p^n(t) &:= \mathscr{L}_z u\big(pT/n, X_{pT/n}^n\big)\big|_{z=X_{pT/n}^n} - \mathscr{L}u\big(t, X_t^n\big)\\ &= \mathscr{L}u\big(pT/n, X_{pT/n}^n\big) - \mathscr{L}u\big(t, X_t^n\big),\end{aligned}$$

we then have

$$\delta_p^n = \mathbb{E}_x\int_{pT/n}^{(p+1)T/n}\big(I_p^n(t) + J_p^n(t)\big)\,dt.$$

Now consider $I_k^n(t)$ and $J_k^n(t)$ as smooth functions of X_t^n and apply Itô's formula; observe also that $\frac{\partial}{\partial t}\mathscr{L}u + \mathscr{L}(\mathscr{L}u) = 0$ since $\frac{\partial}{\partial t}u + \mathscr{L}u = 0$ by (7.40). It is then a simple matter to conclude. □

Proof of Theorem 7.15 The expansion (7.53) can be rewritten as

$$\begin{aligned}\mathbb{E}_x f\big(X_T^n\big) - \mathbb{E}_x f(X_T) &= \frac{T}{n}\int_0^T \mathbb{E}_x \Psi(s, X_s)\, ds \\ &\quad + \frac{T^2}{n^2}\sum_{p=0}^{n-2} \mathbb{E}_x \Psi(pT/n, X_{pT/n}) - \frac{T}{n}\int_0^T \mathbb{E}_x \Psi(s, X_s)\, ds \\ &\quad + \frac{T^2}{n^2}\sum_{p=0}^{n-2} \mathbb{E}_x\big(\Psi\big(pT/n, X_{pT/n}^n\big) - \Psi(pT/n, X_{pT/n})\big) \\ &\quad + \sum_{p=0}^{n-2} R_p^n(x) + \mathbb{E}_x f\big(X_T^n\big) - \mathbb{E}_x\big[(P_{T/n} f)\big(X_{T-T/n}^n\big)\big] \\ &= \frac{T}{n}\int_0^T \mathbb{E}_x \Psi(s, X_s)\, ds + A^n + B^n + \sum_{p=0}^{n-2} R_p^n(x) + C^n. \end{aligned} \tag{7.55}$$

Thus it is reasonable to expect that the error $\mathbb{E}_x f(X_T^n) - \mathbb{E}_x f(X_T)$ is *equal* to

$$\frac{T}{n}\int_0^T \mathbb{E}_x \Psi(s, X_s)\, ds$$

up to a remaining term of order $1/n^2$ since:

- for any function ϕ from $[0, T]$ to $\mathbb{R}$ of class $\mathscr{C}^1$, the approximation error

$$\int_0^T \phi(s)\, ds - \frac{T}{n}\sum_{p=0}^{n-1} \phi(pT/n)$$

  is of order $1/n$ (this convergence rate of Riemann sums results from a first order Taylor expansion),
- the remaining term

$$\sum_{p=0}^{n-2} \mathbb{E}_x\big(\Psi\big(pT/n, X_{pT/n}^n\big) - \Psi(pT/n, X_{pT/n})\big)$$

  should be uniformly bounded w.r.t. n, each term seeming to be of order $1/n$.

We now rigorously prove this last assertion. For all $1 \le p \le n-2$, use the expansion (7.55) and substitute the function

$$f_{n,p}(\cdot) := \Psi(pT/n, \cdot)$$

to f. Set $u_{n,p}(t,x) := P_{pT/n-t} f_{n,p}(x)$, and let $\Psi_{n,p}(t,\cdot)$ be the function obtained by replacing u by $u_{n,p}$ and T by pT/n in (7.51). There exist functions $g_\lambda \in \mathscr{C}_b^\infty(\mathbb{R}^d)$ such that, for all $t \le pT/n$,

$$\Psi_{n,p}(t,\cdot) = \sum_\lambda g_\lambda \partial_\lambda \big[P_{pT/n-t}\Psi(pT/n,\cdot)\big].$$

We thus have obtained

$$\begin{aligned}&\mathbb{E}_x \Psi\big(pT/n, X^n_{pT/n}\big) - \mathbb{E}_x \Psi(pT/n, X_{pT/n}) \\ &\quad = \frac{T^2}{n^2} \sum_{j=0}^{p-2} \mathbb{E}_x \Psi_{n,p}\left(\frac{jT}{n}, X^n_{jT/n}\right) + \sum_{j=0}^{p-1} R_j^{n,p}(x) \end{aligned} \tag{7.56}$$

where the $R_j^{n,p}(x)$ are sums of terms of the type (7.54) with $u_{n,p}$ in place of u.

The key issue therefore is as follows. Let γ and λ be arbitrary multi-indices, and g_γ and g_λ be smooth functions whose partial derivatives of all orders have an at most polynomial growth at infinity. Set

$$\varphi(\theta,\cdot) := g_\gamma(\cdot)\partial_\gamma P_{T-\theta} f(\cdot).$$

We have to prove that any quantity of the type

$$\big|\mathbb{E}_x\big[g_\lambda\big(X_t^n\big)\partial_\lambda P_{\theta-t}\varphi(\theta,\cdot)(z)\big|_{z=X_t^n}\big]\big| \tag{7.57}$$

is bounded uniformly w.r.t. $n \ge 1$ and $\theta \in [0, T - \frac{T}{n}]$ and $t \in [0, \theta - \frac{T}{n}]$.

In Sect. 7.8 we will use a probabilistic technique to show that, for any multi-index α, there exists an increasing function K_α and an integer m_α such that

$$\forall t \in [0,T],\ \forall x \in \mathbb{R}^d, \quad \big|\partial_\alpha^{|\alpha|} u(t,x)\big| \le K_\alpha(T)\big(1+|x|^{m_\alpha}\big). \tag{7.58}$$

Now let $\mathscr{H}_T$ denote the class of the functions $\phi : [0,T]\times\mathbb{R}^d \to \mathbb{R}$ having the following properties: ϕ is of class $\mathscr{C}^\infty([0,t]\times\mathbb{R}^d)$ and, for any multi-index α, there exists an integer s and an increasing function K such that

$$\forall \theta \in [0,T],\ \forall x \in \mathbb{R}^d, \quad \big|\partial_\alpha^{|\alpha|}\phi(\theta,x)\big| \le K(T)\big(1+|x|^s\big). \tag{7.59}$$

Let ϕ be a function in $\mathscr{H}_T$ and θ be a real number in $[0,T]$. The function $u(\theta;t,x)$ defined as

$$u(\theta;t,x) := \mathbb{E}\phi\big(\theta, X_{T-t}(x)\big) = \mathbb{E}_x\phi(\theta, X_{T-t})$$

belongs to $\mathscr{H}_T$ and satisfies

$$\begin{cases} \dfrac{\partial u}{\partial t}(\theta;t,x) + \mathscr{L}u(\theta;t,x) = 0, \quad 0 \le t < T, \\ u(\theta;T,x) = \phi(\theta,x). \end{cases} \tag{7.60}$$

One can easily obtain inequalities similar to (7.58): for any multi-index α, there exist an increasing function K_α and an integer m_α such that

$$\forall \theta \in [0,T], \quad \left|\partial_\alpha^{|\alpha|} u(\theta;t,x)\right| \le K_\alpha(T)\left(1+|x|^{k_\alpha}\right).$$

This result allows us to claim: for any arbitrary multi-indices γ and λ, all continuous functions f and g_γ with an at most polynomial growth at infinity, there exist an increasing function K and an integer m such that

$$\left|\mathbb{E}_x\left[g_\lambda\left(X_t^n\right)\partial_\lambda P_{\theta-t}\varphi(\theta,\cdot)(z)\big|_{z=X_t^n}\right]\right| \le K(T)\left(1+|x|^m\right), \tag{7.61}$$

where

$$\varphi(\theta,\cdot) := g_\gamma\, \partial_\gamma P_{T-\theta} f.$$

Applying this estimate to (7.55) and (7.56) yields the desired result. □

Remark 7.9 The above convergence rate is preserved when using suitable discrete random variables instead of the Brownian increments. Consider the scheme

$$\begin{cases} X_0^n = X_0, \\ X_{(p+1)T/n}^n = X_{pT/n}^n + b\left(X_{pT/n}^n\right)\dfrac{T}{n} + \sigma\left(X_{pT/n}^n\right)\sqrt{\dfrac{T}{n}}\,\xi_{p+1}, \end{cases}$$

where the random vectors ξ_{p+1} are independent and identically distributed, have a finite moment of 6th order, and have the same first five moments as the increments of W. One can easily check that all the above calculations still hold true. The only change concerns the remainder term $Q_T^n(x)$, but the bound (7.52) remains true.

Remark 7.10 The convergence rates of the Euler scheme (7.11) and the Milstein scheme (7.13) for $\mathbb{E} f(X_T)$ are the same. The reason is that the Milstein scheme satisfies (7.47a)–(7.47b) but does not improve the order of the remaining terms. However the functions Ψ in the leading terms of the Euler and Milstein scheme error expansions are different.

7.7 Romberg–Richardson Extrapolation Methods

The expansions (7.50) allow to justify the Romberg–Richardson extrapolation procedure. Note that

$$\mathbb{E}_x f(X_T) - \mathbb{E}_x f\left(X_T^n\right) = \frac{T}{n} e(T) + \mathcal{O}\left(\frac{1}{n^2}\right),$$

$$\mathbb{E}_x f(X_T) - \mathbb{E}_x f\left(X_T^{2n}\right) = \frac{T}{2n} e(T) + \mathcal{O}\left(\frac{1}{n^2}\right),$$

where

$$e(T) := -\frac{T}{n}\int_0^T \mathbb{E}_x \Psi(s, X_s)\, ds,$$

so that the new approximation

$$Z_T^n := 2\mathbb{E}_x f\left(X_T^{2n}\right) - \mathbb{E}_x f\left(X_T^n\right) \tag{7.62}$$

satisfies

$$\mathbb{E}_x f(X_T) - Z_T^n = \mathscr{O}\left(\frac{1}{n^2}\right).$$

Consequently, a convergence rate of order $1/n^2$ is achieved by interpolating linearly the values produced by the Euler scheme with two step-sizes T/n and $T/2n$.

This impressive gain of accuracy is thus obtained with a weak augmentation of the numerical complexity.

7.8 Probabilistic Interpretation and Estimates for Parabolic Partial Differential Equations

The aim of this section is to show the following theorem which has been used in the proof of Theorem 7.15: see (7.58).

Theorem 7.18 *Under the hypotheses* (H) *and* (H'), *for all* $0 \le t \le T$ *and* x *in* $\mathbb{R}^d$,

$$\partial_i u(t,x) = \partial_i \mathbb{E} f\left(X_{T-t}(x)\right) = \mathbb{E}\sum_{j=1}^d \partial_j f\left(X_{T-t}(x)\right)\partial_i X_{T-t}^j(x), \tag{7.63}$$

where the vector-valued process $(\partial_i X_\theta^j(x),\, j = 1, \ldots, d)$ *is the unique solution in* $\mathscr{M}_{\mathscr{F}}^2(0,T)$ *to the system*

$$\begin{aligned}\partial_i X_\theta(x) = e_i &+ \int_0^\theta \partial b\left(X_s(x)\right)\partial_i X_s(x)\, ds \\ &+ \sum_{j=1}^r \int_0^\theta \partial\sigma_j\left(X_s(x)\right)\partial_i X_s(x)\, dW_s^j, \quad 0 \le \theta \le T.\end{aligned} \tag{7.64}$$

Here, e_i *is the vector* $(0, \ldots, 0, 1, 0, \ldots, 0)$ *with* 1 *in position* i, ∂b *is the matrix* $(\frac{\partial b^i}{\partial x_j}, i = 1, \ldots, d;\, j = 1, \ldots, d)$, *and* $\partial\sigma_j$ *is the matrix* $(\frac{\partial \sigma_j^\ell}{\partial x_m}, \ell = 1, \ldots, d; m = 1, \ldots, d)$.

Proof We start by proving that the system (7.64) has a unique strong solution for all x. Under our hypotheses, Theorem 7.6 applies and (7.5a)–(7.5b) has a unique

strong solution. One then constructs the following Picard sequence of successive approximations:

$$Y_\theta^{(0)} = e_i$$

and

$$Y_\theta^{(k+1)} = e_i + \int_0^\theta \partial b(X_s) Y_s^{(k)}\, ds + \sum_{j=1}^r \int_0^\theta \partial \sigma_j(X_s) Y_s^{(k)}\, dW_s^j.$$

As all the matrices ∂b and $\partial \sigma_j$ have bounded coefficients, it is easy to obtain

$$\mathbb{E}\Big(\max_{0\le s\le t} \big|Y_s^{(k+1)} - Y_s^{(k)}\big|^2\Big) \le C \int_0^t \mathbb{E}\big(\big|Y_s^{(k)} - Y_s^{(k-1)}\big|^2\big)\, ds.$$

One can then proceed as in the proof of Theorem 7.6 (see the proof of Theorem 5.2.9 in Karatzas and Shreve [26]): one gets that the sequence $(Y_t^{(k)})$ a.s. converges to a solution to (7.64) and that such a solution is unique.

We now prove the stochastic representation (7.63). To simplify notation, we start by examining the case $d = r = 1$ and write X_t^x instead of $X_t(x)$:

$$X_t^x = x + \int_0^t b\big(X_s^x\big)\, ds + \int_0^t \sigma\big(X_s^x\big)\, dW_s,$$

where (W_t) is a one-dimensional standard Brownian motion. We aim to prove

$$\frac{\partial}{\partial x} u(t,x) = \mathbb{E}\big[\phi'\big(X_t^x\big) Y_t^x\big], \tag{7.65}$$

where (Y_t^x) solves the stochastic differential equation

$$Y_t^x = 1 + \int_0^t b'\big(X_s^x\big) Y_s^x\, ds + \int_0^t \sigma'\big(X_s^x\big) Y_s^x\, dW_s, \tag{7.66}$$

which means that Y_t^x stands for $\partial_x X_t^x$.

Given $\varepsilon > 0$ define

$$Z_t^{x,\varepsilon} := \frac{1}{\varepsilon}\big(X_t^{x+\varepsilon} - X_t^x\big).$$

We are going to prove that $(Z_t^{x,\varepsilon})$ solves an equation which is similar to (7.66). Observe that, for any function f of class $C^1(\mathbb{R})$,

$$f(y) - f(x) = \int_0^1 f'\big(x + \alpha(y-x)\big)(y-x)\, d\alpha$$

for all x and y in $\mathbb{R}$: for this, integrate the function $\alpha \mapsto f'(x + \alpha(y-x))(y-x)$. By explicitly expressing $X_t^{x+\varepsilon}$ and X_t^x and setting

$$\phi_s^\varepsilon := \int_0^1 b'\big(X_s^x + \alpha\big(X_s^{x+\varepsilon} - X_s^x\big)\big)\, d\alpha,$$

$$\psi_s^\varepsilon := \int_0^1 \sigma'\big(X_s^x + \alpha\big(X_s^{x+\varepsilon} - X_s^x\big)\big)\,d\alpha,$$

we obtain

$$Z_t^{x,\varepsilon} = 1 + \int_0^t \phi_s^\varepsilon Z_s^{x,\varepsilon}\,ds + \int_0^t \psi_s^\varepsilon Z_s^{x,\varepsilon}\,dW_s. \tag{7.67}$$

Note that $\int_0^T (\psi_s^\varepsilon Z_s^{x,\varepsilon})^2\,ds < \infty$ so that the stochastic integral is well defined.

Itô's formula leads to

$$\mathbb{E}\big(Z_t^{x,\varepsilon}\big)^2 = 1 + \mathbb{E}\int_0^t \big(2\phi_s^\varepsilon + \big|\psi_s^\varepsilon\big|^2\big)\big(Z_s^{x,\varepsilon}\big)^2\,ds.$$

Observe that the processes (ϕ_t^ε) and (ψ_t^ε) are uniformly bounded w.r.t. ω, t, x and ε. Therefore there exists a positive constant C such that

$$\mathbb{E}\big(Z_t^{x,\varepsilon}\big)^2 \le 1 + C\int_0^t \mathbb{E}\big(Z_s^{x,\varepsilon}\big)^2\,ds.$$

Gronwall's lemma 7.1 then implies

$$\mathbb{E}\big(Z_t^{x,\varepsilon}\big)^2 \le \exp(Ct),$$

from which, using the definition of $Z_t^{x,\varepsilon}$, there exists $C < \infty$ such that

$$\mathbb{E}\big(X_t^{x+\varepsilon} - X_t^x\big)^2 \le C\varepsilon^2, \quad \forall t \in [0,T]. \tag{7.68}$$

Note that

$$Y_t^x - Z_t^{x,\varepsilon} = \int_0^t \big[b'\big(X_s^x\big)Y_s^x - \phi_s^\varepsilon Z_s^{x,\varepsilon}\big]\,ds + \int_0^t \big[\sigma'\big(X_s^x\big)Y_s^x - \psi_s^\varepsilon Z_s^{x,\varepsilon}\big]\,dW_s,$$

from which

$$\begin{aligned}\mathbb{E}\big(Y_t^x - Z_t^{x,\varepsilon}\big)^2 \le{}& 2t\mathbb{E}\int_0^t \big[b'\big(X_s^x\big)Y_s^x - \phi_s^\varepsilon Z_s^{x,\varepsilon}\big]^2\,ds \\ &+ 2\mathbb{E}\int_0^t \big[\sigma'\big(X_s^x\big)Y_s^x - \psi_s^\varepsilon Z_s^{x,\varepsilon}\big]^2\,ds.\end{aligned}$$

For each one of the integrals in the r.h.s., we use the inequality

$$\big(ab - a'b'\big)^2 \le 2a^2\big(b - b'\big)^2 + 2b'^2\big(a - a'\big)^2$$

which results from $ab - a'b' = a(b - b') + b'(a - a')$. We thus obtain that there exists $C < \infty$ such that

$$\mathbb{E}\big(Y_t^x - Z_t^{x,\varepsilon}\big)^2 \le C\mathbb{E}\int_0^t \big(b'\big(X_s^x\big) - \phi_s^\varepsilon\big)^2\big(Y_s^x\big)^2\,ds + C\mathbb{E}\int_0^t \big(\sigma'\big(X_s^x\big) - \psi_s^\varepsilon\big)^2\big(Y_s^x\big)^2\,ds$$
$$+ C\mathbb{E}\int_0^T \big(Y_s^x - Z_s^{x,\varepsilon}\big)^2\,ds, \quad \forall t \in [0,T].$$

We then again apply Gronwall's lemma 7.1: it shows that, if the two first terms in the right-hand side of the preceding equality tend to 0, then

$$\lim_{\varepsilon\to 0} \mathbb{E}\big(Y_t^x - Z_t^{x,\varepsilon}\big)^2 = 0. \tag{7.69}$$

To study the convergence of the two first terms under consideration, observe that $X_t^{x+\varepsilon}$ converges in probability to X_t^x (this is a consequence of (7.68)), and thus

$$\big(b'\big(X_s^x\big) - \phi_s^\varepsilon\big)^2 + \big(\sigma'\big(X_s^x\big) - \psi_s^\varepsilon\big)^2$$

converges in probability to 0. Thus (7.69) results from Lebesgue's Dominated Convergence theorem.

Now, for any bounded function ϕ of class $C^1(\mathbb{R})$ with bounded derivative

$$\begin{aligned}\frac{1}{\varepsilon}\big(\phi\big(X_t^{x+\varepsilon}\big) - \phi\big(X_t^x\big)\big) &= Z_t^{x,\varepsilon}\int_0^1 \phi'\big(X_t^x + \alpha\varepsilon Z_t^{x,\varepsilon}\big)\,d\alpha \\ &= \big(Z_t^{x,\varepsilon} - Y_t^x\big)\int_0^1 \phi'\big(X_t^x + \alpha\varepsilon Z_t^{x,\varepsilon}\big)\,d\alpha \\ &\quad + Y_t^x\int_0^1 \big(\phi'\big(X_t^x + \alpha\varepsilon Z_t^{x,\varepsilon}\big) - \phi'\big(X_t^x\big)\big)d\alpha + \phi'\big(X_t^x\big)Y_t^x \\ &=: A + B + \phi'\big(X_t^x\big)Y_t^x.\end{aligned}$$

In view of (7.69), the Cauchy–Schwarz inequality implies that $\mathbb{E}|A|$ tends to 0 with ε. In addition, again using (7.69), one can show that $\varepsilon Z_t^{x,\varepsilon}$ converges in probability to 0; the Dominated Convergence theorem then implies that $\mathbb{E}|B|$ tends to 0. Consequently,

$$\mathbb{E}\frac{1}{\varepsilon}\big(\phi\big(X_t^{x+\varepsilon}\big) - \phi\big(X_t^x\big)\big)$$

converges when ε tends to 0, and the limit is $\mathbb{E}[\phi'(X_t^x)Y_t^x]$.

In arbitrary dimension, under the hypotheses (H) and (H'), the same technique as above allows to obtain (7.63). In addition, from the hypotheses (H) and (H'), (7.63) and Theorem 7.9, there exist an increasing function K and an integer m such that

$$\big|\partial_t u(t,x)\big| \le K(T)\big(1 + |x|^m\big).$$

□

Remark 7.11 From (7.69) it is easy to prove

$$\forall T > 0, \quad \lim_{\varepsilon \to 0} \mathbb{E} \sup_{0 \le t \le T} \left(Y_t^x - Z_t^{x,\varepsilon} \right)^2 = 0. \tag{7.70}$$

Exercise 7.5 In the case $d = r = 1$, set

$$U_t^x := \int_0^t \left(b'\left(X_s^x\right) - \frac{1}{2}\sigma'\left(X_s^x\right)^2 \right) ds + \int_0^t \sigma'\left(X_s^x\right) dW_s,$$

and apply Itô's formula to $\exp(U_t^x)$. Deduce that (7.66) has solution

$$Y_t^x := \exp\left(U_t^x\right) = \exp\left[\int_0^t \left(b'\left(X_s^x\right) - \frac{1}{2}\sigma'\left(X_s^x\right)^2 \right) ds + \int_0^t \sigma'\left(X_s^x\right) dW_s \right].$$

Observe that (7.64) can be intuited by formally differentiating all the terms in (7.41) and permuting differentiation and integrals. One can iterate the procedure to intuit the systems solved by differentials of order larger than 2, which leads to the probabilistic interpretation of $\partial_\alpha^{|\alpha|} u(t,x)$ for any multi-index α. In addition, this allows one to show by induction that, for any multi-index α,

$$\forall t \le T, \ \forall x \in \mathbb{R}^d, \quad \left| \partial_\alpha^{|\alpha|} u(t,x) \right| \le K_\alpha(T)\left(1 + |x|^{m_\alpha}\right) \tag{7.71}$$

for some increasing function K_α and integer m_α.

The preceding construction of $\partial_i X_t(x)$ and its extension to $\partial_{ij} X_t(x)$ allow one to prove that the function $\mathbb{E} f(X_{T-t}^x)$ is a smooth solution of the partial differential equation (7.40), which closes the statement of Theorem 7.14. The proof is not so easy. A key argument in the proof is the stochastic flow property of the map $x \mapsto X_t(x)$ when the coefficients b and σ_j are Lipschitz, that is, there exists a solution such that, up to an event with null probability which does not depend on x in $\mathbb{R}^d$ and $0 \le s \le t \le T$, this map is continuous at every point x and satisfies the stochastic flow property

$$\mathbb{P}\text{-a.s.}, \quad \forall x \in \mathbb{R}^d, \ \forall 0 \le s \le t \le T, \quad X_{t-s} \circ X_s(x) = X_t(x).$$

If, in addition, the coefficients b and σ_j are differentiable with bounded derivatives then, up to an event with null probability which does not depend on x in $\mathbb{R}^d$ and t in $[0, T]$, the map $x \mapsto X_t(x)$ is differentiable at all point x and its almost sure partial derivatives solve the infinite set of (7.64) parametered by the initial condition x. Of course, these derivatives may (and should) be chosen instead of any other family of solutions $\partial_i X_t(x)$ (defined for each fixed x on an event of full probability which depends on x) which gives sense to (7.63). However, we emphasize the substantial difference between the study of stochastic flows and the way we followed to get (7.63): this latter approach, which is much more limited but suffices to prove the differentiability w.r.t. x of the function $\mathbb{E} f(X_{T-t}(x))$ and obtain its stochastic representation, avoids the delicate question of proving that the stochastic flow $X_t(x)$

is almost surely differentiable w.r.t. x at all points x and at any time t, the critical null event being allowed to be chosen independent of t and x.

We refer to Kunita [30] for the demanding proofs of the almost sure flow property and differentiability of the map $x \mapsto X_t(x)$, and the derivation of a stochastic proof of the existence of a smooth solution to the parabolic problem (7.40).

7.9 Problems

7.1 (Long Time Behavior of the Euler Scheme Error) Let μ and σ be two strictly positive real numbers, and

$$S_t = 1 + \mu \int_0^t S_\theta \, d\theta + \sigma \int_0^t S_\theta \, dW_\theta, \quad t \geq 0. \tag{7.72}$$

1. Write explicitly $\mathbb{E}(S_t)^2$.
2. (a) Discretize the (7.72) by the Euler scheme with step-size T/n. Let $(\bar{S}^n_{pT/n})$ denote the corresponding Markov chain.
   (b) Represent $\bar{S}^n_T$ as a product of terms.
   (c) Compute the exact value of $\mathbb{E}(\bar{S}^n_T)^2$.
   (d) Compute a function $C_1(T)$ such that

$$\mathbb{E}(S_T)^2 - \mathbb{E}\big(\bar{S}^n_T\big)^2 = \frac{C_1(T)}{n} + \mathscr{O}\left(\frac{1}{n^2}\right).$$

   (e) Prove that $|C_1(T)|$ tends to infinity with T.
3. Now consider

$$r_t = -\mu \int_0^t r_\theta \, d\theta + \sigma W_t, \quad t \geq 0.$$

   (a) Compute the exact value of the second moments of $r_{pT/n}$ and of the Euler scheme $\bar{r}^n_{pT/n}$.
   (b) Deduce the value of the approximation error in terms of n and p.
   (c) Prove that, for all n large enough, this error satisfies

$$\exists C > 0, \quad \left|\mathbb{E}(r_{pT/n})^2 - \mathbb{E}\big(\bar{r}^n_{pT/n}\big)^2\right| \leq \frac{C}{n}, \quad \forall p.$$

   (d) Compare with the situation in Question 2.

7.2 (Implicit Euler Scheme) Consider the stochastic differential equation

$$Z_t = -\int_0^t Z_s^3 \, ds + W_t, \quad t \geq 0.$$

Although the drift coefficient is not Lipschitz, one can prove the existence and uniqueness of the solution to this equation. **Admit this result.**

1. (a) Let τ_N be the first exit time of (Z_t) from the interval $[-N, N]$, for $N \geq 1$. Apply Itô's formula to $(Z_{t\wedge\tau_N})^2$.
   (b) Deduce that
$$\mathbb{E}|Z_t|^2 + 2\mathbb{E}\int_0^t |Z_s|^4\, ds \leq t, \quad \forall t.$$
   (c) Come back to the equality obtained from Itô's formula and conclude that
$$\mathbb{E}|Z_t|^2 + 2\int_0^t \mathbb{E}|Z_s|^4\, ds = t, \quad \forall t.$$

2. (a) Prove that
$$z^4 \geq z^2 - \frac{1}{4}, \quad \forall z \in \mathbb{R}.$$
   (b) Deduce a bound from above for $\frac{d}{dt}\mathbb{E}|Z_t|^2$.
   (c) Finally, differentiating $\exp(t)\mathbb{E}|Z_t|^2$, prove that
$$\mathbb{E}|Z_t|^2 \leq \frac{3}{2}, \quad \forall t \geq 0.$$

3. In the following $h > 0$, and for instance can be of the form $h = T/n$.
   (a) Write the Euler scheme $(\bar{Z}^h_{(p+1)h})$.
   (b) Prove that
$$\mathbb{E}\big|\bar{Z}^h_{(p+1)h}\big|^2 = \mathbb{E}\big|\bar{Z}^h_{ph}\big|^2 + h\mathbb{E}\big[\big|\bar{Z}^h_{ph}\big|^6 h - 2\big|\bar{Z}^h_{ph}\big|^4 + 1\big].$$

4. (a) Prove that there exists $\varepsilon > 0$ such that if h large enough then $hz^3 - 2z^2 + 1 > \varepsilon > 0$ for all $z > 0$.
   (b) Deduce that, in this case, there exists no uniform bound w.r.t. p for the error
$$\mathbb{E}\big|Z_{ph}\big|^2 - \mathbb{E}\big|\bar{Z}^h_{ph}\big|^2.$$

5. Consider the implicit Euler scheme
$$\tilde{Z}^h_{(p+1)h} = \tilde{Z}^h_{ph} - \big(\tilde{Z}^h_{(p+1)h}\big)^3 h + W_{(p+1)h} - W_{ph}.$$
   (a) Prove by induction on p that this sequence of random variables is well defined, and that all the moments of $\tilde{Z}^h_{ph}$ exist for all p and h.
   (b) Let $x_p := \mathbb{E}|\tilde{Z}^h_{ph}|^2$. Show that
$$x_{p+1} + 2hx^2_{p+1} \leq x_p + h.$$

(c) Let q be smallest integer such that $x_q > \sqrt{2}/2$. Prove that

$$x_{q+1} \le \frac{\sqrt{2}}{2} + h,$$

and deduce that

$$\mathbb{E}\big|\tilde{Z}^h_{ph}\big|^2 \le \frac{\sqrt{2}}{2} + h, \quad \forall p \in \mathbb{N},\ \forall h \in \mathbb{R}_+.$$

(d) To what property of the implicit Euler scheme can we conclude?

7.3 (Monotonicity and Convexity of $x \mapsto X_t(x)$ ($\star$)) Let be given a filtered probability space $(\Omega, \mathscr{F}, \mathbb{P}, (\mathscr{F}_t))$ and a standard Brownian motion (W_t) on this space. Let b and σ be bounded functions of class $\mathscr{C}^1(\mathbb{R})$ with bounded derivatives. Fix $T > 0$. Let (X^x_t) be the one-dimensional process solution to

$$X_t = x + \int_0^t b(X_\theta)\,d\theta + \int_0^t \sigma(X_\theta)\,dW_\theta, \quad 0 \le t \le T. \tag{7.73}$$

1. Justify that, for all x and t, the derivative w.r.t. x of the map $x \to X_t(x)$ almost surely exists and satisfies the (7.66).

   Hint: Approximate the left-hand side by $\frac{1}{\varepsilon}(X_t(x+\varepsilon) - X_t(x))$.

   Denote this derivative by $\frac{d}{dx}X_t(x)$ *and admit that there exists a process* $(X_t(x))$ *solution to* (7.73) *such that, except on a null event which does not depend on* t *in* $[0, T]$ *and* x *in* $\mathbb{R}$, *the map* $x \to X_t(x)$ *is differentiable and its derivative is equal to* $\frac{d}{dx}X_t(x)$. Using Exercise 7.5, show that, for all $0 \le t \le T$,

$$\frac{d}{dx}X_t(x) = Y^x_t = \exp\left(\int_0^t \left(b'\big(X_\theta(x)\big) - \frac{1}{2}\big(\sigma'\big(X_\theta(x)\big)\big)^2\right)d\theta + \int_0^t \sigma'\big(X_\theta(x)\big)\,dW_\theta\right).$$

   Deduce that the map $x \mapsto X_t(x)$ is almost surely a one-to-one map of x in $\mathbb{R}$.
2. Is the flow $\bar{X}^n_{kT/n}(x)$ of the Euler scheme almost surely a one-to-one map of x?
3. From now on we suppose that b and σ are bounded functions of class $\mathscr{C}^2(\mathbb{R})$ with bounded first and second derivatives. Explicitly state the second derivative w.r.t. x of the flow $X_t(x)$.
4. Let f be a twice continuously differentiable function from $\mathbb{R}$ to $\mathbb{R}$ such that

$$\exists C > 0,\ \exists m \in \mathbb{N}, \quad \big|f''(x)\big| \le C\big(1 + |x|^{2m}\big).$$

   Show that $|f'|$ and $|f|$ can also be bounded from above by polynomial functions.
5. Set $u(t,x) := \mathbb{E}f(X_t(x))$. Show that, for all t, the second derivative w.r.t. x of $u(t,x)$ can be bounded from above by a polynomial function.

   Hint: Apply Theorem 7.18 and use the explicit formula for the second derivative of the flow $X_t(x)$. To solve integrability issues, use the localization technique

with stopping times which has been introduced in this chapter, and get the following inequality for any adapted bounded process (ϕ_t):

$$\exists C > 0,\ \forall 0 \le t \le T, \quad \mathbb{E}\exp\left(\int_0^t \phi_\theta\, dW_\theta\right) \le C.$$

6. Now suppose that b is a *convex* function and σ is a *linear* function. Deduce from the three preceding questions that, if f is *convex and increasing*, then, for all t, $u(t,x)$ is a convex function of x.

7.4 (Estimates for the Derivatives of $x \mapsto X_t(x)$ $(\star)$) Let be given a filtered probability space and a standard one-dimensional Brownian motion (W_t).

1. Let (M_t) be a continuous martingale. Show that, for all $t > 0$ and each positive increasing function $g(t)$,

$$\max_{\theta \le t}(M_\theta) - g(t) \le \max_{0 \le \theta \le t}\big(M_\theta - g(\theta)\big).$$

2. Now fix $M_t := \int_0^t \phi_s dW_s$, where (ϕ_s) is an adapted process satisfying

$$\exists c > 0, \quad \sup_{t \ge 0} |\phi_t|^2 \le c \quad \mathbb{P}\text{-a.s.}$$

Show that, for all real numbers α,

$$\tilde{M}_t := \exp\left(\alpha M_t - \frac{\alpha^2}{2}\int_0^t (\phi_s)^2 ds\right)$$

is a square integrable martingale.

3. Show the following Doob's inequality: for all continuous martingales (N_t),

$$\forall \lambda > 0, \quad \mathbb{P}\Big(\max_{0 \le \theta \le t} |N_\theta| > \lambda\Big) \le \frac{1}{\lambda}\mathbb{E}\big(|N_t|\big).$$

Using now Question 1 and the definition of the martingale $(\tilde{M}_t)$, show that, for all $a > 0$ and all $\alpha > 0$,

$$\mathbb{P}\Big(\max_{0 \le \theta \le t} M_\theta \ge at\Big) \le \exp\left(-\alpha a t + \frac{\alpha^2}{2}ct\right).$$

Deduce that

$$\mathbb{P}\Big(\max_{0 \le \theta \le t} M_\theta \ge at\Big) \le \exp\left(-\frac{a^2 t}{2c}\right),$$

and then

$$\forall y > 0, \quad \mathbb{P}\big(|M_t| \ge y\big) \le 2\exp\left(-\frac{y^2}{2ct}\right).$$

Show that, for all ε small enough,

$$\mathbb{E}\exp\left(\varepsilon\frac{(M_t)^2}{2}\right) < \infty.$$

Hint: For all integer k, find a suitable upper bound for

$$\mathbb{E}\left[\exp\left(\varepsilon\frac{(M_t)^2}{2}\right)\mathbb{1}_{k\le|M_t|<(k+1)}\right].$$

4. We now consider the framework and assumptions of Question 1 in Problem 7.3. We again set $Y_t^x := \frac{\partial}{\partial x}X_t(x)$. Deduce from Question 3 that, for all $T > 0$, there exists a positive constant C_T such that, for all $0 < t \le T$ and all $y > 0$ large enough,

$$\mathbb{P}\big(Y_t^x > y\big) \le 2\exp\left(-\frac{(\log(y) - \mu t)^2}{2C_T t}\right).$$

5. Another application of Question 3 is the following. Let $\psi(x)$ be a Borel function such that

$$\exists m \in \mathbb{N},\ \exists C > 0,\ \forall x \in \mathbb{R},\quad \big|\psi(x)\big| \le C\big(1 + |x|^{2m}\big).$$

Use Question 3 to show that $\mathbb{E}|\psi(X_T^x)Y_T^x|^3$ is finite.

Remark In view of Sect. 3.4 this result is a first step to estimate the number N of Monte Carlo approximations of $\mathbb{E}(\psi(X_T^x)Y_T^x)$ to obtain an accuracy of order ε with confidence interval $1-\delta$ ($0 < \delta < 1$). In view of (7.65), this issue appears when one develops a Monte Carlo method to approximate $\frac{\partial u}{\partial x}(t,x)$.

Part III
Variance Reduction, Girsanov's Theorem, and Stochastic Algorithms

Variance reduction techniques are developed for the use of Monte Carlo methods involving Itô stochastic differential equations and used for the approximation of solutions of certain parabolic partial differential equations.

First a control variate method is described on a simple method. Then an application to sensitivity analysis is given. Lastly, an importance sampling method based on a well-chosen class of Girsanov transforms is developed, the optimal transform being determined by a stochastic algorithm.

The fact that the stochastic algorithm may be implemented simultaneously to the Monte Carlo method, using the same samples, is then explained. The complete algorithm provides us with an example of an adaptive Monte Carlo method.

A simple framework is then developed in order to provide a rigorous study of a class of stochastic algorithms.

Chapter 8
Variance Reduction and Stochastic Differential Equations

Abstract This chapter deepens the variance reduction subject, and focuses on the Monte Carlo methods for deterministic parabolic partial differential equations. This topic requires advanced notions in stochastic calculus, particularly the Girsanov theorem, which we state and discuss first. We strongly emphasize that universal techniques do not exist: most often, effective variance reduction methods depend on the numerical analyst's knowledge and experience. We will see that it is rather easy to construct perfect variance reduction methods which are irrelevant from a numerical point of view; a contrario, the construction of an effective method often lies on the approximation of a perfect method, the approximation method needing to be adapted to each particular case. Interesting examples can be found in Duffie and Glynn (Ann. Appl. Probab. 5(4), 897–905, 1995).

8.1 Preliminary Reminders on the Girsanov Theorem

In this section we present and comment on the Girsanov theorem in the restrictive framework of stochastic differential equations drift changes, see, e.g., Karatzas and Shreve [26, Chap. 3.5]. This will be used later for variance reduction.

Theorem 8.1 (Girsanov Theorem) *Consider the stochastic differential equation in* $\mathbb{R}^d$, *with Lipschitz coefficients*, *driven by a* d*-dimensional standard Brownian motion*,

$$X_t = X_0 + \int_0^t b(X_s)\,ds + \int_0^t \sigma(X_s)\,dW_s.$$

Given $T > 0$, *let* $(\theta_s)_{0\le s\le T}$ *be an* $\mathbb{R}^d$ *valued measurable adapted process such that*

$$\int_0^T |\theta_s|^2\,ds < \infty, \quad a.s., \tag{8.1}$$

and

$$\mathbb{E}\exp\left(\int_0^T \theta_s \cdot dW_s - \frac{1}{2}\int_0^T |\theta_s|^2\,ds\right) = 1. \tag{8.2}$$

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_8, © Springer-Verlag Berlin Heidelberg 2013

The probability measure $\mathbb{P}^\theta$ *on* $(\Omega, \mathscr{F}_T)$ *defined by its Radon–Nikodym derivative w.r.t.* $\mathbb{P}$

$$\frac{d\mathbb{P}^\theta}{d\mathbb{P}} := \exp\left(\int_0^T \theta_s \cdot dW_s - \frac{1}{2}\int_0^T |\theta_s|^2\, ds\right)$$

is equivalent to $\mathbb{P}$. *In addition, the process*

$$W_t^\theta := W_t - \int_0^t \theta_s\, ds, \quad 0 \le t \le T,$$

is an $(\Omega, \mathscr{F}_T, \mathbb{P}^\theta)$ *Brownian motion and,* $\mathbb{P}^\theta$*-a.s., for* $0 \le t \le T$,

$$X_t = X_0 + \int_0^t \big(b(X_s) + \sigma(X_s)\theta_s\big)\, ds + \int_0^t \sigma(X_s)\, dW_s^\theta. \tag{8.3}$$

Remark 8.1 The equality (8.2) is implied by the Novikov condition

$$\mathbb{E}\exp\left(\frac{1}{2}\int_0^T |\theta_s|^2\, ds\right) < \infty.$$

The change of the reference probability measure $\mathbb{P}$ to $\mathbb{P}^\theta$ is called a **Girsanov transform**.

The Girsanov theorem has the following important consequence: for any Borel functional Ψ on the space of the continuous functions from $[0, T]$ to $\mathbb{R}^d$ such that $\mathbb{E}|\Psi(X_\cdot)| < \infty$,

$$\mathbb{E}\Psi(X_\cdot) = \mathbb{E}^\theta\left(\Psi(X_\cdot)\frac{d\mathbb{P}}{d\mathbb{P}^\theta}\right)$$

where $\mathbb{E}^\theta$ stands for the expectation computed under the probability measure $\mathbb{P}^\theta$. Equivalently,

$$\begin{aligned}\mathbb{E}\Psi(X_\cdot) &= \mathbb{E}^\theta\left(\Psi(X_\cdot)\exp\left(-\int_0^T \theta_s \cdot dW_s + \frac{1}{2}\int_0^T |\theta_s|^2\, ds\right)\right)\\ &= \mathbb{E}^\theta\left(\Psi(X_\cdot)\exp\left(-\int_0^T \theta_s \cdot dW_s^\theta - \frac{1}{2}\int_0^T |\theta_s|^2\, ds\right)\right).\end{aligned} \tag{8.4}$$

8.2 Control Variates Method

As in Chap. 7 we consider the second order differential operator

$$\mathscr{L} := \sum_{i=1}^d b_i(\cdot)\partial_i + \frac{1}{2}\sum_{i,j=1}^d a_j^i(\cdot)\partial_{ij}, \tag{8.5}$$

and the parabolic partial differential equation

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + \mathscr{L}u(t,x) = 0, & (t,x) \in [0,T) \times \mathbb{R}^d, \\ u(T,x) = f(x), & x \in \mathbb{R}^d. \end{cases} \tag{8.6}$$

We recall (7.46): under suitable hypotheses,

$$u(t,x) = \mathbb{E} f\big(X_{T-t}(x)\big),$$

where

$$X_t(x) = x + \int_0^t b\big(X_s(x)\big)\, ds + \int_0^t \sigma\big(X_s(x)\big)\, dW_s, \quad 0 \le t \le T. \tag{8.7}$$

In particular, one has $u(0,x) = \mathbb{E} f(X_T(x))$. Consequently, denoting by $(\partial u)(t,x)$ the vector with coordinates $\partial_i u(t,x)$,

$$u\big(T, X_T(x)\big) = f\big(X_T(x)\big) = \mathbb{E} f\big(X_T(x)\big) + \int_0^T (\partial u)\big(t, X_t(x)\big) \cdot \sigma\big(X_t(x)\big)\, dW_t, \quad \text{a.s.}$$

Set

$$Z := f\big(X_T(x)\big) - \int_0^T (\partial u)\big(t, X_t(x)\big) \cdot \sigma\big(X_t(x)\big)\, dW_t.$$

Obviously Z is a zero bias estimator of $\mathbb{E} f(X_T(x))$ with null variance. The Monte Carlo method consisting in simulating Z would then have a null error. Of course, this remark is of no interest since the function $u(t,x)$ is unknown: else, there would be no need to develop a Monte Carlo method to approximate $\mathbb{E} f(X_T(x))$!

Nevertheless, suppose that we can get an explicit approximation $\bar{v}$ of ∂u. Then it is natural to choose $\bar{v}(t,x)$ as a control variable and to simulate the random variable

$$\bar{Z} := f\big(X_T(x)\big) - \int_0^T \bar{v}\big(t, X_t(x)\big) \cdot \sigma\big(X_t(x)\big)\, dW_t.$$

This random variable is a zero bias estimator of $\mathbb{E} f(X_T(x))$. This property is subject to the fact that the stochastic integral in the definition of Z is a martingale. Reinforcing the integrability hypotheses, one can easily check that the Monte Carlo method variance now is different to the variance of $f(X_T(x))$: one has

$$\begin{aligned} &\mathbb{E}\big|\bar{Z} - \mathbb{E} f\big(X_T(x)\big)\big|^2 \\ &\quad = \mathbb{E}\left| f\big(X_T(x)\big) - \int_0^T \bar{v}\big(t, X_t(x)\big) \sigma\big(X_t(x)\big)\, dW_t - \mathbb{E} f\big(X_T(x)\big)\right|^2 \\ &\quad = \mathbb{E} \int_0^T \big|\big((\partial u)\big(t, X_t(x)\big) - \bar{v}\big(t, X_t(x)\big)\big) \cdot \sigma\big(X_t(x)\big)\big|^2\, dt. \end{aligned}$$

Thus the new variance is small when $\bar{v}$ is a good approximation of the function u in the sense that the r.h.s. of the preceding equality is small. In that case, one approximates the quantity

$$\int_0^T (\partial u)\big(t, X_t(x)\big)\sigma\big(X_t(x)\big)\, dW_t$$

by the sum

$$\sum_{p=0}^{n-1} \bar{v}\big(pT/n, X^n_{pT/n}(x)\big)\sigma\big(X^n_{pT/n}(x)\big)(W_{(p+1)T/n} - W_{pT/n}).$$

The main difficulty in this approach introduced by Newton [39] consists in choosing a suitable function $\bar{v}$. There does not exist a universal methodology to achieve this objective. An effective way may be to choose a diffusion process $(\bar{X}_t)$ whose probability distribution at time T is explicitly known and is close to the law of X_T.

For example, consider the computation of the price of a European option in a stochastic volatility model with stochastic interest rate

$$dS_t = r_t S_t\, dt + \sigma_t S_t\, dW_t,$$

where the pair (r_t, σ_t) solves a stochastic differential equation of the type

$$\begin{cases} dr_t = \beta_r(r_t)\, dt + \gamma_r(r_t)\, dB^r_t, \\ d\sigma_t = \beta_\sigma(\sigma_t)\, dt + \gamma_\sigma(\sigma_t)\, dB^\sigma_t. \end{cases}$$

In this model, the process (W_t, B^r_t, B^σ_t) is a Brownian motion whose components are weakly correlated, and the coefficients β_r, γ_r, β_σ, γ_σ are such that r_t is close to an affine model for short-term interest rates and is independent of (S_t), so that (σ_t) can be interpreted as a fluctuation process around a constant and deterministic value.

It then is reasonable to choose $\bar{v}(t, S_t)$ as the product of the price of a zero coupon bond in the affine model under consideration and the price of the option in the Black and Scholes model.

8.3 Variance Reduction for Sensitivity Analysis

This section explains how to compute both the solution and some of its partial derivatives for some of its parameters in an effective way.

8.3.1 Differentiable Terminal Conditions

Again consider the partial differential equation (8.6). It often occurs that one desires to compute both the solution and some of its partial derivatives or its sensitivity to

model parameters. For example, in financial mathematics, the spatial derivative of the solution provides the hedging strategy for the option.

A naive Monte Carlo method consists in using a finite difference approximation of the derivative under consideration, that is,

$$\partial_i u(t,x) \simeq \frac{u(t,x+\varepsilon e_i) - u(t,x)}{\varepsilon},$$

where e_i is the vector $(0,\dots,0,1,0,\dots,0)$ with 1 at position i, and then in approximating each term in the numerator by means of a Monte Carlo method. This algorithm is numerically unstable and is rather ineffective: as ε needs to be chosen small to get a good approximation of the derivative, one needs to approximate the numerator with a high accuracy and therefore to use large numbers of simulations and very small time discretization steps.

To overcome this difficulty and reduce the variance, one can use the same samples of the Brownian increments to approximate $u(t,x+\varepsilon e_i)$ and $u(t,x)$. This is equivalent to perturbing the initial condition of each path of the Euler scheme. It often is more effective to start with Theorem 7.18:

$$\partial_i u(t,x) = \partial_i \mathbb{E} f\big(X_{T-t}(x)\big) = \mathbb{E}\sum_{j=1}^{d} \partial_j f\big(X_{T-t}(x)\big)\partial_i X^j_{T-t}(x). \tag{8.8}$$

We recall that, in this equality, the vector $\partial_i X_\theta(x)$ satisfies

$$\partial_i X_\theta(x) = e_i + \int_0^\theta \partial b\big(X_s(x)\big)\partial_i X_s(x)\,ds + \sum_{j=1}^{r}\int_0^\theta \partial\sigma_j\big(X_s(x)\big)\partial_i X_s(x)\,dW^j_s. \tag{8.9}$$

Consequently, the partial derivatives of $u(t,x)$ can be approximated by means of a Monte Carlo method and the simulation of the Euler scheme which discretizes (8.7) and the linear system

$$Y_t(x) = \mathrm{Id} + \int_0^t \partial b\big(X_s(x)\big)Y_s(x)\,ds + \sum_{\ell=1}^{r}\partial\sigma_\ell\big(X_s(x)\big)Y_s(x)\,dW^\ell_s$$

with matrix valued solution

$$\big\{Y_t(x)\big\}^j_i = \partial_j X^i_t(x).$$

Remark 8.2 In the above we have differentiated $u(t,x)$ w.r.t. x. One can also differentiate $u(t,x)$ w.r.t. parameters in the coefficients b and σ: if these coefficients depend smoothly on their parameters, one can extend the equalities (8.8) and (8.9) by replacing ∂_i by the derivation operator w.r.t. the parameter under consideration. This observation allows one, e.g., to develop Monte Carlo methods to compute Greeks in Financial Mathematics.

8.3.2 Non-differentiable Terminal Conditions

The methodology presented in the preceding subsection supposes that the function f is everywhere differentiable. How to extend it to the case where f is, e.g., piecewise smooth and has a finite number of discontinuities? It often is ineffective to approximate it by a smooth function: for example, consider the smooth approximation $g_\varepsilon * f$ obtained by convoluting f and the Gaussian density g_ε with zero mean and variance ε; the derivatives of $g_\varepsilon * f$ take large values in the neighborhoods of the discontinuity points of f (as expected since, in the limit $\varepsilon = 0$, the derivatives at these points are infinite). To circumvent this difficulty, one can use a kind of integration by parts which extends the following observation: let X be a real-valued random variable with an everywhere differentiable and strictly positive density p; for any function φ of class $\mathscr{C}^1(\mathbb{R})$ such that all the integrals below are well defined,

$$\begin{aligned}\mathbb{E}\varphi'(X) &= \int_{\mathbb{R}} \varphi'(x)p(x)\,dx = -\int_{\mathbb{R}} \varphi(x)p'(x)\,dx = -\int_{\mathbb{R}} \varphi(x)\frac{p'(x)}{p(x)}p(x)\,dx \\ &= -\mathbb{E}\left[\varphi(X)\frac{p'(X)}{p(X)}\right].\end{aligned}$$

Consequently, if φ' locally takes large values, rather than using a Monte Carlo method to approximate $\mathbb{E}\varphi'(X)$, it may be effective to reduce the variance by using a Monte Carlo method to approximate $\mathbb{E}[\varphi(X)\frac{p'(X)}{p(X)}]$.

Let us now come back to the case of the stochastic differential equations. To simplify the presentation, let $d = 1$.

Theorem 8.2 *Let the coefficients b and σ be Lipschitz functions, and σ satisfy*

$$\exists C_1 > 0,\ \exists C_2 > 0, \quad C_1 \le \sigma(x) \le C_2 \quad \textit{for all } x \in \mathbb{R}.$$

Let f be a differentiable function with derivative f' having an at most polynomial growth at infinity. Then

$$\mathbb{E}\big(f'(X_T)Y_T\big) = \mathbb{E}\left(f(X_T)\frac{1}{T}\int_0^T \frac{Y_s}{\sigma(X_s)}\,dW_s\right),$$

where

$$Y_t = 1 + \int_0^t b'(X_s)Y_s\,ds + \sigma'(X_s)Y_s\,dW_s, \quad 0 \le t \le T. \tag{8.10}$$

Proof Let $(\mathscr{F}_t)$ be the Brownian motion (W_t) filtration. Let (θ_s) be an $(\mathscr{F}_t)$ adapted process which satisfies the condition (8.1) and such that, for all $0 \le \varepsilon \le 1$,

$$\mathbb{E}\exp\left(\varepsilon\int_0^T \theta_s \cdot dW_s - \frac{\varepsilon^2}{2}\int_0^T |\theta_s|^2\,ds\right) = 1.$$

For any $\varepsilon > 0$ define a new probability measure on $(\Omega, \mathscr{F}_T)$ by

$$\frac{d\mathbb{P}^\varepsilon}{d\mathbb{P}} := \exp\left(\varepsilon \int_0^T \theta_s \cdot dW_s - \frac{\varepsilon^2}{2} \int_0^T |\theta_s|^2 \, ds\right).$$

Setting

$$W_t^\varepsilon := W_t - \varepsilon \int_0^t \theta_s \, ds,$$

the equality (8.4) becomes (Girsanov transform)

$$\mathbb{E} f(X_T) = \mathbb{E}^\varepsilon \left(f(X_T) \exp\left(-\varepsilon \int_0^T \theta_s \, dW_s^\varepsilon - \frac{\varepsilon^2}{2} \int_0^T |\theta_s|^2 \, ds\right)\right). \tag{8.11}$$

By definition, the joint probability distribution under $\mathbb{P}^\varepsilon$ of the process (W_t^ε, X_t) is identical to the joint probability distribution under $\mathbb{P}$ of the process (W_t, X_t^ε), where

$$X_t^\varepsilon = X_0 + \int_0^t \left(b\left(X_s^\varepsilon\right) + \varepsilon \sigma\left(X_s^\varepsilon\right)\theta_s\right) ds + \int_0^t \sigma\left(X_s^\varepsilon\right) dW_s. \tag{8.12}$$

Therefore

$$\mathbb{E} f(X_T) = \mathbb{E}\left(f\left(X_T^\varepsilon\right) \exp\left(-\varepsilon \int_0^T \theta_s \, dW_s - \frac{\varepsilon^2}{2} \int_0^T |\theta_s|^2 \, ds\right)\right). \tag{8.13}$$

Observe that the r.h.s. does not depend on ε. Consequently the derivative w.r.t. ε of the l.h.s. is null. This derivative can be made explicit by using Remark 8.2. One thus obtains

$$0 = \mathbb{E}\left(f'\left(X_T^\varepsilon\right) \partial_\varepsilon X_T^\varepsilon \exp\left(-\varepsilon \int_0^T \theta_s \, dW_s - \frac{\varepsilon^2}{2} \int_0^T |\theta_s|^2 \, ds\right)\right) + \mathbb{E}\left(f\left(X_T^\varepsilon\right) \xi_T^\varepsilon\right), \tag{8.14}$$

where (see the first question of Problem 7.3 for a similar calculation)

$$\partial_\varepsilon X_t^\varepsilon = \int_0^t \left(b'(X_s) + \varepsilon \sigma'\left(X_s^\varepsilon\right)\theta_s\right) \partial_\varepsilon X_s^\varepsilon \, ds + \int_0^t \sigma\left(X_s^\varepsilon\right)\theta_s \, ds + \int_0^t \sigma'\left(X_s^\varepsilon\right) \partial_\varepsilon X_s^\varepsilon \, dW_s,$$

and

$$\xi_T^\varepsilon = \left(-\int_0^T \theta_s \, dW_s - \varepsilon \int_0^T |\theta_s|^2 \, ds\right) \exp\left(-\varepsilon \int_0^T \theta_s \, dW_s - \frac{\varepsilon^2}{2} \int_0^T |\theta_s|^2 \, ds\right).$$

When ε tends to 0, classical results on the convergence of stochastic integrals (see Sect. 7.1.1) imply that the process (X_t^ε) converges in probability to (X_t^ε), and the random variable ξ_T^ε converges in probability to $-\int_0^T \theta_s \, dW_s$. In addition, the process $(\partial_\varepsilon X_t^\varepsilon)$ converges in probability to the solution (U_t) of

$$U_t = \int_0^t \left(b'(X_s) U_s + \sigma(X_s)\theta_s\right) ds + \int_0^t \sigma'(X_s) U_s \, dW_s.$$

Itô's formula shows that

$$U_T = Y_T \int_0^T \frac{\theta_s \sigma(X_s)}{Y_s}\, ds,$$

where (Y_t) is the process defined in (8.10).

In conclusion, in view of (8.14), we have obtained that

$$0 = \mathbb{E}\left(f'(X_T) Y_T \int_0^T \frac{\theta_s \sigma(X_s)}{Y_s}\, ds - f(X_T) \int_0^T \theta_s\, dW_s \right).$$

It then remains to choose

$$\theta_s := \frac{Y_s}{\sigma(X_s)}$$

to obtain the desired result. □

Exercise 8.1 Identify situations where the process (θ_s) satisfies the hypotheses (8.1) and (8.2).

Remark 8.3 Multi-dimensional extensions of the integration by parts formula in Theorem 8.2 can be proved. However the resulting formula is more complex than in the preceding one-dimensional context and, in full generality, its proof is based on Malliavin calculus which is far beyond the scope of this monograph. For a summary, examples and comments, see Fournié et al. [17, 18]. A classical textbook on Malliavin calculus is Nualart [42].

8.4 Importance Sampling Method

In this section we develop applications of the Girsanov theorem to variance reduction. To simplify the notation, let $d = 1$ and

$$M := \mathbb{E} f(X_T).$$

We aim to approximate M using a Monte Carlo method. Under probability $\mathbb{P}$, the variance of the simulation is

$$\mathbb{E}\left[\left(f(X_T) \right)^2 \right] - M^2.$$

Let us show that, under suitable hypotheses on the function $u(t, x)$, **in theory** one can construct a process (θ_t^*) and the corresponding Girsanov transform which lead one to a **perfect** importance sampling variance reduction method. Actually, suppose that $u(t, x)$ is bounded from below by a strictly positive number for all t (in particular, f is supposed to satisfy this property). Then, for all (θ_t),

$$\mathbb{E} f(X_T) = \mathbb{E}^\theta \left[f(X_T) \frac{\mathbb{E} f(X_T)}{f(X_T)} \right].$$

This equality suggests to look for a process (θ_t) such that the random variable $\frac{\mathbb{E}f(X_T)}{f(X_T)}$ defines a Girsanov transform with Radon–Nikodym density $\frac{d\mathbb{P}}{d\mathbb{P}^\theta}$. As seen in theorem 7.14 we have

$$u(0, X_0) = \mathbb{E}u(T, X_T) = \mathbb{E}f(X_T).$$

We thus also have

$$\frac{\mathbb{E}f(X_T)}{f(X_T)} = \exp\big(\log u(0, x_0) - \log u(T, X_T)\big),$$

from which, applying Itô's formula to $\log u(t, X_t)$ between times 0 and T, using (8.6) and choosing

$$\theta_t := \frac{\sigma(X_t)\frac{\partial}{\partial x}u(t, X_t)}{u(t, X_t)},$$

we get

$$\frac{\mathbb{E}f(X_T)}{f(X_T)} = \exp\bigg(-\int_0^T \theta_s\, dW_s + \frac{1}{2}\int_0^T |\theta_s|^2\, ds\bigg).$$

In view of (8.4), our choice for (θ_t) defines a proper Girsanov transform if the conditions (8.1) and (8.2) are satisfied. The variance of the simulation under $\mathbb{P}^\theta$ then is

$$\begin{aligned}\mathbb{E}^\theta\bigg(\bigg(f(X_T)\frac{d\mathbb{P}}{d\mathbb{P}^\theta}\bigg)^2\bigg) - M^2 &= \mathbb{E}\bigg(\big(f(X_T)\big)^2\frac{d\mathbb{P}}{d\mathbb{P}^\theta}\bigg) - M^2 \\ &= \mathbb{E}\bigg(\big(f(X_T)\big)^2\frac{\mathbb{E}f(X_T)}{f(X_T)}\bigg) - M^2 = 0.\end{aligned}$$

We thus have constructed a **perfect** variance reduction. Unfortunately, it requires to know the function $u(t, x)$ which we aim to approximate...

We thus now limit ourselves to a class of processes (θ_t) which can be simulated. Under any probability measure $\mathbb{P}^\theta$, the variance of the simulation is

$$\mathbb{E}^\theta\bigg(\big(f(X_T)\big)^2\exp\bigg(-2\int_0^T \theta_s \cdot dW_s^\theta - \int_0^T |\theta_s|^2\, ds\bigg)\bigg) - M^2,$$

where (X_t) is the solution to (8.3). Observe that the expression of $\frac{d\mathbb{P}}{d\mathbb{P}^\theta}$ and the definition of W^θ imply that this variance is also equal to

$$\mathbb{E}\bigg(\big(f(X_T)\big)^2\exp\bigg(-\int_0^T \theta_s \cdot dW_s + \frac{1}{2}\int_0^T |\theta_s|^2\, ds\bigg)\bigg) - M^2.$$

Therefore our objective is to seek a process (θ_t) such that

$$\mathbb{E}\bigg(\big(f(X_T)\big)^2\exp\bigg(-\int_0^T \theta_s \cdot dW_s + \frac{1}{2}\int_0^T |\theta_s|^2\, ds\bigg)\bigg) \ll \mathbb{E}\big[\big(f(X_T)\big)^2\big].$$

We formulate this issue as the following optimization problem:

$$\min_{(\theta_t)\in\mathscr{A}} \mathbb{E}\left(\left(f(X_T)\right)^2 \exp\left(-\int_0^T \theta_s \cdot dW_s + \frac{1}{2}\int_0^T |\theta_s|^2\, ds\right)\right),$$

where $\mathscr{A}$ is a class of adapted processes satisfying suitable integrability conditions. Obviously this problem is impossible to solve in practice if the class $\mathscr{A}$ is too large. One thus needs to find a good compromise between restricting the class $\mathscr{A}$ in order to be able to compute a minimum (θ_t) and choosing $\mathscr{A}$ large enough to obtain an effective variance reduction.

In Chap. 3, within the framework of the Black and Scholes model, we have presented a variance reduction method based on constant processes (θ_t), see (3.17). We may extend this approach to more general models and to non-constant deterministic functions (θ_t): the time step T/n being fixed, the Euler scheme may be considered as a **deterministic** (and complex!) function F_n of the initial condition X_0 and the vector of the Brownian increments

$$\Delta W := (W_{(p+1)T/n} - W_{pT/n},\ p = 0, \dots, n-1).$$

Therefore, extending the methodology introduced in Chap. 3 we are led to determine the piecewise constant function

$$\Theta := (\theta_{(p-1)T/n},\ p = 1, \dots, n)$$

which minimizes

$$H(\Theta) := \mathbb{E}\left[\left(f \circ F_n(\Delta W)\right)^2 \exp\left(-\Theta \cdot \Delta W + \frac{T}{2n}\|\Theta\|^2\right)\right].$$

In the next chapter we will see stochastic algorithms which provide the optimal value of Θ. We will also see how to modify the Monte Carlo method to **simultaneously** optimize the choice of Θ and approximate M.

A rigorous analysis shows that the exponential term in the expression of $H(\Theta)$ imposes that these algorithms include a rather complex projection procedure at each time step in order to converge to the desired optimal value.

In order to define a low complexity algorithm, we can reduce our ambition to compute sub-optimal values Θ. To this end, we observe that the preceding optimization problem is equivalent to minimizing a particular entropic distance between $\mathbb{P}$ and $\mathbb{P}^\theta$, namely the distance defined on the set of probability measures as

$$\mathscr{R}_2(\mathbb{P}, \mathscr{Q}) = \begin{cases} \frac{1}{2}\int \left(\left(\frac{d\mathbb{P}}{d\mathscr{Q}}\right)^2 - 1\right) d\mathscr{Q} & \text{if } \mathbb{P} \text{ is absolutely continuous w.r.t. } \mathscr{Q}, \\ +\infty & \text{otherwise.} \end{cases}$$

One may hope to reduce the variance of the simulation by minimizing another entropic distance, for example the Kullback–Leibler distance, between $\mathbb{P}^\theta$ and $\mathbb{P}^{\theta^*}$,

where θ^* is the process θ which minimizes $H(\Theta)$. The Kullback–Leibler distance is defined as

$$\mathscr{R}_0(\mathbb{P}, \mathscr{Q}) = \begin{cases} \int \log\left(\frac{d\mathbb{P}}{d\mathscr{Q}}\right) d\mathbb{P} & \text{if } \mathbb{P} \text{ is absolutely continuous w.r.t. } \mathscr{Q}, \\ +\infty & \text{otherwise.} \end{cases}$$

For details, see the thesis of O. Bardou.[1]

In Chap. 9 we will see that the adaptive Monte Carlo method does not require a projection procedure at each time step.

Exercise 8.2 Check that $\mathscr{R}_0(\mathbb{P}, \mathscr{Q}) \leq 2\mathscr{R}_2(\mathbb{P}, \mathscr{Q})$.

Hint: Consider the function $x^2 - 1 + x - 1 - x\log(x)$. Deduce that solving the preceding minimization problem actually allows one to decrease the variance of the simulation. Exhibit an example where this variance reduction is less effective than the one obtained by minimizing $H(\Theta)$.

8.5 Statistical Romberg Method

In Chap. 7 we have shown that, under suitable hypotheses on the coefficients b and σ, the Euler scheme discretization error satisfies: for any smooth enough function f and time horizon T, there exists a real number C_f^T such that, for all $n \geq 1$,

$$\mathbb{E}f\left(X_T^n\right) - \mathbb{E}f(X_T) = \frac{C_f^T}{n} + \mathscr{O}\left(\frac{1}{n^2}\right).$$

We admit the complementary following weak convergence result, due to Duffie and Glynn [11], and we do not make precise the hypotheses under which it holds true: if $N = N(n) = n^2$, then

$$n\left(\frac{1}{n^2}\sum_{\ell=1}^{n^2} f\left(X_T^{n(\ell)}\right) - \mathbb{E}f(X_T)\right) \xrightarrow[n\to\infty]{\text{in law}} \sigma G + C_T^f,$$

where G is a standard Gaussian random variable, $\sigma^2 := \mathrm{Var}(f(X_T))$ and C_f^T is defined as in the preceding equality.

Therefore, to achieve an accuracy of order $1/n$, the global complexity of the simulation is of order $n^2 \times n = n^3$. This observation suggests to use the Richardson–Romberg extrapolation technique (see Chap. 7) not only to reduce the discretization error but also to reduce the variance of the simulation. The principle of this new method due to Kebaier [27] is as follows.

[1] Bardou, O.: Contrôle dynamique des erreurs de simulation et d'estimation de processus de diffusion. Ph.D. thesis, Université de Provence (2005).

Choose a thin grid with step-size T/n and a coarse grid with step-size T/m_n where $m_n \ll n$. The corresponding Euler schemes are respectively denoted by X^n and X^{m_n}. Suppose that the quantity $\bar{M}^{m_n} := \mathbb{E} f(X_T^{m_n})$ is explicitly known and set

$$Y := f\big(X_T^n\big) - f\big(X_T^{m_n}\big) + \bar{M}^{m_n}.$$

Then $\mathbb{E}Y = \mathbb{E} f(X_T^n)$, and the variance of Y is of order $1/m_n$. It then remains to suitably choose $\bar{M}^{m_n}$. It is effective to estimate it by means of a Monte Carlo method based on N_{m_n} independent simulations of X^{m_n}. One then uses N_n new independent simulations to approximate $\mathbb{E}Y$.

To summarize, the algorithm consists in computing

$$\bar{Y} := \frac{1}{N_{m_n}} \sum_{\ell=1}^{N_{m_n}} f\big(\hat{X}_T^{m_n(\ell)}\big) + \frac{1}{N_n} \sum_{\ell=1}^{N_n} \big\{ f\big(X_T^{n(\ell)}\big) - f\big(X_T^{m_n(\ell)}\big) \big\}.$$

We now have to optimally choose the parameters n, N_{m_n} and N_n to achieve a global accuracy of order $1/n$. A Central Limit Theorem, which we do not state here, shows that good choices are

$$N_{m_n} = n^2, \qquad N_n = n^{3/2}.$$

For $m_n = \sqrt{n}$ the global complexity is of order $n^{5/2}$, which is a substantial reduction of the standard Monte Carlo method complexity, which we have seen is of order n^3.

Remark 8.4 So far we have supposed that the function f is smooth. However one can prove the following result: if the coefficients b and σ are of class $\mathscr{C}^\infty(\mathbb{R}^d)$ with smooth bounded derivatives of all orders, then, for any differentiable function f,

$$\lim_{n\to\infty} \sqrt{n}\big(\mathbb{E} f(X_T) - \mathbb{E} f\big(X_T^n\big)\big) = 0.$$

8.6 Problems

8.1 (Importance Sampling for Stochastic Differential Equations) We are given a time horizon $T > 0$, coefficients b and σ from $\mathbb{R}$ to $\mathbb{R}$, a real-valued Brownian motion (W_t), and a real number x. We suppose that b and σ are bounded functions of class $\mathscr{C}^\infty(\mathbb{R})$ with bounded derivatives of all orders. Consider the solution $(X_t(x))$ to the stochastic differential equation

$$X_t(x) = x + \int_0^t b\big(X_s(x)\big)\, ds + \int_0^t \sigma\big(X_s(x)\big)\, dW_s, \quad 0 \le t \le T.$$

1. Let f be a function of class $\mathscr{C}^\infty(\mathbb{R})$ with bounded derivatives. Identify a function ϕ such that

$$f\big(X_T(x)\big) = \mathbb{E} f\big(X_T(x)\big) + \int_0^T \phi\big(X_s(x)\big)\, dW_s.$$

   Hint: Express ϕ in terms of the function σ and the x-derivative of the solution of a well-chosen partial differential equation.
2. We are given coefficients $\hat{b}$ and $\hat{\sigma}$ which are algebraically simpler than b and σ and satisfy the same smoothness properties. Consider the solution $(\hat{X}_t(x))$ to the stochastic differential equation

$$\hat{X}_t(x) = x + \int_0^t \hat{b}\big(\hat{X}_s(x)\big)\, ds + \int_0^t \hat{\sigma}\big(\hat{X}_s(x)\big)\, dW_s, \quad 0 \le t \le T.$$

   Show that there exists a real number C depending on x, T, b and σ, but not on $\hat{b}$ and $\hat{\sigma}$, such that

$$\mathbb{E}\big|X_T(x) - \hat{X}_T(x)\big|^2 \le C\|b - \hat{b}\|_\infty^2 + C\|\sigma - \hat{\sigma}\|_\infty^2.$$

3. Show that there exists a real number C depending on x, T, b, σ, $\hat{b}$ and $\hat{\sigma}$, such that

$$\mathbb{E}\left|\frac{d}{dx}X_T(x) - \frac{d}{dx}\hat{X}_T(x)\right|^2 \le C\big\|b' - \hat{b}'\big\|_\infty^2 + C\big\|\sigma' - \hat{\sigma}'\big\|_\infty^2.$$

4. Let $\hat{\phi}$ be chosen as above such that

$$f\big(\hat{X}_T(x)\big) = \mathbb{E} f\big(\hat{X}_T(x)\big) + \int_0^T \hat{\phi}(\hat{X}_s)\, dW_s.$$

   Exhibit, in terms of $\|b - \hat{b}\|_\infty$, $\|\sigma - \hat{\sigma}\|_\infty$, $\|b' - \hat{b}'\|_\infty$, and $\|\sigma' - \hat{\sigma}'\|_\infty$, a bound for

$$\mathbb{E}\left|\int_0^T \phi\big(X_s(x)\big)\, dW_s - \int_0^T \hat{\phi}\big(\hat{X}_s(x)\big)\, dW_s\right|^2.$$

5. Develop an application of the preceding estimate to reduce the variance of the standard Monte Carlo method which approximates $\mathbb{E} f(X_T)$.

8.2 (Variance Reduction for the Computation of the Delta of a European Option ($\star$)) Consider the stochastic volatility model for asset prices

$$S_t(\xi) = \xi + r\int_0^T S_\theta\, d\theta + \int_0^t \sigma(S_\theta) S_\theta\, dW_\theta,$$

where r stands for the interest rate which is assumed deterministic and constant, and σ is a function which is bounded from below and from above by strictly positive constants and has bounded derivatives of all orders.

We are given a maturity $T > 0$ and a constant $K > 0$. For $0 \leq t \leq T$ and $\xi > 0$ let

$$u(t, \xi) := \mathbb{E}\big(\mathbb{1}_{\{S_{T-t}(\xi) < K\}}\big) = \mathbb{P}\big(S_{T-t}(\xi) < K\big).$$

We are interested in constructing a Monte Carlo method for $\frac{\partial u}{\partial \xi}(t, \xi)$.

1. Exhibit a probabilistic representation of $\frac{\partial u}{\partial \xi}(t, \xi)$, by approximating the function $x \mapsto \mathbb{1}_{\{x<K\}}$ by a sequence of smooth functions and by applying Theorem 8.2 to $X_t := \log(S_t(\xi))$.
2. Deduce a variance reduction method for the Monte Carlo approximation of $\frac{\partial u}{\partial \xi}(t, \xi)$.
3. Examine the particular case of the log-normal model in which the function σ is constant.

Chapter 9
Stochastic Algorithms

Abstract This chapter addresses a few theoretical and practical issues raised by a class of stochastic optimization or approximation algorithms. These algorithms are efficient numerical tools in various applications and they are at the basis of the variance reduction techniques studied in the preceding chapter. We limit ourselves to a simple framework which allows us to get convergence results by means of dynamical systems and martingales theories without too many technical details. In this, we follow the very pedagogical approach in Benaïm and El Karoui (Chaînes de Markov et simulations; Martingales et stratégies, 2004).

Classic reference books on the topic are, e.g., Kusher and Yin [31] and Duflo [12, 13], see also Asmussen and Glynn [4].

9.1 Introduction

The terminology "stochastic algorithm" (sometimes "optimization" or "approximation" is added between these terms) corresponds to a vast and imprecisely delimited class of stochastic recursive sequences, which can easily be implemented using a computer.

The task on hand is to solve reasonably well a problem featuring a parameter on which there is some control. A stochastic algorithm is devised to update the parameter iteratively, so as to make it close to some "nice" value, which is possibly optimal in some way to be specified.

The structure of the algorithm and the precise sense of optimality may vary according to the applications, and influence the assumptions and mathematical tools for proofs, of convergence or other. The nature of the stochasticity may be very diverse, and can for instance:

- be due to the fact that the observation of the phenomenon is only partial or noisy,
- come from a modelization trying to simplify the complexity of the phenomenon, as for certain statistical physics models, or for certain chaotic behaviors,
- be due to the fact that the function to optimize is ill known, or more generally that one does not know how to devise a deterministic optimization method (which requires precise knowledge and regularity),

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1_9, © Springer-Verlag Berlin Heidelberg 2013

- be introduced, for a function with numerous local optima (tormented energy landscape) that would trap deterministic optimization methods (such as gradient descent), so as to escape local optima and find global ones (stochastic gradient, Hastings–Metropolis algorithms, simulated annealing, ...),
- be inherent to the modelization, as in case of the behavior of individuals or the effects of environment or mutations in ecological or evolutionary scenarios, or for customer behavior in communication networks.

9.2 Study in an Idealized Framework

Given the wide variety of situations and of techniques found in this domain, we study a framework sufficiently simple to remain accessible, and sufficiently wide to illustrate the main ideas in play.

9.2.1 Definitions

The class under study can be defined as follows.

Definition 9.1 A *stochastic algorithm* is a random sequence $(\theta_n)_{n\geq 0}$ with values in $\mathbb{R}^d$, adapted to a filtration $(\mathscr{F}_n)_{n\geq 0}$, that can be written in the form

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\big(F(\theta_n) + U_{n+1}\big), \quad n \geq 0,$$

for a function $F : \mathbb{R}^d \to \mathbb{R}^d$ and a real sequence $(\gamma_n)_{n\geq 1}$ which are *deterministic* and satisfy

$$\gamma_n > 0, \quad \lim_{n\to\infty} \gamma_n = 0, \quad \sum_{n\geq 1} \gamma_n = \infty,$$

and an $(\mathscr{F}_n)_{n\geq 0}$-adapted sequence $(U_n)_{n\geq 1}$ such that $\mathbb{E}(U_{n+1} \mid \mathscr{F}_n) = 0$ for $n \geq 0$.

The γ_n are said to be the *gains* of the algorithm, and the U_n to be *martingale increments*. Regularity assumptions must be made on F, the least of which is that it be *locally Lipschitz*.

Motivation and Examples

The purpose of such an algorithm is the search for a *specific root* of the equation

$$F(\theta) = 0.$$

For instance, in order to reach a desired level v for a function f of a control $\theta \in \mathbb{R}$, one may consider

$$F(\theta) = v - f(\theta) \quad \text{if } f \text{ is non-decreasing,}$$
$$F(\theta) = f(\theta) - v \quad \text{if } f \text{ is non-increasing;}$$

to seek an extremum of a function $V \in C^1$, one may consider

$$F(\theta) = -\nabla V(\theta) \quad \text{for a minimum,} \qquad F(\theta) = \nabla V(\theta) \quad \text{for a maximum.}$$

Varied recursively-defined random sequences may be written in the form in Definition 9.1, as in the next fundamental result. This allows implementing stochastic algorithms in which F and the U_n are *implicit*, but F can be defined by an expectation, which one *does not know how to compute*, of a *simulable* r.v. This is a crucial fact in a Monte Carlo framework.

Lemma 9.1 *Let a function $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ and an i.i.d. sequence $(X_n)_{n\geq 1}$ of r.v.'s with values in $\mathbb{R}^m$ and law P be such that*

$$F(\theta) := \mathbb{E}\big(H(\theta, X_1)\big) = \int H(\theta, x)\, P(dx) \quad \text{is defined for all } \theta \in \mathbb{R}^d,$$

and a sequence $(\gamma_n)_{n\geq 1}$ be as in Definition 9.1. Then, the sequence $(\theta_n)_{n\geq 0}$ of $\mathbb{R}^d$-valued r.v.'s defined by θ_0 independent of $(X_n)_{n\geq 1}$ and

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1}), \quad n \geq 0,$$

is a stochastic algorithm *in the sense of Definition 9.1. Specifically, $\mathscr{F}_n = \sigma(\theta_0, X_1, \ldots, X_n)$ and F is given above and $U_{n+1} = H(\theta_n, X_{n+1}) - F(\theta_n)$.*

Proof Then $\mathbb{E}(H(\theta_n, X_{n+1}) \,|\, \mathscr{F}_n) = F(\theta_n)$ since θ_n is $\mathscr{F}_n$-measurable and X_{n+1} is independent of $\mathscr{F}_n$, hence

$$\begin{aligned}\theta_{n+1} &= \theta_n + \gamma_{n+1}\big(\mathbb{E}\big(H(\theta_n, X_{n+1}) \,|\, \mathscr{F}_n\big) + H(\theta_n, X_{n+1}) - \mathbb{E}\big(H(\theta_n, X_{n+1}) \,|\, \mathscr{F}_n\big)\big)\\ &= \theta_n + \gamma_{n+1}\big(F(\theta_n) + U_{n+1}\big)\end{aligned}$$

with $\mathbb{E}(U_{n+1} \,|\, \mathscr{F}_n) = \mathbb{E}(H(\theta_n, X_{n+1}) - \mathbb{E}(H(\theta_n, X_{n+1}) \,|\, \mathscr{F}_n) \,|\, \mathscr{F}_n) = 0$. □

Results on stochastic algorithms will be given with assumptions bearing on the form given in Definition 9.1. In order to apply them using Lemma 9.1, these assumptions will have to be checked starting from the given P and H.

Instability and Projection

In practice, stochastic algorithms are often unstable, but may be stabilized by adequate *projection operations*, as was done for the variance reduction method previ-

ously described in this book. The parameter θ is then limited to a set Θ of *admissible* parameters, which is usually compact.

There may be other practical reasons for such projections, as when one cannot supply arbitrary values of θ to the actual system (which could be an input rate, a temperature, etc.), or if one seeks an optimum and knows that values outside Θ are far from optimal.

The projection operation consists in not accepting θ_{n+1} if it is outside Θ, and letting the algorithm start anew from a new value $\theta'_{n+1} \in \Theta$. The choice of this value may be varied, and involve θ_n and θ_{n+1}, as for $\theta'_{n+1} = \theta_n$, or when θ'_{n+1} is a boundary point of Θ on the segment between θ_n and θ_{n+1}.

The projection greatly complexifies the mathematical analysis, and we shall not try to give proofs in this case, but it is often intuitive and easy to code.

The results and proofs that we shall give thus correspond to a "standard" framework which does not take into account such features.

9.2.2 The Ordinary Differential Equation Method, Martingale Increments

Consider the ordinary differential equation (ODE) in Definition 6.2 for $t \in \mathbb{R}_+$ with a change of notation, replacing the vector field b by $F : \mathbb{R}^d \to \mathbb{R}^d$, i.e., for $t \in \mathbb{R}_+$,

$$\frac{d}{dt} x_t = F(x_t) \quad \text{and} \quad x_0 = x, \quad \text{and precisely} \quad x_t = x + \int_0^t F(x_s)\, ds. \tag{9.1}$$

Assume that there exists a *flow of homeomorphisms* $(\phi_t)_{t \in \mathbb{R}_+}$: for all $x \in \mathbb{R}^d$ and $t \in \mathbb{R}_+$ it holds that $\phi_t(x) = x + \int_0^t F(\phi_s(x))\, ds$ (see Theorem 6.8, e.g.).

This differential equation plays an important role in the study of the stochastic algorithm in Definition 9.1, through the *continuous* process $(X_t)_{t \in \mathbb{R}_+}$ given by

$$\begin{cases} \tau_0 = 0, \qquad \tau_{n+1} = \tau_n + \gamma_{n+1} = \gamma_1 + \cdots + \gamma_{n+1}, \quad n \in \mathbb{N}, \\ X_t = \theta_n + (t - \tau_n) \dfrac{\theta_{n+1} - \theta_n}{\gamma_{n+1}} = X_{\tau_n} + (t - \tau_n)(F(X_{\tau_n}) + U_{n+1}), \quad t \in [\tau_n, \tau_{n+1}]. \end{cases} \tag{9.2}$$

This is an interpolation for the algorithm in an adequate time-scale. It looks like an *Euler scheme* for the ordinary differential equation, on an irregular grid with mesh going to 0 (see (9.3) later), *perturbed* by martingale increments.

One may then hope to relate its behavior for *large times*, and hence the behavior for a *large number of iterations* of the algorithm, to the behavior for *large times* of the ODE. The latter involves the stable fixed points of the ODE, which are the solutions of $F(\theta) = 0$ such that the vector field locally points toward this root; a sufficient condition is that locally $(x - \theta) \cdot F(x) < 0$.

The study of the stochastic algorithm will thus use tools related to ordinary differential equations and other tools related to martingales, such as the following classic result, see [5, Thm 4.3.3 p. 138], e.g.

Theorem 9.1 (Convergence of L^2-bounded martingales) *Let* $(M_n)_{n\in\mathbb{N}}$ *be a* square integrable martingale *for a filtration* $(\mathscr{F}_n)_{n\in\mathbb{N}}$. *Then*

$$\mathbb{E}\big(M_n^2\big) = \mathbb{E}\big(M_0^2\big) + \sum_{i=1}^{n} \mathbb{E}\big((M_i - M_{i-1})^2\big) \in [0,\infty]$$

and if $(M_n)_{n\in\mathbb{N}}$ *is* bounded *in* L^2, *i.e.*, *if*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\big(M_n^2\big) = \mathbb{E}\big(M_0^2\big) + \sum_{i\geq 1} \mathbb{E}\big((M_i - M_{i-1})^2\big) < \infty,$$

then the martingale converges *a.s. and in* L^2.

9.2.3 Long-Time Behavior of the Algorithm

The following result is given under tractable assumptions, *except* for the a.s. boundedness assumption on $(\theta_n)_{n\geq 0}$, which must be proved on a case-by-case basis and might be tough. One can try a comparison with the ODE, if its solution remains bounded, in a way similar to the proof.

Theorem 9.2 [5, Thm 5.3.5 p. 173] *Consider a stochastic algorithm* $(\theta_n)_{n\in\mathbb{N}}$ *written as in Definition* 9.1, *and its interpolation* $(X_t)_{t\in\mathbb{R}_+}$ *given in* (9.2). *Assume that* F *is* locally Lipschitz, *the ODE* (9.1) *with vector field* F *has a* flow of homeomorphisms $(\phi_t)_{t\in\mathbb{R}_+}$ (*see Theorem* 6.8, *e.g.*), *and one of the following*:

(a) *either*

$$\sup_{n\geq 1} \mathbb{E}\big(|U_n|^2\big) < \infty, \qquad \sum_{n\geq 1} \gamma_n^2 < \infty,$$

(b) *or, the sequence* $(\gamma_n)_{n\geq 1}$ *is non-increasing, and for some* $\Gamma > 0$ *and all* $\theta \in \mathbb{R}^d$ *and* $c > 0$,

$$\sup_{n\geq 1} \mathbb{E}\big(\mathrm{e}^{\theta\cdot U_n} \mid \mathscr{F}_{n-1}\big) \leq \mathrm{e}^{\frac{\Gamma}{2}|\theta|^2}, \qquad \sum_{n\geq 1} \mathrm{e}^{-c/\gamma_n} < \infty.$$

Assume moreover that

$$\sup_{n\geq 0} \theta_n < \infty, \quad \text{a.s.}$$

Then

$$\sup_{0\leq h\leq T} \big|X_{t+h} - \phi_h(X_t)\big| \xrightarrow[t\to\infty]{\text{a.s.}} 0, \quad \forall T \in \mathbb{R}_+.$$

Proof Let the "inverse map" of $n \mapsto \tau_n$ (see (9.2)) be defined by

$$m : t \in \mathbb{R}_+ \mapsto m(t) = \sup\{n \in \mathbb{N} : \tau_n \le t\} \in \mathbb{N}.$$

Observe that (9.2) can be written as

$$X_t = X_0 + \int_0^t \big(F(X_{\tau_{m(s)}}) + U_{m(s)+1}\big)\, ds. \tag{9.3}$$

Let us first prove that

$$\Delta(t,T) := \sup_{0\le h\le T} \left| \int_t^{t+h} U_{m(s)+1}\, ds \right| \xrightarrow[t\to\infty]{\text{a.s.}} 0, \quad \forall T \in \mathbb{R}_+. \tag{9.4}$$

Let $t \in \mathbb{R}_+$ and $T' > T \ge 0$. Since $(\gamma_n)_{n\ge1}$ goes to 0, if t is large enough then

$$\tau_{m(t)} \le t < t + T < \tau_{m(\tau_{m(t)}+T')}$$

so that

$$\Delta(t,T) \le 2 \max_{m(t)\le k\le m(\tau_{m(t)}+T')} \left| \sum_{i=m(t)+1}^{k} \gamma_i U_i \right|.$$

Thus, proof of (9.4) will follow from proof of

$$\max_{n\le k\le m(\tau_n+T)} \left| \sum_{i=n+1}^{k} \gamma_i U_i \right| \xrightarrow[n\to\infty]{\text{a.s.}} 0, \quad \forall T \in \mathbb{R}_+, \tag{9.5}$$

which could replace the assumption "(a) or (b)" in the statement. Under Hypothesis (a), (9.5) follows from Theorem 9.1 applied to the martingale $(M_n)_{n\in\mathbb{N}}$ with $M_n = \sum_{i=1}^n \gamma_i U_i$. Problem 9.2 will show how to deduce it under Hypothesis (b); see also the proof in [5].

Now, the conclusion will follow from (9.4) by a purely *deterministic* study of the ODE. For t and h in $\mathbb{R}_+$, the definition of the flow and (9.3) yield

$$\phi_h(X_t) = X_t + \int_0^h F\big(\phi_s(X_t)\big)\, ds,$$

$$X_{t+h} = X_t + \int_0^h F(X_{t+s})\, ds + A_t(h) + B_t(h),$$

$$\text{with } A_t(h) = \int_t^{t+h} \big(F(X_{\tau_{m(s)}}) - F(X_s)\big)\, ds,\ B_t(h) = \int_t^{t+h} U_{m(s)+1}\, ds.$$

For the Lipschitz constant $L = L_C$ for F on a compact neighborhood C of the sample path $(X_t)_{t\in\mathbb{R}_+}$ (which is bounded, a.s., since $(\theta_n)_{n\ge0}$ is bounded by hypothesis)

it holds that

$$\left|X_{t+h} - \phi(X_t)\right| \le \left|A_t(h)\right| + \left|B_t(h)\right| + L\int_0^h \left|X_{t+s} - \phi_s(X_t)\right| ds$$

and Gronwall's lemma yields

$$\sup_{0\le h\le T} \left|X_{t+h} - \phi(X_t)\right| \le \mathrm{e}^{LT} \sup_{0\le h\le T} \left(\left|A_t(h)\right| + \left|B_t(h)\right|\right). \tag{9.6}$$

Moreover

$$\sup_{0\le h\le T} \left|A_t(h)\right| \le LT \sup_{t\le s\le t+T} \left|X_s - X_{\tau_{m(s)}}\right|$$

and (9.3) implies for $K = \sup_C |F|$ that

$$|X_s - X_{\tau_{m(s)}}| = \left|\int_{\tau_{m(s)}}^s \left(F(X_{\tau_{m(u)}}) + U_{m(u)+1}\right) du\right| \le K\gamma_{m(s)+1} + \left|\int_{\tau_{m(s)}}^s U_{m(u)+1}\, du\right|$$

and for t large enough and $t \le s \le t+T$ it holds that $\gamma_{m(s)+1} \le 1$ and

$$\left|\int_{\tau_{m(s)}}^s U_{m(u)+1}\, du\right| \le \left|\int_{t-1}^{\tau_{m(s)}} U_{m(u)+1}\, du\right| + \left|\int_{t-1}^s U_{m(u)+1}\, du\right| \le 2\Delta(t-1, T+1)$$

and hence

$$\sup_{0\le h\le T} \left|A_t(h)\right| \le KLT \sup_{t\le s\le t+T} \gamma_{m(s)+1} + 2LT\Delta(t-1, T+1)$$

which, jointly with (9.6) and $\sup_{0\le h\le T} |B_t(h)| = \Delta(t, T)$ and (9.4), allows to conclude. □

The preceding result relates the large time behaviors of the interpolation $(X_t)_{t\in\mathbb{R}_+}$ of $(\theta_n)_{n\in\mathbb{N}}$ and of the flow $(\phi_t)_{t\in\mathbb{R}_+}$. The following result completes Theorem 9.2 in a remarkable way, in the sense that it provides crucial information on curves which satisfy its conclusion.

One says that $V \in C^1(\mathbb{R}^d, \mathbb{R})$ is a *strict Lyapunov function* for F if

$$\nabla V(x)\cdot F(x) \le 0, \qquad \nabla V(x)\cdot F(x) = 0 \Leftrightarrow F(x) = 0, \quad \forall x \in \mathbb{R}^d.$$

If there exists such a function and if the ODE (9.1) has a flow of homeomorphisms $(\phi_t)_{t\in\mathbb{R}_+}$, then

$$\frac{d}{dt} V\left(\phi_t(x)\right) = \nabla V\left(\phi_t(x)\right)\cdot F\left(\phi_t(x)\right) \le 0, \qquad \frac{d}{dt} V\left(\phi_t(x)\right) = 0 \Leftrightarrow F\left(\phi_t(x)\right) = 0,$$

so that V *is non-increasing* along the flow, and *strictly* so except if the starting point x is a fixed point for the ODE, i.e., if $F(x) = 0$: assume that for some $t > 0$ it holds that $\phi_t(x) = y$ and $F(y) = 0$, then $\phi_t(x) = y = \phi_t(y)$ and, by injectivity, $x = y$.

Theorem 9.3 [5, Thm 5.3.10 p. 179] *Assume that the ODE* (9.1) *with vector field* F *has a flow of homeomorphisms* $(\phi_t)_{t\in\mathbb{R}_+}$ *(see Theorem* 6.8*, e.g.), that* F *has a strict Lyapunov function* V*, and that the fixed-point set* $\mathscr{E} = \{x \in \mathbb{R}^d : F(x) = 0\}$ *of the ODE is constituted of isolated points. If* $(y_t)_{t\in\mathbb{R}_+}$ *is a continuous bounded curve satisfying*

$$\sup_{0\le h\le T} \big|y_{t+h} - \phi_h(y_t)\big| \xrightarrow[t\to\infty]{} 0, \quad \forall T \in \mathbb{R}_+,$$

then $\lim_{t\to\infty} y_t$ *exists and belongs to* $\mathscr{E}$*.*

Proof (Elements of proof (see [5])) Let the set of accumulation points of (y_t) be

$$L = L\big((y_t)_{t\in\mathbb{R}_+}\big) = \bigcap_{t\in\mathbb{R}_+} \overline{\{(y_{t+h})_{h\in\mathbb{R}_+}\}}.$$

We start by checking that $L \cap \mathscr{E} \neq \emptyset$. For that, for any neighborhood U of $\mathscr{E}$, it holds that $\nabla V(x) \cdot F(x) \le -\delta(U)$ for some $\delta(U) > 0$ on the compact set $L \cap U^c$ (see Lemma 9.2), and it can be proved (by a rather technical argument) that it is not true that $L = L \cap U^c$, since then $\inf_L V = -\infty$ which is impossible, and we conclude from the fact that L is closed. We then prove that $L \subset \mathscr{E}$. For that we prove (by an analogous more technical argument) that if there were some $p_0 \in L \cap \mathscr{E}^c$ then again $\inf_L V = -\infty$ which is impossible. At last, the connectedness of L (see Lemma 9.2 below) and the fact that $\mathscr{E}$ is constituted of isolated points shows that L must contain only one point, and then the boundedness assumption on $(y_t)_{t\in\mathbb{R}_+}$ implies that it converges to its unique accumulation point. □

In the proofs of the two preceding theorems we have used the following lemma.

Lemma 9.2 [5, Lemma 5.3.7 p. 177] *Assume that the ODE* (9.1) *with vector field* F *has a flow of homeomorphisms* $(\phi_t)_{t\in\mathbb{R}_+}$ *(see Theorem* 6.8*, e.g.). If* $(y_t)_{t\in\mathbb{R}_+}$ *is a continuous bounded curve satisfying*

$$\big|y_{t+h} - \phi_h(y_t)\big| \xrightarrow[t\to\infty]{} 0, \quad \forall h \in \mathbb{R}_+,$$

then the set of its accumulation points

$$L = L\big((y_t)_{t\in\mathbb{R}_+}\big) = \bigcap_{t\in\mathbb{R}_+} \overline{\{(y_{t+h})_{h\in\mathbb{R}_+}\}}$$

is non-empty, compact, connected, and flow invariant: $\phi_t(L) = L$*, all* $t \ge 0$*.*

Proof Since $(y_t)_{t\in\mathbb{R}_+}$ is bounded, the closures $\overline{\{(y_{t+h})_{h\in\mathbb{R}_+}\}}$ are decreasing compact sets, therefore their intersection L is non-empty and compact. To prove it is connected, assume by contradiction that U and V are two disjoint open sets such that $L \subset U \cup V$ and $L \cap U \neq \emptyset$ and $L \cap V \neq \emptyset$. The curve $(y_t)_{t\in\mathbb{R}_+}$ would go an infinite number of times from U to V, and by continuity would pass each time through

the closed non-empty set $(U \cup V)^c$, and the boundedness of $(y_t)_{t\in\mathbb{R}_+}$ would imply that $L \cap (U \cup V)^c \neq \emptyset$, which is a contradiction.

We now show that $L = \phi_h(L)$.

Let $y \in L$ and $h \in \mathbb{R}_+$. There is $(t_n)_{n\geq 1}$ increasing to infinity such that $\lim_{n\to\infty} y_{t_n} = y$. The continuity of ϕ_h and the asymptotic assumption on $(y_t)_{t\in\mathbb{R}_+}$ yield

$$\phi_h(y) = \lim_{n\to\infty} \phi_h(y_{t_n}) = \lim_{n\to\infty} y_{t_n+h} \in L,$$

and hence $\phi_h(L) \subset L$.

Reciprocally, for large n the values y_{t_n-h} are defined and bounded and thus have an accumulation point $z \in L$, by assumption

$$\left|y_{t_n} - \phi_h(y_{t_n-h})\right| = \left|y_{t_n-h+h} - \phi_h(y_{t_n-h})\right| \xrightarrow[n\to\infty]{} 0.$$

By continuity $y = \phi_h(z)$ and hence $L \subset \phi_h(L)$. □

9.3 Variance Reduction for Monte Carlo Methods

Variance reduction can be obtained using stochastic algorithms that use the same random variables as the Monte Carlo method, and it is tempting implement both algorithms at the same time with the same samples. This part is largely inspired by Arouna [3].

9.3.1 Searching for an Importance Sampling

We have previously given an example of the use of stochastic algorithms for a *variance reduction* method, in the framework of a Monte Carlo computation. After discretization, the problem can be reduced to the following.

One wishes to find an *approximation* for an expectation $\mathbb{E}(f(X))$ that one does *not know* how to compute exactly. In this, X is a d-dimensional random vector that one *knows* how to simulate, and the function $f : \mathbb{R}^d \to \mathbb{R}$ is *known*. Specifically, X was a standard normal vector.

The problem was to find a *simulable* r.v. Z such that $\mathbb{E}(f(X)) = \mathbb{E}(Z)$ and the *variance reduction*

$$\mathrm{Var}\big(f(X)\big) - \mathrm{Var}(Z) = \mathbb{E}\big(f(X)^2\big) - \mathbb{E}\big(Z^2\big)$$

is relatively large compared to $\mathrm{Var}(f(X))$, or at least positive.

It can be an arduous task to find good candidates for Z, and the corresponding general optimization problem is *trivial* in theory and *ill-posed* in practice, since the optimum is found tautologically for the deterministic $Z = \mathbb{E}(f(X))$, which surely has variance zero, but cannot be simulated since it requires to know what is precisely to be approximated by the Monte Carlo method.

It is reasonable then to limit the search within an adequate family of r.v.'s, possibly parametrized. The family can be given by a family of densities w.r.t. a reference measure, or correspond to some other changes of law.

For instance, let us re-examine the situation presented in Sect. 3.6.2. For X having a density p_X on $\mathbb{R}^d$, and Y a density p_Y on $\mathbb{R}^d$ such that $p_Y(x) > 0$ if $p_X(x) > 0$,

$$\mathbb{E}\big(f(X)\big) = \int f(x)p_X(x)\,dx = \int f(x)\frac{p_X(x)}{p_Y(x)}p_Y(x)\,dx = \mathbb{E}\left(f(Y)\frac{p_X(Y)}{p_Y(Y)}\right)$$

(introducing the Radon–Nikodym derivative p_X/p_Y of the law of X w.r.t. the law of Y), and the problem is to find p_Y such that

$$\mathbb{E}\left(\left(f(Y)\frac{p_X(Y)}{p_Y(Y)}\right)^2\right) = \int\left(f(x)\frac{p_X(x)}{p_Y(x)}\right)^2 p_Y(x)\,dx = \int \frac{f(x)^2 p_X(x)^2}{p_Y(x)}\,dx$$

is small. The minimization problem is again trivial in theory and ill-posed in practice: for $f \geq 0$, the Cauchy–Schwarz inequality and $\int p_Y = 1$ yield

$$\left(\int f(x)p_X(x)\,dx\right)^2 = \left(\int \frac{f(x)p_X(x)}{\sqrt{p_Y(x)}}\sqrt{p_Y(x)}\,dx\right)^2 \leq \int \frac{f(x)^2 p_X(x)^2}{p_Y(x)}\,dx.$$

We have equality if and only if p_Y is proportional to fp_X. Then again the variance is zero, but

$$\int p_Y = 1 \Rightarrow p_Y = \frac{fp_X}{\int fp_X} = \frac{fp_X}{\mathbb{E}(f(X))},$$

so that to simulate according to p_Y one must know $\mathbb{E}(f(X))$. Moreover the function $f\frac{p_X}{p_Y}$ which must be applied to the draws is constant equal to $\mathbb{E}(f(X))$ (hence the vanishing variance), necessitating again to *know* precisely the quantity one tries to *approximate*.

For f taking values in $\mathbb{R}$, this theoretical minimum is reached for $|f|p_X/\int |f|p_X$ and this leads us to search for some p_Y close to this quotient among a family of densities according to which it is easy to simulate and for which p_X/p_Y can be easily computed. This is called an *importance sampling* technique.

For $p_X > 0$, a simple special idea is to take $Y = X + \theta$ for a parameter $\theta \in \mathbb{R}^d$, which has density $p_Y : y \mapsto p_X(y-\theta)$, so that

$$\mathbb{E}\big(f(X)\big) = \mathbb{E}\left(f(Y)\frac{p_X(Y)}{p_X(Y-\theta)}\right) = \mathbb{E}\left(f(X+\theta)\frac{p_X(X+\theta)}{p_X(X)}\right)$$

which can be found directly using the translation invariance

$$\begin{aligned}\mathbb{E}\big(f(X)\big) &= \int f(x)p_X(x)\,dx \\ &= \int f(x+\theta)p_X(x+\theta)\,dx \\ &= \int f(x+\theta)\frac{p_X(x+\theta)}{p_X(x)}p_X(x)\,dx.\end{aligned}$$

9.3.2 Variance Reduction and Stochastic Algorithms

In greater generality, assume that it holds that

$$\mathbb{E}\big(f(X)\big) = \mathbb{E}\big(g(\theta, X)\big), \quad g : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}, \quad \theta \in \mathbb{R}^m.$$

Using a sequence $(X_n)_{n\geq 1}$ of i.i.d. r.v.'s with the same law as X, the Strong Law of Large Numbers

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(\theta, X_n) = \mathbb{E}\big(f(X)\big), \quad \text{a.s.,} \tag{9.7}$$

allows to estimate $\mathbb{E}(f(X))$ without bias. Moreover, if $g(\theta, X)$ is in L^2 and $\sigma_\theta^2 := \mathrm{Var}(g(\theta, X))$ then the Central Limit Theorem yields

$$\lim_{N\to\infty} \sqrt{N}\left(\frac{1}{N} \sum_{n=1}^{N} g(\theta, X_n) - \mathbb{E}\big(f(X)\big)\right) = \mathcal{N}\big(0, \sigma_\theta^2\big), \quad \text{in law,} \tag{9.8}$$

which in theory provides asymptotic confidence intervals for the estimation.

In practice, σ_θ^2 should be unknown as well as $\mathbb{E}(f(X))$, but elementary computations and the Strong Law of Large Numbers

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{j=1}^{N} \left(g(\theta, X_j) - \frac{1}{N} \sum_{n=1}^{N} g(\theta, X_n)\right)^2 = \sigma_\theta^2, \quad \text{a.s.,}$$

allow to estimate without bias σ_θ^2 simultaneously to $\mathbb{E}(f(X))$.

Heuristically, the estimated value of σ_θ^2 can be used for the confidence intervals for (9.7) obtained using (9.8). Obtaining rigorous confidence intervals for this estimated value would necessitate knowledge of fourth-order moments, which are unavailable and should be estimated, and so on.

Finding θ^* which equivalently minimizes the variance or the second moment of $g(\theta, X)$, respectively given by

$$\sigma_\theta^2 = \mathbb{E}\big(g(\theta, X)^2\big) - \mathbb{E}\big(f(X)\big)^2, \quad \mathbb{E}\big(g(\theta, X)^2\big),$$

allows to make the choice $\theta = \theta^*$ in order to have asymptotic confidence intervals for $\mathbb{E}(f(X))$ which are *minimal* in this framework. If this is difficult to attain, the goal is at least to find some θ such that $\mathbb{E}(g(\theta, X)^2) < \mathbb{E}(f(X)^2)$, and better $\mathbb{E}(g(\theta, X)^2) \ll \mathbb{E}(f(X)^2)$.

Deterministic methods, such as gradient descent, involve

$$F(\theta) = -\nabla \mathbb{E}\big(g(\theta, X)^2\big)$$

and there are reasonable assumptions under which differentiation and expectation can be interchanged so that

$$F(\theta) = -\mathbb{E}\big(\nabla_\theta g^2(\theta, X)\big).$$

The computation of F is expected to be at least as difficult as the computation of $\mathbb{E}(f(X)) = \mathbb{E}(g(\theta, X))$, so that deterministic methods cannot be implemented in practice.

On the contrary, the *stochastic algorithm* of Lemma 9.1, for H such that

$$F(\theta) = -\mathbb{E}\big(\nabla_\theta g^2(\theta, X)\big) = \mathbb{E}\big(H(\theta, X)\big)$$

and some i.i.d. r.v.'s X_n with the same law as X, only necessitates the computation of the $H(\theta_n, X_{n+1})$ and *not* of their expectations, the price to pay being the introduction of the "noise" due to the martingale increments U_{n+1}. There are several possible choices for H, one of which being $-\nabla_\theta g^2$. In theory one can write the algorithm as in Definition 9.1, but in practice it is necessary to check the assumptions using the known H and law of X.

The "optimal" Monte Carlo method involves the computation of $g(\theta^*, X_n)$ for some "other" $(X_n)_{n\geq 1}$ which also are i.i.d. with the same law as X. Moreover, the computations of $H(\theta, x)$ and $g(\theta, x)$ will be seen to be rather similar.

It is thus very tempting to implement the stochastic algorithm *simultaneously* to the Monte Carlo method. A theoretical justification for using this adaptive Monte Carlo method and the "optimal" confidence intervals would require to prove that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(\theta_{n-1}, X_n) = \mathbb{E}\big(f(X)\big), \quad \text{a.s.},$$

$$\lim_{N\to\infty} \sqrt{N}\left(\frac{1}{N} \sum_{n=1}^{N} g(\theta_{n-1}, X_n) - \mathbb{E}\big(f(X)\big)\right) = \mathscr{N}\big(0, \tau^2\big), \quad \text{in law},$$

for some τ which should be bounded in terms of the data and not too much larger than σ_{θ^*}.

The r.v.'s $(\theta_n)_{n\geq 0}$ given by the stochastic algorithm are not independent, and we are completely outside the framework for the classical law of large numbers and central limit theorems. Theoretical justifications are available, but are beyond the scope of this book; an example will be given in Problem 9.1.

It is also possible to invoke and implement *numerical experimentation*, and justify the methods by its good results.

The Gaussian Case

We now extend the Gaussian situation in the second part of Sect. 3.6.2 to the multidimensional setting. Consider the special case in which X has standard normal law $\mathscr{N}(0, I_d)$. Then

$$g(\theta, x) = f(x+\theta)\, \mathrm{e}^{-\frac{1}{2}|x+\theta|^2 + \frac{1}{2}|x|^2} = f(x+\theta)\, \mathrm{e}^{-\theta\cdot x - \frac{1}{2}|\theta|^2}$$

(similar to the Girsanov transform). The task is to minimize

$$\mathbb{E}\big(g(\theta, X)^2\big) = \mathbb{E}\big(f(X+\theta)^2 \,\mathrm{e}^{-2\theta\cdot X - |\theta|^2}\big) = \mathbb{E}\big(f(X)^2 \,\mathrm{e}^{-\theta\cdot X + \frac{1}{2}|\theta|^2}\big)$$

in which the second equality follows for the simple change of variables

$$\int f(x+\theta)^2 \,\mathrm{e}^{-2\theta\cdot x - |\theta|^2} \,\mathrm{e}^{-\frac{1}{2}|x|^2}\, dx = \int f(y)^2 \,\mathrm{e}^{-\theta\cdot y + \frac{1}{2}|\theta|^2} \,\mathrm{e}^{-\frac{1}{2}|y|^2}\, dy. \tag{9.9}$$

It can be shown that if $\mathbb{E}(f(X)^{2+\varepsilon}) < \infty$ for some $\varepsilon > 0$ the dominated convergence theorem allows to prove the validity of the interchange

$$\nabla\mathbb{E}\big(g(\theta, X)^2\big) = \mathbb{E}\big(\nabla_\theta g^2(\theta, X)\big) = \mathbb{E}\big(f(X)^2(\theta - X)\,\mathrm{e}^{-\theta\cdot X + \frac{1}{2}|\theta|^2}\big).$$

Moreover, as seen in Proposition 3.3, $\theta \mapsto (\theta - X)\,\mathrm{e}^{-\theta\cdot X + \frac{1}{2}|\theta|^2}$ is increasing in each coordinate, and hence $\theta \mapsto \mathbb{E}(g(\theta, X)^2)$ is a convex function. If $\mathbb{P}(f(X) \neq 0) > 0$ then it is strictly convex and can easily be proved to go to infinity (in norm) at infinity, so that $\min_\theta \mathbb{E}(g(\theta, X)^2)$ is reached at the unique root θ^* of $\mathbb{E}(\nabla_\theta g(\theta, X)^2) = 0$.

Let us assume that we are in this "nice" situation. The computation of $F(\theta) = -\nabla\mathbb{E}(g(\theta, X)^2)$ is at least as difficult as that of $\mathbb{E}(g(\theta, X)^2)$ which was at the start of the study, but a *stochastic algorithm* corresponding to Lemma 9.1 can be devised, with H such that $F(\theta) = \mathbb{E}(H(\theta, X))$ and $(X_n)_{n\geq 1}$ constituted of i.i.d. $\mathcal{N}(0, 1)$ r.v.'s.

To choose H, we may use the fact that, as for (9.9),

$$\begin{aligned} F(\theta) = &-\mathbb{E}\big(f(X)^2(\theta - X)\,\mathrm{e}^{-\theta\cdot X + \frac{1}{2}|\theta|^2}\big) = \mathbb{E}\big(f(X+\theta)^2 X\,\mathrm{e}^{-2\theta\cdot X - |\theta|^2}\big) \\ &:= \mathbb{E}\big(X g(\theta, X)^2\big) \end{aligned}$$

yielding two choices, among which $H(\theta, x) = x g(\theta, x)^2$ for which, in a variance reduction method implemented *simultaneously* to the Monte Carlo method, $g(\theta_n, X_{n+1})$ and then $H(\theta_n, X_{n+1}) = X_{n+1} g(\theta_n, X_{n+1})^2$ would be computed in order to obtain θ_{n+1}.

Moreover, it can be checked that the $\theta \mapsto \mathbb{E}(g(\theta, X)^2)$ is a *strict Lyapunov function* in the sense of Theorem 9.3.

Actually, this algorithm is unstable as presented, and must be stabilized by further *projection operations*, but we shall not detail this very technical issue.

9.4 Problems

9.1 (Adaptive Monte Carlo Method) The goal is to obtain a good approximation by a Monte Carlo method of the quantity $\mathbb{E}(\phi(X))$, where X is a simulable random variable (r.v.), $\phi : \mathbb{R} \to \mathbb{R}$ is a computable function, and $\mathbb{E}(\phi(X)^2) < \infty$.

Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. r.v.'s with the same law as X, and $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ for $n \geq 0$. We assume that, in order to implement a variance reduction method, we have found a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfying:

(a) For every $\theta \in \mathbb{R}$,

$$\mathbb{E}\big(g(\theta, X)\big) = \mathbb{E}\big(\phi(X)\big).$$

(b) For every $\theta \in \mathbb{R}$,

$$m_2(\theta) := \mathbb{E}\big(g(\theta, X)^2\big) < \infty,$$

there exists $\theta^* \in \mathbb{R}$ such that

$$m_2\big(\theta^*\big) = \min_{\theta \in \mathbb{R}} m_2(\theta) < \mathbb{E}\big(\phi(X)^2\big),$$

the function $\theta \mapsto m_2(\theta)$ is continuous at θ^*, and

$$\sigma^2 := \mathrm{Var}\big(g\big(\theta^*, X\big)\big) = \min_{\theta \in \mathbb{R}} \mathrm{Var}\big(g(\theta, X)\big) > 0.$$

(c) There exists an $(\mathscr{F}_n)_{n\geq 0}$-adapted simulable sequence $(\theta_n)_{n\geq 0}$ converging to θ^*, a.s. (For instance given by a stochastic algorithm involving $(X_n)_{n\geq 1}$.)

We wish to prove, under reasonable assumptions, that

$$\frac{1}{n}\sum_{k=1}^{n} g(\theta_{k-1}, X_k) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}\big(\phi(X)\big), \tag{9.10}$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} g(\theta_{k-1}, X_k) - \mathbb{E}\big(\phi(X)\big)\right) \xrightarrow[n\to\infty]{\text{in law}} \mathscr{N}\big(0, \sigma^2\big), \tag{9.11}$$

and moreover to estimate σ^2.

If we succeed, we will be able to estimate $\mathbb{E}(\phi(X))$ with an asymptotic precision as good as the one for the Monte Carlo method with minimal variance among this family, without having to approximate first the optimal parameter θ^*, all this while estimating the minimal variance so as to be able to provide asymptotic confidence intervals. This constitutes an adaptive Monte Carlo method.

We recall the following classic result. Let $(M_n)_{n\geq 0}$ be a square-integrable $(\mathscr{F}_n)_{n\geq 0}$-martingale, and $\langle M\rangle_n$ for $n \geq 0$ be defined by $\langle M\rangle_0 = 0$ and $\langle M\rangle_n - \langle M\rangle_{n-1} = \mathbb{E}((M_n - M_{n-1})^2 \,|\, \mathscr{F}_{n-1})$. The following convergence results hold:

(1) If $\lim_{n\to\infty} \langle M\rangle_n = \infty$ a.s., then $\lim_{n\to\infty} \frac{M_n}{\langle M\rangle_n} = 0$ a.s. ([5, Thm 4.3.9 p. 141]);

(2) If moreover $\sup_{n\geq 1} |M_n - M_{n-1}|$ is bounded, then $\lim_{n\to\infty} \frac{M_n}{\sqrt{\langle M\rangle_n}} = \mathscr{N}(0, 1)$ in law.

1. (a) Prove that a square-integrable martingale $(M_n)_{n\geq 0}$ is defined by

$$M_0 = 0, \qquad M_n = \sum_{k=1}^{n} \big(g(\theta_{k-1}, X_k) - \mathbb{E}\big(\phi(X)\big)\big), \quad n \geq 1.$$

(b) Compute $\langle M\rangle_n$ for $n \geq 0$, and find a simple equivalent for large n.

(c) Prove the convergence result (9.10) for the adaptive Monte Carlo method.

2. From now on, $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is assumed to be uniformly bounded.
   Prove the convergence result (9.11) yielding the rate of convergence.
3. From now on, $\theta \mapsto m_4(\theta) := \mathbb{E}(g(\theta, X)^4) < \infty$ is assumed continuous at θ^*.

   (a) Prove that a square-integrable martingale $(\hat{M}_n)_{n\geq 0}$ is defined by

$$\hat{M}_0 = 0, \qquad \hat{M}_n = \sum_{k=1}^{n} \big(g(\theta_{k-1}, X_k)^2 - m_2(\theta_{k-1})\big), \quad n \geq 1.$$

   (b) Compute $\langle \hat{M} \rangle_n$ for $n \geq 0$, and find a simple equivalent for large n.
   (c) Suggest an estimator for σ^2 and prove it is convergent, a.s.

9.2 (Hypothesis (b) for Theorem 9.2) The goal is to prove (9.5) when Hypothesis (b) is true, but not necessarily Hypothesis (a). We assume the other hypotheses in this theorem are true, and use the notation there and in its proof, with $d = 1$ in order to simplify notation. The following two classic results are *admitted*.

Claim (Doob maximal inequality) *If $(X_k)_{k\in\mathbb{N}}$ is a non-negative sub-martingale then, for all $N \in \mathbb{N}$ and $a \geq 0$,*

$$a\mathbb{P}\Big(\max_{0\leq k\leq N} X_k \geq a\Big) \leq \mathbb{E}(X_N).$$

Claim (Doob decomposition) *If $(X_k)_{k\in\mathbb{N}}$ is a super-martingale then there exists a martingale $(M_k)_{k\in\mathbb{N}}$ and an increasing predictable process $(A_k)_{k\in\mathbb{N}}$ such that $A_0 = 0$ and*

$$X_k = M_k - A_k, \quad k \in \mathbb{N}.$$

1. Let $\theta \in \mathbb{R}$ and $n \in \mathbb{N}$ and, with $Z_0 = 1$ by convention,

$$Z_k = \exp\left(\theta \sum_{i=n+1}^{n+k} \gamma_i U_i - \frac{\Gamma}{2}\theta^2 \sum_{i=n+1}^{n+k} \gamma_i^2\right), \quad k \geq 0.$$

   (a) Prove that $(Z_k)_{k\in\mathbb{N}}$ is a non-negative super-martingale.
   (b) Deduce from this for all $N \in \mathbb{N}$ and $a \geq 0$ that

$$a\mathbb{P}\Big(\max_{0\leq k\leq N} Z_k \geq a\Big) \leq \mathbb{E}(Z_0) = 1.$$

2. Prove for every $\theta \in \mathbb{R}$ and integer $n \leq N$ and $b \in \mathbb{R}$ that

$$\mathbb{P}\left(\max_{n\leq k\leq N} \theta \sum_{i=n+1}^{k} \gamma_i U_i \geq b\right) \leq \exp\left(-b + \frac{\Gamma}{2}\theta^2 \sum_{i=n+1}^{N} \gamma_i^2\right).$$

3. Prove for $e = \pm 1$ and every integer $n \le N$ and $\alpha > 0$ that

$$\mathbb{P}\left(\max_{n \le k \le N} e \sum_{i=n+1}^{k} \gamma_i U_i \ge \alpha \right) \le \exp\left(-\frac{\alpha^2}{2\Gamma \sum_{i=n+1}^{N} \gamma_i^2} \right)$$

and deduce from this that

$$\mathbb{P}\left(\max_{n \le k \le N} \left| \sum_{i=n+1}^{k} \gamma_i U_i \right| \ge \alpha \right) \le 2 \exp\left(-\frac{\alpha^2}{2\Gamma \sum_{i=n+1}^{N} \gamma_i^2} \right).$$

4. Prove that (9.5) is true.

9.3 (Principal Eigenvector and Oja's Algorithm) Let be given vectors $a_k = (a_k^i)_{1 \le i \le d} \in \mathbb{R}^d$ for $1 \le k \le N$, and the symmetric positive matrix

$$C = (C_{ij})_{1 \le i,j \le d}, \quad C_{ij} = \frac{1}{N} \sum_{k=1}^{N} a_k^i a_k^j.$$

We **assume** that the eigenvalues satisfy $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$, and denote by $(v_1, \ldots, v_d)$ an orthonormal basis of corresponding eigenvectors.

For many applications, it is important to obtain approximations for these eigenvalues and eigenvectors, staring with λ_1 (the largest) and v_1 (defined up to sign), which are called the principal eigenvalue and eigenvector. We wish to provide elements of the proof for the following result.

For $x = (x^i)_{1 \le i \le d}$ and $y = (y^i)_{1 \le i \le d}$ in $\mathbb{R}^d$, let $\langle x, y \rangle = \sum_{i=1}^{d} x^i y^i$ denote the scalar product and $\|x\|^2 = \langle x, x \rangle$.

Theorem 9.4 *Let θ_0 be an $\mathbb{R}^d$-valued r.v., $(K_n)_{n \ge 1}$ be a sequence of uniform r.v.'s in $\{1, 2, \ldots, N\}$, all these r.v.'s being independent, and $X_n = a_{K_n}$ (i.e., $X_n = a_k$ if $K_n = k$). Let $(\gamma_n)_{n \ge 1}$ be a deterministic decreasing sequence of positive real numbers, and $(\theta_n)_{n \ge 0}$ be defined by*

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \big(\langle X_{n+1}, \theta_n \rangle X_{n+1} - \langle X_{n+1}, \theta_n \rangle^2 \theta_n \big), \quad n \ge 0.$$

Let $B = \sup_{1 \le i \le N} \|a_i\|^2 < \infty$. Assume that

$$0 < \|\theta_0\|^2 \le 5, \qquad \gamma_1 \le \frac{1}{2B}, \qquad \sum_{n \ge 1} \gamma_n = \infty, \qquad \sum_{n \ge 1} \gamma_n^2 < \infty.$$

Then $(\theta_n)_{n \ge 0}$ converges either to v_1 or to $-v_1$, a.s.

We shall need the next result, which will be proved in the problem.

Theorem 9.5 *Let $(V_n)_{n\geq 0}$, $(\eta_n)_{n\geq 0}$ and $(\varepsilon_n)_{n\geq 0}$ be sequences of r.v.'s and be adapted to a filtration $(\mathscr{F}_n)_{n\geq 0}$. Assume that*

$$V_n \geq 0, \quad \eta_n \geq 0, \quad \mathbb{E}(V_{n+1} \,|\, \mathscr{F}_n) \leq V_n - \eta_n + \varepsilon_n, \quad \sum_{n\geq 0} |\varepsilon_n| < \infty.$$

Then $(V_n)_{n\geq 0}$ converges to a finite r.v. V_∞ and $\sum_{n\geq 0} \eta_n < \infty$, a.s.

Reminders and Results Which Are Admitted A sequence $(Y_n)_{n\geq 0}$ of $\mathbb{R}$-valued r.v.'s is a *super-martingale* for a filtration $(\mathscr{F}_n)_{n\geq 0}$ if it is adapted, integrable, and satisfies $\mathbb{E}(Y_{n+1} \,|\, \mathscr{F}_n) \leq Y_n$ for $n \geq 0$. Let $x^- := \max(-x, 0)$.

We recall Doob's convergence theorem: If $(Y_n)_{n\geq 0}$ is a super-martingale such that $\sup_n \mathbb{E}(Y_n^-) < \infty$, then $(Y_n)_{n\geq 0}$ converges a.s. to an integrable r.v. Y_∞. We recall also Theorem 9.1 on convergence of martingales which are bounded in L^2.

Part I—Proof of Theorem 9.5

1. Prove that $(Y_n)_{n\geq 0}$ defined by $Y_n = V_n - \sum_{k=0}^{n-1}(\varepsilon_k - \eta_k)$ is a super-martingale.
2. For $a \geq 0$ let $T_a = \inf\{n \geq 1 : \sum_{k=0}^{n}(\varepsilon_k - \eta_k) \geq a\}$. Prove that T_a is a stopping time, and that $(Y_n)_{n\geq 0}$ converges on $\{T_a = \infty\}$ to a finite limit, a.s.
   Hint: Introduce the sequence $(Y_{n\wedge T_a})_{n\geq 0}$, where $x \wedge y := \min(x, y)$.
3. On $\{T_a = \infty\}$, prove that $\sum_n \eta_n < \infty$, then that $(V_n)_{n\geq 0}$ converges, a.s.
4. Conclude that Theorem 9.5 is true.

Part II—Proof of Theorem 9.4, the Validity of Theorem 9.5 Being Assumed

1. Prove for $n \geq 0$ that

$$\|\theta_{n+1}\|^2 = \|\theta_n\|^2 + \big(2\gamma_{n+1}\big(1 - \|\theta_n\|^2\big) + \gamma_{n+1}^2\big[\|X_{n+1}\|^2 + \big(\|\theta_n\|^2 - 2\big)\langle X_{n+1}, \theta_n\rangle^2\big]\big)\langle X_{n+1}, \theta_n\rangle^2$$

   and $\|\theta_n\|^2 \leq 5$.
2. (a) Prove for $n \geq 0$ that if $\|\theta_n\|^2 \leq 1$ then $\|\theta_{n+1}\|^2 \geq \|\theta_n\|^2$.
   (b) Assuming that if $1 < \|\theta_n\| \leq 5$ then $\theta_{n+1} \neq 0$, prove that $\|\theta_n\|$ cannot converge to 0, a.s.
3. (a) Write the recurrence relation between θ_{n+1} and θ_n in terms of

$$C_{n+1} = \big(X_{n+1}^i X_{n+1}^j\big)_{1\leq i,j\leq d}.$$

   (b) Justify that it can be written as the stochastic algorithm

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\big(F(\theta_n) + U_{n+1}\big), \quad n \geq 0,$$

   where $F : \theta \in \mathbb{R}^d \mapsto F(\theta) = C\theta - \langle \theta, C\theta\rangle\theta \in \mathbb{R}^d$, and the U_{n+1} are uniformly bounded martingale increments.
4. Prove that $(M_n)_{n\geq 0}$ defined by $M_n = \sum_{k=1}^{n} \gamma_k U_k$ is a martingale which converges a.s. and in L^2, and thus that the series $\sum_{n\geq 0} \gamma_{n+1} U_{n+1}$ converges.

5. (a) Prove for $n \geq 0$ that, a.s.,

$$\|\theta_{n+1}\|^2 = \|\theta_n\|^2 + 2\gamma_{n+1}\big(1 - \|\theta_n\|^2\big)\langle \theta_n, C\theta_n\rangle + 2\langle \theta_n, \gamma_{n+1}U_{n+1}\rangle + O\big(\gamma_{n+1}^2\big),$$

$$\mathbb{E}\Big(\big(\|\theta_{n+1}\|^2 - 1\big)^2 \,|\, \mathscr{F}_n\Big) = \big(\|\theta_n\|^2 - 1\big)^2 - 4\gamma_{n+1}\big(1 - \|\theta_n\|^2\big)^2\langle \theta_n, C\theta_n\rangle + O\big(\gamma_{n+1}^2\big).$$

(b) Deduce from this that

$$\sum_{n\geq 0} \gamma_{n+1}\big(1 - \|\theta_n\|^2\big)^2\langle \theta_n, C\theta_n\rangle < \infty.$$

(c) Deduce from this that $\lim_{n\to\infty} \|\theta_n\|^2 = 1$.
(d) Deduce from this that

$$\sum_{n\geq 0} \gamma_{n+1}\big(1 - \|\theta_n\|^2\big)\langle \theta_n, C\theta_n\rangle \quad \text{converges, a.s.}$$

6. Consider the function

$$V : \theta \in \mathbb{R}^d - \{0\} \mapsto V(\theta) = \frac{\exp(\|\theta\|^2)}{\langle \theta, C\theta\rangle}.$$

(a) Prove that $\nabla V(\theta) = -\frac{2V(\theta)}{\langle \theta, C\theta\rangle} F(\theta)$.
(b) Prove that

$$\mathbb{E}\big(V(\theta_{n+1}) \,|\, \mathscr{F}_n\big) = V(\theta_n) - 2\gamma_{n+1}\frac{2V(\theta_n)}{\langle \theta_n, C\theta_n\rangle}\big\|F(\theta_n)\big\|^2 + O\big(\gamma_{n+1}^2\big), \quad n \geq 0.$$

(c) Prove that $(V(\theta_n))_{n\geq 0}$ and the series $\sum_{n\geq 0} \gamma_{n+1}\frac{2V(\theta_n)}{\langle \theta_n, C\theta_n\rangle}\|F(\theta_n)\|^2$ converge, a.s.
(d) Deduce from this that $\lim_{n\to\infty} F(\theta_n) = 0$, a.s.
(e) Prove that $(\theta_n)_{n\geq 0}$ converges, a.s., to an r.v. θ_∞ with values in $\{v_1, -v_1, \ldots, v_d, -v_d\}$.
(f) Prove that the only local minima of V are at v_1 and $-v_1$.

One can prove from this that θ_∞ takes values in $\{v_1, -v_1\}$, a.s., and hence conclude the proof of Theorem 9.4, but this would lead us too far.

Appendix
Solutions to Selected Problems

2.1 (A Poisson Distribution Simulation Method)

1. From $\mathbb{E}e^{\theta X} = \int_0^\infty e^{\theta x}\lambda e^{-\lambda x}\,dx$ we deduce that the Laplace transform of X is $\frac{\lambda}{\lambda-\theta}$ for all $\theta < \lambda$.
2. A simple computation shows that, for all $\theta < \lambda$,

$$\int_0^\infty e^{\theta x} p_N(x)\,dx = \left(\frac{\lambda}{\lambda-\theta}\right)^N = \mathbb{E}e^{\theta S_N}.$$

3. The result easily follows from

$$\mathbb{P}(M=k) = \mathbb{P}(M \le k) - \mathbb{P}(M \le k-1) = \int_\lambda^\infty \big(p_{k+1}(y) - p_k(y)\big)\,dy$$

and the above explicit formula for $p_k(y)$.

4. Let (U_i) be independent random variables uniformly distributed on $[0,1]$. Since U_i and e^{-X_i} have the same probability distribution, the preceding question shows that the smallest integer M such that $\prod_{i=1}^{M+1} U_i < e^{-\lambda}$ has Poisson distribution.

2.2 (Lyapunov Exponent of Linear Random Recursive Sequences)

1. The integrability at infinity is obvious since $\log|x| \le |x|$ for $|x| \ge 1$.
   To check the integrability around 0, we observe that, for all $z > 0$,

$$\int_0^z \log(x)\,dx = z\log(z) - z,$$

from which an integration by parts leads to

$$\int_0^z \big(\log(x)\big)^2\,dx = \big[\log(x)\big(x\log(x)-x\big)\big]_0^x - \int_0^z \big(\log(x)-1\big)\,dx$$
$$= z\big(\log(z)\big)^2 - 2z\log(z) + 2z.$$

The right-hand side has a finite limit when x tends to 0.

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1, © Springer-Verlag Berlin Heidelberg 2013

2. We have

$$\bar{X}_N^h(x) = x \prod_{p=0}^{N-1} \left(1 + b\sqrt{h}G_{p+1} + \left(a + \frac{b^2}{2}\right)h\right),$$

from which

$$\frac{1}{Nh} \log\left|\bar{X}_N^h(x)\right| = \frac{\log|x|}{Nh} + \frac{1}{Nh} \sum_{p=0}^{N-1} \log\left|1 + b\sqrt{h}G_{p+1} + \left(a + \frac{b^2}{2}\right)h\right|.$$

Note that the random variables in the right-hand side are independent and identically distributed. In addition, as $\log(|x|)$ is integrable around 0, one easily checks that the function

$$x \to \log\left|1 + b\sqrt{h}x + \left(a + \frac{b^2}{2}\right)h\right| \exp\left(-\frac{x^2}{2}\right)$$

is integrable over $\mathbb{R}$. Thus the Strong Law of Large Numbers applies:

$$\exists \bar{\lambda}^h \in \mathbb{R},\ \forall x \in \mathbb{R}^d - \{0\}, \quad \bar{\lambda}^h = \lim_{N \to +\infty} \frac{1}{Nh} \log\left|\bar{X}_N^h(x)\right| \quad \text{a.s.}$$

3. Observe that

$$\left(\sum \alpha_p\right)^2 = \sum \alpha_p^2 + 2 \sum_{p<q} \alpha_p \alpha_q.$$

Since the random variables $\log|1 + b\sqrt{h}G_{p+1} + (a + \frac{b^2}{2})h|$ are independent, it follows that

$$\exists C_h \in \mathbb{R}, \quad \frac{1}{N^2} \mathbb{E}\left[\log\left|\bar{X}_N^h(x)\right|\right]^2 < C_h \quad \text{for all } N \geq 1.$$

To obtain the desired equality for $\bar{\lambda}^h$ it now suffices to use the above representation of $\frac{1}{Nh}(\log|\bar{X}_N^h(x)| - \log|x|)$ as a sum of integrable identically distributed random variables, compute the expectation, and let N go to infinity.

4. Observe that

$$Z_1 = Z_2 + \mathbb{1}_C(Z_1 - Z_2) + \mathbb{1}_{\Omega - C}(Z_1 - Z_2)$$

for all random variables Z_1 and Z_2 and all events C. The desired result follows.

The Cauchy–Schwarz inequality implies that there exist $C > 0$ and $\alpha > 0$ such that, for all h small enough,

$$\left|\mathbb{E}\left[\mathbb{1}_{|Y| \geq 1}\left(\log|1 + Y| - Y + \frac{Y^2}{2} - \frac{Y^3}{3}\right)\right]\right| \leq C \exp\left(-\frac{\alpha}{h}\right).$$

Expanding $\log(1+Y)$ around 0 yields that

$$\left|\mathbb{E}\left[\mathbb{1}_{|Y|<1}\left(\log(1+Y)-Y+\frac{Y^2}{2}-\frac{Y^3}{3}\right)\right]\right|=\mathscr{O}\big(h^2\big).$$

It finally follows that

$$\bar{\lambda}^h=a+\mathscr{O}(h).$$

3.4 (Concentration Inequalities)

1. The desired inequality comes from
$$\frac{d}{dx}K(f)(x)=\big(1-hU''(x)\big)K\big(f'\big)(x).$$
2. By Theorem 3.8, the standard Gaussian law on $\mathbb{R}^d$ satisfies a logarithmic Sobolev inequality with constant $C=2$. In view of the definition of the entropy, one easily deduces that a Gaussian law with mean vector m and covariance matrix $h\,\mathrm{Id}_{\mathbb{R}^d}$ satisfies a logarithmic Sobolev inequality with constant $C=2h$.
3. The right-hand side is a telescopic sum.
4. Observe that
$$\mathrm{Ent}_{\nu^n}\big(f^2\big)=\sum_{i=1}^{n}K^{i-1}\big(\mathrm{Ent}_{\nu^1}\big(g_{n-i}^2\big)\big),$$
and use Question 2.
5. Using Question 1 one obtains
$$\big|g'_{n-i}\big|^2=\frac{|\frac{d}{dx}K^{n-i}(f^2)|^2}{4K^{n-i}(f^2)}\le(1-\lambda h)^2\frac{(K|\frac{d}{dx}K^{n-i-1}(f^2)|)^2}{4K(K^{n-i-1}(f^2))}.$$
The Cauchy–Schwarz inequality implies
$$\frac{(K(f))^2}{K(g)}\le K\left(\frac{f^2}{g}\right),$$
from which
$$\big|g'_{n-i}\big|^2\le(1-\lambda h)^2K\big(\big|g'_{n-i-1}\big|^2\big).$$
Since $K(\phi)\le K(\psi)$ when $\phi\le\psi$, an iterative procedure leads to
$$\big|g'_{n-i}\big|^2\le(1-\lambda h)^{2(n-i)}K^{n-i}\big(\big|f'\big|^2\big).$$
6. Observe that
$$\mathrm{Ent}_{\nu^n}\big(f^2\big)\le h\left(\sum_{i=0}^{n-1}(1-\lambda h)^{2i}\right)K^n\big(\big|f'\big|^2\big)$$

$$= h\frac{1-(1-\lambda h)^{2n}}{1-(1-\lambda h)^2}K^n\big(|f'|^2\big)$$
$$= \frac{2}{\lambda(2-\lambda h)}\big(1-(1-\lambda h)^{2n}\big)K^n\big(|f'|^2\big).$$

7. Apply Theorem 3.10 with

$$C := \frac{2}{\lambda(2-\lambda h)}\big(1-(1-\lambda h)^{2n}\big).$$

4.2 (Central Limit Theorem and Concentration Inequalities for Poisson Laws)

1. Then $\psi_\lambda(z) := \mathbb{E}(\mathrm{e}^{zX_\lambda}) = \mathrm{e}^{-\lambda}\sum_{k\in\mathbb{N}}\mathrm{e}^{zk}\frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda\mathrm{e}^z} = \mathrm{e}^{\lambda(\mathrm{e}^z-1)}$.
2. For every $z\in\mathbb{C}$

$$\mathbb{E}\big(\mathrm{e}^{zW_\lambda}\big) = \mathrm{e}^{-\frac{z\lambda}{\sqrt{\lambda}}}\psi_\lambda\left(\frac{z}{\sqrt{\lambda}}\right) = \mathrm{e}^{\lambda(\mathrm{e}^{\frac{z}{\sqrt{\lambda}}}-1-\frac{z}{\sqrt{\lambda}})} \xrightarrow[\lambda\to\infty]{} \mathrm{e}^{\frac{z^2}{2}},$$

which is the Laplace transform of the $\mathcal{N}(0,1)$ law, hence W_λ converges in law to $\mathcal{N}(0,1)$; it suffices to consider characteristic functions and take $z = iu$, $u\in\mathbb{R}$.
3. The monotonicity of the exponential function and the Markov inequality yield

$$\mathbb{P}(X_\lambda - \lambda \ge a) = \mathbb{P}\big(\mathrm{e}^{\theta X_\lambda} \ge \mathrm{e}^{\theta(\lambda+a)}\big) \le \mathrm{e}^{-\theta(\lambda+a)}\mathbb{E}\big(\mathrm{e}^{\theta X_\lambda}\big)$$

and thus the first inequality. The second follows by optimizing over $\theta > 0$. The last follows using $\log(1+\frac{a}{\lambda}) \ge \frac{a}{\lambda} - \frac{a^2}{2\lambda^2}$. Taking $a = c\sqrt{\lambda}$ for $c \ge 0$ yields that $\mathbb{P}(W_\lambda \ge c) \le \mathrm{e}^{-\frac{c^2}{2}(1-\frac{c}{\sqrt{\lambda}})}$.
4. Similarly $\mathbb{P}(X_\lambda - \lambda \le -b) = \mathbb{P}(\mathrm{e}^{-\theta X_\lambda} \ge \mathrm{e}^{-\theta(\lambda-b)})$ and the first and second inequalities hold. If $-\frac{1}{2} \le \frac{b}{\lambda} < 0$ then the Taylor–Lagrange expansion yields $\log(1-\frac{b}{\lambda}) \ge -\frac{b}{\lambda} - \frac{b^2}{2\lambda^2} - \frac{8b^3}{3\lambda^3}$ and the third inequality follows. Taking $b = d\sqrt{\lambda}$ for $d \le 0$ yields that $\mathbb{P}(W_\lambda \le d) \le \mathrm{e}^{-\frac{d^2}{2}(1-\frac{d}{\sqrt{\lambda}})+(\lambda-b)\frac{8d^3}{3\lambda^{3/2}}}$.

4.3 (Inhomogeneous Poisson Process)

1. The continuity of Θ implies $A(t) = \inf\{u\in\mathbb{R}_+ : \Theta(u) = t\}$ and $\Theta(A(t)) = t$. Also, A is left-hand continuous and $A(\Theta(t)) = \inf\{u\in\mathbb{R}_+ : \Theta(u) = \Theta(t)\} \le t$.

   By definition $N_t - \hat{N}_{\Theta(t)} = N_t - N_{A(\Theta(t))}$ and the law of $N_t - N_{A(\Theta(t))}$ depends only on $\Theta(t) - \Theta(A(\Theta(t)))$. Moreover $\Theta(A(\Theta(t))) = \Theta(t)$ by the previous results. Hence the law of $N_t - \hat{N}_{\Theta(t)}$ is the same as the law of $N_0 := 0$.

   By definition $T_n = \inf\{u\ge 0 : N_u = n\} = \inf\{u\ge 0 : \hat{N}_{\Theta(u)} = n\} = \inf\{u \ge 0 : \Theta(u) = \hat{T}_n\} = A(\hat{T}_n)$. Hence $\Theta(T_n) = \Theta(A(\hat{T}_n)) = \hat{T}_n$.
2. For all $t, h \ge 0$, the r.v. $\hat{N}_{t+h} - \hat{N}_t = N_{A(t+h)} - N_{A(t)}$ is independent of $(N_u)_{0\le u\le A(t)}$, and thus of $(\hat{N}_{s_1},\dots,\hat{N}_{s_n}) = (N_{A(s_1)},\dots,N_{A(s_n)})$ for any $0 \le s_1 < \cdots < s_n \le t$. This implies it is independent of $(\hat{N}_s)_{0\le s\le t}$. Moreover, its law depends only on $\int_{A(t)}^{A(t+h)}\theta(s)\,ds = \Theta(A(t+h)) - \Theta(A(t)) = t+h-t = h$, and

must be the same as the law of N_h. Such a process with independent stationary increments must be a Poisson process.

3. Simulate a Poisson process $(\hat{N}_t)_{t\in\mathbb{R}_+}$ of intensity 1, compute $(\Theta(t))_{t\in\mathbb{R}_+}$, and set $N_t = \hat{N}_{\Theta(t)}$. The problem is to compute $(\Theta(t))_{t\in\mathbb{R}_+}$ in practice.
4. The r.v.'s $N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_n} - N_{t_{n-1}}$ are independent, and $N_{t_i} - N_{t_{i-1}} = \hat{N}_{\Theta(t_i)} - \hat{N}_{\Theta(t_{i-1})}$ has Poisson law with parameter $\Theta(t_i) - \Theta(t_{i-1}) = \int_{t_i}^{t_{i-1}} \theta(s)\,ds$.
5. The strong Markov property of $(\hat{N}_t)_{t\in\mathbb{R}_+}$ yields that

$$\begin{aligned}
\mathbb{P}(T_{n+1} - T_n \geq t \,|\, T_1, \ldots, T_n) &= \mathbb{P}\big(A(\hat{T}_{n+1}) - A(\hat{T}_n) \geq t \,|\, A(\hat{T}_1), \ldots, A(\hat{T}_n)\big) \\
&= \mathbb{P}\big(A(\hat{T}_{n+1}) - A(\hat{T}_n) \geq t \,|\, A(\hat{T}_n)\big) \\
&= \mathbb{P}(T_{n+1} - T_n \geq t \,|\, T_n), \\
&= \int_0^\infty \mathbb{1}_{\{A(\Theta(T_n)+s)-T_n\geq t\}} \mathrm{e}^{-s}\,ds \\
&= \exp\big(-\min\big\{u : A\big(\Theta(T_n)+u\big) \geq T_n + t\big\}\big) \\
&= \exp\big(-\big\{\Theta(T_n + t) - \Theta(T_n)\big\}\big).
\end{aligned}$$

6. This is obvious: $(\hat{N}_t)_{t\in\mathbb{R}_+}$ is indeed a Poisson process with intensity 1 since $\Theta(A(t)) = t$, and $N_t = \hat{N}_{\Theta(t)}$.
7. Use the sampling property for Poisson processes.

4.4 (Brownian Limit for the Poisson Process)

1. (a) Classically $\psi_x(z) = \mathrm{e}^{-x} \sum_{k\in\mathbb{N}} \mathrm{e}^{zk} \frac{x^k}{k!} = \mathrm{e}^{x(\mathrm{e}^z - 1)}$.
   (b) A Taylor expansion of order 4 of the moment generation function $\psi_x(z) = \mathrm{e}^{x(\mathrm{e}^z-1)}$ yields the result.
2. (a) For any $z \in \mathbb{C}$ it holds that $\mathbb{E}(\mathrm{e}^{zW_t^\theta}) = \mathrm{e}^{-\sqrt{\theta}tz}\mathbb{E}(\mathrm{e}^{\frac{z}{\sqrt{\theta}}N_{\theta t}})$, and $N_{\theta t}$ has Poisson law $\mathscr{P}(\theta t)$ and hence

$$\mathbb{E}\big(\mathrm{e}^{zW_t^\theta}\big) = \mathrm{e}^{-\sqrt{\theta}tz}\mathrm{e}^{\theta t(\mathrm{e}^{\frac{z}{\sqrt{\theta}}}-1)} = \mathrm{e}^{\theta t(\mathrm{e}^{\frac{z}{\sqrt{\theta}}}-1-\frac{z}{\sqrt{\theta}})} = \mathrm{e}^{t\frac{z^2}{2}+o(\theta)}.$$

   This implies that W_t^θ converges in law to a $\mathscr{N}(0,t)$ random variable such as W_t; it is enough to consider characteristic functions, and take $z = \mathrm{i}s$ for $s \in \mathbb{R}$.
   (b) The processes have independent and stationary increments. Hence these vectors have independent coordinates, so that the convergence in law of the vectors follows immediately from that of their coordinates (for instance by using multivariate characteristic functions).
   (c) These random vectors are the image by a given linear (hence continuous) function of the previous random vectors.
3. (a) For all $0 \leq s \leq t \leq T$ such that $|t - s| \leq \varepsilon$, there exists $k \in \mathbb{N}$ such that $k \leq K(T,\varepsilon)$ and $k\varepsilon \leq s \leq t \leq k\varepsilon + \varepsilon$ or $k\varepsilon \leq s \leq k\varepsilon + \varepsilon \leq t \leq k\varepsilon + \varepsilon$. We

bound $|W_t^\theta - W_s^\theta|$ in the first case by $|W_t^\theta - W_{k\varepsilon}^\theta| + |W_{k\varepsilon}^\theta - W_s^\theta|$ and in the second by $|W_t^\theta - W_{k\varepsilon+\varepsilon}^\theta| + |W_{k\varepsilon+\varepsilon}^\theta - W_{k\varepsilon}^\theta| + |W_{k\varepsilon}^\theta - W_s^\theta|$, hence the factor 3.

(b) We use the previous result, and set $A_k = \{\sup_{k\varepsilon \le t \le k\varepsilon+\varepsilon} |W_t^\theta - W_{k\varepsilon}^\theta| \ge a\}$. Note that $\mathbb{P}(A_k) = \mathbb{P}(A_0)$ since the process has independent stationary increments. We conclude using the elementary inequalities $K(T,\varepsilon) \le T/\varepsilon$ and

$$\mathbb{P}\left(\bigcup_{k=0}^{K(T,\varepsilon)} A_k\right) \le \sum_{k=0}^{K(T,\varepsilon)} \mathbb{P}(A_k) = \big(K(T,\varepsilon)+1\big)\mathbb{P}(A_0).$$

4. By definition $\mathbb{P}(\sup_{0\le n\le m} M_n \ge c) = \mathbb{P}(M_\tau \ge c)$, and $\mathbb{P}(M_\tau \ge c) \le \frac{1}{c}\mathbb{E}(M_\tau)$ by the Markov inequality. We then note that τ is a stopping time bounded by m, and that thus, classically for a sub-martingale, $\mathbb{E}(M_\tau) \le \mathbb{E}(M_m)$: for instance

$$\mathbb{E}(M_\tau) = \sum_{i=0}^{m} \mathbb{E}(M_i \mathbb{1}_{\{\tau=i\}}) \le \sum_{i=0}^{m} \mathbb{E}\big(\mathbb{E}(M_m \mid M_0, \dots, M_i)\mathbb{1}_{\{\tau=i\}}\big)$$

$$= \sum_{i=0}^{m} \mathbb{E}\big(\mathbb{E}(M_m \mathbb{1}_{\{\tau=i\}} \mid M_0, \dots, M_i)\big) = \mathbb{E}(M_m).$$

5. (a) It is a process with centered independent increments, and such a process is classically a martingale.

(b) Classically, this can be deduced from the previous result by using either the Jensen or the Hölder inequality.

(c) We come back to the bound obtained at the end of Question 3.

We bound $\mathbb{P}(\sup_{0\le t\le\varepsilon} |W_t^\theta| \ge a) = \mathbb{P}(\sup_{0\le t\le\varepsilon} |W_t^\theta|^4 \ge a^4)$ using the previous result for $p=4$, and the lemma for $M_t = |W_t^\theta|^4$ and $c = a^4$ and $u = \varepsilon$.

Note that $\mathbb{E}|W_\varepsilon^\theta|^4 := \frac{1}{\theta^2}\mathbb{E}((X_x - x)^4)$ for $x = \theta\varepsilon$ and that $\mathbb{E}((X_x - x)^4) = \mathbb{E}(X_x^4) - 4x\mathbb{E}(X_x^3) + 6x^2\mathbb{E}(X_x^2) - 4x^3\mathbb{E}(X_x) + x^4 = x + 3x^2$. Thus

$$\mathbb{E}\big|W_\varepsilon^\theta\big|^4 = \frac{1}{\theta^2}\mathbb{E}\big((X_{\theta\varepsilon} - \theta\varepsilon)^4\big) = \frac{1}{\theta^2}\big(\theta\varepsilon + 3\theta^2\varepsilon^2\big) = \left(\frac{\varepsilon}{\theta} + 3\varepsilon^2\right).$$

5.1 (Assymetric Ehrenfest Urn)

1. This is obtained from the description, by superposition of Poisson processes.
2. Clearly $(X_t)_{t\ge0}$ takes values in $\{0,\dots,N\}$, and its jumps are of size $+1$ or -1. When it is in state x, and hence $(Y_i(t), 1 \le i \le N)_{t\ge0}$ is in some state $(y_i, 1 \le i \le N)$ such that $\sum_{1\le i\le N} y_i = x$ (there are many such states):

- it jumps to $x+1$ at rate $\lambda \sum_{1\le k\le N}(1 - y_k) = \lambda(N - x)$;
- it jumps to $x - 1$ at rate $\mu \sum_{1\le k\le N} y_k = \mu x$.

This is indeed the evolution of a Markov process with generator Q.

3. (a) The forward Kolmogorov equation can be written as

$$\begin{cases} \dfrac{d}{dt}\pi_t(0) = -\pi_t(0)\lambda N + \pi_t(1)\mu, \\ \dfrac{d}{dt}\pi_t(x) = \pi_t(x-1)\lambda(N-x+1) - \pi_t(x)\big(\lambda(N-x)+\mu x\big) \\ \qquad\qquad + \pi_t(x+1)\mu(x+1), \\ \dfrac{d}{dt}\pi_t(N) = \pi_t(N-1)\lambda - \pi_t(N)\mu N. \end{cases}$$

   (b) Use $f_t(x) = \pi_t(0) + \cdots + \pi_t(x)$ and $\pi_t(x) = f_t(x) - f_t(x-1)$, $0 \le i \le N$.
   (c) By summation $\frac{d}{dt} f_t(x) = -\pi_t(x)\lambda(N-x) + \pi_t(x+1)\mu(x+1)$.
   (d) It is necessary and sufficient that $\frac{d}{dt}\pi_t = 0$ and hence that $\frac{d}{dt} f_t = 0$, and hence $\tilde\pi(x+1)\mu(x+1) = \tilde\pi(x)\lambda(N-x)$ for $x = 0, \ldots, N-1$.
   (e) Iteration yields $\tilde\pi(x) = \tilde\pi(0)(\frac{\lambda}{\mu})^x \frac{N\cdots(N-x+1)}{x!} := \tilde\pi(0)(\frac{\lambda}{\mu})^x \binom{N}{x}$, and computing $\sum_{x=0}^N (\frac{\lambda}{\mu})^x \binom{N}{x} = \sum_{x=0}^N \binom{N}{x}(\frac{\lambda}{\mu})^x 1^{N-x} = (\frac{\lambda}{\mu}+1)^N$ yields $\tilde\pi(0) = (\frac{\lambda}{\mu}+1)^{-N}$.

5.2 (Chain Reaction)

1. Then $\sum_{k\ne 1}\eta(k) = 1 - \eta(1) > 0$. By sampling from a Poisson process and an elementary conditioning, the process for λ and η has the same stochastic evolution as the process for $\lambda' = \lambda(1-\eta(1)) > 0$ and the law η' on $\mathbb{N}$ given by $\eta'(1) = 0$ and $\eta'(k) = \eta(k)/(1-\eta(1))$ for $k \ne 1$.
2. (a) Non-zero non-diagonal terms of the generator: $Q(x,y) = \lambda x \eta(y-x+1)$ for $x \ne y$ such that $y \ge x-1$ in $\mathbb{N}$. From a state x, the process waits an exponential time of parameter λx and then jumps to $y \ne x$ with probability $\eta(y-x+1)$. This generator is not bounded, and there could be explosion.
   (b) For instance

$$\hat X_n = X_0 + \sum_{k=1}^n (Z_k - 1)\mathbb{1}_{k\ge M}.$$

   Use the strong law of large numbers $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n (Z_k - 1) = m - 1$, a.s.
   (c) Conditional on $(\hat X_k)_{0\le k\le n-1}$, the law of T_n is that of a sum of n independent exponential r.v.'s of parameters $\lambda\hat X_0, \ldots, \lambda\hat X_{n-1}$, and hence

$$\mathbb{E}\big(\mathrm{e}^{-T_n} \mid (\hat X_k)_{0\le k\le n-1}\big) = \prod_{k=0}^{n-1} \frac{\lambda\hat X_k}{1+\lambda\hat X k}.$$

   Taking expectations, using $\hat X_n = \mathscr{O}(n)$, and by dominated convergence,

$$\mathbb{E}\big(\mathrm{e}^{-T_n}\big) = \mathbb{E}\left(\prod_{k=0}^{n-1} \frac{\lambda\hat X_k}{1+\lambda\hat X k}\right) = \mathbb{E}\left(\exp\left(\sum_{k=0}^{n-1} \log\left(1 - \frac{1}{1+\lambda\hat X_k}\right)\right)\right) \xrightarrow[n\to\infty]{} 0$$

   which implies that $T_n \xrightarrow[n\to\infty]{} \infty$, a.s.

3. (a) Conditional on $X_0 = x$, it is enough to consider the sum of x independent copies of the process started at 1. Hence $\mathbb{E}_x(z^{X_t}) = f_t(z)^x$.
   (b) Conditional on X_ε, the process $(X_{t+\varepsilon})_{t\geq 0}$ is a Markov process with the same stochastic evolution as the original process but started at X_ε, so that

$$f_{t+\varepsilon}(z) := \mathbb{E}_1\big(z^{X_{t+\varepsilon}}\big) = \sum_{x\in\mathbb{N}} \mathbb{E}_x\big(z^{X_t}\big)\mathbb{P}_1(X_\varepsilon = x) = \sum_{x\in\mathbb{N}} f_t(z)^x \mathbb{P}_1(X_\varepsilon = x)$$
$$= f_\varepsilon\big(f_t(z)\big).$$

   (c) The process has a jump in $[0,\varepsilon]$ with probability $1 - \mathrm{e}^{-\lambda\varepsilon} = \lambda\varepsilon + \mathscr{O}(\lambda^2\varepsilon^2)$, and conditional on that, its first jump leads it to k with probability $\eta(k)$ and the probability of it jumping another time within $[0,\varepsilon]$ is bounded by

$$\sum_{k\in\mathbb{N}} \big(1 - \mathrm{e}^{-k\lambda\varepsilon}\big)\eta(k) \leq \sum_{k\in\mathbb{N}} k\lambda\varepsilon\eta(k) = m\lambda\varepsilon,$$

   hence the probability that it has two or more jumps within $[0,\varepsilon]$ is $\mathscr{O}(m\lambda^2\varepsilon^2)$.
   (d) The previous results allow us to consider the case $t = 0$ and to justify that

$$f_\varepsilon(z) - f_0(z) = \mathbb{E}_1\big(z^{X_\varepsilon}\big) - z = -\lambda\varepsilon z + \lambda\varepsilon \sum_{k\geq 0} z^k \eta(k) + \mathscr{O}\big((m+1)\lambda^2\varepsilon^2\big)$$
$$= h(z)\varepsilon + \mathscr{O}\big((m+1)\lambda^2\varepsilon^2\big).$$

   From this, the differential equation is obvious.

   It is the line for $x = 1$ of the backward Kolmogorov equation for $f : y \mapsto z^y$:

$$\frac{d}{dt} f_t(z) = \frac{d}{dt} v_t(1) = \sum_{y\in\mathbb{N}} Q(1,y) v_t(y) = \sum_{y\in\mathbb{N}} Q(1,y) f_t(z)^y$$
$$= \lambda g\big(f_t(z)\big) - \lambda f_t(z).$$

4. Since $\mathscr{E} = \bigcup_{t\in\mathbb{R}_+} \uparrow \{X_t = 0\}$, by monotone limit $\mathbb{P}_1(\mathscr{E}) = \lim_{t\to\infty} \uparrow \mathbb{P}_1(X_t = 0)$, and $\mathbb{P}_1(X_t = 0) = f_t(0)$. We conclude by a simple study of $\frac{\partial}{\partial t} f_t(0) = u(f_t(0))$.
5. Then

$$\frac{\partial}{\partial t}\left(\frac{\partial}{\partial z} f_t(z)\right) = \frac{\partial}{\partial z}\frac{\partial}{\partial t} f_t(z) = \frac{\partial}{\partial z}\big(h\big(f_t(z)\big)\big) = \left(\frac{\partial}{\partial z} f_t(z)\right) h'\big(f_t(z)\big),$$
$$\mathbb{E}_1(X_t) = \frac{\partial}{\partial z} f_t(z=0) = \exp\left(\int_0^t h'(1)\, ds\right) = \mathrm{e}^{\lambda(m-1)t}.$$

   If $m = 1$ then $\mathbb{E}_1(X_t) = 1$ and $\lim_{t\to\infty} X_t = 0$ a.s., and the Dominated Convergence Theorem implies that $\mathbb{E}_1(\sup_{t\geq 0} X_t) = \infty$.

6. The function $t \mapsto f_t(z)$ is solution of $f' = \lambda(pf^2 - f + 1 - p)$ with $f_0 = z$, and

$$pf^2 - f + 1 - p = p(f-1)\left(f - \frac{1-p}{p}\right).$$

If $p = 1/2$ then $f' = \frac{\lambda}{2}(f-1)^2$ and hence $(\frac{1}{f-1})' = -\frac{\lambda}{2}$ and we conclude that

$$\frac{1}{f_t(z)-1} = \frac{1}{z-1} - \frac{\lambda}{2}t, \quad f_t(z) = 1 + \frac{1}{\frac{1}{z-1} - \frac{\lambda}{2}t} = 1 - \frac{1-z}{1 + \frac{\lambda}{2}t(1-z)},$$

and $\mathbb{P}_1(\mathscr{E}) = \lim_{t\to\infty} f_t(0) = 1$. If $p \neq 1/2$ then the decomposition

$$\frac{1}{(f-1)(f - \frac{1-p}{p})} = \frac{p}{2p-1}\left(\frac{1}{f-1} - \frac{1}{f - \frac{1-p}{p}}\right)$$

leads to the differential equation $(\log \frac{f-1}{f - \frac{1-p}{p}})' = (2p-1)\lambda$ which is solved by

$$\frac{f_t(z)-1}{f_t(z) - \frac{1-p}{p}} = \frac{z-1}{z - \frac{1-p}{p}} \mathrm{e}^{(2p-1)\lambda t}, \quad z \neq \frac{1-p}{p},$$

and hence

$$f_t(z) = \frac{1 - \frac{1-p}{p}\frac{z-1}{z-\frac{1-p}{p}}\mathrm{e}^{(2p-1)\lambda t}}{1 - \frac{z-1}{z-\frac{1-p}{p}}\mathrm{e}^{(2p-1)\lambda t}} = 1 + \frac{2p-1}{p}\frac{\frac{z-1}{z-\frac{1-p}{p}}\mathrm{e}^{(2p-1)\lambda t}}{1 - \frac{z-1}{z-\frac{1-p}{p}}\mathrm{e}^{(2p-1)\lambda t}}$$
$$= 1 - \frac{2p-1}{p}\frac{1-z}{(z - \frac{1-p}{p})\mathrm{e}^{-(2p-1)\lambda t} + 1 - z}.$$

Hence, if $p < 1/2$ then $\mathbb{P}_1(\mathscr{E}) = \lim_{t\to\infty} f_t(0) = 1$, and if $p > 1/2$ then $\mathbb{P}_1(\mathscr{E}) = \lim_{t\to\infty} f_t(0) = 1 - \frac{2p-1}{p} = \frac{1-p}{p}$.

5.3 (Markov process, generator, and martingales)

1. Since M_t^f is $(X_s)_{0\le s\le t}$-measurable, the Markov property yields

$$\begin{aligned}
&\mathbb{E}\big(M_{t+h}^f \,|\, (X_s)_{0\le s\le t}\big) - M_t^f \\
&\quad = \mathbb{E}\big(M_{t+h}^f - M_t^f \,|\, (X_s)_{0\le s\le t}\big) \\
&\quad = \mathbb{E}\left(f(X_{t+h}) - f(X_t) - \int_t^{t+h} Qf(X_s)\,ds \,|\, (X_s)_{0\le s\le t}\right) \\
&\quad = F_h^f(X_t)
\end{aligned}$$

where

$$F_h^f : x \mapsto \mathbb{E}_x\left(f(X_h) - f(x) - \int_0^h Qf(X_s)\,ds\right)$$

$$= P_h f(x) - f(x) - \int_0^h P_s Q f(X_s)\, ds$$

vanishes considering the backward Kolmogorov equation since $P_s Q = Q P_s$, or by the forward Kolmogorov equation by letting it integrate f.

2. (a) This can be readily deduced from

$$M_t^{f^2} = f^2(X_t) - \int_0^t Q\big(f^2\big)(X_s)\, ds, \qquad M_t^f = f(X_t) - \int_0^t Qf(X_s)\, ds,$$

$$\big(M_t^f\big)^2 = f^2(X_t) - 2f(X_t) \int_0^t Qf(X_s)\, ds + \left(\int_0^t Qf(X_s)\, ds\right)^2.$$

(b) This follows from the fact that a differentiable function can be expressed as the integral of its derivative.

(c) The increment between $t \geq 0$ and $t + h$ for $h \geq 0$ is given by

$$\begin{aligned} & M_{t+h}^f \int_0^{t+h} Qf(X_s)\, ds - \int_0^{t+h} M_s^f Qf(X_s)\, ds \\ & \quad - M_t^f \int_0^t Qf(X_s)\, ds + \int_0^t M_s^f Qf(X_s)\, ds \\ & = \big(M_{t+h}^f - M_t^f\big) \int_0^t Qf(X_s)\, ds + M_{t+h}^f \int_t^{t+h} Qf(X_s)\, ds \\ & \quad - \int_t^{t+h} M_s^f Qf(X_s)\, ds. \end{aligned}$$

Take the conditional expectation w.r.t. $(X_u)_{0 \leq u \leq t}$. The first term vanishes by the martingale property of $(M_t^f)_{t \geq 0}$, and the rest yields, using the Fubini theorem, the tower property of conditional expectations, and the first result,

$$\begin{aligned} & \mathbb{E}\left(\int_t^{t+h} \big(M_{t+h}^f - M_s^f\big) Qf(X_s)\, ds \,|\, (X_u)_{0 \leq u \leq t} \right) \\ & = \int_t^{t+h} \mathbb{E}\big(\big(M_{t+h}^f - M_s^f\big) Qf(X_s) \,|\, (X_u)_{0 \leq u \leq t} \big)\, ds \\ & = \int_t^{t+h} \mathbb{E}\big(\mathbb{E}\big(M_{t+h}^f - M_s^f \,|\, (X_u)_{0 \leq u \leq s}\big) Qf(X_s) \,|\, (X_u)_{0 \leq u \leq t} \big)\, ds = 0. \end{aligned}$$

Another "stochastic calculus type" method uses the integration by parts

$$M_t^f \int_0^t Qf(X_s)\, ds = \int_0^t M_s^f Qf(X_s)\, ds + \int_0^t \int_0^s Qf(X_u)\, du\, dM_s^f$$

where the last integral is of Stieljes type (actually, a series), and can be easily shown to define a martingale.

(d) It is enough to collect the previous results. The expectation of a martingale vanishing at zero is zero.

3. Use that $Qa = 0$ and hence $Q(f-a)^2 - 2(f-a)Q(f-a) = Qf^2 - 2fQf$, then condition on X_0.
4. It holds that

$$Q\big(f^2\big)(x) - 2f(x)Qf(x) = \sum_{y\neq x} Q(x,y)\big(f^2(y) - f^2(x) - 2f(x)\big(f(y)-f(x)\big)\big)$$
$$= \sum_{y\neq x} Q(x,y)\big(f(y)-f(x)\big)^2.$$

6.1 (Binary Energy Exchanges)

1. Let $(e_i)_{1\leq i\leq N}$ be the canonical basis of $\mathbb{R}_+^N$. The generator acting on $f \in L^\infty(\mathbb{R}_+)$ yields the function, of $x = (x_i)_{1\leq i\leq N}$ in $\mathbb{R}_+^N$,

$$\sum_{1\leq i\leq N}\left(\alpha\big(f(x+e_i)-f(x)\big) + \beta x_i \int_0^1 \big(f(x-\theta x_i e_i) - f(x)\big)\,d\theta\right)$$
$$+ \sum_{1\leq i<j\leq N} \gamma x_i x_j \int_0^{x_i+x_j} \big(f\big(x+(z-x_i)e_i + (x_i-z)e_j\big) - f(x)\big)\,dz.$$

This is clearly not bounded.

2. The generator is bounded on each B_k, the jump rate from B_k to B_{k+1} is αN, and the jump rate from B_k to B_{k+m} is null for $m \geq 2$. Thus, explosion is impossible.
3. Let the process be at $x = (x_i)_{1\leq i\leq N}$ in $\mathbb{R}_+^N$. A natural simulation method lets the process stay at x for an exponential duration of parameter

$$q(x) := \alpha N + \beta \sum_{1\leq i\leq N} x_i + \gamma \sum_{1\leq i<j\leq N} x_i x_j.$$

Note that

$$\sum_{1\leq i<j\leq N} x_i x_j = \frac{1}{2}\sum_{1\leq i\neq j\leq N} x_i x_j = \frac{1}{2}\sum_{1\leq i\leq N} x_i \sum_{1\leq j\leq N} x_j - \frac{1}{2}\sum_{1\leq i\leq N} x_i^2$$

allows to compute $q(x)$ in terms of $s(x) = \sum_{1\leq i\leq N} x_i$ and $s_2(x) = \sum_{1\leq i\leq N} x_i^2$.

After this, a coordinate $1 \leq i \leq N$ is chosen with probability $\alpha/q(x)$ to be incremented by 1, with probability $\beta x_i/q(x)$ to be decreased by θx_i for θ drawn uniformly from $[0,1]$, or with probability $\gamma(x_i s(x) - x_i^2)/2q(x)$ to be involved in a collision, with a particle $1 \leq j \leq N$ for $j \neq i$ chosen with probability $x_j/(s(x)-x_j)$ and resulting in the replacement of x_i and x_j by z drawn uniformly in $[0, x_i + x_j]$ and $x_i + x_j - z$.

4. One can avoid computing $s_2(x)$ by replacing $q(x)$ by $\lambda(x) := \alpha N + \beta s(x) + \frac{\gamma}{2}s(x)^2$ in the above and further choosing with probability $\gamma(x_i s(x))/2\lambda(x)$ that

particle i is involved in a collision, with particle $1 \le j \le N$ with probability $x_j/s(x)$, and if $j = i$ then the collision is fictitious.

The sub-domain method with the $B_k - B_{k-1}$ can be tried to accelerate the method.

6.2 (Jump Accumulation)

1. When at a point x in $[a_{n-1}, a_n)$, the process waits an exponential time of parameter λ_n and then jumps to the point of $[a_n, a_{n+1})$ with the same barycentric coordinates. When at 1, the process waits an exponential time of parameter λ and jumps to a uniformly chosen point in $[0, 1]$.
2. The condition $\sum_{n\ge1} \frac{1}{\lambda_n} < \infty$ implies that the jump instants accumulate in finite time (the expectation of the limit is finite), and the evolution implies that 1 is reached in finite time from any state. This implies that the process is defined for all times. Moreover, the description of the evolution is Markovian. Naively, the generator acting on a bounded f yields the function of x given by

$$\sum_{n\ge1} \mathbb{1}_{\{a_{n-1}\le x<a_n\}}\lambda_n\left(f\left(\frac{a_n-x}{a_n-a_{n-1}}a_n+\frac{x-a_{n-1}}{a_n-a_{n-1}}a_{n+1}\right)-f(x)\right)$$
$$-\mathbb{1}_{\{x=1\}}\lambda\int_0^1\big(f(u)-f(x)\big)\,du$$

   which is not necessarily bounded since $\lim_{n\to\infty}\lambda_n = \infty$. If f is assumed to be constant on a neighborhood of 0 then the resulting function is bounded, but depending on the rate of increase of λ_n more general functions can be considered, for instance with vanishing derivatives up to some order at 0.
3. The process cannot be simulated exactly, due to the accumulation of jumps. It can be simulated approximately: select some threshold close enough to 0, and above it replace the actual jumps with the mean evolution of the process, which for $a_{n-1} \le x < a_n$ is at speed

$$\lambda_n\left(\frac{a_n-x}{a_n-a_{n-1}}a_n+\frac{x-a_{n-1}}{a_n-a_{n-1}}a_{n+1}-x\right)$$
$$=\lambda_n\frac{a_n^2-a_{n-1}a_{n+1}+x(a_{n+1}-2a_n+a_{n-1})}{a_n-a_{n-1}}.$$

6.3 (Generalized Kac Equation)

1. This follows from $(v_1\cos\theta - v_2\sin\theta)^2 + (v_1\sin\theta + v_2\cos\theta)^2 = (v_1)^2 + (v_2)^2$.
2. (a) We use the superposition-decomposition properties of the Poisson processes. Iteratively, from a state $v = (v_k)_{1\le k\le N}$, we draw an exponential r.v. of parameter N to determine the duration until the next jump instant, then i uniformly in $\{1,\dots,N\}$, then j uniformly in $\{1,\dots,N\}-\{i\}$, then θ according to $\beta(\theta)\,d\theta$, and then we determine the next state by replacing v_i by $v_i\cos\theta - v_j\sin\theta$ and v_j by $v_i\sin\theta + v_j\cos\theta$ in v. This simulation always functions, since there cannot be any accumulation of jump instants.

(b) Using again the superposition-decomposition properties of the Poisson processes, the random evolution is indeed that of a pure-jump Markov process with said generator $\mathscr{A}$. This generator is bounded by $\frac{2}{N-1} \times \frac{N(N-1)}{2} \times 2 = 2N$, hence the process is well-defined.

3. Let $h : v = (v_k)_{1\le k\le N} \mapsto h(v) = \sum_{k=1}^{N} v_k$. Let $M_t = h(V_t)$ be the impetus of the system, and $m_t = \mathbb{E}(M_t)$ its expectation, $t \ge 0$.

(a) The function β is even and thus $\int_{-\pi}^{\pi} \sin\theta\, \beta(\theta)\, d\theta = 0$. Hence

$$\mathscr{A}h(v) = \frac{2}{N-1}(c-1) \sum_{1\le i<j\le N} (v_i + v_j) = \frac{2}{N-1}(c-1) \sum_{1\le i\ne j\le N} v_i$$
$$= 2(c-1)h(v).$$

(b) This follows from the forward Kolmogorov equation and the preceding result. Solving the differential equation is trivial.

(c) This follows from the fact that $\cos\theta \le 1$ with equality only for $\theta = 0$, which has null Lebesgue measure.

4. Assume here that $0 < \|\beta\|_1 < \infty$.

(a) This is still a Markov process with generator bounded now by $2N\|\beta\|_1$.

(b) It is simulated as above, but with exponential r.v. with parameter $N\|\beta\|_1$ and draws of θ according to the probability density $\frac{\beta(\theta)}{\|\beta\|_1}$.

(c) There will be a large number of iterations for simulating for any reasonable duration of time.

5. (a) Then $\sin\theta = \mathscr{O}(\theta)$ and $\cos\theta - 1 = \mathscr{O}(\theta^2)$ uniformly, and we use a multidimensional Taylor formula of order 1 (considering coordinates i and j for $1 \le i < j \le N$).

(b) Use the above Taylor expansion, and the fact that β is even and $\sin\theta = \mathscr{O}(\theta)$ uneven, so that $\int_0^{\varepsilon} \sin\theta\, \beta(\theta)\, d\theta = 0$, and $\cos\theta = 1 - \frac{\theta^2}{2} + \mathscr{O}(\theta^4)$.

(c) This is a Markov process which follows the ordinary differential equation

$$\frac{d}{dt} v_i(t) = -b_\varepsilon v_i(t), \quad 1 \le i \le N,$$

in between jumps, similar to the above, taken with intensity $N\|\beta(\theta)\mathbb{1}_{|\theta|\ge\varepsilon}\|_1$. It is well-defined due to the bounded jump rate.

(d) The simulation of jumps is as above, replacing $\beta(\theta)$ by $\beta(\theta)\mathbb{1}_{|\theta|\ge\varepsilon}$. In between jumps, use the solution $v_i(t) = v_i(0)\mathrm{e}^{-b_\varepsilon t}$ of the differential equation. Since $\lim_{\varepsilon\to 0}\|\beta(\theta)\mathbb{1}_{|\theta|\ge\varepsilon}\|_1 = \infty$, for ε small the jump rate becomes large.

7.1 (Long Time Behavior of the Euler Scheme Error)

1. From Itô's formula,

$$S_t^2 = 1 + \left(2\mu + \sigma^2\right) \int_0^t S_\theta^2\, d\theta + 2\sigma \int_0^t S_\theta^2\, dW_\theta.$$

Since the coefficients of the stochastic differential equation satisfied by (S_t) are Lipschitz, the solution has finite moments of all order. Thus the stochastic integral in the right-hand side is a martingale and

$$\mathbb{E}S_t^2 = 1 + \big(2\mu + \sigma^2\big) \int_0^t \mathbb{E}S_\theta^2 \, d\theta.$$

We deduce $\mathbb{E}S_T^2 = \exp((2\mu + \sigma^2)T)$.

2. From the Euler scheme definition,

$$\begin{aligned}\bar{S}^n_{(p+1)T/n} &= \bar{S}^n_{pT/n}\left(1 + \mu\frac{T}{n} + \sigma(W_{(p+1)T/n} - W_{pT/n})\right)\\ &= \prod_{j=0}^{p}\left(1 + \mu\frac{T}{n} + \sigma(W_{(j+1)T/n} - W_{jT/n})\right).\end{aligned}$$

Since the Brownian increments are Gaussian and independent random variables,

$$\begin{aligned}\mathbb{E}\big(\bar{S}^n_T\big)^2 &= \left(1 + \big(2\mu + \sigma^2\big)\frac{T}{n} + \mu^2\frac{T^2}{n^2}\right)^n\\ &= \exp\big(\big(2\mu + \sigma^2\big)T\big)\left(1 + C\frac{T^2}{n} + \mathcal{O}\left(\frac{T^2}{n^2}\right)\right),\end{aligned}$$

where C is a constant which only depends on μ and σ. Consequently, one has $|C_1(T)| = |C|T^2\exp((2\mu + \sigma^2)T)$. This function tends to infinity with T.

3. Itô's formula implies that

$$\mathbb{E}r_t^2 = \sigma^2 t - 2\mu\int_0^t \mathbb{E}r_\theta^2 \, d\theta.$$

The solution to this linear differential equation is

$$\mathbb{E}r_t^2 = \frac{\sigma^2}{2\mu}\big(1 - \exp(-2\mu t)\big),$$

from which

$$\mathbb{E}r_{p/n}^2 = \frac{\sigma^2}{2\mu}\big(1 - \exp(-2\mu p/n)\big).$$

Now, in view of the Euler scheme definition,

$$\begin{aligned}\mathbb{E}\big(\bar{r}^n_{p/n}\big)^2 &= \mathbb{E}\big(\bar{r}^n_{(p-1)/n}\big)^2\left(1 - \frac{\mu}{n}\right)^2 + \frac{\sigma^2}{n}\\ &= \frac{\sigma^2}{n}\frac{1 - (1 - \frac{\mu}{n})^{2p}}{1 - (1 - \frac{\mu}{n})^2}\end{aligned}$$

$$= \frac{\sigma^2}{2\mu + \mathcal{O}(\frac{1}{n})}\left(1 - \exp\left(2p\log\left(1 - \frac{\mu}{n}\right)\right)\right).$$

For all $a > 0$ and $b > 0$ one has $|e^{-a} - e^{-b}| \le |b - a|$. Therefore, there exists a positive real number C which only depends on μ such that, for all n such that $1 - \frac{\mu}{n} > 0$, for all integer p,

$$\left|\mathbb{E}r_{p/n}^2 - \mathbb{E}\left(\bar{r}_{p/n}^n\right)^2\right| \le \frac{C}{n}.$$

We observe that, contrary to Question 2, the error constant is bounded uniformly w.r.t. the numerical integration time horizon.

7.3 (Monotonicity and Convexity of $x \mapsto X_t(x)$)

1. In view of (7.70), the sequence of processes $Z^{x,\varepsilon}$ converges in probability uniformly on compact time intervals. Thus a subsequence converges almost surely to the limit explicited in Exercise 7.5. Thus the flow $X_t(x)$ is an almost surely increasing (consequently, one-to-one) function of x.
2. At the first discretization step one has

$$\bar{X}_{T/n}^n(x) = x + b(x)\frac{T}{n} + \sigma(x)W_{T/n}.$$

Choose $b(x) \equiv 0$ and $\sigma(x) = \sin(x)$. In such a case, the flow of the Euler scheme clearly is not almost surely one-to-one.
3. We deduce from the first question that

$$\begin{aligned}\frac{d^2}{dx^2}X_t(x) = \exp&\left(\int_0^t\left(b'\left(X_\theta(x)\right) - \frac{1}{2}\left(\sigma'\left(X_\theta(x)\right)\right)^2\right)d\theta + \int_0^t \sigma'\left(X_\theta(x)\right)dW_\theta\right)\\ &\times\left(\int_0^t\left(b''\left(X_\theta(x)\right) - \sigma'\left(X_\theta(x)\right)\sigma''\left(X_\theta(x)\right)\right)\frac{d}{dx}X_\theta(x)\,d\theta\right.\\ &\left.+\int_0^t \sigma''\left(X_\theta(x)\right)\frac{d}{dx}X_\theta(x)\,dW_\theta\right).\end{aligned}$$

4. Since $f'(x) = f'(0) + \int_0^x f''(y)\,dy$, the assumption on f'' implies that $|f'|$ is bounded from above by a polynomial function in $|x|$. Thus, f does likewise.
5. Theorem 7.18 shows that

$$\frac{\partial u}{\partial x}(t,x) = \mathbb{E}\left[f'\left(X_t(x)\right)\frac{d}{dx}X_t(x)\right].$$

We deduce

$$\frac{\partial^2 u}{\partial x^2}(t,x) = \mathbb{E}\left[f''\left(X_t(x)\right)\left(\frac{d}{dx}X_t(x)\right)^2\right] + \mathbb{E}\left[f'\left(X_t(x)\right)\frac{d^2}{dx^2}X_t(x)\right].$$

Now, Proposition 7.9 shows that all moments of $X_t(x)$ are bounded from above by a polynomial function in $|x|$. In addition, Proposition 7.4 shows that, for all adapted and bounded process (ϕ_t) and all integers m,

$$\exists C > 0, \quad \sup_{0\le t\le T} \mathbb{E}\left| \int_0^t \phi_\theta \, dW_\theta \right|^{2m} \le C.$$

Observe finally

$$\exp\left(\int_0^t \phi_\theta \, dW_\theta\right) = 1 + \frac{1}{2}\int_0^t \exp\left(\int_0^\theta \phi_s \, dW_s\right)\phi_\theta^2 \, d\theta + \int_0^t \exp\left(\int_0^\theta \phi_s \, dW_s\right)\phi_\theta \, dW_\theta.$$

Set $T_N := \inf\{t > 0; \, |\int_0^t \phi_\theta \, dW_\theta| \ge N\}$. From the preceding equality we have

$$\mathbb{E}\exp\left(\int_0^{t\wedge T_N} \phi_\theta \, dW_\theta\right) = 1 + \frac{1}{2}\mathbb{E}\int_0^{t\wedge T_N} \exp\left(\int_0^\theta \phi_s \, dW_s\right)\phi_\theta^2 \, d\theta \le 1 + C\int_0^t \mathbb{E}\exp\left(\int_0^{\theta\wedge T_N} \phi_s \, dW_s\right) d\theta.$$

As at the end of the proof of Proposition 7.8, successively apply Gronwall's lemma and Fatou's lemma. It follows that

$$\exists C > 0, \ \forall 0 \le t \le T, \quad \mathbb{E}\exp\left(\int_0^t \phi_\theta \, dW_\theta\right) \le C.$$

It is then easy to deduce that the function $\frac{\partial^2 u}{\partial x^2}(t, x)$ is bounded from above by a polynomial function in $|x|$.

6. When $\sigma = \lambda x$, one deduces from the third question

$$\frac{d^2}{dx^2} X_t(x) = \exp\left(\int_0^t \left(b'\big(X_\theta(x)\big) - \frac{\lambda^2}{2}\right) d\theta + \lambda W_t\right) \int_0^t b''\big(X_\theta(x)\big) \frac{d}{dx} X_\theta(x) \, d\theta.$$

If in addition b'', f' and f'' are positive functions, it results from the first and fifth questions that $\frac{d^2}{dx^2} X_t(x)$ and thus $\frac{\partial^2 u}{\partial x^2}(t, x)$ are also positive functions.

7.4 (Estimates for the Derivatives of $x \mapsto X_t(x)$)

1. Since the function $g(t)$ is increasing, $g(t) \ge g(\theta)$ for all $0 \le \theta \le t$. Thus

$$\max_{\theta\le t}(M_\theta) - g(t) \le \max_{0\le\theta\le t}\big(M_\theta - g(\theta)\big).$$

2. Itô's formula implies

$$\tilde{M}_t := 1 + \alpha \int_0^t \tilde{M}_s \phi_s \, dW_s.$$

The process $(\tilde{M}_t)$ therefore is a local martingale. In fact, it is a martingale: this fact results from Novikov's criterion (see Remark 8.1) since (ϕ_t) is a bounded process. Alternatively, one can introduce the stopping times $\tau_N := \inf\{\theta > 0; \tilde{M}_\theta \geq N\}$ and observe

$$\begin{aligned}\tilde{M}^2_{t\wedge\tau_N} &= 1 + \alpha^2 \int_0^{t\wedge\tau_N} \tilde{M}_s^2 \phi_s^2 \, ds + 2\alpha \int_0^{t\wedge\tau_N} \tilde{M}_s^2 \phi_s \, dW_s \\ &= 1 + \alpha^2 \int_0^t \tilde{M}^2_{s\wedge\tau_N} \phi^2_{s\wedge\tau_N} \mathbb{1}_{s\leq\tau_N} \, ds + 2\alpha \int_0^t \tilde{M}^2_{s\wedge\tau_N} \phi_{s\wedge\tau_N} \mathbb{1}_{s\leq\tau_N} \, dW_s,\end{aligned}$$

from which

$$\mathbb{E}\tilde{M}^2_{t\wedge\tau_N} \leq 1 + c\alpha^2 \int_0^t \mathbb{E}\tilde{M}^2_{s\wedge\tau_N} \, ds.$$

Using Gronwall and Fatou's lemmae as in the proof of Theorem 7.8, one obtains $\mathbb{E}\tilde{M}_t^2 < \infty$, from which we conclude that (M_t) is a square integrable martingale.

3. Set $\langle M\rangle_t := \int_0^t (\phi_s)^2 \, ds$. We have

$$\begin{aligned}\mathbb{P}\Big(\max_{0\leq\theta\leq t} M_\theta \geq at\Big) &= \mathbb{P}\Big(\max_{0\leq\theta\leq t} \alpha M_\theta - \frac{\alpha^2}{2}\langle M\rangle_t \geq \alpha a t - \frac{\alpha^2}{2}\langle M\rangle_t\Big) \\ &\leq \mathbb{P}\Big(\max_{0\leq\theta\leq t}\Big(\alpha M_\theta - \frac{\alpha^2}{2}\langle M\rangle_\theta\Big) \geq \alpha a t - \frac{\alpha^2}{2} ct\Big) \\ &= \mathbb{P}\Big(\max_{0\leq\theta\leq t}\Big(\tilde{M}_\theta \geq \exp\Big(\alpha a t - \frac{\alpha^2}{2} ct\Big)\Big)\Big) \\ &\leq \exp\Big(-\alpha a t + \frac{\alpha^2}{2} ct\Big)\mathbb{E}\tilde{M}_t,\end{aligned}$$

where we used Theorem 7.3 to obtain the last inequality. Since $(\tilde{M}_t)$ is a martingale, $\mathbb{E}\tilde{M}_t = \mathbb{E}\tilde{M}_0 = 1$, and thus

$$\mathbb{P}\Big(\max_{0\leq\theta\leq t} M_\theta \geq at\Big) \leq \exp\Big(-\alpha a t + \frac{\alpha^2}{2} ct\Big).$$

The right-hand side takes its minimal value at $\alpha = \frac{a}{c}$, from which

$$\mathbb{P}\Big(\max_{0\leq\theta\leq t} M_\theta \geq at\Big) \leq \exp\Big(-\frac{a^2 t}{2c}\Big).$$

For $a := \frac{y}{t}$, where y is a given positive number, we obtain

$$\mathbb{P}\Big(\max_{0\leq\theta\leq t} M_\theta \geq y\Big) \leq \exp\Big(-\frac{y^2}{2ct}\Big).$$

Since $(-M_t)$ is a martingale, we deduce

$$\begin{aligned}\mathbb{P}\big(|M_t| \geq y\big) &\leq \mathbb{P}(M_t \geq y) + \mathbb{P}(-M_t \geq y)\\ &\leq \mathbb{P}\Big(\max_{0\leq\theta\leq t} M_\theta \geq y\Big) + \mathbb{P}\Big(\max_{0\leq\theta\leq t} (-M_\theta) \geq y\Big)\\ &\leq 2\exp\Big(-\frac{y^2}{2ct}\Big).\end{aligned}$$

Since

$$\begin{aligned}\mathbb{E}\exp\Big(\varepsilon\frac{(M_t)^2}{2}\Big) &\leq \sum_{k=0}^{\infty}\mathbb{E}\Big[\exp\Big(\varepsilon\frac{(M_t)^2}{2}\Big)\mathbb{1}_{k\leq|M_t|<(k+1)}\Big]\\ &\leq \sum_{k=0}^{\infty}\exp\Big(\varepsilon\frac{(k+1)^2}{2}\Big)\mathbb{P}\big(k\leq|M_t|\big)\\ &\leq 2\sum_{k=0}^{\infty}\exp\Big(\varepsilon\frac{(k+1)^2}{2}\Big)\exp\Big(-\frac{k^2}{2ct}\Big),\end{aligned}$$

the expectation $\mathbb{E}\exp(\varepsilon\frac{(M_t)^2}{2})$ is finite for all $\varepsilon < \frac{1}{2c}$.

4. As seen in Exercise 7.5, we have

$$Y_t^x = \exp\Big(\int_0^t\Big(b'\big(X_\theta^x\big) - \frac{1}{2}\big(\sigma'\big)^2\big(X_\theta^x\big)\Big)\,d\theta + \int_0^t \sigma'\big(X_\theta^x\big)\,dW_\theta\Big), \quad 0\leq t\leq T.$$

Set $\mu := \|b' - \frac{1}{2}(\sigma')^2\|_\infty$, $C_T := \|(\sigma')^2\|_\infty T$. For all y such that $\log(y) > \mu T$,

$$\begin{aligned}\mathbb{P}\big(Y_t^x > y\big) &\leq \mathbb{P}\Big(\int_0^t\Big(b'\big(X_\theta^x\big) - \frac{1}{2}\sigma^2\big(X_\theta^x\big)\Big)\,d\theta + \int_0^t\sigma'\big(X_\theta^x\big)\,dW_\theta \geq \log(y)\Big)\\ &\leq \mathbb{P}\Big(\Big|\int_0^t\sigma'\big(X_\theta^x\big)\,dW_\theta\Big| \geq \log(y) - \mu t\Big)\\ &\leq 2\exp\Big(-\frac{(\log(y)-\mu t)^2}{2C_T t}\Big).\end{aligned}$$

5. We have

$$\mathbb{E}\big|\psi\big(X_T^x\big)Y_T^x\big|^3 \leq \sqrt{\mathbb{E}\big|\psi\big(X_T^x\big)\big|^6}\sqrt{\mathbb{E}\big(Y_T^x\big)^6}.$$

In view of Theorem 7.9, X_T^x has finite moments of all orders, and Y_T^x inherits the same property (see Question 3).

8.1 (Importance Sampling for Stochastic Differential Equations)

1. As at the beginning of the proof of Theorem 7.14, let $u(t,x)$ be the solution to

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + \mathscr{L}u(t,x) = 0 & \text{in } [0,T)\times\mathbb{R}, \\ u(T,x) = f(x), & x\in\mathbb{R}. \end{cases} \tag{A.1}$$

Denoting the function $\frac{\partial u}{\partial x}(t,x)$ by $(\partial_x u)(t,x)$, Itô's formula implies

$$f\big(X_T(x)\big) = \mathbb{E}f\big(X_T(x)\big) + \int_0^T (\partial_x u)\big(t, X_t(x)\big)\cdot\sigma\big(X_t(x)\big)\,dW_t, \quad \text{a.s.}$$

2. Observe that

$$\begin{aligned} &\mathbb{E}\big|X_T(x) - \hat{X}_T(x)\big|^2 \\ &\quad\le 2\mathbb{E}\bigg(\int_0^T \big|b(X_s) - \hat{b}(\hat{X}_s)\big|\,ds\bigg)^2 + 2\mathbb{E}\int_0^T \big|\sigma(X_s) - \hat{\sigma}(\hat{X}_s)\big|^2\,ds \\ &\quad\le C\mathbb{E}\int_0^T \big|b(X_s) - \hat{b}(\hat{X}_s)\big|^2\,ds + 2\mathbb{E}\int_0^T \big|\sigma(X_s) - \hat{\sigma}(\hat{X}_s)\big|^2\,ds. \end{aligned}$$

In addition, as b is Lipschitz, denoting by C a constant which may change from line to line but does not depend on $\hat{b}$ nor on $\hat{\sigma}$, one gets

$$\begin{aligned} &\int_0^T \big|b(X_s) - \hat{b}(\hat{X}_s)\big|^2\,ds \\ &\quad\le 2\int_0^T \big|b(X_s) - b(\hat{X}_s)\big|^2\,ds + 2\int_0^T \big|b(\hat{X}_s) - \hat{b}(\hat{X}_s)\big|^2\,ds \\ &\quad\le C\int_0^T |X_s - \hat{X}_s|^2\,ds + C\|b-\hat{b}\|_\infty^2. \end{aligned}$$

Similar inequalities hold true when b is replaced by σ. One concludes by applying Gronwall's lemma.

3. For example, one may use

$$\frac{d}{dx}X_T(x) = \exp\bigg(\int_0^t \Big(b'(X_s) - \frac{1}{2}\big|\sigma'(X_s)\big|^2\Big)\,ds + \int_0^t \sigma'(X_s)\,dW_s\bigg),$$

a similar expression for $\frac{d}{dx}\hat{X}_T(x)$, a Taylor expansion of $e^\alpha - e^\beta$, and finally the preceding question.

4. Since f is Lipschitz, one can bound from above

$$\big|\mathbb{E}f\big(X_T(x)\big) - \mathbb{E}f\big(\hat{X}_T(x)\big)\big|^2$$

by using the second question. Then, as the function σ is bounded,

$$\mathbb{E}\left(\int_0^T \big(\phi\big(X_s(x)\big) - \hat{\phi}\big(X_s(x)\big)\big)\, dW_s\right)^2$$
$$\leq C\int_0^T \mathbb{E}\big|\partial_x u\big(s, X_s(x)\big) - \partial_x \hat{u}(s, X_s)\big|^2\, ds$$
$$+ C\|\sigma - \hat{\sigma}\|_\infty^2 \int_0^T \mathbb{E}\big|\partial_x \hat{u}(s, X_s)\big|^2\, ds.$$

To conclude, it then remains to use

$$\partial_x u(t,x) = \mathbb{E}\left(f'\big(X_{T-t}(x)\big)\frac{d}{dx}X_{T-t}(x)\right),$$

and Question 3.

5. The random variable

$$Z := f\big(X_T(x)\big) - \int_0^T (\partial_x u)\big(t, X_t(x)\big)\sigma\big(X_t(x)\big)\, dW_t$$

has the same expectation as $f(X_T(x))$ and has null variance (hence is deterministic). It is impossible to simulate since the function $u(t,x)$ is unknown. One instead chooses a proxy $(\hat{X}_t)$ of (X_t) for which the function $\mathbb{E}f(\hat{X}_T(x))$ is explicitly known, and simulates

$$\hat{Z} := f\big(\hat{X}_T(x)\big) - \int_0^T (\partial_x \hat{u})\big(t, \hat{X}_t(x)\big)\sigma\big(\hat{X}_t(x)\big)\, dW_t.$$

The preceding estimates allow one to estimate the variance reduction.

9.1 (Adaptive Monte Carlo Method)

1. (a) The increments $M_n - M_{n-1} = g(\theta_{n-1}, X_n) - \mathbb{E}(\phi(X))$ are square integrable, hence so are the M_n. Also, $\mathbb{E}(M_n - M_{n-1} \,|\, \mathscr{F}_{n-1}) = \mathbb{E}(g(\theta_{n-1}, X_n) \,|\, \mathscr{F}_{n-1}) - \mathbb{E}(\phi(X))$, and since θ_{n-1} is $\mathscr{F}_{n-1}$-measurable and X_n is independent from $\mathscr{F}_{n-1}$ and has the same law as X, it holds that $\mathbb{E}(g(\theta_{n-1}, X_n) \,|\, \mathscr{F}_{n-1}) = F(\theta_{n-1})$ for $F : \theta \in \mathbb{R} \mapsto F(\theta) = \mathbb{E}(g(\theta, X)) := \mathbb{E}(\phi(X))$.
   (b) Developing the square and using the above computations,

$$\begin{aligned}\mathbb{E}\big((M_n - M_{n-1})^2 \,|\, \mathscr{F}_{n-1}\big) &= \mathbb{E}\big(\big(g(\theta_{n-1}, X_n) - \mathbb{E}\big(\phi(X)\big)\big)^2 \,|\, \mathscr{F}_{n-1}\big)\\ &= \mathbb{E}\big(g(\theta_{n-1}, X_n)^2 \,|\, \mathscr{F}_{n-1}\big) - \mathbb{E}\big(\phi(X)\big)^2\\ &= m_2(\theta_{n-1}) - \mathbb{E}\big(\phi(X)\big)^2\end{aligned}$$

   so that $\langle M\rangle_n = \sum_{k=1}^n (m_2(\theta_{k-1}) - \mathbb{E}(\phi(X))^2)$. A simple equivalent is given by $(m_2(\theta^*) - \mathbb{E}(\phi(X))^2)n = \sigma^2 n$.
   (c) Using (1) of the admitted classical convergence result, $\lim_{n\to\infty} \frac{M_n}{n} = 0$ a.s.

2. Using (2) of the classical convergence result, $\lim_{n\to\infty} \frac{M_n}{\sigma\sqrt{n}} = \mathcal{N}(0,1)$ in law.
3. (a) Same arguments as for $(M_n)_{n\geq 0}$.
   (b) Same arguments as for $(M_n)_{n\geq 0}$ yield $\langle \bar{M}\rangle_n = \sum_{k=1}^n (m_4(\theta_{k-1}) - m_2(\theta_{k-1})^2)$. A simple equivalent is $(m_4(\theta^*) - m_2(\theta^*)^2)n$.
   (c) The natural estimator is (with perhaps a factor $\frac{n}{n-1}$)

$$\frac{1}{n}\sum_{k=1}^n g(\theta_{k-1}, X_k)^2 - \left(\sum_{k=1}^n g(\theta_{k-1}, X_k)\right)^2.$$

The preceding results yield that $\frac{1}{n}\sum_{k=1}^n g(\theta_{k-1}, X_k)^2$ and $\frac{1}{n}\sum_{k=1}^n m_2(\theta_{k-1})$ converge to the same limit, a.s. Clearly $\frac{1}{n}\sum_{k=1}^n m_2(\theta_{k-1})$ converges to $m_2(\theta^*)$. Hence, this estimator converges to $\sigma^2 = m_2(\theta^*) - \mathbb{E}(\phi(X))^2$, a.s.

References

1. Anderson, W.J.: Continuous-Time Markov Chains. Springer Series in Statistics: Probability and Its Applications. Springer, New York (1991)
2. Ané, C., Blachère, S., Chafaï, D., Fougères, P., Gentil, I., Malrieu, F., Roberto, C., Scheffer, G.: Sur les inégalités de Sobolev logarithmiques. Panoramas et Synthèses, vol. 10. Société Mathématique de France, Paris (2000)
3. Arouna, B.: Robbins-Monro algorithms and variance reduction in finance. J. Comput. Finance **7**(2), 35–61 (2003)
4. Asmussen, S., Glynn, P.W.: Stochastic Simulation: Algorithms and Analysis. Stochastic Modelling and Applied Probability, vol. 57. Springer, New York (2007)
5. Benaïm, M., El Karoui, N.: Chaînes de Markov et simulations; Martingales et stratégies. Éditions de l'École Polytechnique, Palaiseau (2004)
6. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley Series in Probability and Statistics: Probability and Statistics. Wiley, New York (1999)
7. Cercignani, C., Illner, R., Pulvirenti, M.: The Mathematical Theory of Dilute Gases. Applied Mathematical Sciences, vol. 106. Springer, New York (1994)
8. Clark, J.M.C., Cameron, R.J.: The maximum rate of convergence of discrete approximations for stochastic differential equations. In: Stochastic Differential Systems. Lecture Notes in Control and Information Sciences, vol. 25, pp. 162–171. Springer, Berlin (1980)
9. Davis, M.H.A.: Markov Models and Optimization. Monographs on Statistics and Applied Probability, vol. 49. Chapman & Hall, London (1993)
10. Devroye, L.: Nonuniform Random Variate Generation. Springer, New York (1986)
11. Duffie, D., Glynn, P.: Efficient Monte Carlo simulation of security prices. Ann. Appl. Probab. **5**(4), 897–905 (1995)
12. Duflo, M.: Algorithmes Stochastiques. Mathématiques & Applications, vol. 23. Springer, Berlin (1996)
13. Duflo, M.: Random Iterative Models. Applications of Mathematics, vol. 34. Springer, Berlin (1997)
14. Faure, O.: Simulation du mouvement brownien et des diffusions. Ph.D. thesis, École Nationale des Ponts et Chaussées (1992)
15. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. I, 3rd edn. Wiley, New York (1968)
16. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. II, 2nd edn. Wiley, New York (1971)
17. Fournié, E., Lasry, J.M., Lebuchoux, J., Lions, P.L.: Applications of Malliavin calculus to Monte-Carlo methods in finance. II. Finance Stoch. **5**(2), 201–236 (2001)
18. Fournié, E., Lasry, J.M., Lebuchoux, J., Lions, P.L., Touzi, N.: Applications of Malliavin calculus to Monte Carlo methods in finance. Finance Stoch. **3**(4), 391–412 (1999)

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1, © Springer-Verlag Berlin Heidelberg 2013

19. Gentle, J.E.: Random Number Generation and Monte Carlo Methods, 2nd edn. Statistics and Computing. Springer, New York (2003)
20. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Applications of Mathematics, Stochastic Modelling and Applied Probability, vol. 53. Springer, New York (2004)
21. Graham, C., Kurtz, T.G., Méléard, S., Protter, P.E., Pulvirenti, M., Talay, D.: In: Talay, D., Tubaro, L. (eds.) Probabilistic Models for Nonlinear Partial Differential Equations, Lectures given at the 1st Session and Summer School Held in Montecatini Terme, May 22–30, 1995. Lecture Notes in Mathematics, vol. 1627. Springer, Berlin (1996). Fondazione C.I.M.E. [C.I.M.E. Foundation]
22. Graham, C., Méléard, S.: Stochastic particle approximations for generalized Boltzmann models and convergence estimates. Ann. Probab. **25**(1), 115–132 (1997)
23. Ikeda, N., Watanabe, S.: Stochastic Differential Equations and Diffusion Processes, 2nd edn. North-Holland Mathematical Library, vol. 24. North-Holland, Amsterdam (1989)
24. Jacod, J., Protter, P.: Probability Essentials, 2nd edn. Universitext. Springer, Berlin (2004)
25. Kanagawa, S.: On the rate of convergence for Maruyama's approximate solutions of stochastic differential equations. Yokohama Math. J. **36**(1) (1988)
26. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus, 2nd edn. Graduate Texts in Mathematics, vol. 113. Springer, New York (1991)
27. Kebaier, A.: Statistical Romberg extrapolation: A new variance reduction method and applications to option pricing. Ann. Appl. Probab. **15**(4), 2681–2705 (2005)
28. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Applications of Mathematics, vol. 23. Springer, Berlin (1992)
29. Kuipers, L., Niederreiter, H.: Uniform Distribution of Sequences. Pure and Applied Mathematics. Wiley-Interscience, New York (1974)
30. Kunita, H.: Stochastic differential equations and stochastic flows of diffeomorphisms. In: École d'été de probabilités de Saint-Flour, XII—1982. Lecture Notes in Math., vol. 1097, pp. 143–303. Springer, Berlin (1984)
31. Kushner, H.J., Yin, G.G.: Stochastic Approximation Algorithms and Applications. Applications of Mathematics, vol. 35. Springer, New York (1997)
32. Lapeyre, B., Pardoux, É., Sentis, R.: Introduction to Monte-Carlo Methods for Transport and Diffusion Equations. Oxford Texts in Applied and Engineering Mathematics, vol. 6. Oxford University Press, Oxford (2003). Translated from the 1998 French original by Alan Craig and Fionn Craig
33. L'Ecuyer, P.: Random number generation. In: Handbook of Computational Statistics, pp. 35–70. Springer, Berlin (2004)
34. L'Ecuyer, P., Hellekalek, P.: Random number generators: Selection criteria and testing. In: Random and Quasi-random Point Sets. Lecture Notes in Statist., vol. 138, pp. 223–265. Springer, New York (1998)
35. L'Ecuyer, P., Simard, R.: TestU01: A C library for empirical testing of random number generators. ACM Trans. Math. Softw. **33**(4), 40 (2007), Art. 22. http://dx.doi.org/10.1145/1268776.1268777
36. Milstein, G.N.: A method with second order accuracy for the integration of stochastic differential equations. Teor. Verojatnost. i Primenen. **23**(2), 414–419 (1978)
37. Milstein, G.N.: Numerical Integration of Stochastic Differential Equations. Mathematics and Its Applications, vol. 313. Kluwer Academic, Dordrecht (1995). Translated and revised from the 1988 Russian original
38. Milstein, G.N., Tretyakov, M.V.: Stochastic Numerics for Mathematical Physics. Scientific Computation. Springer, Berlin (2004)
39. Newton, N.J.: Variance reduction for simulated diffusions. SIAM J. Appl. Math. **54**(6), 1780–1805 (1994)
40. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 63. Society for Industrial and Applied Mathematics, Philadelphia (1992)

41. Niederreiter, H., Shparlinski, I.E.: Recent advances in the theory of nonlinear pseudorandom number generators. In: Monte Carlo and Quasi-Monte Carlo methods, Hong Kong, 2000, pp. 86–102. Springer, Berlin (2002)
42. Nualart, D.: Malliavin Calculus and Its Applications. CBMS Regional Conference Series in Mathematics, vol. 110. Conference Board of the Mathematical Sciences, Washington (2009)
43. Petrov, V.V.: Sums of Independent Random Variables. Springer, New York (1975). Translated from the Russian by A.A. Brown, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82
44. Shiryayev, A.N.: Probability. Graduate Texts in Mathematics, vol. 95. Springer, New York (1984). Translated from the Russian by R.P. Boas
45. Sznitman, A.S.: Topics in propagation of chaos. In: Hennequin, P.L. (ed.) Ecole d'été de Probabilités de Saint-Flour XIX–1989. Lecture Notes in Mathematics, vol. 1464, Chap. 3, pp. 165–251. Springer, Berlin (1991)
46. Talay, D., Tubaro, L.: Expansion of the global error for numerical schemes solving stochastic differential equations. Stoch. Anal. Appl. **8**(4), 483–509 (1990)
47. Williams, D.: Probability with Martingales. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge (1991)

Index

C. Graham, D. Talay, *Stochastic Simulation and Monte Carlo Methods*,
Stochastic Modelling and Applied Probability 68,
DOI 10.1007/978-3-642-39363-1, © Springer-Verlag Berlin Heidelberg 2013