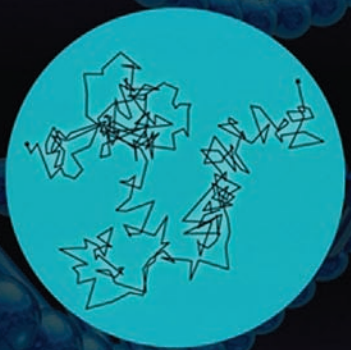
DUNOD

WILEY

Markov Processes and Applications

Algorithms, Networks, Genome and Finance



Etienne Pardoux

WILEY SERIES IN PROBABILITY AND STATISTICS

Markov Processes and Applications

Algorithms, Networks, Genome and Finance

Étienne Pardoux

Laboratoire d'Analyse, Topologie, Probabilités
Centre de Mathématiques et d'Informatique
Université de Provence, Marseille, France.

A John Wiley and Sons, Ltd., Publication

This work is in the Wiley-Dunod Series co-published between Dunod and John Wiley & Sons, Ltd.

Markov Processes and Applications

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by WALTER A. SHEWHART and SAMUEL S. WILKS
Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Jozef L. Teugels*

A complete list of titles in this series appears at the end of the volume.

This work is in the Wiley-Dunod Series co-published between Dunod and John Wiley & Sons, Ltd.

This edition first published 2008

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Pardoux, E. (Etienne), 1947-
Markov processes and applications : algorithms, networks, genome, and finance / Etienne Pardoux.
p. cm. – (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 978-0-470-77271-3 (cloth)
1. Markov processes. I. Title.
QA274.7.P375 2008
519.2′33–dc22

2008043729

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-77271-3

Typeset in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall

Contents

Notations

The following notations will be used throughout this book.
IN = {0, 1, 2,. . .} stands for the set of positive integers, including 0.
IN* = {1, 2,. . .} stands for the set of positive integers, 0 excluded.

PREFACE

The core parts of this book are Chapter 1 on Monte Carlo methods, Chapter 2 on discrete time Markov chains with values in a finite or countable set, and Chapters 6 and 7 on the Poisson process and continuous time jump Markov processes, likewise with values in a finite or countable set.

With these chapters are their starting point, this book presents applications in several fields. Chapter 3 deals with stochastic algorithms. Specifically, we present the Markov chain Monte Carlo method invented in the 1950s for applications in statistical mechanics. Used in image processing, it has become an essential algorithm in Bayesian statistics when the data to hand are complex and numerous. We also present a stochastic optimization algorithm, namely simulated annealing.

Another application concerns molecular biology, with two distinct examples. One, presented in Chapter 4, concerns the annotation of DNA sequences and sequence alignment. The main tools here are hidden Markov models, which are also very commonly used in signal processing, in particular in speech recognition. A second biological application is concerned with phylogeny, which is the study of the relations between living species, those relations being illustrated by the 'phylogenetic tree'. Several phylogenetic tree reconstruction methods are based on probabilistic models of evolution of genomic sequences. These models are continuous time jump Markov processes on trees. This application is in Chapter 7.

Chapter 5 presents an introduction to control and filtering, including the famous Kalman–Bucy filter, an algorithm which is frequently used for guiding satellites.

The subject of Chapter 8 is queues and networks.

Finally, Chapter 9 gives an introduction to financial mathematics. It presents both discrete and continuous time models. In particular, it contains a presentation of Itô's stochastic calculus and diffusion processes, which again are Markov processes, but this time both in continuous time and taking their values in the Euclidian space $\mathbb{R}^d$. Note that this chapter is the only one where several proofs of basic results are omitted. Including them would have made the book too long, and they are available elsewhere.

Each chapter is followed by a number of exercises. Some of these bear the label 'Programming'. This means that they suggest simulations, for example with Matlab, in most cases with the idea of visualizing the results graphically. Solutions to more than half of the exercises are given in Chapter 10. Students are urged to try to solve the exercises by themselves, without immediate recourse to the solutions. This is

essential for mastering the content of the book. While most exercises are designed for understanding the content of the book, a few present additional applications.

The content of this book was taught in several courses both at the Université de Provence in Marseille, and at the École Centrale de Marseille.

A complete reading of this book (including Chapter 9) requires reader to have a knowledge of probability theory, including measure theory and conditional expectation. However, most of the book uses only discrete laws, together with some laws with density, and the two basic limit theorems: the law of large numbers and the central limit theorem. Hence, large parts of the book will be accessible to mathematicians who have only studied probability at undergraduate level, as well as by computer scientists, statisticians, economists, physicists, biologists and engineers.

I am grateful to Geneviève Foissac, who typed most of the French version of this book, and my colleagues Julien Berestycki, Fabienne Castell, Yuri Golubev, Arnaud Guillin, Stéphanie Leocard, Laurent Miclo and Rémi Rhodes, who read parts of the manuscript and whose comments and criticisms helped me improve the original version.

This is a translation of the original French version *Processus de Markov et applications: algorithmes, réseaux, génome et finance*, published by Dunod in 2007. I have added one section (Section 2.8). I wish to thank Judith R. Miller, who kindly read and improved my English translation. She could not of course make my English perfect, but thanks to her I hope it is readable.

Marseille

1

Simulations and the Monte Carlo method

Introduction

In order to introduce the Monte Carlo method, let us consider a problem of numerical integration. There exist several numerical methods for the approximate computation of the integral

$$\int_{[0,1]} f(x)dx,$$

based on formulae of the type $\sum_{i=1}^{n} w_i f(x_i)$, where the w_i are positive numbers whose sum equals 1 and the x_i are points in the interval $[0, 1]$. For example, if $w_i = 1/n$, $1 \leq i \leq n$, and $x_i = i/n$, this is the trapezoid rule. But there exist other approximations, such as Simpson's rule and the Gaussian quadrature formula. A Monte Carlo method is of the same type: we choose $w_i = 1/n$, and we choose the x_i 'at random' (meaning here according to the uniform law on $[0, 1]$, later denoted by $\mathcal{U}(0, 1)$). As we shall see below, the convergence is guaranteed by the law of large numbers, and the rate of convergence, of order $n^{-1/2}$, is given by the central limit theorem. Clearly, that rate of convergence may seem rather slow, if we compare it with the rate of other numerical integration methods in dimension 1. But all these numerical methods collapse if we go to higher dimensions. Indeed, in all these methods, the precision is a function of the distance between two contiguous points of the discretization. But if we use n points for the discretization of $[0, 1]^d$, the distance between two contiguous points is of order $n^{-1/d}$, hence if we want a precision of order $1/n$ with a 'first-order' method of approximation of an integral over $[0, 1]^d$, the number of points we need is of order n^d. On the other hand, the Monte Carlo method is essentially unaffected by the dimension.

Historically, the method goes back to Count Buffon who described in 1777 a method for the approximate computation of π, based on the realization of repeated

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

experiments. But the true birth of the Monte Carlo method is linked to the appearance of the first computers. The first papers describing methods of this type date back from the late 1940s and early 1950s. These methods continue to grow more and more popular. This is in large part due to the simplicity with which one can program them, as well as the ability of today's computers to perform a huge number of random draws in a reasonable length of time.

1.1 Description of the method

If we wish to use a Monte Carlo method, we need first to write the quantity of interest as the expectation of a random variable. This is often easy, as in the case of the computation of an integral, but it might be much more involved, as when we wish to solve a parabolic or elliptic partial differential equation (see Sections 7.9 and 9.3 below).

The next step is to compute a quantity of the form $\mathbb{E}(X)$, where X is a random variable. In order do so, we need to be able to simulate mutually independent random variables $X_1, \ldots, X_n$, all having the law of X. It then remains to approximate $\mathbb{E}(X)$ by

$$\mathbb{E}(X) \approx \frac{1}{n}(X_1 + \ldots + X_n).$$

Let us describe one example of the application of the Monte Carlo method, to the computation of an integral. We will explain in detail the two steps presented above: how to write the integral as an expectation, and how to simulate the random variables. Suppose that we wish to compute an integral of the form

$$I = \int_{[0,1]^d} f(u_1, \ldots, u_d) du_1 \ldots du_d.$$

We set $X = f(U_1, \ldots, U_d)$, where the U_i, $i = 1, \ldots, d$, are independent and identically distributed (i.i.d.) random variables, each one having the law $\mathcal{U}(0, 1)$. We have

$$\mathbb{E}(X) = \mathbb{E}\left(f(U_1, \ldots, U_d)\right) = \int_{[0,1]^d} f(u_1, \ldots, u_d) du_1 \ldots du_d.$$

We have just completed the first step – our integral is written as an expectation.

For the simulation, suppose we can produce a sequence $(U_i, i \geq 1)$ of i.i.d. random variables whose common law is $\mathcal{U}(0, 1)$. We define $X_1 = f(U_1, \ldots, U_d)$, $X_2 = f(U_{d+1}, \ldots, U_{2d})$, etc. Then the sequence $(X_i, i \geq 1)$ is an i.i.d. sequence of random variables, all having the same law as X. We can now implement the Monte Carlo method.

It is important to note the simplicity with which the corresponding program can be written. Note also that no specific regularity of f is required. f need only be integrable.

One often needs to compute a more general type of integral, namely

$$I = \int_{\mathbb{R}^d} g(x) f(x) dx = \int_{\mathbb{R}^d} g(x_1, \ldots, x_d) f(x_1, \ldots, x_d) dx_1 \ldots dx_d,$$

with $f(x)$ non-negative and $\int f(x)dx = 1$. Then I equals $\mathbb{E}(g(X))$ if X is an $\mathbb{R}^d$-valued random variable whose law is $f(x)dx$. The problem now is to simulate random vectors having that probability law. Some answers, related to commonly used probability laws, will be given in Section 1.3 below.

But let us first answer the two questions:

- When and why does this algorithm converge?
- Can we get a precise idea of the accuracy of this algorithm?

1.2 Convergence theorems

The answers to the two above questions are given by the two most fundamental theorems in the calculus of probability, namely the law of large numbers, which permits us to establish the convergence of the method, and the central limit theorem, which gives a precise indication of its rate of convergence.

Theorem 2.1 *Let $(X_n, n \geq 1)$ be a sequence of i.i.d. random variables, all having the law of X. If $\mathbb{E}(|X|) < +\infty$, then, for $\mathbb{P}$ almost all ω (this means that there exists $N \subset \Omega$, with $\mathbb{P}(N) = 0$ and such that whenever $\omega \notin N$),*

$$\mathbb{E}(X) = \lim_{n \to +\infty} \frac{1}{n}(X_1 + \ldots + X_n)(\omega).$$

The evaluation of the method relies upon estimating the error

$$\varepsilon_n = \mathbb{E}(X) - \frac{1}{n}(X_1 + \ldots + X_n).$$

The central limit theorem gives the asymptotic behaviour of the quantity ε_n, which has a random nature. It says that the law of ε_n tends to look like a centred Gaussian law.

Theorem 2.2 *Let $(X_n, n \geq 1)$ be a sequence of i.i.d. random variables, all having the law of X. Assume that $\mathbb{E}(X^2) < +\infty$. Let σ^2 denote the variance of X:*

$$\sigma^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}\left((X - \mathbb{E}(X))^2\right).$$

Then

$$\frac{\sqrt{n}}{\sigma}\varepsilon_n \text{ converges in law towards } Z \simeq N(0, 1).$$

In other words, for all $a < b$,

$$\lim_{n \to +\infty} \mathbb{P}\left(\frac{\sigma}{\sqrt{n}}a \leq \varepsilon_n \leq \frac{\sigma}{\sqrt{n}}b\right) = \int_a^b e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

In practice, if n is not too small (which will always be the case in the situation of a Monte Carlo computation), the above probability can be replaced by its limit, hence we may act as if ε_n were a centred Gaussian random variable with variance σ^2/n.

Remark 2.3 *1. This result is extremely powerful, since it gives us a rate of convergence which can be easily estimated with the help of the simulations which have already been realized. The fact that we have a reliable estimate of the error, without any further computation, is a real strength of the method.*

2. However, the central limit theorem never provides a bound for the error, since the support of a Gaussian random variable is $\mathbb{R}$*. One way to describe the error in the Monte Carlo method is either by providing the standard deviation of* ε_n*, which is equal to* $\sigma/\sqrt{n}$*, or else by providing a* 95% *confidence interval for the result. This means that there is a* 0.95 *chance that the quantity of interest is in the given interval (and hence there is a* 0.05 *chance that it is outside that interval). Clearly* 0.95 *can be replaced buy any value close to* 1.

Note the important role played by the variance of X in the estimation of the error. Since we can choose the law of X, with the restriction that $\mathbb{E}(X)$ be the quantity which we are interested in, we may wish to replace X by another random variable with the same expectation and a smaller variance. Such a procedure is called a variance reduction method (see Section 1.4 below).

We should also note that the rate at which the error goes to 0 is not very fast. However, there are several situations where this slowly converging method is the only available one (e.g. integral or parabolic partial differential equations in dimension higher than 4). It is also remarkable that the rate of convergence does not depend upon the smoothness of the function f.

We will now describe the use of the central limit theorem for analysing the rate of convergence of the Monte Carlo method, in two examples. This will allow us to present a limitation of the use of the Monte Carlo method.

A good case Suppose we wish to compute $p = \mathbb{P}(Y \leq \lambda)$, where Y is a random variable with an arbitrary law. Define $X = \mathbf{1}_{\{Y \leq \lambda\}}$. Then $\mathbb{E}(X) = p$, and $\sigma^2 = \mathrm{var}(X) = p(1-p)$. Consequently, after n independent draws $X_1, \ldots, X_n$ of X, we have

$$p_n = \frac{X_1 + \ldots + X_n}{n} \approx p + \frac{\sigma}{\sqrt{n}} Z.$$

Since $p(1-p) \leq 1/4$, if we want the standard deviation $\sigma/\sqrt{n}$ of the error to be of the order of 0.01, we should choose n of the order of 2500. If we choose $n = 2500$, the 0.95 confidence interval for p is then, according to the central limit theorem, $[p_n - 1.96 \times 0.01, p_n + 1.96 \times 0.01]$. If the true unknown value p is of the order of 0.50, this leads to an acceptable error.

However, if the true value of p is very small, the above value of n may be insufficient, if we want the error to be smaller than the quantity to be estimated. We need a number of simulations of the order of $1/p$.

A tough case Imagine that we wish to compute $\mathbb{E}(\exp(\beta Z))$, where Z is an $N(0, 1)$ random variable. Clearly

$$E = \mathbb{E}\left(e^{\beta Z}\right) = e^{\beta^2/2}.$$

If we apply a Monte Carlo method to this case, we let $X = e^{\beta Z}$. The variance of X is $\sigma^2 = e^{2\beta^2} - e^{\beta^2}$. After n simulations $X_1, \ldots, X_n$ according to the law of X, we have

$$E_n = \frac{X_1 + \ldots + X_n}{n} \approx E + \frac{\sigma}{\sqrt{n}} Z.$$

The standard deviation of the relative error is

$$\frac{\sigma}{E\sqrt{n}} = \sqrt{\frac{e^{\beta^2} - 1}{n}}.$$

If we want that quantity to be smaller than a given $\varepsilon > 0$, then we should choose $n \approx \varepsilon^{-2}(e^{\beta^2} - 1)$. If $\varepsilon = 1$ and $\beta = 5$, this means $n = 7 \times 10^{10}$, which is far too high. After 10^5 simulations, the 0.95 confidence interval might be $[-467\,647, 2\,176\,181]$, which is a disaster. The only positive point is that we are aware of the fact that our estimate is terrible, at least if we have a good estimate of the variance of the X_n.

This example shows a practical limitation of the Monte Carlo method, when we use random variables with large variances. This leads us to formulate the following rule: in any Monte Carlo computation, one must exploit the simulations, in order to estimate the variance of the random variable whose expectation we wish to compute.

Note that reducing the variance of the random variable to be simulated is often a crucial step in making a Monte Carlo computation efficient. We shall discuss this issue in Section 1.4.

1.3 Simulation of random variables

Simulation of $\mathcal{U}(0, 1)$ Any programming language today possesses a *pseudo-random* number generator. Such a program produces as output a perfectly deterministic (and also periodic) sequence, but whose statistical properties resemble those of a sequence of independent realizations of the law $\mathcal{U}(0, 1)$. The problem of inventing a good 'random number generator' is to create a recurrence formula which, in a reasonable time, produces a sequence of numbers which looks as much as possible like a sequence of realizations of independent $\mathcal{U}(0, 1)$ random variables, with a period which should be as large as possible. The study of those generators is part of the theory of dynamical systems. Most classical algorithms generating pseudo-random numbers are presented in [23] and [32], among others. More recently, Matsumoto and Nishimura [26] proposed a generator with period $2^{19\,937} - 1$!

Note that all random number generators try in fact to deliver draws from a uniform law on $\{1/M, 2/M, \ldots, (M-1)/M, 1\}$, with M very, very large.

It remains to simulate laws other than the uniform law.

Simulation of a Bernoulli random variable Let $0 < p < 1$. If U is a $\mathcal{U}(0, 1)$ random variable, $X = \mathbf{1}_{\{U \leq p\}}$ is a Bernoulli random variable with parameter p.

Simulation of a binomial random variable If $U_1, \ldots, U_n$ are independent $\mathcal{U}(0, 1)$ random variables, then

$$X = \mathbf{1}_{\{U_1 \leq p\}} + \ldots + \mathbf{1}_{\{U_n \leq p\}}$$

is a $B(n, p)$ random variable (binomial with parameters n and p).

Simulation of a geometric random variable $X = \inf\{k \geq 1;\ U_k \leq p\}$ is a geometric random variable with parameter p. A more efficient simulation procedure, based on the next lemma, is proposed in Exercise 5.1.

Inversion of the distribution function Recall the following classical result:

Lemma 3.1 *Let X be a random variable, and F its distribution function (i.e. $F(x) = \mathbb{P}(X \leq x)$). Define, for $0 \leq t \leq 1$,*

$$F^{-1}(t) = \inf\{x;\ F(x) > t\}.$$

Then if U has the law $\mathcal{U}[0, 1]$, $F^{-1}(U)$ has the same law as X.

PROOF This is immediate:

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

Indeed, $\{t; F^{-1}(t) \leq x\} \subset \{t; t \leq F(x)\}$, and the difference between those two sets is at most a one point set.

□

This method can be used whenever we have an explicit expression for the inverse of F. This is particularly the case for the exponential probability law.

Simulation of an exponential random variable Recall that a random variable X has the exponential law with parameter λ whenever, for all $t \in \mathbb{R}_+$,

$$\mathbb{P}(X > t) = \exp(-\lambda t).$$

Hence, if F is the distribution function of X, $F(t) = 1 - e^{-\lambda t}$, and

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

If $U \simeq \mathcal{U}[0, 1]$, the same is true with $1 - U$, and

$$-\frac{\log U}{\lambda} \simeq \mathcal{E}(\lambda).$$

Simulation of Gaussian random variables (Box–Müller algorithm) A classical method for the simulation of Gaussian random variables is based on the remark that, if U and V are two independent $\mathcal{U}(0, 1)$ random variables,

$$\sqrt{-2\log(U)}\cos(2\pi V) \quad \text{and} \quad \sqrt{-2\log(U)}\sin(2\pi V)$$

are independent $N(0, 1)$ random variables. One can check this result as follows. If X and Y are independent $N(0, 1)$ random variables, $f : \mathbb{R}^2 \to \mathbb{R}_+$,

$$\begin{aligned}
\mathbb{E}f(X, Y) &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \exp\left(-\frac{x^2+y^2}{2}\right) f(x, y)dxdy \\
&= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} r\exp\left(-\frac{r^2}{2}\right) f(r\cos\theta, r\sin\theta)drd\theta \\
&= \int_0^1\int_0^1 f\left(\sqrt{-2\log u}\cos(2\pi v), \sqrt{-2\log u}\sin(2\pi v)\right)dudv \\
&= \mathbb{E}f\left(\sqrt{-2\log U}\cos(2\pi V), \sqrt{-2\log U}\sin(2\pi V)\right).
\end{aligned}$$

For the simulation of a Gaussian random variable with mean μ and variance σ^2, it suffices to define $X = \mu + \sigma Y$, where $Y \simeq N(0, 1)$.

Simulation of a Poisson random variable A Poisson random variable with parameter λ is an $\mathbb{N}$-valued random variable such that

$$\mathbb{P}(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}, \quad \text{for} \quad n \geq 0.$$

We shall see in Chapter 6 that whenever $\{T_i; i \geq 1\}$ is a sequence of i.i.d. random variables, all being exponential with parameter λ, then the law of $N_t = \sum_{n\geq 1} n\mathbf{1}_{\{T_1+\ldots+T_n\leq t<T_1+\ldots+T_{n+1}\}}$ is Poisson with parameter λt. Hence N_1 has the law which we want to simulate. On the other hand, any exponential random variable T_i can be written in the form $-\log(U_i)/\lambda$, where the $(U_i)_{i\geq 1}$ are mutually independent $\mathcal{U}(0, 1)$ random variables. Hence N_1 can be written

$$N_1 = \sum_{n\geq 1} n\mathbf{1}_{\{U_1U_2\cdots U_{n+1}<e^{-\lambda}\leq U_1U_2\cdots U_n\}}.$$

This gives an algorithm for the simulation of Poisson random variables.

The rejection method Suppose we wish to simulate a random variable with density f (e.g. with respect to Lebesgue measure on $\mathbb{R}^d$), and suppose that there is an easily simulable density g, such that, for all $x \in \mathbb{R}^d$,

$$f(x) \leq k\, g(x), \quad g(x) > 0 \Leftrightarrow f(x) > 0,$$

where k is a real constant. Define

$$\alpha(x) = \frac{f(x)}{k\, g(x)}$$

on the set $\{g(x) > 0\}$.

Proposition 3.2 *Let $(X_n, U_n)_{n\geq 1}$ be a sequence of independent random vectors where, for each $n \geq 1$, X_n and U_n are independent, X_n has the density g and $U_n \simeq \mathcal{U}(0,1)$. Let $N = \inf\{k \geq 1;\ U_k \leq \alpha(X_k)\}$ and $X = X_N$. The random variable X has the density f.*

Remark 3.3 *1. The probability of acceptance at the first step is*

$$\begin{aligned} p_1 &= \mathbb{P}(U_1 \leq \alpha(X_1)) \\ &= \int \mathbb{P}(U_1 \leq \alpha(x))\mathbb{P}_{X_1}(dx) \\ &= \int \alpha(x)g(x)dx \\ &= \frac{1}{k}, \end{aligned}$$

since U_1 and X_1 are independent.

If we wish to reduce the number of rejections while simulating X, we need to maximize the acceptance probability p_1, hence to minimize k. Given that f and g are probability densities and that $f \leq kg$, necessarily $k \geq 1$. Note that the number of rejections is limited if $f(x)/kg(x)$ is close to 1, that is, if the function g is similar to f.

2. *The above algorithm is still valid if X has a density f with respect to an arbitrary positive measure μ, which is bounded from above by kg, where g is the density with respect to μ of an easily simulable random variable Y. In other words,*

$$\mathbb{P}(X \in A) = \int_A f(x)\mu(dx) \leq \int_A kg(x)\mu(dx) = k\mathbb{P}(Y \in A).$$

If the law of X is supported by a discrete set E, we can choose for μ the counting measure of the points of E. The rejection method can be used for laws on a discrete set. In this case, $f(x) = \mathbb{P}(X = x)$.

Proof of Proposition 3.2 Note that the inequality $U_k \leq \alpha(X_k)$ will be satisfied after a finite number of steps. Indeed,

$$\begin{aligned} \mathbb{P}(\forall k \geq 1,\ X \neq X_k) &= \lim_{n\to\infty} \mathbb{P}(\cap_{k\leq n}\{X \neq X_k\}) \\ &= \lim_{n\to\infty} \mathbb{P}(\cap_{k\leq n}\{U_k > \alpha(X_k)\}) \\ &= \lim_{n\to\infty} \mathbb{P}(U_1 > \alpha(X_1))^n \\ &= \lim_{n\to\infty} (1 - p_1)^n = 0, \end{aligned}$$

since the random variables (X_k, U_k) are i.i.d. Consequently,

$$\begin{aligned}
\mathbb{P}[X \in A] &= \sum_{n\geq 1} \mathbb{P}[N = n, X \in A] \\
&= \sum_{n\geq 1} \mathbb{P}[\cap_{k\leq n-1}\{U_k > \alpha(X_k)\} \cap \{U_n \leq \alpha(X_n)\} \cap \{X_n \in A\}] \\
&= \sum_{n\geq 1} (1-p_1)^{n-1} \mathbb{P}[\{U_1 \leq \alpha(X_1)\} \cap \{X_1 \in A\}] \\
&= \frac{1}{p_1} \mathbb{P}[\{U_1 \leq \alpha(X_1)\} \cap \{X_1 \in A\}] \\
&= \mathbb{P}[X_1 \in A | U_1 \leq \alpha(X_1)].
\end{aligned}$$

The law of X is then the law of X_1, conditioned upon the acceptation set $\{U_1 \leq \alpha(X_1)\}$. From the independence of X_1 and U_1,

$$\begin{aligned}
\mathbb{P}[X \in A] &= \frac{1}{p_1} \int_A \mathbb{P}(U_1 \leq \alpha(x)) P_{X_1}(dx) \\
&= k \int_A \alpha(x) g(x) dx \\
&= \int_A f(x) dx.
\end{aligned}$$

□

For the simulation of other laws, or other simulation methods of the above laws, one can consult, among others, [7], [8], [13] and [35].

1.4 Variance reduction techniques

We have seen that the rate of convergence of the Monte Carlo method is of order $\sigma/\sqrt{n}$. Clearly, the convergence is accelerated if the variance is reduced. We now present several variance reduction methods.

Importance sampling Suppose that we try to compute $\mathbb{E}(g(X))$, where the law of X is $f(x)dx$ (on $\mathbb{R}$, for the sake of argument). We have

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) f(x) dx.$$

But if $\tilde{f}$ is the density of a probability such that $\tilde{f} > 0$, then one can rewrite $\mathbb{E}(g(X))$ as

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} \frac{g(x) f(x)}{\tilde{f}(x)} \tilde{f}(x) dx.$$

This means that $\mathbb{E}(g(X)) = \mathbb{E}\left(g(Y)f(Y)/\tilde{f}(Y)\right)$, where Y has the law $\tilde{f}(x)dx$. Hence, there is another method for computing $\mathbb{E}(g(X))$, using n simulations $Y_1, \ldots, Y_n$ of Y, and approximating $\mathbb{E}(g(X))$ by

$$\frac{1}{n}\left(\frac{g(Y_1)f(Y_1)}{\tilde{f}(Y_1)} + \ldots + \frac{g(Y_n)f(Y_n)}{\tilde{f}(Y_n)}\right).$$

If we let $Z = g(Y)f(Y)/\tilde{f}(Y)$, then this alternative method improves the convergence provided $\mathrm{var}(Z) < \mathrm{var}(g(X))$. It is easy to compute the variance of Z:

$$\mathrm{var}(Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2 = \int_{\mathbb{R}} \frac{g^2(x)f^2(x)}{\tilde{f}(x)}dx - \mathbb{E}(g(X))^2.$$

If $g(x) \geq 0$, it is easy to see that choosing $\tilde{f}(x) = g(x)f(x)/\mathbb{E}g(X)$ makes $\mathrm{var}(Z) = 0$. Of course, this relies on the fact that we can compute $\mathbb{E}(g(X))$ exactly.

This justifies the following heuristic: choose $\tilde{f}(x)$ as close as possible to $|g(x)f(x)|$, then normalize (divide by $\int \tilde{f}(x)dx$) so as to obtain a density of an easily simulable probability law. Of course, these constraints are largely contradictory.

Let us give one simple example. Suppose that we seek to compute

$$\int_0^1 \cos(\pi x/2)\, dx.$$

Let us replace the function cos by a polynomial of degree 2. Since the integrand is even and equals 0 at $x = 1$ and 1 at $x = 0$, it is natural to choose $\tilde{f}(x)$ of the form $\lambda(1 - x^2)$. If we normalize, we get $\tilde{f}(x) = 3(1 - x^2)/2$. If we compute the variances, we can verify that the method has reduced the variance by a factor of 100.

Control variate This method involves writing $\mathbb{E}(f(X))$ in the form

$$\mathbb{E}(f(X)) = \mathbb{E}(f(X) - h(X)) + \mathbb{E}(h(X)),$$

where $\mathbb{E}(h(X))$ can be explicitly computed, and $\mathrm{var}(f(X) - h(X))$ is significantly smaller than $\mathrm{var}(f(X))$. We then use a Monte Carlo method for the computation of $\mathbb{E}(f(X) - h(X))$ and a direct computation for $\mathbb{E}(h(X))$.

Let us start with a simple example. Suppose we wish to compute $\int_0^1 e^x dx$. Since near $x = 0$, $e^x \approx 1 + x$, we can write

$$\int_0^1 e^x dx = \int_0^1 (e^x - 1 - x)dx + \frac{3}{2}.$$

It is easy to see that the variance is significantly reduced.

In applications to finance (see Chapter 9), one needs to evaluate quantities of the type

$$C = \mathbb{E}\left(\left(e^{\sigma Z} - K\right)_+\right), \tag{1.1}$$

where Z is standard normal random variable and $x_+ = \max(0, x)$. Such a quantity is the price of a call option. Of course, in this precise case, there is an explicit formula for the above quantity, namely the celebrated Black–Scholes formula,

$$\mathbb{E}\left(\left(e^{\sigma Z} - K\right)_+\right) = e^{\sigma^2/2} F\left(\sigma - \frac{\log(K)}{\sigma}\right) - K F\left(-\frac{\log(K)}{\sigma}\right), \tag{1.2}$$

where

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du.$$

However there are variants of this problem which can be solved only by the Monte Carlo method (see Chapter 9). Suppose that we wish to compute the above quantity by the Monte Carlo method, that is, we approximate that quantity by

$$C \simeq n^{-1}\left[\left(e^{\sigma Z_1} - K\right)_+ + \ldots + \left(e^{\sigma Z_n} - K\right)_+\right].$$

Suppose now that we wish to evaluate the price of a put option,

$$P = \mathbb{E}\left(\left(K - e^{\sigma Z}\right)_+\right), \tag{1.3}$$

hence

$$P \simeq n^{-1}\left[\left(K - e^{\sigma Z_1}\right)_+ + \ldots + \left(K - e^{\sigma Z_n}\right)_+\right].$$

At least whenever $K^2 << \exp(\sigma^2/2)$,

$$\mathrm{var}\left[\left(K - e^{\sigma Z}\right)_+\right] < \mathrm{var}\left[\left(e^{\sigma Z} - K\right)_+\right].$$

The put–call parity relationship (which follows from C and P, and the relation $x = x^+ - x^-$) says that

$$C - P = e^{-\sigma^2/2} - K,$$

hence we should instead compute P by a Monte Carlo procedure, and use the put–call parity relationship in order to get C, rather than computing C directly by Monte Carlo (see Exercise 5.9 below).

Antithetic variables Suppose we wish to compute

$$I = \int_0^1 f(x) dx.$$

Since $x \to 1 - x$ leaves the measure dx invariant on $[0, 1]$,

$$I = \frac{1}{2} \int_0^1 (f(x) + f(1 - x)) dx.$$

We can then compute I as follows. We simulate n i.i.d. $\mathcal{U}(0,1)$ random variables $U_1, \ldots, U_n$, and we approximate I by

$$I_{2n} = \frac{1}{n}\left(\frac{1}{2}(f(U_1) + f(1-U_1)) + \ldots + \frac{1}{2}(f(U_n) + f(1-U_n))\right)$$
$$= \frac{1}{2n}\left(f(U_1) + f(1-U_1) + \ldots + f(U_n) + f(1-U_n)\right).$$

If we compare this method with a direct Monte Carlo method after n simulations, we note that the approximation is improved provided

$$\mathbb{E}f(U)f(1-U) < \mathbb{E}f^2(U),$$

which holds true provided the random variables $f(U)$ and $f(1-U)$ are linearly independent.

The method can be generalized to higher dimensions, and to other transformations which leave the law of the random variable to be simulated invariant.

For example, if we try to compute the price of a put option (1.3), we can use the fact that the law of Z is identical to that of $-Z$ and reduce the variance by a factor of at least 2. Indeed, if $f(x) = [K - e^{\sigma x}]_+$, $\sigma > 0$, f is monotone decreasing, hence

$$\operatorname{var}\left(\frac{f(Z) + f(-Z)}{2}\right) = \frac{1}{2}\operatorname{var}(f(Z)) + \frac{1}{2}\operatorname{cov}(f(Z), f(-Z))$$
$$\leq \frac{1}{2}\operatorname{var}(f(Z)),$$

since

$$\operatorname{cov}(f(Z), f(-Z)) \leq \mathbb{E}\left([f(Z) - f(0)][f(-Z) - f(0)]\right)$$
$$\leq 0.$$

Stratification method This method is well known in the context of survey sample design. Suppose we seek to compute

$$I = \mathbb{E}(g(X)) = \int g(x)f(x)dx,$$

where X has the law $f(x)dx$. We start by decomposing I into

$$I = \sum_{i=1}^{m} I_i = \sum_{i=1}^{m} \mathbb{E}(\mathbf{1}_{\{X \in D_i\}} g(X)),$$

where D_i is a partition of the integration set. We then use n_i simulations for the computation of I_i. Define $\sigma_i^2 = \operatorname{var}(\mathbf{1}_{\{X \in D_i\}} g(X))$. Then the variance of the approximation is

$$\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}.$$

If we minimize this quantity with the constraint that $\sum_{i=1}^m n_i = n$ is fixed, we get $n_i = n\sigma_i / \sum_{i=1}^m \sigma_i$. The minimum equals $n^{-1} \left(\sum_{i=1}^m \sigma_i \right)^2$. We can show that it is smaller that the variance obtained with n simulations of a standard Monte Carlo procedure. Of course, one can rarely compute the σ_i, which limits the use of this technique (but we can estimates the σ_i via a Monte Carlo procedure!). To learn more about this procedure, see [10].

Mean value Suppose we wish to compute

$$\mathbb{E}(g(X, Y)) = \int g(x, y) f(x, y) dx dy,$$

where $f(x, y)dxdy$ is the law of the pair (X, Y).

If we let

$$h(x) = \frac{1}{m(x)} \int g(x, y) f(x, y) dy,$$

with $m(x) = \int f(x, y)dy$, it is easy to check that

$$\mathbb{E}(g(X, Y)) = \mathbb{E}(h(X)).$$

Indeed, the law of X is $m(x)dx$, hence

$$\mathbb{E}(h(X)) = \int m(x)h(x)dx = \int dx \int g(x, y) f(x, y) dy = \mathbb{E}(g(X, Y)).$$

On the other hand, interpreting $h(X)$ as a conditional expectation, we can show that

$$\text{var}(h(X)) \leq \text{var}(g(X, Y)).$$

Consequently, if we can compute the function h explicitly, it is preferable to use a Monte Carlo procedure for $h(X)$.

Remark 4.1 *We wrote in the introduction to this chapter that the Monte Carlo method is particularly well suited to the computation of multiple integrals. We shall see a typical example of such a situation, for a mathematical finance problem, in Exercise 7.5. of Chapter 9.*

1.5 Exercises

Exercise 5.1 *Let X be a geometric random variable with parameter p, that is, $\mathbb{P}(X = k) = p(1 - p)^{k-1}, k \geq 1$.*

1. *Describe a method for simulating X based on a sequence of Bernoulli trials.*
2. *Give another method for simulating this law based on the formula $\mathbb{P}(X > k) = (1 - p)^k$, $k \geq 0$, and compare the two methods.*

Exercise 5.2 *1. Describe a standard method for simulating the Gaussian* $N(0, 1)$ *law.*

2. Propose a rejection algorithm for the simulation of a Gaussian random variable, based upon the simulation of doubly exponential random variables with density $(\lambda/2)\exp(-\lambda|x|)$.

3. Let X *and* Y *be two independent random variables, both exponential with parameter 1.*

(a) Give the conditional law of X, *given that* $\{Y > (1-X)^2/2\}$.

(b) Let Z *be a random variable having the above conditional law, and* S *an independent random variable taking the values* ± 1 *with probability* $1/2$. *Give the law of* SZ.

(c) Deduce another method for simulating the Gaussian $N(0, 1)$ *law.*

Exercise 5.3 *A process* $\{X(t);\ t \geq 0\}$ *with continuous trajectories is said to be a Brownian motion if it possesses the two following properties:*

(i) For any $n \geq 1$, $0 = t_0 < t_1 < t_2 < \ldots < t_n$, *the random variables* $X(t_k) - X(t_{k-1})(1 \leq k \leq n)$ *are mutually independent (we say that X(t) has independent increments).*

(ii) $X(0) = 0$ *and the law of* $X(t+h) - X(t)$ *is the Gaussian law* $N(0, h)$, *for all* $t \geq 0$, $h > 0$.

1. Propose a method for simulating $\{X(kh);\ k \geq 1\}$, *for a given* $h > 0$.

2. Give the conditional law of $X(t)$, *given that* $X(t-a) = x$ *and* $X(t+a) = y$. *Deduce a method for simulating* $\{X(kh/2);\ k \geq 1\}$ *which avoids the need to redo the simulations of part 1.*

Exercise 5.4 *Let* (X_1, X_2) *be a Gaussian random vector, with correlation coefficient* ρ *and such that, for* $i = 1, 2$, *the random variable* X_i *has the law* $N(\mu_i, \sigma_i^2)$.

1. Show that if (Y_1, Y_2) *is a pair of* $N(0, 1)$ *independent random variables, then the pair* $Z_1 = \mu_1 + \sigma_1 Y_1$, $Z_2 = \mu_2 + \sigma_2(\rho Y_1 + \sqrt{1-\rho^2}Y_2)$ *has the same law as* (X_1, X_2). *Deduce a method for simulating this random vector.*

2. Generalize to the case of an arbitrary dimension.

Exercise 5.5 *Let* X *denote a random variable with the distribution function* F. *Assume that* F *is one-to-one, and denote its inverse by* F^{-1}.

1. Give a method for simulating X *conditionally upon* $X > m$, *based on a rejection method. Discuss the efficiency of the method. What happens when* m *is large?*

2. *For a $\mathcal{U}(0,1)$ random variable U, define*

$$Z = F^{-1}\left(F(m) + (1 - F(m))U\right).$$

Compute the distribution function of Z and deduce a method of simulating X, conditionally upon $X > m$. Compare with the above rejection method.

3. *Generalize the previous method to the case where one seeks to simulate X conditionally upon $a < X < b$.*

4. *Suppose we now try to simulate a Gaussian $N(\mu, \sigma^2)$ random variable X, conditionally upon $X > m$. Show that we can restrict ourselves to the case of a standard normal random variable, provided we modify the value of m.*

5. *Propose, for the problem of part 4, a rejection method based upon a translated exponential law with the density $\theta e^{-\theta(x-m)}\mathbf{1}_{\{x > m\}}$. How should one choose the parameter θ?*

Exercise 5.6 (Importance sampling) *Suppose we wish to compute by a Monte Carlo method the quantity*

$$p_\ell = \mathbb{P}(X \in [\ell,\ \ell + 1]),$$

where X is an exponential random variable with parameter 1.

1. *Give the standard estimator of p_ℓ and compute its variance.*

2. *Propose an importance sampling method, such that the new simulations all belong to the interval $[\ell,\ \ell + 1]$. Compute the variance of this new estimator and discuss the case of large values of ℓ.*

Exercise 5.7 (Variance reduction)

1. *Propose an importance sampling method for the computation of*

$$I = \mathbb{E}\left(\mathbf{1}_{\{X > 0\}} \exp \beta X\right),$$

where X is a Gaussian $N(0,1)$ random variable and $\beta = 5$.

2. *Propose a control variate method for the same computation.*

3. *Improve the method with the help of an antithetic variable method.*

Exercise 5.8 *The aim of this exercise is to prove that the method of antithetic variables reduces the variance whenever we have a function which is monotone in each of its variables.*

1. *Suppose that f and g are both bounded and increasing from $\mathbb{R}$ into $\mathbb{R}$. Show that for any real-valued random variables X and Y,*

$$\mathbb{E}(f(X)g(X)) + \mathbb{E}(f(Y)g(Y)) \geq \mathbb{E}(f(X)g(Y)) + \mathbb{E}(f(Y)g(X)).$$

Deduce that for any real random variable X,

$$\mathbb{E}(f(X)g(X)) \geq \mathbb{E}(f(X))\,\mathbb{E}(g(X)) \Leftrightarrow \operatorname{cov}(f(X), g(X)) \geq 0.$$

2. *Show that if $X_1, \ldots, X_n$ are mutually independent real random variables,*

$$\mathbb{E}(f(X_1, \ldots, X_n)g(X_1, \ldots, X_n)|X_n) = \Phi(X_n),$$

where Φ is a function to be expressed as an expectation. Deduce that whenever f and g are increasing in each of their arguments,

$$\mathbb{E}(f(X_1, \ldots, X_n)g(X_1, \ldots, X_n)) \geq \mathbb{E}(f(X_1, \ldots, X_n))\,\mathbb{E}(g(X_1, \ldots, X_n)).$$

3. *Let h be a mapping from $[0,1]^n$ into $\mathbb{R}$, which is monotone in each of its arguments, and let $U_1, \ldots, U_n$ be independent $\mathcal{U}(0,1)$ random variables. Show that*

$$\operatorname{cov}(h(U_1, \ldots, U_n), h(1 - U_1, \ldots, 1 - U_n)) \leq 0,$$

and show that the method of antithetic random variables reduces the variance in this case.

Exercise 5.9 (Programming) *Recall the formula (1.1) for the price of a call option, and (1.3) for the price of a put option. Deduce from the identity $x = x^+ - (-x)^+$ the put–call parity relationship*

$$C - P = \mathbb{E}e^{\sigma Z} - K,$$

where the expectation $\mathbb{E}e^{\sigma Z}$ can be computed explicitly and equals $\exp(\sigma^2/2)$. Deduce from this identity a control variate method, and show that it reduces the variance.

Since Z and $-Z$ have the same law, one can apply a method of antithetic random variables to the two Monte Carlo computations of the call and of the put.

Choose for the simulation $\sigma = 1.5$ and $K = 1$. Do the Monte Carlo computations with sample sizes $N = 1000, 10\,000$ and $100\,000$. For each computation, give the estimate deduced from the Monte Carlo simulations, and a 95 % confidence interval, based on the central limit theorem and an estimate of the variance.

1. *Compute the value deduced from the Black–Scholes formula (1.2).*

2. *Compute C by a Monte Carlo procedure, using first the formula (1.1), and then the put–call parity relationship and (1.3) for the computation of P by a Monte Carlo procedure.*

3. *Repeat the same two computations, using an antithetic variable method.*

2

Markov chains

Introduction

A Markov chain is a sequence of random variables $\{X_n; n = 0, 1, 2, \dots\}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking its values in a set E which could be arbitrary, but which for us will be either finite or countable, and which possesses the Markov property. Intuitively, a Markov chain has the property that, knowing the present state X_n, one can forget the past if one wants to predict the future. One way to construct a Markov chain is as follows. Let $\{Y_n;\ n \geq 1\}$ be mutually independent F-valued random variables, which are globally independent of X_0. Given a mapping $f : \mathbb{N} \times E \times F \to E$, we define $\{X_n; n \geq 1\}$ recursively by

$$X_n = f(n, X_{n-1}, Y_n).$$

In a way, this is the simplest model of non-mutually independent random variables.

The next two chapters will present many applications of Markov chains. Note that we shall restrict our presentation to *homogeneous* Markov chains (in the above recurrence relation, f does not depend upon n, and the Y_n all have the same law), even though non-homogeneous chains are necessary in many applications. Even in those cases, understanding the long time behaviour of the homogeneous chains is crucial.

2.1 Definitions and elementary properties

We wish to define and study Markov chains $\{X_n; n \in \mathbb{N}\}$ with values in a (finite or) countable state space E. We shall denote by $x, y, \dots$ generic points of E. We shall use the convention that whenever a condition involves a conditional probability $\mathbb{P}(A|B)$, that condition is assumed to be satisfied only when $\mathbb{P}(B) > 0$.

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

Definition 1.1 *The E-valued stochastic process $\{X_n; n \in \mathbb{N}\}$ is called a Markov chain whenever, for all $n \in \mathbb{N}$, the conditional law of X_{n+1} given $X_0, X_1, \ldots, X_n$ equals its conditional law given X_n, that is, for all $x_0, \ldots, x_{n+1} \in E$,*

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n).$$

A simple criterion, which allows us in many cases to verify that a given process is a Markov chain, is given by the following lemma:

Lemma 1.2 *Let E and F be two countable sets, and let f be a mapping from $\mathbb{N} \times E \times F$ into E. Let $X_0, Y_1, Y_2, \ldots$ be mutually independent random variables, X_0 being E-valued, and the Y_n being F-valued. Let $\{X_n;\ n \geq 1\}$ be the E-valued process defined by*

$$X_{n+1} = f(n, X_n, Y_{n+1}), \quad n \in \mathbb{N}.$$

Then $\{X_n;\ n \in \mathbb{N}\}$ is a Markov chain.

PROOF

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n) \\
&\quad = \frac{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, X_{n+1} = x_{n+1})}{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n)} \\
&\quad = \sum_{\{z; f(n,x_n,z) = x_{n+1}\}} \frac{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n, Y_{n+1} = z)}{\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n)} \\
&\quad = \sum_{\{z; f(n,x_n,z) = x_{n+1}\}} \mathbb{P}(Y_{n+1} = z) \\
&\quad = \frac{\mathbb{P}(X_n = x_n, X_{n+1} = x_{n+1})}{\mathbb{P}(X_n = x_n)}.
\end{aligned}$$

□

A Markov chain is the analogue of a deterministic sequence which is defined by a recurrence relation of the type

$$x_{n+1} = f(n, x_n),$$

as opposed to a system 'with memory', of the type

$$x_{n+1} = f(n, x_n, x_{n-1}, \ldots, x_1, x_0).$$

Here the function $f(n, \cdot)$ is replaced by the 'transition matrix'

$$P_{xy} = \mathbb{P}(X_{n+1} = y | X_n = x).$$

From now on, this matrix $P = (P_{xy};\ x, y \in E)$ will be assumed to be independent of the time variable n. One then says that the Markov chain is *homogeneous*.

The chain constructed in Lemma 1.2 is homogeneous whenever f does not depend upon n, and the Y_n all have the same law. We now state a variant of Lemma 1.2, whose proof is essentially identical, and which will be useful below.

Lemma 1.3 *Let E be a countable set, and f be a mapping from $E \times [0,1]$ into E, such that for all $x, y \in E$, the set $\{u \in [0,1];\ f(x,u) = y\}$ is a Borel subset of $[0,1]$. Let $X_0, Y_1, Y_2, \ldots$ be mutually independent random variables, with X_0 taking its values in E, and the Y_n being uniform on $[0,1]$, and let $\{X_n;\ n \geq 1\}$ be the E-valued random sequence defined by*

$$X_{n+1} = f(X_n, Y_{n+1}), \quad n \in \mathbb{N}.$$

Then $\{X_n;\ n \in \mathbb{N}\}$ is a Markov chain.

The matrix P is called *Markovian* (or *stochastic*), in the sense that it has the property that, for all $x \in E$, the row vector $(P_{xy};\ y \in E)$ is a probability measure on E, or in other words,

$$P_{xy} \geq 0,\ \forall y \in E; \quad \sum_{y \in E} P_{xy} = 1.$$

Remark 1.4 *P_{xy} is the entry in row x and column y of the matrix P. This notation may surprise the reader, but it is very convenient. It is more common to enumerate rows and columns, and hence to index them by $1, 2, \ldots$. We note, moreover, that our matrices are square matrices, with possibly an infinite number of rows and columns, in the case where $|E| = \infty$.*

As we will now see, the law of a Markov chain is entirely determined by the 'initial law' $(\mu_x;\ x \in E)$, which is the law of X_0, and the transition matrix of the chain.

Definition 1.5 *Let μ be a probability on E, and P a Markovian matrix. An E-valued random sequence $\{X_n;\ n \in \mathbb{N}\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is called a (μ, P) Markov chain (i.e. with initial law μ and transition matrix P) if:*

(i) $\mathbb{P}(X_0 = x) = \mu_x,\ \forall x \in E;$

(ii) $\mathbb{P}(X_{n+1} = y | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x) = P_{xy},$
$\forall x_0, \ldots, x_{n-1}, x, y \in E.$

Proposition 1.6 *A necessary and sufficient condition for an E-valued random sequence $\{X_n;\ n \in \mathbb{N}\}$ to be a (μ, P) Markov chain is that, for all $n \in \mathbb{N}$, the law of the random variable $(X_0, X_1, \ldots, X_n)$ be given by*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mu_{x_0} P_{x_0 x_1} \times \cdots \times P_{x_{n-1} x_n}.$$

Proof For the necessary condition, if $\mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) > 0$, then

$$\begin{aligned}\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) =& \mathbb{P}(X_n = x_n | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \\ & \times \cdots \times \mathbb{P}(X_1 = x_1 | X_0 = x_0) \mathbb{P}(X_0 = x_0),\end{aligned}$$

and the above identity follows from the definition. Otherwise, both sides of the identity in the statement are zero (consider the smallest index k such that $\mathbb{P}(X_0 = x_0, \ldots, X_k = x_k) = 0$).

We now turn to the sufficient condition. (i) The identity in the statement follows from the definition. Let us prove more than (ii):

$$\mathbb{P}(X_{n+1} = x_{n+1}, \ldots, X_{n+p} = x_{n+p} | X_0 = x_0, \ldots, X_n = x_n) = \frac{\mu_{x_0} P_{x_0 x_1} \times \cdots \times P_{x_{n+p-1} x_{n+p}}}{\mu_{x_0} P_{x_0 x_1} \times \cdots \times P_{x_{n-1} x_n}}$$

and (ii) now follows if we choose $p = 1$. □

We have in fact established the following result:

Corollary 1.7 *If* $\{X_n;\ n \in \mathbb{N}\}$ *is a* (μ, P) *Markov chain, then for all* n, p, $x_0, \ldots, x_{n+p}$,

$$\mathbb{P}(X_{n+1} = x_{n+1}, \ldots, X_{n+p} = x_{n+p} | X_0 = x_0, \ldots, X_n = x_n) = P_{x_n x_{n+1}} \times \cdots \times P_{x_{n+p-1} x_{n+p}}.$$

A probability μ on E is considered to be a row vector, a mapping $g : E \to \mathbb{R}$ as a column vector, which justifies the notation

$$(\mu P)_y = \sum_{x \in E} \mu_x P_{xy},$$

$$(Pg)_x = \sum_{y \in E} P_{xy} g_y,$$

and the integral of a function g with respect to a measure μ is written (whenever the sum converges absolutely) as the product of a row vector on the left with a column vector on the right:

$$\mu g = \sum_{x \in E} \mu_x g_x.$$

Proposition 1.8 *Let* $\{X_n; n \in \mathbb{N}\}$ *be a* (μ, P) *Markov chain. Then*

(i) $\mathbb{P}(X_n = y | X_0 = x) = \mathbb{P}(X_{n+p} = y | X_p = x) = (P^n)_{xy}$,

(ii) $\mathbb{P}(X_n = y) = (\mu P^n)_y$,

(iii) $\mathbb{E}[g(X_n) | X_0 = x] = (P^n g)_x$.

PROOF

(i)

$$\mathbb{P}(X_n = y | X_0 = x) = \sum_{x_1, \ldots, x_{n-1}} \mathbb{P}(X_n = y, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1 | X_0 = x) = \sum_{x_1, \ldots, x_{n-1}} \frac{\mu_x P_{x x_1} \times \cdots \times P_{x_{n-1} y}}{\mu_x}$$

$$= \sum_{x_1,\ldots,x_{n-1}} P_{xx_1} \times \cdots \times P_{x_{n-1}y}$$
$$= (P^n)_{xy}.$$

(ii) We note that

$$\mathbb{P}(X_n = y) = \sum_{x\in E} \mathbb{P}(X_n = y, X_0 = x)$$
$$= \sum_{x\in E} \mathbb{P}(X_n = y|X_0 = x)\mu_x,$$

and we use (i).

(iii) We again use (i) starting from

$$\mathbb{E}[g(X_n)|X_0 = x] = \sum_{y\in E} g_y \mathbb{P}(X_n = y|X_0 = x).$$

□

2.2 Examples

2.2.1 Random walk in $E = \mathbb{Z}^d$

Let $\{Y_n; n \in \mathbb{N}^*\}$ denote an i.i.d. $\mathbb{Z}^d$-valued random sequence, with the common law λ, and let X_0 be a $\mathbb{Z}^d$-valued random variable, independent of the Y_n. Then the random sequence $\{X_n; n \geq 0\}$ defined by

$$X_{n+1} = X_n + Y_{n+1}, \quad n \in \mathbb{N},$$

is a (μ, P) Markov chain, with μ the law of X_0, and $P_{xy} = \lambda_{y-x}$. The classical case is that of the symmetric random walk starting from 0, that is,

$$\mu = \delta_0, \quad \lambda_{\pm e_i} = \frac{1}{2d},$$

where $(e_1, \ldots, e_d)$ is an orthonormal basis of $\mathbb{R}^d$.

2.2.2 Bienaymé–Galton–Watson process

This is a branching process $\{Z_n; n \in \mathbb{N}\}$ where Z_n denotes the number of males in the nth generation with a certain name, those individuals being all descendants of a common ancestor, the unique male in generation 0 ($Z_0 = 1$ almost surely). We assume that the ith male from the nth generation has ξ_i^n male children ($1 \leq i \leq Z_n$), in such a way that

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_i^n.$$

Our main assumption is that the random variables $\{\xi_i^n; i = 1, 2, \ldots, n = 0, 1, 2, \ldots\}$ are i.i.d., so that in particular Z_n and $\{\xi_1^n, \ldots, \xi_p^n, \ldots\}$ are independent.

The random sequence $\{Z_n;\ n \in \mathbb{N}\}$ is a (μ, P) $\mathbb{N}$-valued Markov chain, with $\mu = \delta_1$ and

$$P_{xy} = (p^{*x})_y,$$

where p^{*x} denotes the xth convolution power of the joint law p on $\mathbb{N}$ of the ξ_n^k, that is, the law of the sum of x i.i.d. random variables, all having the law p.

2.2.3 A discrete time queue

We consider a queue at a counter. X_n denotes the number of customers who are either waiting or being served at time n. Between time n and time $n + 1$, Y_{n+1} new customers join the queue, and whenever $X_n > 0$, Z_{n+1} customers leave the queue (with $Z_{n+1} = 0$ or 1). We assume that $X_0, Y_1, Z_1, Y_2, Z_2 \ldots$ are mutually independent, with $0 < \mathbb{P}(Y_n = 0) < 1$, and moreover $\mathbb{P}(Z_n = 1) = p = 1 - \mathbb{P}(Z_n = 0)$. We have

$$X_{n+1} = X_n + Y_{n+1} - \mathbf{1}_{\{X_n > 0\}} Z_{n+1}.$$

2.3 Strong Markov property

Let us first reformulate the Markov property. Let $\{X_n; n \in \mathbb{N}\}$ be an E-valued Markov chain defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given a probability measure μ on E, we shall use the notation $\mathbb{P}_\mu$ to denote any probability on $(\Omega, \mathcal{F})$ such that under $\mathbb{P}_\mu$ the sequence $\{X_n; n \geq 0\}$ is a Markov chain with initial law μ; in other words, μ is the law of X_0, that is,

$$\mathbb{P}_\mu(X_0 = x) = \mu_x,\ x \in E.$$

Whenever $\mu = \delta_x$, we shall write $\mathbb{P}_x$ instead of $\mathbb{P}_{\delta_x}$. $\mathbb{P}_x$ can be interpreted as the conditional law of X, given that $X_0 = x$. For any $n \geq 0$, we define $\mathcal{F}_n$ to be the sigma-algebra of those events which are 'determined by $X_0, X_1, \ldots, X_n$', that is,

$$\mathcal{F}_n = \left\{\{\omega; (X_0(\omega), \ldots, X_n(\omega)) \in B_n\};\ B_n \in \mathcal{P}(E^{n+1})\right\},$$

where $\mathcal{P}(F)$ denotes the collection of all the subsets of F.

Theorem 3.1 *Let $\{X_n; n \geq 0\}$ be a (μ, P) Markov chain. Then for any $n \in \mathbb{N}$, $x \in E$, conditionally upon $\{X_n = x\}$, $\{X_{n+p}; p \geq 0\}$ is a (δ_x, P) Markov chain, which is independent of $(X_0, \ldots, X_n)$. In other words, for all $A \in \mathcal{F}_n$ and any $m > 0$, $x_1, \ldots, x_m \in E$,*

$$\begin{aligned}&\mathbb{P}\left(A \cap \{X_{n+1} = x_1, \ldots, X_{n+m} = x_m\} | X_n = x\right)\\ &\quad = \mathbb{P}(A | X_n = x)\mathbb{P}_x(X_1 = x_1, \ldots, X_m = x_m).\end{aligned}$$

PROOF It suffices to prove the result in the case where $A = \{X_0 = y_0, X_1 = y_1, \ldots, X_n = y_n\}$ (A is a finite or countable union of disjoint sets of that form, and the result in the general case will then follow from the σ-additivity of $\mathbb{P}$). It suffices to consider the case $y_n = x$, since otherwise both sides of the equality vanish. The left-hand side of the identity in the statement equals

$$\frac{\mathbb{P}(X_0 = y_0, \ldots, X_n = x, X_{n+1} = x_1, \ldots, X_{n+m} = x_m)}{\mathbb{P}(X_n = x)},$$

which, applying Proposition 1.6 twice, is shown to equal

$$\frac{\mathbb{P}(A)}{\mathbb{P}(X_n = x)} \times P_{xx_1} \times P_{x_1x_2} \times \cdots \times P_{x_{m-1}x_m},$$

or, in other words,

$$\mathbb{P}(A|X_n = x)\mathbb{P}_x(X_1 = x_1, \ldots, X_m = x_m).$$

□

The preceding result says in particular that the past and the future of the chain are conditionally independent, given the position of the chain at the present time n.

We now wish to extend the Markov property, replacing the fixed time n by a random time (but not any random time).

Definition 3.2 *A random variable T taking values in the set $\mathbb{N} \cup \{+\infty\}$ is called a stopping time if, for all $n \in \mathbb{N}$,*

$$\{T = n\} \in \mathcal{F}_n.$$

In other words, the observation of $X_0, X_1, \ldots, X_n$, the trajectory of the chain up to time n, is enough to decide whether or not T equals n.

Example 3.3 *(i)For all $x \in E$, the first passage time at state x,*

$$S_x = \begin{cases} \inf\{n \geq 0;\, X_n = x\} & \text{if such an } n \text{ exists}, \\ +\infty, & \text{otherwise}, \end{cases}$$

and the time of the first return to state x,

$$T_x = \begin{cases} \inf\{n \geq 1;\, X_n = x\} & \text{if such an } n \text{ exists}, \\ +\infty, & \text{otherwise}, \end{cases}$$

are stopping times. (With the convention that the infimum of the empty set is $+\infty$, it is sufficient to write: $T_x = \inf\{n \geq 1;\ X_n = x\}$.) In the case of T_x this is because

$$\{T_x = n\} = \{X_1 \neq x\} \cap \ldots \cap \{X_{n-1} \neq x\} \cap \{X_n = x\}.$$

(ii) For all $A \subset E$, the time of the first visit to the set A,

$$T_A = \inf\{n \geq 1;\, X_n \in A\},$$

is a stopping time.

(iii) On the other hand, the time of the last visit to A,

$$L_A = \sup\{n \geq 1;\, X_n \in A\},$$

is not a stopping time, since we need to know the trajectory after time n in order to decide whether or not $L_A = n$.

We shall denote by $\mathcal{F}_T$ the σ-algebra of events which are 'determined by $X_0, X_1, \ldots, X_T$', which is defined as the σ-algebra of those events $B \in \mathcal{F}$ which are such that for all $n \in \mathbb{N}$,

$$B \cap \{T = n\} \in \mathcal{F}_n.$$

Theorem 3.4 (Strong Markov property) *Let $\{X_n; n \geq 0\}$ be a (μ, P) Markov chain, and T a stopping time. Conditionally upon $\{T < \infty\} \cap \{X_T = x\}$, $\{X_{T+n}; n \geq 0\}$ is a (δ_x, P) Markov chain, which is independent of $\mathcal{F}_T$. In other words, for all $A \in \mathcal{F}_T$ and all $m > 0$, $x_1, \ldots, x_m \in E$,*

$$\begin{aligned}&\mathbb{P}(A \cap \{X_{T+1} = x_1, \ldots, X_{T+m} = x_m\} | X_T = x, T < \infty)\\ &= \mathbb{P}(A | X_T = x,\ T < \infty) \times \mathbb{P}_x(X_1 = x_1, \ldots, X_m = x_m).\end{aligned}$$

PROOF It suffices to show that, for all $n \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{P}(A \cap \{T = n\} \cap \{X_{T+1} = x_1, \ldots, X_{T+m} = x_m\} | X_T = x)\\ &= \mathbb{P}(A \cap \{T = n\} | X_T = x) \mathbb{P}_x(X_1 = x_1, \ldots, X_m = x_m),\end{aligned}$$

which follows from Theorem 3.1, and then to sum over all possible values of n. □

2.4 Recurrent and transient states

Define $T_x = \inf\{n \geq 1;\, X_n = x\}$ as in Example 3.3.

Definition 4.1 *$x \in E$ is said to be recurrent if $\mathbb{P}_x(T_x < \infty) = 1$, and transient otherwise (i.e. if $\mathbb{P}_x(T_x < \infty) < 1$).*

We define the number of returns to the state x:

$$N_x = \sum_{n \geq 1} \mathbf{1}_{\{X_n = x\}}.$$

Proposition 4.2 *(a) If x is recurrent, then*

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

(b) If x is transient, then

$$\mathbb{P}_x(N_x = k) = (1 - \Pi_x)\Pi_x^k,\ k \geq 0,$$

where $\Pi_x = \mathbb{P}_x(T_x < \infty)$ (in particular, $N_x < \infty$, $\mathbb{P}_x$ almost surely).

PROOF Let

$$\begin{aligned} T_x^2 &= \inf\{n > T_x;\, X_n = x\} \\ &= T_x + \inf\{n \geq 1;\, X_{T_x+n} = x\}. \end{aligned}$$

It is not hard to show that T_x^2 is a stopping time:

$$\begin{aligned} \mathbb{P}_x(T_x^2 < \infty) &= \mathbb{P}_x(T_x^2 < \infty | T_x < \infty)\mathbb{P}_x(T_x < \infty) \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x(T_x^2 = T_x + n | T_x < \infty)\mathbb{P}_x(T_x < \infty). \end{aligned}$$

But from Theorem 3.4 we deduce that

$$\begin{aligned} &\mathbb{P}_x(T_x^2 = T_x + n | T_x < \infty) \\ &\quad = \mathbb{P}_x(X_{T_x+1} \neq x, \ldots, X_{T_x+n-1} \neq x, X_{T_x+n} = x | T_x < \infty) \\ &\quad = \mathbb{P}_x(X_1 \neq x, \ldots, X_{n-1} \neq x, X_n = x) \\ &\quad = \mathbb{P}_x(T_x = n). \end{aligned}$$

Finally,

$$\mathbb{P}_x(T_x^2 < \infty) = (\mathbb{P}_x(T_x < \infty))^2$$

or

$$\mathbb{P}_x(N_x \geq 2) = (\mathbb{P}_x(T_x < \infty))^2$$

and, iterating the same argument, we deduce that

$$\mathbb{P}_x(N_x \geq k) = (\mathbb{P}_x(T_x < \infty))^k, \quad k \in \mathbb{N}.$$

Both statements of the proposition follow easily from this identity. □

Corollary 4.3 *x is recurrent if and only if*

$$\sum_{n=0}^{\infty} (P^n)_{xx} = +\infty.$$

PROOF

$$\mathbb{E}_x(N_x) = \sum_{n\geq 1} \mathbb{P}_x(X_n = x)$$
$$= \sum_{n\geq 1} (P^n)_{xx}.$$

It follows from Proposition 4.2 that this quantity is infinite whenever x is recurrent. On the other hand, if x is transient, then

$$\mathbb{E}_x(N_x) = \sum_{k=1}^{\infty} k(1 - \Pi_x)\Pi_x^k$$
$$= \frac{\Pi_x}{1 - \Pi_x} < \infty.$$

□

Definition 4.4 *We say that the state y is accessible from x (denoted by $x \to y$) whenever there exists $n \geq 0$ such that $\mathbb{P}_x(X_n = y) > 0$. We say that x and y communicate (written $\leftrightarrow$) whenever both $x \to y$ and $y \to x$.*

The relation $x \leftrightarrow y$ is an equivalence relation, and we can partition E into equivalence classes modulo the relation $\leftrightarrow$.

Note that $x \to y \Leftrightarrow \exists\, n \geq 0$ such that $(P^n)_{xy} > 0$, since $\mathbb{P}_x(X_n = y) = (P^n)_{xy}$ (Proposition 1.8(i)).

Theorem 4.5 *Let $C \subset E$ be an equivalence class for the relation $\leftrightarrow$. Then all states in C either are recurrent, or else they all are transient.*

PROOF Let $x, y \in C$. It suffices to show that x transient $\Rightarrow$ y transient (since then y recurrent $\Rightarrow$ x recurrent). Since $x \leftrightarrow y$, there exist $n, m > 0$ such that $(P^n)_{xy} > 0$ and $(P^m)_{yx} > 0$. But for all $r \geq 0$,

$$(P^{n+r+m})_{xx} \geq (P^n)_{xy}(P^r)_{yy}(P^m)_{yx}$$

and

$$\sum_{r=0}^{\infty} (P^r)_{yy} \leq \frac{1}{(P^n)_{xy}(P^m)_{yx}} \sum_{n=0}^{\infty} (P^{n+r+m})_{xx} < \infty.$$

□

Definition 4.6 *A (μ, P) Markov chain is said to be irreducible whenever E consists of a single equivalence class. It is said to be irreducible and recurrent if it is irreducible and all states are recurrent.*

Proposition 4.7 *Any irreducible Markov chain on a finite state space E is recurrent.*

PROOF Whenever E is finite, at least one state must be visited infinitely many times with positive probability, hence almost surely by Proposition 4.2, and that state (as well as all states) is (are) recurrent. □

2.5 The irreducible and recurrent case

In this section, we assume that the chain is both irreducible and recurrent. We start by studying the *excursions* of the chain between two successive returns to state x:

$$\mathcal{E}_k = (X_{T_x^k}, X_{T_x^k+1}, \ldots, X_{T_x^{k+1}}), \quad k \geq 0.$$

These excursions are random sequences whose length is random, at least 2 and finite, composed of elements of $E\backslash\{x\}$, except for the first and the last, which are equal to x. Denote by U the set of sequences

$$u = (x, x_1, \ldots, x_n, x),$$

with $n \geq 0$, $x_\ell \neq x$, $1 \leq \ell \leq n$. U is countable, and it is the set of all possible excursions $\mathcal{E}_0, \mathcal{E}_1, \ldots$. Hence these random variables take their values in a countable set, and their probability law is characterized by the quantities

$$\mathbb{P}(\mathcal{E}_k = u), \quad u \in U.$$

Proposition 5.1 *Under $\mathbb{P}_x$, the sequence $(\mathcal{E}_0, \mathcal{E}_1, \ldots)$ of excursions is i.i.d.; in other words, there exists a probability $\{p_u;\ u \in U\}$ on U such that, for all $k > 0$, $u_0, \ldots, u_k \in U$,*

$$\mathbb{P}_x(\mathcal{E}_0 = u_0, \mathcal{E}_1 = u_1, \ldots, \mathcal{E}_k = u_k) = \prod_{\ell=0}^{k} p_{u_\ell}.$$

Proof This is a consequence of the strong Markov property. Indeed, $\{\mathcal{E}_0 = u_0\} \in \mathcal{F}_{T_x}$, and the event

$$\{\mathcal{E}_1 = u_1, \ldots, \mathcal{E}_k = u_k\}$$

is of the form

$$\{X_{T_x+1} = x_1, \ldots, X_{T_x+p} = x_p\},$$

for some $p > 0$, $x_1, \ldots, x_p \in E$. Consequently,

$$\begin{aligned}
&\mathbb{P}_x(\mathcal{E}_0 = u_0, \mathcal{E}_1 = u_1, \ldots, \mathcal{E}_k = u_k) \\
&= \mathbb{P}_x(\{\mathcal{E}_0 = u_0\} \cap \{X_{T_x+1} = x_1, \ldots, X_{T_x+p} = x_p\} | T_x < \infty) \\
&= \mathbb{P}_x(\mathcal{E}_0 = u_0)\mathbb{P}_x(X_1 = x_1, \ldots, X_p = x_p) \\
&= \mathbb{P}_x(\mathcal{E}_0 = u_0)\mathbb{P}_x(\mathcal{E}_0 = u_1, \ldots, \mathcal{E}_{k-1} = u_k) \\
&= \mathbb{P}_x(\mathcal{E}_0 = u_0)\mathbb{P}_x(\mathcal{E}_0 = u_1) \times \ldots \times \mathbb{P}_x(\mathcal{E}_0 = u_k) \\
&= p_{u_0} p_{u_1} \cdots p_{u_k},
\end{aligned}$$

where $\{p_u; u \in U\}$ is the law of $\mathcal{E}_0$ under $\mathbb{P}_x$. □

A measure on the set E is a 'row vector' $\{\gamma_x;\ x \in E\}$ such that $0 \leq \gamma_x < \infty$, for all x. Whenever the measure is finite, $\sum_{x \in E} \gamma_x < \infty$, we can normalize it, to

make it a probability on E, $\left(\gamma_x / \sum_z \gamma_z,\ x \in E\right)$. A measure γ is said to be invariant (with respect to the transition matrix P) whenever

$$\gamma P = \gamma,$$

that is,

$$\sum_{y \in E} \gamma_y P_{yx} = \gamma_x,\ x \in E.$$

A measure γ is said to be strictly positive if $\gamma_x > 0$, for all $x \in E$.

A probability measure γ is invariant if and only if the chain (γ, P) has the property that γ is the law of X_n, for all $n \in \mathbb{N}$, hence for all n, $\{X_{n+m};\ m \in \mathbb{N}\}$ is a (γ, P) Markov chain.

Remark 5.2 *An invariant probability is a probability π which satisfies $\pi P = \pi$, or equivalently, for all $x \in E$,*

$$\sum_{y \neq x} \pi_y P_{yx} = \pi_x (1 - P_{xx}),$$

that is,

$$\mathbb{P}(X_n \neq x, X_{n+1} = x) = \mathbb{P}(X_n = x, X_{n+1} \neq x),$$

which means that at equilibrium, the mean number of departures from state x between time n and time $n+1$ equals the mean number of arrivals at state x between time n and time $n+1$. The relation which characterizes the invariant probability is very intuitive.

Theorem 5.3 *Let $\{X_n; n \in \mathbb{N}\}$ be a Markov chain with transition matrix P, which we assume to be irreducible and recurrent. Then there exists a strictly positive invariant measure γ, which is unique up to a multiplicative constant.*

Proof To prove existence, let γ_y^x denote the mean number of visits to state y during the excursion $\mathcal{E}_0$ starting at x, that is,

$$\begin{aligned}
\gamma_y^x &= \mathbb{E}_x \sum_{n=1}^{T_x} \mathbf{1}_{\{X_n = y\}} \\
&= \sum_{n=1}^{\infty} \mathbb{P}_x(X_n = y, n \leq T_x) \\
&= \sum_{z \in E} \sum_{n=1}^{\infty} \mathbb{P}_x(\{X_{n-1} = z, n-1 < T_x\} \cap \{X_n = y\}) \\
&= \sum_{z \in E} \left(\sum_{n=2}^{\infty} \mathbb{P}_x(X_{n-1} = z, n-1 \leq T_x) \right) P_{zy} \\
&= (\gamma^x P)_y.
\end{aligned}$$

Note that we have used recurrence to obtain the penultimate equality. We now exploit the irreducibility of the chain. There exist n, m such that $(P^n)_{xy} > 0$, $(P^m)_{yx} > 0$. Hence, since $\gamma_x^x = 1$,

$$0 < (P^n)_{xy} = \gamma_x^x (P^n)_{xy} \leq (\gamma^x P^n)_y = \gamma_y^x,$$
$$\gamma_y^x (P^m)_{yx} \leq (\gamma^x P^m)_x = \gamma_x^x = 1.$$

Consequently, γ^x is a strictly positive measure, which satisfies $\gamma_x^x = 1$.

Turning to uniqueness, let λ denote an invariant measure such that $\lambda_x = 1$. We shall first prove that $\lambda \geq \gamma^x$, then that $\lambda = \gamma^x$. Note that this part of the proof of the theorem exploits only irreducibility (and not recurrence). We have

$$\begin{aligned}
\lambda_y &= P_{xy} + \sum_{z_1 \neq x} \lambda_{z_1} P_{z_1 y} \\
&= P_{xy} + \sum_{z_1 \neq x} P_{x z_1} P_{z_1 y} + \sum_{z_1, z_2 \neq x} \lambda_{z_2} P_{z_2 z_1} P_{z_1 y} \\
&\geq \sum_{n=0}^{\infty} \sum_{z_1, \ldots, z_n \neq x} P_{x z_n} P_{z_n z_{n-1}} \cdots P_{z_1 y} \\
&= \sum_{n=0}^{\infty} \mathbb{P}_x (X_{n+1} = y, T_x \geq n+1) \\
&= \gamma_y^x.
\end{aligned}$$

Hence $\mu = \lambda - \gamma^x$ is also an invariant measure, and $\mu_x = 0$. Let $y \in E$, and n be such that $(P^n)_{yx} > 0$. Then

$$0 = \mu_x = \sum_{z \in E} \mu_z (P^n)_{zx} \geq \mu_y (P^n)_{yx}.$$

Hence $\mu_y = 0$, and this holds for all $y \in E$. □

We have seen that a state x is recurrent whenever

$$\mathbb{P}_x (T_x < \infty) = 1.$$

Let $m_x = \mathbb{E}_x (T_x)$. If this quantity is finite, then x is called *positive recurrent*, and otherwise it is called *null recurrent*.

Theorem 5.4 *Assume again that the chain is irreducible. A state x is positive recurrent if and only if all the states are positive recurrent, if and only if there exists an invariant probability $\pi = (\pi_x = m_x^{-1}, \ x \in E)$.*

Proof Note that

$$m_x = \sum_{y \in E} \gamma_y^x.$$

Hence if x is positive recurrent, then the probability $\pi = (\pi_y = \gamma_y^x/m_x,\ y \in E)$ is invariant.

Conversely, if π is an invariant probability, from the irreducibility (see the end of the proof of existence in Theorem 5.3), π is strictly positive, hence if x is an arbitrary state, $\lambda = \left(\lambda_y = \pi_y/\pi_x,\ y \in E\right)$ is an invariant measure which satisfies $\lambda_x = 1$. From the irreducibility and the proof of uniqueness in Theorem 5.3,

$$m_x = \sum_{y \in E} \gamma_y^x = \sum_{y \in E} \frac{\pi_y}{\pi_x} = \frac{1}{\pi_x} < \infty.$$

Hence x is positive recurrent, as are all the states. □

The following dichotomy follows from the two preceding theorems: in the *irreducible and recurrent* case, the chain is *positive recurrent* whenever there exists an *invariant probability*, *null recurrent* if *one (hence all) invariant measure(s)* has *infinite* total mass ($\sum_i \pi_i = +\infty$). In particular, if $|E| < \infty$, there do not exist null recurrent states, rather, any recurrent state is positive recurrent.

Corollary 5.5 *Let $\{X_n\}$ be an irreducible Markov chain which is positive recurrent. With any $x \in E$ we associate $T_x = \inf\{n > 0;\ X_n = x\}$. Then for all $y \in E$,*

$$\mathbb{E}_y(T_x) < \infty.$$

PROOF Note that

$$T_x \geq T_x \mathbf{1}_{\{T_y < T_x\}},$$

whence, taking the expectation under $\mathbb{P}_x$,

$$m_x \geq \mathbb{E}_x(T_x | T_y < T_x)\mathbb{P}_x(T_y < T_x).$$

But it follows from the strong Markov property that $\mathbb{E}_x(T_x | T_y < T_x) > \mathbb{E}_y(T_x)$, and from the irreducibility that $\mathbb{P}_x(T_y < T_x) > 0$, and the proof is complete. □

Remark 5.6 (The non-irreducible case) *For simplicity, we consider here only the case $|E| < \infty$. There exists at least one recurrent class (which is positive recurrent), hence there exists a least one invariant probability. Any invariant probability charges only recurrent states. If there is only one recurrent class, then the chain possesses one and only one invariant probability. Otherwise, we can associate with each recurrent class a unique invariant probability whose support is that class, and all invariant measures are convex linear combinations of these, which are the extremal ones. Hence, if there are at least two different recurrent classes, there are an uncountable number of invariant probabilities.*

We restrict ourself again to the irreducible case. We can now establish the ergodic theorem, which is a generalization of the law of large numbers.

Theorem 5.7 *Suppose that the chain is irreducible and positive recurrent. Let* $\pi = (\pi_x,\ x \in E)$ *denote its unique invariant probability. If* $f : E \to \mathbb{R}$ *is bounded, then* $\mathbb{P}$ *almost surely, as* $n \to \infty$,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \sum_{x\in E} \pi_x f(x).$$

PROOF By assumption, there exists c such that $|f(x)| \le c$, for all $x \in E$.

Let

$$N_x(n) = \sum_{1\le k\le n} \mathbf{1}_{\{X_k = x\}}$$

denote the number of returns to state x before time n. We wish to study the limit as $n \to \infty$ of

$$\frac{N_x(n)}{n}.$$

Let $S_x^0, S_x^1, \ldots, S_x^k, \ldots$ denote the lengths of the excursions $\mathcal{E}_0, \mathcal{E}_1, \ldots, \mathcal{E}_k, \ldots$ starting at x. We have

$$S_x^0 + \ldots + S_x^{N_x(n)-1} \le n < S_x^0 + \ldots + S_x^{N_x(n)}.$$

Hence

$$\frac{S_x^0 + \ldots + S_x^{N_x(n)-1}}{N_x(n)} \le \frac{n}{N_x(n)} \le \frac{S_x^0 + \ldots + S_x^{N_x(n)}}{N_x(n)}.$$

But since the random variables $\mathcal{E}_k$ are i.i.d. (hence the same is true for the S_x^k), as $n \to \infty$,

$$\frac{S_x^0 + \ldots + S_x^{N_x(n)}}{N_x(n)} \to \mathbb{E}_x(T_x) = m_x \ \ \mathbb{P}_x \text{ a.s.},$$

since $N_x(n) \to +\infty \ \ \mathbb{P}_x$ almost surely. Again from the law of large numbers,

$$\frac{n}{N_x(n)} \to m_x \ \ \mathbb{P}_x \text{ a.s.},$$

that is

$$\frac{N_x(n)}{n} \to \frac{1}{m_x} \ \ \mathbb{P}_x \text{ a.s.}$$

This convergence is also true $\mathbb{P}_\mu$ almost surely, for any initial law μ , since the limit of $N_x(n)/n$ is the same for the chains $\{X_n; n \ge 0\}$ and $\{X_{T_x+n}; n \ge 0\}$.

Now let $F \subset E$. We define $\bar{f} = \sum_{x \in E} \pi_x f(x)$, $c = \sup_x |f(x)|$. We have

$$\begin{aligned}\left| \frac{1}{n} \sum_{k=1}^{n} f(X_k) - \bar{f} \right| &= \left| \sum_{x \in E} \left(\frac{N_x(n)}{n} - \pi_x \right) f(x) \right| \\ &\leq c \sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| + c \sum_{x \notin F} \left(\frac{N_x(n)}{n} + \pi_x \right) \\ &= c \sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| + c \sum_{x \in F} \left(\pi_x - \frac{N_x(n)}{n} \right) + 2c \sum_{x \notin F} \pi_x \\ &\leq 2c \sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| + 2c \sum_{x \notin F} \pi_x .\end{aligned}$$

We choose a finite F such that $\sum_{x \notin F} \pi_x \leq \varepsilon/4c$, and then $N(\omega)$ such that, for all $n \geq N(\omega)$,

$$\sum_{x \in F} \left| \frac{N_x(n)}{n} - \pi_x \right| \leq \frac{\varepsilon}{4c},$$

which proves the result. □

We shall state a central limit theorem in the next section.

2.6 The aperiodic case

We have just shown that in the irreducible, positive recurrent case,

$$\frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k = y\}} \to \pi_y \text{ a.s.},$$

as $n \to \infty$. Taking the expectation under $\mathbb{P}_x$, we deduce that

$$\frac{1}{n} \sum_{k=1}^{n} (P^k)_{xy} \to \pi_y, \quad \forall x, y \in E.$$

We see that the Cesaro means of the $(P^k)_{xy}$ converge. This raises the natural question whether it is true under the above assumptions that, as $n \to \infty$,

$$(P^n)_{xy} \to \pi_y, \quad \forall x, y \in E.$$

It is easily seen this is not the case.

Consider a random walk on $E = \mathbb{Z}/N$, where N is an even integer (we identify 0 and N),

$$X_n = X_0 + Y_1 + \ldots + Y_n,$$

with the Y_n i.i.d. with values in $\{-1, 1\}$; in other words,

$$X_n = (X_0 + Y_1 + \ldots + Y_n) \bmod N.$$

This chain is irreducible, and positive recurrent since E is finite. But $(P^{2k+1})_{xx} = 0$, for all $x \in E$. In the particular case $N = 2$, we have $P^{2k} = I$ and $P^{2k+1} = P$.

In order for the desired convergence to be true, we need an additional assumption:

Definition 6.1 *A state $x \in E$ is said to be aperiodic if there exists N such that*

$$(P^n)_{xx} > 0, \quad \forall n \geq N.$$

Lemma 6.2 *If P is irreducible and there exists an aperiodic state x, then for all $y, z \in E$, there exists M such that $(P^n)_{yz} > 0$, for all $n \geq M$. In particular, all states are aperiodic.*

Proof From the irreducibility, there exist $r, s \in \mathbb{N}$ such that $(P^r)_{yx} > 0$, $(P^s)_{xz} > 0$. Moreover,

$$\left(P^{r+n+s}\right)_{yz} \geq (P^r)_{yx}(P^n)_{xx}(P^s)_{xz} > 0$$

if $n \geq N$. Hence we have the desired property with $M = N + r + s$. □

Remark 6.3 *Suppose we are in the irreducible, positive recurrent case. Let π be the invariant probability, so that $\pi_y > 0$, for all $y \in E$. Hence the fact that there exists N such that, for all $n \geq N$, $(P^n)_{xy} > 0$ is a necessary condition for the convergence $(P^n)_{xy} \to \pi_y$ to hold. We shall now see that it is a sufficient condition.*

Theorem 6.4 *Suppose that P is irreducible, positive recurrent and aperiodic. Let π denote the unique invariant probability. If $\{X_n; n \in \mathbb{N}\}$ is a (μ, P) Markov chain, for all $y \in E$,*

$$\mathbb{P}(X_n = y) \to \pi_y, \quad n \to \infty;$$

in other words,

$$(\mu P^n)_y \to \pi_y,$$

for any initial law μ. In particular, for all $x, y \in E$,

$$(P^n)_{xy} \to \pi_y.$$

Proof We shall use a coupling argument. Let $\{Y_n; n \in \mathbb{N}\}$ be a (π, P) Markov chain, independent of $\{X_n; n \in \mathbb{N}\}$, and $x \in E$ be arbitrary. Let

$$T = \inf\{n \geq 0; \ \ X_n = Y_n = x\}.$$

Step 1. We show that $\mathbb{P}(T < \infty) = 1$. $\{W_n = (X_n, Y_n); n \in \mathbb{N}\}$ is an $(E \times E)$-valued Markov chain, with initial law λ (where $\lambda_{(x,u)} = \mu_x \pi_u$) and transition matrix $\tilde{P}_{(x,u)(y,v)} = P_{xy} P_{uv}$. Since P is aperiodic, for all x, u, y, v, for all n large enough,

$$(\tilde{P}^n)_{(x,u)(y,v)} = (P^n)_{xy}(P^n)_{uv} > 0.$$

Hence $\tilde{P}$ is irreducible. Moreover, $\tilde{P}$ possesses an invariant probability

$$\tilde{\pi}_{(x,u)} = \pi_x \pi_u.$$

Hence, from Theorem 5.4, $\tilde{P}$ is positive recurrent. T is the first passage time of the chain $\{W_n\}$ at the point (x, x); it is finite almost surely.
Step 2. Define

$$Z_n = \begin{cases} X_n, & n \leq T; \\ Y_n, & n > T. \end{cases}$$

By the strong Markov property, both processes $\{X_{T+n}; n \geq 0\}$ and $\{Y_{T+n}; n \geq 0\}$ are (δ_x, P) Markov chains, independent of $(X_0, \ldots, X_T)$. Consequently, $\{Z_n; n \in \mathbb{N}\}$ is, like $\{X_n\}$, a (μ, P) Markov chain.
Step 3. We now conclude. We have the three identities

$$\mathbb{P}(Z_n = y) = \mathbb{P}(X_n = y),$$
$$\mathbb{P}(Y_n = y) = \pi_y,$$
$$\mathbb{P}(Z_n = y) = \mathbb{P}(X_n = y, n \leq T) + \mathbb{P}(Y_n = y, n > T).$$

Hence

$$|\mathbb{P}(X_n = y) - \pi_y| = |\mathbb{P}(Z_n = y) - \mathbb{P}(Y_n = y)| \leq \mathbb{P}(n < T) \to 0,$$

as $n \to \infty$. □

Remark 6.5 *One can define the period of a state $x \in E$ as the greatest common divisor of the integers n such that $(P^n)_{xx} > 0$. One can show with an argument very close to that of Lemma 6.2 that whenever P is irreducible, all states have the same period. A state is said to be aperiodic if its period is 1. The equivalence of the two definitions of aperiodicity is proved in Exercise 10.6.*

We now make precise the rate of convergence in the preceding theorem, under an additional assumption, called Doeblin's condition : there exist $n_0 \in \mathbb{N}$, $\beta > 0$ and a probability ν on E such that

$$(D) \quad (P^{n_0})_{xy} \geq \beta \nu_y, \ \forall x, y \in E.$$

Remark 6.6 *Condition (D) is equivalent to the condition*

$$\exists x \in E, n_0 \geq 1 \text{ such that } \inf_{y \in E} (P^{n_0})_{yx} > 0.$$

This implies that this state x is aperiodic. But it does not imply irreducibility (it is easy to construct a counterexample). We shall see in Exercise 10.4 that this condition implies existence of a unique recurrence class, and of a unique invariant probability.

Lemma 6.7 *If P is irreducible and aperiodic, and E is finite, then condition (D) is satisfied.*

PROOF Choose $x \in E$. For all $y \in E$, there exists n_y such that $n \geq n_y \Rightarrow (P^n)_{yx} > 0$. Let

$$\bar{n} = \sup_{y \in E} n_y, \quad \alpha = \inf_y (P^{\bar{n}})_{yx}.$$

Then $\alpha > 0$, and for all $y \in E$,

$$(P^{\bar{n}})_{yx} \geq \alpha.$$

Hence condition (D) is satisfied with $n_0 = \bar{n}$, $\beta = \alpha$, $\nu = \delta_x$. □

On the other hand, Doeblin's condition is rarely satisfied in the case where $\text{card}(E) = +\infty$, since then typically, for all $n \in \mathbb{N}$, $y \in E$,

$$\inf_{x \in E} (P^n)_{xy} = 0.$$

Theorem 6.8 *Suppose that P is irreducible and satisfies Doeblin's condition (D). Then P is aperiodic, positive recurrent, and if π denotes its invariant probability,*

$$\sum_{y \in E} |(P^n)_{xy} - \pi_y| \leq 2\,(1-\beta)^{[n/n_0]}, \quad \forall x \in E,\ n \in \mathbb{N},$$

where $[n/n_0]$ stands for the integer part of n/n_0.

Let us first introduce a tool which will be useful in the proof of this theorem.

Definition 6.9 *A coupling of two probabilities p and q on E is any pair (X, Y) of E-valued random variables such that p is the law of X and q is the law of Y.*

Lemma 6.10 *Let p and q denote two probabilities on E. We have the identity*

$$||p - q||_1 = 2 \inf_{(X,Y) \text{ coupling of } p,\,q} \mathbb{P}(X \neq Y).$$

PROOF First, note that whenever (X, Y) is a coupling of p and q,

$$\mathbb{P}(X = Y) = \sum_{x \in E} \mathbb{P}(X = Y = x) \leq \sum_{x \in E} p_x \wedge q_x,$$

whence

$$\mathbb{P}(X \neq Y) \geq 1 - \sum_{x \in E} p_x \wedge q_x = \sum_{x \in E} (p_x - q_x)^+$$

and

$$\|p - q\|_1 = \sum_{x \in E} |p_x - q_x| \leq 2\mathbb{P}(X \neq Y).$$

On the other hand, define $\alpha = \sum_{x \in E} p_x \wedge q_x$. If ξ, U, V and W are mutually independent random variables satisfying $\mathbb{P}(\xi = 1) = 1 - \mathbb{P}(\xi = 0) = \alpha$, the law

of U is r defined by $r_x = \alpha^{-1} p_x \wedge q_x$, the law of V is $\bar{p}$ defined by $\bar{p}_x = (1-\alpha)^{-1}(p_x - q_x)^+$, and the law of W is $\bar{q}$ defined by $\bar{q}_x = (1-\alpha)^{-1}(q_x - p_x)^+$, then

$$X = \xi U + (1-\xi)V,$$
$$Y = \xi U + (1-\xi)W$$

is a coupling (X, Y) of p and q, such that $2\mathbb{P}(X \neq Y) = \|p - q\|_1$. □

Proof of Theorem 6.8 The chain being irreducible, Doeblin's condition (D) clearly implies that it is aperiodic.
Step 1. We first show that for any two probabilities μ and ν on E,

$$\|\mu P^n - \nu P^n\|_1 \leq 2(1-\beta)^{[n/n_0]}. \tag{2.1}$$

To prove this, by Lemma 6.10, it suffices to construct a coupling (X_n, Y_n) of the probabilities μP^n and νP^n such that

$$\mathbb{P}(X_n \neq Y_n) \leq (1-\beta)^{[n/n_0]}.$$

Suppose that $n = kn_0 + m$, with $m < n_0$. Given (X_0, Y_0) with the law $\mu \times \nu$ on $E \times E$, for $\ell = 0, 1, \ldots, k-1$, we define $(X_{(\ell+1)n_0}, Y_{(\ell+1)n_0})$ in terms of $(X_{\ell n_0}, Y_{\ell n_0})$ as follows. Let $\{\xi_\ell, U_\ell, V_\ell;\ \ell \geq 0\}$ be a sequence of mutually independent random variables, the ξ_ℓ being Bernoulli with $\mathbb{P}(\xi_\ell = 1) = \beta = 1 - \mathbb{P}(\xi_\ell = 0)$, the law of the U_ℓ being $\bar{m} = \beta^{-1} m$ and the V_ℓ uniform on $[0, 1]$. Define

$$Q_{xy} = (1-\beta)^{-1}((P^{n_0})_{xy} - m_y)$$

and $f : E \times [0,1] \to E$ such that, for all $x, y \in E$, $\{u;\ f(x,u) = y\}$ is a Borel subset of $[0, 1]$, and provided V is uniform on $[0, 1]$, the law of $f(x, V)$ is $Q_{x\cdot}$, $x \in E$. We now let

$$X_{(\ell+1)n_0} = \xi_\ell U_\ell + (1-\xi_\ell) f(X_{\ell n_0}, V_\ell),$$
$$Y_{(\ell+1)n_0} = \xi_\ell U_\ell + (1-\xi_\ell) f(Y_{\ell n_0}, V_\ell).$$

Note that we have really constructed a coupling $(X_{\ell n_0}, Y_{\ell n_0})$ of $\mu P^{\ell n_0}$ and $\nu P^{\ell n_0}$, for $\ell = 0, \ldots, k$, which is such that

$$\mathbb{P}(X_{\ell n_0} \neq Y_{\ell n_0}) \leq \mathbb{P}(\cap_{m=0}^{\ell} \xi_m = 0) = (1-\beta)^\ell.$$

It remains to construct a coupling (X_n, Y_n) of μP^n and νP^n, such that $\{X_n \neq Y_n\} \subset \{X_{kn_0} \neq Y_{kn_0}\}$, which is easy.
Step 2. We now show that for any probability μ on E, $\{\mu P^n;\ n \geq 0\}$ is a Cauchy sequence in the Banach space $\ell^1(E)$. If $\nu = \mu P^m$, it follows from (2.1) that

$$\|\mu P^{n+m} - \mu P^n\|_1 = \|\nu P^n - \mu P^n\|_1 \leq 2c^{n-n_0},$$

where $c = (1-\beta)^{1/n_0}$. The result follows.

Step 3. It follows from the second step that the sequence of probabilities $\{\mu P^n;\ n \geq 0\}$ converges in $\ell^1(E)$, towards a probability π on E. But

$$\pi P = \lim_{n\to\infty} \mu P^{n+1} = \pi,$$

hence π is invariant, and the chain is positive recurrent. Consequently, from (2.1), for any probability μ on E,

$$\|\mu P^n - \pi\|_1 \leq 2(1-\beta)^{[n/n_0]},$$

which establishes the claimed rate of convergence, together with aperiodicity. □

We now state a central limit theorem for irreducible, positive recurrent and aperiodic Markov chains. Such a chain, if it also satisfies

$$\sum_{y\in E} |(P^n)_{xy} - \pi_y| \leq M t^n, \quad x \in E, n \in \mathbb{N}$$

with $M \in \mathbb{R}$ and $0 < t < 1$, is said to be *uniformly ergodic*. We have just shown that Doeblin's condition implies uniform ergodicity. That property implies the central limit theorem.

Theorem 6.11 *Let $\{X_n; n \in \mathbb{N}\}$ be an E-valued Markov chain, with an irreducible transition matrix P, which is moreover uniformly ergodic and aperiodic. Let π denote the unique invariant probability of the chain, and $f : E \to \mathbb{R}$ be such that*

$$\sum_{x\in E} \pi_x f^2(x) < \infty \quad \text{and} \quad \sum_{x\in E} \pi_x f(x) = 0.$$

Then as $n \to \infty$,

$$\frac{1}{\sqrt{n}} \sum_{1}^{n} f(X_k) \text{ converges in law to } \sigma_f Z,$$

where $Z \simeq N(0,1)$ and

$$\begin{aligned}\sigma_f^2 &= \sum_{x\in E} \pi_x (Qf)_x^2 - \sum_x \pi_x (PQf)_x^2 \\ &= 2\sum_x \pi_x (Qf)_x f_x - \sum_x \pi_x f_x^2,\end{aligned}$$

with

$$(Qf)_x = \sum_{n=0}^{\infty} \mathbb{E}_x[f(X_n)], \quad x \in E.$$

Note that the uniform ergodicity property implies that the series which defines the operator Q converges. The reader may consult [22], Corollary 5 and the references in that paper, for a proof, and other conditions under which the theorem holds. One of the other versions (without the uniform ergodicity, but with a stronger moment condition on f) is established in [12], Theorem 3.18. of Chapter 4.

2.7 Reversible Markov chain

Consider the irreducible, positive recurrent case. The Markov property – that the past and future are conditionally independent given the present – tells us that whenever $\{X_n; n \in \mathbb{N}\}$ is a Markov chain, it follows that, for all N, $\{\hat{X}_n^N = X_{N-n}; 0 \leq n \leq N\}$ is also a Markov chain. In general, the time-reversed chain is not homogeneous, unless $\{X_n\}$ is initialized with its invariant probability π.

Proposition 7.1 *Let $\{X_n; n \in \mathbb{N}\}$ be a (π, P) Markov chain whose transition matrix P is supposed to be irreducible, and π be its invariant probability. Then the time-reversed chain $\{\hat{X}_n^N; 0 \leq n \leq N\}$ is a $(\pi, \hat{P})$ Markov chain, with*

$$\pi_y \hat{P}_{yx} = \pi_x P_{xy}, \quad \forall\, x, y \in E.$$

PROOF

$$\begin{aligned}
&\mathbb{P}(\hat{X}_{p+1} = x | \hat{X}_p = y) \\
&= \mathbb{P}(X_n = x | X_{n+1} = y) \\
&= \mathbb{P}(X_{n+1} = y | X_n = x) \times \frac{\mathbb{P}(X_n = x)}{\mathbb{P}(X_{n+1} = y)}.
\end{aligned}$$

□

We say that the chain $\{X_n; n \in \mathbb{N}\}$ is *reversible* if $\hat{P} = P$, which holds if and only if the following *detailed balance equation* is satisfied:

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall\, x, y \in E,$$

where π denotes the invariant probability. It is easily checked that whenever a probability π satisfies this relation, then it is P-invariant. The converse need not be true.

Remark 7.2 *If π is the invariant probability of an irreducible (and hence also positive recurrent) Markov chain, the chain need not be reversible. Suppose that $card(E) \geq 3$. Then there may exist $x \neq y$ such that $P_{xy} = 0 \neq P_{yx}$. Consequently, $\pi_x P_{xy} = 0 \neq \pi_y P_{yx}$. The transitions from y to x of the original chain correspond to the transitions from x to y of the time-reversed chain, hence $P_{yx} \neq 0 \Rightarrow \hat{P}_{xy} \neq 0$, whence $\hat{P} \neq P$.*

Remark 7.3 *Given the transition matrix P of an irreducible positive recurrent Markov chain, one might wish to compute its invariant probability. This problem is not always solvable.*

Another problem, which will appear in the next chapter, is to determine an irreducible transition matrix P whose associated Markov chain admits a given probability π as its invariant probability.

The second problem is rather easy to solve. In fact there are always many solutions. The simplest way to solve it is to look for P such that the associated chain

is reversible with respect to π. In other words, it suffices to find an irreducible transition matrix P such that the quantity $\pi_x P_{xy}$ is symmetric in x, y.

In order to solve the first problem, one can try to find π such that

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x,\ y \in E,$$

which, unlike solving $\pi P = \pi$, implies no summation with respect to x. But that equation has a solution only if the chain is reversible with respect to its unique invariant probability measure, which need not be the case.

Suppose now that we are given a pair (P, π), and that we wish to check whether or not π is the invariant probability of the chain with the irreducible transition matrix P. If the quantity $\pi_x P_{xy}$ is symmetric in x, y, then the answer is yes, and we have an additional property, namely the reversibility. If this is not the case, one needs to check whether or not $\pi P = \pi$. One way to carry out that verification is given by the next proposition, whose elementary proof is left to the reader.

Proposition 7.4 *Let P be an irreducible transition matrix, and π a strictly positive probability on E. For each pair x, $y \in E$, we define*

$$\hat{P}_{xy} = \begin{cases} \frac{\pi_y}{\pi_x} P_{yx}, & \text{if } x \neq y, \\ P_{xx}, & \text{if } x = y. \end{cases}$$

π is the invariant probability of the chain having the transition matrix P, and $\hat{P}$ is the transition matrix of the time-reversed chain if and only if, for all $x \in E$,

$$\sum_{y \in E} \hat{P}_{xy} = 1.$$

2.8 Rate of convergence to equilibrium

Suppose we are in the irreducible, positive recurrent and aperiodic case. We then know that for all x, $y \in E$, $(P^n)_{x,y} \to \pi_y$ as $n \to \infty$, where π denotes the unique invariant probability measure. More generally, we expect that for a large class of functions $f : E \to \mathbb{R}$, $(P^n f)_x \to \langle f, \pi \rangle$ as $n \to \infty$ for all $x \in E$, where, here and below,

$$\langle f, \pi \rangle = \sum_{x \in E} f(x) \pi_x.$$

In this section, we discuss the rate at which the above convergence holds.

2.8.1 The reversible finite state case

Let us first consider the simplest case, in which we assume that E is finite (we write $d = |E|$) and that the process is reversible. We first note that we can identify $L^2(\pi)$ with $\mathbb{R}^d$, equipped with the scalar product

$$\langle f, g \rangle_\pi = \sum_{x \in E} f(x) g(x) \pi_x.$$

Next the reversibility of P is equivalent to the fact that P, as an element of $\mathcal{L}(L^2(\pi))$, is a self-adjoint operator, in the sense that

$$\begin{aligned}\langle Pf, g\rangle_\pi &= \sum_{x,y\in E} P_{x,y} f(y) g(x) \pi_x \\ &= \sum_{x,y\in E} P_{y,x} f(y) g(x) \pi_y \\ &= \langle f, Pg\rangle_\pi,\end{aligned}$$

where we have used the detailed balance equation for the second identity. We now check that the operator norm of P, as an element of $\mathcal{L}(L^2(\pi))$, is at most 1. Indeed, if $\|\cdot\|_\pi$ denotes the usual norm in $L^2(\pi)$,

$$\begin{aligned}\|Pf\|_\pi^2 &= \sum_{x\in E} \left[(Pf)_x\right]^2 \pi_x \\ &= \sum_{x\in E} \left(\mathbb{E}[f(X_t)|X_0 = x]\right)^2 \pi_x \\ &\le \mathbb{E}[f^2(X_t)|X_0 = x]\pi_x \\ &= \sum_{x\in E} f^2(x)\pi_x,\end{aligned}$$

where we have used Schwarz's (or equivalently Jensen's) inequality for the inequality, and the invariance of π for the last identity.

In order to be able to work in $\mathbb{R}^d$ equipped with the Euclidean norm, let us introduce the new $d \times d$ matrix

$$\tilde{P}_{x,y} := \sqrt{\frac{\pi_x}{\pi_y}} P_{x,y}.$$

In matrix notation, $\tilde{P} = \Pi^{1/2} P \Pi^{-1/2}$, where $\Pi_{x,y} = \delta_{x,y}\pi_x$ is a diagonal matrix. Moreover, if we denote by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$, for any $f : E \to \mathbb{R}$ (i.e. f is a collection of real numbers indexed by the d elements of the set E, in other words an element of $\mathbb{R}^d$), denoting $g = \Pi^{-1/2} f$, we have

$$\|\tilde{P} f\|^2 = \sum_{x\in E} (P\Pi^{-1/2} f)_x^2 = \|Pg\|_\pi^2 \le \|g\|_\pi^2 = \|f\|^2.$$

First, note that f is an eigenvector of $\tilde{P}$ if and only if $g = \Pi^{-1/2} f$ is a right eigenvector of P, and $g' = \Pi^{1/2} f$ is a left eigenvector of P associated with the same eigenvalue. We have that $\tilde{P}$ is a symmetric $d \times d$ matrix, whose norm is bounded by 1. Hence, from elementary results in linear algebra, $\tilde{P}$ admits the eigenvalues $-1 \le \lambda_d \le \lambda_{d-1} \le \lambda_2 \le \lambda_1 \le 1$. Let us establish the following lemma.

Lemma 8.1 *We have* $\lambda_2 < \lambda_1 = 1$ *and* $-1 < \lambda_d$.

PROOF If e denotes the vector whose xth component equals $\sqrt{\pi_x}$, we have $(\tilde{P}e)_x = \sqrt{\pi_x}\sum_{y\in E} P_{x,y} = e_x$, and we have an eigenvector for the eigenvalue $\lambda_1 = 1$. This is also an eigenvalue of P, the associated right eigenvector being the vector $\Pi^{-1/2}e = (1, 1, \ldots, 1)$ and the associated left eigenvector being the vector $\Pi^{1/2}e = \pi$.

The equality $\lambda_2 = \lambda_1$ would mean that the eigenspace associated with the eigenvalue 1 would be two-dimensional; in other words, there would exist f linearly independent of e such that $\tilde{P}f = f$, which would imply that $f' = \Pi^{1/2}f$, considered as a row vector, would be such that $f'P = f'$. Now there would exist $\alpha \in \mathbb{R}$ such that $(f' + \alpha\pi)_x \geq 0$, for all $x \in E$. We would have a second invariant measure linearly independent of π, which contradicts irreducibility.

Finally, if -1 were an eigenvalue of $\tilde{P}$, it would also be an eigenvalue of P, hence there would exist f such that $Pf = -f$, then we would have $P^{2n}f = f$, hence $f = \lim_{n\to\infty} P^{2n}f = \langle f, \pi\rangle$. But $g = \Pi^{1/2}f$ is an eigenvector of $\tilde{P}$ associated with the eigenvalue -1, hence it is orthogonal to e, in other words $\langle f, \pi\rangle = 0$, hence $f \equiv 0$, and -1 is not an eigenvalue. □

Denote by $g_1, \ldots, g_d$ the orthonormal basis of $L^2(\pi)$ made of right eigenvectors of P, corresponding respectively to the eigenvalues $1, \lambda_2, \ldots, \lambda_d$. For any $f \in L^2(\pi)$, since $g_1 = (1, \ldots, 1)$,

$$
\begin{aligned}
f - \langle f, \pi\rangle &= \sum_{\ell=2}^{d} \langle f, g_\ell\rangle_\pi g_\ell, \\
Pf - \langle f, \pi\rangle &= \sum_{\ell=2}^{d} \lambda_\ell \langle f, g_\ell\rangle_\pi g_\ell, \\
P^n f - \langle f, \pi\rangle &= \sum_{\ell=2}^{d} \lambda_\ell^n \langle f, g_\ell\rangle_\pi g_\ell, \\
\|P^n f - \langle f, \pi\rangle\|_\pi^2 &= \sum_{\ell=2}^{d} \lambda_\ell^{2n} \langle f, g_\ell\rangle_\pi^2, \\
&\leq \sup_{2\leq\ell\leq d} \lambda_\ell^{2n} \|f - \langle f, \pi\rangle\|_\pi^2,
\end{aligned}
$$

hence we have the following proposition:

Proposition 8.2

$$\|P^n f - \langle f, \pi\rangle\|_\pi \leq (1-\beta)^n \|f - \langle f, \pi\rangle\|_\pi,$$

where $\beta := (1 - \lambda_2) \wedge (1 + \lambda_d)$ *is the spectral gap.*

2.8.2 The general case

More generally, the same is true with

$$\beta := 1 - \sup_{f \in L^2(\pi),\ \|f\|_\pi = 1} \|Pf - \langle f, \pi \rangle\|_\pi .$$

Indeed, with this β, considering only the case $f \neq 0$, since all inequalities below are clearly true for $f = 0$, we have

$$\begin{aligned} \|Pf - \langle f, \pi \rangle\|_\pi &= \left\| P\left(\frac{f}{\|f\|_\pi}\right) - \left\langle \frac{f}{\|f\|_\pi}, \pi \right\rangle \right\|_\pi \times \|f\|_\pi \\ &\leq (1 - \beta)\|f\|_\pi . \end{aligned}$$

Finally, we check that Proposition 8.2 still holds in the general case, with β defined above. Note that

$$\begin{aligned} \|P^{n+1} f - \langle f, \pi \rangle\|_\pi &= \|P[P^n f - \langle f, \pi \rangle]\|_\pi \\ &\leq (1 - \beta)\|P^n f - \langle f, \pi \rangle\|_\pi . \end{aligned}$$

The result follows by induction.

In practice the problem is to estimate the spectral gap β precisely. We shall describe one such result in Section 3.3. The notion of spectral gap will appear again in Section 7.10.

The content of this section was inspired by the treatment in [37]. For a more complete introduction to this topic, see [36].

2.9 Statistics of Markov chains

The aim of this section is to introduce the basic notions for the estimation of the parameters of a Markov chain.

We have seen that, for all $n > 0$, the law of the random vector $(X_0, X_1, \ldots, X_n)$ depends only on the initial law μ and on the transition matrix P. We are interested in the conditions under which one can estimate the pair (μ, P), given the observation of $(X_0, X_1, \ldots, X_n)$, in such a way that the error tends to zero, as $n \to \infty$.

Let us first discuss the estimation of the invariant probability μ. For any $x \in E$,

$$\hat{\mu}_x^n = \frac{1}{n+1} \sum_{\ell=0}^{n} \mathbf{1}_{\{X_\ell = x\}}$$

is a consistent estimator of μ_x, since the following is an immediate consequence of the ergodic theorem:

Proposition 9.1 *For any $x \in E$, $\hat{\mu}_x^n \to \mu_x$ almost surely, as $n \to \infty$.*

Let us now discuss the estimation of the P_{xy}, $x, y \in E$. We choose the estimator

$$\hat{P}_{xy}^n = \frac{\sum_{\ell=0}^{n-1} \mathbf{1}_{\{X_\ell = x, X_{\ell+1} = y\}}}{\sum_{\ell=0}^{n-1} \mathbf{1}_{\{X_\ell = x\}}}.$$

We have the following proposition.

Proposition 9.2 *For any $x, y \in E$, $\hat{P}_{xy}^n \to P_{xy}$ almost surely as $n \to \infty$.*

Proof We clearly have

$$\hat{P}_{xy}^n = \left(\frac{1}{n} \sum_{\ell=0}^{n-1} \mathbf{1}_{\{X_\ell = x\}} \right)^{-1} \frac{1}{n} \sum_{\ell=0}^{n-1} \mathbf{1}_{\{X_\ell = x, X_{\ell+1} = y\}}.$$

We know that

$$\frac{1}{n} \sum_{\ell=0}^{n-1} \mathbf{1}_{\{X_\ell = x\}} \to \mu_x.$$

For $n \geq 0$, define $\widetilde{X}_n = (X_n, X_{n+1})$. It is not very hard to check that $\{\widetilde{X}_n;\ n \geq 0\}$ is an irreducible and positive recurrent $\widetilde{E} = \{(x, y) \in E \times E;\ P_{xy} > 0\}$-valued Markov chain, with transition matrix $\widetilde{P}_{(x,y)(u,v)} = \delta_{yu} P_{uv}$. It admits the invariant probability $\widetilde{\mu}_{(x,y)} = \mu_x P_{xy}$. The ergodic theorem applied to the chain $\{\widetilde{X}_n\}$ implies that almost surely, as $n \to \infty$,

$$\frac{1}{n} \sum_{\ell=0}^{n-1} \mathbf{1}_{\{X_\ell = x, X_{\ell+1} = y\}} \to \mu_x P_{xy}.$$

□

2.10 Exercises

Exercise 10.1 *Show that the $E = \{1, 2, 3\}$-valued Markov chain $\{X_n;\ n \in \mathbb{N}\}$ whose transition matrix is*

$$P = \begin{pmatrix} 1 & 0 & 0 \\ p & 1 - p - q & q \\ 0 & 0 & 1 \end{pmatrix}, \quad p, q > 0,\ p + q < 1,$$

starting at $X_0 = 2$, first changes its value at a random time $T \geq 1$ whose law is geometric. Show also that X_T is independent of T, and give the law of X_T. Finally, show that $X_t = X_T$ if $t \geq T$.

Exercise 10.2 *Let $\{X_n;\ n \in \mathbb{N}\}$ be an $E = \{1, 2, 3, 4, 5\}$-valued Markov chain, with transition matrix*

$$P = \begin{pmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{pmatrix}.$$

Find the equivalence classes, the transient and recurrent states, and all invariant measures of $\{X_n\}$.

Exercise 10.3 *Consider a Markov chain $\{X_n;\ n \in \mathbb{N}\}$ taking values in the finite state $E = \{1, 2, 3, 4, 5, 6\}$, with a transition matrix P whose off-diagonal entries are given by*

$$P = \begin{pmatrix} \cdot & 2/3 & 1/3 & 0 & 0 & 0 \\ 1/4 & \cdot & 0 & 0 & 1/5 & 2/5 \\ 0 & 0 & \cdot & 1/2 & 0 & 0 \\ 0 & 0 & 2/3 & \cdot & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdot & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & \cdot \end{pmatrix}.$$

1. *Find the diagonal entries of P.*
2. *Show that E can be partitioned into three equivalence classes to be specified, of which one ($\mathcal{T}$) is transient and two ($\mathcal{R}_1$ and $\mathcal{R}_2$) are recurrent.*
3. *Find an invariant probability whose support is $\mathcal{R}_1$ and another whose support is $\mathcal{R}_2$. Find all invariant probabilities.*

Exercise 10.4 *Let P be a Markovian matrix over a finite or countable set E, which satisfies Doeblin's condition (D) of Section 2.6.*

1. *Suppose first that condition (D) is satisfied with $n_0 = 1$. Show that there exists at least one recurrent state, which is visited infinitely often by the chain, for any starting point. Deduce that the chain has a unique recurrent class. (Hint: first show that there exist $x \in E$, $\beta > 0$ such that the chain can be simulated by setting, at each time n, $X_n = x$ with probability β, and following a certain Markovian transition with probability $1 - \beta$.)*
2. *Show that the result is still true in the general case of condition (D). (Hint: consider the subchain $\{X_{kn_0};\ k = 0, 1, \ldots\}$.)*

Exercise 10.5 *Show that whenever x is recurrent, $\sum_{n\geq 0}(P^n)_{xy}$ equals $+\infty$ if and only if $x \leftrightarrow y$, and equals 0 if and only if $x \not\to y$.*

Exercise 10.6 (Equivalence of the two definitions of aperiodicity) *Let $x \in E$. Define $N_x = \{n;\ (P^n)_{xx} > 0\}$.*

1. *Show that whenever N_x contains two consecutive integers, the greatest common divisor of the elements of N_x is* 1.

2. *Show that if $n, n+1 \in N_x$, then $\{n^2, n^2+1, n^2+2, \ldots\} \subset N_x$.*

3. *Show that if the greatest common divisor of the elements of N_x is* 1*, then there exists $n \in \mathbb{N}$ such that $\{n, n+1\} \subset N_x$.*

4. *Conclude that the two definitions of aperiodicity of a state x are equivalent.*

Exercise 10.7 *Consider an $E = \{1, 2, 3, 4, 5, 6\}$-valued Markov chain $\{X_n;\ n \in \mathbb{N}\}$ with transition matrix P, whose off-diagonal entries are specified by*

$$P = \begin{pmatrix} \cdot & 1/2 & 0 & 0 & 0 & 0 \\ 1/3 & \cdot & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 7/8 & 0 \\ 1/4 & 1/4 & 0 & \cdot & 1/4 & 1/4 \\ 0 & 0 & 3/4 & 0 & \cdot & 0 \\ 0 & 1/5 & 0 & 1/5 & 1/5 & \cdot \end{pmatrix}.$$

1. *Find the diagonal terms of P.*

2. *Find the equivalence classes of the chain.*

3. *Show that* 4 *and* 6 *are transient states, and that the other states can be grouped into two recurrent classes to be specified. In the sequel, we let $\mathcal{T} = \{4, 6\}$, $\mathcal{C}$ be the recurrent class containing* 1*, and $\mathcal{C}'$ the other recurrent class. For all $x, y \in E$, define $\rho_x := \mathbb{P}_x(T < \infty)$, where $T := \inf\{n \geq 0;\ X_n \in \mathcal{C}\}$.*

4. *Show that*

$$\rho_x = \begin{cases} 1, & \text{if } x \in \mathcal{C}, \\ 0, & \text{if } x \in \mathcal{C}', \end{cases}$$

and that $0 < \rho_x < 1$ if $x \in \mathcal{T}$.

5. *Using the decomposition $\{T < \infty\} = \{T = 0\} \cup \{T = 1\} \cup \{2 \leq T < \infty\}$ and conditioning in the computation of $\mathbb{P}_x(2 \leq T < \infty)$ by the value of X_1, establish the formula*

$$\rho_x = \sum_{y \in E} P_{xy} \rho_y, \qquad \text{if } x \in \mathcal{T}.$$

6. *Compute ρ_4 and ρ_6.*

7. *Deduce (without any serious computation!) the values of $\mathbb{P}_4(T_{\mathcal{C}'} < \infty)$ and $\mathbb{P}_6(T_{\mathcal{C}'} < \infty)$, where $T_{\mathcal{C}'} := \inf\{n \geq 0;\ X_n \in \mathcal{C}'\}$.*

Exercise 10.8 *Consider an $E = \{1, 2, 3, 4, 5, 6\}$-valued Markov chain $\{X_n;\ n \in \mathbb{N}\}$ with transition matrix P, whose off-diagonal entries are given by*

$$P = \begin{pmatrix} \cdot & 1/4 & 1/3 & 0 & 0 & 0 \\ 1/4 & \cdot & 0 & 1/4 & 1/3 & 0 \\ 1/2 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdot & 1/2 & 1/3 \\ 0 & 0 & 0 & 1/2 & \cdot & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/4 & \cdot \end{pmatrix}.$$

1. *Find the diagonal entries of P.*
2. *Show that E is the union of two equivalence classes to be specified, one ($\mathcal{R}$) being recurrent and the other ($\mathcal{T}$) transient.*
3. *Define $T := \inf\{n \geq 0;\ X_n \in \mathcal{R}\}$ and $h_x = \mathbb{E}_x(T)$, for $x \in E$. Show that $h_x = 0$ for $x \in \mathcal{R}$, and that $1 < h_x < \infty$ for $x \in \mathcal{T}$.*
4. *Show that, for all $x \in \mathcal{T}$,*

$$h_x = 1 + \sum_{y \in E} P_{xy} h_y.$$

Deduce the values of h_x, $x \in \mathcal{T}$.

Exercise 10.9 *Given $0 < p < 1$, we consider an $E = \{1, 2, 3, 4\}$-valued Markov chain $\{X_n;\ n \in \mathbb{N}\}$ with transition matrix P given by*

$$P = \begin{pmatrix} p & 1-p & 0 & 0 \\ 0 & 0 & p & 1-p \\ p & 1-p & 0 & 0 \\ 0 & 0 & p & 1-p \end{pmatrix}.$$

1. *Show that the chain $\{X_n\}$ is irreducible and recurrent.*
2. *Compute its unique invariant probability π.*
3. *Show that the chain is aperiodic. Deduce that P^n tends, as $n \to \infty$, towards the matrix*

$$\begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{pmatrix}.$$

4. *Compute P^2. Show that this transition matrix coincides with the above limit. Determine the law of X_2, as well as that of X_n, $n \geq 2$.*
5. *Define $T_4 = \inf\{n \geq 1;\ X_n = 4\}$. Compute $\mathbb{E}_4(T_4)$.*

Exercise 10.10 *Consider an $E = \{0, 1, 2, 3, 4\}$-valued Markov chain $\{X_n;\ n \in \mathbb{N}\}$ with transition matrix*

$$P = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ p & 0 & \frac{1-p}{2} & 0 & \frac{1-p}{2} \\ p & \frac{1-p}{2} & 0 & \frac{1-p}{2} & 0 \\ p & 0 & \frac{1-p}{2} & 0 & \frac{1-p}{2} \\ p & \frac{1-p}{2} & 0 & \frac{1-p}{2} & 0 \end{pmatrix},$$

where $0 < p < 1$. Let $T := \inf\{n \geq 1;\ X_n = 0\}$.

1. *Show that the chain $\{X_n\}$ is irreducible and recurrent. Denote its invariant probability by π.*

2. *Show that under $\mathbb{P}_0$, the law of T is a geometric law to be specified. Show that $\mathbb{E}_0(T) = (p + 1)/p$.*

3. *Let*

$$N_n = \sum_{k=1}^{n} \mathbf{1}_{\{X_k = 0\}}, \quad M_n = \sum_{k=1}^{n} \mathbf{1}_{\{X_k \neq 0\}}.$$

   Compute the limits as $n \to \infty$ of $n^{-1}N_n$ and $n^{-1}M_n$.

4. *Give an intuitive argument to support the identity*

$$\pi_1 = \pi_2 = \pi_3 = \pi_4.$$

   Deduce the probability π, exploiting this identity.

5. *Show the following general result. If there exists a one-to-one mapping τ from E into itself, such that*

$$P_{\tau x, \tau y} = P_{xy}, \quad \forall x, y \in E,$$

   then the invariant probability π has the property $\pi_{\tau x} = \pi_x$, $x \in E$. Deduce a rigorous argument for the result in part 4.

Exercise 10.11 (Random walk in $\mathbb{Z}$) *Let*

$$X_n = X_0 + Y_1 + \ldots + Y_n,$$

where the X_n take their values in $\mathbb{Z}$, the Y_n in $\{-1, 1\}$, $X_0, Y_1, \ldots, Y_n, \ldots$ being a sequence of independent random variables, and for all n,

$$\mathbb{P}(Y_n = 1) = p = 1 - \mathbb{P}(Y_n = -1), \quad 0 < p < 1.$$

1. *Show that the chain $\{X_n\}$ is irreducible.*

2. *Show that if $p \neq 1/2$, the chain is transient (use the law of large numbers).*

3. *Consider the case $p = 1/2$. Show that the chain is recurrent (evaluate $\sum_{n\geq 1}(P^n)_{00}$ using Stirling's formula $n! \simeq \sqrt{2\pi n}(n/e)^n$). Show that the chain is null recurrent (look for an invariant measure). Determine the quantities*

$$\limsup_{n\to\infty} X_n \quad \text{and} \quad \liminf_{n\to\infty} X_n.$$

Exercise 10.12 (Random walk in $\mathbb{Z}^d$) *Let*

$$X_n = X_0 + Y_1 + \ldots + Y_n,$$

where the X_n take their values in $\mathbb{Z}^d$,the Y_n being i.i.d., globally independent of X_0, and their law specified by

$$\mathbb{P}(Y_n = \pm e_i) = (2d)^{-1}, \quad 1 \leq i \leq d,$$

where $\{e_1, \ldots, e_d\}$ is the canonical basis of $\mathbb{Z}^d$.

1. *Show that the common characteristic function of the Y_n is given by*

$$\phi(t) = d^{-1}\sum_{j=1}^{d} \cos(t_j),$$

and that

$$(P^n)_{00} = (2\pi)^{-d}\int_{[-\pi,\pi]^d} \phi^n(t)dt.$$

2. *Deduce that, for all $0 < r < 1$,*

$$\sum_{n\geq 0} r^n (P^n)_{00} = (2\pi)^{-d}\int_{[-\pi,\pi]^d} (1 - r\phi(t))^{-1}dt.$$

3. *Show that, for all $\alpha > 0$, the mapping*

$$(r, t) \to (1 - r\phi(t))^{-1}$$

is bounded on $]0, 1] \times ([-\pi, \pi]^d \backslash C_\alpha)$, where $C_\alpha = \{t \in \mathbb{R}^d; \|t\| \leq \alpha\}$, and that whenever $\|t\|$ is sufficiently small, $r \to (1 - r\phi(t))^{-1}$ is positive and increasing.

4. *Deduce from the fact that $1 - \phi(t) \simeq \|t\|^2/2$, as $t \to 0$, that $\{X_n\}$ is an irreducible $\mathbb{Z}^d$-valued Markov chain, which is null recurrent if $d = 1, 2$, and transient if $d \geq 3$.*

Exercise 10.13 *Consider again the $\mathbb{Z}$-valued random walk from Exercise 10.11 in the symmetric case ($p = 1/2$). The goal of this exercise is to establish the null recurrence of the walk by a method which is completely different from that of Exercise 10.11. Suppose for simplicity that $X_0 = x \in \mathbb{Z}$.*

For all $a, b \in \mathbb{Z}$ with $a < x < b$, let

$$T_{a,b} = \inf\{n \geq 0;\ X_n \notin]a, b[\},$$

$$T_a = \inf\{n \geq 0;\ X_n = a\},$$

$$T_b = \inf\{n \geq 0;\ X_n = b\}.$$

We note that

$$X_{n \wedge T_{a,b}} = x + \sum_{k=1}^{n} Y_k \mathbf{1}_{\{T_{a,b} > k-1\}}.$$

1. *Show that the random variables Y_k and $\mathbf{1}_{\{T_{a,b} > k-1\}}$ are independent. Deduce that*
$$\mathbb{E}X_{n \wedge T_{a,b}} = x.$$

2. *Show that $|X_{n \wedge T_{a,b}}| \leq \sup(|a|, |b|)$, $T_{a,b} < \infty$ almost surely, and*
$$\mathbb{E}X_{T_{a,b}} = x.$$

3. *Establish the identities*
$$\mathbb{P}(X_{T_{a,b}} = a) = \frac{b-x}{b-a}, \quad \mathbb{P}(X_{T_{a,b}} = b) = \frac{x-a}{b-a}.$$

4. *Show that $\mathbb{P}(T_a < T_n) \to 1$, as $n \to \infty$.*

5. *Show that $T_a < \infty$ almost surely, and similarly that $T_b < \infty$ almost surely. Deduce that the chain is recurrent.*

6. *In the sequel we consider without loss of generality the case $x = 0$, for the sake of notational simplicity. Show that, for all $n \geq 1$,*
$$X^2_{n \wedge T_{a,b}} = \sum_{k=1}^{n} (1 - 2X_{k-1}Y_k) \mathbf{1}_{\{T_{a,b} > k-1\}}.$$

7. *Deduce that $\mathbb{E}(X^2_{T_{a,b}}) = \mathbb{E}(T_{a,b}) = -ab$ and that, for all $a \in \mathbb{Z}$, $\mathbb{E}(T_a) = +\infty$, which shows that the chain is null recurrent.*

Exercise 10.14 (Reflected random walk) *With the $\{Y_n\}$ as in Exercise 10.11, define the $\mathbb{N}$-valued Markov chain $\{X_n\}$ by the recurrence formula*

$$X_{n+1} = X_n + \mathbf{1}_{\{X_n > 0\}} Y_{n+1} + \mathbf{1}_{\{X_n = 0\}}.$$

Assume that $X_0 \in \mathbb{N}$. Denote by $\{X'_n\}$ the (unreflected) random walk from Exercise 10.11, with the same X_0 and the same $\{Y_n\}$. Below we shall use freely the results from Exercise 10.11.

1. *Show that the chain $\{X_n\}$ is irreducible, as an $\mathbb{N}$-valued chain. Give its transition matrix.*

2. *Show that $X_n \geq X'_n$ almost surely, for all n. Conclude that $\{X_n\}$ is transient in the case $p > 1/2$.*

3. *Let $T = \inf\{n \geq 0;\ X_n = 0\}$. Show that $X_n = X'_n$ whenever $T \geq n$. Conclude that the chain is recurrent in the case $p \leq 1/2$ (one can, for example, show that the state 1 is recurrent).*

4. *Show that the chain is null recurrent in the case $p = 1/2$, and positive recurrent in the case $p < 1/2$. (Hint: check that in the first case $(1/2, 1, 1, 1, \ldots)$ is an invariant measure, and in the second case that the probability μ defined by*

$$\mu_0 = \frac{1-2p}{2(1-p)}, \quad \mu_x = \frac{1-2p}{2}\frac{p^{x-1}}{(1-p)^{x+1}}, x \geq 1,$$

is an invariant measure.)

Exercise 10.15 (Birth and death Markov chain) *Let $\{X_n\}$ be an $E = \mathbb{N}$-valued Markov chain with transition P given by*

$$P_{x,x-1} = q_x, \quad P_{x,x} = r_x, \quad P_{x,x+1} = p_x,$$

where for all $x \in \mathbb{N}$, $p_x + r_x + q_x = 1$, $q_0 = 0$, $q_x > 0$ if $x > 0$, and $p_x > 0$ for all $x \in \mathbb{N}$.

For $x \in \mathbb{N}$, let $\tau_x = \inf\{n \geq 0;\ X_n = x\}$. Given three states a, x and b such that $a \leq x \leq b$, define $u(x) = \mathbb{P}_x(\tau_a < \tau_b)$. Let $\{\gamma_x;\ x \in \mathbb{N}\}$ be defined by $\gamma_0 = 1$ and, for $x > 0$, $\gamma_x = (q_1 \cdots q_x)/(p_1 \cdots p_x)$.

1. *Show that the chain is irreducible.*

2. *For $a < x < b$, establish a relation between $u(x) - u(x+1)$ and $u(x-1) - u(x)$. Compute $u(a) - u(b)$ in terms of the γ_x, and deduce that, for $a < x < b$,*

$$u(x) = \frac{\sum_{y=x}^{y=b-1} \gamma_y}{\sum_{y=a}^{y=b-1} \gamma_y}.$$

Consider the particular case where $p_x = q_x$ for all $x > 0$.

3. *Compute $\mathbb{P}_1(\tau_0 = \infty)$ and show that the chain is recurrent if and only if $\sum_0^\infty \gamma_y = +\infty$.*

4. *Find the invariant measures, and deduce that the chain is positive recurrent if and only if*

$$\sum_{x=1}^{\infty} \frac{p_0 p_1 \cdots p_{x-1}}{q_1 q_2 \cdots q_x} < \infty.$$

5. *Show that in the positive recurrent case, the chain is reversible. (Hint: first note that, for $x > 0$, the relation $\pi_x = (\pi P)_x$ can be written*

$$\pi_x P_{x,x-1} + \pi_x P_{x,x+1} = \pi_{x-1} P_{x-1,x} + \pi_{x+1} P_{x+1,x};$$

then consider the case $x = 0$, and show by recurrence that

$$\pi_x P_{x,x+1} = \pi_{x+1} P_{x+1,x}, \quad \forall x \geq 0.)$$

Exercise 10.16 (Queue) *Consider a discrete time queue such that, at each time $n \in \mathbb{N}$, one customer arrives with probability p ($0 < p < 1$) and no customer arrives with probability $1 - p$. During each unit time interval when at least one customer is present, one customer is served and leaves the queue with probability q, $0 < q < 1$, and no customer leaves the queue with probability $1 - q$ (a customer who arrives at time n leaves at the earliest at time $n + 1$). All the above events are mutually independent. Denote by X_n the number of customers in the queue at time n.*

1. *Show that $\{X_n;\ n \in \mathbb{N}\}$ is an irreducible $E = \mathbb{N}$-valued Markov chain. Determine its transition matrix P_{xy}, $x, y \in \mathbb{N}$.*

2. *Give a necessary and sufficient condition on p and q for the chain $\{X_n\}$ to possess an invariant probability. We assume below that this condition is satisfied. Specify the unique invariant probability $\{\pi_x;\ x \in \mathbb{N}\}$ of the chain $\{X_n\}$.*

3. *Compute $\mathbb{E}_\pi(X_n)$.*

4. *Customers are served according to the order in which they arrive. Denote by T the sojourn time in the queue of a customer who arrives at an arbitrary fixed time. Assuming that the queue is initialized with its invariant probability, what is the expectation of T?*

Exercise 10.17 (Queue) *Consider a queue at a counter. X_n denotes the number of customers in the queue at time n. Between times n and $n + 1$, Y_{n+1} new customers join the queue, and provided $X_n > 0$, Z_{n+1} customers leave the queue. Assume that X_0, Y_1, Z_1, Y_2, Z_2, ... are mutually independent, the Y_n all having the same law, such that $0 < \mathbb{P}(Y_n = 0) < 1$, and the Z_n satisfying $\mathbb{P}(Z_n = 1) = p = 1 - \mathbb{P}(Z_n = 0)$.*

1. *Show that $(X_n; n \in \mathbb{N})$ is a Markov chain, and give its transition matrix.*

2. *Let φ denote the common characteristic function of the Y_n, ρ that of the Z_n, Ψ_n that of X_n. Compute Ψ_{n+1} in terms of Ψ_n, φ and ρ.*

3. *Show that there is a unique invariant probability if and only if $\mathbb{E}(Y_1) < p$, and determine its characteristic function.*

Exercise 10.18 (Queue) *Let X denote the random number of individuals in a given population, and $\phi(u) = \mathbb{E}[u^X]$, $0 \leq u \leq 1$, its generating function. Each individual is selected with probability q ($0 < q < 1$), independently of the others. Let Y denote the number of individuals selected in the initial population of X individuals.*

1. *Show that the generating function ψ of Y (defined as $\psi(u) = \mathbb{E}[u^Y]$) is given by*

$$\psi(u) = \phi(1 - q + qu).$$

Consider a service system (equipped with an infinite number of servers), and denote by X_n ($n = 0, 1, 2, \ldots$) the number of customers present in the system at time n. Assume that, at time $n + \frac{1}{3}$, each of the X_n customers leaves the system with probability $1 - p$, and stays with probability p (independently of the others, and of all the other events), denote by X'_n the number of remaining customers, and assume that, at time $n + \frac{2}{3}$, Y_{n+1} new customers join the queue. Assume that the random variables $X_0, Y_1, Y_2, \ldots$ are mutually independent, and globally independent of the service times, and that the joint law of the Y_n is the Poisson distribution with parameter $\lambda > 0$ (i.e. $\mathbb{P}(Y = k) = e^{-\lambda}\lambda^k/k!$ and $\mathbb{E}[u^{Y_n}] = \exp[\lambda(u - 1)]$).

2. *Show that $\{X_n;\ n \geq 0\}$ is an irreducible $E = \mathbb{N}$-valued Markov chain.*

3. *Compute $\mathbb{E}[u^{X_{n+1}} | X_n = x]$ in terms of u, p, λ and x.*

4. *Denote by $\phi_n(u) = \mathbb{E}[u^{X_n}]$ the generating function of X_n. Compute ϕ_{n+1} in terms of ϕ_n, and show that*

$$\phi_n(u) = \exp\left[\lambda(u - 1)\sum_0^{n-1} p^k\right]\phi_0(1 - p^n + p^n u).$$

5. *Show that $\rho(u) = \lim_{n\to\infty} \phi_n(u)$ exists and does not depend on ϕ_0, and that ρ is the generating function of a Poisson distribution whose parameter is to be specified in terms of λ and p.*

6. *Show that $\{X_n;\ n \geq 0\}$ is positive recurrent and specify its invariant probability.*

Exercise 10.19 *Let $X_0, A_0, D_0, A_1, D_1, \ldots$ be $\mathbb{N}$-valued mutually independent random variables. The D_n are Bernoulli random variables with parameter q, that is, $\mathbb{P}(D_n = 1) = 1 - \mathbb{P}(D_n = 0) = q$, $0 < q < 1$. The A_n all have the same law defined by $\mathbb{P}(A_n = k) = r_k$, $k \in \mathbb{N}$, where $0 \leq r_k < 1$, $0 < r_0 < 1$ and $\sum_{k=0}^{\infty} r_k = 1$.*

Assume that $p = \sum_k kr_k < \infty$. *Consider the sequence of random variables* $\{X_n; n \in \mathbb{N}\}$ *defined by*

$$X_{n+1} = (X_n + A_n - D_n)^+, \quad n \geq 0,$$

with the usual notation $x^+ = \sup(x, 0)$.

1. *Show that* $\{X_n;\ n \in \mathbb{N}\}$ *is an* $E = \mathbb{N}$*-valued Markov chain. Give its transition matrix* P, *and show that the chain is irreducible.*

Assume from now on that $X_0 = 0$. *Let* $T = \inf\{n > 0;\ X_n = 0\}$. *Define* $S_n = \sum_{k=0}^{n-1}(A_k - D_k)$.

2. *Show that* $X_n \geq S_n$, *and that* $X_{n+1} = S_{n+1}$ *on the event* $\{T > n\}$.

3. *Show that* $S_n/n \to p - q$ *almost surely, as* $n \to \infty$.

4. *Show that whenever* $p < q$, $T < \infty$ *almost surely.*

5. *Assume that* $p > q$. *Show that* $\{X_n; n \in \mathbb{N}\}$ *visits* 0 *at most a finite number of times.*

6. *In the case* $p \neq q$, *specify when the chain is recurrent, and when it is transient.*

Assume from now on that $\mathbb{P}(A_n = 1) = 1 - \mathbb{P}(A_n = 0) = p$, *where* $0 < p < 1$ *(*p *is again the expectation of* A_n*).*

7. *Specify the transition matrix* P *in this case.*

8. *Show that if* $p = q$, *the chain is null recurrent. (Hint: use the result of Exercise 10.11, part 3, in order to show the recurrence, and then look for an invariant measure.)*

9. *Assume that* $p < q$. *Show that the chain has a unique invariant probability* π *on* $\mathbb{N}$, *and that* $\pi_k = (1 - a)a^k$, *with* $a = p(1 - q)/q(1 - p)$. *(Hint: first establish a recurrence relation for the sequence* $\Delta_k = \pi_k - \pi_{k+1}$.*) Show that the chain is positive recurrent.*

Exercise 10.20 (Discrete Aloha) *The aim of this exercise is to study the following communication protocol: users arrive at times* $\{1, 2, \ldots, n, \ldots\}$ *in order to transmit a message through a channel, which has the capacity to transmit only one message at a time. If several users try to transmit a message at the same time, no message is transmitted, and each user knows this and makes a new attempt later. We look for a 'distributed' retransmission policy, such that each user may decide when to try to retransmit, without knowing the intentions of other users. The 'discrete Aloha' protocol prescribes that each user whose message has been blocked at time* n *makes a new attempt at time* $n + 1$ *with probability* p. *If he decides not to try at time* $n + 1$, *he again makes an attempt at time* $n + 2$ *with probability* p, *and so on until by chance he does try. Let* Y_n *denote the number of 'new' messages (i.e. which*

have not been presented before) arriving at time n. We assume that the Y_n are i.i.d., with $\mathbb{P}(Y_n = i) = a_i$, $i \in \mathbb{N}$, and $\mathbb{E}(Y_n) > 0$. Let X_n denote the number of delayed messages which are waiting to be transmitted at time n.

1. *Show that $\{X_n\}$ is a Markov chain, and give its transition matrix.*

2. *Show that $\{X_n\}$ is irreducible, but not positive recurrent.*

Exercise 10.21 (Programming) *Consider again the queue in Exercise 10.16.*

1. *Simulate and plot a trajectory $\{X_n;\ n \geq 0\}$ from $n = 1$ to $n = 1000$, with $p = 1/2$ and successively $q = 3/5, 7/13, 15/29, 1/2$.*

2. *Since $\{X_n\}$ is irreducible, positive recurrent and aperiodic, $(P^n)_{yx} \to \pi_x$. Plot either the empirical histogram or the empirical distribution function of $(P^n)_{y\cdot}$, for $n = 100, 500, 1000$, and a sample size of 10^4. Show the histogram (or the distribution function) of π on the same plot. Treat the cases $p = 1/2$, $q = 3/5, 7/13$.*

3. *Graphically compare the quantities*

$$n^{-1} \sum_{k=1}^{n} \mathbf{1}_{\{X_k = x\}}, \quad x \in \mathbb{N},$$

and the histogram of π, for $n = 10^3, 10^4, 10^5$. Treat the cases $p = 1/2$, $q = 3/5, 7/13$. For each value of q, choose the interval of values of x from the previous results.

Exercise 10.22 (Ordering a database) *Suppose that a computer memory contains n items $1, 2, \ldots, n$. The memory receives successive requests, each consisting of one of the items. The closer the item is to the top of the list, the faster the access is. Assume that the successive requests are i.i.d. random variables. If the common law of those random variables were known, the best choice would be to order the data in decreasing order of their associated probability of being requested. But this probability $(p_1, p_2, \ldots, p_n)$ is either unknown or slowly varying. Assume that $p_k > 0$, for all $k \in \{1, 2, \ldots, n\}$.*

We need to choose a method of replacement of the data after they are requested, in such a way that in the long run the time taken to get the requested data will be as small as possible.

We will compare two such methods. The first involves systematically replacing any item which has been requested at the top of the list. The second involves moving each item which has been requested one step ahead. In both cases, we have an irreducible Markov chain with values in the set E of all permutations of the set $\{1, 2, \ldots, n\}$. Denote by Q (P) the transition matrix of the first (second) chain, and

by π (μ) the associated invariant measure. We associate with the Markov chain with transition matrix Q, the quantity

$$J_Q \stackrel{def}{=} \sum_{k=1}^{n} \pi(\textit{position of } k)\, p_k,$$

where $\pi(\textit{position of } k)$ is the expectation under π of the position of the element k. We associate with the Markov chain with transition matrix P, the quantity

$$J_P \stackrel{def}{=} \sum_{k=1}^{n} \mu(\textit{position of } k)\, p_k.$$

1. *Show that the chain with transition matrix Q is not reversible.*
2. *Show that any irreducible and positive recurrent Markov chain which satisfies the following conditions is reversible:*

   (i) $P_{k\ell} > 0 \Leftrightarrow P_{\ell k} > 0$;

   (ii) for any excursion $k, k_1, k_2, \ldots, k_m, k$,

$$P_{kk_1} \prod_{i=2}^{m} P_{k_{i-1}k_i} P_{k_m k} = P_{kk_m} \prod_{i=m-1}^{1} P_{k_{i+1}k_i} P_{k_1 k}.$$

   This is known as the 'Kolmogorov cycle condition'.

3. *Show that P satisfies (i) and (ii).*
4. *Show that the second procedure is preferable, in the sense that $J_P < J_Q$.*

3

Stochastic algorithms

Introduction

The aim of this chapter is to present some algorithms which are based on the simulation of a Markov chain. Sections 3.1–3.3 are devoted to the study of the Markov chain Monte Carlo (MCMC) method, which is very much used for simulating random variables with values in a finite but very large set, in cases where the usual Monte Carlo method is not feasible. In particular, it has in recent years become the main numerical method in Bayesian statistics. Section 3.4 is devoted to simulated annealing, a random optimization algorithm.

In this chapter, all the Markov chains take their values in a finite set E.

3.1 Markov chain Monte Carlo

We have seen in Chapter 1 that one way to compute a sum of the type

$$\sum_{x \in E} f(x)\pi_x,$$

where $\{\pi_x; x \in E\}$ is a probability, is to approximate it by

$$\frac{1}{n}\sum_{k=1}^{n} f(U_k),$$

where $(U_1, U_2, \ldots)$ is a sequence of i.i.d. random variables with common law π, and the convergence of the algorithm is guaranteed by the strong law of large numbers. In a number of important applications, it is very hard (if not impossible) to simulate random variables following the law π, even in the case of a finite (but very large) set E. A typical case is where the π_x are known up to a multiplicative constant, and where the straightforward computation of the normalization constant

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

is impractical, because it would imply summing a huge number of terms. It might be much simpler to find an irreducible Markovian matrix P which admits π as its invariant probability, and such that simulating a Markov chain having the transition matrix P is easy. This is what Metropolis et al. [27] were the first to propose in 1953. How can one determine an irreducible matrix P which admits π as its invariant probability measure? By determining a matrix P such that the pair (π, P) satisfies the 'detailed balance equation'. Note that knowing π up to a multiplicative constant is sufficient for that purpose. It remains to simulate a Markov chain $\{X_n; n \geq 0\}$ having the transition matrix P, and use the ergodic theorem in order to justify the approximation of

$$\sum_{x \in E} f(x)\pi_x \quad \text{by} \quad \frac{1}{N}\sum_{n=1}^{N} f(X_n).$$

Given a probability π on E such that $\pi_x > 0$, for all $x \in E$, how can we find a Markovian matrix P such that the pair (π, P) satisfies the detailed balance equation? Let R be an arbitrary Markovian transition matrix on E. Then the formula

$$\begin{cases} P_{xy} = R_{xy} \wedge \left(\dfrac{\pi_y}{\pi_x} R_{yx}\right), \quad x \neq y, \\ P_{xx} = 1 - \displaystyle\sum_{y \neq x} P_{xy} \end{cases} \tag{3.1}$$

(where $a \wedge b = \inf(a, b)$) defines a Markovian matrix P such that $\pi_x P_{xy} = \pi_y P_{yx}$, for all $x, y \in E$. The irreducibility of R is not sufficient for ensuring that of P. For P to be irreducible, we need, for any $x \neq y$, that there exist $n \geq 1$ and $\{x_0, \ldots, x_n\} \subset E$ with $x_0 = x$ and $x_n = y$ such that

$$R_{x_{k-1}x_k} \wedge R_{x_k x_{k-1}} > 0, \quad \forall 1 \leq k \leq n.$$

How will we choose R in practice? We first choose a non-oriented graph G on E, such that for all $x, y \in E$, there exist $n \in \mathbb{N}$, $x_1, \ldots, x_{n+1}$ such that $x_1 = x$, $x_{n+1} = y$ and for all $1 \leq k \leq n$, $(x_k, x_{k+1}) \in G$, and we choose R such that

$$R_{xy} > 0 \Leftrightarrow (x, y) \in G.$$

Then the matrix P defined by (3.1) is irreducible.

There are two 'classical' choices for R once the graph G has been chosen. The first choice, which is known as the *Gibbs sampler* , involves choosing

$$R_{xy} = \begin{cases} \left(\sum_{\{z;(x,z)\in G\}} \pi_z\right)^{-1} \pi_y, & \text{if } (x, y) \in G, \\ 0, & \text{if } (x, y) \notin G. \end{cases}$$

The second choice, known as the *Metropolis algorithm*, involves choosing

$$R_{xy} = \begin{cases} (n_x)^{-1}, & \text{if}(x, y) \in G, \\ 0, & \text{if}(x, y) \notin G, \end{cases}$$

where $n_x = |\{z; (x, z) \in G\}|$.

Note that the MCMC procedure was invented for problems where E is very large in size. The idea is to choose G in such a way that n_x will be significantly smaller than the cardinality of E. Then simulating the transitions of a Markov chain with transition matrix R is relatively easy in both the Gibbs and Metropolis cases. It remains to combine that simulation with *Hastings' algorithm*, in order to perform a transition according to the matrix P, from X_n to X_{n+1}. That algorithm makes formula (3.1) practical. Given that $X_n = x$, we first choose Y_n according to the probability law $R_{x\cdot}$, and given that $Y_n = y$, we make a new random simulation (independently of all preceding simulations) which results in choosing

$$X_{n+1} = \begin{cases} Y_n, & \text{with probability } \frac{\pi_y R_{yx}}{\pi_x R_{xy}} \wedge 1, \\ X_n, & \text{with probability } 1 - \frac{\pi_y R_{yx}}{\pi_x R_{xy}} \wedge 1. \end{cases}$$

One way to do this is to draw a random variable U_n, distributed uniformly on $[0, 1]$ and independently of all other random variables, and let

$$X_{n+1} = \mathbf{1}_{\{U_n \leq \pi_y R_{yx}/\pi_x R_{xy}\}} Y_n + \mathbf{1}_{\{U_n \,>\, \pi_y R_{yx}/\pi_x R_{xy}\}} X_n.$$

3.1.1 An application

We now describe the classical application of the MCMC method in statistical physics and in image processing. Another very popular application of the MCMC algorithm is in Bayesian statistics (see Section 8.3 of Chapter 7).

We choose E of the form

$$E = S^{\Lambda},$$

where a point $x \in E$ is an application

$$m \in \Lambda \to x(m) \in S.$$

Λ is the set of 'sites' (set of points (or 'pixels') of the discretized image). Λ and S are finite, hence so is E.

Typically, Λ is very large. On the other hand, S (the set of grey levels or of colours of the image) is a much smaller set. In some applications, $S = \{-1, +1\}$. Even in this simplest case, $\operatorname{card} E = 2^{\operatorname{card}(\Lambda)}$, so clearly E is very large if Λ is large.

Each random variable X_n takes its values in the set S^{Λ}. It is a map from Λ into S. For each $m \in \Lambda$, $X_n(m)$ is an S-valued random variable.

The Markov chain $\{X_n; n \in \mathbb{N}\}$ evolves in such a way that between times n and $n + 1$ only one component of X is modified. In other words, there exists $m \in \Lambda$ such that

$$X_{n+1} \overset{m}{\sim} X_n,$$

in the sense that $X_{n+1}(m') = X_n(m')$, for all $m \neq m'$. This means that the graph G mentioned above is such that $(x, y) \in G$ if and only if x and y differ at one site only. Simulating the Markov chain will involve choosing at each step one site

m where the value of X_n should be modified, and changing the value of $X_n(m)$ randomly.

Let us first describe how we wish to change $X_n(m)$ into $X_{n+1}(m)$. We shall explain afterwards how we wish to 'visit' the various sites, that is, which site m should be chosen at time n. The way to change $X_n(m)$ into $X_{n+1}(m)$ is described by a Markovian matrix $P^{(m)}$, which possesses the property that $P^{(m)}_{xy} = 0$ unless $x \overset{m}{\sim} y$ (i.e. all components of x and y coincide, except perhaps that called m). We want the measure π to be invariant under $P^{(m)}$.

For that purpose, we want to ensure that

$$\pi_x P^{(m)}_{xy} = \pi_y P^{(m)}_{yx}, \quad \forall\, x, y \in E,\ m \in \Lambda.$$

We choose $P^{(m)}$ as follows. Given a Markovian matrix $R^{(m)}$ such that

$$R^{(m)}_{xy} \neq 0 \Leftrightarrow x \overset{m}{\sim} y,$$

we let, for $x \neq y$,

$$\pi_x P^{(m)}_{xy} = (\pi_x R^{(m)}_{xy}) \wedge (\pi_y R^{(m)}_{yx})$$

and

$$P^{(m)}_{xx} = 1 - \sum_{y \neq x} P^{(m)}_{xy} \geq 0.$$

In other words (this is Hastings' algorithm), given that $X_n = x$, we simulate Y_n according to the probability law $R^{(m)}_{x\cdot}$, and given that $Y_n = y$, we make a new random draw (independently of all the others) so as to choose

$$X_{n+1} = \begin{cases} Y_n, & \text{with probability } \dfrac{\pi_y R^{(m)}_{yx}}{\pi_x R^{(m)}_{xy}} \wedge 1, \\ X_n, & \text{with probability } 1 - \dfrac{\pi_y R^{(m)}_{yx}}{\pi_x R^{(m)}_{xy}} \wedge 1. \end{cases}$$

Again the two 'classical' choices for $R^{(m)}$ are the Gibbs sampler, which corresponds to the choice

$$P^{(m)}_{xy} = R^{(m)}_{xy} = \left(\sum_{z \overset{m}{\sim} x} \pi_z \right)^{-1} \pi_y, \quad \text{if } x \overset{m}{\sim} y,$$

and involves choosing $X_{n+1}(m)$ according to its conditional law under π, given the other components (except m); and the Metropolis algorithm, which in its simplest form involves choosing

$$R^{(m)}_{xy} = (|S| - 1)^{-1}, \quad \text{if } x \overset{m}{\sim} y,$$

so that

$$P^{(m)}_{xy} = (|S| - 1)^{-1} \left(\frac{\pi_y}{\pi_x} \wedge 1 \right), \quad x \overset{m}{\sim} y,\ x \neq y.$$

In the Metropolis algorithm we first choose a new value y_m at site m, uniformly in the set $S\backslash\{X_n(m)\}$, and then choose $X_{n+1}(m) = y_m$ if $\pi_y \geq \pi_x$, otherwise with probability π_y/π_x(whenever that quantity is less than 1).

It remains to describe the 'program of visits to the successive sites', that is, to decide at each time n which component m of X_n should be modified. One method involves visiting Λ in a fixed order, and then repeating again and again. Another, which we will adopt because of its simple mathematical formulation, involves choosing the site m at random at each time n, according to the uniform law on Λ, independently of all the rest. In that case, $\{X_n; n \in \mathbb{N}\}$ is a homogeneous Markov chain with transition matrix

$$P = |\Lambda|^{-1} \sum_{m \in \Lambda} P^{(m)}.$$

We then clearly have

$$\pi_x P_{xy} = \pi_y P_{yx},$$

hence, provided P is irreducible, π is the unique invariant probability of the Markov chain $\{X_n; n \in \mathbb{N}\}$. Note that irreducibility is ensured by any of the choices described above for $R^{(m)}$, and a random uniform choice of the site at each step.

3.1.2 The Ising model

This is one of the most popular models of statistical physics. Given $N \in \mathbb{N}$ (N is supposed to be 'large'), let (we have chosen to describe the two-dimensional model)

$$\Lambda = \{-N, \ldots, -1, 0, 1, \ldots, N\}^2 \subset \mathbb{Z}^2,$$

($\Lambda = \Lambda_N$), whose boundary is $\partial\Lambda = \Lambda_N\backslash\Lambda_{N-1}$, and define the *space of configurations* as

$$E = \{-1, 1\}^{\Lambda}.$$

For $x \in E$, we set

$$H(x) = \frac{1}{2} \sum_{\substack{m,m' \in \Lambda \\ |m-m'|=1}} |x(m) - x(m')|^2.$$

Note that H is small whenever x is essentially constant at neighbouring sites. We define

$$E^+ = \{x \in E; x(m) = 1, \ \forall m \in \partial\Lambda\}.$$

For all $\beta > 0$ ($1/\beta$ might be interpreted as a temperature), we define the probability on E^+:

$$\pi(x) = \frac{1}{Z(\beta)} e^{-\beta H(x)}, \quad x \in E^+,$$

with

$$Z(\beta) = \sum_{x \in E^+} e^{-\beta H(x)}.$$

As $\beta \to 0$, π converges towards the uniform measure on E^+, while as $\beta \to +\infty$, π converges towards the uniform measure on the global minima of H, in the present case towards the Dirac measure at the point whose coordinates are all equal to 1.

Physicists have been very interested in realizing simulations under the measure π for large N. But if N is really large, it is impossible to simulate directly under the law π. It is in fact essentially impossible to compute the normalizing constant $Z(\beta)$. We will now describe a version of the MCMC method which exploits the particular form of the Ising model, in such a way that the algorithm is easy to parallelize – that is to say, most of the computation involved can be distributed between independent processors which need not exchange much information.

Let us first describe the Gibbs sampler. Consider a partition of Λ according to whether the sum of the two coordinates of the point m is even or odd:

$$\Lambda^+ = \{(m_1, m_2) \in \Lambda; m_1 + m_2 \text{ even}\},$$
$$\Lambda^- = \{(m_1, m_2) \in \Lambda; m_1 + m_2 \text{ odd}\}.$$

For $x \in E$, denote

$$x^+ = (x(m), m \in \Lambda^+),$$
$$x^- = (x(m), m \in \Lambda^-).$$

It follows from the form of the Ising model that $\pi_{+-}(x^+|x^-)$, the conditional probability of the event $X^+ = x^+$, given that $X^- = x^-$, in the case where the law of X is π, is of the form

$$\pi_{+-}(x^+|x^-) \propto \prod_{m \in \Lambda^+ \setminus \partial\Lambda} e^{\beta x(m) s(m)},$$

(we use the notation $\propto$ if two functions are equal up to a multiplicative normalization constant) where, if $m \in \Lambda^+ \setminus \partial\Lambda$,

$$s(m) = \sum_{m'; |m'-m|=1} x^-(m').$$

We have an analogous formula for $\pi_{-+}(x^-|x^+)$.

Performing simulations under those two laws is easy, thanks to their product form, and the normalizing constant of each factor is explicit. The procedure is as follows. We first choose an arbitrary configuration X_0 in E^+. We next use the following recurrence. Given X_n, we first simulate X^+_{n+1} according to the law $\pi_{+-}(\cdot|X^-_n)$, then X^-_{n+1} according to the law $\pi_{-+}(\cdot|X^+_{n+1})$.

This procedure is precisely the Gibbs sampler of the previous subsection, where we visit alternately all sites of $\Lambda^+ \setminus \partial\Lambda$, then those of $\Lambda^- \setminus \partial\Lambda$ (with a different labelling of the sequence X_n). The convergence follows from the arguments given in the general discussion above. It is not hard to verify that we simulate an irreducible Markov chain whose invariant probability measure is indeed π.

Let us now describe the Metropolis algorithm. Given that $X_n = x$, independently for each $m \in \Lambda^+ \backslash \partial\Lambda$, we change the sign of $x(m)$ with probability

$$p(m, x) = \frac{\pi(x_m)}{\pi(x)} \wedge 1 = e^{-2\beta x(m)s(m)} \wedge 1,$$

with $x_m \overset{m}{\sim} x$, $x_m(m) = -x(m)$, and

$$s(m) = \sum_{|m'-m|=1} x^-(m').$$

We have thus obtained $X^+_{n+1}(m)$. We then simulate $X^-_{n+1}(m)$ by conditioning upon the value of $X^+_{n+1}(m)$. The process $\{X_n; n \in \mathbb{N}\}$ which we obtain in this way is an irreducible Markov chain whose invariant probability is π.

3.1.3 Bayesian analysis of images

One can use the Ising model (or other similar models) as the a priori law of a two-dimensional digitized picture. Each point $m \in \Lambda$ is called a 'pixel'. $x(m)$ is the grey level or colour of the pixel m (in the case of the Ising model there are just two colours, black and white). By changing the parameter β of the Ising model, we change the 'texture' of the image: for large values of β, the image will consist of large white patches and large black patches, while smaller values of β will produce images where smaller white and black patches coexist.

We observe the colour (white or black) of each pixel, and the observation gives the exact colour of each pixel with probability $p \in]0, 1[$, the possible errors being independent among the pixels.

Then the a posteriori law, or more precisely the conditional probability of the event $X = x$, given that we have observed the configuration y , equals

$$\pi(x|y) \propto e^{-\beta H(x)} p^{a(x,y)} (1-p)^{d(x,y)},$$

where $a(x, y)$ is the number of sites where the configurations x and y coincide and $d(x, y)$ the number of sites where they differ.

We obtain an image where the observation errors have been suppressed by simulating according to the probability $\pi(x|y)$. Although we are no longer dealing strictly with the Ising model, the same methods apply. Let us describe the Metropolis algorithm. Given that $X_n = x$, independently for each site $m \in \Lambda^+ \backslash \partial\Lambda$, we change the sign of $X^+_n(m)$ with probability

$$\begin{aligned} p(m, x, y) &= \frac{\pi(x_m|y)}{\pi(x|y)} \wedge 1 \\ &= e^{-2\beta x(m)s(m)} \left(\frac{1-p}{p}\right)^{x(m)y(m)} \wedge 1, \end{aligned}$$

thus producing X^+_{n+1}. We then simulate X^-_{n+1} using the values thus obtained on $\Lambda^+ \backslash \partial\Lambda$. In this manner we construct an irreducible Markov chain with the invariant probability $\pi(\cdot)$.

3.1.4 Heated chains

The convergence of the MCMC algorithm requires the simulated chain to visit each state sufficiently often. But the shape of the transition probability distribution may be such that the chain tends to get stuck for long periods of time in certain regions of the state space. Suppose, for example, that $E = \mathbb{Z}$ (or an interval of $\mathbb{Z}$, or $\mathbb{Z}/N$), and that $R_{x,x+1} = R_{x,x-1} = 1/2$. Let

$$H(x) = -\log(\pi_x).$$

We might choose for the transition matrix P of the Markov chain the matrix

$$P_{xy} = \begin{cases} \frac{\exp[H(x)-H(y)]}{2}, & \text{if } y = x \pm 1, \\ 1 - \frac{\exp[H(x)-H(x+1)]}{2} - \frac{\exp[H(x)-H(x-1)]}{2}, & \text{if } y = x, \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that there are two regions A and B of $\mathbb{Z}$ where π takes relatively large values, which are separated by an interval where π is extremely small (or equivalently, where H takes huge values). In that case, the chain will very rarely move from A to B (or from B to A). One solution is to introduce a family of transition matrices indexed by β:

$$P_{\beta,xy} = \frac{1}{2}\exp[\beta(H(x) - H(y))], \quad \text{if } y = x \pm 1,$$

with $0 < \beta < 1$. The corresponding chains are called 'heated' (β can be thought of as the inverse of a temperature). One simulates in parallel the chain $\{X_n\}$ corresponding to $\beta = 1$, and several chains $\{X_n^1\}, \ldots, \{X_n^k\}$, corresponding to values $1 > \beta_1 > \beta_2 > \ldots > \beta_k > 0$. Clearly, the smaller β is, the less the chain tends to get stuck in certain regions of the state space. The trick is then to permute randomly from time to time the values of $(X_n, X_n^1, \ldots, X_n^k)$, in such a way that X_n visits the state space E more rapidly. Of course, in the final computation, only the values of $\{X_n; n \geq 0\}$ are retained.

3.2 Simulation of the invariant probability

One problem in MCMC algorithms is the choice of the number of times one should iterate the Markov chain. The difference from a standard Monte Carlo computation is that we start the chain from an arbitrary point, that is, the chain is not started from the invariant probability. In a sense one may think that there is an 'initial phase' of the algorithm, during which the law of the chain gets close to the invariant probability. Then, during the second phase of the algorithm, we should control the rate of convergence in the ergodic theorem, which – as in a standard Monte Carlo method – can be done with the help of the corresponding central limit theorem (see Theorem 6.11 of Chapter 2).

We shall discuss the rate of convergence towards the invariant measure in the next section. Let us first discuss ideas due to Propp and Wilson [31], which permit a 'perfect' (in the sense of 'exact', as opposed to 'approximate') simulation under the invariant probability measure. The idea is that it can be reached in a finite (but random) number of steps.

We assume in this section that $\mathrm{card}(E) < \infty$ and, to fix notation, specify that $E = \{1, 2, \ldots, N\}$.

3.2.1 Perfect simulation

We assume here that

$$\beta(P) = \sum_{y \in E} \inf_{x \in E} P_{xy} > 0,$$

or in other words that there exists $y \in E$ such that $P_{xy} > 0$, for all $x \in E$. This condition is nothing but condition (D) from Section 2.6, but with the restriction that $n_0 = 1$. Clearly $\beta(P) \leq 1$. We set

$$\nu_y = \frac{\inf_x P_{xy}}{\beta(P)}, \quad y \in E,$$

hence ν is a probability on E.

Remark 2.1 *One could choose another pair (β, ν), with $\beta > 0$ and ν a probability on E such that $P_{xy} \geq \beta \nu_y$, but the above choice is optimal in the sense that it maximizes β.*

Remark 2.2 *The assumption $\beta(P) > 0$ entails the existence of a unique recurrent class (see Exercise 10.4 of Chapter 2), hence P possesses a unique invariant probability, which we denote by π.*

We choose a function

$$F : E \times [0, 1] \to E$$

such that if U is a $\mathcal{U}([0, 1])$ random variable,

$$\mathbb{P}(F(x, U) = y) = P_{xy}, \quad x, y \in E.$$

Hence if $\{U_n;\ n \in \mathbb{N}\}$ is a sequence of i.i.d. $\mathcal{U}([0, 1])$ random variables, which are independent of X_0,

$$X_n = F(X_{n-1}, U_n), \quad n \geq 1,$$

defines a Markov chain with transition matrix P.

We define $\ell : \{0\} \cup E \to [0, \beta(P)]$ by

$$\ell(0) = 0,$$

$$\ell(y) = \ell(y - 1) + \inf_x P_{xy}, \quad 1 \leq y \leq N,$$

and we let $J(y) = [\ell(y-1), \ell(y))$. We further define $k : E \times (\{0\} \cup E) \to [\beta(P), 1]$ by

$$k(x, 0) = \beta(P),$$
$$k(x, y) = k(x, y-1) + P_{xy} - \inf_z P_{zy}, \quad 1 \le y \le N,$$

and $K(x, y) = [k(x, y-1), k(x, y))$, $1 \le x, y \le N$. We have

$$I(x, y) = J(y) \cup K(x, y).$$

We note that $|I(x, y)| = P_{xy}$. Finally, let

$$F(x, u) = \sum_{y \in E} y \mathbf{1}_{\{u \in I(x,y)\}}, \quad 1 \le x \le N, \ \ u \in [0, 1].$$

Note that $\mathbb{P}(F(x, U) = y) = P_{xy}$, so that

$$\mathbb{P}(F(X_{n-1}, U_n) = y | X_{n-1} = x) = P_{xy}.$$

The crucial point about this construction is that if at time n, $U_n < \beta(P)$, then the value of X_n does not depend on X_{n-1}. In other words, if we run this algorithm in parallel with the same sequence $\{U_n\}$ and different starting points X_0, then the various sequences coalesce at the first time n when $U_n < \beta(P)$. We have the following proposition.

Proposition 2.3 *Let $T = \inf\{n \ge 1; \ U_n < \beta(P)\}$. Then T and X_T are independent, T follows the geometric law with parameter $\beta(P)$, and the law of X_T is ν.*

PROOF

$$\{X_T = x, T = n\} = \{U_1 \ge \beta(P), \ldots, U_{n-1} \ge \beta(P), U_n \in J(x)\}.$$

Hence

$$\mathbb{P}(X_T = x, T = n) = (1 - \beta(P))^n \beta(P) \nu_x.$$

□

Let us now construct a stationary chain with the transition matrix P. Let $\{U_n; \ n \in \mathbb{Z}\}$ be an i.i.d. sequence of $\mathcal{U}([0, 1])$ random variables, and let

$$N_k = \mathbf{1}_{\{U_k < \beta(P)\}}, \ \ k \in \mathbb{Z}.$$

The $\{N_k\}$ are mutually independent Bernoulli random variables.

For all $n \in \mathbb{Z}$, we let

$$\tau(n) = \max\{k \le n; U_k < \beta(P)\}.$$

Note that $\tau(k) = \tau(n)$, for all $k \in [\tau(n), n]$. Moreover, $\mathbb{P}(n - \tau(n) > k) = (1-\beta(P))^k$, or equivalently $\mathbb{P}(n - \tau(n) = k) = (1-\beta(P))^{k-1}\beta(P)$.

We then define the process $\{X_n;\ n \in \mathbb{Z}\}$ as follows. For all $k \in \mathbb{Z}$ such that $N_k = 1$, we let

$$X_k = \sum_{y \in E} y \mathbf{1}_{\{U_k \in J(y)\}}.$$

Now let k be such that $N_k = 0$. $X_{\tau(k)}$ is defined by the above formula. Moreover

$$X_{\tau(k)+1} = F(X_{\tau(k)}, U_{\tau(k)+1}), \ldots, X_k = F(X_{k-1}, U_k).$$

Proposition 2.4 *The process $\{X_n; n \in \mathbb{Z}\}$ thus defined is stationary (this means that, for all $\ell \in \mathbb{Z}$, $k \in \mathbb{N}$, $(X_{\ell+1}, \ldots, X_{\ell+k}) \simeq (X_1, \ldots, X_k)$, in the sense that these two random vectors have the same law). In particular, the law of X_0 is the unique invariant probability π of the Markov chain with transition matrix P.*

Proof It suffices to establish the last point. Define $\hat{P}$ by

$$P_{xy} = (1 - \beta)\hat{P}_{xy} + \beta \nu_y,$$

where $\beta = \beta(P)$. Note that

$$\hat{P}_{xy} = \mathbb{P}(X_n = y | X_{n-1} = x, U_n \geq \beta).$$

Consequently,

$$\begin{aligned}
\mathbb{P}(X_0 = x) &= \sum_{k=0}^{\infty} \mathbb{P}(X_0 = x, \tau(0) = -k) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_0 = x | \tau(0) = -k)(1 - \beta)^k \beta \\
&= \beta \sum_{k=0}^{\infty} (\nu \hat{P}^k)_x (1 - \beta)^k.
\end{aligned}$$

But if we let $\mu_x = \mathbb{P}(X_0 = x)$, we have

$$\begin{aligned}
\mu P &= \beta \sum_{x \in E} \sum_{k=0}^{\infty} (\nu \hat{P}^k)_x (1 - \beta)^k P_{x\cdot} \\
&= \beta \sum_{k=0}^{\infty} (\nu \hat{P}^{k+1})_x (1 - \beta)^{k+1} + \beta \nu \\
&= \mu,
\end{aligned}$$

hence the law of X_0 is the unique probability which is P-invariant. □

Perfect simulation algorithm. The algorithm works as follows:

1. Simulate $U_0,\ U_{-1}, \ldots, U_{\tau(0)}$ (this requires a number of simulations which follows a geometric law).
2. Compute $X_{\tau(0)} = \sum_y y \mathbf{1}_{\{U_{\tau(0)} \in J(y)\}}$.
3. Compute $X_{\tau(0)+1}, \ldots, X_0$ using the formula $X_n = F(X_{n-1}, U_n)$, and the $U_{\tau(0)+1}, \ldots, U_0$ simulated above.
4. The law of the random variable X_0 thus simulated is the probability which is invariant under P.

Remark 2.5 *This approach requires that $\beta(P) > 0$. One might wish to generalize the approach to the case where there exists $k \geq 1$ such that $\beta(P^k) > 0$, but it does not seem that this leads to an effective algorithm.*

3.2.2 Coupling from the past

We now assume only that P is irreducible. Suppose we are given a mapping

$$F : E \times [0, 1] \to E$$

such that if U has the law $\mathcal{U}([0, 1])$, then

$$\mathbb{P}(F(x, U) = y) = P_{xy}, \quad \forall x, y \in E.$$

Let us now define a 'multiple coupling'. Let $\{U_n^i; 1 \leq i \leq N,\ n \in \mathbb{Z}\}$ be a sequence of i.i.d. random variables, all having the law $\mathcal{U}([0, 1])$, and, for $k \in \mathbb{Z}$, let

$$X_n^{1,k} = \begin{cases} 1, & \text{if } n = k, \\ \mathcal{F}(X_{n-1}^{1,k}, U_n^1), & \text{if } n > k; \end{cases}$$

$$X_n^{2,k} = \begin{cases} 2, & \text{if } n = k, \\ \mathcal{F}(X_{n-1}^{2,k}, U_n^2), & \text{if } n > k \text{ and } X_{n-1}^{2,k} \neq X_{n-1}^{1,k}, \\ \mathcal{F}(X_{n-1}^{2,k}, U_n^1), & \text{if} n > k \text{ and } X_{n-1}^{2,k} = X_{n-1}^{1,k}; \end{cases}$$

$$\ldots$$

$$X_n^{N,k} = \begin{cases} N, & \text{if } n = k, \\ F(X_{n-1}^{N,k}, U_n^N), & \text{if } n > k \text{ and } X_{n-1}^{N,k} \notin \{X_{n-1}^{1,k}, \ldots, X_{n-1}^{N-1,k}\}, \\ F(X_{n-1}^{N,k}, U_n^{i_N}), & \text{if } n > k,\ X_{n-1}^{N,k} \in \{X_{n-1}^{1,k}, \ldots, X_{n-1}^{N-1,k}\} \\ & \text{and } i_N = \inf\{i;\ X_{n-1}^{i,k} = X_{n-1}^{N,k}\}. \end{cases}$$

Define the stopping time

$$S_k = \inf\{\ell > k;\ X_\ell^{1,k} = X_\ell^{2,k} = \ldots = X_\ell^{N,k}\}$$

and, for all $n \in \mathbb{Z}$, let

$$\tau(n) = \sup\{k \leq n;\ \ S_k \leq n\},$$

the largest of those times which are such that the state of the process at time n does not depend upon its values before time $\tau(n)$ (and this for all k between 1 and N), since at time $\tau(n)$ the kth process had the value k, and at time n its value does not depend upon its index.

Theorem 2.6 *For $k \in \mathbb{Z}$, let $(X_n^{1,k}, \ldots, X_n^{N,k};\ \ n \geq k)$ denote the multiply coupled process defined above. Suppose that $\mathbb{P}(\tau(0) > -\infty) = 1$. Then, for all $x \in E$, the law of $X_0^{x,\tau(0)}$ is the P-invariant probability.*

Proof For $k \in \mathbb{Z}$, $y \in E$,

$$\mathbb{P}(\{X_0^{x,\tau(0)} = y\} \cap \{\tau(0) > k\}) = \mathbb{P}(\{X_0^{x,k} = y\} \cap \{\tau(0) > k\}).$$

Hence

$$\mathbb{P}(X_0^{x,\tau(0)} = y) = \lim_{k \to -\infty} \mathbb{P}(X_0^{x,k} = y).$$

Let π denote the P-invariant probability. Then

$$\begin{aligned}
|\mathbb{P}(X_0^{x,k} = y) - \pi_y| &= |\mathbb{P}(X_0 = y | X_k = x) - \sum_{z \in E} \pi_z \mathbb{P}(X_0 = y | X_k = z)| \\
&\leq \sum_{z \in E} \pi_z |\mathbb{P}(X_0 = y | X_k = x) - \mathbb{P}(X_0 = y | X_k = z)| \\
&\leq \left(\sum_{z \in E} \pi_z \right) \mathbb{P}(\tau(0) < k) \\
&\to 0, \quad \text{as } k \to -\infty.
\end{aligned}$$

□

Coupling from the past: the algorithm. Choose some $k \in \mathbb{Z}_-$.

1. Simulate $U_k^i, U_{k+1}^i, \ldots,\ U_0^i,\ \ 1 \leq i \leq N$.
2. Construct the sequence

   $$(X_\ell^{1,k}, \ldots, X_\ell^{N,k},\ \ \ell = k, k+1, \ldots, 0)$$

   with the above Us and the above algorithm.
3. Check whether or not

   $$X_0^{1,k} = X_0^{2,k} = \ldots = X_0^{N,k}.$$

   If this is the case, then $X_0^{1,k}$ is a realization of the invariant probability μ. If not, we need to go back further in the past. In that case, we simulate

   $$U_{2k}^i, U_{2k+1}^i, \ldots, U_{k-1}^i, \quad 1 \leq i \leq N.$$

With the help of that sequence and the previous one, we construct

$$(X_\ell^{1,2k}, \ldots, X_\ell^{N,2k};\ \ell = 2k, \ldots, 0)$$

and proceed as above. Note that it is essential to combine the two sequences of Us as indicated above, since we must go back into the past until we reach $\tau(0)$.

It remains to give conditions under which $\mathbb{P}(\tau(0) > -\infty) = 1$.

Theorem 2.7 *If P is aperiodic, then the above algorithm satisfies $\mathbb{P}(\tau(0) > -\infty) = 1$.*

PROOF This result can be proved by following step 1 of the proof of Theorem 6.4. of Chapter 2 □

Remark 2.8 *In the case $\beta(P) > 0$, this is a variant of the perfect simulation algorithm of Section 3.1.1 applying the procedure of the present section, simplified as follows. We simulate a unique sequence $(U_n, n \in \mathbb{Z})$, and we use the recurrence relation*

$$(X_n^{1,k}, \ldots, X_n^{N,k}) = \begin{cases} (1, \ldots, N), & \text{if } n = k, \\ (F(X_{n-1}^{1,k}, U_n), \ldots, F(X_{n-1}^{N,k}, U_n), & \text{if } n > k. \end{cases}$$

3.3 Rate of convergence towards the invariant probability

The 'coupling from the past' algorithm cannot be used in most concrete situations. Most users of the MCMC algorithm reject the first n iterations (this is the so-called 'burn-in' procedure), hoping that the law of X_{n+1} is 'close' to the invariant probability. The problem is clearly to choose the number n correctly. The statements in Section 2.8 are useful, provided one can compute the spectral gap β.

Several authors, in particular Diaconis [14] (see also Saloff-Coste [36]), have shown that in many situations the total variation distance between the law of X_n and the invariant probability evolves as follows. During an initial phase, the decrease is very slow, and the distance stays close to 1. Then there is a fast decrease in a few iterations, followed by a third phase, with slow convergence towards zero. In such a situation, after the few first iterations, the result remains very poor, then it improves quickly, and after the phase of rapid decrease it is not really worth continuing the iterations. The problem then is to know precisely how many iterations are necessary in order to reach the end of the rapid decrease phase. Diaconis and Saloff-Coste have given very precise results in certain specific situations, such as the problem of card shuffling, but of course in most practical situations we do not have such precise

information. Let us describe a result due to Diaconis, Khare and Saloff-Coste [15], which provides an opportunity to give an interesting example of the use of the MCMC method, namely simulating the 'Glauber dynamics'.

Suppose that we wish to simulate realizations of an $E = E_1 \times \cdots \times E_d$-valued random vector, with the law

$$\mu(dx_1, \ldots, dx_d) = f(x_1, \ldots, x_d)\nu(dx_1, \ldots, dx_d),$$

where the measure ν on E is a product measure,

$$\nu(dx_1, \ldots, dx_d) = \nu_1(dx_1) \times \cdots \times \nu_d(dx_d).$$

Suppose that there is no easy direct way to perform simulations under μ, but that, for all $x = (x_1, \ldots, x_n)$ and each $1 \leq i \leq d$, we know how to simulate X_i with the law

$$\frac{f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d)\nu_i(dy)}{\int_{E_i} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_d)\nu_i(dz)}.$$

We will construct a Markov chain (if E is uncountable, such an E-valued Markov chain is outside the scope of this book, but we can restrict ourselves to the case of a Markov chain with values in a finite or countable state space by discretization) whose transition is given by the transformation of the random vector X into the random vector X', which we now specify by describing the conditional law of X' given that $X = x$.

The conditional law of X'_1, given that $X = x$, is the law

$$\frac{f(y, x_2, \ldots, x_d)\nu_1(dy)}{\int_{E_1} f(z, x_2, \ldots, x_d)\nu_1(dz)}$$

and, for $2 \leq i \leq d$, the conditional law of X'_i, given that $X = x$ and $(X'_1, \ldots, X'_{i-1}) = (x'_1, \ldots, x'_{i-1})$, is the law

$$\frac{f(x'_1, \ldots, x'_{i-1}, y, x_{i+1}, \ldots, x_d)\nu_i(dy)}{\int_{E_i} f(x'_1, \ldots, x'_{i-1}, z, x_{i+1}, \ldots, x_d)\nu_i(dz)}.$$

In the following proposition we assume that E is countable, and we write $\mu(x)$ (resp. $\nu(x)$) for μ_x (resp. ν_x).

Proposition 3.1 *If E is countable and $\mu(x) > 0$ for all $x \in E$, then the Markov chain with transition matrix*

$$P_{xy} = \mathbb{P}(X' = y | X = x)$$

is irreducible and positive recurrent, and has μ as its unique invariant probability.

PROOF The condition $\mu(x) > 0$, for all $x \in E$, implies that the chain is irreducible. It remains to show that the probability μ is invariant by P. We have

$$\begin{aligned}
\sum_{x\in E} \mu(x) P_{xy} = & \sum_{x_1,\ldots,x_d} \frac{f(y_1, x_2, \ldots, x_d)\nu_1(y_1)}{\sum_{z_1} f(z_1, x_2, \ldots, x_d)\nu_1(z_1)} \\
& \times \frac{f(y_1, y_2, x_3 \ldots, x_d)\nu_2(y_2)}{\sum_{z_2} f(y_1, z_2, x_3, \ldots, x_d)\nu_2(z_2)} \\
& \times \cdots \\
& \times \frac{f(y_1, \ldots, y_{d-1}, x_d)\nu_{d-1}(y_{d-1})}{\sum_{z_{d-1}} f(y_1, \ldots, z_{d-1}, x_d)\nu_{d-1}(z_{d-1})} \\
& \times \frac{f(y_1, \ldots, y_{d-1}, y_d)\nu_d(y_d)}{\sum_{z_d} f(y_1, \ldots, y_{d-1}, z_d)\nu_d(z_d)} \\
& \times f(x_1, \ldots, x_d)\nu_1(x_1) \times \cdots \times \nu_d(x_d) \\
= & \sum_{x_1,\ldots,x_d} \frac{f(x_1, x_2, \ldots, x_d)\nu_1(x_1)}{\sum_{z_1} f(z_1, x_2, \ldots, x_d)\nu_1(z_1)} \\
& \times \frac{f(y_1, x_2, x_3, \ldots, x_d)\nu_2(x_2)}{\sum_{z_2} f(y_1, z_2, x_3, \ldots, x_d)\nu_2(z_2)} \\
& \times \cdots \\
& \times \frac{f(y_1, \ldots, y_{d-1}, x_d)\nu_d(x_d)}{\sum_{z_d} f(y_1, \ldots, y_{d-1}, z_d)\nu_d(z_d)} \\
& \times f(y_1, \ldots, y_{d-1}, y_d)\nu_1(y_1) \times \cdots \times \nu_d(y_d) \\
= & \mu(y),
\end{aligned}$$

where the second equality is obtained by rearranging terms. The third equality follows from the fact that only the first factor depends on x_1 and its sum over $x_1 \in E_1$ is 1, and the fact that in the remaining expression only the first factor depends upon x_2, and its sum over $x_2 \in E_2$ is 1, etc. □

Diaconis, Khare and Saloff-Coste [15] study in particular the following case where $d = 2$:

- $E_1 = \{0, 1, \ldots, m\}$, $E_2 = [0, 1]$, ν_1 (ν_2) is the uniform probability on E_1 (E_2),
- $f(k, p) = C_m^k p^k (1-p)^{m-k}$,

where, $m \in \mathbb{N}^*$ is arbitrary. Their result for this example is that the number of iterations one needs to perform for the law of the chain to be close to its invariant probability is of the order of m. Specifically, they show that if f_n denotes the

density with respect to the measure $\nu = \nu_1 \times \nu_2$ of the law of the nth iterate in the above procedure, starting from the point (m, p), then

$$\left(1 - \frac{2}{m+2}\right)^n \leq \|f_n - f\|_{L^1(\nu)} \leq 6\left(1 - \frac{2}{m+2}\right)^{n-1/2}.$$

3.4 Simulated annealing

The search for the global maxima of a function is one of the most important problems in applied mathematics. In the case of a differentiable function defined on $\mathbb{R}^d$, one can start at an arbitrary point and move in the direction of the gradient as far as the function decreases. Unfortunately such a method leads to a local minimum, which need not be a global one. In the case of a function defined on a finite set E, one could in principle compute all the values $f(x)$ for all x in E, but in the interesting cases such a procedure is impossible to implement because of the size of the set E.

In this section we present the 'simulated annealing' algorithm which, compared with the classical gradient method, introduces random perturbations which allow us to escape from the basins of attraction of the local minima. While the computations progress, the random perturbations are reduced in such a way that one may hope that the algorithm reaches one of the global minima. The name for this algorithm comes from the analogy with chemical processes for manufacturing certain crystals, which reach the desired state at the end of a process which involves slow cooling, possibly interrupted by phases where the crystal is heated up.

We now present the simulated annealing algorithm for the minimization of a function defined on a finite set E. Let us first present two particular minimization problems of a function on a finite but very large set E.

Example 4.1 (The travelling salesman problem) *Let* $\{1, \ldots, N\}$ *be a set of* N *towns. The salesman must visit each of the towns, starting and finishing at* 1. E *is the set of all the possible routes, and* $\mathrm{card}(E) = (N-1)!$. *Each route is an ordered sequence*

$$x = (x_1, \ldots, x_N)$$

such that $x_1 = 1$ *and* $(x_2, \ldots, x_N)$ *constitutes a permutation of the set* $\{2, \ldots, N\}$. *The cost function to be minimized is*

$$V(x) = \sum_{k=1}^{N} d(x_k, x_{k+1}),$$

where $d(n, m)$ *denotes the distance from town* n *to town* m *and* $x_{N+1} = 1$. *The search for a global minimum of* V *is one of the classical problems in operational research.*

Example 4.2 (Image restoration) *Let us return to a model presented in Section 3.1.3. Here we wish, in order to restore the image, to find the maximum of the a posteriori law – that is, in the notation of Section 3.1.3, for a given y, we look for*

$$\hat{x} = \arg\max_x e^{-\beta H(x)} p^{a(x,y)} (1-p)^{d(x,y)}.$$

Suppose that we wish to maximize a function

$$U : E \to \mathbb{R}_-$$

such that, in order to fix ideas,

$$\max_{x \in E} U_x = 0.$$

We seek an x such that $U_x = 0$.

For any $\beta > 0$, we define the probability π_β on E by

$$\pi_{\beta,x} = Z_\beta^{-1} e^{\beta U_x}, \quad x \in E,$$

with $Z_\beta = \sum_{x \in E} e^{\beta U_x}$. The parameter β will eventually go to $+\infty$. As $\beta \to +\infty$, the probability π_β converges towards the uniform probability on the maxima of U.

We associate with each $\beta > 0$ the transition matrix of an irreducible and aperiodic Markov chain, with invariant probability π_β. The transition matrix can be chosen as follows. Let G denote a non-oriented graph in E, that is, a collection of pairs of points in E. We suppose that G has the property that, for all $x, y \in E$, there exist n and $x = x_1, x_2, \ldots, x_n = y \in E$ such that $(x_k, x_{k+1}) \in G$, $1 \leq k \leq n-1$. Let

$$n_x = |\{y; (x, y) \in G\}|.$$

Then the matrix P_β whose off-diagonal elements are given by

$$P_{\beta,xy} = \mathbf{1}_{\{(x,y) \in G\}} n_x^{-1} \left[e^{\beta(U_y - U_x)} \wedge 1 \right],$$

with appropriate diagonal elements, has the required properties. Note that the transitions which decrease U become less probable as β gets larger. Provided that the choice of the graph G does not make the chain periodic, for a fixed value of β, if $\{X_n^\beta;\ n \geq 0\}$ denotes a Markov chain with transition matrix P_β, then the law of X_n^β converges towards π_β as $n \to \infty$. The idea of the annealing algorithm is to let β depend upon n, in such a way that $\beta_n \to +\infty$ as $n \to \infty$, in the hope that X_n converges towards the maximum (or the set of maxima) of the function U. This is true if β_n converges sufficiently slowly to $+\infty$. We shall present in Section 7.10 below a result in that direction, for an analogue in continuous time of the chain $\{X_n^{\beta_n};\ n \geq 0\}$.

3.5 Exercises

Exercise 5.1 *Let p and q be two probabilities on a countable set E with $0 < p \leq cq$, q being easily simulable. Let $\{Y_n; n \geq 1\}$ be i.i.d. random variables with the common law q, jointly independent of the random variable X_0. Define*

$$X_{n+1} = \begin{cases} Y_{n+1}, & \text{with probability } \frac{p(Y_{n+1})}{cq(Y_{n+1})}, \\ X_n, & \text{with probability } 1 - \frac{p(Y_{n+1})}{cq(Y_{n+1})}. \end{cases}$$

1. *Write X_{n+1} explicitly in the form $f(X_n, U_{n+1}, Y_{n+1})$, where the U_n are i.i.d. $\mathcal{U}[0,1]$ random variables, and deduce that $\{X_n; n \geq 0\}$ is a Markov chain.*
2. *Compute the transition probability P_{xy} of X_n.*
3. *Compute μP for a probability μ, as well as $(\mu - p)P^n$. Conclude that the law of X_n converges as $n \to \infty$ towards the unique invariant probability p.*
4. *How can you compare this chain and the classical rejection method?*

Exercise 5.2 *Let P_{xy} be the transition matrix of an E-valued Markov chain (where E is at most countable). We assume that*

$$P_{xy} \geq \alpha c_y, \quad \forall x \in E, \tag{3.2}$$

where c is a probability and $\alpha > 0$. We identify the set of all bounded signed measures on E with $\ell^1(E)$, equipped with the norm $|\nu| = \sum_{x \in E} |\nu_x|$.

1. *Let ν be a bounded signed measure with total mass equal to zero. Show that $|\nu P| \leq (1-\alpha)|\nu|$. Deduce that if μ are μ' two probability measures on E, then*
$$\left|\mu P - \mu' P\right| \leq (1-\alpha)\left|\mu - \mu'\right|.$$
2. *Show that if an invariant probability measure exists, then it is necessarily unique, and for any probability μ the sequence $(\mu P^n, n \geq 0)$ is a Cauchy sequence in $\ell^1(E)$.*
3. *Let $(X_n, n \geq 0)$ be a Markov chain with transition matrix P. Show that, whatever the initial law μ of X_0, the law of X_n converges towards the unique invariant probability π, and that*
$$|\mu P^n - \pi| \leq C\rho^n,$$
where C is a finite constant and $0 < \rho < 1$.
4. *Show that the above results still hold if there exists $\ell \geq 1$ such that*
$$P^{\ell}_{xy} \geq \alpha c_y, \quad \forall x, y \in E.$$

5. *We now consider the Metropolis algorithm on a finite set E. We assume that $P_{xy} = P_{yx}$ and that relation (3.2) is satisfied. We try to simulate a law μ given up to a multiplicative constant by*

$$\mu_x = Ce^{-\beta H(x)}, \quad \text{with } H(x) \geq x, \forall x \in E.$$

Describe the transition matrix of the chain which combines the transitions of P and Hastings' algorithm.

6. *Show that this new transition matrix satisfies a relation of type (3.2), and state it precisely. Propose an approximate simulation method under the law μ.*

Exercise 5.3 *We wish to solve in $\mathbb{R}^d$ the equation*

$$(I - A)x = b, \tag{3.3}$$

where A is a matrix whose norm is strictly less than 1. To do so, we consider a Markov chain $\{X_n\}$ on $E = \{1, 2, \ldots, d\}$, with the strictly positive initial law μ, and the transition matrix P, which is assumed to be strictly positive on $E \times E$.

1. *Show that the solution x can be written in the form*

$$x = \sum_{n=0}^{\infty} A^n b.$$

2. *For $n \geq 1$ and $b \in \mathbb{R}^d$, define*

$$W_n = b(X_0)\frac{A(X_0, X_1)\cdots A(X_{n-1}, X_n)}{\mu(X_0)P(X_0, X_1)\cdots P(X_{n-1}, X_n)}b(X_n).$$

Compute $\mathbb{E}(W_n)$. Deduce a method for the computation of an approximate solution of equation (3.3).

3. *We let $\widetilde{E} = E \cup \{\delta\}$ and now consider a chain with a cemetery δ. By this we mean a chain $\widetilde{X}_n$ with the initial law $\widetilde{\mu}$ supported by E, strictly positive on E, with transition matrix $\widetilde{P}$ defined by*

$$\widetilde{P}(x, y) = \begin{cases} (1-p)P(x, y), & \text{if } x, y \in E, \\ p, & \text{if } x \in E, y = \delta, \\ 1, & \text{if } x, y = \delta, \\ 0, & \text{if } x = \delta, y \in E, \end{cases}$$

where $0 < p < 1$. Define $T = \inf\{n \geq 1;\ \widetilde{X}_n = \delta\}$.

(a) Show that T is finite almost surely. Let

$$W = b(\widetilde{X}_0)\frac{A(\widetilde{X}_0, \widetilde{X}_1)\cdots A(\widetilde{X}_{T-2}, \widetilde{X}_{T-1})}{\widetilde{\mu}(\widetilde{X}_0)\widetilde{P}(\widetilde{X}_0, \widetilde{X}_1)\cdots \widetilde{P}(\widetilde{X}_{T-2}, \widetilde{X}_{T-1})\widetilde{P}(\widetilde{X}_{T-1}, \widetilde{X}_T)}b(\widetilde{X}_{T-1}).$$

4. *Compute $\mathbb{E}(W)$. Deduce a method for the computation of an approximate solution of equation (3.3).*

4

Markov chains and the genome

Introduction

This chapter is devoted to the presentation of algorithms for the annotation of the genome (and especially the search for coding regions as well as genes) and for sequence alignment which make use of Markov chains. It discusses the so-called *hidden Markov model* and associated algorithms. This concept is used in many applications outside bioinformatics, for example in speech recognition, see Jelinek [21]. For further reading on the subject of this chapter, one can consult among other sources [17], [19] and [28].

4.1 Reading DNA

Consider a fragment of DNA, a single string composed of a succession of nucleotides, which we will view as letters in the alphabet {A, C, G, T}, where A stands for *adenine*, C for *cytosine*, G for *guanine*, and T for *thymine*, for example,

A C C G T A A T T C G G A ... T T G C

To 'read' or 'annotate' such a sequence mainly involves decomposing it into *forward coding regions*, *backward coding regions* (the DNA macro-molecule is made of two complementary strings, with an A always coupled with a T, a C with a G, which are read in opposite directions), and *non-coding regions*; in the case of a eukaryotic genome, within the *coding regions* one also needs to separate the *introns* from the *exons*. Note that the coding region is read codon by codon, where a codon is made of three nucleotides, each codon then being translated into an

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

amino acid. The succession of amino acids constitutes a *protein*. It is therefore essential to read each coding region with the right *phase*. Skipping one codon may not necessarily be too serious, but shifting the reading frame by one or two nucleotides is catastrophic!

The presence of a *start* (*stop*) codon at the start (end) of each coding region is a great help. But a potential start need not be a start codon. In contrast, a potential stop codon, in a coding region, encountered in the correct phase, is a stop. There are no very precise signals for the transition between an *intron* and an *exon*.

In order to distinguish between coding and non-coding regions, a first possibility is that the proportions of As, Cs, Gs and Ts differ significantly between coding and non-coding regions. A second possibility is that these proportions do not differ significantly, and that one should count the *di-* or the *trinucleotides*.

In the first case, we should try to distinguish between coding and non-coding regions by comparing the proportions of As, Cs, Gs and Ts. In the second case, one should count pairs or triplets. And whatever the criteria might be, the hardest part of the job is to localize precisely the boundaries of the different regions.

The methods which we have just discussed for decomposing a DNA sequence into different regions – the goal being to detect the genes – can be seen as statistical procedures associated with a given probabilistic model. This model is not the same if one counts *nucleotides*, *di-* or *trinucleotides*. At any rate, in the sequel we shall consider the sequence of nucleotides as a sequence of random variables with values in the set $E = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

Before we discuss the various possible probabilistic models for a DNA sequence, let us discuss two of the simplest possible problems associated with the analysis of DNA sequences.

4.1.1 CpG islands

We denote by CpG the dinucleotide C *followed by* G (the notation C G or C-G denotes the pair of coupled bases C and G, one on each of the two complementary strings). In the human genome, those dinucleotides tend to disappear, because whenever a cytosine C is followed by a guanine G, it tends to be modified by methylation, and methyl-C mutates easily to thymine T. For that reason, the dinucleotides CpG are less frequent than the product of the two frequencies of Cs and Gs would predict. On the other hand, the methylation process is inhibited in certain portions of the genome, around the promotor regions and the start codons. In those regions one finds many CpGs (in fact more than what the product of the frequencies of Cs and of Gs predicts). Such regions are called ‘CpG islands’.

One can formulate two kinds of questions. Given a small piece of the genome, how do we decide whether or not it is a part of a CpG island? Given a long sequence, how do we locate the CpG islands?

4.1.2 Detection of the genes in a prokaryotic genome

In a prokaryotic genome, a gene is a collection of codons (i.e. of trinucleotides), each one coding for an amino acid, flanked by a start and a stop codon. The start codon is almost always the codon ATG, while there are three possible stop codons. But while a potential stop codon encountered in the correct phase is always a stop, the codon ATG in the middle of a non-coding region need not be a start codon.

Hence in a prokaryotic genome we have *potential genes* made of a start codon, a given number of codons (i.e. a multiple of 3 nucleotides), and a stop codon. How do we distinguish, among a collection of *potential genes*, the *true genes* from the *false genes*? How do we find the genes in a prokaryotic genome?

Remark 1.1 *In the previous chapters, the random sequences (in particular, the Markov chains) were indexed by the time parameter* $n = 0, 1, 2, \ldots$. *In this chapter, we consider random sequences indexed by the position* n *on a genome sequence. We shall let* $n = 1, 2, \ldots$, *that is, the index* n *starts at* 1, *and not at* 0.

4.2 The i.i.d. model

Let X_1 denote the first nucleotide of our sequence. Its *probability law* is defined by the vector $p = (p_{\mathrm{A}}, p_{\mathrm{C}}, p_{\mathrm{G}}, p_{\mathrm{T}})$ given by

$$p_{\mathrm{A}} = \mathbb{P}(X_1 = \mathrm{A}),\ p_{\mathrm{C}} = \mathbb{P}(X_1 = \mathrm{C}),\ p_{\mathrm{G}} = \mathbb{P}(X_1 = \mathrm{G}),\ p_{\mathrm{T}} = \mathbb{P}(X_1 = \mathrm{T})$$

Note that $p_{\mathrm{A}}, p_{\mathrm{C}}, p_{\mathrm{G}}, p_{\mathrm{T}} \geq 0$ and $p_{\mathrm{A}} + p_{\mathrm{C}} + p_{\mathrm{G}} + p_{\mathrm{T}} = 1$.

We shall say that the random variables $(X_1, \ldots, X_n)$ are i.i.d. if they are independent and have the same law. One might say (in the language of statistics) that the sequence $(X_1, \ldots, X_n)$ is a sample of size n of the common law of the X_i. With that sample we associate its empirical probability defined by

$$p_{\mathrm{A}}^n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathrm{A}\}}, \quad p_{\mathrm{C}}^n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathrm{C}\}},$$

$$p_{\mathrm{G}}^n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathrm{G}\}}, \quad p_{\mathrm{T}}^n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathrm{T}\}}.$$

$p^n = (p_{\mathrm{A}}^n, p_{\mathrm{C}}^n, p_{\mathrm{G}}^n, p_{\mathrm{T}}^n)$ is a probability on E.

In practice, the common law $p = (p_{\mathrm{A}}, p_{\mathrm{C}}, p_{\mathrm{G}}, p_{\mathrm{T}})$ of the X_i is unknown. At least if n is sufficiently large, p^n is a good approximation of p. Indeed, it follows from the law of large numbers that

$$p_{\mathrm{A}}^n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathrm{A}\}} \to \mathbb{E}(\mathbf{1}_{\{X_1 = \mathrm{A}\}}) = \mathbb{P}(X_1 = \mathrm{A})$$

as $n \to \infty$ (and similarly for C, G, T), and, moreover, from the central limit theorem,

$$\mathbb{P}\left(-\delta\sqrt{\frac{p_{\mathrm{A}}(1-p_{\mathrm{A}})}{n}} \leq p_{\mathrm{A}} - p_{\mathrm{A}}^n \leq \delta\sqrt{\frac{p_{\mathrm{A}}(1-p_{\mathrm{A}})}{n}}\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\delta}^{\delta} e^{-x^2/2}dx,$$

and hence since $\sqrt{p_{\mathrm{A}}(1-p_{\mathrm{A}})} \leq 1/2$,

$$\mathbb{P}\left(|p_{\mathrm{A}} - p_{\mathrm{A}}^n| > \frac{\delta}{2\sqrt{n}}\right) \leq \sqrt{\frac{2}{\pi}}\int_{\delta}^{\infty} e^{-x^2/2}dx.$$

One then can estimate p, under the assumption that the nucleotides are i.i.d., hence in particular that the region considered is *homogeneous*.

The assumption of independence need not be correct, and in fact is not really necessary, in order to justify the above procedure.

4.3 The Markov model

It is not very reasonable to assume that the nucleotides are mutually independent. For example, in a coding region, the law of the second nucleotide of a codon may depend upon which is the first nucleotide.

It is then natural to think of the sequence $(X_1, \ldots, X_n)$ as a Markov chain. However, it is useful for applications to genome sequences to consider a more general Markov property than we have considered so far.

Definition 3.1 *The sequence $(X_1, \ldots, X_n)$ is called an ℓth-order Markov chain ($\ell \geq 1$), (this is the $M\ell$ model) if for all $k > \ell$,*

$$\begin{aligned}\mathbb{P}(X_k = x_k | X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) \\ = \mathbb{P}(X_k = x_k | X_{k-\ell} = x_{k-\ell}, \ldots, X_{k-1} = x_{k-1}).\end{aligned}$$

An independent sequence follows a model $M0$. The model $M1$ is the usual Markov model, which has been studied in the previous chapters. An E-valued Mk model can be viewed as an E^k-valued $M1$ model.

4.3.1 Application to CpG islands

The data below are taken from [17]. We estimate two Markov transition matrices ($M1$ model), one for CpG islands, and one for sequences which are not in a CpG island. For the first case we obtain the estimated transition matrix

$$P^+ = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ 0.180 & 0.274 & 0.426 & 0.120 \\ 0.171 & 0.368 & 0.274 & 0.188 \\ 0.161 & 0.339 & 0.375 & 0.125 \\ 0.079 & 0.355 & 0.384 & 0.182 \end{array}$$

and for the second case we obtain

$$P^- = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ 0.300 & 0.205 & 0.285 & 0.210 \\ 0.322 & 0.298 & 0.078 & 0.302 \\ 0.248 & 0.246 & 0.298 & 0.208 \\ 0.177 & 0.239 & 0.292 & 0.292 \end{array}$$

Given a sequence $x = (x_1, \ldots, x_n)$ of nucleotides, we compute a *score* which is in fact a *log-likelihood ratio* of the form

$$S(x) = \log \frac{\mathbb{P}_{\text{model}+}(X = x)}{\mathbb{P}_{\text{model}-}(X = x)} = \sum_{i=2}^{n} \log \left(\frac{P^+_{x_{i-1}x_i}}{P^-_{x_{i-1}x_i}} \right) = \sum_{i=2}^{n} R_{x_{i-1}x_i}.$$

One can normalize this score by dividing by the length n of the sequence. The R matrix is given by

$$R = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{A} \end{array} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ -0.740 & 0.419 & 0.580 & -0.803 \\ -0.913 & 0.302 & 1.812 & -0.685 \\ -0.624 & 0.461 & 0.331 & -0.730 \\ -1.169 & 0.573 & 0.393 & -0.679 \end{array}$$

If one compares the scores $S(x)$ for various sequences for which we know whether or not they belong to a CpG island, the above score distinguishes between CpG islands and other sequences fairly well.

4.3.2 Search for genes in a prokaryotic genome

We proceed as we did in the previous subsection. Suppose that we have at our disposal *true genes* and *false genes*. We use part of the data for a *learning step*, and the rest of the data for testing whether or not our *score* distinguishes true from false genes well.

We use a first subset of the *true genes* for estimating the transition matrix of a Markov chain. Denote by P^+ the estimate obtained. The model '–' is an i.i.d. model, where the probability of each of the four nucleotides is given by its frequency, computed from a set of sequences containing both *true genes* and *false genes*. Denote by π the estimate obtained.

If x is a *potential gene*, we compute its score

$$S(x) = \log \frac{\mathbb{P}_{\text{model}+}(X = x)}{\mathbb{P}_{\text{model}-}(X = x)} = \sum_{i=2}^{n} \log \left(\frac{P^+_{x_{i-1}x_i}}{\pi_{x_i}} \right) = \sum_{i=2}^{n} R_{x_{i-1}x_i}.$$

It turns out that this statistic distinguishes *true genes* from *false genes* very poorly. However, if we choose for the + model a Markov model $M1$ for *codons*, and proceed as above, then the new statistic $S(x)$distinguishes between *true genes* and *false genes* rather well.

4.3.3 Statistics of Markov chains Mk

In order to simplify the exposition, let us restrict ourselves to the $M2$ model. In this case, what replaces the usual transition matrix P from E to E is a transition matrix from $E \times E$ to E, which gives the law of X_{k+1}, given the pair (X_{k-1}, X_k). Here $E = \{\text{A, C, G, T}\}$, hence we have a transition matrix with 16 rows (indexed by the dinucleotides $\{\text{AA, AC, } \ldots, \text{GT, TT}\}$) and 4 columns (indexed by $\{\text{A, C, G, T}\}$).

Remark 3.2 *We can also reduce to an $M1$ model on the state space $E \times E$, since if $(X_1, X_2, \ldots, X_n)$ is an E-valued $M2$ Markov chain, then $((X_1, X_2), (X_2, X_3), \ldots, (X_{n-1}, X_n))$ is an $M1$ $E \times E$-valued Markov chain. Then the transition matrix is a square matrix, and we can study the associated invariant measure.*

We estimate the transition matrix $P_{xy,z}$ with the help of the quantity

$$\frac{\sum_{k=1}^{n-2} \mathbf{1}_{\{X_k=x, X_{k+1}=y, X_{k+2}=z\}}}{\sum_{k=1}^{n-2} \mathbf{1}_{\{X_k=x, X_{k+1}=y\}}},$$

which converges almost surely towards $P_{xy,z}$ as $n \to \infty$. Note that this statistic includes a counting of trinucleotides, and so in particular of codons, so that $M2$ Markov chains are very commonly used as models for coding regions of DNA.

4.3.4 Phased Markov chains

In a 'coding region', one might guess that the transition matrix should not be taken as constant, but rather as periodic, with period 3. Since the notion of a 'periodic Markov chain' has already been used for something very different (a periodic chain is one which is not aperiodic in the sense of Definition 2.6.1), we shall use, following [34], the name *phased Markov chain* to denote a Markov chain $(X_n,\ 1 \leq n \leq N)$ such that, for all $x, y \in E$, the mapping $n \to P(X_{n+1} = y | X_n = x)$ is periodic. In the case we are discussing here, one might think of an $M2$ Markov chain, which is such that, for all $y \in E$, the quantity $P(X_{n+1} = y | X_n = x, X_{n-1} = x\prime)$ does not depend upon $x, x\prime$ for $n = 3k$, but rather only upon x for $n = 3k + 1$ and upon $x, x\prime$ for $n = 3k + 2$. This would imply, in particular, that successive codons are i.i.d. More generally, we could also assume that the codons constitute an $M1$ Markov chain.

4.3.5 Locally homogeneous Markov chains

If we consider the genomic sequence at a larger scale, we might expect that the Markov chain described above is homogeneous in the union of the non-coding

regions, as well as in the forward coding regions, in the backward coding regions, and in the union of the introns, but the chain is probably not globally homogeneous, and this is precisely what should help us to annotate the genome. The main problem is precisely to detect 'changes of model'.

There exists a substantial statistical literature concerning such change-of-model problems, but it is not clear that those algorithms would be useful for solving our problem, where it is essential to take advantage of the homogeneity of the chain on the union of regions of the same type (non-coding, forward coding, ...), and not just on one such region alone.

We shall present an algorithm due to Audic and Claverie [1] for the annotation of prokaryotic genomes. For an explanation of why this algorithm converges, and a connection with the EM algorithm (which we shall present below for the estimation of HMM parameters), see Baldi [3]. We assume that our model (which could be of type $M0$, $M1$, $M2$, ...) is described by a parameter θ (which is a probability on E in the $M0$ case, a transition probability in the $M1$ case, ...), which can take three distinct (and unknown!) values $(\theta_0, \theta_1, \theta_2)$, depending upon whether we are currently reading a non-coding region, a forward coding region, or a backward coding region.

1. *Initialization step.* We cut the sequence into intervals of length 100 (the last interval might be of length greater than 100). We put each of these intervals at random, independently of the others, into one of the three 'boxes' 0, 1, and 2. Based upon all the nucleotides which are in the union of the intervals located in box 0, we estimate the parameter θ, giving (say) an estimate $\theta_0^{(1)}$. We estimate similarly the values $\theta_1^{(1)}$ and $\theta_2^{(1)}$.

2. *Iteration step.* Suppose that each of the three 'boxes' 0, 1, and 2 contains distinct intervals of length 100 or more, which have been used to compute the estimated values $\theta_0^{(n)}$, $\theta_1^{(n)}$ and $\theta_2^{(n)}$. We first empty these boxes, and consider once more the whole sequence $\{X_n;\ 1 \leq n \leq N\}$. We extract the subsequence $\{X_n;\ 1 \leq n \leq 100\}$. We estimate the parameter θ on the basis of this subsequence, and decide which of the three values $\theta_0^{(n)}$, $\theta_1^{(n)}$ and $\theta_2^{(n)}$ is the closest to this newly estimated value. Next we ask the same question with the sequence $\{X_n;\ 10 \leq n \leq 110\}$, with the sequence $\{X_n;\ 20 \leq n \leq 120\}$, and so on until the estimated value is closest to another of the three values $\theta_0^{(n)}$, $\theta_1^{(n)}$ and $\theta_2^{(n)}$. Then we go back 50 nucleotides, and we put the selected interval (which starts at 0) in the box which was selected first. We next start afresh, with an interval of length 100, adjacent to the interval which we have just put in one of the boxes, and we repeat the above procedure. When the sequence has been exhausted, we have three boxes each containing (we hope) intervals of length at least 100. We then compute three new estimates $\theta_0^{(n+1)}$, $\theta_1^{(n+1)}$ and $\theta_2^{(n+1)}$, on the basis of the content of the boxes 0, 1, and 2 respectively.

If the sequence is effectively an aggregate of intervals whose composition is of three different types, then the algorithm converges rapidly, and when we stop, our

sequence is split into subsequences of three different types. All that remains is to decide 'which is which', and this requires some a priori knowledge, which can be obtained from already annotated sequences.

4.4 Hidden Markov models

The Bayesian point of view on a locally homogeneous Markov chain involves giving oneself an a priori law on the unknown parameter values θ_i and on the evolution of θ. More precisely, we consider another Markov chain $(Y_1, \ldots, Y_N)$, which is said to be 'hidden' since we do not observe it. For instance, in the case of a prokaryotic genome, the chain (Y_n) takes its values in the state $F = \{0, 1, 2\}$, and in the eukaryotic case one must also split the states 1 and 2 into *introns* and *exons*. In fact the situation is slightly more complicated, since one should take into account the start and stop codons, but we will discuss that further below. The advantage of this approach is that we have at our disposal algorithms which are well adapted to our problems. Denote by F the state space into which the hidden chain takes its values, and let $d = \text{card}(F)$. Recall that in the application which we have in mind, $d \geq 3$.

In order to simplify the presentation of the hidden Markov model algorithms, we will assume that $(Y_1, \ldots, Y_N)$ is a $(\mu, P)F$-valued Markov chain and that, for each n, conditionally upon $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$, the sequence of nucleotides $(X_1, \ldots, X_n)$ is independent, the law of each X_k being a given function of y_k alone. In other words, for all $1 \leq n \leq N$,

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_n = y_n) &= \prod_{k=1}^{n} \mathbb{P}(X_k = x_k | Y_k = y_k) \\ &= \prod_{k=1}^{n} Q_{y_k x_k}.\end{aligned}$$

Our problem is as follows: having observed the sequence of nucleotides $(x_1, \ldots, x_N)$, which sequence of hidden states $(y_1^*, \ldots, y_N^*)$ 'best explains' those observations? In other words, we should find the sequence which maximizes the a posteriori law, given the observations, which is the sequence

$$(y_1^*, \ldots, y_N^*) = \arg\max_{y_1, \ldots, y_N} \mathbb{P}(Y_1 = y_1, \ldots, Y_N = y_N | X_1 = x_1, \ldots, X_N = x_N).$$

Note that in this model we have as unknown parameters (μ, P, Q). In order to solve the above problem, one first needs to estimate those parameters (but we shall discuss this at the end). Given that we know the values of those parameters, our problem will be solved by the Viterbi algorithm.

But let us first describe the computation of the likelihood.

4.4.1 Computation of the likelihood

It should be clear, from the above description, that

$$\mathbb{P}_\theta(Y_1 = y_1, X_1 = x_1, Y_2 = y_2, X_2 = x_2, \ldots, Y_N = y_N, X_N = x_N)$$
$$= \mu_{y_1} P_{y_1 y_2} \times \cdots \times P_{y_{N-1} y_N} \times Q_{y_1 x_1} Q_{y_2 x_2} \times \cdots \times Q_{y_N x_N}.$$

And so

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_N = x_N)$$
$$= \sum_{y_1, y_2, \ldots, y_n \in F} \mu_{y_1} P_{y_1 y_2} \times \cdots \times P_{y_{N-1} y_N} Q_{y_1 x_1} \times \cdots \times Q_{y_N x_N}$$

But this formula cannot be used in practice, since that would involve of the order of Nd^N operations, which, for large N, is unrealistic. We now describe a recursive procedure for the computation of the likelihood, namely the forward algorithm.

Forward algorithm. Consider the sequence (indexed by n) of row vectors $\alpha(n)$ defined by

$$\alpha_y(n) = \mathbb{P}_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, Y_n = y), \quad y \in F.$$

This sequence can be computed by a 'forward' algorithm as follows:

1. Initialization:
$$\alpha_y(1) = \mu_y Q_{y x_1}, \quad y \in F.$$
2. Recurrence:
$$\alpha_y(n+1) = (\alpha(n)P)_y Q_{y x_{n+1}}, \quad y \in F.$$
The desired quantity is
$$\sum_{y \in F} \alpha_y(N).$$
This computation requires of the order of $d^2 N$ operations.

The same quantity can also be computed by the backward algorithm.

Backward algorithm. Consider the column vectors $\beta(n)$, defined by

$$\beta_y(n) = \mathbb{P}_\theta(X_{n+1} = x_{n+1}, \ldots, X_N = x_N | Y_n = y), \quad y \in F.$$

This sequence can be computed by a 'backward' recurrence as follows:

1. Initialization:
$$\beta_y(N) = 1, \quad y \in F.$$
2. Recurrence:

(a) $\tilde{\beta}_y(n) = \beta_y(n) Q_{y x_n}$, $y \in F$.

(b) $\beta(n-1) = P\tilde{\beta}(n)$.

Finally, the desired quantity is the real number (product of a row vector on the left by a column vector on the right)

$$\mu\tilde{\beta}(1).$$

Again, the number of operations necessary is of the order of d^2N. Note that

$$\tilde{\beta}_y(n) = \mathbb{P}_\theta(X_n = x_n, X_{n+1} = x_{n+1}, \ldots, X_N = x_N | X_n = y), \quad y \in F.$$

4.4.2 The Viterbi algorithm

Once the value of the parameter θ is known, the Viterbi algorithm computes the sequence $(y_1^*, \ldots, y_N^*)$ which maximizes the likelihood. Define the sequence of row vectors $\delta(n)$ by

$$\delta_y(n) = \max_{y_1, y_2, \ldots, y_{n-1}} \mathbb{P}_\theta(Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}, Y_n = y, X_1 = x_1, \ldots, X_n = x_n).$$

$\delta_y(n)$ is the highest probability of a trajectory $\{Y_k; 1 \leq k \leq n-1\}$ ending with $Y_n = y$ and corresponding to the observed nucleotides $x_1, \ldots, x_n$. We have the following recurrence formula for the $\delta(n)$:

$$\delta_y(n+1) = (\delta(n) * P)_y Q_{yx_{n+1}}$$

where the operation $*$ associates to a d-dimensional row vector δ and a $d \times d$ matrix P a d-dimensional row vector defined by

$$(\delta * P)_y = \sup_{z \in F} \delta_z P_{zy}.$$

The Viterbi algorithm involves computing $\delta(n)$ from $n = 1$ to $n = N$, then finding the optimal trajectory backward step by step as follows: given y_n^*, we deduce y_{n-1}^* from the formula

$$y_{n-1}^* = \psi_{y_n^*}(n),$$

where

$$\psi_y(n) = \arg\max_{z \in F} \delta_z(n-1) P_{zy}.$$

The Viterbi algorithm can be described as follows.

1. *Initialization:*

$$\delta_y(1) = \mu_y Q_{yx_1}, \quad y \in F;$$
$$\psi(1) = 0.$$

2. *Recurrence:* for $1 < n \leq N$,

$$\delta_y(n) = (\delta(n-1) * P)_y Q_{yx_n},$$
$$\psi_y(n) = \arg\max_{z \in F} \delta_z(n-1) P_{zy}, \quad y \in F.$$

3. *Final step:*

$$\delta^* = \max_{y \in F} \delta_y(N);$$

$$y_N^* = \arg\max_{y \in F} \delta_y(N).$$

4. *Backward recurrence*

$$y_n^* = \psi_{y_{n+1}^*}(n+1), \quad 1 \leq n < N.$$

Remark 4.1 *Note that the Viterbi algorithm belongs to the class of dynamic programming algorithms. The dynamic programming algorithm was invented by Richard Bellman in 1950, as an algorithm for control optimization; see Chapter 5 below.*

4.4.3 Parameter estimation

There are two possible strategies. The first involves estimating the parameters with the help of a 'learning' sequence which has already been annotated. In that case, one estimates the parameters of a model, where the whole sequence $\{(X_n, Y_n);\ 1 \leq n \leq N\}$ is observed. One uses the well-known estimation algorithms presented in Section 2.9.

The second strategy involves estimating the parameters on the sole basis of observation of the sequence of nucleotides. The advantage is that we do the estimation from the genome under study, and not from a different one. The drawback, of course, is that we are estimating the parameters of a model on the basis of very partial observations. However, there are classical algorithms (the EM algorithm, and its variant SEM) which allow us to solve this problem.

EM algorithm. Mathematical statistics teaches us that a good estimator of θ is the maximum likelihood estimator

$$\hat{\theta}_N = \arg\max_\theta \mathbb{P}_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N).$$

Let us henceforth use on the notations

$$O_N = \{X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N\},$$

$$Y^N = (Y_1, Y_2, \ldots, Y_N), \quad y^N = (y_1, y_2, \ldots, y_N).$$

There is no known algorithm for the computation of a global maximum of the function

$$\theta \to \mathbb{P}_\theta(O_N).$$

We now present an iterative algorithm which converges towards a local maximum of that function, the Baum–Welch algorithm. First, note that

$$\mathbb{P}_\theta(O_N) = \frac{\mathbb{P}_\theta(Y^N = y^N, O_N)}{\mathbb{P}_\theta(Y^N = y^N | O_N)},$$

hence

$$\mathbb{P}_{\theta_0}(Y^N = y^N | O_N) \log \mathbb{P}_\theta(O_N) = \mathbb{P}_{\theta_0}(Y^N = y^N | O_N) \log \mathbb{P}_\theta(Y^N = y^N, O_N) - \mathbb{P}_{\theta_0}(Y^N = y^N | O_N) \log \mathbb{P}_\theta(Y^N = y^N | O_N).$$

Summing over $y^N \in F^N$, we obtain

$$\log \mathbb{P}_\theta(O_N) = \frac{1}{\mathbb{P}_{\theta_0}(O_N)} \sum_{y^N} \mathbb{P}_{\theta_0}(Y^N = y^N, O_N) \log \mathbb{P}_\theta(Y^N = y^N, O_N) - \sum_{y^N} \mathbb{P}_{\theta_0}(Y^N = y^N | O_N) \log \mathbb{P}_\theta(Y^N = y^N | O_N)$$

from which, subtracting the same identity with $\theta = \theta_0$, we deduce

$$\begin{aligned} &\log \mathbb{P}_\theta(O_N) - \log \mathbb{P}_{\theta_0}(O_N) \\ &\quad = \frac{1}{\mathbb{P}_{\theta_0}(O_N)} \left[\sum_{y^N} \mathbb{P}_{\theta_0}(Y^N = y^N, O_N) \log \mathbb{P}_\theta(Y^N = y^N, O_N) \right. \\ &\qquad \left. - \sum_{y^N} \mathbb{P}_{\theta_0}(Y^N = y^N, O_N) \log \mathbb{P}_{\theta_0}(Y^N = y^N, O_N) \right] \\ &\qquad + \sum_{y^N} \mathbb{P}_{\theta_0}(Y^N = y^N | O_N) \log \frac{\mathbb{P}_{\theta_0}(Y^N = y^N | O_N)}{\mathbb{P}_\theta(Y^N = y^N | O_N)}. \end{aligned}$$

From the convexity of the negative log function and Jensen's inequality, the last term of the above identity is non-negative. Define

$$Q(\theta_0, \theta) = \sum_{y^N \in F^N} \mathbb{P}_{\theta_0}(Y^N = y^N, O_N) \log \mathbb{P}_\theta(Y^N = y^N, O_N).$$

It follows from the above computation that

$$Q(\theta_0, \theta) \geq Q(\theta_0, \theta_0) \Rightarrow \mathbb{P}_\theta(O_N) \geq \mathbb{P}_{\theta_0}(O_N).$$

The iterative Baum–Welch algorithm involves, at each iteration, computing θ_{n+1} as a function of θ_n according to the formula

$$\theta_{n+1} = \arg_\theta \max Q(\theta_n, \theta).$$

This algorithm can be interpreted as an EM algorithm, adapted to our problem.

The EM algorithm is well known in statistics. The expectation step (E-step) here involves computing the function $\theta \to Q(\theta_0, \theta)$, and the maximization step (M-step) involves the search for the point $\bar{\theta}$ where the function achieves its maximum.

Note that

$$\mathbb{P}_\theta(Y^N = y^N, O_N) = \mu_{y_1} Q_{y_1 x_1} \prod_{n=2}^{N} P_{y_{n-1} y_n} Q_{y_n x_n},$$

$$\log \mathbb{P}_\theta(Y^N = y^N, O_N) = \log \mu_{y_1} + \sum_{n=2}^{N} \log P_{y_{n-1} y_n} + \sum_{n=1}^{N} \log Q_{y_n x_n}$$

It is then easy to see, with the notation $\theta = (\mu, P, Q)$, that

$$Q(\theta_0, \theta) = Q_0(\theta_0, \mu) + \sum_{y \in F} Q_1(\theta_0, P_{y\cdot}) + \sum_{y \in F} Q_2(\theta_0, Q_{y\cdot}),$$

where

$$Q_0(\theta_0, \mu) = \sum_{y \in F} \mathbb{P}_{\theta_0}(O_N, Y_1 = y) \log \mu_y,$$

$$Q_1(\theta_0, P_{y\cdot}) = \sum_{n=2}^{N} \sum_{x \in F} \mathbb{P}_{\theta_0}(O_N, Y_{n-1} = y, Y_n = x) \log P_{yx},$$

$$Q_2(\theta_0, Q_{y\cdot}) = \sum_{n=1}^{N} \sum_{z \in E} \delta_{x_n z} \mathbb{P}_{\theta_0}(O_N, Y_n = y) \log Q_{yz}.$$

We see that the search for the maximum decomposes into $1 + 2\,\mathrm{card}(F)$ searches for a maximum, each of the form

$$\arg \max_{0 \le v_y \le 1; y \in F, \sum_y v_y = 1} \sum_{y \in F} w_y \log v_y,$$

where the $w_j \in [0, 1]$. We can get rid of the constraint $\sum_y v_y = 1$ by expressing one of the v_y in terms of the others, and setting the gradient to zero leads to the solution

$$v_y = \frac{w_y}{\sum_{y' \in F} w_{y'}}.$$

From this come the following formulae for the maximum $\bar{\theta} = (\bar{\mu}, \bar{P}, \bar{Q})$:

$$\bar{\mu}_y = \frac{\mathbb{P}_{\theta_0}(O_N, Y_1 = y)}{\mathbb{P}_{\theta_0}(O_N)} = \mathbb{P}_{\theta_0}(Y_1 = y | O_N), \quad y \in F,$$

$$\bar{P}_{xy} = \frac{\sum_{n=2}^{N} \mathbb{P}_{\theta_0}(O_N, Y_{n-1} = x, Y_n = y)}{\sum_{n=2}^{N} \mathbb{P}_{\theta_0}(O_N, Y_{n-1} = x)}, \quad x, y \in F,$$

$$\bar{Q}_{xz} = \frac{\sum_{n=1}^{N} \mathbb{P}_{\theta_0}(O_N, Y_n = x) \delta_{x_n z}}{\sum_{n=1}^{N} \mathbb{P}_{\theta_0}(O_N, Y_n = x)}, \quad x \in F, z \in E,$$

where x_n denotes the observed value of X_n, and δ stands for the Kronecker delta ($\delta_{u'u} = 1$ or 0, according to whether $u' = u$ or $u' \neq u$). All that remains is to see how one can compute the probabilities which appear above. In order to avoid supplementary indices, we write $\theta_0 = (\mu, P, Q)$. We have

$$\bar{\mu}_y = \frac{\tilde{\beta}_y(1)\mu_y}{\sum_{y\in F} \tilde{\beta}_y(1)\mu_y}.$$

Define $A = \{X_1 = x_1, \ldots, X_{n-1} = y_{n-1}, Y_{n-1} = x\}$, $B = \{Y_n = y, X_n = x_n, \ldots, X_N = x_N\}$. We have

$$\mathbb{P}_{\theta_0}(A \cap B) = \mathbb{P}_{\theta_0}(B|A)\mathbb{P}_{\theta_0}(A),$$
$$\mathbb{P}_{\theta_0}(A) = \alpha_x(n-1),$$

and thanks to the Markov property and the conditional independence of the X_k, given the Y_k,

$$\begin{aligned}\mathbb{P}_{\theta_0}(B|A) &= \mathbb{P}_{\theta_0}(B|Y_{n-1} = x)\\ &= \mathbb{P}_{\theta_0}(X_n = x_n, \ldots, X_N = x_N|Y_n = y)\mathbb{P}_{\theta_0}(Y_n = y|Y_{n-1} = x)\\ &= P_{xy}\tilde{\beta}_y(n).\end{aligned}$$

For the computation of the denominator, we note that $\beta_x(n-1) = \sum_{y\in F} P_{xy}\tilde{\beta}_y(n)$. We deduce from the above the formula

$$\bar{P}_{xy} = \frac{\sum_{n=2}^{N} \alpha_x(n-1)P_{xy}\tilde{\beta}_y(n)}{\sum_{n=2}^{N} \alpha_x(n-1)\beta_x(n-1)}.$$

Finally, an analogous computation produces

$$\bar{Q}_{xz} = \frac{\sum_{n=1}^{N} \alpha_x(n)\beta_x(n)\delta_{x_n z}}{\sum_{n=1}^{N} \alpha_x(n)\beta_x(n)}.$$

Remarks on implementation The Viterbi and Baum–Welch algorithms above cannot be directly implemented on a computer. The reason is that one manipulates products of a large (if N is large) number of terms which are less than 1, which produces microscopic quantities.

The Viterbi algorithm contains only products and maxima. In this case the solution is to replace the quantities of interest by their logarithm (which replaces products by sums). The final step in the search for the optimal trajectory is unchanged, since the function log is increasing.

The Baum–Welch algorithm involves sums. In this case one should instead use normalization constants.

In practice, we replace the α's by $\hat{\alpha}$'s defined by

$$\hat{\alpha}_x(n) = \left(\sum_{y\in F} \alpha_y(n)\right)^{-1} \alpha_x(n),$$

denoting

$$C(n) = \left(\sum_{y \in F} \alpha_y(n) \right)^{-1}$$
$$= c_1 c_2 \times \cdots \times c_n,$$

where

$$c_n = \left(\sum_{y \in F} (\hat{\alpha}(n-1)P)_y Q_{yx_n} \right)^{-1}.$$

We define analogously

$$\hat{\tilde{\beta}}_y(n) = c_n \times c_{n+1} \times \cdots \times c_N \tilde{\beta}_y(n),$$
$$\hat{\beta}_y(n) = c_n c_{n+1} \times \cdots \times c_N \beta_y(n),$$

and it is clear how one should rewrite $\bar{P}$ and $\bar{Q}$ in terms of the $\hat{\alpha}$s, $\hat{\beta}$s and $\hat{\tilde{\beta}}$s, in such a way that each term in the sum appearing in the numerator, as well as in the denominator, is multiplied by $C(N)$.

Note that with this notation,

$$\log \mathbb{P}_\theta(O_N) = -\log C(N) = -\sum_{n=1}^{N} \log c_n.$$

SEM algorithm. We now finally present the main ideas of the SEM algorithm, which is the most efficient one in situations to be described below. With each value of the unknown parameter θ we associate the conditional law of the hidden states, given the sequence of nucleotides, denoted

$$\mathbb{P}_\theta \left(Y_1 = y_1, \ldots, Y_N = y_N | X_1 = x_1, \ldots, X_N = x_N \right),$$

or rather

$$\mathbb{P}_\theta \left(Y_1^N = y_1^N | X_1^N = x_1^N \right).$$

The SEM algorithm is an iterative algorithm, which can be initialized with an arbitrary value θ_0. The iteration which replaces θ_n by θ_{n+1} can be decomposed into two steps as follows:

- *Simulation* We draw at random a realization of the random sequence Y_1^N, according to the law $\mathbb{P}_{\theta_n}(Y_1^N = \cdot | X_1^N = x_1^N)$. Denote by $y_1^N(n)$ the sequence thus obtained.

- *Maximization* We choose

$$\theta_{n+1} = \arg\max_\theta \mathbb{P}_\theta \left(Y_1^N = y_1^N(n), X_1^N = x_1^N \right).$$

Note that in the EM algorithm, the simulation step is replaced by the computation of the conditional expectation $\mathbb{E}_{\theta_n}(Y_1^N | X_1^N = x_1^N)$.

4.5 Hidden semi-Markov model

4.5.1 Limitations of the hidden Markov model

One consequence of the Markov property is that the time spent by a Markov chain in each of the states that it visits follows a geometric law. Consequently the model from Section 4.4 implies that the lengths of the coding and non-coding regions of a prokaryotic genome follow geometric laws. But that assumption is inconsistent with some of the data at our disposal. This is a first argument for considering a more general model. We now give another argument for abandoning the hidden Markov model.

Let us consider our problem more precisely, restricting ourselves for the sake of simplicity to eukaryotic genomes. It is of course essential to take into account the information contained in the start and stop codons. Ignoring the phased models for now, we are forced to introduce three start states, three coding states, and three stop states, each one corresponding to its place in the corresponding codon, all this being multiplied by 2 if we wish to take into account the complementary strand. We add one non-coding state. This adds up to 19 states. Of course, most of the entries of the transition matrix will be zeros, but that still means a lot of states, and the situation is much worse in the eukaryotic case. We can reduce this number by using a phased model, but then we multiply by 3 the number of transition matrices to be estimated. We might also think of working with the sequence of codons rather than nucleotides, but this would work only for the coding regions.

We will see below that a semi-Markov model allows us to reduce the number of states to 3 in the prokaryotic case, while it also allows us to choose a distribution which is more realistic than the geometric, for the length of the coding regions.

4.5.2 What is a semi-Markov chain?

Let us give one possible definition. As its name indicates, a semi-Markov chain is 'less Markovian' (i.e. forgets less about its past) than a Markov chain. Given a random sequence $(X_1, \ldots, X_N)$, and $1 < n < N$, we define, for each $1 < n < N$, the random variable η_n by

$$\eta_n = \sup\{k \geq 0;\ X_{n-k} = X_{n-k+1} = \ldots = X_n\}.$$

In the application which we have in mind, this is the number of sites located on the left of the site n, which belong to the same region as it. Of course, if one knows the realization of the sequence $(X_1, \ldots, X_n)$, one knows the value of η_n. Denote by $\varphi_n(x_1, \ldots, x_n)$ the value of η_n when $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$. In other words,

$$\varphi_n(x_1, \ldots, x_n) = \sup\{k; x_{n-k} = \ldots = x_n\},$$

whence $\eta_n = \varphi_n(X_1, \ldots, X_n)$.

Definition 5.1 *An E-valued random sequence $(X_1, \ldots, X_N)$ is called semi-Markov if and only if, for all $1 < n \leq N$, all $(x_1, \ldots, x_{n-1}, x, y) \in E^{n+1}$,*

$$\mathbb{P}(X_{n+1} = y | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x)$$
$$= \mathbb{P}(X_{n+1} = y | X_n = x, \eta_n = \varphi_n(x_1, \ldots, x_{n-1}, x)).$$

The fact that the next state of a semi-Markov chain depends not only upon the current state, but also upon the time which has already been spent in that state, allows the law of the time spent by the chain in any state to be completely arbitrary.

More precisely, a 'generic' way to specify the law of a semi-Markov chain (and also to give a method of simulation for it) is as follows.

1. We associate with each state $x \in E$ a probability distribution $(d_x(n),\ n \in \mathbb{N}\backslash\{0\})$ on $\mathbb{N}^* := \mathbb{N}\backslash\{0\}$, which is the law of the time spent by the chain at x.

2. We specify a transition matrix P on $E \times E$ of a Markov chain, whose diagonal entries are zero. This matrix describes how the chain moves from one state to another.

Let us now see how we can simulate a semi-Markov chain whose law is specified by the following data: for each $x \in E$, d_x denotes the probability distribution of the time spent by the chain at x, and $P_{x\cdot}$ is the probability distribution of the next state which is visited on leaving x. If x denotes the starting point $(X_1 = x)$, we simulate a random variable T_1 with values in $\mathbb{N}^*$, whose probability distribution is d_x. Let n denote the result of this simulation. Then $X_1 = X_2 = X_3 = \ldots = X_n = x$. We next simulate a random variable Z_1 whose probability distribution is $P_{x\cdot}$ on $E\backslash\{x\}$. Suppose that we obtain $Z_1 = y$. Then $X_{n+1} = y$, and we now simulate a random variable T_2 whose probability distribution is d_y, a random variable Z_2 whose probability distribution is $P_{y\cdot}$, and so on. These successive simulations must clearly be mutually independent.

4.5.3 The hidden semi-Markov model

Let us restrict ourselves once more to the prokaryotic case. Consider three hidden states, state 0 standing for *non-coding*, state 1 for *forward coding*, state 2 for *backward coding*. State 0 is a *Markovian* state (i.e. the future after visiting state 0 does not depend upon the past – we shall see at the end of the next subsection why we make this restriction), in other words the law of the lengths of the non-coding regions is geometric with parameter q (to be estimated). States 1 and 2 are said to be *semi-Markov*. We will choose for the law of the lengths of the forward and backward coding regions the image under the mapping $x \to 3x$ of a negative binomial law with parameters $m \in \mathbb{N}^*$ and $0 < p < 1$ (i.e. this law concerns the number of codons rather than the number of nucleotides).

Definition 5.2 *We say that the law of the random variable T is negative binomial with parameters m and p if T is the number of times one must toss a coin in order*

to obtain exactly m heads, p being the probability of obtaining heads at each toss. In other words,

$$\mathbb{P}(T = k) = \begin{pmatrix} k-1 \\ m-1 \end{pmatrix} (1-p)^{k-m} p^m,$$

which is the probability of obtaining $m-1$ heads in the first $k-1$ tosses, times the probability of obtaining head at the kth toss. T can also be considered as the sum of m mutually independent geometric (p) random variables.

It would be natural to choose for the value of the parameter m the smallest number of amino acids which a gene contains, plus 2 (for the start and stop codons). Unfortunately, this minimal number is extremely small in a few exceptional cases, while it is of the order of 10 in almost all cases. A reasonable choice seems to be $m = 10$, but this choice should be validated (and confronted with the sequence under study).

The parameter p must be estimated.

The transition probability from one state to another is chosen as follows. We admit that each coding region (forward as well as backward) is followed by a non-coding region. Hence $P_{10} = P_{20} = 1$. Moreover, $P_{01} + P_{02} = 1$, and one can either choose $P_{01} = 1/2$ or estimate that quantity.

Let us now discuss the law of the nucleotides, given the hidden state. One might allow the nucleotides in the non-coding region to be i.i.d., the common law having to be estimated. In a coding region, we assume, for example, that the codons are mutually independent, the first being a start codon, the last having the uniform law on the possible stop codons, and the rest being i.i.d. with values in the set of codons which code for an amino acid (in particular, those codons cannot take as value a possible stop codon). The description of the law of the backward coding regions can easily be deduced from that of the forward coding regions.

4.5.4 The semi-Markov Viterbi algorithm

Let us now describe the Viterbi algorithm in the present situation (which is a mixture of hidden Markov and hidden semi-Markov models). As for hidden Markov chains, the idea is to compute the quantities $\delta_y(n)$, for each hidden state y and $1 \leq n \leq N$. For $y = 0$ we define, as before,

$$\delta_y(n) = \max_{y_1, y_2, \ldots, y_{n-1}} \mathbb{P}_\theta(Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}, Y_n = y, X_1 = x_1, \ldots, X_n = x_n).$$

For $y = 1$ (and similarly for $y = 2$), we define

$$\delta_y(n) = \max_{y_1, \ldots, y_{n-1}} P_\theta(Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}, Y_n = y, Y_{n+1} \neq y, X_1 = x_1, \ldots, X_n = x_n).$$

The recurrence formula is

$$\delta_y(n) = \max\Big\{ P_\theta(X_1^n = x_1^n | Y_1^n = y, Y_{n+1} \neq y) d_y(n) \mu_y,$$
$$\max_{1 \leq k \leq n-1} \big[P_\theta(X_{n-k+1}^n = x_{n-k+1}^n | Y_{n-k} \neq y, Y_{n-k+1}^n = y, Y_{n+1} \neq y),$$
$$\max_{z \neq y} \left(\delta_z(n-k) P_{zy}\right) d_y(k) \big] \Big\}.$$

Note that the fact that we need to compute a maximum over $1 \leq k \leq n$ at each step makes the algorithm quadratic in N, which is bad news. However, the range of this maximum can be restricted, since a coding region terminates at the first stop codon. This remark is still valid for an exon in the case of a eukaryotic genome: it ends at the latest at the first stop codon (either since this codon effectively marks the end of the gene, or else since this codon is located in an intron, or possibly in the wrong reading phase, but beyond the first intron). This remark does not apply to non-coding regions. This is why we choose the state 0 to be a Markov state.

4.5.5 Search for genes in a prokaryotic genome

The a priori law of the Y's

We have three hidden states: coding $+ = 1$, coding $- = 2$, non-coding $= 0$.

- The law of Y_1 is $\mu = (\mu_0, \mu_1, \mu_2)$.
- If $Y_n = 0$, we choose Y_{n+1} according to the probability $p = (p_0, p_1, p_2)$, with $p_0 > 0$, $p_1 > 0$, $p_2 > 0$ and $p_0 + p_1 + p_2 = 1$.
- If $Y_{n-1} = 0$ and $Y_n = 1$ (or if $Y_{n-1} = 0$ and $Y_n = 2$), we choose the length of the coding region according to a law whose support is included in the set of multiples of 3. Just after a coding region, Y is in the non-coding state (state 0).

The conditional law of the X's given the Y's

- Given that $Y_n = 0$, X_n is independent of all other random variables, and its law is $q = (q_A, q_C, q_G, q_T)$, with $q_A > 0$, $q_C > 0$, $q_G > 0$, $q_T > 0$ and $q_A + q_C + q_G + q_T = 1$).
- If $Y_n = 1$ (and analogously for $Y_n = 2$), it depends upon where we are in the coding region.

  1. The first codon is a start codon.
  2. The next codons are chosing according to a law which charges the various codons which code for an amino acid, the possible stop codons being excluded.
  3. The last codon is chosen according to a law whose support consists of the stop codons.

Computing the δ's

For $y = 1$ and $y = 2$, we define

$$\delta_y(n) = \max_{y_1,\ldots,y_{n-1}} \mathbb{P}(Y_1^{n-1} = y_1^{n-1}, Y_n = y, Y_{n+1} = 0, X_n^1 = x_n^1).$$

For $y = 0$, we define

$$\delta_0(n) = \max_{y_1,\ldots,y_{n-1}} \mathbb{P}(Y_1^{n-1} = y_1^{n-1}, Y_n = 0, X_n^1 = x_n^1).$$

For $n < 3m$, $\delta_1(n) = \delta_2(n) = 0$. Moreover, if $Y_n = 0$, then the whole sequence $Y_1, \ldots, Y_n$ is constant and equals 0. In that case

$$\delta_0(n) = \mathbb{P}(Y_1^n = 0, X_1^n = x_1^n) = \mu_0 p_0^{n-1} \prod_{k=1}^{n} q_{x_k}.$$

Now suppose that $n \geq 3m$.

- Consider first the recurrence formula in the case $y = 0$. If $n > 3m$, the set of $y_1, \ldots, y_{n-1}$ over which one maximizes is subdivided into three classes, depending upon the value of y_{n-1}. We have

$$\delta_0(n) = \max[\delta_0(n-1)p_0, \delta_1(n-1), \delta_2(n-1)]q_{x_n}.$$

- Consider now the case $y = 1$ (the case $y = 2$ is treated similarly).

  1. Either $Y_1^n = 1$ and $Y_{n+1} \neq 1$ (in this case n is a multiple of 3). This implies that

$$\mathbb{P}(Y_1^n = 1, Y_{n+1} \neq 1, X_1^n = x_1^n) = \mathbb{P}(X_1^n = x_1^n \mid Y_1^n = 1, Y_{n+1} \neq 1) d_1(n) \mu_1.$$

  2. Or the hidden sequence is constant and equal to 1 on the interval $[n - k + 1, n]$, hence (with $y_{n-k} \neq 1$, consequently $y_{n-k} = 0$, which implies that k is a multiple of 3; and the following quantity is zero unless both $(x_{n-k}, x_{n-k+1}, x_{n-k+2})$ is the start codon and (x_{n-2}, x_{n-1}, x_n) is a stop codon)

$$\begin{aligned}
&\max_{y_1,\ldots,y_{n-k-1}} \mathbb{P}(Y_1^{n-k} = y_1^{n-k}, Y_{n-k+1}^n = 1, Y_{n+1} = 0, X_1^n = x_1^n) \\
&= \delta_0(n-k)p_1 \\
&\quad \times \mathbb{P}(X_{n-k+1}^n = x_{n-k+1}^n, Y_{n-k+1}^n = 1, Y_{n+1} = 0 | Y_{n-k+1} = 1, Y_{n-k} = 0) \\
&= \delta_0(n-k)p_1 \\
&\quad \times \mathbb{P}(X_{n-k+1}^n = x_{n-k+1}^n | Y_{n-k} = 0, Y_{n-k+1}^n = 1, Y_{n+1} = 0) d_1(k).
\end{aligned}$$

The above conditional probability factorizes as

$$\prod_{j=0}^{k/3} \mathbb{P}(X_{n-k+3j+1}^{n-k+3j+3} = x_{n-k+3j+1}^{n-k+3j+3} | Y_{n-k} = 0, Y_{n-k+1}^{n} = 1, Y_{n+1} \neq 1)$$

$$= \mathbb{P}_{\text{start}}(x_{n-k+1}^{n-k+3}) \prod_{j=1}^{k/3-1} \mathbb{P}_{\text{coding}}(x_{n-k+3j+1}^{n-k+3j+3}) \mathbb{P}_{\text{stop}}(x_{n-2}^{n})$$

$$:= \mathbb{P}_{\text{C+}}(x_{n-k+1}^{n}).$$

Finally, with the convention $\delta_0(0) = \mu_1/p_1$,

$$\delta_1(n) = \max_{k \text{ multiple of } 3;\, 3m \leq k \leq n} \delta_0(n-k) p_1 \mathbb{P}_{\text{C+}}(x_{n-k+1}^{n}) d_1(k).$$

4.6 Alignment of two sequences

Consider two nucleotide sequences (or amino acid sequences)

$$x_1^n = (x_1, \ldots, x_n),$$
$$y_1^m = (y_1, \ldots, y_m).$$

Suppose we are considering nucleotide sequences, such as

A A C G G T T C C C A G T T

A C G T T T C C A G T C

We wish to align them as follows:

A A C G G T T C C C A G T T

A – C G T T T C C – A G T C

To do this, we need to create *gaps* in one of the sequences (one might in other examples wish to create gaps in both sequences). Note that whenever $n \neq m$, there are at least $|n - m|$ gaps.

We can associate with each alignment of two sequences x_1^n and y_1^m a *score* which measures the quality of the alignment. Suppose that the common length of the aligned sequences (counting the gaps) is T. Clearly $T \geq \sup(n, m)$. For $1 \leq i \leq T$, if the nucleotide a_i is at position i in the first sequence and the nucleotide b_i in the second sequence ($a_i = x_i$ only when there is no gap in the first sequence at the left of position i; the same remark is valid for the second sequence), then position i contributes $s(a_i, b_i)$ to the global score of the alignment. If a gap opens at position i in one of the two sequences (in the sense that there is a gap in the first (or second)

sequence at position i, and no gap in the same sequence at position $i-1$), then that position contributes $-d$ to the global score, and if a gap continues at position i, then i contributes $-e$ to the global score. Finally, the global score of an alignment of length T is

$$\sum_{\substack{1\le i\le T,\\ \text{no gap at } i}} s(a_i, b_i) - \sum_{\substack{\text{gaps of the}\\ \text{first sequence}}} (d + (\ell_j - 1)e) - \sum_{\substack{\text{gaps of the}\\ \text{second sequence}}} (d + (\ell\prime_k - 1)e),$$

if ℓ_j ($\ell\prime_k$) denotes the length of the jth (kth) gap of the first (second) sequence. s is a map from $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^2$ into $\mathbb{R}$, which is maximal on the diagonal. In the case of amino acid sequences, we replace $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ by the set of 20 amino acids. In both cases, s is chosen in such a way that, for $a \neq b$, the more probable a mutation from a to b (or vice versa), the bigger $s(a, b)$.

4.6.1 The Needleman–Wunsch algorithm

The search for an optimal alignment can be done with the help of a dynamic programming algorithm. Define $M(i, j)$ as the best possible score among all partial alignments which end with x_i aligned with y_j. $I_x(i, j)$ is the best score among all partial alignments which end with x_i aligned with a gap in the second sequence, y_j being the last nucleotide in the second sequence on the left of the gap. Finally, $I_y(i, j)$ is the best score among all partial alignments which end with y_j aligned with a gap in the first sequence, x_i being the last nucleotide of the first sequence on the left of the gap. Then we have the following recurrence formulae:

$$M(i, j) = \max \begin{cases} M(i-1, j-1) + s(x_i, y_j), \\ I_x(i-1, j-1) + s(x_i, y_j), \\ I_y(i-1, j-1) + s(x_i, y_j); \end{cases}$$

$$I_x(i, j) = \max \begin{cases} M(i-1, j) - d, \\ I_x(i-1, j) - e; \end{cases}$$

$$I_y(i, j) = \max \begin{cases} M(i, j-1) - d, \\ I_y(i, j-1) - e. \end{cases}$$

We have excluded in these formulae the possibility that a gap in one sequence could be immediately followed by a gap in the other one, which is certainly the case for the optimal alignment provided $-d - e < \inf_{a,b} s(a, b)$. In the case $e = d$, the above triple recurrence reduces to the unique recurrence

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d. \end{cases}$$

There exist many variants of this algorithm. In particular, one can search for the best local (and not necessarily global) alignment. This is obtained in the case $d = e$ by modifying the recurrence as follows:

$$F(i, j) = \max \begin{cases} 0, \\ F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d. \end{cases}$$

A maximum of 0 indicates the beginning of a local alignment. Note that in this case, $F(\cdot, 0) = F(0, \cdot) \equiv 0$. The local alignment algorithm which we have just described is called the Smith–Waterman algorithm.

4.6.2 Hidden Markov model alignment algorithm

We now formulate the search for an optimal alignment of two sequences in the framework of hidden Markov models. This model is often called 'pair HMM' (see in particular [17]), since the hidden chain is in fact a pair of processes. as we shall see. Here the hidden chain will be denoted by $Z_1, \ldots, Z_T$. Note that T is not given a priori, it is random. This chain takes its values in a space containing four states, which we shall denote by $\{A, I, S, E\}$: A for *align*, I for *insert* a gap in the first sequence, S for *suppress*, which means inserting a gap in the second sequence, and E for *end*, which is an absorbing state of the chain.

We can now specify both the a priori law of the chain $Z_1, \ldots, Z_T$, and the conditional law of the double sequence

$$\left(\begin{pmatrix} \hat{X}_1 \\ \hat{Y}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \hat{X}_{T-1} \\ \hat{Y}_{T-1} \end{pmatrix} \right)$$

with values in the set $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}, -\}$, given the sequence $Z_1, \ldots, Z_T$ (the value of T, the length of the sequence, being part of the unknowns which are specified by the alignment).

We now consider $\{Z_t; t \geq 1\}$ as a Markov chain with values in the space $\{A, I, S, E\}$, and $T := \inf\{t \geq 1; Z_t = E\}$. The transition matrix of this chain takes the form

$$\begin{pmatrix} 1-2\delta-\tau & \delta & \delta & \tau \\ 1-\varepsilon-\tau & \varepsilon & 0 & \tau \\ 1-\varepsilon-\tau & 0 & \varepsilon & \tau \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We choose for the law of Z_1 the probability

$$(1-\tau)^{-1}(1-2\delta-\tau \quad \delta \quad \delta \quad 0).$$

The law of $(Z_1, \ldots, Z_T)$ is now completely specified.

Note that the law of T is given by

$$\mathbb{P}(T = t) = (1-\tau)^{t-2}\tau, \quad t \geq 2.$$

One can also specify the law of the transitions, conditioned upon the fact that they do not produce E. If we let

$$P_{GH} := \mathbb{P}(Z_{t+1} = H | Z_t = G, T > t+1),$$

$G, H \in \{A, I, S\}$, we have

$$P = \begin{pmatrix} \frac{1-2\delta-\tau}{1-\tau} & \frac{\delta}{1-\tau} & \frac{\delta}{1-\tau} \\ \frac{1-\varepsilon-\tau}{1-\tau} & \frac{\varepsilon}{1-\tau} & 0 \\ \frac{1-\varepsilon-\tau}{1-\tau} & 0 & \frac{\varepsilon}{1-\tau} \end{pmatrix}.$$

We can now describe the conditional law of the $\hat{X}$s and the $\hat{Y}$s, given $(Z_1, \ldots, Z_T)$. Let us again consider the case of sequences of nucleotides. There is no difficulty in translating the formulae to the case of amino acid sequences. We specify both a probability $\{p_{a,b}; (a,b) \in \{\mathrm{A, C, G, T}\}^2\}$ and a probability $\{q_a; a \in \{\mathrm{A, C, G, T}\}\}$. Conditionally upon $(Z_1, \ldots, Z_T)$, the random variables $(\hat{X}_t, \hat{Y}_t)$, $1 \le t < T$, are independent. Given that $1 \le t < T$ and $Z_t = A$, the law of $(\hat{X}_t, \hat{Y}_t)$ is p. Given that $1 \le t < T$ and $Z_t = I$, $\hat{X}_t = -$ and the law of $\hat{Y}_t$ is q. Given that $1 \le t < T$ and $Z_t = S$, the law of $\hat{X}_t$ is q and $\hat{Y}_t = -$. $(\hat{X}_t, \hat{Y}_t)$ is not defined for $t \ge T$. In other words,

$$\mathbb{P}\left(\begin{pmatrix} \hat{X}_1^{T-1} \\ \hat{Y}_1^{T-1} \end{pmatrix} = \begin{pmatrix} x_1^{t-1} \\ y_1^{t-1} \end{pmatrix} \middle| Z_1^T = z_1^t \right) = \prod_{1 \le i < t; z_i = A} p_{x_i y_i} \prod_{1 \le j < t; z_j = I} \mathbf{1}_{\{x_j = -\}} q_{y_j} \times \prod_{1 \le k < t; z_k = S} q_{x_k} \mathbf{1}_{\{y_k = -\}}$$

We can compute this as

$$\begin{aligned} &\mathbb{P}\left(\begin{pmatrix} \hat{X}_1^{T-1} \\ \hat{Y}_1^{T-1} \end{pmatrix} = \begin{pmatrix} x_1^{t-1} \\ y_1^{t-1} \end{pmatrix}, Z_1^T = z_1^t \right) \\ &\quad = \mu_{z_1} (1-\tau)^{t-2} \tau \prod_{1 \le i \le n \ x_i \ne -, y_i \ne -} p_{x_i y_i} \prod_{1 \le j \le n \ x_j = -} q_{y_j} \prod_{1 \le k \le n \ y_k = -} q_{x_k} \\ &\qquad \times \left(\frac{1-2\delta-\tau}{1-\tau} \right)^{\ell_1} \left(\frac{\delta}{1-\tau} \right)^{\ell_2} \left(\frac{\varepsilon}{1-\tau} \right)^{\ell_3} \left(\frac{1-\varepsilon-\tau}{1-\tau} \right)^{\ell_4} \\ &\quad = \mu_{z_1} \tau \prod_{1 \le i \le n; x_i \ne -, y_i \ne -} p_{x_i y_i} \prod_{1 \le j \le n; x_j = -} q_{y_j} \prod_{1 \le k \le n; y_k = -} q_{x_k} \\ &\qquad \times (1-2\delta-\tau)^{\ell_1} \delta^{\ell_2} \varepsilon^{\ell_3} (1-\varepsilon-\tau)^{\ell_4}, \end{aligned}$$

where ℓ_1 is the number of transitions of the chain $\{Z_\cdot\}$ from A to A, ℓ_2 is number of transitions from A to I or S (opening of a gap), ℓ_3 the number of transitions from I or S to itself (continuation of a gap), and ℓ_4 the number of transitions from I or S to A (closure of a gap).

Note that the observation is made of realizations of the two sequences $(X_1, \ldots, X_N)$ and $(Y_1, \ldots, Y_M)$ obtained from the double sequence

$$\left(\begin{pmatrix}\hat{X}_1\\ \hat{Y}_1\end{pmatrix}, \ldots, \begin{pmatrix}\hat{X}_{T-1}\\ \hat{Y}_{T-1}\end{pmatrix}\right),$$

by deleting the gaps.

We still need to specify the quantity

$$\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t).$$

In fact we will instead compute the following quantity, whose maximization with respect to z_t^1 will produce the same result:

$$\mathbb{P}^*(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t) = \frac{\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t)}{\prod_{i=1}^n q_{x_i} \prod_{j=1}^m q_{y_j}}.$$

We have

$$\mathbb{P}^*(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t) = \frac{\tau}{1-\tau} \prod_{i=1}^{t-1} v(i),$$

with

$$v(i) = \begin{cases} (1-2\delta-\tau)\frac{p_{x_{k(i)}y_{\ell(i)}}}{q_{x_{k(i)}} q_{y_{\ell(i)}}}, & \text{if } z_i = A \text{ and either } i = 1 \text{ or } z_{i-1} = A, \\ (1-\varepsilon-\tau)\frac{p_{x_{k(i)}y_{\ell(i)}}}{q_{x_{k(i)}} q_{y_{\ell(i)}}}, & \text{if } z_i = A, i > 1 \text{ and } z_{i-1} = I \text{ or } S, \\ \delta\frac{q_{y_{\ell(i)}}}{q_{y_{\ell(i)}}} = \delta, & \text{if } z_i = I, \text{ and either } i = 1 \text{ or } z_{i-1} = A, \\ \delta\frac{q_{x_{k(i)}}}{q_{x_{k(i)}}} = \delta, & \text{if } z_i = S, \text{ and either } i = 1 \text{or} z_{i-1} = A, \\ \varepsilon\frac{q_{y_{\ell(i)}}}{q_{y_{\ell(i)}}} = \varepsilon, & \text{if } z_i = I, i > 1 \text{ and } z_{i-1} = I, \\ \varepsilon\frac{q_{x_{k(i)}}}{q_{x_{k(i)}}} = \varepsilon, & \text{if } z_i = S, i > 1 \text{ and } z_{i-1} = S, \end{cases}$$

where

$$k(i) = i + \sum_{j=1}^{i-1} \mathbf{1}_{\{z_j = I\}}, \quad \ell(i) = i + \sum_{j=1}^{i-1} \mathbf{1}_{\{z_j = S\}}.$$

Note that $\prod_{i=1}^{t-1} v(i) = \prod_{i=1}^{t} v\prime(i)$, where

$$v\prime(i) = \begin{cases} (1-2\delta-\tau)\frac{p_{x_{k(i)}y_{\ell(i)}}}{q_{x_{k(i)}} q_{y_{\ell(i)}}}, & \text{if } z_i = A, \\ \delta\frac{1-\varepsilon-\tau}{1-2\delta-\tau}, & \text{if } z_i = I, \text{and either } i = 1 \text{or} z_{i-1} = A, \\ \delta\frac{1-\varepsilon-\tau}{1-2\delta-\tau}, & \text{if } z_i = S, \text{and either } i = 1 \text{or} z_{i-1} = A, \\ \varepsilon, & \text{if } z_i = I, i > 1 \text{ and } z_{i-1} = I, \\ \varepsilon, & \text{if } z_i = S, i > 1 \text{ and } z_{i-1} = S, \end{cases}$$

for $1 \leq i < t$ and

$$v\prime(t) = \begin{cases} 1, & \text{if } z_{t-1} = A, \\ \frac{1-\varepsilon-\tau}{1-2\delta-\tau}, & \text{if } z_{t-1} = I \, or \, S. \end{cases}$$

Looking for a sequence $(z^*)_1^t$ which maximizes $\mathbb{P}^*(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t)$ is equivalent to look for a sequence $(z^*)_1^t$ which maximizes

$$\log \mathbb{P}^*(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t) \log\left(\frac{\tau}{1-\tau}\right) + \sum_{i=1}^{t} \log v\prime(i).$$

It is now easy to see that the Viterbi algorithm for solving this problem coincides with the Needleman–Wunsch algorithm if we let

$$s(a,b) = \log \frac{p_{ab}}{q_a q_b} + \log(1-2\delta-\tau),$$
$$d = -\log \frac{\delta(1-\varepsilon-\tau)}{1-2\delta-\tau},$$
$$e = -\log \varepsilon.$$

Note that the algorithm computes a trajectory z_1^t which maximizes the a posteriori probability. In particular, the value t of the length of the optimal alignment is given by

$$t = \inf\{s; k(s-1) = n, \ell(s-1) = m\}.$$

4.6.3 A posteriori probability distribution of the alignment

Having just put the Needleman–Wunsch algorithm in a probabilistic framework, we will now exploit this framework in order to introduce new concepts.

Trying to align two sequences is related to the belief that there is a similarity between these two sequences, for example because they are the result of evolution from the same ancestral sequence. If an alignment is of poor quality, it may be because the alignment is not the right one, or because no good alignment exists, for example because the two sequences have nothing in common. It may be interesting to have a criterion for deciding how well two sequences x_1^n and y_1^m can be aligned. Such a criterion can be given by the probability that our hidden Markov model produces the pair of observed sequences, i.e. by the quantity

$$\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m) = \sum_{z_1^t \in \text{ alignments}} \mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t).$$

What is the *set of alignments* in which z_1^t takes its values? It is the set of sequences of arbitrary length t, whose first $t-1$ entries belong to the set $\{A, I, S\}$ and whose

last entry is E. Clearly, for the corresponding probability to be non-zero, a very strong constraint relating z_1^t to n and m must be satisfied, namely

$$t = \inf\{s; k(s-1) = n, \ell(s-1) = m\}.$$

Necessarily $t \geq \sup(n, m) + 1$.

We now describe a *forward* algorithm which computes the quantity $\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m)$. This is just the forward part of the Viterbi–Needelman–Wunsch algorithm, with the maximization replaced by a sum. That is, one computes $(f^A(i, j), f^I(i, j), f^S(i, j))$ by a *forward recurrence* on (i, j) as follows:

$$\begin{aligned}
f^{\cdot}(i, -1) &= f^{\cdot}(-1, j) = 0, \quad \forall i, j, \\
f^A(0, 0) &= 1, \ f^I(0, 0) = 0, \ f^S(0, 0) = 0, \\
f^A(i, j) &= p_{x_i y_j}[(1 - 2\delta - \tau) f^A(i-1, j-1), \\
&\qquad + (1 - \varepsilon - \tau)(f^I(i-1, j-1) + f^S(i-1, j-1))], \\
f^I(i, j) &= q_{y_j}[\delta f^A(i, j-1) + \varepsilon f^I(i, j-1)], \\
f^S(i, j) &= q_{x_i}[\delta f^A(i-1, j) + \varepsilon f^S(i-1, j)].
\end{aligned}$$

Finally,

$$\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m) = \tau[f^A(n, m) + f^I(n, m) + f^S(n, m)].$$

We can now consider the a posteriori law of the alignment, that is, the conditional probability

$$\mathbb{P}(Z_1^T = z_1^t | X_1^N = x_1^n, Y_1^M = y_1^m) = \frac{\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t)}{\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m)}.$$

The most favourable case is when that law is very much concentrated around the optimal alignment, that is, when the quantity $\mathbb{P}(Z_1^T = (z^*)_1^t | X_1^N = x_1^n, Y_1^M = y_1^m)$ is close to 1, or at least significantly non-zero. If this is not the case, it can be interesting to know whether the set of alignments which are *close to the optimal one* carries a significant mass of the a posteriori law. The a posteriori law contains much information about the quality and pertinence of the optimal alignment. We now describe a *backward* algorithm for simulating according to the a posteriori law, which makes use of the above computations of the quantities $(f^A(i, j), f^I(i, j), f^S(i, j))$.

We first choose at random, according to the probability

$$(f^A(n, m) + f^I(n, m) + f^S(n, m))^{-1}(f^A(n, m) \quad f^I(n, m) \quad f^S(n, m)),$$

whether the alignment terminates at site $t - 1$ by the alignment of x_n with y_m (choice of A), by the alignment of y_m with a gap in the first sequence (choice of I), or by the alignment of x_n with a gap in the second sequence (choice of S).

Let us now describe the iteration step of this algorithm. Suppose that at some point the algorithm led us to align x_i and y_j. Let us now examine the quantity

$$f^A(i,j) = p_{x_i y_j}[(1-2\delta-\tau) f^A(i-1,j-1) + (1-\varepsilon-\tau)(f^I(i-1,j-1) + f^S(i-1,j-1))].$$

We now decide to

- align x_{i-1} with y_{j-1} with probability $\frac{p_{x_i y_j}(1-2\delta-\tau) f^A(i-1,j-1)}{f^A(i,j)}$,
- align y_{j-1} with a gap in the first sequence sequence with probability $\frac{p_{x_i y_j}(1-\varepsilon-\tau) f^I(i-1,j-1)}{f^A(i,j)}$,
- align x_{i-1} with a gap in the second sequence with probability $\frac{p_{x_i y_j}(1-\varepsilon-\tau) f^S(i-1,j-1)}{f^A(i,j)}$.

If, however, the algorithm had led us to align y_j with a gap in the first sequence, x_i being the first *as yet unaligned* nucleotide of that sequence, we decide to

- align x_i with y_{j-1} with probability $\frac{q_{y_j} \delta f^A(i,j-1)}{f^I(i,j)}$,
- align y_{j-1} with a gap in the first sequence with probability $\frac{q_{y_j} \varepsilon f^I(i,j-1)}{f^I(i,j)}$.

The case where x_i is aligned with a gap in the second sequence is described analogously.

4.6.4 A posteriori probability of a given match

If the probability of any alignment is small, it might be the case that that of certain partial alignments is high. We will see that one can compute the probability that the alignment matches certain pairs (x_i, y_j). Denote by $x_i \diamond y_j$ the set of those alignments z_1^t which put x_i in front of y_j. We now compute the probability

$$\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, x_i \diamond y_j) = \sum_{z_1^t \in x_i \diamond y_j} \mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, Z_1^T = z_1^t).$$

In fact

$$\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, x_i \diamond y_j) = \mathbb{P}(X_1^i = x_1^i, Y_1^j = y_1^j, x_i \diamond y_j) \times \mathbb{P}(X_i^N = x_i^n, Y_j^M = y_j^m | x_i \diamond y_j).$$

The first factor in this formula is precisely the quantity $f^A(i,j)$ which was computed by a forward algorithm in the preceding subsection. The second factor is the quantity $b^A(i,j)$, which can be computed by a *backward recurrence* as follows:

$$b^A(n,m) = b^I(n,m) = b^S(n,m) = \tau;$$

$$b^{\cdot}(\cdot, m+1) \equiv b^{\cdot}(n+1, \cdot) \equiv 0.$$

For all $(i, j) \neq (n, m)$,

$$\begin{aligned}
b^A(i, j) &= (1 - 2\delta - \tau) p_{x_{i+1} y_{j+1}} b^A(i + 1, j + 1) \\
&\quad + \delta[q_{x_{i+1}} b^S(i + 1, j) + q_{y_{j+1}} b^I(i, j + 1)], \\
b^I(i, j) &= (1 - \varepsilon - \tau) p_{x_{i+1} y_{j+1}} b^A(i + 1, j + 1) + \varepsilon q_{y_{j+1}} b^I(i, j + 1), \\
b^S(i, j) &= (1 - \varepsilon - \tau) p_{x_{i+1} y_{j+1}} b^A(i + 1, j + 1) + \varepsilon q_{x_{i+1}} b^S(i + 1, j).
\end{aligned}$$

We define

$$\mathbb{P}(x_i \diamond y_j | X_1^N = x_1^n, Y_1^M = y_1^m) = \frac{\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m, x_i \diamond y_j)}{\mathbb{P}(X_1^N = x_1^n, Y_1^M = y_1^m)},$$

which it will be convenient to write as $\mathbb{P}(x_i \diamond y_j)$.

Given two sequences x_1^n and y_1^m, and an alignment z_1^t of those sequences, we shall write $(i, j) \in z_1^t$ whenever $z_1^t \in x_i \diamond y_j$, that is, if the alignment z_1^t matches x_i and y_j. The following quantity is a sort of expectation, for two given sequences x_1^n and y_1^m, of the overlap between the alignment z_1^t and an alignment taken at random according to the a posteriori law of the alignments:

$$\mathcal{A}_{x_1^n, y_1^m}(z_1^t) = \sum_{(i,j) \in z_1^t} \mathbb{P}(x_i \diamond y_j).$$

This is a new criterion for the quality of an alignment, for which we can compute the optimal alignment, by the classical dynamic programming algorithm associated with the following progressive recurrence (we abandon below the notation (x_1^n, y_1^m)):

$$A(i, j) = \max \begin{cases} A(i - 1, j - 1) + \mathbb{P}(x_i \diamond y_j), \\ A(i - 1, j), \\ A(i, j - 1). \end{cases}$$

Note that all the $\mathbb{P}(x_i \diamond y_j)$ can be readily deduced from the quantities computed by the above forward and backward algorithms.

4.7 A multiple alignment algorithm

It is often necessary to align more than two sequences. This is a very difficult problem, and most of the algorithms which solve it in a reasonable time (i.e. those which are used in practice) proceed by successive alignments of two sequences.

ProbCons is a recent algorithm which is to date one of the best in terms of precision, as well as being faster than some of its competitors; see [16]. It is one of those algorithms which measure the 'consistency' of each alignment of a pair of sequences (x, y) by confronting them with the $m - 2$ alignments of (x, z) and

(z, y) (see the matrix $P\prime$ below). The same idea was already present in the T-Coffee algorithm (see [30]), which can be considered the prototype of algorithms based upon consistency; see the survey [18]. We prefer to present ProbCons since it uses, as we shall see, the pair HMM framework which we met in the previous section. We shall introduce a change in the notation introduced above: henceforth the alignments will be denoted by a rather than z. The reason is that we will be forced to denote three sequences at the same time, for which we will write $x_1^{|x|}$, $y_1^{|y|}$ and $z_1^{|z|}$.

Given m sequences $\{s^1, \ldots, s^m\}$, the algorithm does the following successive operations:

1. Computation of the a posteriori probability matrices. For each pair $x, y \in S$, we compute the matrix

$$P(x, y) = \left(P_{ij}(x, y)\right), \quad 1 \leq i \leq |x|,\ 1 \leq j \leq |y|,$$

given by

$$P_{ij}(x, y) = \mathbb{P}(x_i \diamondsuit y_j | X_1^N = x_1^{|x|}, Y_1^M = y_1^{|y|}).$$

2. Two by two alignments. For each pair $x, y \in S$, we compute the alignment a^* which maximizes the quantity

$$\mathcal{A}_{x_1^{|x|}, y_1^{|y|}}(a),$$

and we let

$$E(x, y) = \frac{1}{\min(|x|, |y|)} \mathcal{A}_{x_1^{|x|}, y_1^{|y|}}(a^*).$$

3. Transformation by consistency. We associate with the three sequences $x, y, z \in S$ in particular the $|x| \times |z|$ matrix $P(x, z)$ and the $|z| \times |y|$ matrix $P(z, y)$. Then, for any $1 \leq i \leq |x|$ and $1 \leq j \leq |y|$,

$$(P(x, z)P(z, y))_{ij} = \sum_{k=1}^{|z|} P_{ik}(x, z) P_{kj}(z, y),$$

and we define a new matrix $P\prime(x, y)$ by

$$P\prime(x, y) = \frac{1}{|S|} \sum_{z \in S} P(x, z)P(z, y).$$

Note that, by definition,

$$P_{ij}(x, x) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

In each of the above matrices, many entries are very small. All entries which are smaller than a given value are set to zero, and an algorithm for multiplication of sparse matrices is used. The transformation $P \to P\prime$ can be iterated an arbitrary number of times. Unless otherwise specified, ProbCons does it twice.

4. A tree construction. Here we make use of the quantities $E(x, y)$ computed in step 2. The construction is done by the following iterative method. To start with, each sequence is identified with a group. With each pair of groups x and y we associate the quantity $E(x, y)$ computed in step 2. We look for the pair of groups (x, y) which maximizes the quantity $E(x, y)$. We then merge those two groups into a unique one, denoted xy, we define for any other group z the quantity

$$E(xy, z) := E(x, y)\frac{E(x, z) + E(y, z)}{2},$$

and we continue until all the sequences constitute a unique group.

5. Progressive alignment. We align all sequences two by two, using the Needleman–Wunsch algorithm, with the score

$$s(x_i, y_j) := P\prime_{ij}(x, y),$$

and gap penalties (d and e) equal to zero. Two groups are aligned following the best alignment of two sequences, one taken in each group.

6. Iterative refinement. We randomly partition the set S into two subgroups, to which we apply steps 4 and 5.

4.8 Exercises

Exercise 8.1 *We wish to use the Viterbi algorithm from hidden Markov models for the detection of CpG islands in the human genome. We choose as hidden Markov model a* $\{0, 1\}$*-valued Markov chain* $\{Y_n; 1 \leq n \leq N\}$ *(*1 *for 'CpG island',* 0 *for 'non-CpG island') with transition matrix*

$$\mathbb{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

Assume that $Y_1 = 0$*. We assume that the sequence* $\{X_1, \ldots, X_N\}$ *of observed nucleotides has been produced as follows. Let* π *denote the probability* $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ *on the set* $E = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$*. Suppose that the law of the random variable* X_1 *is* π*. Moreover, for* $n \geq 1$*, if* $X_n \neq \mathrm{C}$*, the law of* X_{n+1} *is* π*, and* X_{n+1} *is independent of all the other random variables; if* $X_n = \mathrm{C}$*,*

$$\text{the law of}(X_{n+1}) = \begin{cases} (0.3, 0.3, 0.1, 0.3), & \text{if } Y_{n+1} = 0, \\ (0.2, 0.2, 0.4, 0.2), & \text{if } Y_{n+1} = 1. \end{cases}$$

Explain how the Viterbi algorithm can be written in this case, which is slightly different from that treated above (the X_n *are not mutually independent, given* $\{Y_n; 1 \leq n \leq N\}$*).*

In particular, give explicitly the formula for the vector $\delta(n+1)$ *in terms of* $\delta(n)$*, according to whether* $X_n \neq \mathrm{C}$ *or* $X_n = \mathrm{C}$*. Give the formula for* $\delta(n_2)$ *in terms of* $\delta(n_1)$*, in the case where* $X_n \neq \mathrm{C}$*, for all* $n_1 \leq n < n_2$*, and assuming that the ratio between both coordinates of* $\delta(n_1)$ *lies between* $1/9$ *and* 9*.*

Exercise 8.2 (Programming)

1. *Simulate a non-homogeneous* $E = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$*-valued Markov chain* $\{X_n\}$*, from* $n = 1$ *to* 5000*, using three different transition matrices* P_1*,* P_2 *and* P_3*, changing the transition matrix each 200 to 400 nucleotides (the length of the homogeneous regions should vary).*

2. *Find the three matrices and the homogeneous regions from the simulated data, using the Audic–Claverie algorithm.*

3. *Test the hidden Markov model algorithm on the same data set, with and without knowledge of the three matrices* P_i*.*

4. *Repeat these operations with matrices estimated from the three types of regions of an annotated prokaryotic genome sequence. Then apply the same algorithms to that sequence, and compare the results with the annotation given in the database.*

5. *Now apply to that sequence a hidden semi-Markov model algorithm, taking into account the fact that a coding region starts with a start codon and ends with a stop codon.*

5

Control and filtering of Markov chains

Introduction

Consider a moving body (e.g. an airplane or satellite) whose trajectory, which is subject to random perturbations, must be controlled. The aim might be to keep the trajectory close to a desired one. The problem will be very different, depending upon whether or not the perturbed trajectory is directly observed.

After a short section on optimal deterministic control, we present some notions on the control of Markov chains, then on the control of Gaussian sequences, with a quadratic cost. We will then consider the problem of 'filtering' (i.e. trying to follow a randomly moving vehicle whose trajectory is partially observed) of Markov chains, and of Gaussian sequences, whose solution is given by the famous Kalman–Bucy filter. We will close this chapter with the linear quadratic partially observed control problem, whose solution combines the Kalman filter and the linear control problem with quadratic cost.

5.1 Deterministic optimal control

We consider the controlled discrete time dynamical system

$$X_n = f(X_{n-1}, u_n), \quad n \geq 1, \quad X_0 \text{ given},$$

where $X_n \in E \subset \mathbb{R}^d$, $u_n \in U \subset \mathbb{R}^k$, $f : E \times U \to E$. The sequence $\{u_n; n = 1, 2, \ldots\}$ is the 'control' which can be chosen at each time n from the set of admissible controls U, the goal being to minimize a cost of the form

$$J(u) = \sum_{n=1}^{N} L(X_n, u_n).$$

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

This means that one looks for a control $u^* = (u_1^*, \ldots, u_N^*)$ such that

$$J(u^*) = \min_{u \in U^N} J(u).$$

Assuming that such an optimal control exists (one can give conditions under which this is the case – existence is, for instance, obvious if U is a finite set), let us now give an algorithm for computing it. For this purpose, we introduce the quantities

$$\Phi(n, x) = \min_{\{u_n, \ldots, u_N \in U\}} \sum_{k=n}^{N} L(X_k, u_k),$$

where $X_{n-1} = x$. We have the *dynamic programming equation*

$$\Phi(n, x) = \min_{u \in U} \{L(f(x, u), u) + \Phi(n+1, f(x, u))\}$$

due to R. Bellman. This equation follows from the fact that if $u^{*,n} = (u_n^*, u_{n+1}^*, \ldots, u_N^*)$ is optimal for the control problem between times n and N, with the initial condition $X_{n-1} = x$, then $u^{*,n+1} \stackrel{\text{def}}{=} (u_{n+1}^*, \ldots, u_N^*)$ is optimal for the control problem between times $n+1$ and N, with the initial condition $X_n = f(x, u_n^*)$.

Dynamic programming algorithm. Suppose that E is a finite set. The algorithm first progresses backward, starting from the final time N, for computing $\Phi(n, x)$, for $n = N, N-1, \ldots, 1$, and all $x \in E$.

- At instant N, for each $x \in E$, we compute

$$\Phi(N, x) = \min_{u \in U} L(f(x, u), u),$$

  and we denote by $u^*(N, x)$ one of the arguments which realize that minimum.

- From $n+1$ to n, for each $x \in E$, we compute

$$\Phi(n, x) = \min_{u \in U} \{L(f(x, u), u) + \Phi(n+1, f(x, u))\},$$

  and denote by $u^*(n, x)$ one of the arguments which realize that minimum.

After these computations, we have at our disposal the quantities $\{\Phi(n, x);\ 1 \leq n \leq N,\ \ x \in E\}$ and, in particular,

$$\Phi(1, x) = \min_{u_1, \ldots, u_N \in U} J(u), \quad \text{if } X_0 = x,$$

for all $x \in E$.

Provided we have conserved all values $\{u^*(n, x);\ 1 \leq n \leq N,\ \ x \in E\}$, we can now construct the optimal trajectory (and – this is the most important thing – determine an optimal control) by the following forward recursion:

$$X_1^* = f(X_0, u^*(1, X_0)),$$
$$X_n^* = f(X_{n-1}^*, u^*(n, X_{n-1}^*)),$$
$$X_N^* = f(X_{N-1}^*, u^*(N, X_{N-1}^*)).$$

5.2 Control of Markov chains

Suppose now that $\{X_n; n = 0, 1, 2, \ldots\}$ is a controlled E-valued Markov chain (where E is either finite or countable), that is, its evolution is described by:

$$X_n = f(X_{n-1}, Y_n, u_n), \quad n \geq 1, \quad X_0 \in E \text{ given},$$

where $\{Y_n;\ n \geq 1\}$ is a sequence of i.i.d. F-valued random variables, $\{u_n;\ n \geq 1\}$ is a sequence of U-valued random variables, such that for all $n \geq 1$, u_n is $\sigma(X_{n-1})$ measurable and $f : E \times F \times U \to E$. If F and/or U is uncountable, the relevant measurability assumption must be imposed upon f. It is easy to verify that the sequence $\{X_n; n \geq 0\}$ thus defined is a Markov chain. We seek to minimize the cost

$$J(u) = \mathbb{E} \sum_{n=1}^{N} L(X_n, u_n).$$

We introduce as in the previous section the quantities ($1 \leq n \leq N,\ x \in E$)

$$\Phi(n, x) = \min_{\{u_n \in \mathcal{U}_n, \ldots, u_N \in \mathcal{U}_N\}} \mathbb{E}_{n,x} \sum_{k=n}^{N} L(X_k, u_k),$$

where $\mathcal{U}_k$ denotes the set of the U-valued $\sigma(X_{k-1})$-measurable random variables, and $\mathbb{E}_{n,x}$ stands for the conditional expectation $\mathbb{E}(\cdot | X_{n-1} = x)$. We then have the *dynamic programming equation*

$$\Phi(n, x) = \min_{u \in U} \mathbb{E}\{L(f(x, Y_n, u), u) + \Phi(n + 1, f(x, Y_n, u))\}$$

and we denote by $u^*(n, x)$ any of the values of $u \in U$ which realizes this minimum (such a value exists at least if U is finite).

The dynamic programming algorithm then involves, as in the deterministic case, computing the $\Phi(n, x)$ and the $u^*(n, x)$ by a backward recursion, starting with $n = N$ and finishing with $n = 1$. We can then activate the optimal control by using the control $u^*(n, X_{n-1})$ at time n, starting at time $n = 1$.

Remark 2.1 *More generally, one could look for an optimal control in the class of controls $u = (u_1, \ldots, u_N)$ such that, for each n, u_n may depend upon $(X_0, X_1, \ldots, X_{n-1})$. One would then show that there exists an optimal 'Markovian' control, that is, a control such that u_n depends only upon X_{n-1}, so that the sequence $\{X_n\}$ is a Markov chain.*

5.3 Linear quadratic optimal control

In this section, we assume that $\{X_n; n \in \mathbb{N}\}$ is an $\mathbb{R}^d$-valued Gaussian sequence which satisfies the linear recursion

$$X_n = AX_{n-1} + Bu_n + f_n + \eta_n, \quad n \geq 1,$$

where $X_0, \eta_1, \eta_2, \ldots$ is a sequence of mutually independent Gaussian random vectors, $X_0 \simeq N(\bar{X}_0, \lambda_0)$, $\eta_n \simeq N(0, Q)$, and the f_n are given vectors in $\mathbb{R}^d$. We suppose that the controls u_n take their values in $U = \mathbb{R}^k$, A is a $d \times d$ matrix, and B a $d \times k$ matrix. We seek to minimize the functional

$$J(u) = \mathbb{E} \sum_{n=1}^{N} [\langle F X_n, X_n \rangle + \langle R u_n, u_n \rangle],$$

where F (R) is a symmetric $d \times d$ ($k \times k$) matrix and R is positive definite.

The linear quadratic control problem is particularly interesting because it has an explicit solution, in the sense that we will give an explicit formula for the optimal control.

Remark 3.1 *One could introduce a more general cost functional of the form*

$$J(u) = \sum_{n=1}^{N} [\langle F(X_n - x_n), X_n - x_n \rangle + \langle R u_n, u_n \rangle],$$

where $(x_1, \ldots, x_N)$ is a 'nominal trajectory' which we wish $\{X_n\}$ to stay close to, the cost leading to a compromise between having the X_n close to the x_n, and using a control with a small norm. In any case, such a problem can be put in the form of the problem above, choosing $\tilde{X}_n = X_n - x_n$, and changing the f_n in the recurrence formula.

Note that, provided each u_n depends only upon X_{n-1}, the sequence $\{X_n;\ n \geq 0\}$ is a Markov chain, but with values in $\mathbb{R}^d$, which is not countable, so we are no longer in the framework studied so far in this book!

With the notation of Section 5.1, we have the following characterization of an optimal control:

Theorem 3.2 *For all $1 \leq n \leq N$, $x \in \mathbb{R}^d$,*

$$\Phi(n, x) = \langle G_n x, x \rangle + 2\langle h_n, x \rangle + c_n,$$

and $u^*(n, x) = -(R + B^*(F + G_{n+1})B)^{-1} B^*[(F + G_{n+1})(Ax + f_n) + h_{n+1}]$ *where the sequence $\{(G_n, h_n, c_n); 1 \leq n \leq N\}, \rangle$ is defined by the following backward recurrence:*

$$\begin{aligned}
G_n &= A^*(I - (F + G_{n+1})B(R + B^*(F + G_{n+1})B)^{-1} B^*)(F + G_{n+1})A, \\
h_n &= A^*(I - (F + G_{n+1})B(R + B^*(F + G_{n+1})B)^{-1} B^*) \\
&\quad \times [(F + G_{n+1}) f_n + h_{n+1}] \\
c_n &= c_{n+1} + \mathrm{tr}[(F + G_{n+1})Q] + \langle (F + G_{n+1}) f + 2h_{n+1}, f_n \rangle \\
&\quad - \langle B(R + B^*(F + G_{n+1})B)^{-1} B^*[(F + G_{n+1}) f_n + h_{n+1}] \\
&\quad \times (F + G_{n+1}) f_n + h_{n+1} \rangle
\end{aligned}$$

and $G_{N+1} = 0,\ \ h_{N+1} = 0,\ \ c_{N+1} = 0.$

The proof of this theorem is based upon the following elementary result:

Lemma 3.3 *Let P be a $k \times k$ self-adjoint positive definite matrix and $g \in \mathbb{R}^k$. Then*

$$\min_{u \in \mathbb{R}^k} [\langle Pu, u\rangle + 2\langle u, g\rangle] = -\langle P^{-1}g, g\rangle,$$

and the minimum is achieved by $u^* = -Pg$.

Proof of Theorem 3.2 First suppose that the formula of the statement holds for $\Phi(n+1, x)$. Then

$$\begin{aligned}
\Phi(n, x) &= \min_{u \in \mathbb{R}^k} \mathbb{E}_{n,x}[\langle FX_n, X_n\rangle + \langle Ru, u\rangle + \langle G_{n+1}X_n, X_n\rangle + 2\langle h_{n+1}, X_n\rangle] + c_{n+1} \\
&= \langle (F + G_{n+1})(Ax + f_n), Ax + f_n\rangle + 2\langle h_{n+1}, Ax + f_n\rangle + c_{n+1} \\
&\quad + \mathrm{tr}[(F + G_{n+1})Q] + \min_u \left\{ \langle (R + B^*(F + G_{n+1})B)u, u\rangle \right. \\
&\quad \left. + 2\langle B^*[(F + G_{n+1})(Ax + f_n) + h_{n+1}], u\rangle \right\}.
\end{aligned}$$

The formula for $u^*(n, x)$ follows from Lemma 3.3, and moreover

$$\begin{aligned}
\Phi(n, x) = &\langle A^*(I - (F + G_{n+1})B(R + B^*(F + G_{n+1})B)^{-1}B^*)(F + G_{n+1})Ax, x\rangle \\
&+ 2\langle A^*(I - (F + G_{n+1})B(R + B^*(F + G_{n+1})B)^{-1}B^*)[(F + G_{n+1})f_n \\
&+ h_{n+1}], x\rangle + c_{n+1} + \mathrm{tr}[(F + G_{n+1})Q] + \langle (F + G_{n+1})f_n + 2h_{n+1}, f_n\rangle \\
&- \langle B(R + B^*(F + G_{n+1})B)^{-1}B^*[(F + G_{n+1})f_n + h_{n+1}], (F + G_{n+1})f_n \\
&+ h_{n+1}\rangle,
\end{aligned}$$

from which one deduces the recurrence formulae of the statement. Indeed, since the above formulae are correct for $n = N$, if we let $G_{N+1} = 0$, $h_{N+1} = 0$, $c_{N+1} = 0$, the result is established by recurrence. □

Remark 3.4 *Only the values of the constants c_n depend upon the covariance matrices Λ_0 and Q; in particular, the formula for the optimal control does not depend upon them. It is the same as in the case $\Lambda_0 = Q = 0$, which is the deterministic case.*

5.4 Filtering of Markov chains

Let $\{X_n; n \in \mathbb{N}\}$ be an E-valued Markov chain, with transition matrix P, which is not observed. One observes the sequence $\{Y_n; n \in \mathbb{N}\}$ of F-valued random variables, given by

$$Y_n = h(X_n, \xi_n),$$

where the ξ_n are i.i.d. G-valued random variables, globally independent of the chain $\{X_n\}$, and $h : E \times G \to F$ is given. E and G are assumed to be either finite or countable.

The situation is in fact that of *hidden Markov models*, but here we are in a dynamical situation, where one seeks at each time n to 'estimate' X_n, given the past and present observations $Y_1, Y_2, \ldots, Y_n$. We want a 'recursive' algorithm, such that at time $n+1$ the computation makes use of the result of the computation at time n, as well as the new observation Y_{n+1}, without having to use the observations $Y_1, Y_2, \ldots, Y_n$ again. The idea is to be able to have the algorithm run in 'real time'. It turns out that the right quantity to compute at each time n is the conditional law Π_n of X_n, given $Y_1, Y_2, \ldots, Y_n$. We also define the conditional law $\Pi_{n|n-1}$ of X_n, given $Y_1, Y_2, \ldots, Y_{n-1}$. Recall that these probability distributions are defined as row vectors.

For any $x \in E$, $y \in F$, we let

$$g(x, y) = \mathbb{P}(h(x, \xi_n) = y) = \mathbb{P}(Y_n = y | X_n = x).$$

The evolution of the conditional laws $\{\Pi_n\}$ is given by the following theorem.

Theorem 4.1 *For all $n \geq 1$,*

$$\Pi_{n|n-1}(x) = (\Pi_{n-1} P)_x,$$

$$\Pi_n(x) = \frac{\Pi_{n|n-1}(x) g(x, Y_n)}{\sum_{x' \in F} \Pi_{n|n-1}(x') g(x', Y_n)}.$$

PROOF We may assume that the chain $\{X_n\}$ has been constructed by the recurrence formula

$$X_n = f(X_{n-1}, \eta_n),$$

where the $\{\eta_n\}$ are i.i.d., globally independent of the $\{\xi_n\}$. It follows in particular that

$$\mathbb{P}(X_n = x | Y_1, \ldots, Y_{n-1}, X_{n-1}) = \mathbb{P}(X_n = x | X_{n-1}),$$

whence

$$\begin{aligned} \Pi_{n|n-1}(x) &= \mathbb{P}(X_n = x | Y_1, \ldots, Y_{n-1}) \\ &= \mathbb{E}\left[\mathbb{P}(X_n = x | X_{n-1}) | Y_1, \ldots, Y_{n-1}\right] \\ &= (\Pi_{n-1} P)_x. \end{aligned}$$

In order to establish the second relation, first note that if $\mathbb{P}^{Y_1, \ldots, Y_{n-1}}$ denotes the conditional probability $\mathbb{P}(\cdot | Y_1, \ldots, Y_{n-1})$, we have

$$\mathbb{P}(X_n = x | Y_1, \ldots, Y_{n-1}, Y_n) = \mathbb{P}^{Y_1, \ldots, Y_{n-1}}(X_n = x | Y_n).$$

But for any probability $\mathbb{Q}$,

$$\mathbb{Q}(X_n = x | Y_n) = H(x, Y_n),$$

where

$$H(x, y) = \mathbb{Q}(X_n = x | Y_n = y)$$
$$= \frac{\mathbb{Q}(Y_n = y | X_n = x)\mathbb{Q}(X_n = x)}{\sum_{x'} \mathbb{Q}(Y_n = y | X_n = x')\mathbb{Q}(X_n = x')}$$

from Bayes's formula. The second relation follows, since

$$\mathbb{P}^{Y_1,\ldots,Y_{n-1}}(Y_n = y | X_n = x) = \mathbb{P}(h(x, \xi_n) = y) = g(x, y).$$

□

5.5 The Kalman–Bucy filter

In the Gaussian case, the computations become simpler, and we obtain the famous 'Kalman–Bucy filter', which is very widely used in practice.

5.5.1 Motivation

Consider a satellite which moves in space according to the differential equation

$$\frac{dx}{dt}(t) = f(x(t)), \quad x(0) = x_0.$$

Since numerical computations impose a discretization on the equations (and we do not wish to discuss the continuous time Kalman filter), we first replace the above ordinary differential equation by a recurrence formula for the $x_n = x(n\tau)$, where τ is a time step (in other words, $1/\tau$ is the sampling frequency). The exact formula reads

$$x_{n+1} = x_n + \int_{n\tau}^{(n+1)\tau} f(x(s))ds,$$

which we approximate by the simplest numerical scheme, the Euler scheme, whose precision is acceptable if τ is small enough:

$$x_{n+1} = x_n + \tau f(x_n),\ n \geq 0, \quad x_0 \text{ given.}$$

In fact this trajectory is a theoretical one; let us call it the 'nominal' trajectory. The point is that the above equation is not exact, since we cannot pretend that we have taken into account the attractions due to all celestial bodies (for some of them, the fact that they get close to the satellite at some give time cannot be predicted, and can be considered as random), the exact shape of the earth, etc. These arguments justify the introduction of a random model for the evolution of the x_n:

$$x_{n+1} = x_n + \tau f(x_n) + V_n,\ n \geq 0, \quad x_0 \text{ given,}$$

where the V_n are centred Gaussian vectors. In most applications, the random perturbations are small enough that we can approximate the difference $X_n = x_n - x_n^0$ between x_n and the nominal trajectory x_n^0 (note that this will be even more true if the Kalman filter is used to maintain the true trajectory close to the nominal one) by the solution of the linearized difference equation

$$X_{n+1} = A_n X_n + V_n,$$

where

$$A_n = I + \tau \nabla_x f(x_n^0).$$

When this approximation is not valid, more complex filtering algorithms need to be implemented.

In order to track the satellite, radar observations are made from the earth. But these observations need not give all components of the vector X_n (note that typically X_n denotes the six-dimensional vector corresponding to the position and speed of the satellite at time $n\tau$, while the radar might return information about the position only), and moreover that measurement contains some error (like any real measurement!). Moreover, depending upon the position of the satellite above the earth, it might be observed by one or no radar sites. Those measurements can be modelled by the observation at each time n of

$$Y_n = H_n X_n + W_n,$$

where $H_n = 0$ whenever the satellite is invisible to the various radar sites, and the W_n are again centred Gaussian random variables. Moreover, the initial condition is also a random vector, which we again assume to be Gaussian.

5.5.2 Solution of the filtering problem

We assume that X_n takes its values in $\mathbb{R}^d$, Y_n in $\mathbb{R}^k$, and that

$$X_n = AX_{n-1} + \eta_n, \quad n \geq 1,$$
$$Y_n = HX_n + \xi_n, \quad n \geq 1,$$

with $X_0, \eta_1, \xi_1, \eta_2, \xi_2, \ldots, \eta_n, \xi_n, \ldots$ mutually independent, the law of X_0 being $N(\bar{X}_0, P_0)$, the common law of the η_n being $N(0, Q)$ and that of the ξ_n being $N(0, R)$. We assume that the matrix R is invertible. Of course, we are no longer talking about Markov chains with values in a finite or countable state space. But we will not use the theory of Markov chains.

We give the formulae for the Kalman filter in the stationary case (A and H are assumed independent of n) in order to simplify notation. But the extension to the non-stationary case is both obvious and essential for the applications (see above). This ability to treat non-stationary signals is one of the main reasons for the success of the Kalman filter. It explains its superiority over older algorithms, notably the Wiener filter.

Let Π_n again denote the conditional law of X_n, given $Y_1, \ldots, Y_n$.

Theorem 5.1 *Π_n equals $N(\hat{X}_n, \Lambda_n)$, where $(\hat{X}_n, \Lambda_n)_{n\geq 0}$ is given by the recurrence formula*

$$\begin{aligned}
\hat{X}_{n+1} &= A\hat{X}_n + \Sigma_n H^\star (H\Sigma_n H^\star + R)^{-1}(Y_{n+1} - HA\hat{X}_n),\\
\Sigma_n &= A\Lambda_n A^\star + Q,\\
\Lambda_{n+1} &= \Sigma_n - \Sigma_n H^\star (H\Sigma_n H^\star + R)^{-1} H\Sigma_n,\\
\hat{X}_0 &= \bar{X}_0, \quad \Lambda_0 = P_0.
\end{aligned}$$

The theorem could be deduced from the results of the preceding section, once we had generalized them to $\mathbb{R}^d$-valued Markov chains. But we will give an independent derivation. First recall two results concerning conditioning in the Gaussian case.

Proposition 5.2 *Let $\begin{pmatrix} X \\ Y \end{pmatrix}$ be a Gaussian $(d+k)$-dimensional random vector, having the law $N\left[\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]$. We assume that $\Sigma_{22} > 0$. Then the law $\hat{\mu}_Y = N(\hat{X}, \hat{\Sigma})$, where*

(i) $\hat{X} = \bar{X} + \Sigma_{12}\Sigma_{22}^{-1}(Y - \bar{Y})$,

(ii) $\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$,

is a regular conditional probability distribution of X, given Y, that is, for all $B \in \mathcal{B}_d$,

$$\mathbb{P}(X \in B|Y) = \hat{\mu}_Y(B) \quad a.s.$$

Moreover $\hat{\Sigma} = \operatorname{cov}(X - \hat{X})$.

Proof If $\hat{X}$ denotes the random vector defined by (i), we let

$$\tilde{X} = X - \hat{X}.$$

It is easily seen that $\begin{pmatrix} \tilde{X} \\ Y \end{pmatrix}$ is a Gaussian random vector, and that

$$\operatorname{cov}(\tilde{X}, Y) = 0.$$

Hence $\tilde{X}$ and $Y$are independent, while $\hat{X}$ is a function of Y. Consequently, whenever $\varphi \in C_b(\mathbb{R}^d)$,

$$\begin{aligned}
\mathbb{E}[\varphi(X)|Y] &= \mathbb{E}[\varphi(\hat{X} + \tilde{X})|Y]\\
&= \int_{\mathbb{R}^d} \varphi(\hat{X} + x)\mathbb{P}_{\tilde{X}}(dx)\\
&= \int_{\mathbb{R}^d} \varphi(x)\hat{\mu}_Y(dx),
\end{aligned}$$

where $\hat{\mu}_Y = N(\hat{X}, \hat{\Sigma})$, $\hat{\Sigma} = \mathrm{cov}(\tilde{X})$.

Finally,

$$\begin{aligned}\mathrm{cov}(\tilde{X}) &= \mathrm{cov}(X - \hat{X}) \\ &= \mathrm{cov}(X - \bar{X} - \Sigma_{12}\Sigma_{22}^{-1}(Y - \bar{Y})) \\ &= \Sigma_{11} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.\end{aligned}$$

□

Proposition 5.3 *Let* $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ *be a Gaussian* $(d + k + \ell)$*-dimensional random vector, such that* Y *and* Z *are independent. We denote by* $\hat{\mu}_Y = N(\hat{X}, \hat{\Sigma})$ *the conditional law of* X *given* Y*, and* $\hat{\hat{\mu}}_{Y,Z} = N(\hat{\hat{X}}, \hat{\hat{\Sigma}})$ *the conditional law of* X *given* (Y, Z)*. Again we let* $\bar{X} = \mathbb{E}(X)$*. Then*

(i) $\hat{\hat{X}} = \hat{X} + \mathbb{E}(X - \bar{X}|Z) = \hat{X} + \hat{\tilde{X}}$,

(ii) $\hat{\hat{\Sigma}} = \hat{\Sigma} - \mathrm{cov}(\hat{\tilde{X}})$.

Proof We may assume without loss of generality that $\mathbb{E}(Z) = 0$. Denote by $\mathcal{U}$, $\mathcal{Y}$ and $\mathcal{Z}$ the closed subvector spaces of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ respectively generated by: the constants and the coordinates of Y, Z; the constants and the coordinates of Y; and the coordinates of Z. Then $\mathcal{U} = \mathcal{Y} \bigoplus \mathcal{Z}$, and $\mathcal{Y} \perp \mathcal{Z}$. Hence, for all $1 \leq i \leq d$, $\hat{\hat{X}}_i$, the orthogonal projection of X_i on $\mathcal{U}$, is the sum of $\hat{X}_i$, the orthogonal projection of X_i on $\mathcal{Y}$, and $\mathbb{E}(X_i - \bar{X}_i|Z)$, the orthogonal projection of X_i on $\mathcal{Z}$. Thus (i) is proved.

Consequently,

$$X - \hat{X} = X - \hat{\hat{X}} + \mathbb{E}(X - \bar{X}|Z)$$

for all $1 \leq i, j \leq d$, $\mathbb{E}(X_i - \bar{X}_i|Z) \in \mathcal{Z} \subset \mathcal{U}$, hence it is orthogonal in $L^2(\Omega, Z, \mathbb{P})$ to $X_j - \hat{\hat{X}}_j$. This implies that

$$\mathrm{cov}(X - \hat{X}) = \mathrm{cov}(X - \hat{\hat{X}}) + \mathrm{cov}(\mathbb{E}(X - \bar{X}|Z)),$$

which, together with the last assertion of Proposition 5.5.2, implies (ii). □

Proof of Theorem 5.1 Since $(X_n, Y_1, Y_2, \ldots, Y_n)$ is a Gaussian random vector, it follows from Proposition 5.5.2 that $\Pi_n = N(\hat{X}_n, \Lambda_n)$, where $\hat{X}_n$ is an affine function of $Y_1, \ldots, Y_n$, and $\Lambda_n = \mathrm{cov}(X_n - \hat{X}_n)$. It remains to compute $(\hat{X}_{n+1}, \Lambda_{n+1})$ as a function of $(\hat{X}_n, \Lambda_n)$. Since

$$X_{n+1} = AX_n + \eta_{n+1},$$

where η_{n+1} is centred and independent of $(X_n, Y_1, \ldots, Y_n)$, hence also of $\hat{X}_n$,

$$\mathbb{E}(X_{n+1}|Y_1, \ldots, Y_n) = A\hat{X}_n$$

and, moreover,

$$\operatorname{cov}(X_{n+1} - A\hat{X}_n) = A\Lambda_n A^* + Q.$$

It remains to apply Proposition 5.3, in order to 'add the conditioning by Y_{n+1}'. For that purpose, we need to define the'innovation'

$$\begin{aligned} I_{n+1} &= Y_{n+1} - \mathbb{E}(Y_{n+1}|Y_1, \ldots, Y_n) \\ &= Y_{n+1} - HA\hat{X}_n \\ &= HA\tilde{X}_n + H\eta_{n+1} + \xi_{n+1}, \end{aligned}$$

where $\tilde{X}_n = X_n - \hat{X}_n$.

Note that $(Y_1, \ldots, Y_n, I_{n+1})$ is a Gaussian random vector, and that the coordinates of I_{n+1} are orthogonal in $L^2(\Omega, Z, \mathbb{P})$ to those of $Y_1, \ldots, Y_n$. Hence, $(Y_1, \ldots, Y_n)$ and I_{n+1} are independent. Moreover, I_{n+1} is centred. Since in addition $\sigma(Y_1, \ldots, Y_n, Y_{n+1}) = \sigma(Y_1, \ldots, Y_n, I_{n+1})$, we can use Proposition 5.3, which tells us that

$$\begin{aligned} \hat{X}_{n+1} &= A\hat{X}_n + \mathbb{E}(X_{n+1} - \mathbb{E}X_{n+1}|I_{n+1}), \\ \Lambda_{n+1} &= A\Lambda_n A^* + Q - \operatorname{cov}(\hat{X}_{n+1} - A\hat{X}_n). \end{aligned}$$

The above conditional expectation can be computed with the help of Proposition 5.2. Now

$$\begin{aligned} \mathbb{E}(X_{n+1}I_{n+1}^*) &= A\mathbb{E}\left[X_n(X_n - \hat{X}_n)^*\right]A^*H^* + QH^* \\ &= A\Lambda_n A^*H^* + QH^*, \\ \mathbb{E}I_{n+1}I_{n+1}^* &= HA\Lambda_n A^*H^* + HQH^* + R. \end{aligned}$$

Hence,

$$\begin{aligned} \hat{X}_{n+1} =& A\hat{X}_n + (A\Lambda_n A^* + Q)H^*\left[H(A\Lambda_n A^* + Q)H^* + R\right]^{-1}(Y_{n+1} - HA\hat{X}_n), \\ \Lambda_{n+1} =& A\Lambda_n A^* + Q \\ & - (A\Lambda_n A^* + Q)H^*\left[H(A\Lambda_n A^* + Q)H^* + R\right]^{-1}H(A\Lambda_n A^* + Q), \end{aligned}$$

which proves the theorem. □

5.6 Linear–quadratic control with partial observation

We now consider the following problem:

$$X_n = AX_{n-1} + Bu_n + f_n + \eta_n, \quad n \geq 1,$$
$$Y_n = HX_n + \xi_n,$$

where again $X_0, \eta_1, \xi_1, \eta_2, \xi_2, \ldots$ is a sequence of mutually independent Gaussian random vectors, X_0 and the η_n being d-dimensional, the ξ_n k-dimensional, with an invertible covariance matrix R, and u_n takes values in $\mathbb{R}^\ell$ and is a function of $Y_1, \ldots, Y_{n-1}$. The aim is to minimize the criterion

$$J(u) = \mathbb{E}\sum_{n=1}^{N}[\langle FX_nX_n + \langle Qu_n, u_n\rangle].$$

We shall need the following technical result. Define the sequences

$$X_n^0 = AX_{n-1}^0 + f_n + \eta_n, \ n \geq 1, \quad X_0^0 = X_0,$$
$$Y_n^0 = HX_n^0 + \xi_n, \quad n \geq 1,$$

and let $\mathcal{Y}_n = \sigma(Y_1, \ldots, Y_n)$, $\mathcal{Y}_n^0 = \sigma(Y_1^0, \ldots, Y_n^0)$, $n \geq 1$, $\mathcal{Y}_0 = \mathcal{Y}_0^0 = \{\emptyset, \Omega\}$.

Lemma 6.1 *For all n, $\mathcal{Y}_n = \mathcal{Y}_n^0$.*

Proof Let

$$X_n^u = AX_{n-1}^u + Bu_n, \ n \geq 1, \quad X_0^u = 0,$$
$$Y_n^u = HX_n^u, \quad n \geq 1.$$

Since $(u_1, \ldots, u_n)$ is $\mathcal{Y}_{n-1}$-measurable, the same is true with $(Y_1^u, \ldots, Y_n^u)$. But

$$Y_n = Y_n^0 + Y_n^u, \quad n \geq 1.$$

Hence $(Y_1^0, \ldots, Y_n^0)$ is $\mathcal{Y}_n$-measurable and $\mathcal{Y}_n^0 \subset \mathcal{Y}_n$. Now since u_1 is deterministic (because it is $\mathcal{Y}_0$-measurable), Y_1^u is known, hence $Y_1 = Y_1^0 + Y_1^u$ is a function of Y_1^0, hence $\mathcal{Y}_1 \subset \mathcal{Y}_1^0$. Suppose that $\mathcal{Y}_n \subset \mathcal{Y}_n^0$. Then $(u_1, \ldots, u_{n+1})$ is $\mathcal{Y}_n^0$-measurable, and the same is true for Y_{n+1}^u, hence

$$Y_{n+1} = Y_{n+1}^0 + Y_{n+1}^u$$

is $\mathcal{Y}_{n+1}^0$-measurable. □

A consequence of Lemma 6.1 is that the sequence of sigma-algebras $(\mathcal{Y}_1, \ldots, \mathcal{Y}_N)$ does not depend upon the choice of the control. It is now clear that

$$J(u) = \mathbb{E}\sum_{n=1}^{N}\left[\mathbb{E}^{\mathcal{Y}_n}\langle FX_nX_n + \langle Ru_n, u_n\rangle\right].$$

But

$$X_n = \hat{X}_n + \tilde{X}_n,$$

where $\hat{X}_n = \mathbb{E}(X_n|\mathcal{Y}_n)$, and $\tilde{X}_n$ is independent of $\mathcal{Y}_n$. Hence, if we let $\Lambda_n = \mathrm{cov}(\tilde{X}_n)$,

$$J(u) = \mathbb{E}\sum_{n=1}^{N}\left[\langle F\hat{X}_n, \hat{X}_n\rangle + \langle Ru_n, u_n\rangle + \mathrm{tr}F\Lambda_n\right].$$

It remains to solve the problem (see Section 5.4)

$$\begin{aligned}\hat{X}_n =& (I - \Sigma_{n-1}H^*(H\Sigma_{n-1}H^* + R)^{-1}H)A\hat{X}_{n-1} + Bu_n + f_n\\ &+ \Sigma_{n-1}H^*(H\Sigma_{n-1}H^* + R)^{-1}Y_n,\\ \hat{X}_0 =& \bar{X}_0.\end{aligned}$$

We seek to minimize

$$\hat{J}(u) = \mathbb{E}\sum_{n=1}^{N}\left[\langle F\hat{X}_n, \hat{X}_n\rangle + \langle Ru_n, u_n\rangle\right].$$

This is a 'deterministic' optimal control problem (since the Y_n's are observed, hence known), whose solution is given by Theorem 5.3.2.

5.7 Exercises

Exercise 7.1 (Programming) *Consider the Kalman–Bucy filtering model, with $d = k = 1$, $A = 0.98$, $X_0 = 1$, $Q = 2$, $R = 2$, $H_n = \cos(\pi n/13)$ (or some other periodic function to be chosen).*

1. *Simulate X_n and Y_n, for $1 \le n \le 100$.*
2. *Write a program which computes the Kalman estimate $\hat{X}_n$ of X_n, for $1 \le n \le N$. Graph $(X_n, \hat{X}_n)$ and the estimation error.*
3. *Do the second step with other values of A, and possibly also of Q and R. To what extent can the Kalman filter tolerate errors in the model?*

Exercise 7.2 1. *Give the Kalman filter formulae in the non-homogeneous case, where the matrices A and H depend upon n. In other words, the model from the beginning of Section 5.5.2 is replaced by*

$$\begin{aligned}X_n &= A_nX_{n-1} + \eta_n, \quad n \ge 1,\\ Y_n &= H_nX_n + \xi_n, \quad n \ge 1.\end{aligned}$$

2. *Now suppose that we specify, for each* $n \geq 1$, *Borel measurable and bounded mappings*

$$(y_1, \ldots, y_{n-1}) \to A(y_1, \ldots, y_{n-1}) \text{ from } \mathbb{R}^{(n-1)\times k} \text{ into } \mathbb{R}^{d\times d},$$
$$(y_1, \ldots, y_{n-1}) \to H(y_1, \ldots, y_{n-1}) \text{ from } \mathbb{R}^{(n-1)\times k} \text{ into } \mathbb{R}^{k\times d},$$

and we now assume that the sequences $\{X_n\}$ *and* $\{Y_n\}$ *are related by the formulae*

$$X_n = A(Y_1, \ldots, Y_{n-1})X_{n-1} + \eta_n, \quad n \geq 1,$$
$$Y_n = H(Y_1, \ldots, Y_{n-1})X_n + \xi_n, \quad n \geq 1,$$

the assumptions on $(X_0, \eta_1, \xi_1, \ldots, \eta_n, \xi_n, \ldots)$ *being the same as in Section 5.2. Clearly the sequence of random vectors* $\{(X_n, Y_n); n \geq 1\}$ *is no longer Gaussian. Show that for each* $n \geq 1$, *the conditional law of* X_n, *given* $Y_1, \ldots, Y_n$ *is Gaussian, and give the recurrence formulae for the mean and covariance matrix of that law. This model is called 'conditionally Gaussian'.*

6

The Poisson process

Introduction

In this chapter and the two following, we will study Markov processes which are indexed by $\mathbb{R}_+$, with values in a finite or countable set E, and which are constant between their jumps, which happen at random times. These are called *jump Markov processes*.

In this chapter, we shall introduce the 'prototype' of jump Markov processes, namely the Poisson process. This process models random distributions of points on $\mathbb{R}_+$, which could be times of collisions of particles, times of arrivals of customers in a queue, times of arrivals of telephone calls, etc.

6.1 Point processes and counting processes

A *point process* on $\mathbb{R}_+$ can be described as an increasing sequence of random points

$$0 < T_1 < T_2 < \cdots < T_n < \cdots$$

which are random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In addition to the above inequalities, we assume that $T_n \uparrow \infty$, $n \to \infty$. Let

$$S_1 = T_1, S_2 = T_2 - T_1, \ldots, S_n = T_n - T_{n-1}, \ldots.$$

The T_n are the times when events happen, and the S_n are waiting times between successive events.

We define the random *counting function* $\{N_t; t \geq 0\}$ of the point process $\{T_n; n \in \mathbb{N}\}$ as follows:

$$N_t = \sup\{n; T_n \leq t\} = \sum_{j \geq 1} \mathbf{1}_{\{T_j \leq t\}}.$$

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

Thus N_t is the number of events which have happened before time t. Note that $N_0 = 0$, since $T_1 > 0$; and for all $t > 0$, $N_t < \infty$ since $T_n \uparrow \infty$, $n \to \infty$. For $0 \leq s < t$, $N_t - N_s$ is the number of events which have happened during the time interval $]s, t]$. A typical trajectory of the process $\{N_t; t \geq 0\}$ is shown in Figure 6.1. Note that the trajectories of $\{N_t\}$ are *right continuous*.

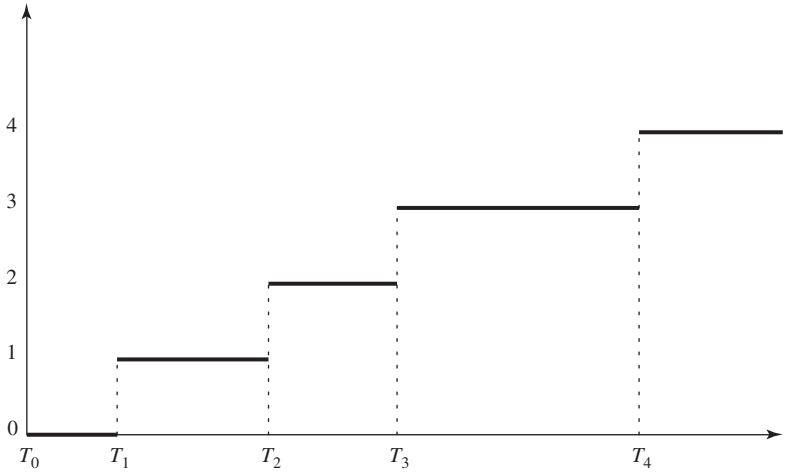


Figure 6.1 Trajectory of a Poisson process.

The knowledge of $\{N_t; t \geq 0\}$ is equivalent to that of the sequence $\{T_n; n \in \mathbb{N}\}$, and we have the identities

$$\{N_t \geq n\} = \{T_n \leq t\},$$
$$\{N_t = n\} = \{T_n \leq t < T_{n+1}\},$$
$$\{N_s < n \leq N_t\} = \{s < T_n \leq t\}.$$

6.2 The Poisson process

Definition 2.1 *We will say that the point process $\{T_n; n \in \mathbb{N}\}$ or its counting function $\{N_t; t \geq 0\}$ is a Poisson process if $\{N_t; t \geq 0\}$ is a process with stationary independent increments, that is, whenever*

(a) for all $n \geq 2$, $0 \leq t_0 < t_1 < \cdots < t_n$, the increments $\{N_{t_j} - N_{t_{j-1}}; 1 \leq j \leq n\}$ are mutually independent, and

(b) for all $0 \leq s < t$, the law of $N_t - N_s$ depends upon the pair (s, t) only through the difference $t - s$.

Property (b) is called the 'stationarity of the increments' of $\{N_t\}$.

The term 'Poisson process' is justified by the following proposition.

Proposition 2.2 *Let $\{N_t; t \geq 0\}$ be the counting function of a Poisson process. There exists $\lambda > 0$ such that, for all $0 \leq s < t$, the law of $N_t - N_s$ is the Poisson distribution with parameter $\lambda(t-s)$, that is,*

$$\mathbb{P}(N_t - N_s = k) = e^{-\lambda(t-s)}[\lambda(t-s)]^k/k\,!, \quad k \in \mathbb{N}.$$

Remark 2.3 *The parameter λ is called the intensity of the Poisson process $\{N_t; t \geq 0\}$. It is equal to the mean number of events which happen during a time interval of unit length,*

$$\mathbb{E}[N_{t+1} - N_t] = \lambda.$$

Proof of Proposition 2.2 For all $0 \leq s < t$, consider the generating function of the random variable $N_t - N_s$, which is the mapping $u \to f_{t-s}(u)$ from $[0, 1]$ into itself and is defined by

$$f_{t-s}(u) = \mathbb{E}[u^{N_t - N_s}] = \sum_{k \geq 0} \mathbb{P}(N_t - N_s = k)u^k.$$

From property (a) of Definition 2.1,

$$f_t(u) = f_s(u) f_{t-s}(u), \quad 0 \leq s < t, \ \ u \in [0, 1].$$

It follows from this identity that

$$f_t(u) = [f_1(u)]^t$$

first for t rational, then for all t in $\mathbb{R}_+$ since $t \to f_t(u)$ is decreasing. Since, moreover,

$$\begin{aligned} f_t(u) &\geq P(N_t = 0) \\ &= P(T_1 > t) \\ &\nearrow 1, \quad \text{as } t \downarrow 0, \end{aligned}$$

$f_1(u) \neq 0$, hence there exists $\lambda(u) \in \mathbb{R}_+$ such that

$$f_t(u) = e^{-t\lambda(u)}.$$

Since $u \to \exp(-\theta(1-u))$ is the generating function of the Poisson distribution with parameter θ, it just remains to show that

$$\lambda(u) = \lambda(0)(1-u).$$

But clearly

$$\begin{aligned} \lambda(u) &= \lim_{t \downarrow 0} \frac{1}{t}(1 - f_t(u)) \\ &= \lim_{t \downarrow 0} \sum_{k \geq 1} \frac{1}{t}\mathbb{P}(N_t = k)(1 - u^k). \end{aligned}$$

Since $0 \le u \le 1$,

$$0 \le \sum_{k\ge 2} \frac{1}{t}\mathbb{P}(N_t = k)(1-u^k) \le \frac{1}{t}\mathbb{P}(N_t \ge 2)$$

and the result follows from the identity

$$\lambda(u) = \lim_{t\downarrow 0}\left[\frac{1}{t}\mathbb{P}(N_t = 1)\right](1-u),$$

provided we have

$$\frac{1}{t}\mathbb{P}(N_t \ge 2) \to 0, \quad \text{as } t \downarrow 0. \tag{6.1}$$

But

$$\bigcup_{n\in\mathbb{N}} \{N_{nt} = 0,\ \ N_{(n+1)t} \ge 2\} \subset \{T_2 < T_1 + t\}.$$

Since $\mathbb{P}(N_t = 0) = f_t(0) = \exp(-\lambda(0)t)$, we deduce from this inclusion and property (a) of Definition 2.1 that

$$\sum_{n\in\mathbb{N}} \exp(-\lambda(0)nt)\mathbb{P}(N_t \ge 2) = [1-\exp(-\lambda(0)t)]^{-1}\mathbb{P}(N_t \ge 2)$$
$$\le \mathbb{P}(T_2 < T_1 + t).$$

As $t \downarrow 0$,

$$P(T_2 < T_1 + t) \to P(T_2 \le T_1) = 0,$$

and for all t sufficiently small,

$$(\lambda(0)t)^{-1} < (1-\exp(-\lambda(0)t))^{-1},$$

hence (6.1) is established. □

Remark 2.4 *We can give an intuitive interpretation of the preceding result. From the last part of the above proof,*

$$\mathbb{P}(N_{t+h} - N_t = 0) = 1 - \lambda h + o(h),$$
$$\mathbb{P}(N_{t+h} - N_t = 1) = \lambda h + o(h),$$
$$\mathbb{P}(N_{t+h} - N_t \ge 2) = o(h).$$

Then up to probabilities which are small compared with h, $N(t+h) - N(t)$ is a Bernoulli random variable taking the value 0 *with probability* $1 - \lambda h$ *and the value* 1 *with probability* λh*. This, together with the independence of the increments and the formula*

$$N_{t+s} - N_t = \sum_{j=1}^{n} [N_{t+jh} - N_{t+(j-1)h}], \quad \textit{with } h = \frac{s}{n},$$

implies that $N_{t+s} - N_t$ is approximately a binomial random variable with parameters $(n, \lambda s/n)$. But as $n \to \infty$, that law converges towards Poisson with parameter λs.

Note that, for all $n \geq 2$, $0 < t_1 < t_2 < \ldots < t_n$, the law of the random vector $(N_{t_1}, N_{t_2} \ldots, N_{t_n})$ is determined by Proposition 2.2 and condition (a) in Definition 2.1.

Corollary 2.5 *The law of the time T_1 of the first event is exponential with parameter λ (i.e. the law on $\mathbb{R}_+$ with density $\lambda e^{-\lambda t}$). The same is true for the law of $T_{N_s+1} - s$, which is the waiting time after s of the next event, for all $s > 0$.*

Proof It suffices to note that, for $t > 0$,

$$\mathbb{P}(T_1 > t) = \mathbb{P}(N_t = 0) = e^{-\lambda t}$$

and similarly

$$\mathbb{P}(T_{N_s+1} - s > t) = \mathbb{P}(N_{s+t} - N_s = 0) = P(N_t = 0).$$

□

6.3 The Markov property

Let $\{N_t; t \geq 0\}$ be a Poisson process with intensity λ. For all $s, t > 0$, let

$$N_t^s = N_{s+t} - N_s.$$

It follows from Definition 2.1 that $\{N_t^s; t \geq 0\}$ is a Poisson process with intensity λ, independent of $\{N_r; 0 \leq r \leq s\}$. Note that the knowledge of $\{N_t; 0 \leq t \leq s\}$ is equivalent to that of $(N_s, T_1, T_2, \ldots, T_{N_s})$. The above independence is equivalent to that of the random vectors $(N_s, T_1, T_2, \ldots, T_{N_s})$ and $(T_{N_s+1}, T_{N_s+2}, \ldots, T_{N_s+p})$, for all p.

Since the increments $\{N_{s+t} - N_s; t \geq 0\}$ after s are independent of the past $\{N_t; 0 \leq t \leq s\}$, clearly the future $\{N_{s+t}; t \geq 0\}$ after s depends upon the past $\{N_t; 0 \leq t \leq s\}$ only through the present value N_s; in other words, the past and the future are conditionally independent, given the present. This is the Markov property, and we shall return to it in the next chapter.

We shall now generalize the above property to the case where s is a certain type of random time. Let us first recall some notation, and state a definition.

A σ-algebra of subsets of a set $\mathcal{E}$ is a class of subsets of $\mathcal{E}$ which is closed under complement, countable unions and intersections. One can always speak of the 'smallest σ-algebra containing the class $\mathcal{C} \subset \mathcal{P}(\mathcal{E})$', since it is the intersection of all σ-algebras containing $\mathcal{C}$ (there exists at least one such σ algebra, namely $\mathcal{P}(\mathcal{E})$, and an arbitrary intersection of σ-algebras is a σ-algebra, as can be easily checked). For example, the Borel σ-algebra of subsets of $\mathbb{R}^d$, denoted $\mathcal{B}_d$, is the smallest σ-algebra of subsets of $\mathbb{R}^d$ containing all the open sets.

In the case of a random variable with values in a countable state space E, $\sigma(X) = \{X^{-1}(F);\ F \subset E\}$. Given a d-dimensional random vector X (i.e. an $\mathbb{R}^d$-valued random variable), we denote by $\sigma(X) = \{X^{-1}(B);\ B \in \mathcal{B}_d\}$ the smallest σ-algebra of subsets of Ω which makes X measurable. This is the set of events for which we know whether or not they are realized provided that we know the value of X. Given an arbitrary collection $\{X_i;\ i \in I\}$ (of arbitrary dimensions), we denote by $\sigma\{X_i;\ i \in I\}$ the smallest σ-algebra containing $\sigma(X_i)$, for all $i \in I$.

It will be convenient in this chapter to use the following notation: for any $t \geq 0$,

$$\mathcal{F}_t^N = \sigma\{N_s;\ 0 \leq s \leq t\} = \sigma\{N_t, T_1, T_2, \ldots, T_{N_t}\}.$$

Definition 3.1 *Given a Poisson process $\{N_t;\ t \geq 0\}$, a stopping time (associated with $\{N_t\}$) is a random variable S taking values in $\mathbb{R}_+ \cup \{+\infty\}$ such that, for all t in $\mathbb{R}_+$,*

$$\{S \leq t\} \in \mathcal{F}_t^N.$$

For all s in $\mathbb{R}_+$, $S \equiv s$ is a stopping time. For all n, T_n is a stopping time. T_{N_s+1} is also a stopping time. But T_{N_s} is not a stopping time, since whenever $t < s$,

$$\{T_{N_s} \leq t\} = \{N_s - N_t = 0\} \notin \mathcal{F}_t^N, \quad 0 \leq t < s.$$

With any stopping time S associated with $\{N_t\}$, we associate the σ-algebra of those events which are 'determined by the trajectory $\{N_{t\wedge S};\ t \geq 0\}$ stopped at time S':

$$\mathcal{F}_S^N \stackrel{\text{def}}{=} \{A \in \mathcal{F}_\infty^N;\ A \cap \{S \leq t\} \in \mathcal{F}_t^N,\ \forall t \geq 0\}.$$

We have the following proposition:

Proposition 3.2 *Let $\{N_t;\ t \geq 0\}$ be a Poisson process with intensity λ, and S a stopping time associated with $\{N_t\}$. On the event $\{S < \infty\}$ we define, for $t \geq 0$,*

$$N_t^S = N_{S+t} - N_S.$$

Conditionally upon $\{S < \infty\}$, $\{N_t^S;\ t \geq 0\}$ is a Poisson process with intensity λ, independent of $\mathcal{F}_S^N$.

PROOF We already know that the result holds if S is constant. Suppose next that S takes its values in an increasing sequence $(s_j,\ j \geq 1)$ of positive real numbers. Note that, since S is a stopping time,

$$\{S = s_j\} = \{S \leq s_j\} \backslash \{S \leq s_{j-1}\} \in \mathcal{F}_{s_j}^N.$$

Let $A \in \mathcal{F}_S^N$, $0 < t_1 < t_2 < \ldots < t_\ell$ and $n_1, \ldots, n_\ell$ belong to $\mathbb{N}$. We have

$$\begin{aligned}
&\mathbb{P}\left(A \cap \left(\bigcap_{k=1}^{\ell} \{N_{t_k}^S = n_k\}\right)\right) \\
&= \sum_j \mathbb{P}\left(\{S = s_j\} \cap A \cap \left(\bigcap_{k=1}^{\ell} \{N_{s_j+t_k} - N_{s_j} = n_k\}\right)\right) \\
&= \sum_j \mathbb{P}(\{S = s_j\} \cap A)\mathbb{P}\left(\bigcap_{k=1}^{\ell} \{N_{s_j+t_k} - N_{s_j} = n_k\}\right) \\
&= \mathbb{P}(A)\mathbb{P}\left(\bigcap_{k=1}^{\ell} \{N_{t_k} = n_k\}\right),
\end{aligned}$$

where we have used the property $\{S = s_j\} \cap A \in \mathcal{F}_{s_j}^N$ for the second equality, and the fact that the second factor of the penultimate expression does not depend upon s_j, by the stationarity of the increments of $\{N_t\}$.

The result is thus established in the case of a stopping time which takes its values in an increasing sequence. But any stopping time S can be approximated by a decreasing sequence of stopping times of this form. Indeed, for all n, define

$$S_n = \sum_{k \in \mathbb{N}} k2^{-n} \mathbf{1}_{\{(k-1)2^{-n} < S \leq k2^{-n}\}}.$$

The above identity is true with S replaced by S_n, since

$$S \leq S_n \Rightarrow \mathcal{F}_S^N \subset \mathcal{F}_{S_n}^N.$$

We can now easily take the limit in the above identity, with S replaced by S_n, since from the right continuity of the trajectories of $\{N_t; t \geq 0\}$,

$$\mathbb{P}\left(A \cap \left(\bigcap_{k=1}^{\ell} \{N_{t_k}^{S_n} = n_k\}\right)\right) \to \mathbb{P}\left(A \cap \left(\bigcap_{k=1}^{\ell} \{N_{t_k}^S = n_k\}\right)\right).$$

□

Corollary 3.3 *Let $\{N_t; t \geq 0\}$ be a Poisson process with intensity λ, and $(T_n)_{n \geq 1}$ its jump times. We let $S_1 = T_1$, $S_2 = T_2 - T_1, \ldots, S_n = T_n - T_{n-1}, \ldots$. The random variables $S_1, S_2, \ldots, S_n, \ldots$ are i.i.d., their common law being exponential with the parameter λ.*

Proof We already know that the law of T_1, the first jump time of a Poisson process with intensity λ, is exponential with parameter λ. It follows from Proposition 3.2 with $S = T_n$ that $S_{n+1} = T_{n+1} - T_n$ is the first jump time of a Poisson process with

intensity λ, hence its law is exponential with parameter λ, and is independent of $T_1, T_2, \ldots, T_n$, hence also of $S_1, S_2, \ldots, S_n$. The result follows from the fact that this statement is true for all $n \geq 1$. □

In the other direction, we have the following result:

Proposition 3.4 *Let $\{S_n; n \geq 1\}$ be a sequence of i.i.d. random variables, their common law being the exponential law with parameter $\lambda > 0$. We define*

$$T_n = S_1 + \ldots + S_n, \quad n \geq 1,$$

$$N_t = \sup\{n; T_n \leq t\}, \quad t \geq 0.$$

Then $\{N_t; t \geq 0\}$ is a Poisson process with intensity λ.

We then have a way to 'construct' a Poisson process, which in particular shows that there does exist a process satisfying the requirements of Definition 2.1! We also have here a way to *simulate* a Poisson process.

6.4 Large time behaviour

Again, let $\{N_t; t \geq 0\}$ be a Poisson process with intensity λ. Then

$$\mathbb{E}[N_t] = \lambda t, \quad \text{var}[N_t] = \lambda t.$$

In particular $\mathbb{E}[N_t/t] = \lambda$, $\text{var}[N_t/t] = \lambda/t$, hence $N(t)/t \to \lambda$ in mean square, as $t \to \infty$. In fact we have the 'strong law of large numbers':

Proposition 4.1 *Let $\{N_t; t \geq 0\}$ be a Poisson process with intensity $\lambda > 0$. Then $N_t/t \to \lambda$ almost surely as $t \to \infty$.*

PROOF First, note that

$$N_n = \sum_{1 \leq i \leq n} [N_i - N_{i-1}]$$

is the sum of n i.i.d. random variables, their common law being Poisson with parameter λ (hence, they are integrable). It then follows from the strong law of large numbers that

$$\frac{N_n}{n} \to \lambda, \quad \text{a.s., as } n \to \infty.$$

But with $[t]$ denoting the integer part of t,

$$\frac{N_t}{t} = \frac{N_{[t]}}{[t]} \times \frac{[t]}{t} + \frac{N_t - N_{[t]}}{t}.$$

It then suffices to show that

$$\sup_{n<t<n+1} \frac{N_t - N_n}{n} \to 0, \quad \text{as } n \to \infty.$$

Let

$$\xi_n \stackrel{\text{def}}{=} \sup_{n<t\leq n+1} N_t - N_n = N_{n+1} - N_n.$$

The $\{\xi_n\}$ are i.i.d. and integrable. Then $(\xi_1 + \ldots + \xi_n)/n \to \lambda$ almost surely, consequently

$$\frac{\xi_n}{n} \to 0 \text{ a.s.}$$

□

We have the following 'central limit theorem':

Proposition 4.2 *Let $\{N_t; t \geq 0\}$ be a Poisson process with intensity λ. Then*

$$\frac{N_t - \lambda t}{\sqrt{\lambda t}} \to Z \text{ in law}, \quad \text{as } t \to \infty,$$

where Z is a centred Gaussian random variable with unit variance.

PROOF We essentially argue as in the preceding proof:

$$\frac{N_n - \lambda n}{\sqrt{\lambda n}} \to Z \text{ in law}, \quad \text{as } n \to \infty,$$

from the 'classical' central limit theorem. And

$$\frac{N_t - N_{[t]}}{\sqrt{\lambda [t]}} \leq \frac{\xi_{[t]}}{\sqrt{\lambda [t]}}$$

converges to 0 in probability as $t \to \infty$ since

$$\begin{aligned} P\left(\xi_n/\sqrt{\lambda n} > \varepsilon\right) &= P\left(\xi_n > \varepsilon\sqrt{\lambda n}\right) \\ &= P\left(\xi_1 > \varepsilon\sqrt{\lambda n}\right) \\ &\to 0, \quad \text{as } n \to \infty. \end{aligned}$$

Then also $(N_t - N_{[t]})/\sqrt{\lambda [t]} \to 0$ in probability as $t \to \infty$. Finally,

$$\frac{N_t - \lambda t}{\sqrt{\lambda t}} = \frac{N_{[t]} - \lambda [t]}{\sqrt{\lambda [t]}} \sqrt{\frac{[t]}{t}} + \frac{N_t - N_{[t]}}{\sqrt{\lambda t}} \sqrt{\frac{[t]}{t}} + \sqrt{\lambda} \frac{[t] - t}{\sqrt{t}},$$

and we know that whenever $X_n \to X$ in law, $Y_n \to 0$ in probability, then

$$X_n + Y_n \to X \text{ in law}.$$

□

One could in fact establish a 'functional central limit theorem' which we now briefly describe. A proof very similar to that of Proposition 4.2 shows that, for all $t > 0$,

$$\frac{N_{tu} - \lambda t u}{\sqrt{\lambda u}} \to B_t \text{ in law}, \quad \text{as } u \to \infty,$$

where B_t is a centred Gaussian random variable with variance t. Note that, for each $u > 0$, $\{[N_{tu} - \lambda tu]/\sqrt{\lambda u}; t \geq 0\}$ is a process with independent increments, whose jumps are of size $(\lambda u)^{-1/2}$. It follows that one can take the above limit as $u \to \infty$ jointly for the various values of t, in such a way that the limit $\{B_t; t \geq 0\}$ is a centred Gaussian process with independent increments and continuous trajectories, satisfying $E[B_t^2] = t$. $\{B_t; t \geq 0\}$ is called a *Brownian motion*, and will be discussed in Chapter 9.

6.5 Exercises

Exercise 5.1 *Let X be an $\mathbb{R}_+$-valued random variable such that $\mathbb{P}(X > t) > 0$, for all $t > 0$. We assume, moreover, that for all $s, t > 0$,*

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s).$$

Conclude that the law of X is exponential with some parameter $\lambda > 0$.

Exercise 5.2 *Three persons, A, B and C, get to a post office at the same time. They each want to make a telephone call. There are two telephone booths, which are immediately occupied by A and B. C makes her call after whoever finishes first. They leave the post office as soon as they have completed their calls.*

We denote by X, Y and Z the length of the telephone calls made by A, B and C, respectively. These three random variables are assumed to be i.i.d., their common law being exponential with parameter $\lambda > 0$.

1. *Compute the probability that C leaves last.*
2. *Give the probability distribution of the total time T spent by C in the post office.*
3. *With 0 being the time of arrival of the three persons at the post office, give the probability distribution of the time of the last departure.*

(Hint: first give the probability distribution of the random vector $(X \wedge Y, X \vee Y - X \wedge Y)$, where $\wedge = \inf$ and $\vee = \sup$.)

Exercise 5.3 *A machine has a lifespan τ_1 whose law is exponential with parameter θ. As soon as it breaks down, it is instantly replaced by an identical machine with lifespan τ_2, and so on. We assume that the random variables $(\tau_n; n \in \mathbb{N})$ are i.i.d. The first machine starts running at time 0; the successive machine failure times T_n ($n \geq 1$ – in other words, $T_1 = \tau_1$, $T_2 = \tau_1 + \tau_2, \ldots$) constitute a Poisson point process.*

1. *Given $t > 0$, let D_t denote the elapsed time since the machine running at time t started to run. In which set does the random variable D_t take its values? What is the law of D_t? Show that as $t \to \infty$, this law has a limit.*

2. *Let S_t be such that $t + S_t$ is the failure time of the machine running at time t. What is the law of S_t? What is the law of the pair (D_t, S_t) and what is the limit of that law as $t \to \infty$? Why are D_t and S_t not identically distributed, and why do they tend to be identically distributed as $t \to \infty$?*

3. *What is the law of $D_t + S_t$, the lifespan of the machine which is running at time t? Compare the limit of that law as $t \to \infty$ with the joint law of the τ_n. Explain the apparent contradiction.*

Exercise 5.4 1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, whose common law is the uniform law on $[0, t]$, and $Y_1, Y_2, \ldots, Y_n$ be the same sequence, but in increasing order. In other words, the Y_k are defined by*

$$Y_1 = \inf_{1 \le i \le n} X_i = X_{i_1},$$

$$Y_2 = \inf_{1 \le i \le n, i \ne i_1} X_i,$$

and so on. Give the probability distribution of the random vector $(Y_1, Y_2, \ldots, Y_n)$.

2. *Let $\{N_t; t \ge 0\}$ be a Poisson process with intensity λ. Show that the conditional law of the random vector $(T_1, T_2, \ldots, T_n)$, given that $N_t = n$, is the distribution which was identified in part 1.*

Exercise 5.5 *Let $\{N_t^1;\ t \ge 0\}$ and $\{N_t^2;\ t \ge 0\}$ be two independent Poisson processes of intensity λ_1 and λ_2, respectively. Show that $\{N_t^1 + N_t^2;\ t \ge 0\}$ is a Poisson process with intensity $\lambda_1 + \lambda_2$.*

Exercise 5.6 *Suppose that the number of individuals infected by HIV follows a Poisson process of a given intensity λ. We denote by N_t the number of individuals infected at time t. We do not take deaths into account.*

Each infected individual has an incubation period between the time of infection and the time when the symptoms of AIDS appear. The length of this incubation period is random. The incubation periods for various individuals are i.i.d., their common law on $\mathbb{R}_+$ having a given distribution function G. We denote by $\bar{G}$ the function $\bar{G}(t) = 1 - G(t)$.

Let N_t^1 denote the number of individuals who have AIDS symptoms at time t, and let N_t^2 denote the number of individuals who at time t are infected by HIV but do not yet have AIDS symptoms. Of course

$$N_t = N_t^1 + N_t^2.$$

Show that, for each $t > 0$, N_t^1 and N_t^2 are independent, the probability distribution of N_t^1 is Poisson with parameter $\lambda \int_0^t G(s)ds$, and that of N_t^2 Poisson with parameter $\lambda \int_0^t \bar{G}(s)ds$. You can make use of the result from Exercises 5.4, which says that conditionally upon $N_t = n$, the infection times between 0 and t which are counted by a Poisson process have the same law as an i.i.d. sequence of n uniform $[0, t]$ random variables.

Exercise 5.7 (Programming) *Define the failure rate of an $\mathbb{R}_+$-valued random variable X with density f and distribution function F to be the function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ defined by*

$$\lambda(t) = \frac{f(t)}{1 - F(t)}.$$

Exercises 5.1 proves that the only probability distribution with a constant failure rate is the exponential.

The Weibull distribution with parameters $\lambda, \alpha > 0$ is the distribution on $\mathbb{R}_+$ with survival function

$$\bar{F}(t) = 1 - F(t) = e^{-(\lambda t)^\alpha}$$

and failure rate

$$\lambda(t) = \alpha\lambda(\lambda t)^{\alpha-1}.$$

The Weibull distribution has an increasing failure rate if $\alpha > 1$ and a decreasing rate if $\alpha < 1$. It reduces to the exponential distribution with parameter λ when $\alpha = 1$.

The gamma distribution $\Gamma(\alpha, \lambda)$ is the distribution on $\mathbb{R}_+$ with density

$$f(t) = \frac{\lambda}{\Gamma(\alpha)} e^{-\lambda t} (\lambda t)^{\alpha-1},$$

where $\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt$. Again the gamma distribution has an increasing failure rate if $\alpha > 1$ and a decreasing rate if $\alpha < 1$. Note that the sum of n i.i.d. exponential random variables with parameter λ follows the $\Gamma(n, \lambda)$ distribution with an increasing failure rate.

Suppose that two machines function in parallel, and need a fragile part M. Suppose that we have only one spare part, which immediately replaces whichever breaks down first. The three parts M (the two in place at the beginning, plus the spare) have i.i.d. lifetimes. The second failure is fatal for the machine that suffers it. If the lifetimes are exponential, then Exercises 5.2 shows that the two machines have the same probability of suffering the fatal failure.

Suppose that we replace the exponential law by a law with an increasing failure rate. Then the machine whose part has already been replaced has a better chance of functioning longer than its counterpart, and the reverse is true in the case of a decreasing failure rate.

Illustrate by a Monte Carlo computation the result from Exercises 5.2, and the two conjectures which we have just formulated. More precisely, with $\mathbb{P}$ given successively by the exponential distribution with parameter 1, the $\Gamma(3, 1)$ distribution and the Weibull $(1, 0.5)$ distribution (easy to simulate by inversion of its distribution function), simulate a matrix $3 \times N$ of i.i.d. random variables with the law $\mathbb{P}$, denoted by X. Graph, for n from 1 to N, the three quantities

$$n^{-1} \sum_{k=1}^{n} \{\min[X(1,k), X(2,k)] + X(3,k) - \max[X(1,k), X(2,k)]\}.$$

You can choose $N = 10^3$ or $N = 10^4$.

7

Jump Markov processes

Introduction

In this chapter, we present the theory of continuous time jump Markov processes, with values in a finite or countable state space E. As we will see in Section 7.4, these processes are in a sense combinations of a Poisson process and a discrete time Markov chain (the 'embedded chain'). Sections 7.8–7.10 are devoted to applications to phylogeny, to discretized partial differential equations, and to the annealing algorithm. The proof of convergence of the annealed algorithm which we present here is due to Francis Comets (private communication). Applications to queues will be developed in the next chapter.

7.1 General facts

We wish to study continuous time Markov processes with values in a finite or countable state space E. We will assume that their trajectories are constant in between jumps, and that the latter are isolated. Moreover, we will assume that the trajectories are right continuous. They will have left limits at each point. The jumps of such a process $\{X_t;\ t \geq 0\}$ happen at random times $T_1(\omega), T_2(\omega), \ldots, T_n(\omega), \ldots$. The main difference with the Poisson process of the preceding chapter is that, given the time of the jump and the position before the jump, the position after the jump is random. If we denote by $Z_n(\omega)$ the value of $\{X_t\}$ just after the nth jump $T_n(\omega),\ n \geq 1$, a typical trajectory of the process $\{X_t;\ t \geq 0\}$ is shown in Figure 7.1. The knowledge of $\{X_t;\ t \geq 0\}$ is equivalent to that of the double sequence $\{T_n, Z_n;\ n \geq 0\}$.

For certain applications, it is convenient to make certain states absorbing (for instance, in a model describing the evolution of the size of a population without immigration, 0 is an absorbing state). $x \in E$ is absorbing if $X_{T_n}(\omega) = x \Rightarrow T_{n+1}(\omega) = +\infty$.

We will then assume that the jump times constitute an increasing sequence

$$0 = T_0 < T_1 \leq T_2 \leq \ldots \leq T_n \leq \ldots \tag{7.1}$$

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

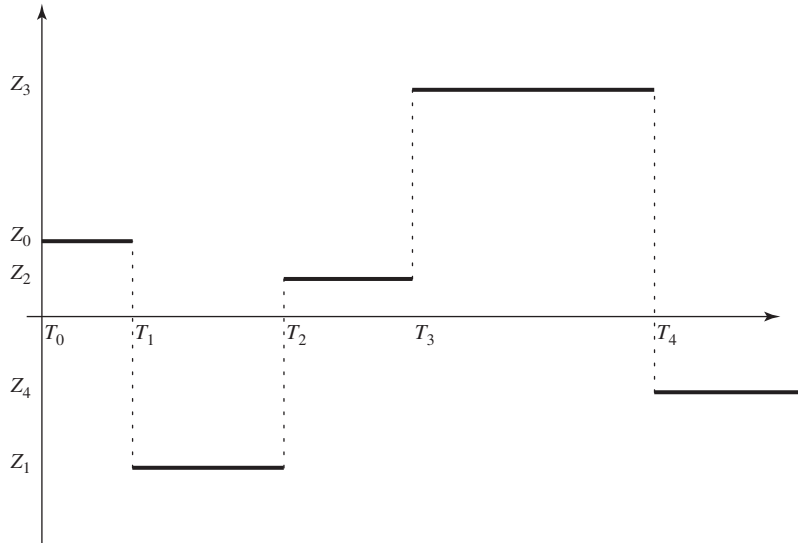


Figure 7.1 Trajectory of a continuous time jump Markov process.

with $T_n \in \mathbb{R}_+ \cup \{+\infty\}$ and

$$T_n(\omega) < T_{n+1}(\omega) \quad \text{if } T_n(\omega) < \infty. \tag{7.2}$$

We assume, moreover, that there is no explosion, that is, jump times do not accumulate at finite distance; in other words,

$$T_n(\omega) \to +\infty \text{ a.s.}, \quad \text{as } n \to \infty. \tag{7.3}$$

In what follows, (7.1)–(7.3) will implicitly be assumed to hold.

An E-valued random function $\{X_t;\ t \geq 0\}$ is called a random jump function if it is of the form

$$X_t(\omega) = \sum_{\{n \geq 0;\ T_n(\omega) < \infty\}} Z_n(\omega) \mathbf{1}_{[T_n(\omega), T_{n+1}(\omega)[}(t)$$

where the random variables Z_n take their values in E.

Definition 1.1 *An E-valued random jump function $\{X_t;\ t \geq 0\}$ is called a jump Markov process (or a continuous time Markov chain) if, for all $0 < s < t$, the conditional law of the random variable X_t given $\{X_u;\ 0 \leq u \leq s\}$ depends upon X_s only, that is, for all $n \in \mathbb{N}$, $0 \leq t_0 < t_1 < \ldots < t_n < s$, $x_0, x_1, \ldots, x_n,\ x, y \in E$, Note that this condition makes sense only when*

$$\mathbb{P}(X(t_0) = x_0,\ X(t_1) = x_1, \ldots, X(t_n) = x_n,\ X(s) = x) > 0.$$

In this condition we shall disregard the values $n, x_0, x_1, \ldots, x_n, x$ for which that inequality does not hold.

$$\mathbb{P}(X_t = y, X_{t_0} = x_0, X_{t_1} = x_1, \ldots, X_{t_n} = x_n,\ X_s = x) = \mathbb{P}(X_t = y | X_s = x).$$

We shall say that the jump Markov process $\{X_t;\ t \geq 0\}$ is homogeneous if the quantity $P(X_t = y | X_s = x)$ depends upon s and t only through the difference $t - s$.

We restrict ourselves to the study of homogeneous Markov processes. For $s < t$ let us write

$$\mathbb{P}(X_t = y | X_s = x) = P_{xy}(t - s),$$

where, for all $t > 0$, $P(t)$ is a 'Markovian matrix' on $E \times E$, which is called the transition matrix from time 0 to time t. We shall denote below by $\mu(t)$ the probability distribution of X_t on E, $t \geq 0$; $\mu(0)$ is called the 'initial law' of the process $\{X_t;\ t \geq 0\}$.

Proposition 1.2 *Let $\{X_t;\ t \geq 0\}$ be a jump Markov process, with initial law μ and transition matrices $\{P(t);\ t > 0\}$. For all $n \in \mathbb{N}$, $0 < t_1 < \ldots < t_n$, the law of the random vector $(X_0, X_{t_1}, \ldots, X_{t_n})$ is given, for all $x_0, x_1, \ldots, x_n \in E$, by*

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n)\\ &= \mu_{x_0} P_{x_0 x_1}(t_1) P_{x_1 x_2}(t_2 - t_1) \cdots P_{x_{n-1} x_n}(t_n - t_{n-1}).\end{aligned}$$

Consequently, for all $t > 0$,

$$\mu(t) = \mu(0) P(t)$$

in the sense that $\mu_y(t) = \sum_{x \in E} \mu_x(0) P_{xy}(t)$, and for any positive or bounded function $g : E \to \mathbb{R}$,

$$\mathbb{E}[g(X_t) | X_0 = x] = (P(t)g)_x = \sum_{y \in E} P_{xy}(t) g_y.$$

Moreover, the transition matrices $\{P(t);\ t > 0\}$ satisfy the semigroup property (also called the Chapmann–Kolmogorov equation)

$$P(s + t) = P(s) P(t)$$

in the sense that, for all x, y in E,

$$P_{xy}(t + s) = \sum_{z \in E} P_{xz}(t) P_{zy}(s).$$

Proof It follows from the definition of conditional probability and the Markov property that

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, X_{t_1} = x_1, \ldots, X_{t_n} = x_n)\\ =&\mathbb{P}(X_0 = x_0) P(X_{t_1} = x_1 | X_0 = x_0) \mathbb{P}(X_{t_2} = x_2 | X_0 = x_0, X_{t_1} = x_1)\\ &\times \cdots \times \mathbb{P}(X_{t_n} = x_n | X_0 = x_0, X_{t_1} = x_1, \ldots, X_{t_{n-1}} = x_{n-1})\\ =&\mu_{x_0} P_{x_0 x_1}(t_1) P_{x_1 x_2}(t_2 - t_1) \cdots P_{x_{n-1} x_n}(t_n - t_{n-1}).\end{aligned}$$

In the case $n = 1$, this formula reduces to

$$\mathbb{P}(X_0 = x,\ X_t = y) = \mu_x P_{xy}(t),$$

and the second result follows by summing over $x \in E$. From the definition of $P(t)$,

$$\mathbb{P}(X_t = y | X_0 = x) = P_{xy}(t),$$

the third result follows by multiplying by g_y and summing over $y \in E$.

Finally, the above formula in the case $n = 2$ gives, after division by μ_{x_0},

$$\mathbb{P}(X_s = z,\ X_{s+t} = y | X_0 = x) = P_{xz}(s) P_{zy}(t).$$

The last result follows by summing over $z \in E$. □

We now present some examples of jump Markov processes.

Example 1.3 *A Poisson processes $\{N_t;\ t \geq 0\}$ of intensity λ is an $\mathbb{N}$-valued Markov process, with transition matrix*

$$P_{xy}(t) = \begin{cases} e^{-\lambda t}(\lambda t)^{y-x}/(y-x)!, & \text{if } y \geq x, \\ 0, & \text{otherwise.} \end{cases}$$

Example 1.4 (The telegraph process) *Given a Poisson process $\{N_t\}$ of intensity λ, and an $E = \{-1, +1\}$-valued random variable X_0, independent of $\{N_t;\ t \geq 0\}$, we define*

$$X_t = X_0 (-1)^{N_t}, \quad t \geq 0.$$

$\{X_t;\ t \geq 0\}$ is a Markov process with transition matrix

$$P_{+1+1}(t) = P_{-1-1}(t) = e^{-\lambda t} \sum_{n \geq 0} \frac{(\lambda t)^{2n}}{(2n)!},$$

$$P_{-1+1}(t) = P_{+1-1}(t) = e^{-\lambda t} \sum_{n \geq 0} \frac{(\lambda t)^{2n+1}}{(2n+1)!}.$$

Example 1.5 *Let $\{N_t;\ t \geq 0\}$ be a Poisson process with intensity λ, and jump times $0 < T_1 < T_2 < T_3 < \ldots < T_n < \ldots$. Also let $\{Z_n;\ n \in \mathbb{N}\}$ be an E-valued discrete time Markov chain, with transition matrix $\{P_{xy};\ x, y \in E\}$, independent of $\{N_t;\ t \geq 0\}$. One can show (see Exercises 11.1) that*

$$X_t = \sum_{n=0}^{\infty} Z_n \mathbf{1}_{[T_n, T_{n+1}[}(t), \quad t \geq 0,$$

is a jump Markov process.

7.2 Infinitesimal generator

It follows from the semigroup property that $P(t)$ is known for all $t > 0$ if it is known for all small enough t. In fact, we will see that it is completely determined by its right derivative at $t = 0$ (we know that $P(0) = I$).

Theorem 2.1 *Let $\{P(t);\ t > 0\}$ be the semigroup of transition matrices of a jump Markov process $\{X_t;\ t \geq 0\}$. There exists a matrix $\{Q_{xy}; x, y \in E\}$ (called the infinitesimal generator of the semigroup $\{P(t);\ t \geq 0\}$ or of the Markov process $\{X_t;\ t \geq 0\}$) which satisfies*

$$Q_{xy} \geq 0, \quad \text{if } x \neq y,$$
$$Q_{xx} = -\sum_{y \in E \setminus \{x\}} Q_{xy} \leq 0$$

(this last inequality being strict, unless the state x is absorbing) and such that, as $h \downarrow 0$,

$$P_{xy}(h) = hQ_{xy} + o(h), \quad \text{if } x \neq y,$$
$$P_{xx}(h) = 1 + hQ_{xx} + o(h).$$

Moreover, conditioned upon $X_0 = x$, the time T_1 of the first jump and the position $Z_1 = X_{T_1}$ after that jump are independent, the law of T_1 being exponential with parameter $q_x = -Q_{xx}$, and the law of Z_1 on E being given by $\{Q_{xy}/q_x; y \neq x\}$.

Proof First, note that

$$\{T_1 > nh\} \subset \{X_0 = X_h = \ldots = X_{nh}\} \subset \{T_1 > nh\} \cup \{T_2 - T_1 \leq h\}.$$

Since $P(T_2 - T_1 \leq h) \to 0$ as $h \to 0$, we have that as $h \to 0$, $nh \to t$ (with $nh \geq t$),

$$\mathbb{P}(T_1 > t | X_0 = x) = \lim \mathbb{P}(X_0 = X_h = \ldots = X_{nh} | X_0 = x)$$
$$= \lim [P_{xx}(h)]^n.$$

Existence of this limit implies that

$$\frac{1}{h}[1 - P_{xx}(h)] \to q_x \in [0, +\infty],$$

as $h \to 0$, and consequently

$$\mathbb{P}(T_1 > t | X_0 = x) = e^{-q_x t}.$$

Hence necessarily $q_x < \infty$ and $q_x = 0$ if and only if x is absorbing. Define $Q_{xx} = -q_x$.

The proof of existence of the limits of $h^{-1}P_{xy}(h)$ for $x \neq y$ is done similarly: we have

$$\{T_1 \leq t,\ Z_0 = x,\ Z_1 = y\}$$
$$= \lim_{h\to 0, nh\to t} \bigcup_{1\leq m\leq n} \{X_0 = X_h = \ldots = X_{(m-1)h} = x, X_{mh} = y\}$$

and

$$\mathbb{P}(T_1 \leq t,\ Z_1 = y | X_0 = x) = \lim \frac{1 - P_{xx}(h)^n}{1 - P_{xx}(h)} P_{xy}(h)$$
$$= \frac{1 - e^{-q_x t}}{q_x} \lim \frac{1}{h} P_{xy}(h).$$

Hence, $Q_{xy} = \lim h^{-1}P_{xy}(h)$ exists for $x \neq y$ and

$$\mathbb{P}(T_1 \leq t,\ Z_1 = y | X_0 = x) = (1 - e^{-q_x t})\frac{Q_{xy}}{q_x},$$

whence

$$\mathbb{P}(T_1 \leq t,\ Z_1 = y | X_0 = x) = \mathbb{P}(T_1 \leq t | X_0 = x)\mathbb{P}(Z_1 = y | X_0 = x)$$

and

$$\mathbb{P}(Z_1 = y | X_0 = x) = \frac{Q_{xy}}{q_x}.$$

□

In the case where E is a finite set, we immediately deduce from the theorem the following corollary.

Corollary 2.2 *(i)* $\{P(t);\ t \geq 0\}$ *is the unique solution of Kolmogorov's backward equation*

$$\frac{dP}{dt}(t) = QP(t),\ \ t > 0, \quad P(0) = I.$$

Moreover, $u(t, x) := \mathbb{E}[g(X_t)|X_0 = x]$ *also solves a Kolmogorov backward equation*

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, x) = \displaystyle\sum_{y\in E} Q_{xy} u(t, y), \quad t > 0,\ x \in E, \\ \\ u(0, x) = g(x), \quad x \in E. \end{cases}$$

(ii) $\{P(t); t \geq 0\}$ *is also the unique solution of the forward Kolmogorov equation*

$$\frac{dP(t)}{dt} = P(t)Q,\ t > 0, \quad P(0) = I.$$

Moreover, the family of marginal probability distributions $\{\mu(t);\ t \geq 0\}$ *of the random variables* $\{X_t; t \geq 0\}$ *satisfies the Fokker–Planck equation*

$$\frac{\partial \mu_x(t)}{\partial t} = \sum_{y \in E} \mu_y(t) Q_{yx}, \quad t > 0,\ x \in E.$$

PROOF The Kolmogorov backward equation follows by differentiating $P_{xy}(t)$, exploiting the semigroup property in the form

$$P(t+h) = P(h)P(t).$$

The equation for u then follows from the equation just obtained by multiplying it on the right by the column vector $\{g_x\}$.

The forward equation is obtained by differentiating, starting with the identity

$$P(t+h) = P(t)P(h).$$

The Fokker–Planck equation then follows by multiplying on the left by the row vector $\{\mu(0)\}$. □

Remark 2.3 *Let us explain the terms 'forward equation' and 'backward equation'. The backward equation is an equation for the function* $(t, x) \to P_{xy}(t)$*, where* $y \in E$ *is fixed. The variables are* t *and the 'backward' variable* x*:* x *is the position at the initial time – the position in the past. In contrast, the forward equation is an equation for the function* $(t, y) \to P_{xy}(t)$*, with* $x \in E$ *fixed. The variable* y *denotes the position of the process at time* t *– the position at the present time.*

Consider now the backward equation for the quantity $u(t, x) = \mathbb{E}[g(X_t)|X_0 = x]$*. Fix* $T > 0$ *and define, for* $0 \leq t \leq T$*,* $v(t, x) = u(T - t, x) = \mathbb{E}[g(X_T)|X_t = x]$*.* v *satisfies the equation*

$$\begin{cases} \dfrac{\partial v}{\partial t}(t, x) + \displaystyle\sum_{y \in E} Q_{xy} u(t, y) = 0, \quad t > 0,\ x \in E, \\ \\ v(T, x) = g(x), \quad x \in E. \end{cases}$$

The equation for v *is a backward equation in the sense that it is solved in the backward direction of time, from* $t = T$ *to* $t = 0$*. Note that in the non-homogeneous case, where the infinitesimal generator* Q *depends upon* t*, the quantity* $v(t, x) = \mathbb{E}[g(X_T)|X_t = x]$ *solves that same equation, while we no longer have an equation for* u*.*

The proof of the corollary is not rigorous in the case where E is countable, since it implies interchanging a derivation and an infinite sum. The backward Kolmogorov equation will be established in the general case in the next section.

7.3 The strong Markov property

The notion of a stopping time S and the associated σ-field $\mathcal{F}_S^X$ are defined as in Section 6.3, replacing $\{N_t; t \geq 0\}$ by $\{X_t; t \geq 0\}$.

Theorem 3.1 *Let S be a stopping time of the jump Markov process $\{X_t;\ t \geq 0\}$. Conditionally upon $\{S < \infty\}$ and $\{X_S = x\}$, $\{X_{S+t};\ t \geq 0\}$ is independent of $\mathcal{F}_S^X$, and its law is that of $\{X_t;\ t \geq 0\}$ given that $X_0 = x$.*

Proof It suffices to prove the theorem in the case of a constant stopping time $S \equiv s$. The general case then follows as for the Poisson process (see the proof of Proposition 3.2 of Chaptet 6). Let $0 \leq s_1 < s_2 < \ldots < s_k < s;\ \ 0 < t_1 < t_2 < \ldots < t_\ell;\ x, x_1, \ldots, x_z, y_1, \ldots, y_\ell \in S$. We have

$$\begin{aligned}
&\mathbb{P}(X_{s+t_1} = y_1, \ldots, X_{s+t_\ell} = y_\ell | X_{s_1} = x_1, \ldots, X_{s_k} = x_k, X_s = x) \\
&= \frac{\mathbb{P}(X_{s_1} = x_1, \ldots, X_{s_k} = x_k, X_s = x, X_{s+t_1} = y_1, \ldots, X_{s+t_\ell} = y_\ell)}{\mathbb{P}(X_{s_1} = x_1, \ldots, X_{s_k} = x_k, X_s = x)} \\
&= P_{xy_1}(t_1) P_{y_1 y_2}(t_2 - t_1) \cdots P_{y_{\ell-1} y_\ell}(t_\ell - t_{\ell-1}) \\
&= \mathbb{P}(X_{t_1} = y_1, \ldots, X_{t_\ell} = y_\ell | X_0 = x).
\end{aligned}$$

□

We now establish the backward Kolmogorov equation in the general case.

Theorem 3.2 *For all $x, y \in E$, the function $t \to P_{xy}(t)$ is differentiable and*

$$\frac{d}{dt} P_{xy}(t) = (QP)_{xy}(t).$$

Proof Define for all $n \in \mathbb{N}$ the conditional law of (Z_n, T_n), given that $X_0 = Z_0 = x$:

$$R_n(x; y, B) = \mathbb{P}(Z_n = y, T_n \in B | Z_0 = x), \quad B \text{ Borel subset of } \mathbb{R}_+.$$

Note that

$$R_0(x; y, B) = \begin{cases} 1, & \text{if } x = y,\ 0 \in B, \\ 0, & \text{otherwise.} \end{cases}$$

and it follows from Theorem 2.1 that

$$R_1(x; y, B) = \begin{cases} Q_{xy} \displaystyle\int_B e^{-q_x t} dt, & \text{if } x \neq y; \\ 0, & \text{if } x = y. \end{cases}$$

The strong Markov property at time T_m implies that

$$\mathbb{P}(Z_{m+n} = z, T_{m+n} \in B | \mathcal{F}_{T_m}^X) = R_n(X_{T_m};\ z, B - T_m),$$

where we have used the notation

$$B - t = \{s \in \mathbb{R}_+;\ s + t \in B\}.$$

Hence,

$$\begin{aligned}\mathbb{P}(Z_{m+n} = z, T_{m+n} \in B | X_0 = x) &= \mathbb{E}[R_n(Z_m; z, B - T_m) | X_0 = x] \\ &= \sum_{y \in E} \int_{\mathbb{R}_+} R_m(x; y; dt) R_n(y; z, B - t).\end{aligned}$$

In other words,

$$R_{m+n}(x; z, B) = \sum_{y \in E} \int_B \int_{R_+} R_m(x; y, dt) R_n(y; z, du - t),$$

hence also

$$R_{m+n}(x; z, du) = \sum_y \int_0^u R_m(x; y, dt) R_n(y; z, du - t),$$

where the measure $R_n(y; z, du - t)$ is defined by

$$\int_{\mathbb{R}_+} R_n(y; z, du - t) f(u) = \int_{\mathbb{R}_+} R_n(y; z, du) f(t + u).$$

Clearly the $\{R_n;\ n \geq 1\}$ are completely determined by R_1 and this equation.

Note that

$$\begin{aligned}P_{xy}(s) &= \sum_{m \geq 0} \mathbb{P}(Z_m = y, T_m \leq s < T_{m+1} | Z_0 = x) \\ &= \sum_{m \geq 0} P(Z_m = y, T_m \leq s, T_{m+1} - T_m > s - T_m | Z_0 = x) \\ &= \sum_{m \geq 0} \mathbb{E}[\mathbb{P}(T_{m+1} - T_m > s - T_m | Z_m, T_m) \mathbf{1}_{\{Z_m = y, T_m \leq s\}} | Z_0 = x] \\ &= \sum_{m \geq 0} \mathbb{E}[e^{-q Z_m (s - T_m)} \mathbf{1}_{\{Z_m = y, T_m \leq s\}} | Z_0 = x] \\ &= \sum_{m \geq 0} \int_0^s e^{-q_y (s-t)} R_m(x; y, dt),\end{aligned}$$

where we have used the strong Markov property at time T_m for the third equality. Hence, from the above equation,

$$\begin{aligned}P_{xy}(s) &= \delta_{xy} e^{-q_x s} + \sum_{m \geq 1} \int_0^s e^{-q_y (s-t)} R_m(x; y, dt) \\ &= \delta_{xy} e^{-q_x s} + \sum_{m \geq 0, z \in E} \int_0^s e^{-q_y (s-t)} \int_0^t R_1(x; z, du) R_m(z; y, dt - u)\end{aligned}$$

or, equivalently,

$$P_{xy}(t) = \delta_{xy} e^{-q_x t} + \sum_{z \in E} \int_0^t R_1(x; z, ds) P_{zy}(t-s)$$

$$e^{q_x t} P_{xy}(t) = \delta_{xy} + \int_0^t e^{q_x s} \sum_{z \neq x} Q_{xz} P_{zy}(s) ds.$$

Hence the function $t \to P_{xy}(t)$ is differentiable and

$$\begin{aligned} \frac{d}{dt} P_{xy}(t) &= \sum_{z \neq x} Q_{xz} P_{zy}(t) - q_x P_{xy}(t) \\ &= \sum_z Q_{xz} P_{zy}(t). \end{aligned}$$

□

The above argument shows that

$$\begin{aligned} &\mathbb{P}(Z_1 = y, T_1 \in B, Z_2 = z, T_2 - T_1 \in C | Z_0 = x) \\ &\qquad = \int_B \int_C R_1(x, y, dt) R_1(y, z, du). \end{aligned}$$

This formula generalizes to the law of $((Z_1, T_1), \ldots, (Z_n, T_n))$. One can deduce the Markov property of the corresponding process $\{X_t;\ t \geq 0\}$ from that joint law.

Remark 3.3 *If we allow an arbitrary generator Q, one can always define R_1, and hence the law of the (Z_n, T_n). But the sequence $\{T_n\}$ does not necessarily satisfy the non-explosion condition (7.3), that is, the corresponding process $\{X_t\}$ need not be defined for all $t \geq 0$. In the next section we shall give sufficient conditions on Q for the non-explosion condition to hold.*

7.4 Embedded Markov chain

Let $\{X_t; t \geq 0\}$ be a jump Markov process, whose jump times $T_1, T_2, \ldots, T_n, \ldots$ satisfy the non-explosion condition (7.3). The sequence $\{Z_n; n \in \mathbb{N}\}$ defined by

$$Z_n = X_{T_n} \quad (\text{with } T_0 = 0)$$

is a discrete time Markov chain (this is a consequence of the strong Markov property of $\{X_t\}$), called the 'embedded chain', which has the property that $Z_{n+1} \neq Z_n$ almost surely, for all $n \geq 0$. Its transition matrix P is easily computed in terms of the infinitesimal generator Q of $\{X_t\}$:

$$P_{xy} = \begin{cases} (-Q_{xx})^{-1} Q_{xy}, & \text{if } y \neq x, \\ 0, & \text{if } y = x. \end{cases}$$

Define, for $n \geq 1$,

$$S_n = q_{Z_{n-1}}(T_n - T_{n-1}) \quad (\text{where } q_x = -Q_{xx})$$

and, for $t \geq 0$,

$$N_t = \sup\left\{n; \sum_{k=1}^{n} S_k \leq t\right\}.$$

Then $\{N_t;\ t \geq 0\}$ is a Poisson process with intensity 1 (this follows from the strong Markov property of $\{X_t\}$, and the fact that if $U \simeq$ exponential (λ), then $\lambda U \simeq$ exponential (1)).

Now let Q be an infinitesimal generator, that is, a matrix indexed by $E \times E$, such that, for all $x \in E$,

$$Q_{xy} \geq 0, \quad y \neq x,$$

$$Q_{xx} = -\sum_{y \neq x} Q_{xy} < 0.$$

We let $q_x = -Q_{xx}$, and we define the transition matrix P by

$$P_{xy} = \begin{cases} \frac{Q_{xy}}{q_x}, & \text{if} y \neq x, \\ 0, & \text{if} y = x, \end{cases} \tag{7.4}$$

with the convention that if $Q_{xx} = 0$, $P_{xy} = 0$ for all $y \neq x$, $P_{xx} = 1$. With any initial condition $x \in E$ we associate the Markov chain $\{Z_n;\ n \geq 0\}$, with transition matrix P. Now let $\{N_t;\ t \geq 0\}$ be a Poisson process with intensity 1, independent of the chain $\{Z_n;\ n \in \mathbb{N}\}$. Denote by $0 = T_0 < T_1 < T_2 < \ldots$ the times of the jumps of the Poisson process, and define, for $n \geq 1$,

$$S_n = \frac{T_n - T_{n-1}}{q(Z_{n-1})},$$

$$T'_n = S_1 + \ldots + S_n.$$

If the non-explosion condition (7.3) is satisfied by the sequence $\{T'_n\}$, then

$$X_t \stackrel{\text{def}}{=} \sum_{n \geq 0} Z_n \mathbf{1}_{[T'_n, T'_{n+1}[}(t), \quad t \geq 0, \tag{7.5}$$

is a jump Markov process with the infinitesimal generator Q.

It remains to answer the question: given an infinitesimal generator Q, when does the associated sequence of stopping times $\{T'_n;\ n \geq 0\}$ satisfy the non-explosion condition, that is, when does (7.5) define X_t for all $t \geq 0$? Let us establish the following proposition.

Proposition 4.1 *The non-explosion condition (7.3) is satisfied if and only if*

$$\sum_{n \geq 0} q_{Z_n}^{-1} = +\infty \ a.s. \tag{7.6}$$

Let us first state a corollary.

Corollary 4.2 *A sufficient condition for the infinitesimal generator Q to be the infinitesimal generator of a Markov process which satisfies condition (7.3) is that one of the two following conditions holds:*

1. $\sup_{x\in E} q_x < \infty$.
2. *The Markov chain $\{Z_n\}$ with transition matrix P defined by (7.4) is recurrent.*

It is clear that each of the two conditions in this corollary implies (7.6). Proposition 4.1 follows from the following lemma if we let

$$A_n = T_{n+1} - T_n, \quad B_n = \frac{1}{q_{Z_n}}, \quad n \geq 0.$$

Lemma 4.3 *Let $\{A_n;\ n \geq 1\}$ and $\{B_n;\ n \geq 1\}$ be two mutually independent sequences of $\mathbb{R}_+^*$-valued random variables, the sequence $\{A_n\}$ being i.i.d., and the common law being the exponential distribution with parameter* 1. *Then the following two statements are equivalent:*

1. $\sum_{n=1}^{\infty} A_n B_n = +\infty$ *a.s.*
2. $\sum_{n=1}^{\infty} B_n = +\infty$ *a.s.*

Proof Since the two sequences are mutually independent, the lemma will follow from the fact that, for any sequence $\{b_n;\ n \geq 1\}$ of strictly positive real numbers,

$$\sum_{n=1}^{\infty} A_n b_n = +\infty \text{ a.s.} \Longleftrightarrow \sum_{n=1}^{\infty} b_n = +\infty. \tag{7.7}$$

If $\sum_n b_n < \infty$, then $\mathbb{E}\sum_n A_n b_n = \sum_n b_n < \infty$ and a fortiori $\sum_n A_n b_n < \infty$ almost surely. It remains to prove that if $\sum_n b_n = +\infty$, then

$$\Lambda_n := \sum_{k=1}^{n} A_k b_k \to +\infty \text{ a.s.}, \quad \text{as } n \to \infty.$$

In the case where there is a subsequence n_j such that $b_{n_j} \to +\infty$, clearly $\sum_n A_n b_n \geq \sum_j A_{n_j} b_{n_j} = +\infty$, since, the A_{n_j} being i.i.d. exponential random variables, infinitely many of them are greater than 1. It thus remains to consider the case where $0 \leq b_n \leq C$ and $\sum_n b_n = +\infty$. In this case, for all $M > 0$, if n is big enough such that $\mathbb{E}\Lambda_n > 2M$, then

$$\begin{aligned}
\mathbb{P}(\Lambda_n \leq M) &\leq \mathbb{P}\left(|\Lambda_n - \mathbb{E}\Lambda_n| \geq \frac{\mathbb{E}\Lambda_n}{2}\right) \\
&\leq 4\frac{\operatorname{var}(\Lambda_n)}{(\mathbb{E}\Lambda_n)^2} = 4\frac{\sum_1^n b_k^2}{(\sum_1^n b_k)^2} \\
&\leq \frac{4C}{\sum_1^n b_k} \to 0,
\end{aligned}$$

hence $\Lambda_n \to +\infty$ in probability, and also almost surely since the sequence is monotone. □

Remark 4.4 *In the next chapter, we shall specify jump Markov processes by describing their infinitesimal generator Q. The reader can check that in each of the examples considered, one (usually the first) of the two sufficient conditions of Corollary 4.2 is satisfied.*

7.5 Recurrent and transient states

In what follows, as in the discrete time case, we shall denote by $\mathbb{P}_x$ the conditional law of $\{X_t;\ t \geq 0\}$ given that $X_0 = x$. The equivalence classes of the jump Markov process $\{X_t; t \geq 0\}$ are those of the embedded chain. Note that if $\{X_t; t \geq 0\}$ is irreducible,

$$P_{xy}(t) > 0, \quad \forall x, y \in E, t > 0. \tag{7.8}$$

Indeed, for all $x, y \in E$, there exist $n \geq 1$ and $x_0 = x, x_1, \ldots, x_{n-1}, x_n = y$ such that $Q_{x_{k-1}x_k} > 0$, $1 \leq k \leq n$, and it follows from the property of the exponential law that $P_{xy}(t) \geq P_{xx_1}(t/n) \times \cdots \times P_{x_{n-1}y}(t/n) > 0$.

A state $x \in E$ is called recurrent (transient) for $\{X_t; t \geq 0\}$ if it is recurrent (transient) for the embedded chain. Then, in particular in the irreducible case, either all states are recurrent or all are transient.

As in the case of discrete time Markov chains, we have the following theorem.

Theorem 5.1 *Let $\{X_t; t \geq 0\}$ be an irreducible and recurrent jump Markov process. Then there exists a strictly positive invariant measure π on E which solves the equation $\pi Q = 0$ and is unique up to a multiplicative constant. Moreover, such a measure is invariant for the semigroup $\{P(t)\}$ in the sense that $\pi P(t) = \pi$, for all $t \geq 0$.*

Proof We note that if Q is the infinitesimal generator of the jump Markov process $\{X_t\}$ and P is the transition matrix of its embedded chain, then

$$Q = \mathbf{q}(P - I),$$

where $\mathbf{q}$ is the diagonal matrix defined by

$$\mathbf{q}_{xy} = \delta_{xy} q_x, \quad x, y \in E.$$

Note that the assumption that the process is irreducible implies that $q_x > 0$, for all $x \in E$. Hence, we can multiply by $\mathbf{q}^{-1}$. Our assumption is that the embedded chain is irreducible and recurrent. Hence, the measure γ^x defined in the proof of Theorem 5.3 of Chapter 2 is strictly positive, and it is the unique solution (up to a multiplicative constant) of the equation $\gamma^x P = \gamma^x$. Hence, the strictly positive

measure $\mu^x = \mathbf{q}^{-1}\gamma^x$ satisfies $\mu^x Q = 0$, and any other solution μ' of the same equation is such that $\mathbf{q}\mu'$ is P-invariant, hence there exists a constant c such that $\mu' = c\mu^x$. It remains to check that μ^x is invariant by $P(t)$, for all $t \geq 0$.

Since γ_y^x is the expectation of the number of visits to state y by the embedded chain during an excursion starting at x, q_y^{-1} is the expectation of the time spent at state y by the process $\{X_t\}$ at each of the visits to y of the embedded chain, and the embedded chain is independent of the times spent at each state by the jump Markov process,

$$\mu_y^x = \frac{\gamma_y^x}{q_y} = \mathbb{E}_x \int_0^{R_x} \mathbf{1}_{\{X_s = y\}} ds$$
$$= \mathbb{E}_x \int_0^{\infty} \mathbf{1}_{\{X_s = y, s < R_x\}} ds,$$

where $R_x = \inf\{t > T_1;\ X_t = x\}$ denotes the time of the first return to x. But if $t > 0$, by the strong Markov property,

$$\mathbb{E}_x \int_0^t \mathbf{1}_{\{X_s = y\}} ds = \mathbb{E}_x \int_{R_x}^{R_x + t} \mathbf{1}_{\{X_s = y\}} ds.$$

Hence,

$$\mu_y^x = \mathbb{E}_x \int_t^{R_x + t} \mathbf{1}_{\{X_s = y\}} ds$$
$$= \mathbb{E}_x \int_0^{R_x} \mathbf{1}_{\{X_{t+s} = y\}} ds$$
$$= \int_0^{\infty} \mathbb{P}_x(X_{t+s} = y, s < R_x) ds$$
$$= \int_0^{\infty} \sum_z \mathbb{P}_x(X_s = z, s < R_x) P_{zy}(t) ds$$
$$= \sum_z \mu_z^x P_{zy}(t).$$

□

7.6 The irreducible recurrent case

In order to distinguish between the positive and null recurrent cases (this is an open question only if $|E| = +\infty$), it is not sufficient to consider the property of the embedded chain, as we shall now see. As in the previous section, define $R_x = \inf\{t \geq T_1; X_t = x\}$ again as the time of the first return to state x.

Definition 6.1 *The state x is said to be positive recurrent if it is recurrent and $\mathbb{E}_x(R_x) < \infty$, and null recurrent if it is recurrent and $\mathbb{E}_x(R_x) = +\infty$.*

Again in the irreducible recurrent case, states are either all null recurrent or all positive recurrent, and we say accordingly that the process $\{X_t\}$ is null recurrent or positive recurrent. We now prove that the positive recurrent case is equivalent to the existence of a unique invariant probability distribution.

Theorem 6.2 *Let $\{X_t;\ t \geq 0\}$ be an irreducible jump Markov process. A state $x \in E$ is positive recurrent if and only if all states are positive recurrent, if and only if there exists a unique invariant probability distribution π, and in that case*

$$\mathbb{E}_x R_x = \frac{1}{\pi_x q_x}, \quad \forall x \in E.$$

PROOF If state x is positive recurrent for $\{X_t\}$, then x is recurrent for the embedded chain $\{Z_n;\ n \geq 0\}$. Denote by γ_y^x the mean number of visits to state y during an excursion of $\{Z_n\}$ starting at x. Since the time spent at y by $\{X_t\}$ at each visit of $\{Z_n\}$ is independent of the embedded chain and has expectation q_y^{-1},

$$\mathbb{E}_x R_x = \sum_{y \in E} \frac{\gamma_y^x}{q_y}.$$

But we saw in the proof of Theorem 5.1 that the measure μ^x defined by

$$\mu_y^x = \frac{\gamma_y^x}{q_y}$$

satisfies $\mu^x Q = 0$. The condition that x is positive recurrent therefore implies the existence of an invariant measure with finite mass, hence of an invariant probability distribution, whose uniqueness follows from Theorem 5.1. Suppose now that there exists a probability distribution π which solves $\pi Q = 0$. Then the measure $\mathbf{q}\pi$ is P-invariant, and, for all $x, y \in E$,

$$\frac{q_y \pi_y}{q_x \pi_x}$$

is the mean number of visits to state y during an excursion of $\{Z_n\}$ starting at x. Hence

$$\mathbb{E}_x R_x = \sum_{y \in E} \frac{\pi_y}{q_x \pi_x} < \infty, \quad \forall x \in E,$$

and any state $x \in E$ is positive recurrent. □

Remark 6.3 *An invariant measure π of the jump Marov process $\{X_t\}$ is a solution of $\pi Q = 0$. An invariant measure μ of the embedded chain is a solution of $\mu(P - I) = 0$. Hence, π is invariant for $\{X_t\}$ if and only if $\mu = \pi \mathbf{q}$ is invariant for the embedded chain. It is then easy to choose $\mathbf{q}$ (which specifies the expectations of the lengths of the visits of $\{X_t\}$ to the various states) in such a way that π has a finite mass while μ has infinite mass (hence $\{X_t\}$ is positive recurrent while the embedded chain is null recurrent), or vice versa. See part 8 of Exercises 11.8 and the explanations at the end of the solution.*

We now restrict ourselves to the positive recurrent case, establishing an ergodic theorem and convergence of transition probabilities towards the invariant probability.

Theorem 6.4 *Let $\{X_t; t \geq 0\}$ be an E-valued irreducible, positive recurrent jump Markov process. Let Q be its infinitesimal generator, and π the unique invariant probability distribution. Then if $f : E \to \mathbb{R}$ is bounded,*

$$\frac{1}{t}\int_0^t f(X_s)ds \to \sum_{x\in E} f(x)\pi_x$$

almost surely as $t \to \infty$.

Proof It suffices to consider the case $f(y) = \mathbf{1}_{\{y=x\}}$ and to work under $\mathbb{P}_x$ (see the proof of Theorem 5.7 of Chapter 2). As in the discrete time case, the successive excursions starting at x are i.i.d. Let $N^x(t)$ denote the number of visits to state x between time 0 and time t, and let T_k^x denote the time spent at state x by the process $\{X_t\}$ during its kth visit. Since the $N^x(t)$th visit to state x need not be terminated at time t, we have

$$\frac{1}{t}\sum_{k=1}^{N^x(t)-1} T_k^x < \frac{1}{t}\int_0^t \mathbf{1}_{\{X_s=x\}}ds \leq \frac{1}{t}\sum_{k=1}^{N^x(t)} T_k^x.$$

But clearly $T^x_{N^x(t)}/t \to 0$ almost surely as $t \to \infty$, and

$$\begin{aligned}\frac{1}{t}\sum_{k=1}^{N^x(t)} T_k^x &= \frac{N^x(t)}{t} \times \frac{1}{N^x(t)}\sum_{k=1}^{N^x(t)} T_k^x \\ &\to \frac{1}{\mathbb{E}_x R_x} \times \frac{1}{q_x} \\ &= \pi_x.\end{aligned}$$

Indeed, since the sequence $\{T_k^x;\ k \geq 1\}$ is i.i.d., and since from recurrence $N^x(t) \to \infty$ almost surely as $t \to \infty$,

$$\frac{1}{N^x(t)}\sum_{k=1}^{N^x(t)} T_k^x \to \mathbb{E}_x(T_1^x) = \frac{1}{q_x},$$

and the proof of the fact that

$$\frac{t}{N^x(t)} \to \mathbb{E}_x(R_x)$$

follows the same argument as in the proof of Theorem 5.7 of Chapter 2. □

In the continuous time case, the convergence of the probability distribution of X_t towards the invariant distribution as $t \to \infty$ holds in the irreducible and positive recurrent case, without further restriction.

Theorem 6.5 *Let $\{X_t; t \geq 0\}$ be an E-valued irreducible, positive recurrent jump Markov process, and π its unique invariant probability distribution solving the stationary Fokker–Planck equation $\pi Q = 0$. Then, for any probability distribution μ on E and $x \in E$, $(\mu P)_x(t) \to \pi_x$ as $t \to \infty$.*

PROOF One could imitate the proof of Theorem 6.4 of Chapter 2, but instead we shall use that result.

If we sample the process $\{X_t\}$ by letting $Y_n = X_{nh}$, $n = 0, 1, \ldots$, where $h > 0$ is arbitrary, then clearly $\{Y_n;\ n \in \mathbb{N}\}$ is an irreducible and aperiodic (see (7.8)) Markov chain, whose unique invariant probability distribution, which does not depend upon h, is π. Let us assume the following lemma, which we will establish once the proof of the present theorem is complete.

Lemma 6.6 *For all $t, h > 0$, $x, y \in E$,*

$$|P_{xy}(t+h) - P_{xy}(t)| \leq 1 - e^{-q_x h}.$$

Fix $\varepsilon > 0$ and $x, y \in E$. We first choose $h > 0$ sufficiently small, in such a way that

$$1 - e^{-q_x s} \leq \varepsilon/2, \quad \text{if } 0 \leq s \leq h.$$

Then we choose N large enough such that

$$|P_{xy}(nh) - \pi_y| \leq \varepsilon/2, \quad \text{if } n \geq N.$$

We conclude that if $t \geq Nh$, denoting by n the integer such that $nh \leq t < (n+1)h$,

$$|P_{xy}(t) - \pi_y| \leq |P_{xy}(t) - P_{xy}(nh)| + |P_{xy}(nh) - \pi_h| \leq \varepsilon.$$

The theorem follows easily from this result, if we decompose the set of starting points into a finite set which supports the mass of μ except for δ, and its complement. □

PROOF OF LEMMA 6.6 It suffices to note that

$$\begin{aligned}
|P_{xy}(t+h) - P_{xy}(t)| &= |\sum_z P_{xz}(h) P_{zy}(t) - P_{xy}(t)| \\
&= \left| \sum_{z \neq x} P_{xz}(h) P_{zy}(t) - (1 - P_{xx}(h)) P_{xy}(t) \right| \\
&= \left| \sum_{z \neq x} P_{xz}(h) (P_{zy}(t) - P_{xy}(t)) \right| \\
&\leq \sum_{z \neq x} P_{xz}(h) \\
&= 1 - P_{xx}(h).
\end{aligned}$$

□

Remark 6.7 *The convergence in Theorem 6.5 is in the sense of weak convergence of probability distributions on E. This follows from the fact that, as $t \to \infty$, $(\mu P)_x(t) \to \pi_x$ for all $x \in E$, hence also for any finite subset $F \subset E$, $\sum_{x\in F} (\mu P)_x(t) \to \sum_{x\in F} \pi_x$. Since, moreover, $\mu P(t)$ and π are probability distributions on E, it is not hard to show that, for any bounded $f : E \to \mathbb{R}$,*

$$\sum_x (\mu P)_x(t) f(x) \to \sum_x \pi_x f(x), \quad \text{as } t \to \infty.$$

Note that if we choose the invariant probability distribution π as the law of X_0, then the process $\{X_t; t \geq 0\}$ is *stationary* in the sense that, for all $n \in \mathbb{N}$, $0 \leq t_1 < t_2 < \ldots < t_n$, the law of the random vector $(X_{t_1+s}, X_{t_2+s}, \ldots, X_{t_n+s})$ does not depend upon $s \geq 0$.

Remark 6.8 *For all $x \in E$, the equation $\pi Q = 0$ reads*

$$\sum_{y\neq x} \pi_y Q_{yx} = \pi_x \sum_{y\neq x} Q_{xy}.$$

The left-hand side of this equality is the mean flux entering state x at equilibrium, coming from the various other states, and the right-hand side is the mean flux leaving x at equilibrium, towards the other states. The equation $\pi Q = 0$ says that at equilibrium the mean numbers of departures and arrivals per unit time are equal.

We also have a generalization of the central limit theorem. The next result is a special case of Theorem 2.1 in [5].

Theorem 6.9 *Suppose that the jump Markov process $\{X_t; t \geq 0\}$ is irreducible, and that it has an invariant probability distribution π. Let $f \in L^2(E, \pi)$ be of the type $f = Qg$, where $g \in L^2(E, \pi)$ (this implies that $\langle \pi f \rangle = \sum_{x\in E} \pi_x f_x = \sum_{x,y\in E} \pi_x Q_{xy} g_y = 0$). Let*

$$C(f) := -2 \sum_{x\in E} f_x g_x \pi_x,$$

which we suppose not to be equal to zero (consequently, $C(f) > 0$). Then

$$\frac{1}{\sqrt{tC(f)}} \int_0^t f(X_s) ds \to Z$$

in law, as $t \to \infty$, where Z is a centred Gaussian random variable with unit variance.

Finally, we have the convergence of

$$\left\{ \frac{1}{\sqrt{uC(f)}} \int_0^{tu} f(X_s) ds, \; t \geq 0 \right\}$$

towards a Brownian motion $\{B_t; \; t \geq 0\}$, as $u \to \infty$.

7.7 Reversibility

Given a jump Markov process $\{X_t;\ t \geq 0\}$ and $T > 0$, $\{\hat{X}_t^T = X_{T-t},\ 0 \leq t \leq T\}$ is also a Markov process. If the law of X_0 is an invariant probability distribution π, then $\hat{X}^T$ is time-homogeneous. Denote by $\hat{Q}$ its infinitesimal generator.

Theorem 7.1 *$\hat{Q} = Q$ if and only if the detailed balance equation*

$$\pi_x Q_{xy} = \pi_y Q_{yx}, \quad \forall x, y \in E,$$

is satisfied. In this case, we say that the process $\{X_t\}$ is reversible (with respect to the probability distribution π, which then is invariant).

Proof The same argument as in the discrete time case implies that, for all $t > 0$, $x, y \in E$,

$$\hat{P}_{xy}(t) = \frac{\pi_y}{\pi_x} P_{yx}(t),$$

from which it follows, by taking the t-derivative at $t = 0$, that

$$\hat{Q}_{xy} = \frac{\pi_y}{\pi_x} Q_{yx}.$$

The result is now obvious. □

Remark 7.2 *As in the case of discrete time Markov chains, a jump Markov process which is irreducible and positive recurrent need not be reversible. Again, a counter-example is provided by a Q-matrix such that, for a given pair $x \neq y$, $Q_{xy} = 0 \neq Q_{yx}$, which does not contradict irreducibility if $|E| \geq 3$.*

Remark 7.3 *As in the case of discrete time Markov chains, to find a generator Q such that a given distribution π is Q-invariant is not difficult. The easiest approach is to look for Q such that the associated process is reversible with respect to π, hence to look for an infinitesimal generator Q such that the quantity $\pi_x Q_{xy}$ is symmetric in x and y.*

To find the invariant probability distribution, given an irreducible generator, is in general more difficult. One can try to solve

$$\pi_x Q_{xy} = \pi_y Q_{yx}, \quad \forall x, y \in E,$$

but this equation has a solution only in the reversible case. In the non-reversible case, one should solve the stationary Fokker–Planck equation $\pi Q = 0$. If one can guess π up to a multiplicative constant, then one can take advantage of the following result.

Theorem 7.4 *Given a probability distribution π on E, define, for $x, y \in E$,*

$$\hat{Q}_{xy} = \frac{\pi_y}{\pi_x} Q_{yx}.$$

If

$$\sum_{y \neq x} \hat{Q}_{xy} = \sum_{y \neq x} Q_{xy},$$

then π is an invariant probability distribution and $\hat{Q}$ is the generator of the time-reversed process.

PROOF Using the first and then the second identity from the statement, we deduce that

$$\sum_{y \neq x} \pi_y Q_{yx} = \pi_x \sum_{y \neq x} \hat{Q}_{xy} = \pi_x \sum_{y \neq x} Q_{xy} = -\pi_x Q_{xx},$$

which implies that $\pi Q = 0$. The second part of the statement is then a consequence of the formula which appears in the proof of Theorem 7.1. □

Note that if we can guess the generator of the time-reversed process, we can deduce the invariant probability distribution up to the normalization constant.

7.8 Markov models of evolution and phylogeny

This section is devoted to the presentation of probabilistic models of molecular evolution, and maximum likelihood and Bayesian methods for the construction of phylogenetic trees. In this section only, we depart from our usual practice of listing works cited at the back of the book. Instead we give the names of the author(s) followed by a date, and direct the reader to the excellent bibliographies in Felsenstein [20] and Yang [38]. Note that Felsenstein (F84) does not correspond to a paper; see the details on page 220 of [20]. We also refer to those two books for further reading on the topics of this section.

We shall define Markov processes on trees, which is a model frequently used in phylogeny. We shall consider rooted and unrooted binary trees. Figure 7.2 shows a rooted binary tree (the root is at the top and the leaves are at the bottom) and Figure 7.3 an unrooted binary tree.

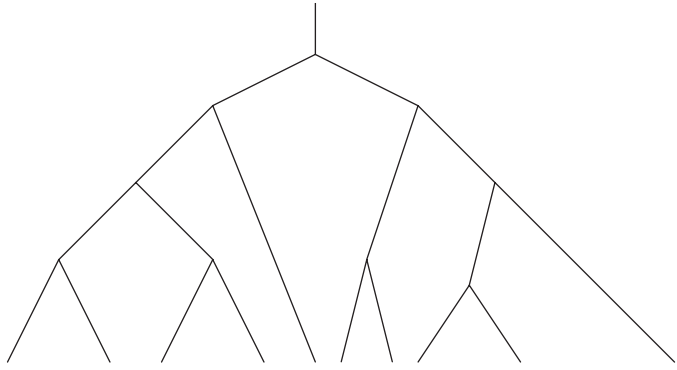


Figure 7.2 Rooted binary tree.

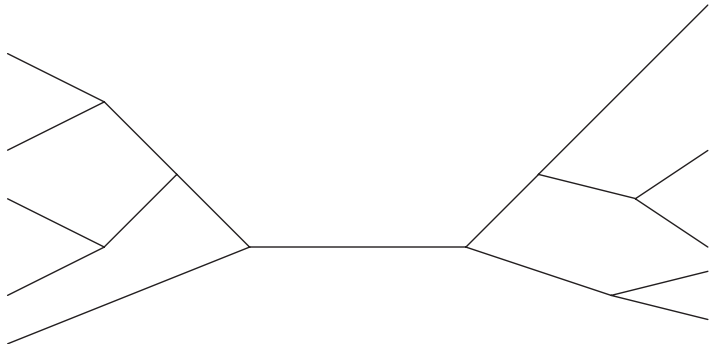


Figure 7.3 The unrooted version of the tree in Figure 7.2.

The Markov process on a *rooted* binary tree starts at the root (which plays the role of the initial time 0) in a certain state, say x. It evolves up to the first node which is located at distance r from the root, as a continuous time jump Markov process during a time interval of length r. Denote by y the state of the process at that node. On each branch which starts from that node a jump Markov process runs, starting at y, in such a way that the processes on the two branches are independent, up to the next node, and so on down to the leaves of the tree. Note that we shall only consider irreducible processes with values in a finite state space, hence the process will be positive recurrent, and we shall choose the invariant probability distribution as the law at the root. We can then, without changing the law of the process, suppress the branch between the root and the first node.

Turning to the Markov process on an *unrooted* tree, let us suppress the branch between the root and the first node. This means that the process starts at the root (located at the first node) under the invariant probability distribution, and evolves independently on the two branches, until it encounters the next node, etc. Consider Figure 7.3. The root is no longer indicated on the central branch. One can still imagine that it is there, and consider that two processes start at that point, one in each direction, towards the two nodes located at the two ends of the central branch. Suppose now that we move the root on the central branch either to the right or to the left. It is easy to convince oneself that the law of the resulting process on the tree is not modified, provided the process is *reversible*. Indeed, the difference between the two constructions with the root at two different points on the central branch is that a portion of that branch is run in the two different directions by the two versions of the process. In the reversible case, the root can equivalently be located at either end of the central branch, or even at any node of the tree, or at any point of any branch of the tree. This means that one can define a jump Markov process on an unrooted tree, provided the dynamics is time-reversible, by locating the starting point anywhere on the tree.

7.8.1 Models of evolution

In order to compute the likelihood of a given tree as a function of the data, we need to choose a model of evolution which tells us how the data have been 'manufactured' by evolution along the branches of the tree, for each DNA site. We shall describe several Markovian models of DNA evolution, by describing the transition rate from one nucleotide to another. This means that we shall prescribe the matrix Q in the form

$$Q = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \end{array}$$

The Jukes and Cantor (1969) model This is the simplest one, which assumes that all mutations happen at the same rate, that is, for a given $\alpha > 0$,

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}.$$

The associated invariant probability distribution is uniform over the four nucleotides. The transition probabilities are easy to compute: $P(t)$ is given by

$$\begin{pmatrix} 0.25 + 0.75e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} \\ 0.25 - 0.25e^{-4\alpha t} & 0.25 + 0.75e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} \\ 0.25 - 0.25e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} & 0.25 + 0.75e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} \\ 0.25 - 0.25e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} & 0.25 - 0.25e^{-4\alpha t} & 0.25 + 0.75e^{-4\alpha t} \end{pmatrix}.$$

The Kimura (1980, 1981) models Among the four nucleotides, cytosine and thymine are pyrimidines, while adenine and guanine are purines. It is reasonable to assume that transitions (replacement of one purine by the other, or of one pyrimidine by the other) are more frequent than transversions (replacement of a purine by a pyrimidine or vice versa). One is then led to assume that the substitution rates between A and G or between C and T are greater than all others, hence the model (with $\beta > \alpha$)

$$Q = \begin{pmatrix} -2\alpha - \beta & \alpha & \beta & \alpha \\ \alpha & -2\alpha - \beta & \alpha & \beta \\ \beta & \alpha & -2\alpha - \beta & \alpha \\ \alpha & \beta & \alpha & -2\alpha - \beta \end{pmatrix}.$$

The invariant probability distribution is again uniform. The transition probabilities are given by

$$P_{xx}(t) = 0.25 + 0.25e^{-4\beta t} + 0.5e^{-2(\alpha+\beta)t},$$

$$P_{xy}(t) = 0.25 + 0.25e^{-4\beta t} - 0.5e^{-2(\alpha+\beta)t},$$

if $x \neq y$ are either both purines or both pyrimidines, or

$$P_{xy}(t) = 0.5 - 0.5e^{-4\beta t}$$

in the other case.

Kimura proposed a second model, of the form

$$Q = \begin{pmatrix} -\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & -\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & -\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & -\alpha-\beta-\gamma \end{pmatrix},$$

for which the invariant probability distribution is still uniform.

The Felsenstein (F84) model Given a probability distribution π on $E = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, and a positive number u, Felsenstein proposed the model

$$Q = \begin{pmatrix} u(\pi_{\mathrm{A}}-1) & u\pi_{\mathrm{C}} & u\pi_{\mathrm{G}} & u\pi_{\mathrm{T}} \\ u\pi_{\mathrm{A}} & u(\pi_{\mathrm{C}}-1) & u\pi_{\mathrm{G}} & u\pi_{\mathrm{T}} \\ u\pi_{\mathrm{A}} & u\pi_{\mathrm{C}} & u(\pi_{\mathrm{G}}-1) & u\pi_{\mathrm{T}} \\ u\pi_{\mathrm{A}} & u\pi_{\mathrm{C}} & u\pi_{\mathrm{G}} & u(\pi_{\mathrm{T}}-1) \end{pmatrix}.$$

Clearly, for $x \neq y$,

$$\pi_x Q_{xy} = \pi_y Q_{yx},$$

hence π is the invariant probability distribution and the chain is reversible. The matrix Q has two eigenvalues: $-u$, whose associated eigenspace is composed of those vectors which are orthogonal to π in $\mathbb{R}^4$; and 0, whose associated eigenspace is composed of those vectors which are collinear to $(1, 1, 1, 1)$. One can then show that

$$P_{xy}(t) = \left(e^{tQ}\right)_{xy} = e^{-ut}\delta_{xy} + (1 - e^{-ut})\pi_y.$$

In the special case where $\pi = (1/4, 1/4, 1/4, 1/4)$, this model reduces to the Jukes–Cantor model.

The Hasegawa, Kishino and Yano (1985) model This is a generalization of both the first Kimura model and Felsenstein's. Given again an arbitrary probability distribution π on E, and u, v two positive numbers, let

$$Q = \begin{pmatrix} -u\pi_{\mathrm{G}} - v\pi_2 & v\pi_{\mathrm{C}} & u\pi_{\mathrm{G}} & v\pi_{\mathrm{T}} \\ v\pi_{\mathrm{A}} & -u\pi_{\mathrm{T}} - v\pi_1 & v\pi_{\mathrm{G}} & u\pi_{\mathrm{T}} \\ u\pi_{\mathrm{A}} & v\pi_{\mathrm{C}} & -u\pi_{\mathrm{A}} - v\pi_2 & v\pi_{\mathrm{T}} \\ v\pi_{\mathrm{A}} & u\pi_{\mathrm{C}} & v\pi_{\mathrm{G}} & -u\pi_{\mathrm{C}} - v\pi_1 \end{pmatrix},$$

where $\pi_1 = \pi_{\mathrm{A}} + \pi_{\mathrm{G}}$, $\pi_2 = \pi_{\mathrm{C}} + \pi_{\mathrm{T}}$. Again π is the invariant probability distribution, and again it is possible to deduce an explicit expression for $P(t)$.

There are good reasons to assume that $\pi_{\mathrm{C}} = \pi_{\mathrm{G}}$ and $\pi_{\mathrm{A}} = \pi_{\mathrm{T}}$, since DNA is a molecule with two strands, made of pairs C : G and A : T. The above identities are

a consequence of the effect of this constraint on evolution. With this restriction, the HKY model becomes a three-parameter model, namely u, v and $\theta = \pi_C + \pi_G$, as proposed by Tamura (1992). It takes the form

$$Q = \frac{1}{2}\begin{pmatrix} -u\theta - v & v\theta & u\theta & v(1-\theta) \\ v(1-\theta) & -u(1-\theta) - v & v\theta & u(1-\theta) \\ u(1-\theta) & v\theta & -u(1-\theta) - v & v(1-\theta) \\ v(1-\theta) & u\theta & v\theta & -u\theta - v \end{pmatrix}.$$

The general reversible model Since $|E|$ is very small, one can try to use the most general model. It is easy to parameterize the most general reversible model in the form

$$Q = \begin{pmatrix} -uW & uA\pi_C & uB\pi_G & uC\pi_T \\ uD\pi_A & -uX & uE\pi_G & uF\pi_T \\ uG\pi_A & uH\pi_C & -uY & uI\pi_T \\ uJ\pi_A & uK\pi_C & uL\pi_G & -uZ \end{pmatrix},$$

where u is a positive number, π the invariant probability distribution,

$$\begin{aligned} W &= A\pi_C + B\pi_G + C\pi_T \\ X &= D\pi_A + E\pi_G + F\pi_T \\ Y &= G\pi_A + H\pi_G + I\pi_T \\ Z &= J\pi_A + K\pi_C + L\pi_G, \end{aligned}$$

and the parameters $A, B, \ldots, L$ are to be chosen. As we shall see below, it is useful for the computation of the likelihood that the model be reversible. The constraint of reversibility imposes six relations, namely

$$A = D,\ B = G,\ C = J,\ E = H,\ F = K,\ I = L.$$

There remain six parameters to choose, for example A, B, C, E, F and I. There are, moreover, three free parameters describing the invariant probability distribution. Thus, nine parameters need to be chosen.

Codon models A codon is a triplet of nucleotides which codes for an amino acid. Among the $4^3 = 64$ possible codons, three are possible stop codons, while the 61 others code for the 20 amino acids. Note that the genetic code (the translation rule from codons to amino acids) is *degenerate*, in the sense that several distinct codons code for the same amino acid. Hence, among the possible codon mutations, one must distinguish the synonymous mutations (which transform one codon into another which codes for the same amino acid) from the non-synonymous mutations. The latter are either slowed down or favoured by selection, while the former accumulate at the rate of the mutations. In general, the ratio of synonymous to non-synonymous mutations is greater than 1.

Goldman and Yang (1994) proposed a model with 63 parameters, namely 60 for the frequencies π_{xyz}, the remaining three being for the transition rate α, the transversion rate β, and the ratio of synonymous to non-synonymous mutations ω. The GY model can be written

$$Q_{(x_1y_1z_1)(x_2y_2z_2)} = \begin{cases} 0, & \text{if 1 and 2 differ by more than one base,} \\ \alpha\pi_{x_2y_2z_2}, & \text{for a synonymous transition,} \\ \beta\pi_{x_2y_2z_2}, & \text{for a synonymous transversion,} \\ \omega\alpha\pi_{x_2y_2z_2}, & \text{for a non-synonymous transition,} \\ \omega\beta\pi_{x_2y_2z_2}, & \text{for a non-synonymous transversion.} \end{cases}$$

Note that among the 63 parameters to be estimated, the 60 which determine the invariant probability distribution π are usually estimated not by a maximum likelihood procedure, but from the empirical frequencies of the various codons present in our data. Another possibility is to estimate π_{xyz} by the product $\pi_x^1\pi_y^2\pi_z^3$ of the frequencies of the various nucleotides at positions 1, 2 and 3 of the codons.

Non-homogeneous models An implicit assumption in all the Markovian models we have considered so far is their stationarity. The infinitesimal generator is the same on all the branches of the phylogenetic tree. Hence, the invariant probability distribution is the same on the various branches, which implies that the various sequences must have roughly the same composition in bases. Some data contradict this assumption. One can then relax the homogeneity assumption of the Markov process on the whole tree. For example, Galtier and Gouy (1998) adopt the Tamura model, with parameters α and β which are constant on the tree, and a parameter θ (which regulates the proportion of G + C) which is allowed to vary from one branch to another.

Dependence or independence between sites Almost all Markovian models assume that mutations at the various sites are mutually independent. This assumption is of course not reasonable, but it makes the computations (in particular of the likelihood, see below) feasible. There has so far been very little work on Markovian models where the evolutions at the various sites are correlated.

Consider a model of the type

$$Q_{xy} = s_{xy}\pi_y,$$

which is a reversible model provided that $s_{xy} = s_{yx}$. Pollock, Taylor and Goldman (1999) model the evolution of a pair of proteins by choosing an infinitesimal generator of the form

$$\begin{aligned} Q_{xx',yx'} &= s_{xy}\overline{\pi}_{yx'}, \\ Q_{xx',xy'} &= s_{x'y'}\overline{\pi}_{xy'}, \\ Q_{xx',yy'} &= 0, \quad \text{if } x \neq y \text{ and } x' \neq y', \end{aligned}$$

where $\overline{\pi}$ is an invariant probability distribution on the set of pairs of proteins.

Variation of the evolution rate between the branches Given an infinitesimal generator Q, for all $u > 0$, uQ is also an infinitesimal generator. Suppose that Q is constant on the tree. If u is constant as well, since the leaves (the species living today) are all at the same distance from the common ancestor, located at the root of the tree (distances are measured in elapsed time), this means that we are making the assumption of a 'molecular clock'. Certain sets of data are incompatible with such an assumption. One should then, in order to use a model which is coherent with such data, allow the parameter u to take a different value on each branch of the tree. We then have a new parameter on each branch of the tree, which all together means a lot of parameters.

Another point of view involves assuming that the u's are the values taken by a stochastic process, which evolves on the tree as a Markov process, either in continuous time, or else in discrete time (in which case the process is constant on each branch, the transitions taking place at the nodes). Conditionally upon the values taken by that process, the various nucleotides evolve as a non-homogeneous Markov process along the tree. We then are in a Bayesian framework, for which the MCMC algorithm (see Sections 3.1 and 7.8.3) makes the required simulations feasible.

Variation of the evolution rate between sites The most popular model for the variation of rate between sites is to assume that the rates associated to the various sites are i.i.d., the common distribution being a gamma distribution (or a discretized version of the same).

Another approach, due to Felsenstein and Churchill (1996), assumes that they form a Markov chain along the DNA sequence (which is 'hidden'), taking its values in a set which, for practical reasons, is taken to be of very small cardinality.

'Covarion' models In the concomitantly variable codon or covarion model the rate of evolution not only differs from one site to another, but also, at a given site, from one branch of the tree to another. Write $E = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and let G be the set of all possible values of the rate u. Galtier (2001) considers an independent $E \times G$-valued Markov process at each site.

7.8.2 Likelihood methods in phylogeny

Comparison of the genomes of the various species is at present the main tool for the reconstruction of phylogenetic trees. Several algorithms exist for the construction of such trees. We shall now give some comments concerning the maximum likelihood method.

Note that we can compare either genes (i.e. amino acid sequences) or DNA sequences. In order to fix ideas, we shall consider DNA sequences, which we assume to be already aligned.

Computation of the likelihood of a tree Suppose we use the Felsenstein (F84) model. Time t corresponds here to distance along the tree. Note that the only

parameter of interest is the product $u \times t$. If we modify the lengths of the branches of the tree accordingly, we can and shall from now on assume that $u = 1$. We shall consider below only *binary trees*.

For the remainder of this section, we shall assume that each of the various sites evolves independently of the others, and that all evolve at the same rate, this rate being constant along the tree. This assumption is not very realistic, and several recent works concentrate on the detection of those sites which evolve faster than others, possibly only on a portion of the tree. However, this simplifying assumption is natural for starting the analysis and constructing a first tree. Another popular assumption would be that the evolution rates of the various sites are i.i.d. random variables, with a common law gamma.

The information at our disposal, the *data*, consists of a set of k aligned sequences of length m; that is, for each site s, $1 \leq s \leq m$, we have k letters from the alphabet A, C, G, T, one for each leaf of the tree. With each rooted binary tree T with k leaves, we shall associate its likelihood $L(T)$, which is a function of the data. The likelihood $L(T)$ is a product from $s = 1$ to m of the likelihoods associated with each site s:

$$L(T) = \prod_{s=1}^{m} L_s(T).$$

The computation of each factor $L_s(T)$ takes advantage of the Markov property, as we shall now see. Let T denote a rooted tree. We can, for example, code as follows the nodes of such a tree, starting from the root, towards the leaves (see Figure 7.4):

- 0 denotes the root;
- 1, 2 are the 'sons' of the root, that is, the nodes which are directly connected with the root by one branch;
- 1.1, 1.2 denote the sons of 1; 2.1, 2.2 those of 2;
- and so on up to the leaves.

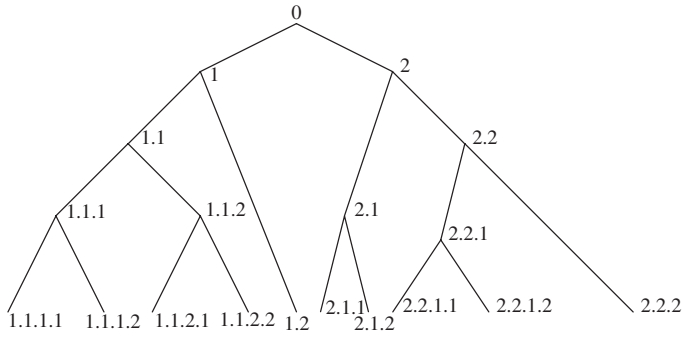


Figure 7.4 Rooted binary tree with coded nodes.

For each node $\alpha \in T\backslash\{0\}$, denote by ℓ_α the length of the branch joining α and the 'father' of α. We associate with α the set Λ_α of the leaves of the subtree whose root is α. In particular, Λ_0 denotes the set of the leaves of the tree. If $\alpha \in \Lambda_0$, $\Lambda_\alpha = \{\alpha\}$. For $\alpha \in T\backslash\Lambda_0$, we denote by $\Gamma_\alpha = \{\alpha.1, \alpha.2\}$ the two 'sons' of α.

Let $\{X_\alpha;\ \alpha \in T\}$ denote the nucleotides at the nodes of the tree. We assume that they are the values at these nodes of a Markov process on the tree whose infinitesimal generator is Q. Only the values of $\{X_\alpha;\ \alpha \in \Lambda_0\}$ are observed. We denote by x_α the observed value of X_α, for $\alpha \in \Lambda_0$. The likelihood of the tree, based upon the nucleotides at site s, is

$$L_s(T) = \mathbb{P}_T\left(\bigcap_{\alpha\in\Lambda_0}\{X_\alpha = x_\alpha\}\right).$$

We shall explain how to compute this quantity, and we shall show how it depends upon the tree T.

For each $\alpha \in T$, $x \in E$, we define $L_{s,x}^{(\alpha)}$, the conditional likelihood of the subtree whose α is the root, conditioned upon $X_\alpha = x$, which we compute by the following upward recurrence: for $\alpha \in \Lambda_0$,

$$L_{s,x}^{(\alpha)} = \begin{cases} 1, & \text{if } x = x_\alpha, \\ 0, & \text{otherwise;} \end{cases}$$

in all other cases,

$$L_{s,x}^{(\alpha)} = \sum_{x_{\alpha.1},x_{\alpha.2}\in E} P_{xx_{\alpha.1}}(\ell_{\alpha.1})L_{s,x_{\alpha.1}}^{(\alpha.1)} \times P_{xx_{\alpha.2}}(\ell_{\alpha.2})L_{s,x_{\alpha.2}}^{(\alpha.2)}.$$

This computation eventually specifies the quantities $L_{s,x}^{(0)}$, $x \in E$. Finally,

$$L_s(T) = \sum_{x\in E} \pi_x L_{s,x}^{(0)}$$

and

$$L(T) = \prod_{s=1}^{m} L_s(T).$$

We could also have described each $L_s(T)$ as a sum of $4^{|T\backslash\Lambda_0|}$ terms. But the above formulae describe the so-called pruning algorithm, due to Felsenstein, which should be used in practice.

Maximum likelihood The computation of a global maximum of the likelihood over all possible trees is complex. The easiest part is the maximization over branch lengths. However, it is not clear that the algorithm commonly used for that subproblem leads to a global maximum. The idea is to maximize successively over each branch length, and to iterate as long as the likelihood increases. We shall now

see explicitly how the likelihood depends upon a particular branch length. It is then easy to see how to maximize the likelihood with respect to that particular branch length.

We shall assume that the values $\{X_\alpha;\ \alpha \in T\}$ are the values at the nodes of the tree of a *reversible* Markov process. Hence, the law of the $\{X_\alpha\}$ does not depend upon the choice of of a root at any node of the tree (or, more generally, anywhere on the tree).

Consider two neighbouring nodes α and β in the tree. Denote by $\ell_{\alpha\beta}$ the length of the branch which connects them. If we put the root anywhere on that branch, we define as above the quantities $L^{(\alpha)}_{s,x}$ and $L^{(\beta)}_{s,y}$, $x, y \in E$. Note that the first term is the likelihood of a subtree whose root is α, taken as a subtree whose root has been displaced. For example, in the tree of Figure 7.4, if $\alpha = 2$, $\beta = 2.2$, and we locate the root between α and β, the subtree whose root is α contains the nodes 2, 0, 1, 1.1, 1.1.1, 1.1.1.1, 1.1.1.2, 1.1.2, 1.1.2.1, 1.1.2.2, 1.2, 2.1, 2.1.1, 2.1.2, and the subtree whose root is β contains the nodes 2.2.1, 2.2.1.1, 2.2.1.2, 2.2.2.

Then

$$\begin{aligned} L_s(T) &= \sum_{x,y\in E} \pi_x P_{xy}(\ell_{\alpha\beta}) L^{(\alpha)}_{s,x} L^{(\beta)}_{s,y} \\ &= \sum_{x,y\in E} \pi_y P_{yx}(\ell_{\alpha\beta}) L^{(\alpha)}_{s,x} L^{(\beta)}_{s,y}. \end{aligned}$$

This formula makes explicit the dependence of $L_s(T)$ and $L(T)$ upon the length of a given branch, and allows us to compute the maximum of the likelihood with respect to that branch. The search for that maximum is rather easy in the case of the evolution model described above (one maximizes the logarithm of $L(T)$, which replaces the product of the $L_s(T)$ by a sum, and thus simplifies the maximization).

Remark 8.1 *Not all evolution models are reversible. It is still possible to give explicitly the dependence of the likelihood with respect to the length of a given branch, but one has to be careful using the transition probability of the time-reversed process whenever the displacement of the root makes the process starting at the new root run along a branch in the direction opposite to that of the initial one.*

7.8.3 The Bayesian approach to phylogeny

Let us return to the expression for the likelihood. Denote by D the vector of the observed random variables, and by d the vector of the observed values; that is, d contains the various aligned sequences. We now describe the various parameters upon which the likelihood depends. Among the unknown parameters (which we wish to estimate), we have:

- on the one hand, the shape of the tree, which we shall denote by τ, which is an unknown in a finite set $\mathcal{T}$ (whose cardinality is $(2n-3)!!$ in the case of a rooted tree with n leaves and $(2n-5)!!$ in the case of an unrooted tree, where $k!! = 1 \times 3 \times 5 \times \cdots \times k$ for an odd number k);

- on the other hand the lengths of the various branches, and the infinitesimal generator Q of the evolution model (at least the parameters of that matrix other than the invariant probability distribution). The branch lengths and the unknown parameters of the matrix Q vary in a subset V of a Euclidian space $\mathbb{R}^d$. We shall denote this set of parameters by λ.

Thus the unknown parameter is the pair $\theta = (\tau, \lambda)$ whose value is arbitrary in the set $\Theta = \mathcal{T} \times V$, and the likelihood is the function

$$L(\theta) = \mathbb{P}_\theta(D = d).$$

The likelihood of the value θ of the unknown parameter is the probability of observing the data which we have in our computer, if θ is the true value of that parameter.

In the Bayesian framework, the unknown parameter θ is the realization of a random variable, that is, (τ, λ) is the realization of a random vector (T, Λ). This point of view forces us to choose an *a priori probability distribution*, which gives us the chance to incorporate a priori information about the unknown parameter, which the anti-Bayesians refuse to do, claiming that the only information one should use is that contained in the data.

We shall thus obtain an a priori probability distribution for the random vector (T, Λ) as follows:

- We specify the law of T, which is a probability distribution on the finite set $\mathcal{T}$, hence we specify the $\alpha_\tau = \mathbb{P}(T = \tau)$, $\tau \in \mathcal{T}$.

- We specify the conditional law of Λ, given T, and we assume that, for all $\tau \in \mathcal{T}$, the conditional law of Λ, given that $T = \tau$, has a density $q_\tau(\lambda)$; in other words, for any Borel measurable function $f : \mathcal{T} \times V \to \mathbb{R}_+$,

$$\mathbb{E}[f(T, \Lambda)] = \sum_{\tau \in \mathcal{T}} \int_V f(\tau, \lambda) p_\tau(\lambda) d\lambda,$$

  where $p_\tau(\lambda) = \alpha_\tau \times q_\tau(\lambda)$.

In this context, we have a random pair, consisting of both a 'parameter' (T, Λ), and the data D. The law of this pair is specified by, on the one hand, the a priori law of (T, Λ), and on the other hand, the conditional law of the data, given the parameter. More precisely, in this Bayesian framework, the likelihood is interpreted as the conditional probability distribution of the data, given the parameters:

$$L(\tau, \lambda) = \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right).$$

The rule of the game is to compute the *a posteriori* law of the parameter, which is the conditional law of the 'parameter' (T, Λ), given the data, that is, given that $D = d$. This conditional probability distribution is given by the famous 'Bayes formula', which in our case specifies the joint law of (T, Λ) given that $D = d$ in the form

$$p_\tau(\lambda | D = d) = \frac{\mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda)}{\sum_{\tau \in \mathcal{T}} \int_V \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda) d\lambda}.$$

In other words, again if $f : \mathcal{T} \times V \to \mathbb{R}_+$,

$$\mathbb{E}\left[f(T, \Lambda)|D = d\right] = \frac{\sum_{\tau\in\mathcal{T}}\int_V f(\tau, \lambda)\mathbb{P}\left(D = d|(T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda)d\lambda}{\sum_{\tau\in\mathcal{T}}\int_V \mathbb{P}\left(D = d|(T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda)d\lambda}.$$

For example, we might wish to specify the *a posteriori* probability distribution of the shape of the tree, that is, of the random variable T. It is given, for all $\tau \in \mathcal{T}$, by

$$\mathbb{P}\left(T = \tau|D = d\right) = \frac{\int_V \mathbb{P}\left(D = d|(T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda)d\lambda}{\sum_{\tau\in\mathcal{T}}\int_V \mathbb{P}\left(D = d|(T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda)d\lambda}.$$

The MCMC algorithm Suppose that we wish to compute the last quantity for a small number of values of τ. An explicit computation is pointless, because of the size of the data (the number of species we consider) and the complexity of the models we might use. One is thus led to use a Monte Carlo type of method, using random draws. However, it is not really possible to simulate under the a posteriori probability distribution of (T, Λ), given the data. Indeed, in order to identify this probability distribution, it would be necessary to compute the denominator in the formula above. If the cardinality of $\mathcal{T}$ is very large, this task becomes impossible. We are exactly in the situation described at the beginning of Section 3.1.

Let us recall the Metropolis–Hastings algorithm. Denote by π the a posteriori probability distribution. Let Q denote a transition matrix on F (which has nothing to do with the probability distribution π), whose transitions are easy to simulate. We choose as transition matrix P the matrix whose off-diagonal entries are given by

$$P_{xy} = \min\left(Q_{xy}, \frac{\pi_y}{\pi_x}Q_{yx}\right),$$

and whose diagonal entries are

$$P_{xx} = 1 - \sum_{y\neq x} P_{xy},$$

provided the matrix P thus defined is irreducible, which is the case if, for instance, Q is irreducible and satisfies the property that, for all x, y, $Q_{xy} > 0 \Leftrightarrow Q_{yx} > 0$. This matrix P is clearly a transition matrix ($P_{xy} \leq Q_{xy}$, $x \neq y$, implies that $P_{xx} \geq 0$), and π is P-invariant, since the detailed balance equation

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x \neq y,$$

holds. For $x, y \in F$, let

$$r(x, y) = \frac{P_{xy}}{Q_{xy}} = \min\left(1, \frac{\pi_y Q_{yx}}{\pi_x Q_{xy}}\right).$$

One way of simulating a transition of the chain $\{X_k\}$ with the transition matrix P is as follows. Suppose that $X_k = x$, and we wish to simulate X_{k+1}. We first simulate

a transition of the chain $\{Y_k\}$ with the transition matrix Q, starting from $Y_k = x$. Suppose that the result of this simulation is $Y_{k+1} = y$. We *accept this transition* (and in this case $X_{k+1} = y$) with probability $r(x, y)$; we *reject this transition* (and in this case $X_{k+1} = x$) with probability $1 - r(x, y)$. Note that $r(x, x) = 1$, hence whenever $y = x$, $X_{k+1} = x$.

In other words, the transition from $X_k = x$ to X_{k+1} is computed as follows:

- we draw a realization Y_{k+1} of the probability distribution $Q_{x\cdot}$;
- we draw U_{k+1} with the uniform law on [0, 1];

and we let

$$X_{k+1} = Y_{k+1}\mathbf{1}_{\{U_{k+1} \le r(x, Y_{k+1})\}} + X_k\mathbf{1}_{\{U_{k+1} > r(x, Y_{k+1})\}}.$$

Implementation of the MCMC algorithm The implementation of the MCMC algorithm poses delicate questions, for which we essentially have no satisfactory answers, in particular in the application to phylogeny. We have already discussed this issue in a general framework in Section 3.3. Recall that one should eliminate the first simulations (*burn-in*). Moreover, in order to obtain a sample of the a posteriori law, one keeps only one iteration among n, where the choice of n depends upon the rate of decorrelation of the chain, which might be estimated from simulations. Some implementations involve the simulation of several chains in parallel, some of them possibly being 'heated' (see Section 3.1.4).

7.9 Application to discretized partial differential equations

Let D be a bounded domain in $\mathbb{R}^2$ (we could also treat a problem in higher dimensions) whose boundary ∂D is Lipschitz continuous. Suppose that $0 \in D$. Consider the Dirichlet problem

$$\begin{cases} \Delta u(x) = 0, & x \in D, \\ u(x) = f(x), & x \in \partial D, \end{cases}$$

where $f \in C(\partial D)$. It is well known that this equation has a unique solution u in $C(D)$.

Given $h>0$, let $h\mathbb{Z}^2$ denote the set of points in the plane whose coordinates are multiples of h. Define $D_h = D \cap h\mathbb{Z}^2$. ∂D_h consists of those points in $D^c \cap h\mathbb{Z}^2$ which are at distance h from at least one point in D_h, and $\bar{D}_h = D_h \cup \partial D_h$. Let e_1 and e_2 be the two vectors of an orthonormal basis. We define the approximate operator Δ_h as

$$(\Delta_h v)(x) = \frac{1}{4}\sum_{i=1}^{2}(v(x + he_i) + v(x - he_i)) - v(x).$$

From Exercises 11.4 below, the solution of the discretized Dirichlet problem

$$\begin{cases} \Delta_h u_h(x) = 0, & x \in D_h, \\ u_h(x) = f(x), & x \in \partial D_h, \end{cases}$$

is given by the formula

$$u_h(x) = \mathbb{E}[f(X^h_{T^h_{D_h^c}})|X^h_0 = x],$$

where $\{X^h_t; t \geq 0\}$ is an $h\mathbb{Z}^2$-valued jump Markov process with infinitesimal generator $\frac{1}{2}\Delta_h$, and

$$T^h_{D_h^c} = \inf\{t \geq 0; X^h_t \in D_h^c\}.$$

Note that $\{X^h_t; t \geq 0\}$ has the same law as $\{hX^1_{h^{-2}t}; t \geq 0\}$, and we explained just after Theorem 6.9, that this process converges towards a standard one-dimensional Brownian motion, as $h \to 0$. It is not too hard to deduce that

$$u_h(x) \to u(x) = \mathbb{E}[f(B_{T_{D^c}})|B_0 = x],$$

where $T_{D^c} = \inf\{t \geq 0;\ B_t \in D^c\}$. This formula gives a probabilistic interpretation of the Dirichlet problem. We then have a proof (which is an alternative to classical arguments from numerical analysis) of the convergence of u_h towards u.

Note that the discretized Dirichlet problem could also be interpreted in terms of a discrete time Markov chain with transition matrix $\Delta_h + I$.

Such a probabilistic interpretation justifies the use of Monte Carlo numerical methods for the approximate computation of solutions of partial differential equations. These methods are mainly used when 'classical' numerical analysis algorithms cannot be used (in particular, in high-dimensional problems); see [25]. They are also very popular because of the simplicity with which the associated programs can be written, a few lines of code being sufficient to program the computation of the approximate solution of a partial differential equation! Even if we need to let the computer run a bit longer than would be needed to obtain the same precision with a finite difference, finite element or finite volume method, the fact that the program is very easy to write is greatly appreciated by users, especially in situations where a program written by one person might have to be modified later by another.

7.10 Simulated annealing

In this section, which follows on from Section 3.4, E is assumed to be finite. Recall that we wish to maximize a function

$$U : E \to \mathbb{R}_-,$$

such that

$$\max_{x\in E} U_x = 0.$$

In other words, we are looking for an x which is such that $U_x = 0$.

With each $\beta > 0$ we associate the infinitesimal generator $Q = \{Q_{xy};\ x, y \in E\}$ where, for $x \neq y$,

$$Q_{xy} = \mathbf{1}_{\{(x,y)\in G\}} \exp\left[\frac{\beta}{2}(U_y - U_x)\right].$$

Let G be a non-oriented graph in E, that is, a collection of pairs of points in E, chosen in such a way that the jump Markov process with infinitesimal generator Q is irreducible (which means that, for all $x, y \in E$, there exist n and $x_1, x_2, \ldots, x_n$ such that $(x, x_1) \in G$, $(x_1, x_2) \in G, \ldots, (x_n, y) \in G$).

The jump Markov process with infinitesimal generator Q is clearly reversible with respect to its invariant probability distribution π_β defined by

$$\pi_{\beta,x} = Z_\beta^{-1} e^{\beta U_x}, \quad x \in E.$$

We define the *Dirichlet form* associated with Q as the bilinear form on $\mathbb{R}^E$:

$$\begin{aligned}\mathcal{E}(\varphi, \varphi) &= \langle \varphi, -Q\varphi\rangle_\pi \\ &= -\sum_{x,y} \varphi_x Q_{xy} \varphi_y \pi_x \\ &= \frac{1}{2}\sum_{x,y} |\varphi_x - \varphi_y|^2 Q_{xy}\pi_x,\end{aligned}$$

where we have exploited reversibility and (twice) the identity $\sum_y Q_{xy} = 0$, hence $-Q : \ell^2(\pi) \to \ell^2(\pi)$ is a self-adjoint positive semi-definite operator.

Definition 10.1 *We call the quantity*

$$\lambda \stackrel{def}{=} \inf_{\varphi \text{ non-constant}} \frac{\mathcal{E}(\varphi, \varphi)}{\mathrm{var}_\pi(\varphi)}$$

the spectral gap of Q, where

$$\underset{\pi}{\mathrm{var}}(\varphi) = \sum_{x\in E} \varphi_x^2 \pi_x - \left(\sum_{x\in E} \varphi_x \pi_x\right)^2.$$

Lemma 10.2 *Since Q is the infinitesimal generator of an irreducible jump Markov process with values in a finite set E, its spectral gap is strictly positive.*

Proof From the above formula for $\mathcal{E}(\varphi, \varphi)$, the ratio $\mathcal{E}(\varphi, \varphi)/\mathrm{var}_\pi(\varphi)$ is not modified if we add a constant to φ. We can then minimize this ratio over those φ that are such that $\mathbb{E}_\pi(\varphi) = 0$, whence

$$\lambda = \inf_{\varphi\neq 0;\, \mathbb{E}_\pi(\varphi)=0} \frac{\langle \varphi, -Q\varphi\rangle\pi}{\langle \varphi, \varphi_\pi\rangle},$$

and λ is the smallest eigenvalue of $-Q$, considered as a linear operator on $\ell^2(\pi)$, restristed to the subvector space orthogonal to the constants. Since Q is a self-adjoint positive semi-definite operator, it suffices to show that the eigenspace associated with the eigenvalue 0 is the set of constant functions. But if φ belongs to that eigenspace,

$$\sum_y Q_{xy}\varphi_y = 0, \quad \forall x \in E,$$

then a fortiori

$$0 = -\sum_{x,y} \varphi_x Q_{xy}\varphi_y \pi_x = \frac{1}{2}\sum_{x,y} |\varphi_x - \varphi_y|^2 Q_{xy}\pi_x,$$

which, since Q is irreducible, does imply that φ is constant. □

Now let $\{X_t; t \geq 0\}$ be a jump Markov process with infinitesimal generator Q. For each $t > 0$, let $\mu(t) = (\mu_x(t); x \in E)$ denote the law of X_t. We let

$$\varepsilon(t) = \sum_{x\in E} \left(\frac{\mu_x(t)}{\pi_x} - 1\right)^2 \pi_x,$$

and remark that $\varepsilon(t) = 0$ if and only if $\mu(t) = \pi$.

Lemma 10.3 *If λ denotes the spectral gap of Q, then*

$$\varepsilon(t) \leq \varepsilon(0)e^{-2\lambda t}.$$

PROOF We first remark that

$$\begin{aligned}\varepsilon(t) &= \sum_x \left(\frac{\mu_x(t)}{\pi_x} - 1\right)^2 \pi_x \\ &= \sum_x \left(\frac{\mu_x(t)}{\pi_x}\right)^2 \pi_x - 1.\end{aligned}$$

Hence,

$$\begin{aligned}\frac{d\varepsilon}{dt}(t) &= 2\sum_x \frac{\mu_x(t)\mu_x'(t)}{\pi_x} \\ &= 2\sum_{x,y} \frac{\mu_x(t)\mu_y(t)Q_{yx}}{\pi_x} \\ &= 2\sum_{x,y} \frac{\mu_x(t)}{\pi_x} \times \frac{\mu_y(t)}{\pi_y} \times \pi_y Q_{yx} \\ &= 2\left\langle Q\left(\frac{\mu(t)}{\pi}\right), \frac{\mu(t)}{\pi}\right\rangle_\pi\end{aligned}$$

$$\leq -2\lambda \operatorname*{var}_{\pi}\left(\frac{\mu(t)}{\pi}\right)$$
$$= -2\lambda\varepsilon(t),$$

and consequently

$$\frac{d}{dt}\log\varepsilon(t) \leq -2\lambda.$$

□

We have shown that $\mu_t \to \pi$ at exponential rate (cf. Theorem 6.8 of Chapter 2).

We now come to 'annealing'. Let β depend upon t, and let it go to infinity (and hence the 'temperature', its inverse, towards zero) as $t \to \infty$. More precisely, Δ being a constant which will be specified below, we choose

$$\beta(t) = \frac{1}{\Delta}\log(1+t),$$

hence $\beta(0) = 0$ and $\beta(t) \to +\infty$ as $t \to \infty$. Of course, the chain is no longer time-homogeneous, since the infinitesimal generator Q, the spectral gap λ, the invariant measure π, and the normalization constant Z all become functions of t: $Q(t)$, $\lambda(t)$, $\pi(t)$, $Z(t)$. Note that $\pi(0)$ is the uniform measure on E, $Z(0) = |E|^{-1}$, while $\pi(\infty) = \lim_{t\to\infty}\pi(t)$ is the uniform measure on the zeros of U (i.e. on the maxima of U).

Let $M \stackrel{\text{def}}{=} \sup_{x\in E}(-U_x)$. Our goal is to prove the following theorem.

Theorem 10.4 *If $\Delta > M$, then $\mu(t) \to \pi(\infty)$ as $t \to \infty$.*

We first establish the following lemma.

Lemma 10.5

$$\lambda(t) \geq \lambda(0)\left(\frac{1}{1+t}\right)^{M/\Delta}.$$

PROOF Choose φ such that $\mathbb{E}_{\pi(0)}[\varphi] = 0$. Then from the definition of $\lambda(0)$,

$$\frac{1}{2}\sum_{x,y}|\varphi_x - \varphi_y|^2 Q_{xy}(0)\pi_x(0) \geq \lambda(0)\sum_x \varphi_x^2\pi_x(0).$$

On the other hand,

$$Q_{xy}(t)\pi_x(t) = Q_{xy}(0)\frac{e^{\beta(t)(U_x+U_y)/2}}{Z(t)}$$
$$\geq Q_{xy}(0)\frac{e^{-\beta(t)M}|E|}{Z(t)}\pi_x(0).$$

Hence,

$$\begin{aligned}
\mathcal{E}_t(\varphi,\varphi) &= \frac{1}{2}\sum_{x,y}|\varphi_x-\varphi_y|^2 Q_{xy}(t)\pi_x(t)\\
&\geq \frac{e^{-\beta(t)M}|E|}{2Z(t)}\sum_{x,y}|\varphi_x-\varphi_y|^2 Q_{xy}(0)\pi_x(0)\\
&\geq \lambda(0)e^{-\beta(t)M}\sum_x \varphi_x^2\frac{1}{Z(t)}\\
&\geq \lambda(0)e^{-\beta(t)M}\sum_x \varphi_x^2\pi_x(t)\\
&\geq \lambda(0)e^{-\beta(t)M}\operatorname*{var}_{\pi(t)}(\varphi),
\end{aligned}$$

where we have used for the third inequality the fact that $U_x \leq 0$, hence $Z^{-1}(t) \leq \pi_x(t)$. The two extreme expressions of this set of inequalities being invariant under the addition of a constant to φ, the resulting inequality still holds without the restriction that $\mathbb{E}_{\pi(0)}[\varphi]=0$. The lemma is established, since

$$e^{-\beta(t)M} = \left(\frac{1}{1+t}\right)^{M/\Delta}.$$

□

Proof of Theorem 10.4 It suffices to show that $\varepsilon(t)\to 0$, where

$$\begin{aligned}
\varepsilon(t) &= \sum_x\left(\frac{\mu_x(t)}{\pi_x(t)}-1\right)^2\pi_x(t)\\
&= \sum_x\frac{\mu_x(t)^2}{\pi_x(t)}-1.
\end{aligned}$$

Note that $\varepsilon(t)$ is an upper bound on the square of the L^1-norm of the difference $\mu(t)-\pi(t)$. Indeed, from Cauchy–Schwarz,

$$\sum_x|\mu_x(t)-\pi_x(t)| = \sum_x\frac{|\mu_x(t)-\pi_x(t)|}{\sqrt{\pi_x(t)}}\sqrt{\pi_x(t)} \leq \sqrt{\varepsilon(t)}.$$

We have that

$$\begin{aligned}
\frac{d\varepsilon}{dt}(t) &= \sum_x\frac{d}{dt}\left[\frac{\mu_x(t)^2}{\pi_x(t)}\right]\\
&= \sum_x\pi_x(t)^{-2}\left[2\mu_x(t)\frac{d\mu_x}{dt}(t)\pi_x(t)-\mu_x(t)^2\frac{d\pi_x}{dt}(t)\right]\\
&= -2\mathcal{E}_t\left(\frac{\mu(t)}{\pi(t)},\frac{\mu(t)}{\pi(t)}\right)-\beta'(t)\sum_x U_x\frac{\mu_x^2(t)}{\pi_x^2(t)}\pi_x(t)
\end{aligned}$$

$$
\begin{aligned}
&+ \beta'(t) \sum_{x,y} \frac{U_y e^{\beta(t)U_y}}{Z^2(t)} \times \frac{\mu_x^2(t)}{\pi_x^2(t)} e^{\beta(t)U_x} \\
&\leq -2\lambda(t)\varepsilon(t) + \beta'(t)M(\varepsilon(t)+1) \\
&\leq -(2\lambda(t) - M\beta'(t))\varepsilon(t) + M\beta'(t) \\
&\leq -\left(\frac{2\lambda(0)}{(1+t)^{M/\Delta}} - \frac{M}{\Delta(1+t)}\right)\varepsilon(t) + \frac{M}{\Delta(1+t)}
\end{aligned}
$$

where the last inequality uses Lemma 10.5. Since $\Delta > M$, $(1+t)^{-M/\Delta} \gg (1+t)^{-1}$ as $t \to \infty$, and there exists $c > 0$ such that, for t large enough,

$$
\frac{d\varepsilon}{dt}(t) \leq -c(1+t)^{-M/\Delta}\varepsilon(t) + \frac{M}{\Delta}(1+t)^{-1},
$$

and the theorem follows from the next lemma. □

Lemma 10.6 *Let $x, b \in C^1(\mathbb{R}_+; \mathbb{R}_+)$, $a \in C(\mathbb{R}_+; \mathbb{R}_+)$ be such that:*

(i) $\displaystyle\int_0^\infty a(t)\,dt = +\infty$;

(ii) $b(t) \searrow 0$ *as* $t \to \infty$;

(iii) $\displaystyle\frac{dx}{dt}(t) \leq -a(t)(x(t) - b(t))$.

Then $x(t) \to 0$ *as* $t \to \infty$.

PROOF

$$
\begin{aligned}
\frac{d}{dt}\left(x(t)\exp\left(\int_0^t a(s)ds\right)\right) &= e^{\int_0^t a(s)ds}\left(\frac{dx}{dt}(t) + a(t)x(t)\right) \\
&\leq e^{\int_0^t a(s)ds} a(t)b(t).
\end{aligned}
$$

Integrating, we obtain

$$
x(t) \leq x(0)e^{-\int_0^t a(s)ds} + \int_0^t e^{-\int_s^t a(r)dr} a(s)b(s)ds.
$$

The right-hand side of this inequality is the solution of a linear ordinary differential equation, which majorizes $x(t)$. Hence, it suffices to establish the result with $x(t)$ as solution of the ODE; in other words, we can also assume that

$$
\frac{dx}{dt}(t) = -a(t)(x(t) - b(t)).
$$

Let $y(t) = x(t) - b(t)$. It suffices to show that $y(t) \to 0$ as $t \to \infty$. We have that

$$
\frac{dy}{dt}(t) = -a(t)y(t) - b'(t).
$$

Note that $b'(t) \le 0$, and $-\int_t^\infty b'(s)ds = b(t) < \infty$. Hence, for $t \ge N$,

$$y(t) = e^{-\int_0^t a(s)ds} y(0) - \int_0^t e^{-\int_s^t a(r)dr} b'(s)ds$$

$$\le e^{-\int_0^N a(s)ds} y(0) + \int_N^\infty |b'(s)|ds + e^{-\int_N^t a(r)dr} \int_0^N e^{-\int_s^N a(r)dr} |b'(s)|ds.$$

Let $\delta > 0$ be arbitrary. We choose N large enough that the sum of the first two terms on the right-hand side is less than $\delta/2$. Now choosing t large enough, the third term is less than $\delta/2$. The lemma is established. □

Remark 10.7 *The function $\beta(t) = \Delta^{-1}\log(1+t)$ tends too slowly to infinity as $t \to \infty$ to be used in practice. One can prove some results which are weaker than $\mu(t) \to \pi(\infty)$ with a function β which grows faster than a logarithm (a power function). If, on the other hand, we ask how to achieve the best possible result on a fixed finite horizon, it can be shown that some β's growing at exponential rate are close to the optimum.*

7.11 Exercises

Exercise 11.1 *Let $\{T_n;\ n \ge 1\}$ be a Poisson point process with intensity λ, and $\{Z_n;\ n \ge 0\}$ be an E-valued Markov chain which is independent of $\{T_n;\ n \ge 1\}$, with transition matrix P_{xy}, $x, y \in E$. Let*

$$X_t = \sum_{n=0}^{\infty} Z_n \mathbf{1}_{[T_n, T_{n+1}[}(t), \quad t \ge 0.$$

Show that $\{X_t; t \ge 0\}$ is a jump Markov process, and give its transition matrices, its infinitesimal generator, and the law of its first jump time.

Exercise 11.2 *Let $\{X_t; t \ge 0\}$ be a jump Markov process with values in a finite or countable state space E, with infinitesimal generator $\{Q_{xy};\ x, y \in E\}$. Assume that $\lambda := \sup_x -Q_{xx} < \infty$. Let $\{N_t;\ t \ge 0\}$ denote the counting process of the jumps of $\{X_t\}$, and $\{N'_t;\ t \ge 0\}$ a Poisson process with intensity λ. Compare $\mathbb{P}(N_t \ge n)$ and $\mathbb{P}(N'_t \ge n)$, $\mathbb{E}[f(N_t)]$ and $\mathbb{E}[f(N'_t)]$, where the function f is increasing from $\mathbb{N}$ into $\mathbb{R}$. Show that Exercises 11.1 gives another proof of this result.*

Exercise 11.3 *Let $\{N_t;\ t \ge 0\}$ and $\{P_t;\ t \ge 0\}$ be two mutually independent Poisson processes with intensities λ and μ, respectively.*

1. *Show that $\{X_t;\ t \ge 0\}$, defined by*

$$X_t = N_t - P_t,$$

is a $\mathbb{Z}$-valued irreducible jump Markov process, and give its infinitesimal generator.

2. *Assume that $\lambda \neq \mu$. Show that $\{X_t/t\}$ and $\{X_t\}$ converge almost surely in $\bar{\mathbb{R}}$ as $t \to \infty$. What is the limit of $\{X_t\}$, depending upon the sign of $\lambda - \mu$? Show that $\{X_t\}$ is transient.*

3. *Assume that $\lambda = \mu$. Give the transition matrix of the embedded chain. Deduce from Exercises 10.11 and 10.13 of Chapter 2 that $\{X_t\}$ is null recurrent.*

Exercise 11.4 *1. Let $\{X_t; t \geq 0\}$ be an E-valued jump Markov process, with infinitesimal generator $\{Q_{xy};\ x, y \in E\}$. Let $F \subset E$. Define*

$$T_F = \begin{cases} \inf\{t;\ X_t \in F\}, & \text{if such } t \text{ exists;} \\ \infty, & \text{otherwise,} \end{cases}$$

the function $u : E \to \mathbb{R}$ by

$$u(x) = \mathbb{E}[h(X_{T_F})\mathbf{1}_{\{T_F < \infty\}} | X_0 = x],$$

where h is a bounded mapping from F into $\mathbb{R}$, and the function $v : E \to \mathbb{R} \cup \{+\infty\}$ by

$$v(x) := \mathbb{E}[T_F | X_0 = x].$$

Show that T_F is a stopping time. Show that u and v solve respectively the equations:

$$Qu(x) = 0, x \in E \backslash F,$$
$$u(x) = h(x), x \in F;$$

$$Qv(x) + 1 = 0, x \in E \backslash F,$$
$$v(x) = 0, x \in F.$$

(Hint: condition upon $(T_1, X(T_1))$.)

2. *Consider next the case of an $E = \mathbb{Z}$-valued birth and death process, whose infinitesimal generator Q satisfies $Q_{x,x+1} = \alpha(x)$, $Q_{x,x-1} = \beta(x)$, $Q_{x,x} = -\alpha(x) - \beta(x)$, in the particular case $\alpha(x) = \alpha$, $\beta(x) = \beta$, $x \in \mathbb{Z}$ $(\alpha, \beta > 0)$. Let $F = \{1, 2, \ldots, N-1\}^c$, where N is a positive integer. Compute $u(x) = \mathbb{E}[X_{T_F} | X_0 = x]$, $x \in \mathbb{Z}$. Show that T_F is almost surely finite. Give the conditional law of the random variable X_{T_F}, given that $X_0 = x$.*

Exercise 11.5 *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $\{\mathcal{F}_t\}$ (i.e. an increasing collection indexed by $t \in \mathbb{R}_+$ of sub-σ-fields of $\mathcal{A}$), a martingale (with respect to the filtration $\{\mathcal{F}_t\}$) is a stochastic process $\{M_t; t \in \mathbb{R}_+\}$ such that*

$$M_t \text{ is integrable, } \forall t \geq 0; \quad \mathbb{E}[M_t | \mathcal{F}_s] = M_s, \ \ \forall 0 \leq s \leq t.$$

1. *Let $\{M_t;\ t \in \mathbb{R}_+\}$ be a martingale which is continuous on the right, and S a stopping time which is bounded by a constant t. Show that $\mathbb{E}[M_S] = \mathbb{E}[M_t] = \mathbb{E}[M_0]$.*

2. *Let $\{X_t;\ t \geq 0\}$ be an E-valued jump Markov process, with the infinitesimal generator $\{Q_{xy};\ x, y \in E\}$ satisfying $\sup_x Q_{xx} < \infty$, and f be a bounded mapping from E into $\mathbb{R}$. Show that $\{M_t;\ t \in \mathbb{R}_+\}$ defined by*

$$M_t = f(X_t) - \int_0^t Qf(X_s)ds$$

is a martingale with respect to the filtration $\{\mathcal{F}_t^X\}$ (accept the fact that for any bounded function f from E into $\mathbb{R}$, and any $r > 0$, $P_r Qf = QP_r f$).

3. *Using the notation from part 2 of Exercises 11.4, compute $\mathbb{E}[T_F | X_0 = x]$ in terms of the law of X_{T_F}. (Hint: in the case $\alpha \neq \beta$, use the results of parts 1 and 2 of the present exercise with the function $f(x) = x$ and the stopping time $S = \inf(T_F, t)$, then let t tend to infinity; in the case $\alpha = \beta$, do the same computations with $f(x) = x^2$.) You should assume that the result in part 2 of the present exercise applies to these two functions, even though they are not bounded.*

Exercise 11.6 *Let $\{X_t;\ t \geq 0\}$ be an $\mathbb{N}$-valued jump Markov process with infinitesimal generator*

$$Q = \begin{pmatrix} -\mu & \mu & 0 & 0 & 0 & \dots \\ \lambda & -(\lambda+\mu) & \mu & 0 & 0 & \dots \\ 0 & \lambda & -(\lambda+\mu) & \mu & 0 & \dots \\ 0 & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

where $\lambda, \mu > 0$.

1. *Specify the embedded chain. Show that $\{X_t;\ t \geq 0\}$ is irreducible.*

2. *Show that $\{X_t;\ t \geq 0\}$ is recurrent in the case $\lambda \geq \mu$, and transient in the case $\lambda < \mu$.*

3. *Show that whenever $\lambda > \mu$, $\{X_t;\ t \geq 0\}$ possesses a unique invariant probability distribution π and specify it.*

4. *Show that $\{X_t;\ t \geq 0\}$ is positive recurrent in the case $\lambda > \mu$, and null recurrent in the case $\lambda = \mu$.*

Exercise 11.7 *With P denoting the transition matrix defined in Exercise 10.9 of Chapter 2, let $Q = P - I$ and consider a continuous time jump Markov process $\{X_t;\ t \geq 0\}$ whose infinitesimal generator is Q.*

1. *Give the transition matrix P' of the embedded chain.*

2. *Describe the trajectories of the process $\{X_t\}$, and specify the parameters of the exponential laws of the time spent in the various states.*

3. *Show that the process $\{X_t\}$ is irreducible and positive recurrent. Determine its invariant probability distribution.*

4. *Determine the invariant probability distribution of the embedded chain.*

Exercise 11.8 *Consider both the $E = \mathbb{N}$-valued discrete time Markov chain $\{X_n;\ n \in \mathbb{N}\}$ whose transition matrix is*

$$P = \begin{pmatrix} q & p & 0 & 0 & \dots \\ q & 0 & p & 0 & \dots \\ 0 & q & 0 & p & \dots \\ 0 & 0 & q & 0 & p \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and the $E = \mathbb{N}$-valued continuous time jump Markov process $\{X_t;\ t \geq 0\}$ whose infinitesimal generator is

$$Q = \begin{pmatrix} -p & p & 0 & 0 & \dots \\ q & -1 & p & 0 & \dots \\ 0 & q & -1 & p & \dots \\ 0 & 0 & q & -1 & p \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $0 < p, q < 1$, $p + q = 1$. You might compare the chain $\{X_n;\ n \in \mathbb{N}\}$ and the $\mathbb{Z}$-valued random walk which was studied in Exercise 10.11 of Chapter 2 (which we proved to be transient if $p \neq q$, and null recurrent if $p = q$), and note that the process $\{X_t;\ t \geq 0\}$ is an $M/M/1$ queue as studied in the next chapter.

1. *Is the Markov chain $\{X_n;\ n \in \mathbb{N}\}$ the embedded chain associated with the jump Markov process $\{X_t;\ t \geq 0\}$?*

2. *Show that the two processes $\{X_n;\ n \in \mathbb{N}\}$ and $\{X_t;\ t \geq 0\}$ are irreducible.*

3. *Show that any invariant probability distribution for $\{X_n;\ n \in \mathbb{N}\}$ is also invariant for $\{X_t;\ t \geq 0\}$, and vice versa. Show that the two processes are either both transient, both null recurrent, or both positive recurrent.*

4. *Show that the two processes are transient in the case $p > q$.*

5. *Show that they are recurrent in the case $p = q = 1/2$ (you can compare with the chain studied in Exercise 10.11 of Chapter 2). Specify for this case an invariant measure with infinite mass, and deduce that the processes are null recurrent.*

6. *Now assume that $p < q$. Set $\lambda = p/q$, and note that $q^{-1}(\lambda - p) = \lambda^2$. Show that there exists a geometric probability distribution π which is invariant for the two processes (compute π_1 in terms of π_0, π_2 in terms of $\pi_0, \ldots$).*

The rest of this exercise studies two variants of the continuous time jump Markov process $\{X_t;\ t \geq 0\}$.

7. *Modify the infinitesimal generator Q by multiplying p and q by the same constant $c > 0$. Show that neither the classification of the process (transient, null recurrent or positive recurrent) nor the possible invariant measure is modified by the presence of the constant c. What is changed in the process?*

8. *Suppose now that $p < q$; still using the notation $\lambda = p/q$, we now consider the jump Markov process $\{Y_t;\ t \geq 0\}$ whose infinitesimal generator Q' is defined by*

$$Q'_{0y} = \begin{cases} -p, & \text{if } y = 0, \\ p, & \text{if } y = 1, \\ 0, & \text{otherwise} \end{cases}$$

and, for $x \geq 1$,

$$Q'_{xy} = \begin{cases} \lambda^x q, & \text{if } y = x - 1, \\ -\lambda^x, & \text{if } y = x, \\ \lambda^x p, & \text{if } y = x + 1, \\ 0, & \text{otherwise}. \end{cases}$$

Compare the embedded chains of $\{X_t;\ t \geq 0\}$ and $\{Y_t;\ t \geq 0\}$. Show that $\{\pi_x = 1;\ x \in \mathbb{N}\}$ is an invariant measure, and deduce that $\{Y_t;\ t \geq 0\}$ is null recurrent. Explain why the mean time taken by $\{Y_t\}$ to return to x, starting from x, is longer than it is for $\{X_t\}$.

Exercise 11.9 *Consider an $\mathbb{N}$-valued random sequence $\{X_n;\ n \geq 0\}$, defined as follows: $X_0 = x_0 \in \mathbb{N}$ and, for all $n \in \mathbb{N}$,*

$$X_{n+1} = \begin{cases} (X_n + U_{n+1})^+, & \text{if } V_{n+1} = 1, \\ 0, & \text{if } V_{n+1} = 0, \end{cases}$$

where the sequence $(U_1, V_1, U_2, V_2, \ldots)$ is independent and, for all $n \geq 1$,

$$\mathbb{P}(U_n = 1) = \mathbb{P}(U_n = -1) = 1/2, \qquad \mathbb{P}(V_n = 1) = 1 - \mathbb{P}(V_n = 0) = p,$$

with $0 < p < 1$.

1. *Show that $\{X_n;\ n \geq 0\}$ is an $\mathbb{N}$-valued irreducible Markov chain, and give its transition matrix.*

2. *Show that $\{X_n;\ n \geq 0\}$ is positive recurrent (do not look for an invariant probability distribution; instead use a more 'probabilistic' argument).*

3. *Show that if* α *solves the equation* $p(1+\alpha^2) = 2\alpha$ *and* $0 < \alpha < 1$, *then the geometric probability distribution* π *on* $\mathbb{N}$ *given by* $\pi_x = (1-\alpha)\alpha^x$, $x \in \mathbb{N}$, *is the unique invariant probability distribution of the chain.*

   Consider now an $\mathbb{N}$*-valued continuous time jump Markov process* $\{X_t;\ t \geq 0\}$ *whose infinitesimal generator is given by*

$$Q = \begin{pmatrix} -p/2 & p/2 & 0 & 0 & 0 & \dots \\ 1-p/2 & -1 & p/2 & 0 & 0 & \dots \\ 1-p & p/2 & -1 & p/2 & 0 & \dots \\ 1-p & 0 & p/2 & -1 & p/2 & 0\dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \dots \end{pmatrix}.$$

4. *Specify the transition matrix of the associated embedded chain (compare with the chain from parts 1–3). Show that* $\{X_t;\ t \geq 0\}$ *is irreducible and recurrent. Show that the probability distribution from part 3 is invariant for* X_t. *Deduce that* $\{X_t;\ t \geq 0\}$ *is positive recurrent.*

8

Queues and networks

Introduction

Jump Markov processes are used to model queues. These were first studied in the context of the telephone, then for operational research. Later mathematicians studied networks of queues, which are used to evaluate the performance of computing systems and also production systems.

The basic mathematical model of queues is as follows. Customers arrive according to a certain point process. When a customer arrives, if a server is free the customer is served straight away. If no server is available, then the customer waits, and is served as soon as a server is free (and the customers who arrived before him have been served – unless another priority policy is in order). The various service times are mutually independent and follow a certain probability distribution (which in certain cases may depend upon the type of customer). We shall implicitly assume that the 'waiting room' has unlimited capacity and that no customer is rejected, unless we explicitly say otherwise.

The queue will be characterized by the law of the arrival times, the law of the service times, and the number of servers. We shall always assume that the service times are i.i.d. and independent of the process of the arrival times.

See [2] and [33] for more complete treatments of this subject.

8.1 $M/M/1$ queue

Assume that the arrivals constitute a Poisson process with intensity λ, that the service times are i.i.d., their common law being exponential (μ) – M stands for 'memoryless' or 'Markovian' – and that 1 is the number of servers. The number of customers present in the queue (either being served or waiting) is then an $\mathbb{N}$-valued jump Markov process whose infinitesimal generator Q is given by

$$Q_{x,x+1} = \lambda,\ x \in \mathbb{N}; \quad Q_{x,x-1} = \mu,\ x \geq 1; \quad Q_{xy} = 0 \text{ if } |x-y| \geq 2;$$

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

hence

$$Q_{00} = -\lambda, \quad Q_{xx} = -(\lambda + \mu),\ x \geq 1.$$

Indeed, let $t \geq 0$ and $x \geq 1$. Let $\{N_s\}$ be the arrivals process and S the waiting time after t for the current service to be completed. We have

$$\begin{aligned}
\mathbb{P}(X_{t+h} = x + 1 | X_t = x) &= \mathbb{P}(N_{t+h} - N_t = 1, S > h) + o(h) \\
&= e^{-\lambda h} \lambda h e^{-\mu h} + o(h), \\
h^{-1} P_{x,x+1}(h) &\to \lambda, \quad h \to 0. \\
\mathbb{P}(X_{t+h} = x - 1 | X_t = x) &= \mathbb{P}(N_{t+h} - N_t = 0,\ S \leq h) + o(h) \\
&= e^{-\lambda h}(1 - e^{-\mu h}) + o(h), \\
h^{-1} P_{x,x-1}(h) &\to \mu, \quad h \to 0.
\end{aligned}$$

We have used independence between the arrivals process and the service times. If we conditioned upon past values of $\{X_s\}$ before time t, the result would not be affected. This follows from both the Markov property of $\{N_s\}$ and the properties of the exponential law of S (see Exercise 5.1 of Chapter 6).

Moreover, by the same argument, for $|y-x| \geq 2$,

$$\mathbb{P}(X_{t+h} = y | X_t = x) = o(h).$$

The $M/M/1$ queue is an irreducible $\mathbb{N}$-valued continuous time 'birth and death process'. If $\lambda > \mu$, the mean number of arrivals per unit time is greater than the mean number of departures per unit time, and $X_t \to +\infty$ almost surely. The process is transient.

If $\lambda = \mu$, the process is null recurrent; if $\lambda < \mu$, it is positive recurrent. Its invariant probability distribution is given by

$$\pi_x = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^x.$$

The detailed justification of these facts is left to Exercises 13.1. At equilibrium, the mean number of customers present in the queue is

$$\mathbb{E}_\pi(X_t) = \sum_{x=1}^{\infty} x\pi_x = \frac{\lambda}{\mu - \lambda}.$$

The expectation of the return time to 0 is

$$\mathbb{E}_0(R_0) = \frac{1}{q_0 \pi_0} = \frac{\mu}{\lambda(\mu - \lambda)}.$$

The expected time between two periods when the queue is empty is

$$\mathbb{E}_0(R_0) - \frac{1}{q_0} = \frac{1}{\mu - \lambda}.$$

Let us compute the mean time spent by a customer in the queue. Conditioned upon the fact that there are x customers in front of him when he arrives, the mean time spent by a customer is $(x+1)/\mu$. Hence the mean time is

$$\mathbb{E}_\pi (X_t + 1)/\mu = \frac{1}{\mu - \lambda}.$$

The above argument, which may seem contradictory (why should the state of the queue at an arrival time be in the stationary regime?), is correct asymptotically, in the sense that the law of the number of customers that the nth arriving customer finds in front of him converges towards the invariant probability distribution of $\{X_t\}$, as $n \to \infty$ (see Exercise 13.2). It is, moreover, a consequence of the very intuitive Theorem 1.1, where $\{X_t\}$ denotes the number of customers present in a queue, which need not be of $M/M/1$ type. Let us formulate three assumptions, which in particular are satisfied in the positive recurrent Markovian case. First,

$$t^{-1} \int_0^t X(s) ds \to \bar{X} \text{ a.s.}, \quad \text{as } t \to \infty, \tag{H1}$$

where $\bar{X}$ is a constant. Second,

$$\exists \text{ a random sequence } t_n \to \infty, \text{ as } n \to \infty, \text{ such that } X_{t_n} = 0. \tag{H2}$$

Third, the time spent in the queue by each customer is finite, and if D_n denotes the time spent in the queue by the nth arriving customer, there exists a constant $\bar{D}$ such that

$$\bar{D} = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} D_k. \tag{H3}$$

Theorem 1.1 (Little's formula) *Consider a queuing system such that the mean number of arrivals per unit time equals λ, and which satisfies assumptions (H1)–(H3). Then*

$$\bar{X} = \lambda \, \bar{D}.$$

Proof Denote by $N(t)$ the number of customers who arrived before t, and by $\{t_n;\ n \in \mathbb{N}\}$ a sequence which is such that $X_{t_n} = 0$, for all n and $t_n \to \infty$, as $n \to \infty$. Then, if $X_0 = 0$, it is not hard to check that

$$\sum_{k=1}^{N(t_n)} D_k = \int_0^{t_n} X_s ds.$$

Hence,

$$\frac{1}{t_n} \int_0^{t_n} X_s ds = \frac{N(t_n)}{t_n} \times \frac{1}{N(t_n)} \sum_{k=1}^{N(t_n)} D_k.$$

It remains to let $n \to \infty$ and exploit the assumptions. The case $X_0 \neq 0$ then follows, since the lengths of the visits of the customers present at time 0 are finite. □

8.2 $M/M/1/K$ queue

In practice, only a finite number of customers can be waiting. In the $M/M/1/K$ model, arrivals are modelled by a Poisson point process with intensity λ, service times are exponential with parameter μ, there is one server, and the 'waiting room' contains at most K customers (including the one being served). This means that any customer who arrives while the waiting room is full is rejected.

The number of customers present in the queue is then a $\{0, 1, 2 \ldots, K\}$-valued jump Markov process. Its infinitesimal generator is given by

$$Q_{x,x+1} = \lambda, \quad 0 \le x \le K-1; \quad Q_{x,x-1} = \mu, \quad 1 \le x \le K;$$
$$Q_{xy} = 0, \quad \text{if } |x-y|>2.$$
$$Q_{00} = -\lambda, \quad Q_{xx} = -(\lambda+\mu), \quad 1 \le x \le K-1; \quad Q_{KK} = -\mu.$$

This Markov process is irreducible, with values in a finite set, hence it is positive recurrent. Its invariant probability distribution π can be easily computed. If $\lambda \neq \mu$, we have

$$\pi_x = \left(\frac{\lambda}{\mu}\right)^x \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{K+1}}, \quad 0 \le x \le K.$$

If $\lambda = \mu$, we have

$$\pi_x = \frac{1}{K+1}.$$

The probability that an arriving customer is rejected equals the proportion of the time the waiting room is full, that is,

$$\pi_K = \begin{cases} \left(\frac{\lambda}{\mu}\right)^K \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{K+1}}, & \text{if } \lambda \neq \mu, \\ \pi_K = \frac{1}{K+1}, & \text{if } \lambda = \mu. \end{cases}$$

8.3 $M/M/s$ queue

Let us again assume that the 'waiting room' has infinite capacity, but now there are s available servers. The service times at the various counters are of course assumed to be i.i.d. It is not hard to see that

$$Q_{01} = \lambda, \quad Q_{0x} = 0, \; x > 1;$$

for $1 \le x \le s$,

$$Q_{x,x+1} = \lambda, \quad Q_{x,x-1} = x\mu, \quad Q_{xy} = 0, \; |x-y| > 1;$$

for $x \ge s$,

$$Q_{x,x+1} = \lambda, \quad Q_{x,x-1} = s\mu, \quad Q_{xy} = 0, \; |x-y| > 1.$$

The only new computation needed in order to justify those statements is: if $S_1, \ldots, S_x$ are i.i.d. exponential (μ) random variables, then

$$\mathbb{P}(S_1 \wedge \cdots \wedge S_x > h) = (e^{-\mu h})^x.$$

Hence, the probability that there are at least two departures during a time interval of length h, while x counters are busy, is

$$1 - e^{-x\mu h},$$

and

$$\frac{1 - e^{-x\mu h}}{h} \to x\mu.$$

Moreover, the probability that at least two departures take place during an interval of time of length h is of order $o(h)$.

$\{X_t\}$ is positive recurrent if $\lambda < \mu s$. In that case, one can look for an invariant probability distribution by looking for a solution of the detailed balance equation

$$\pi_x Q_{x,x+1} = \pi_{x+1} Q_{x+1,x}.$$

We find that

$$\frac{\pi_x}{\pi_0} = \begin{cases} (\lambda/\mu)^x / x!, & \text{if } 0 \le x \le s; \\ (\lambda/\mu)^x / s^{x-s} s!, & \text{if } x > s. \end{cases}$$

The two cases which lead to a simple formula are $s = 1$ (already treated – geometric law) and $s = \infty$, in which case $\pi_0 = e^{-\lambda/\mu}$ and

$$\pi_x = e^{-\lambda/\mu} (\lambda/\mu)^x / x!,$$

that is, the invariant probability distribution is Poisson (λ/μ).

We have the following theorem.

Theorem 3.1 (Burke) *If $\lambda < s\mu$, at equilibrium the departures process is a Poisson process with intensity λ.*

PROOF Since $\{X_t\}$ is time-reversible with respect to π, at equilibrium (i.e.under $\mathbb{P}_\pi$), $\{X_{T-t};\ 0 \le t \le T\}$ has the same law as $\{X_t; 0 \le t \le T\}$, and the departures of $\{X_t\}$ are the arrivals of $\{X_{T-t}\}$. □

Let us now compute the probability that a customer arrives while all servers are busy (and hence that he has to wait before being served). This equals the proportion of time during which all servers are busy, hence

$$\sum_{x=s}^{\infty} \pi_x = \pi_0 \frac{s^s}{s!} \sum_{x=s}^{\infty} \left(\frac{\lambda}{\mu s}\right)^x = \pi_0 \frac{(\lambda/\mu)^s}{s!} \times \frac{\mu s}{\mu s - \lambda}, \tag{8.1}$$

with

$$\pi_0 = \left[\sum_{x=0}^{s-1} \frac{(\lambda/\mu)^x}{x!} + \frac{(\lambda/\mu)^s}{s!} \times \frac{\mu s}{\mu s - \lambda} \right]^{-1}.$$

Equation (8.1) is known as Erlang's formula.

Finally, we wish to know whether it is more efficient, when arrivals happen with intensity 2λ, to have two servers in parallel, each with service time following the exponential law with parameter μ (solution 1), or one server, with service time following the exponential law with parameter 2μ (solution 2) ($\lambda < \mu$). Let us compute the mean time spent by a customer in the system in each case. For solution 1, at equilibrium the law of the number of customers is given by

$$\pi_0 = \frac{1-\lambda/\mu}{1+\lambda/\mu}; \quad \pi_x = 2\left(\frac{\lambda}{\mu}\right)^x \frac{1-\lambda/\mu}{1+\lambda/\mu}, \ x \geq 1.$$

So the mean number of customers present in the queue equals

$$\bar{X}_1 = 2\frac{1-\lambda/\mu}{1+\lambda/\mu}\sum_{x=1}^{\infty} x\left(\frac{\lambda}{\mu}\right)^x = \frac{2\lambda/\mu}{(1-\lambda/\mu)(1+\lambda/\mu)}.$$

Hence, by Little's formula, the mean time spent by each customer in the queue equals

$$\bar{D} = \frac{\bar{X}_1}{2\lambda} = \frac{1}{(\mu-\lambda)(1+\lambda/\mu)},$$

For solution 2,

$$\bar{D}_2 = \frac{1}{(\mu-\lambda)2}.$$

But $1+\lambda/\mu < 2$, hence $\bar{D}_2 < \bar{D}_1$, and the solution of a unique server, with service time following the exponential law with parameter 2μ, is preferable.

8.4 *M/M/s/s* queue

We now consider an $M/M/s$ queue with no waiting room, where any customer arriving while all s counters are busy is rejected. This is how a telephone exchange works.

The number X_t of customers in the system at time t is a $\{0, 1, \ldots, s\}$-valued jump Markov process, with infinitesimal generator

$$Q_{01} = \lambda, \quad Q_{0x} = 0, \ x > 1;$$

if $0 < x < s$,

$$Q_{x,x+1} = \lambda, \quad Q_{x,x-1} = x\mu, \quad Q_{xy} = 0, |x-y| > 1;$$

$$Q_{s,s-1} = s\mu, \quad Q_{sy} = 0, \ y \neq s, y \neq s-1.$$

The invariant probability distribution is the 'truncated Poisson',

$$\pi_x = \frac{(\lambda/\mu)^x/x!}{\sum_{y=0}^{s}(\lambda/\mu)^y/y!}, \quad 0 \leq x \leq s.$$

From the ergodic theorem, the proportion of time during which the system is full, which is also the proportion of lost calls, equals

$$\pi_s = \frac{(\lambda/\mu)^s/s!}{\sum_{y=0}^{s}(\lambda/\mu)^y/y!}$$

which is another way of writing Erlang's formula.

8.5 Repair shop

Suppose that s machines are working in a factory. Each has a failure rate λ (i.e. the failure times are i.i.d. exponential random variables with parameter λ). The factory is equipped with a repair shop. The repair times are exponential with parameter μ.

Denote by X_t the number of machines in working condition at time t, taking values in the set $E = \{0, 1, \ldots, s\}$. $\{X_t\}$ is a jump Markov process. Its infinitesimal generator is the same as that of the queue $M/M/s/s$, but with λ and μ interchanged. Hence, the invariant probability distribution of this process is

$$\pi_x = \frac{(\mu/\lambda)^x/x!}{\sum_{y=0}^{s}(\mu/\lambda)^y/y!}, \quad 0 \le x \le s.$$

The global failure rate is

$$\lambda_g = \sum_{x=1}^{s} \lambda_x \pi_x = \mu \frac{\sum_{x=0}^{s-1}(\mu/\lambda)^x/x!}{\sum_{y=0}^{s}(\mu/\lambda)^y/y!}.$$

The mean number of machines which are repared per unit time equals

$$n_r = \mu(1 - \pi_s) = \lambda_b.$$

8.6 Queues in series

Suppose that customers require two services: they queue first at counter A, then at counter B, whose service times are independent. The service time at A follows an exponential (α) distribution, and that at B an exponential (β). What is the mean length of the queue at B? Denote by $\{X_t\}$ the number of customers (being served or waiting) at A, $\{Y_t\}$ the number of customers at B.

Arrivals at A are supposed to constitute a Poisson point process with intensity λ. If $\lambda > \alpha$, $X_t \to \infty$, we might as well assume that there are always customers in A, and they exit according to a Poisson process with intensity α. If $\lambda < \alpha$, by Burke's theorem the departures from A follow a Poisson process with intensity λ. It is then natural to pretend that the process of arrivals at B is Poisson with intensity $\lambda \wedge \alpha$. Hence, $\{Y_t\}$ is positive recurrent if and only if $\lambda \wedge \alpha < \beta$, and in

this case the mean length of the queue at B is $(\alpha \wedge \lambda)/(\beta - \alpha \wedge \lambda)$. The two queues are positive recurrent if and only if $\lambda < \alpha \wedge \beta$. In that case, at equilibrium, the mean time spent by a customer in the system A+B is

$$\frac{1}{\alpha - \lambda} + \frac{1}{\beta - \lambda}.$$

8.7 $M/G/\infty$ queue

Customers arrive according to a Poisson point process with intensity λ, as before. Service times are i.i.d., the common law being completely arbitrary ('general'), with distribution function $F(t) = \mathbb{P}(T \leq t)$, $t \geq 0$. We assume here that the number of servers is infinite, which simplifies the analysis a great deal, since there is no interaction between the customers.

The number N_t of customers who arrived before time t follows a Poisson distribution with parameter λt. Let us condition upon $N_t = n$. If we label these n customers randomly, then their arrival times $A_1, \ldots, A_n$ are i.i.d., their common law being uniform on $[0, t]$ (see Exercise 5.4 of Chapter 6).

For each of these customers, service is not completed at time t with probability

$$p = \frac{1}{t} \int_0^t \mathbb{P}(T > s) ds = \frac{1}{t} \int_0^t (1 - F(s)) ds.$$

Hence the conditional law of X_t, given that $N_t = n$, is binomial (n, p).

Finally,

$$\begin{aligned}
\mathbb{P}(X_t = x) &= \sum_{n=0}^{\infty} \mathbb{P}(X_t = x | N_t = n) \mathbb{P}(N_t = n) \\
&= \sum_{n=x}^{\infty} C_x^n p^x (1-p)^{n-x} e^{-\lambda t} (\lambda t)^n / n! \\
&= e^{-\lambda t} \frac{(\lambda p t)^x}{x!} \sum_{n=x}^{\infty} \frac{(\lambda (1-p) t)^{n-x}}{(n-x)!} \\
&= \frac{e^{-\lambda p t} (\lambda p t)^x}{x!}.
\end{aligned}$$

Hence, the law of X_t is Poisson with parameter $\lambda \int_0^t (1 - F(s)) ds$. Note that

$$\int_0^{\infty} (1 - F(s)) ds = \int_0^{\infty} \mathbb{P}(T > s) ds = \mathbb{E}(T).$$

If $\mathbb{E}(T) < \infty$, the asymptotic probability distribution of X_t as $t \to \infty$ is Poisson with parameter $\lambda \mathbb{E}(T)$.

8.8 $M/G/1$ queue

Customers arrive according to a Poisson point process with intensity $\lambda > 0$. Successive service times are i.i.d. random variables, which are globally independent of the process of arrivals. We denote by ν the law of service times, and L its Laplace transform,

$$L(u) = \int_0^\infty e^{-ut}\nu(dt), \quad u > 0.$$

Let μ denote the expectation of the probability ν, which is assumed to be finite.

8.8.1 An embedded chain

For all $n \geq 0$, let X_n denote the number of customers waiting (or being served) just after the nth departure. For $n \geq 0$,

$$X_{n+1} = X_n + Y_{n+1} - \mathbf{1}_{X_n > 0},$$

where Y_n is the number of arrivals while the nth customer is being served. The fact that $\{X_n;\ n \geq 0\}$ is a Markov chain is a consequence of the following lemma.

Lemma 8.1 *The random variables $\{Y_n;\ n \geq 1\}$ are i.i.d., independent of X_0, and their joint law has the generating function*

$$A(z) = L(\lambda(1 - z)).$$

PROOF Independence between the Y_n and X_0 is clear. Let $n \geq 1$. Denote by S_i the length of service of the ith customer. Then

$$\begin{aligned}
&\mathbb{E}\left(z_1^{Y_1}\cdots z_n^{Y_n}\right)\mathbb{E}\left[\mathbb{E}\left(z_1^{Y_1}\cdots z_n^{Y_n}|S_1,\ldots,S_n\right)\right] \\
&\qquad = \int_{\mathbb{R}_+^n} \mathbb{E}\left(z_1^{Y_1}\cdots z_n^{Y_n}|S_1 = t_1,\ldots,S_n = t_n\right)\nu(dt_1)\cdots\nu(dt_n) \\
&\qquad = \prod_{i=1}^n \int_{\mathbb{R}_+} e^{-\lambda t_i(1-z_i)}\nu(dt_i) \\
&\qquad = \prod_{i=1}^n L(\lambda(1 - z_i)).
\end{aligned}$$

□

Clearly the $\mathbb{N}$-valued Markov chain $\{X_n\}$ is irreducible. Note that

$$\rho \stackrel{\text{def}}{=} \mathbb{E}(Y_n) = \lambda\mu.$$

8.8.2 The positive recurrent case

Proposition 8.2 *If $\rho < 1$, the chain $\{X_n\}$ is positive recurrent.*

PROOF Denote by Z_n the number of times the chain visits 0 before the nth departure. Note that

$$X_n = X_0 + Y_1 + \ldots + Y_n - n + Z_n.$$

Suppose that the chain is transient. Then $X_n \to \infty$ almost surely as $n \to \infty$, and $Z_n/n \to 0$ as $n \to \infty$, since there are almost surely at most a finite number of returns to 0. From the law of large numbers,

$$n^{-1}(Y_1 + \ldots + Y_n) \to \rho < 1,$$

hence

$$Y_1 + \ldots + Y_n - n \to -\infty,$$

as $n \to \infty$, which contradicts the fact that $X_n \geq 0$, and the chain is recurrent. Let us take the expectation in the above recurrence formula. Since $X_n \geq 0$,

$$0 < 1 - \rho \leq n^{-1}\mathbb{E}X_0 + n^{-1}\mathbb{E}Z_n \to m_0^{-1}$$

as $n \to \infty$, where m_0 is the mean return time to 0, from the ergodic theorem. Hence

$$m_0 \leq \frac{1}{1-\rho} < \infty.$$

□

Denote by π the invariant probability of $\{X_n\}$, and by G its generating function. Then

$$\begin{aligned} zG(z) &= \mathbb{E}\left(z^{X_{n+1}+1}\right) \\ &= \mathbb{E}\left(z^{X_n+Y_{n+1}+\mathbf{1}_{X_n=0}}\right) \\ &= A(z)(\pi_0 z + G(z) - \pi_0), \end{aligned}$$

hence

$$(A(z) - z)G(z) = \pi_0 A(z)(1-z). \tag{8.2}$$

As $z \to 1$,

$$\frac{A(z)-z}{1-z} \to 1 - A'(1-) = 1 - \rho,$$

hence $\pi_0 = 1-\rho$ and

$$G(z) = (1-\rho)\frac{(1-z)A(z)}{A(z)-z}.$$

Let us compute the *mean length of the queue*. Differentiating (8.2), we get

$$(A(z)-z)G'(z)+(A'(z)-1)G(z)=(1-\rho)[A'(z)(1-z)-A(z)],$$

hence, substituting for $G(z)$,

$$G'(z)=(1-\rho)A'(z)\frac{1-z}{A(z)-z}-(1-\rho)A(z)\frac{(A'(z)-1)(1-z)+A(z)-z}{(A(z)-z)^2}.$$

But as $z\to 1$,

$$\begin{aligned}\frac{(A'(z)-1)(1-z)+A(z)-z}{(A(z)-z)^2}&=\frac{A(z)-1-A'(z)(z-1)}{(A(z)-z)^2}\\&\simeq -\frac{1}{2}A''(z)\frac{(1-z)^2}{(A(z)-z)^2}\\&\to -\frac{A''(1-)}{2(1-\rho)^2}.\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}_\pi(X_n)&=G'(1-)\\&=\rho+\frac{A''(1-)}{2(1-\rho)}\\&=\rho+\frac{\lambda^2L''(0+)}{2(1-\rho)}\\&=\rho+\frac{\lambda^2\mathbb{E}S^2}{2(1-\rho)},\end{aligned}$$

where S denotes the service time (i.e. a random variable whose probability distribution is ν).

We now compute the *mean time spent by a customer waiting* before being served. Consider the queue at equilibrium. A customer who leaves has spent time Q waiting and time S being served. Conditionally upon the fact that $Q+S=t$, the number of customers left behind him follows the Poisson distribution with parameter λt. Hence, if $M(u)=\mathbb{E}(e^{-uQ})$ denotes the Laplace transform of the law of Q,

$$\begin{aligned}G(z)&=\mathbb{E}\left(e^{-\lambda(Q+S)(1-z)}\right)\\&=M(\lambda(1-z))L(\lambda(1-z)).\end{aligned}$$

Hence

$$M(u)=\frac{(1-\rho)u}{u-\lambda(1-L(u))}.$$

It then follows from the independence of Q and S that

$$\mathbb{E}Q=-M'(0+)=\frac{\lambda\mathbb{E}(S^2)}{2(1-\rho)}.$$

Finally, we wish to compute the *mean length of the periods of activity* O, during which the server is busy. Denote by

$$B(u) = \mathbb{E}(e^{-uO})$$

its Laplace transform. Let S denote the service time of the first customer during the period of activity. Conditionally upon $S = t$,

$$O = t + O_1 + \ldots + O_N,$$

where N is the number of customers who arrive while the first customer is being served, whose conditional law given that $S = t$ is Poisson with parameter λt, and $O_1, O_2, \ldots$ are i.i.d., with the same law as O, jointly independent of N. Hence,

$$\begin{aligned} B(u) &= \int_0^\infty \mathbb{E}(e^{-uO|S=t})\nu(dt) \\ &= \int_0^\infty e^{-ut} e^{-\lambda t(1-B(u))}\nu(dt) \\ &= L(u + \lambda(1 - B(u))). \end{aligned}$$

One can deduce from these calculations the expectation of O:

$$\mathbb{E}(O) = \mu(1 + \lambda\mathbb{E}(O)) = \frac{\mu}{1-\rho}.$$

8.9 Open Jackson network

Consider a network consisting of N interconnected service stations. Each station p $(1 \le p \le N)$ must serve customers arriving from outside, as well as those sent from the other stations.

In the Jackson model, exogenous arrivals at station p constitute a Poisson point process with intensity λ_p^0. Each station is a queue. The rate at which customers leave the station p is $\mu_p(n)$, if n customers are present at station p. For example,

$$\mu_p(n) = \begin{cases} \mu_p \mathbf{1}_{\{n>0\}}, & \text{for an } M/M/1 \text{ server,} \\ n \wedge s_p\mu_p, & \text{for an } M/M/s \text{ server.} \end{cases}$$

A customer leaving station p goes next to station q with probability r_{pq} $(1 \le q \le N)$, and leaves the network with probability r_{p0}. We assume for simplicity that $r_{pp} = 0$, $1 \le p \le N$; one can easily dispense with this assumption.

The process $X_t = (X_t^1, \ldots, X_t^N)$ of the numbers of customers present at stations $1, \ldots, N$ at time t is an $E = \mathbb{N}^N$-valued jump Markov process.

Denote by $e_p = (0, \ldots, 0, 1, 0, \ldots, 0)$ the vector whose pth entry is 1, all others being zero.

The only non-zero off-diagonal entries of the infinitesimal generator of the process $\{X_t\}$ are given by

$$\begin{cases} Q_{x,x+e_p} = \lambda_p^0, & x \in E, \quad 1 \le p \le N, \\ Q_{x+e_p,x} = \mu_p(x_p+1)r_{p0}, & x \in E, \quad 1 \le p \le N, \\ Q_{x+e_p,x+e_q} = \mu_p(x_p+1)r_{pq}, & 1 \le p \ne q \le N, \quad x \in E. \end{cases}$$

We say that the network is 'capture-free' if, for all $p \in \{1, \ldots, N\}$, there exist $n \ge 0$, $p_1, \ldots, p_n \in \{1, \ldots, N\}$, such that

$$r_{pp_1}r_{p_1p_2}\cdots r_{p_{n-1}p_n}r_{p_n0} > 0. \tag{8.3}$$

We consider the equation

$$\lambda_p^0 + \sum_{q \ne p} \lambda_q r_{qp} = \lambda_p, \quad 1 \le p \le N, \tag{8.4}$$

whose unknowns are λ_p, $1 \le p \le N$.

Lemma 9.1 *Under assumption (8.3), equation (8.4) has a unique positive and finite solution.*

PROOF Let R be the $N \times N$ matrix whose entries are $R_{pq} = r_{pq}$. Equation (8.4) can be written in the vector form

$$\lambda(I - R) = \lambda^0,$$

which has the unique finite solution

$$\lambda^0 = \lambda K,$$

with $K = \sum_{n=0}^{\infty} R^n$, provided this series converges in $\mathbb{R}^{N\times N}$.

Consider the Markovian matrix

$$P = \begin{pmatrix} r_{10} & & \\ r_{20} & & R \\ r_{N0} & & \\ 1 & 0\cdots 0 & \end{pmatrix}.$$

Assumption (8.3) implies that the states $1, 2, \ldots, N$ are transient for the $\{0, 1, 2, \ldots, N\}$-valued chain whose transition matrix is P, hence, by Exercise 10.5 of Chapter 2,

$$\sum_{n=0}^{\infty} (P^n)_{xy} < \infty, \quad 1 \le x, y \le N.$$

But $(P^n)_{xy} = (R^n)_{xy}$, $1 \le x$, $y \le N$, from which the result follows. □

We can now prove the following result.

Theorem 9.2 *Under assumption (8.3), if, for all* $1 \le p \le N$,

$$\sum_{n=1}^{\infty} \prod_{r=1}^{n} \frac{\lambda_p}{\mu_p(r)} < \infty,$$

then the E*-valued jump Markov process* $\{X_t\}$ *possesses the unique invariant probability distribution*

$$\pi_x = \prod_{p=1}^{N} \pi_{x_p}^p,$$

where, for $1 \le p \le N$, *the probability distribution* π^p *on* $\mathbb{N}$ *is defined by*

$$\pi_n^p = b_p \frac{\lambda_p^n}{\prod_{r=1}^{n} \mu_p(r)},$$

with

$$b_p = \left(1 + \sum_{n=1}^{\infty} \prod_{r=1}^{n} \frac{\lambda_p}{\mu_p(r)}\right)^{-1}.$$

Before proceeding, let us first state the result more precisely in two particular cases.

Corollary 9.3 *In the case* $\mu_p(n) = \mu_p \mathbf{1}_{\{n > 0\}}$ *($M/M/1$ queue network), if (8.3) is satisfied and* $\lambda_p < \mu_p$ *for all* p, *then the process* $\{X_t\}$ *possesses the invariant probability* $\pi_x = \prod_{p=1}^{N} \pi_{xp}^p$, *where*

$$\pi_n^p = \left(1 - \frac{\lambda_p}{\mu_p}\right)\left(\frac{\lambda_p}{\mu_p}\right)^n, \quad 1 \le p \le N, \ n \in \mathbb{N}.$$

Corollary 9.4 *In the case* $\mu_p(n) = \mu_p \times \left(n \wedge s_p\right)$ *($M/M/s_p$ queue network), if (8.3) is satisfied and* $\lambda_p < \mu_p s_p$, *then the process* $\{X_t\}$ *possesses the invariant probability* $\pi_x = \prod_{p-1}^{N} \pi_{x_p}^p$, *with*

$$\pi_n^p = b_p \frac{(\lambda_p/\mu_p)^n}{(n \wedge s_p)!}$$

and

$$b_p = \left[\sum_{r=0}^{s_p - 1} \frac{(\lambda_p/\mu_p)^r}{r!} + \frac{(\lambda_p/\mu_p)^{s_p}}{s_p!}\left(\frac{1}{1 - \lambda_p/\mu_p s_p}\right)\right]^{-1}.$$

Proof of Theorem 9.2 If an invariant probability π exists, and if $\hat{Q}$ denotes the infinitesimal generator of the time-reversed process, then

$$\pi_x Q_{x,x+e_p} = \pi_{x+e_p} \hat{Q}_{x+e_p,x},$$

$$\pi_{x+e_p} Q_{x+e_p,x} = \pi_x \hat{Q}_{x,x+e_p},$$

$$\pi_{x+e_q} Q_{x+e_p,x+e_p} = \pi_{x+e_p} \hat{Q}_{x+e_p,x+e_q}.$$

If π exists and is given by the formula in the statement, then

$$\hat{Q}_{x+e_p,x} = \frac{\lambda_p^0}{\lambda_p}\mu_p(x_p+1),$$

$$\hat{Q}_{x,x+e_p} = \lambda_p r_{p0},$$

$$\hat{Q}_{x+e_p,x+e_q} = \frac{\lambda_q}{\lambda_p} r_{pq}\mu_p(x_p+1).$$

The formula for π then follows from Theorem 7.4 of Chapter 7, provided that, for all $x \in E$,

$$\sum_{y\neq x} \hat{Q}_{xy} = \sum_{y\neq x} Q_{xy}.$$

But

$$\sum_{y\neq x} Q_{xy} = \sum_{p=1}^{N} \left(\lambda^0 + \mu_p(x_p)\right),$$

and

$$\begin{aligned}\sum_{y\neq x} \hat{Q}_{xy} &= \sum_{p=1}^{N} \left(\lambda_p r_{p0} + \frac{\lambda_p^0}{\lambda_p}\mu_p(x_p) + \sum_{q\neq p} \lambda_q r_{qp} \frac{\mu_p(x_p)}{\lambda_p}\right) \\ &= \sum_{p=1}^{N} \left(\lambda_p r_{p0} + \mu_p(x_p)\right),\end{aligned}$$

where we have used (8.4). But summing (8.4) over p, we get

$$\sum_p \lambda_p^0 = \sum_p \lambda_p \left(1 - \sum_q r_{pq}\right) = \sum_p \lambda_p r_{p0}$$

from which the desired identity follows. □

The formula for the infinitesimal generator of the time-reversed process follows from the proof of the theorem, hence we have the following corollary.

Corollary 9.5 *Under the conditions of Theorem 9.2, if $\{X_t\}$ is initialized with its invariant probability distribution, then the process reversed at time T, $\hat{X}_t = X_{T-t}$, $0 \le t \le T$, is a jump Markov process, which can be interpreted as a Jackson network with N interconnected queues, with the following parameters: intensity of the exogenous arrivals at station p,*

$$\hat{\lambda}_p^0 = \lambda_p r_{p0};$$

routing probability from p to q,

$$\hat{r}_{pq} = \frac{\lambda_q}{\lambda_p} r_{qp};$$

exit probability when leaving station p,

$$\hat{r}_{p0} = \frac{\lambda_p^0}{\lambda_p};$$

and service rate at station p,

$$\hat{\mu}_p(x_p) = \mu_p(x_p).$$

Moreover, the exit processes from the network for the initial process from stations $\{1, \ldots, N\}$ *are mutually independent Poisson point processes, each having the intensity* $\lambda_p r_{p0}$.

8.10 Closed Jackson network

Let us again consider the model of the preceding section, except that there is no exogenous arrival ($\lambda_p^0 = 0$, forall p), and no exit from the network ($r_{p0} = 0$, for all p), that is,

$$r_{pp} = 0, \quad \sum_{q=1}^{N} r_{pq} = 1.$$

The matrix $R = (r_{pq})_{1 \le p,q \in N}$ is then Markovian. We suppose in what follows that it is *irreducible*. Under these assumptions, there is clearly conservation of the total number of customers in the network. Let $I(>0)$ stand for this number.

Here $\{X_t\}$ is a jump Markov process with values in

$$E(I) = \{x \in \mathbb{N}_+^N,\ x_1 + \ldots + x_p = I\}.$$

The infinitesimal generator of $\{X_t\}$ is given by

$$Q_{x+e_p, x+e_q} = \mu_p(x_p + 1) r_{pq}, \quad p \neq q,\ x \in E(I).$$

Let $\{\lambda_p;\ 1 \le p \le N\}$ denote the probability distribution on $\{1, \ldots, N\}$ which is R-invariant. We then show, for example by the same method as in the preceding section, the following result.

Theorem 10.1 *If R is irreducible,then the process* $\{X_t\}$ *admits the unique invariant probability*

$$\pi_x^I = G(I)^{-1} \prod_{p=1}^{N} \frac{\lambda_p^{x_p}}{\prod_{r=1}^{x_p} \mu_p(r)},$$

with

$$G(I) = \sum_{x \in E(I)} \prod_{p=1}^{N} \frac{\lambda_p^{x_p}}{\prod_{r=1}^{x_p} \mu_p(r)}.$$

The main difficulty is in the computation of the normalization constant (also called the 'partition function') $G(I)$. If the values of I and N are not small, the cardinality of $E(I)$ (which is equal to C_{I+N-1}^{N-1}) makes the computation of $G(I)$ by direct summation impossible.

Suppose now for the rest of this section that $\mu_p(r) = \mu_p \mathbf{1}_{\{r>0\}}$. Define $\rho_p = \lambda_p/\mu_p$.

Mean number of customers Let $1 \leq k < I$. For all $y \in E(I-k)$, note that

$$\pi^I_{y+ke_p} G(I) = \pi^{I-k}_y \rho_p^k G(I-k),$$

if we let π^I denote the invariant probability on $E(I)$. Hence if $\{X_t\}$ is initialized with the invariant probability π^I, and if $1 \leq k < I$, then

$$\begin{aligned}
\mathbb{P}(X_t^p \geq k) &= \sum_{y \in E(I-k)} \pi^I_{y+ke_p} \\
&= \frac{G(I-k)}{G(I)} \rho_p^k \sum_{y \in E(I-k)} \pi_y^{I-k} \\
&= \frac{G(I-k)}{G(I)} \rho_p^k.
\end{aligned}$$

Hence,

$$\mathbb{E}X_t^p = \sum_{k=1}^{I} \rho_p^k \frac{G(I-k)}{G(I)},$$

where $G(0) = 1$.

Intensity of the arrivals At equilibrium, the intensity of departures from station p towards station q equals

$$d_{pq} = \sum_{x \in E(I), x_p > 0} \pi_x Q_{x, x-e_p+e_q}.$$

From the bijection between $\{x \in E(I),\ x_p > 0\}$ and $E(I-1)$, and the identity

$$\pi^I_x G(I) = \pi^{I-1}_{x-e_p} G(I-1)\, \rho_p,$$

we deduce the formula

$$d_{pq} = \frac{\lambda_p r_{pq} G(I-1)}{G(I)}.$$

We then obtain, by an argument based upon time reversal, the following formula for the intensity of the arrivals at station p:

$$a_p = \frac{\lambda_p G(I-1)}{G(I)}.$$

Computation of the partition function Consider the generating functions

$$\Phi_p(z) = \sum_{n=0}^{\infty} (\rho_p z)^n \frac{1}{1-\rho_p z}, \quad 1 \le p \le N, \ |z| \le 1.$$

Comparing the coefficients of the z^J on both sides, we obtain the identity

$$\sum_{n=0}^{\infty} G(n) z^n = \prod_{p=1}^{N} \Phi_p(z).$$

Define

$$\beta_p(z) = \prod_{q=1}^{p} \Phi_q(z), \quad 1 \le p \le N, \ |z| \le 1.$$

We have

$$\beta_p(z)(1-\rho_p z) = \beta_{p-1}(z), \quad p > 1.$$

Denote by $T(k, p)$ the coefficient of z^k in $\beta_p(z)$. We have

$$T(k, p) - \rho_p T(k-1, p) = T(k, p-1), \quad k \ge 1, \ p > 1,$$

with $T(0, p) = 1$, $p \ge 1$ and $T(k, 1) = \rho_1^k$, $n \ge 0$.

We deduce from those formulae an algorithm for computing $G(I) = T(I, N)$, of complexity $O(I \times N)$, which involves computing the $T(k, p)$ row by row for $1 \le p \le N$ (i.e. for increasing values of k).

8.11 Telephone network

We consider a telephone network which consists of N interconnected channels. Establishing a communication requires the exclusive use of (for example) n channels. If one or more of the requested channels is busy, the call is rejected. Otherwise, communication is established and lasts a random length of time, whose law is exponential with parameter μ_n. The calls requesting n channels arrive according to a Poisson process with intensity λ_n. The flow of arrivals of different types and the durations of the calls are mutually independent.

Let $X_t = (X_t^1, \ldots, X_t^P)$ where X_t^n represents the number of communications established at time t, which request n channels. X_t is a jump Markov process with values in

$$E = \{x = (x_1, \ldots, x_P) \in \mathbb{N}^P | \bar{x} \le N\},$$

where $\bar{x} = x_1 + 2x_2 + \ldots + Px_P$, with infinitesimal generator Q, whose only non-zero off-diagonal entries are given by

$$Q_{x+e_n,x} = \mu_n \times (x_n + 1),$$

$$Q_{x,x+e_n} = \lambda_n \frac{C_{N-\bar{x}}^n}{C_N^n},$$

where $e_n = (0, \ldots, 0, 1, 0, \ldots, 0)$, the vector in E whose entries are all zero, except for the nth which is 1. Note that $C^n_{N-\bar{x}}/C^n_N$ is the probability that a customer is not rejected if $X_t = x$. Define $\tau_n = (C^n_N)^{-1}\lambda_n$.

Invariant probability

Theorem 11.1 *The jump Markov process $\{X_t\}$ described above is time-reversible with respect to its invariant probability*

$$\pi_x = \pi_0 \frac{N!}{(N-\bar{x})!} \prod_{n=1}^{P} \left\{ \left(\frac{\tau_n}{\mu_n n!} \right)^{x_n} \frac{1}{x_n!} \right\}.$$

PROOF Let us look for a solution to the detailed balance equation

$$\pi_x \tau_n \frac{(N-\bar{x})!}{n!(N-\bar{x}-n)!} = \pi_{x+e_n} \mu_n (x_n + 1), \quad x \in E,\ 1 \le n \le P,\ \bar{x} + e_n \in E;$$

that is,

$$\pi_{x+e_n} = \frac{\tau_n}{\mu_n} \frac{1}{x_n + 1} \frac{(N-\bar{x})!}{n!(N-\bar{x}-n)!} \pi_x.$$

Hence, if $x_n \ge 1$,

$$\pi_x = \frac{\tau_n}{\mu_{nn}!} \frac{1}{x_n} \frac{(N-\bar{x}+n)!}{(N-\bar{x})!} \pi_{x-e_n}.$$

So

$$\pi_x = \pi_{x-x_1 e_1} \left(\frac{\lambda_1}{\mu_1} \right)^{x_1} \frac{1}{x_1!} \frac{(N-\bar{x}+x_1)!}{(N-\bar{x})!},$$

$$\pi_x = \pi_{x-x_1 e_1 - x_2 e_2} \left(\frac{\lambda_1}{\mu_1 1!} \right)^{x_1} \frac{1}{x_1!} \left(\frac{\lambda_2}{\mu_2 2!} \right)^{x_2} \frac{1}{x_2!} \frac{(N-\bar{x}+x_1+2x_2)!}{(N-\bar{x}!}.$$

The formula in the statement can be obtained by completing this computation. □

It remains to compute π_0. Define

$$y_n = \frac{\tau_n}{\mu_n n!}, \quad y = (y_1, \ldots, y_P),$$

$$G(N, x) = \sum_{x \in E} \frac{N!}{(N-\bar{x})!} \prod_{n=1}^{P} \frac{y_n^{x_n}}{x_n!}.$$

Then $\pi_0 = G(N, y)^{-1}$. We shall see below an algorithm for the computation of the partition function $G(N, y)$.

Probability of rejection In the stationary regime, the successive instants when communications requesting n channels end constitute a Poisson point process with intensity

$$\sum_{x \in E} \mu_n x_n \pi_x = \mathbb{E}(X_t^n)\mu_n.$$

Reversibility implies that this quantity equals the intensity of accepted communications requesting n channels. Hence the ratio between the rate of acceptance of and rate of demand for communications requesting n channels, which represents the probability of acceptance, equals $\mathbb{E}(X_t^n)\mu_n/\lambda_n$. It then remains to compute the mean number of communications requesting n channels.

Mean number of communications requesting n channels We have

$$\sum_{x \in E} x_n \pi_x = G(N, y)^{-1} \sum_{x \in E} \frac{N!}{(N - \bar{x})!} x_n \prod_{m=1}^{P} \frac{y_m^{x_n}}{x_m!}.$$

But

$$\frac{\partial}{\partial y_n} G(N, y) = \sum_{x \in E} \frac{N!}{(N - \bar{x})!} \frac{y_n^{x_n - 1}}{(x_n - 1)!} \prod_{m \neq n} \frac{y_m^{x_m}}{x_m!}.$$

Hence,

$$\sum_{x \in E} x_n \pi_x = y_n \; G(N, y)^{-1} \frac{\partial}{\partial y_n} G(N, y).$$

With the convention that $n! = -\infty$ if $n < 0$, we have

$$\frac{\partial}{\partial y_n} G(N, y) = \sum_{x_1=0}^{\infty} \cdots \sum_{x_P=0}^{\infty} \frac{N!}{(N - \bar{x})!} \frac{y_n^{x_n - 1}}{(x_n - 1)!} \prod_{m \neq n} \frac{y_m^{x_n}}{x_m!},$$

that is,

$$\frac{\partial}{\partial y_n} G(N, y) = \frac{N!}{(N - n)!} \sum_{x_1=0}^{\infty} \cdots \sum_{x_P=0}^{\infty} \frac{(N - n)!}{(N - n - \bar{x})!} \prod_{m} \frac{y_m^{x_n}}{x_m!},$$

hence

$$\frac{\partial}{\partial y_n} \; G(N, y) = \frac{N!}{(N - n)!} \; G(N - n, y),$$

and

$$\mathbb{E}X_t^n = \frac{\lambda_n}{\mu_n} \; \frac{G(N - n, y)}{G(N, y)}.$$

Probability of rejection (continued) The probability of rejection of a communication requesting n channels then equals

$$p_n = 1 - \frac{G(N - n, y)}{G(N, y)}.$$

Computation of the partition function One can establish recurrence relations for the $G(N, y)$, which are similar to those obtained in the case of closed Jackson networks. Consider, for $z \in \mathbb{C}$, the generating function

$$g(z, y) = \sum_{N=0}^{\infty} \frac{z^N}{N!} G(N, y),$$

with $G(0, x) = 1$. One can show that

$$g(z, y) = \exp\left[z + \sum_{n=1}^{P} y_n z^n \right],$$

from which one deduces the recurrence formulae

$$G(0, y) = 1,$$

$$G(N, y) = G(N-1, y) + \sum_{n=1}^{N} n y_n \frac{(N-1)!}{(N-n)!} G(N-n, y), \quad N = 1, 2, \ldots, P;$$

$$G(N, y) = G(N-1, y) + \sum_{n=1}^{P} n y_n \frac{(N-1)!}{(N-n)!} G(N-n, y), \quad N \geq P.$$

8.12 Kelly networks

We now consider 'multi-class' networks, also called 'Kelly networks'. The main difference between Kelly and Jackson networks is that a given customer no longer follows a trajectory that is the result of random drawings that are the same for all, but rather a deterministic trajectory which depends upon the class to which the customer belongs. We consider customers of class $j = 1, 2, \ldots, J$, where $J \in \mathbb{N}$.

Since now all customers are no longer the same, in order to study the evolution of fluxes in the network it becomes crucial to specify the priority policy at each queue, which hitherto has not really been important (since we have only been interested in global fluxes, and not in what happens to a particular customer).

Let us consider first the case of a single queue.

8.12.1 Single queue

Arrivals consist of J Poisson point processes with intensity $\lambda_1, \ldots, \lambda_J$, λ_j denoting the intensity of arrivals of customers of class j. Whatever his class may be, a customer who joins a queue which already has n customers will be placed in position $\ell = 1, \ldots, n+1$, with probability $\gamma(\ell, n)$, where $\sum_{\ell=1}^{n+1} \gamma(\ell, n) = 1$.

If n customers are present in the queue, the customer in position ℓ receives a service which follows the exponential law with parameter $\Phi(n)\delta(\ell, n)$, where $\sum_{\ell=1}^{n} \delta(\ell, n) = 1$. Globally, the server then works with 'intensity' $\Phi(n) > 0$ when $n > 0$.

The state of the queue is described by a jump Markov process with values in

$$E = \emptyset \cup \bigcup_{n=1}^{\infty} \{1, \ldots, J\}^n.$$

A point $c \in E$ is a sequence of length $n = |c|$,

$$c = (c_1, \ldots, c_n),$$

where $c_i \in \{1, \ldots, J\}$, $1 \le i \le n$.

Let us specify the possible transitions of the E-valued process X_t which describes the state of the queue at time t. Two types of transition are possible from state c:

(i) addition of a customer of class j, added between the $(i-1)$th and the ith customers ($1 \le j \le J$, $1 \le i \le n+1$, $n \ge 0$), that is, transition from state c to state

$$A_i^j(c) = (c_1, \ldots, c_{i-1}, j, \ c_{i+1} \ldots, c_n);$$

(ii) departure of a customer who was at rank i in the queue, that is, transition from state c to state

$$S_i(c) = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n).$$

The transition matrix Q is completely characterized by its off-diagonal entries, given by:

$$Q_{c, A_i^j(c)} = \lambda_j \gamma(i)|c|), \quad 1 \le i \le |c| + 1, \ 1 \le j \le J;$$

$$Q_{c, S_i(c)} = \Phi(|c|)\delta(i, |c|), \quad 1 \le i \le |c|.$$

Note that a customer's class has no bearing on the way he is placed in the queue, nor upon his service.

Example 12.1 *For an $M/M/K/FIFO$ queue, characterized as 'first in, first out', we have:*

$$\Phi(n) = n \wedge K;$$

$$\delta(\ell, n) = \begin{cases} 1/n \wedge K, & \text{if } 1 \le \ell \le n \wedge K, \\ 0, & \text{if } \ell > n \wedge K; \end{cases}$$

$$\gamma(\ell, n) = \begin{cases} 0, & \text{if } \ell = 1, \ldots, n, \\ 1, & \text{if } \ell = n+1. \end{cases}$$

Example 12.2 *For an $M/M/K/LIFO$ queue, characterized as 'last in, first out', Φ and γ are as in Example 12.1, but now*

$$\delta(\ell, n) = \begin{cases} 1/n \wedge K, & \text{if } (n-K)^+ \le \ell \le n, \ n \ne 0, \\ 0, & \text{if } \ell \le (n-K)^+. \end{cases}$$

Example 12.3 *For an $M/M/1$ queue with process sharing, $\Phi(n) = \Phi > 0$, the choice of γ is unimportant, and*

$$\delta(\ell, n) = \frac{1}{n}, \quad 1 \le \ell \le n.$$

The jump Markov process $\{X_t\}$ with the above generator Q is clearly irreducible. Moreover, we have the following theorem.

Theorem 12.4 *A queue with several classes of customers is positive recurrent if and only if*

$$Z = \sum_{c \in E} \prod_{\ell=1}^{|c|} \frac{\lambda_{c_\ell}}{\Phi(\ell)} < \infty,$$

and in this case the invariant probability distribution is

$$\pi_c = Z^{-1} \prod_{\ell=1}^{|c|} \frac{\lambda_{c_\ell}}{\Phi(\ell)}, \quad c \in E.$$

Furthermore, at equilibrium, for $1 \le j \le J$, the departures process of class j customers is a Poisson point process with intensity λ_j.

Proof We shall exploit Theorem 7.4. of Chapter 7 Let

$$\begin{aligned}
\hat{Q}_{c, A_i^j(c)} &= \frac{\pi_{A_i^j(c)}}{\pi_c} Q_{A_i^j(c), c} \\
&= \frac{\lambda_j}{\Phi(|c|+1)} \Phi(|c|+1) \delta(i, |c|+1) \\
&= \lambda_j \delta(i, |c|+1), \\
\hat{Q}_{c, S_i(c)} &= \frac{\Phi(|c|)}{\lambda_i} Q_{S_i(c), c} \\
&= \frac{\Phi(|c|)}{\lambda_i} \lambda_i \gamma(i, |c|-1) \\
&= \Phi(|c|) \gamma(i, |c|-1).
\end{aligned}$$

Note that

$$\sum_{i=1}^{|c|+1} \hat{Q}_{c, A_i^j(c)} = \lambda_j = \sum_{i=1}^{|c|+1} Q_{c, A_i^j(c)},$$

$$\sum_{i=1}^{|c|} \hat{Q}_{c, S_i(c)} = \Phi(|c|) = \sum_{i=1}^{|c|} Q_{c, S_i(c)}.$$

It then follows from Theorem 7.4 of Chapter 7 that under the assumption of the present theorem, π is the invariant probability distribution, $\hat{Q}$ is the generator of

the time-reversed process, and arrivals of class j customers constitute a Poisson point process with intensity $\lambda_j = \sum_{i=1}^{|c|+1} \hat{Q}_{c,A_i^j(c)}$. □

Note that the total number of customers in the queue is a Markov birth and death process with generator Q characterized by

$$Q_{i,i+1} = \sum_{j=1}^{J} \lambda_j, \quad Q_{i,i-1} = \Phi(i).$$

However, the above detailed description will be necessary in what follows.

8.12.2 Multi-class network

Consider now a network consisting of N nodes, each being a queue of the above type. For each $1 \leq j \leq J$, class j customers arrive in the network according to a Poisson point process with intensity λ_j. Each class j customer first joins queue $f_1^j \in \{1, \ldots, N\}$, then when he leaves that queue he joins queue f_2^j, and so on until $f_{n_j}^j$, and then he leaves the network. This means that the jth flow follows the route $f_1^j, f_2^j, \ldots, f_{n_j}^j$ in the network. At each node i, $1 \leq i \leq N$, the functions Φ, δ, γ associated with the node i are denoted by Φ_i, δ_i, γ_i.

There is no reason to forbid tours $f_1^j, \ldots, f_{n_j}^{j'}$ from visiting certain nodes of the network several times. For this reason, and in order to make the process Markovian, we need to attach to each customer present in the network not only his class, but also the number of queues already visited.

The state of the ith queue ($1 \leq i \leq N$) is then described by the vector

$$x_i = ((c_{i_1}, s_{i_1}), \ldots, (c_{im_i}, s_{im_i})),$$

where c_{ik} denotes the class of the customer who occupies the kth position, and s_{ik} the number of queues already visited (including the one in which he presently stands). Hence, $1 \leq c_{ik} \leq J$ and $1 \leq s_{ik} \leq n_{c_{ik}}$. The state space describing the queue at node i is

$$E_i = \emptyset \cup \bigcup_{n=1}^{\infty} \{1, \ldots, J\}^n, \quad X_t = (X_t^1, \ldots, X_t^N),$$

where $X_t^i \in E_i$ is the state of the queue at node i at time t. The jump Markov process $\{X_t\}$ takes its values in $E = \prod_{i=1}^{N} E_i$. It is an irreducible process. The possible transitions from the state $x = (x_1, \ldots, x_N)$ are as follows:

1. A class j customer arrives, according to a Poisson point process with intensity λ_j, in the network at node f_1^j. The pair $(j, 1)$ is inserted in the ℓth position in that queue with probability $\gamma_{f_1^j}(\ell, |c_{f_1^j}|)$. This happens with intensity

$$\lambda_j \gamma_{f_1^j}(\ell, |c_{f_1^j}|).$$

2. A class j customer at step $s < n_j$ of his journey, at position ℓ in the queue at node f_s^j, leaves that queue for position m in the queue f_{s+1}^j, with intensity
$$\Phi_{f_s^j}(|c_{f_s^j}|)\delta_{f_s^j}(\ell, |c_{f_s^j}|)\gamma_{f_{s+1}^j}(m, |c_{f_{s+1}^j}|).$$
After that transition, the pair $(j, s+1)$ is at position m in the queue at node f_{s+1}^j.

3. A class j customer at step n_j of his journey, at position ℓ in the queue at node f_{nj}^j, leaves the network. This happens with intensity
$$\Phi_{f_{n_j}^j}(|c_{f_{n_j}^j}|)\delta_{f_{n_j}^j}(\ell, |c_{f_{n_j}^j}|).$$

Let
$$Z = \sum_{x \in E} \prod_{i=1}^{N} \prod_{k=1}^{|x_i|} \frac{\lambda_{c_{ik}}}{\Phi_i(k)}.$$

We have the following result, which can be proved similarly to the previous theorem.

Theorem 12.5 *If $Z < \infty$, then $\{\pi_x;\ x \in E\}$ defined by*
$$\pi_x = Z^{-1} \prod_{i=1}^{N} \pi_{x_i}^i, \quad \text{with } \pi_{x_i}^i = \prod_{k=1}^{|x_i|} \frac{\lambda_{c_{ik}}}{\Phi_i(k)},$$
is the invariant probability distribution of the process $\{X_t\}$ which describes the customers present in the network of multi-class queues. Moreover, at equilibrium the departures process of class j customers is a Poisson point process with intensity λ_j, $1 \le j \le J$.

8.13 Exercises

Exercise 13.1 *We study the $\mathbb{N}$-valued jump Markov process $\{X_t\}$ which models the $M/M/1$ queue, that is, the jump Markov process with infinitesimal generator*
$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

1. *Show that the embedded chain is the reflected random walk from Exercise 10.14 of Chapter 2, with $p = \lambda/(\lambda+\mu)$.*

2. *Deduce that the process* $\{X_t\}$ *is transient in the case* $\lambda > \mu$, *and recurrent in the case* $\lambda \leq \mu$.

3. *Show that* $\{X_t\}$ *is null recurrent in the case* $\lambda = \mu$, *positive recurrent in the case* $\lambda < \mu$. *Note that in the former (latter) case, the measure* $(1, 1, 1, 1, \ldots)$ *(the geometric measure with the parameter* λ/μ*) is invariant.*

Exercise 13.2 *Consider the* $M/M/1$ *queue* $\{X_t;\ t \geq 0\}$, *and define the random sequence* $\{Y_n = X_{T_n^-};\ n \geq 1\}$, *where* $\{T_1, T_2, \ldots\}$ *denotes the successive arrival times of customers starting at time* 0, *and* $X_{T_n^-} \geq 0$ *is then the number of customers which the nth arriving customer finds in front of him in the queue.*

1. *Show that* $\{Y_n;\ n \geq 1\}$ *is an* $\mathbb{N}$*-valued Markov chain. Specify its transition matrix. Show that this chain is irreducible and aperiodic.*

2. *Consider the case* $\lambda < \mu$. *Show that the geometric probability distribution* π *with parameter* λ/μ *(which is the invariant probability distribution of the jump Markov process* $\{X_t;\ t \geq 0\}$*) is the invariant probability distribution of the chain* $\{Y_n;\ n \geq 1\}$, *and that the law of* Y_n *converges to* π *as* $n \to \infty$.

3. *Show that Little's formula in the particular case of the* $M/M/1$ *queue is a consequence of the result of part 2.*

4. *Suppose now that the queue is initialized with its invariant probability distribution (i.e. the law of* X_0 *is* π*). Compute the law of* $X_{T_1^-}$. *Show that for any increasing function* f *from* $\mathbb{N}$ *into* $\mathbb{R}$, $\mathbb{E}f(X_{T_1^-}) \leq \sum_{x=0}^{\infty} \pi_x f(x)$, *and that the inequality is strict if* f *is not constant. Is this result in accordance with your intuition? Why does the law of* $X_{T_1^-}$ *differ from* π*? Compare with the result of Exercise 5.3. of Chapter 6*

9

Introduction to mathematical finance

Introduction

The goal of this chapter is to present mathematical models which allow us to solve the problem of pricing European and American options, as well as to specify the associated hedging strategies. We shall present in particular the famous Black–Scholes formula; see [6]. In addition to the continuous model of Black and Scholes, we shall develop the discrete model of Cox, Ross and Rubinstein [11].

The nice feature of the discrete model is that it allows elementary proofs of the results; that of the continuous model is that it leads to formulae that are in constant use by finance practitioners. We shall introduce the necessary tools from stochastic calculus, which will allow us to describe diffusion processes which are continuous time $\mathbb{R}^d$-valued Markov processes. Note that this is the only part of this book where a few fundamental results are stated without proof. Including all proofs would have made the book too long. There are by now plenty of well-written manuals on stochastic calculus.

The chapter closes with an introduction to interest rate and bond models.

We found our inspiration for this chapter in the book by Lamberton and Lapeyre [24], as well as in [4] and [29].

In this chapter, we shall always use t to denote time, whether it is discrete ($t = 0, 1, 2, \ldots, T$ or $t \in \mathbb{N}$) or continuous ($0 \leq t \leq T$ or $t \geq 0$), unlike in Chapters 1–5.

9.1 Fundamental concepts

In this chapter, we consider an investor who can split his assets between two types of investments: a bank account deposit, with a fixed and guaranteed return (constant

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

interest rate), and a risky investment, which involves, for example, buying shares on the stock market. We shall mainly be concerned with the simple case where there is only one type of available stock (except in Section 9.3.10). The price at time t of that asset will be denoted by S_t. We shall present two probabilistic models for the fluctuations of $\{S_t\}$, one in discrete time and the other in continuous time. In our model, the economic agent acts only as an investor, he does not consume. He is a 'small' investor, in the sense that his investment choices have no effect on the evolution of the price of the risky asset. Moreover, we allow him to buy an option based on the risky asset. We should also mention that in our models we neglect transaction costs.

9.1.1 Option

An *option* is a contract which gives its holder the right (but not the obligation) to buy (in the case of a *call* option) or to sell (in the case of a *put* option) a fixed quantity of a given asset (which can be a stock, a bond, a currency, a raw material, ...) at a price which is fixed by the contract (called the *strike price*), at a time (the *exercise time*) which is fixed by the contract in the case of a European option, or at any time between when the contract is signed and when it expires in the case of an American option. An option is also called a *derivative*, since its value is derived from an underlying asset.

In the case of a European call option with exercise time T, based on an asset whose price at time t is S_t, and with strike price K, the option holder makes at time T a profit equal to $(S_T - K)_+$. Indeed, he earns $S_T - K$ per unit of the asset by exercising the option if $S_T > K$ (buying at price K and selling at the market price S_T), and he neither gains nor loses anything by not exercising the option if $S_T \leq K$. An analogous argument tells us that, in the case of a put, the option holder's profit at time T is $(K - S_T)_+$. The profit of the holder (i.e. the buyer) of the option is the loss of the seller of the same option. The price of this option, also called the *premium*, is supposed to compensate for this loss.

The mathematical theory of options treats two problems:

(a) *option pricing*, that is, the premium that the buyer of the option pays the seller;

(b) *hedging*, that is, what the seller of the option should do with the money he receives for selling the option, in order to compensate for a loss at time T of $(S_T - K)_+$ ($(K - S_T)_+$) in the case of a European call (put) option.

9.1.2 Arbitrage

One of the basic assumptions which will be needed for our model is the *absence of arbitrage opportunity* – it should not be possible to earn money without taking any risk; in other words, 'there is no such thing as a free lunch'. This assumption implies the so-called *call–put parity relation* for European options:

Proposition 1.1 *If there is no arbitrage opportunity, then the prices C_t and P_t at time t of a call and a put option with the same exercise time T and strike price K satisfy the relation*

$$C_t - P_t = S_t - Ke^{-r(T-t)},$$

where r is the bank deposit interest rate.

Remark 1.2 *Here and below, we assume that the interest rates for depositing and borrowing money are the same, and they equal the constant r. Of course this assumption is not realistic. This is crucial for our model to be linear, and for us to be able to deduce the explicit Black–Scholes formula. A generalized Black–Scholes model which avoids this assumption will be presented in Section 9.3.6.*

Proof Suppose that the parity relationship is not satisfied, that is, assume that at time t we have, for example,

$$C_t - P_t > S_t - Ke^{-r(T-t)}$$

(an analogous argument applies to the 'less than' case). We will deduce from this inequality an arbitrage opportunity. At time t we buy one stock (or bond, or ...), and also buy a put and sell a call. The combination of these operations gives a net profit of

$$X_t = C_t - P_t - S_t.$$

If $X_t > 0$, we deposit X_t at the bank (i.e. at rate r) until time T; otherwise, we borrow $-X_t$ at the same rate until time T.

At time T, either of two things can happen:

1. $S_T > K$. In this case we exercise the call (and not the put) – we receive K, withdraw the deposit (or repay the loan), so that we end up with a wealth of
$$K + e^{r(T-t)}(C_t - P_t - S_t) > 0.$$
2. $S_T \leq K$. In this case we exercise the put (and not the call), and close the bank account as above, so that we end up with the same wealth as above.

In both cases we realize a strictly positive profit at time T without investing any money at time t – this is an example of arbitrage. □

9.1.3 Viable and complete markets

A market is said to be *viable* if there exists no arbitrage opportunity.

A market is said to be *complete* if any asset at time T (i.e. any function of $\{S_t;\ 0 \leq t \leq T\}$, in particular of S_T, for example $(S_T - K)_+$ or $(K - S_t)_+$) can be realized, in the sense that there exists an admissible strategy which, starting with a given initial wealth at time 0, produces exactly that wealth almost surely at time T.

The notion of an admissible strategy will be made precise below, in both the discrete and continuous models. It is a self-financing strategy, which follows decisions based upon past information only. The fair price of a European call (put) will then be the initial value of an admissible strategy with final value equal to $(S_T - K)_+$ $((K - S_T)_+)$. Such a strategy acts as a *hedge* for the option, and is called the *replication strategy*. It allows the seller to make sure that he does not lose money (and does not make any profit either), no matter how the market fluctuates.

9.2 European options in the discrete model

9.2.1 The model

We consider a discrete time model with a single risky asset, whose price at time t is denoted by S_t, $t = 0, 1, \ldots, T$, and a riskless asset whose price at time t is denoted by R_t. We assume that there exists $r>0$ such that

$$R_{t+1} = R_t(1 + r).$$

For simplicity we take $R_0 = 1$, hence

$$R_t = (1 + r)^t, \quad 0 \leq t \leq T.$$

We assume that S_0 is constant, and that there exist i.i.d. random variables ξ_t, $1 \leq t \leq T$, taking their values in the set $\{d, u\}$, where $0 < d < u$, such that

$$S_{t+1} = S_t\xi_{t+1}, \quad t = 0, 1, \ldots, T - 1.$$

Our probability space here is $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \{d, u\}^T$, $\mathcal{F} = \mathcal{P}(\Omega)$, and $\mathbb{P}$ is such that the random variables ξ_t, $1 \leq t \leq T$, are i.i.d. with $\mathbb{P}(\xi_1 = d)>0$ and $\mathbb{P}(\xi_1 = u)>0$.

We define the *discounted price* of the risky asset at time t as the quantity

$$\widetilde{S}_t = \frac{S_t}{R_t}, \quad t = 0, 1, \ldots, T.$$

9.2.2 Admissible strategy

A *strategy* is an $\mathbb{R}^2$-valued random sequence $\{(X_t, Y_t);\ t = 0, 1, \ldots, T\}$ such that

$$\mathcal{F}_{-1} = \mathcal{F}_0 = \{\emptyset, \Omega\},$$
$$\mathcal{F}_t = \sigma\{\xi_1, \ldots, \xi_t\}, \quad t \geq 1;$$

(X_t, Y_t) is assumed to be $\mathcal{F}_{t-1}$-measurable, for each $0 \leq t \leq T$. We say that the sequence $\{(X_t, Y_t)\}$ is *predictable*.

The value of the *portfolio* at time t is given by

$$V_t(X, Y) = X_t R_t + Y_t S_t,$$

and its *discounted value* is the quantity

$$\widetilde{V}_t(X, Y) = \frac{V_t(X, Y)}{R_t} = X_t + Y_t \widetilde{S}_t.$$

The strategy is said to be *self-financing* (with neither addition nor subtraction of money) if

$$X_t R_t + Y_t S_t = X_{t+1} R_t + Y_{t+1} S_t$$

or, equivalently,

$$V_{t+1}(X, Y) - V_t(X, Y) = X_{t+1}(R_{t+1} - R_t) + Y_{t+1}(S_{t+1} - S_t);$$

in other words,

$$X_t + Y_t \widetilde{S}_t = X_{t+1} + Y_{t+1} \widetilde{S}_t.$$

This can be rewritten as

$$\widetilde{V}_{t+1}(X, Y) - \widetilde{V}_t(X, Y) = Y_{t+1}(\widetilde{S}_{t+1} - \widetilde{S}_t).$$

With the notation $\Delta S_t = S_t - S_{t-1}$, $\Delta \widetilde{S}_t = \widetilde{S}_t - \widetilde{S}_{t-1}$, we have the following proposition.

Proposition 2.1 *The following three conditions are equivalent:*

(i) The strategy $\{(X_t, Y_t); 0 \le t \le T\}$ *is self-financing.*

(ii) For all $1 \le t \le T$,

$$V_t(X, Y) = V_0(X, Y) + \sum_{s=1}^{t} (X_s \Delta R_s + Y_s \Delta S_s).$$

(iii) For all $1 \le t \le T$,

$$\widetilde{V}_t(X, Y) = \widetilde{V}_0(X, Y) + \sum_{s=1}^{t} Y_s \Delta \widetilde{S}_s.$$

Proposition 2.2 *For any predictable process* $\{Y_t;\ 0 \le t \le T\}$ *and any deterministic initial condition* V_0 *of the portfolio, there exists a unique predictable process* $\{X_t;\ 0 \le t \le T\}$ *such that the strategy* $\{(X_t, Y_t);\ 0 \le t \le T\}$ *is self-financing and corresponds to a portfolio with initial value* V_0.

PROOF The self-financing condition implies that, for all $0 \leq t \leq T$,

$$\begin{aligned}\widetilde{V}_t(X, Y) &= X_t + Y_t \widetilde{S}_t \\ &= V_0 + \sum_{s=1}^{t} Y_s \Delta \widetilde{S}_s,\end{aligned}$$

which defines X_t. Predictability can easily be verified. □

Definition 2.3 *A strategy (X, Y) is said to be admissible if it is self-financing and satisfies $V_t(X, Y) \geq 0$, for all $0 \leq t \leq T$.*

Definition 2.4 *An arbitrage strategy (X, Y) is an admissible strategy such that $V_0(X, Y) = 0$ and $V_T(X, Y) \not\equiv 0$, or equivalently $V_0(X, Y) = 0$ and $\widetilde{V}_T(X, Y) \not\equiv 0$.*

9.2.3 Martingales

Definition 2.5 *A sequence $\{M_t;\ 0 \leq t \leq T\}$ is said to be adapted if M_t is $\mathcal{F}_t$-measurable, $0 \leq t \leq T$; it is a martingale if it is adapted and, for all $1 \leq t \leq T$,*

$$\mathbb{E}[M_t | \mathcal{F}_{t-1}] = M_{t-1}.$$

Proposition 2.6 *Let $\{M_t;\ 0 \leq t \leq T\}$ be a martingale, and $\{Y_t;\ 0 \leq t \leq T\}$ a predictable sequence. Then the sequence $\{M(Y)_t;\ \ 0 \leq t \leq T\}$ defined by*

$$\begin{aligned}M(Y)_0 &= Y_0 M_0, \\ M(Y)_t &= Y_0 M_0 + \sum_{1 \leq s \leq t} Y_s \Delta M_s, \quad t \geq 1,\end{aligned}$$

is a martingale.

PROOF It suffices to note that

$$\mathbb{E}[Y_t \Delta M_t | \mathcal{F}_{t-1}] = Y_t \ \mathbb{E}[\Delta M_t | \mathcal{F}_{t-1}] = 0,$$

where we have used first the predictability of Y and then the martingale property of M. □

Corollary 2.7 (Optional stopping theorem) *Let $\{M_t;\ 0 \leq t \leq T\}$ be a martingale, and τ a stopping time which is bounded by T, that is, an $\{0, 1, 2, \ldots, T\}$-valued random variable such that, for all $0 \leq t \leq T$, $\{\tau = t\} \in \mathcal{F}_t$. Then*

$$\mathbb{E}M_\tau = \mathbb{E}M_0.$$

PROOF It suffices to note that $M_\tau = M(Y)_T$ if Y is defined by

$$Y_t = \mathbf{1}_{\{\tau > t-1\}}, \quad 0 \le t \le T,$$

and that if M is a martingale, $\mathbb{E}M_t$ is a constant (independent of t). □

Remark 2.8 *Although it plays a very important part in the theory of stochastic processes, the martingale appears here for only the second time in this book – the first appearance having been in Exercise 11.5 of Chapter 7. As that exercise shows, there is a very strong connection between the notions of Markov process and martingale. Constructing the law of a Markov process is equivalent to solving a martingale problem, that is, finding a probability distribution under which a large class of processes are martingales. Martingales also appear naturally in the formulation of rather simple games. But their introduction becomes really useful when one wishes to prove difficult results, which is not what we are doing in this book. Martingales have very important mathematical properties, notably that one knows how to estimate their fluctuations. A simple but deep result concerning martingales is the optional stopping theorem which we have just proved. At the relatively elementary level of this book, almost the only the property of martingales which we use is the fact that if* $\{M_t;\ t \ge 0\}$ *is a martingale, then* $\mathbb{E}[M_t] = \mathbb{E}[M_0]$*, which gives* $\mathbb{E}[M_t]$ *since* $\mathbb{E}[M_0]$ *is usually known.*

9.2.4 Viable and complete market

Theorem 2.9 *The market defined above is viable (i.e. there is no arbitrage strategy) if and only if* $a < 1 + r < b$.

PROOF It is not hard to show (and we leave it as an exercise) that if $1 + r \notin]a, b[$, then there exists an arbitrage strategy.

On the other hand, if $a < 1 + r < b$, the probability distribution $\mathbb{P}^*$ on $(\Omega, \mathcal{F})$ such that the random variables ξ_t are i.i.d. with

$$\mathbb{E}^*(\xi_t) = 1 + r$$

(called the *risk-neutral probability*) is equivalent to $\mathbb{P}$ (since $\mathbb{P}^*(\xi_1 = a) > 0$ and $\mathbb{P}^*(\xi_1 = b) > 0$). But under $\mathbb{P}^*$, $\{\widetilde{S}_t\}$ is a martingale, hence by Proposition 2.6, $\widetilde{V}(X, Y)$ is a martingale for all strategies (X, Y). Consequently if $V_0(X, Y) = 0$ then $\mathbb{E}^* \widetilde{V}_T(X, Y) = 0$. The admissibility condition implies that $\widetilde{V}_T(X, Y) \ge 0$ almost surely, hence $\widetilde{V}_T(X, Y) \equiv 0$. □

For notational convenience we write $c = 1 + r$. It is easy to check that

$$\mathbb{P}^*(\xi_1 = a) = \frac{b - c}{b - a}, \quad \mathbb{P}^*(\xi_1 = b) = \frac{c - a}{b - a}.$$

Theorem 2.10 *If $a < 1 + r < b$, then the market defined above is complete, that is, for any $\mathcal{F}_T$-measurable random variable $H \geq 0$, there exists an admissible strategy (X, Y) such that $V_T(X, Y) = H$. Moreover, for all $0 \leq t < T$,*

$$V_t(X, Y) = \frac{R_t}{R_T}\mathbb{E}^*(H|\mathcal{F}_t).$$

PROOF If there exists an admissible strategy such that $V_T(X, Y) = H$, then by Proposition 2.1(iii), for all $0 \leq t < T$,

$$\frac{H}{R_T} = \widetilde{V}_t(X, Y) + \sum_{s=t+1}^{T} Y_s \Delta \widetilde{S}_s.$$

From the same computation as in Proposition 2.6 it follows that, for $s \geq t + 1$, $\mathbb{E}^*(Y_s \Delta \widetilde{S}_s|\mathcal{F}_t) = 0$, hence

$$\widetilde{V}_t(X, Y) = \mathbb{E}^*\left(\frac{H}{R_T}|\mathcal{F}_t\right)$$

and also

$$V_t(X, Y) = \frac{R_t}{R_T}\mathbb{E}^*(H|\mathcal{F}_t).$$

Note in particular that $H \geq 0$ implies that $V_t(X, Y) \geq 0$, hence if there exists a self-financing strategy which produces the above sequence $\{V_t(X, Y);\ 0 \leq t \leq T\}$, then it is admissible. In view of Proposition 2.2, it remains to show that there exists a predictable sequence $\{Y_t;\ 0 \leq t \leq T\}$ such that

$$\sum_{s=1}^{T} Y_s \Delta \widetilde{S}_s = \frac{H}{R_T} - \mathbb{E}^*\left(\frac{H}{R_T}\right).$$

Taking first $\mathbb{E}^*(\cdot|\mathcal{F}_t)$ and then $\mathbb{E}^*(\cdot|\mathcal{F}_{t-1})$ in this formula, and computing the difference between the two expressions thus obtained, we note that Y_t must satisfy

$$Y_t \widetilde{S}_{t-1}\left(\frac{\xi_t}{c} - 1\right) = \mathbb{E}^*(\widetilde{H}|\mathcal{F}_t) - \mathbb{E}^*(\widetilde{H}|\mathcal{F}_{t-1}),$$

where $\widetilde{H} := H/R_T$ (recall that $\widetilde{S}_t = \widetilde{S}_{t-1}\xi_t/c$), and

$$Y_t = \frac{c(\mathbb{E}^*(\widetilde{H}|\mathcal{F}_t) - \mathbb{E}^*(\widetilde{H}|\mathcal{F}_{t-1}))}{\widetilde{S}_{t-1}(\xi_t - c)}.$$

It remains to show that Y_t is $\mathcal{F}_{t-1}$-measurable (i.e. does not depend upon ξ_t).

Write $\xi^{t-1} := (\xi_1, \ldots, \xi_{t-1})$. Then the random variable $\mathbb{E}^*(\widetilde{H}|\mathcal{F}_t)/\widetilde{S}_{t-1}$ is a function of the pair (ξ^{t-1}, ξ_t).

Write

$$g_t(\xi^{t-1}, \xi_t) := c\frac{\mathbb{E}^*(\widetilde{H}|F_t)}{\widetilde{S}_{t-1}}.$$

We have

$$Y_t = \frac{g_t(\xi^{t-1}, \xi_t) - \mathbb{E}^*(g_t(\xi^{t-1}, \xi_t)|\mathcal{F}_{t-1})}{\xi_t - c}.$$

Note that

$$\mathbb{E}^*\left(g_t(\xi^{t-1}, \xi_t)|\mathcal{F}_{t-1}\right) = g_t(\xi^{t-1}, a)\frac{b-c}{b-a} + g_t(\xi^{t-1}, b)\frac{c-a}{b-a},$$

hence

$$Y_t = \frac{\left(g_t(\xi^{t-1}, \xi_t) - g_t(\xi^{t-1}, a)\right)\frac{b-c}{b-a} + \left(g_t(\xi^{t-1}, \xi_t) - g_t(\xi^{t-1}, b)\right)\frac{c-a}{b-a}}{\xi_t - c},$$

and in both cases $\xi_t = a$ and $\xi_t = b$, we have

$$Y_t = \frac{g_t(\xi^{t-1}, b) - g_t(\xi^{t-1}, a)}{b-a}.$$

□

Remark 2.11 *The formula above gives the fraction of the portfolio to be invested in the risky asset, at each time t, in order to produce a hedging strategy. Note the particular form of the right-hand side, which looks like an 'approximate derivative'. In the case of the continuous model of the next section, this will be a derivative.*

9.2.5 Call and put pricing

We shall now write the formula for $V_t(X, Y)$ explicitly in the two cases of a call and a put option. Write

$$p = \frac{b-c}{b-a} = \mathbb{P}^*(\xi_1 = a),$$

hence

$$1 - p = \mathbb{P}^*(\xi_1 = b).$$

In the case of a European call option,

$$V_t(X, Y) = c^{-(T-t)}\mathbb{E}^*\left[\left(S_t \prod_{s=t+1}^{T} \xi_s - K\right)_+ |\mathcal{F}_t\right].$$

But for all $0 \leq k \leq T - t$,

$$\mathbb{P}^*\left(\prod_{s=t+1}^{T} \xi_s = a^k b^{T-t-k}\right) = \binom{T-t}{k} p^k (1-p)^{T-t-k}.$$

Hence,

$$V_t(X,Y) = c^{-(T-t)} \sum_{k=0}^{T-t} \binom{T-t}{k} p^k (1-p)^{T-t-k} (S_t k - K)_+;$$

in particular,

$$\begin{aligned} V_0(X,Y) &= c^{-T} \sum_{k=0}^{T} \binom{T}{k} p^k (1-p)^{T-k} (S_0 k - K)_+ \\ &= c^{-T} \mathbb{E}^* \left[(S_T - K)_+ \right]. \end{aligned}$$

In the case of a European put option,

$$V_t(X,Y) = c^{-(T-t)} \sum_{k=0}^{T-t} \binom{T-t}{k} p^k (1-p)^{T-t-k} (K - S_t k)_+.$$

9.2.6 The Black–Scholes formula

We shall now establish the formulae for the price of an option in the continuous model which we shall present at Section 9.3, namely the celebrated Black–Scholes formula, by taking the limit on discrete models. We shall later do the same thing again by two different methods directly from the continuous model.

Let T be an arbitrary positive real number, and t take its values in the set

$$0, \frac{1}{N}, \ldots, \frac{[NT]}{N}.$$

Suppose now (here and below the superscript N is an index not an exponent) that

$$S_t = S_0 \prod_{k=1}^{[Nt]} \xi_k^N,$$

$$\tilde{S}_t = S_0 \exp\left(\sum_{k=1}^{[Nt]} \eta_k^N \right),$$

with

$$\eta_k^N = \log \xi_k^N - \frac{r}{N}.$$

We assume that the random variables η_k^N take their values in the set $\{-\frac{\sigma}{\sqrt{N}}, \frac{\sigma}{\sqrt{N}}\}$. This means, in comparison with the previous notation, that

$$c^N = \exp(r/N), \quad a^N = \exp(r/N - \sigma/\sqrt{N}), \quad b^N = \exp(r/N + \sigma/\sqrt{N}).$$

The formula for the price of a call option then becomes, if $Z_t^N := \sum_{k=1}^{[Nt]} \eta_k^N$,

$$\mathbb{E}^*\left[\left(S_0 \exp(Z_T^N) - Ke^{-rT}\right)_+\right],$$

while that of a put option becomes

$$\mathbb{E}^*\left[\left(Ke^{-rT} - S_0 \exp(Z_T^N)\right)_+\right].$$

It remains to find the limit law of Z_T^N as $N \to \infty$ under $\mathbb{P}^*$.

Theorem 2.12 *If $Z_t^N := \sum_{k=1}^{[Nt]} \eta_k^N$ and for each N the random variables $\{\eta_k^N,\ k \geq 0\}$ are i.i.d., taking their values in the set $\{-\frac{\sigma}{\sqrt{N}}, \frac{\sigma}{\sqrt{N}}\}$, with $\mathbb{E}\eta_k^N = \lambda_N$, and $N\lambda_N \to \lambda$ as $N \to \infty$, then under $\mathbb{P}$, as $N \to \infty$,*

$$Z_t^N \Rightarrow \lambda t + \sigma B_t, \quad t \geq 0,$$

where $\{B_t;\ t \geq 0\}$ is a Brownian motion (see Definition 3.1 below).

Proof We know (see [9], p. 180) that if a random variable X admits a moment of order 3, then for all $r \in \mathbb{R}$,

$$\mathbb{E}(\exp(irX)) = 1 + ir\mathbb{E}(X) - \frac{r^2}{2}\mathbb{E}(X^2) - i\frac{r^3}{6}(\mathbb{E}(X^3) + \delta(X, r)),$$

where $|\delta(X, r)| \leq 3\mathbb{E}(|X|^3)$ and $\delta(X, r) \to 0$, as $r \to 0$. Hence

$$\mathbb{E}(\exp(ir\eta_k^N)) = 1 + ir\lambda_N - \frac{r^2\sigma^2}{2N} + O(N^{-3/2}),$$

then

$$\begin{aligned}\mathbb{E}(\exp(irZ_t^N)) &= \left(1 + ir\lambda_N - \frac{r^2\sigma^2}{2N} + O(N^{-3/2})\right)^{[Nt]} \\ &\to \exp\left(ir\lambda t - \frac{r^2\sigma^2 t}{2}\right)\end{aligned}$$

as $N \to \infty$. We recognize the value at r of the characteristic function of the Gaussian $N(\lambda t, \sigma^2 t)$ law, which is the law of $\lambda t + \sigma B_t$, since the law of B_t is $N(0, t)$. □

In order to apply this theorem, it remains to compute the expectation of η_k^N under $\mathbb{P}^*$. This last probability distribution is characterized by the identity

$$\mathbb{E}^* \exp(\eta_k^N) = 1,$$

that is, with $p_a := \mathbb{P}^*(\eta_k^N = -\frac{\sigma}{\sqrt{N}})$, $p_b = \mathbb{P}^*(\eta_k^N = \frac{\sigma}{\sqrt{N}})$,

$$\exp\left(-\frac{\sigma}{\sqrt{N}}\right) p_a + \exp\left(\frac{\sigma}{\sqrt{N}}\right) p_b = 1,$$

hence

$$p_a = \frac{\exp\left(\frac{\sigma}{\sqrt{N}}\right) - 1}{\exp\left(\frac{\sigma}{\sqrt{N}}\right) - \exp\left(-\frac{\sigma}{\sqrt{N}}\right)}, \quad p_b = \frac{1 - \exp\left(-\frac{\sigma}{\sqrt{N}}\right)}{\exp\left(\frac{\sigma}{\sqrt{N}}\right) - \exp\left(-\frac{\sigma}{\sqrt{N}}\right)},$$

and

$$\mathbb{E}^* \eta_k^N = -\frac{\sigma^2}{2N} + o\left(\frac{1}{N}\right).$$

It then follows from Theorem 2.12 that under $\mathbb{P}^*$,

$$Z_t^N \Rightarrow -\frac{\sigma^2}{2} t + \sigma B_t.$$

We now deduce the limiting formula for the price of a call option,

$$C_0 = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \left(S_0 e^{-\sigma^2 T/2 + \sigma\sqrt{T} y} - K e^{-rT}\right)_+ e^{-y^2/2} dy,$$

and that of a put option,

$$P_0 = (2\pi)^{-1/2} \int_{-\infty}^{+\infty} \left(K e^{-rT} - S_0 e^{-\sigma^2 T/2 + \sigma\sqrt{T} y}\right)_+ e^{-y^2/2} dy.$$

These formulae can be written as follows in terms of the distribution function F of the centred normalized Gaussian probability distribution $N(0, 1)$:

$$C_0 = S_0 F(d_1) - K e^{-rT} F(d_2),$$
$$P_0 = K e^{-rT} F(-d_2) - S_0 F(-d_1),$$

where

$$d_1 = \frac{1}{\sigma\sqrt{T}} \log\left(\frac{S_0}{K}\right) + \frac{r\sqrt{T}}{\sigma} + \frac{\sigma\sqrt{T}}{2},$$
$$d_2 = \frac{1}{\sigma\sqrt{T}} \log\left(\frac{S_0}{K}\right) + \frac{r\sqrt{T}}{\sigma} - \frac{\sigma\sqrt{T}}{2}.$$

Observe that the call–put parity formula follows from these expressions, noting that $F(d_i) + F(-d_i) = 1$, $i = 1, 2$.

9.3 The Black–Scholes model and formula

We now consider a model in which the price of the underlying asset $S_t \in \mathbb{R}_+$ varies in continuous time, $t \in \mathbb{R}_+$. From now on, all functions will be defined on some subset of a Euclidean space $\mathbb{R}^k$ into $\mathbb{R}$. All the functions we shall introduce will be assumed to be Borel measurable, that is, the inverse image of any open subset of $\mathbb{R}$ must be a Borel subset of $\mathbb{R}^k$. Recall that the class of the Borel subsets of $\mathbb{R}^k$ is the smallest class which contains all open balls and is closed under countable union and complementation.

9.3.1 Introduction to stochastic calculus

All random variables and processes below will be defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The Black–Scholes model stipulates that

$$S_t = S_0 \exp(\lambda t + \sigma B_t),$$

which by Itô's formula in Theorem 3.4 below is equivalent to

$$dS_t = \mu S_t dt + \sigma S_t dB_t,$$

with $\lambda = \mu - \sigma^2/2$, where $\mu \in \mathbb{R}$ is the *drift* coefficient and $\sigma \in \mathbb{R}$ is called the *volatility*. $\{B_t; t \geq 0\}$ is a (standard) *Brownian motion*.

Definition 3.1 *A stochastic process $\{B_t; t \geq 0\}$ is called a Brownian motion if its trajectories are continuous, $B_0 = 0$, and*

(i) for all $n \in \mathbb{N}$, $0 = t_0 < t_1 < \ldots < t_n$, the sequence $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}$ is a sequence of independent random variables;

(ii) for all $0 \leq s < t$, the law of $B_t - B_s$ is the Gaussian probability distribution $N(0, t-s)$.

We then deduce that the process $\{\log(S_t/S_0);\ t \geq 0\}$ has independent increments (i.e. also possesses property (i) from the definition), and that for all $0 \leq s < t$, the law of $\log(S_t/S_s)$ is the Gaussian law $N(\lambda(t-s), \sigma^2(t-s))$. The process $\{S_t\}$ is called a *geometric Brownian motion*.

The following proposition states a fundamental property of Brownian motion.

Proposition 3.2 *Let $t>0$, and $0 = t_0^n < t_1^n < \ldots < t_n^n = t$ be a sequence of subdivisions of the interval $[0, t]$ such that $\sup_{k \leq n}(t_k^n - t_{k-1}^n) \to 0$ as $n \to \infty$. Then*

$$\sum_{k=1}^{n} (B_{t_k^n} - B_{t_{k-1}^n})^2 \to t,$$

in mean square, as $n \to \infty$.

Proof We have the identity

$$\mathbb{E} \sum_{k=1}^{n} \left(B_{t_k^n} - B_{t_{k-1}^n} \right)^2 = t.$$

But

$$\begin{aligned} \mathrm{Var}\left[\sum_{k=1}^{n} (B_{t_k^n} - B_{t_{k-1}^n})^2 \right] &= \sum_{k=1}^{n} \mathrm{Var}\left[(B_{t_k^n} - B_{t_{k-1}^n})^2 \right] \\ &= 2 \sum_{k=1}^{n} (t_k^n - t_{k-1}^n)^2 \end{aligned}$$

$$\leq 2t \sup_{k \leq n} (t_k^n - t_{k-1}^n)$$
$$\to 0,$$

as $n \to \infty$. □

This result shows that the trajectories of Brownian motion are very irregular. Were they differentiable (with an integrable derivative), then the above limit could not be t but would have to be 0 (exercise). We will, however, define a *stochastic integral* of the type

$$\int_0^t \varphi_s dB_s, \quad t \geq 0$$

(which cannot be written as $\int_0^t \varphi_s (dB_s/ds) ds$, since the derivative dB_s/ds does not exist as a function). We shall distinguish two types of stochastic integrals: the Wiener integral, in the case where the integrand φ is deterministic; and the Itô integral, in the case where the integrand φ is a stochastic process.

Let us first construct the *Wiener integral*. Consider a *deterministic* function $\{f(s);\ 0 \leq s \leq T\}$ such that

$$\int_0^T f^2(s) ds < \infty.$$

Suppose first that f is a step function, that is,

$$f(s) = \sum_{k=1}^n f_k \mathbf{1}_{]t_k t_{k+1}]}$$

with $0 \leq t_0 < t_1 < \ldots < t_n \leq T$. Then a natural definition of the Wiener integral is

$$\int_0^t f(s) dB_s = \sum_{k=1}^n f_k (B_{t \wedge t_{k+1}} - B_{t \wedge t_k}).$$

We easily deduce from the properties of Brownian motion that

$$\mathbb{E} \int_0^t f(s) dB_s = 0,$$

$$\mathbb{E}\left[\left(\int_0^t f(s) dB_s \right)^2 \right] = \mathbb{E} \left| \sum_{k=1}^n f_k \left(B_{t \wedge t_{k+1}} - B_{t \wedge t_k} \right) \right|^2$$
$$= \sum_{k=1}^n f_k^2 (t \wedge t_{k+1} - t \wedge t_k)$$
$$= \int_0^t f^2(s) ds.$$

The isometry formula

$$\mathbb{E}\left[\left(\int_0^t f(s)dB_s\right)^2\right] = \int_0^t f^2(s)ds$$

allows us to extend the definition of the Wiener integral from the class of step functions to that of all Borel measurable and square integrable functions. One can easily verify that the process

$$\left\{\int_0^t f(s)dB_s;\ \ 0 \le t \le T\right\}$$

is a centred Gaussian process with independent increments, which satisfies

$$\mathbb{E}\left[\left(\int_0^t f(s)dB_s\right)^2\right] = \int_0^t f^2(s)ds, \quad 0 \le t \le T.$$

In order to present the construction of the *Itô integral*, we need to introduce the natural filtration of the Brownian motion,

$$\mathcal{F}_t \triangleq \mathcal{F}_t^B = \sigma\{B_s;\ \ 0 \le s \le t\} \vee \mathcal{N},$$

that is, $\mathcal{F}_t$ is the smallest σ-field which makes all the random variables B_s measurable, for $0 \le s \le t$, and which contains the $\mathbb{P}$-null sets of the σ-algebra $\mathcal{F}$.

Denote by $M^2(0, T)$ the sub-Hilbert space of

$$L^2(\Omega \times [0, T], \mathcal{F} \otimes B([0, T]), d\mathbb{P} \times dt)$$

consisting of equivalence classes of square integrable processes $\{\varphi_t(\omega);\ \omega \in \Omega,\ 0 \le t \le T\}$ such that, for all $0 \le t \le T$, the random variable φ_t is $\mathcal{F}_t$-measurable. We call such a process $\{\varphi_t\}$ an *adapted* process.

Our construction of the Wiener integral extends to the Itô integral as follows. First consider processes of the form

$$\varphi_s(\omega) = \sum_{k=1}^n \varphi_k(\omega)\mathbf{1}_{]t_k, t_{k+1}]}(s),$$

where φ_k is assumed to be $\mathcal{F}_{t_k}$-measurable and square integrable, $1 \le k \le n$. For such a φ,

$$\int_0^t \varphi_s dB_s \triangleq \sum_{k=1}^n \varphi_k \left(B_{t\wedge t_{k+1}} - B_{t\wedge t_k}\right).$$

We shall make repeated use of the following fact which follows from property (i) of the Brownian motion and the definition of $\mathcal{F}_t$ (exercise): for all $0 \le s < t$, $\mathcal{F}_s$ and $B_t - B_s$ are independent. Then

$$\mathbb{E}\int_0^t \varphi_s dB_s = \sum_{k=1}^n \mathbb{E}\varphi_k \mathbb{E}(B_{t\wedge t_{k+1}} - B_{t\wedge t_k}) = 0$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^t \varphi_s dB_s\right)^2\right] &= \sum_k \mathbb{E}\left[\varphi_k^2 \left(B_{t\wedge t_{k+1}} - B_{t\wedge t_k}\right)^2\right] \\ &\quad + 2\sum_{\ell<k} \mathbb{E}\left[\varphi_\ell \left(B_{t\wedge t_{\ell+1}} - B_{t\wedge t_\ell}\right) \varphi_k \left(B_{t\wedge t_{k+1}} - B_{t\wedge t_k}\right)\right] \\ &= \sum_k \mathbb{E}(\varphi_k^2)(t \wedge t_{k+1} - t \wedge t_k) \\ &= \mathbb{E}\int_0^t \varphi_s^2 ds.\end{aligned}$$

Again the isometry property which we have just established allows us to extend the Itô integral to all integrands $\varphi \in M^2(0, T)$.

Theorem 3.3 *For all $\varphi \in M^2(0, T)$, $0 \le t \le T$, the Itô integral satisfies*

$$\mathbb{E}\int_0^t \varphi_s dB_s = 0,$$

$$\mathbb{E}\left[\left(\int_0^t \varphi_s dB_s\right)^2\right] = \mathbb{E}\int_0^t \varphi_s^2 ds.$$

Moreover, the process $\{\int_0^t \varphi_s dB_s;\ 0 \le t \le T\}$ is a martingale, since if $0 < s < t$, then

$$\mathbb{E}\left[\int_0^t \varphi_r dB_r | \mathcal{F}_s\right] = \int_0^s \varphi_r dB_r.$$

We shall assume that the Itô integral can be extended to adapted $\{\varphi_t\}$ which only satisfy

$$\int_0^T \varphi_t^2 dt < \infty \text{ a.s.}$$

However, for such a $\{\varphi_t\}$, the random variable $\int_0^t \varphi_s dB_s$ need not be integrable, and the three formulae in Theorem 3.3 need not be true. Nevertheless the following inequality is always valid (exercise):

$$\mathbb{E}\left[\left(\int_0^t \varphi_s dB_s\right)^2\right] \le \mathbb{E}\int_0^t \varphi_s^2 ds, \quad 0 \le t \le T.$$

We can now establish the celebrated *Itô formula*:

Theorem 3.4 *If $\Phi \in C^{1,2}(\mathbb{R}_+ \times \mathbb{R})$, then, for all $t>0$,*

$$\begin{aligned}\Phi(t, B_t) &= \Phi(0, B_0) + \int_0^t \Phi'_s(s, B_s) ds \\ &\quad + \int_0^t \Phi'_x(s, B_s) dB_s + \frac{1}{2}\int_0^t \Phi''_{xx}(s, B_s) ds.\end{aligned}$$

Remark 3.5 *The term $\frac{1}{2}\Phi''_{xx}$ is new, if we compare with the formula from standard differential calculus. Its presence is due to the irregularity of the trajectories of Brownian motion, and specifically to the result of Proposition 3.2.*

PROOF We shall prove the formula for a function Φ depending upon x only (and not upon t), the general case being left as an exercise. The formula is correct if and only if it is correct with t replaced by $t \wedge \tau_n$, where $\tau_n = \inf\{s;\ |X_s| \geq n\}$. Hence, it suffices to prove the result for $\Phi \in C_b^2(\mathbb{R})$ and we will show that for such a function

$$\Phi(B_t) = \Phi(0) + \int_0^t \Phi'(B_s)dB_s + \frac{1}{2}\int_0^t \Phi''(B_s)ds.$$

Write $t_k^n = (k/n)t$, $n \in \mathbb{N}$, $0 \leq k \leq n$. Then

$$\begin{aligned}\Phi(B_t) - \Phi(0) &= \sum_{k=1}^n \left(\Phi(B_{t_k^n}) - \Phi(B_{t_{k-1}^n})\right.\\ &= \sum_{k=1}^n \Phi'(B_{t_{k-1}^n})(B_{t_k^n} - B_{t_{k-1}^n}) + \frac{1}{2}\sum_{k=1}^n \Phi''(\Theta_k^n)(B_{t_k^n} - B_{t_{k-1}^n})^2\end{aligned}$$

by the second-order Taylor expansion, with Θ_k^n belonging to the interval $[B_{t_{k-1}^n}, B_{t_k^n}]$ (or $[B_{t_k^n}, B_{t_{k-1}^n}]$, depending on how the two quantities compare). It follows from the isometry formula for the Itô integral that

$$\begin{aligned}&\mathbb{E}\left(\left|\int_0^t \Phi'(B_s)dB_s - \sum_{k=1}^n \Phi'(B_{t_{k-1}}^n)(B_{t_k^n} - B_{t_{k-1}^n})\right|^2\right)\\ &= \mathbb{E}\sum_{k=1}^n \int_{t_{k-1}^n}^{t_k^n} |\Phi'(B_s) - \Phi'(B_{t_{k-1}^n})|^2 ds\\ &\to 0\end{aligned}$$

by dominated convergence.

We next note that

$$\begin{aligned}&\left|\sum_{k=1}^n \left(\Phi''(B_{t_{k-1}^n}) - \Phi''(\Theta_k^n)\right)(B_{t_k^n} - B_{t_{k-1}^n})^2\right|\\ &\leq \sup_k \left|\Phi''(B_{t_{k-1}^n}) - \Phi''(\Theta_k^n)\right| \sum_{k=1}^n (B_{t_k^n} - B_{t_{k-1}^n})^2\\ &\to 0\end{aligned}$$

in probability, as $n \to \infty$, since

$$\sup_k \left|\Phi''(B_{t_{k-1}^n}) - \Phi''(\Theta_k^n)\right| \to 0 \quad \text{and} \quad \sum_{k=1}^n (B_{t_k^n} - B_{t_{k-1}^n})^2 \to t.$$

Finally, a variant of the argument in the proof of Proposition 3.2 allows us to show that

$$\sum_{k=1}^{n} \Phi''(B_{t_{k-1}^n})(B_{t_k^n} - B_{t_{k-1}^n})^2 \to \int_0^t \Phi''(B_s)ds,$$

as $n \to \infty$. Specifically, we show both that

$$\mathbb{E}\left[\left(\sum_{k=1}^{n} \Phi''(B_{t_{k-1}^n})\left[(B_{t_k^n} - B_{t_{k-1}^n})^2 - (t_k^n - t_{k-1}^n)\right]\right)^2\right] \to 0$$

and that

$$\sum_{k=1}^{n} \Phi''(B_{t_{k-1}^n})(t_k^n - t_{k-1}^n) \to \int_0^t \Phi''(B_s)ds.$$

□

The above Itô formula generalizes as follows. We define an *Itô process* to be a process $\{X_t; 0 \le t \le T\}$ of the form

$$X_t = x + \int_0^t \psi_s ds + \int_0^t \varphi_s dB_s, \tag{9.1}$$

where $x \in \mathbb{R}$, and ψ and φ are adapted processes such that

$$\int_0^T (|\psi_t| + |\varphi_t|^2)dt < \infty \text{ a.s.}$$

We then have the following result, whose proof is analogous to that of Theorem 3.4:

Theorem 3.6 *If $\{X_t;\ 0 \le t \le T\}$ is an Itô process of the form (9.1), and $\Phi \in C^{1,2}(\mathbb{R}_+ \times \mathbb{R})$, then, for all $t>0$,*

$$\begin{aligned}\Phi(t, X_t) = \Phi(0, x) &+ \int_0^t \Phi'_s(s, X_s)ds \\ &+ \int_0^t \Phi'_x(s, X_s)\psi_s ds + \int_0^t \Phi'_x(s, X_s)\varphi_s dB_s \\ &+ \frac{1}{2}\int_0^t \Phi''_{xx}(s, X_s)\varphi_s^2 ds.\end{aligned}$$

We shall in due course need the multidimensional Itô formula: consider an $\mathbb{R}^k$-valued Brownian motion $\{B_t;\ t \ge 0\}$ (whose coordinates are k scalar mutually independent Brownian motions), $\psi_t \mathbb{R}^d$-valued and adapted, $\varphi_t \mathbb{R}^{d\times k}$-valued and adapted with

$$\int_0^T \left(\|\psi_s\| + \|\varphi_s\|^2\right) ds < \infty \text{ a.s.}$$

Then if $x \in \mathbb{R}^d$, the process

$$X_t = x + \int_0^t \psi_s ds + \int_0^t \varphi_s dB_s, \quad 0 \le t \le T,$$

is an $\mathbb{R}^d$-valued Itô process.

If $\Phi \in C^{1,2}([0, T] \times \mathbb{R}^d)$, we have the Itô formula,

$$\Phi(t, X_t) = \Phi(0, x) + \int_0^t \Phi'_s(s, X_s)ds + \int_0^t < \Phi'_x(s, X_s), \psi_s > ds + \int_0^t ¡\Phi'_x(s, X_s), \varphi_s dB_s > + \frac{1}{2}\int_0^t \mathrm{tr}[\Phi''_{xx}(s, X_s)\varphi_s\varphi_s^*]ds.$$

9.3.2 Stochastic differential equations

Let $f,\ g : [0, T] \times \mathbb{R} \to \mathbb{R}$ be such that

$$\sup_{0 \le t \le T} (|f(t, 0)| + |g(t, 0)|) < \infty.$$

We assume that, for some $K > 0$,

$$|f(t, x) - f(t, y)| + |g(t, x) - g(t, y)| \le K|x - y|, \quad \forall x, y \in \mathbb{R},\ t \in [0, T]. \tag{9.2}$$

Condition (9.2) is called the Lipschitz condition.

Theorem 3.7 *Under the above conditions, in particular (9.2), for all $x \in \mathbb{R}$, the stochastic differential equation (SDE)*

$$X_t = x + \int_0^t f(s, X_s)ds + \int_0^t g(s, X_s)dB_s, \quad 0 \le t \le T,$$

admits a unique solution $X \in M^2(0, T)$.

Proof Denote by F the mapping from $M^2(0, T)$ into itself defined by

$$F(X)_t = x + \int_0^t f(s, X_s)ds + \int_0^t g(s, X_s)dB_s, \quad 0 \le t \le T.$$

A solution of the SDE is a fixed point of F. But for F to have a unique fixed point, it suffices that F be a strict contraction for a norm on $M^2(0, T)$ to be chosen.

Applying Itô's formula to the Itô process $F(X)_t - F(Y)_t$ and to the function $\Phi(t,x) = e^{-\alpha t}|x|^2$ ($\alpha > 0$ will be chosen below), we obtain

$$\begin{aligned} e^{-\alpha T}|F(X)_T - F(Y)_T|^2 + \alpha \int_0^T e^{-\alpha t}|F(X)_t - F(Y)_t|^2 dt \\ = 2\int_0^T e^{-\alpha t}(F(X)_t - F(Y)_t)(f(t,X_t) - f(t,Y_t))dt \\ + 2\int_0^T e^{-\alpha t}(F(X)_t - F(Y)_t)(g(t,X_t) - g(t,Y_t))dB_t \\ + \int_0^T e^{-\alpha t}|g(t,X_t) - g(t,Y_t)|^2 dt. \end{aligned}$$

We now wish to take expectations; we shall omit the proof of the fact that the stochastic integral is integrable and has zero expectation. It follows, making use of the Lipschitz property, that

$$\begin{aligned} & e^{-\alpha T}\mathbb{E}|F(X)_T - F(Y)_T|^2 + \alpha\mathbb{E}\int_0^T e^{-\alpha t}|F(X)_t - F(Y)_t|^2 dt \\ & = \mathbb{E}\int_0^T e^{-\alpha t}\left[2(F(X)_t - F(Y)_t)(f(t,X_t) - f(t,Y_t))\right. \\ & \qquad \left. + |g(t,X_t) - g(t,Y_t)|^2\right] dt \\ & \le \mathbb{E}\int_0^T e^{-\alpha t}\left[2K|F(X)_t - F(Y)_t| \times |X_t - Y_t| + K^2|X_t - Y_t|^2\right] dt. \end{aligned}$$

It now follows from Schwarz's inequality that

$$\begin{aligned} & 2K\mathbb{E}\int_0^T e^{-\alpha t}|F(X)_t - F(Y)_t| \times |X_t - Y_t| dt \\ & \le \mathbb{E}\int_0^T e^{-\alpha t}|F(X)_t - F(Y)_t|^2 dt + K^2\mathbb{E}\int_0^T e^{-\alpha t}|X_t - Y_t|^2 dt. \end{aligned}$$

Hence,

$$(\alpha - 1)\mathbb{E}\int_0^T e^{-\alpha t}|F(X)_t - F(Y)_t|^2 dt \le 2K^2\mathbb{E}\int_0^T e^{-\alpha t}|X_t - Y_t|^2 dt.$$

We now choose $\alpha = 2K^2 + 2$, whence

$$\mathbb{E}\int_0^T e^{-(2K^2+2)t}|F(X)_t - F(Y)_t|^2 dt \le \frac{2K^2}{2K^2+1}\mathbb{E}\int_0^T e^{-(2K^2+2)t}|X_t - Y_t|^2 dt.$$

□

9.3.3 The Feynman–Kac formula

Now consider the backward parabolic partial differential equation (PDE)

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + f(x)\dfrac{\partial u}{\partial x}(t,x) + \dfrac{1}{2}g^2(x)\dfrac{\partial^2 u}{\partial x^2}(t,x) = c(x)u(t,x), \\ 0 \le t \le T, \ \ x \in \mathbb{R}; \qquad u(T,x) = h(x), \ \ x \in \mathbb{R}. \end{cases} \tag{9.3}$$

We assume that f and g satisfy condition (9.2) and that c and φ are continuous and bounded on $\mathbb{R}$. For each $0 < t < T$, $x \in \mathbb{R}$, we denote by $\{X_s^{t,x};\ t \le s \le T\}$ the solution of the SDE

$$X_s^{t,x} = x + \int_t^s f(X_r^{t,x})dr + \int_t^s g(X_r^{t,x})dB_r, \quad t \le s \le T.$$

Theorem 3.8 *Suppose that $u \in C_b^{1,2}((0,T) \times \mathbb{R})$ is a solution of the PDE (9.3). Then u is given by the Feynman–Kac formula,*

$$u(t,x) = \mathbb{E}\left[h(X_T^{t,x}) \exp\left(-\int_t^T c(X_s^{t,x})ds\right)\right].$$

Proof We apply Itô's formula to the process $(X_s^{t,x}, Y_s)$, with $Y_s = -\int_t^s c(X_r^{t,x})dr$, and to the function $\Phi(s,x,y) = u(s,x)\exp(y)$, and obtain

$$\begin{aligned} u(T, X_T^{t,x}) \exp\left(-\int_t^T c(X_s^{t,x})ds\right) &= u(t,x) \\ &+ \int_t^T \exp\left(-\int_t^s c(X_r^{t,x})dr\right) g\left(X_s^{t,x}\right) \frac{\partial u}{\partial x}(s, X_s^{t,x})dB_s \\ &+ \int_t^T \left(\frac{\partial u}{\partial s} + Lu - cu\right)(s, X_s^{t,x}) \exp(Y_s)ds, \end{aligned}$$

where

$$Lu(s,x) = f(x)\frac{\partial u}{\partial x}(s,x) + \frac{1}{2}g^2(x)\frac{\partial^2 u}{\partial x^2}(s,x).$$

It remains to take the expectation, and to exploit the fact that u solves (9.3) in order to obtain the theorem. □

9.3.4 The Black–Scholes partial differential equation

We now present a first solution to the problems of *pricing* and *hedging* a European contingent claim, through the derivation of a PDE.

Consider a European option, which pays its holder $H = h(S(T))$ at exercise time T. Once again we have in mind the two examples $h(x) = (x-K)_+$ and $h(x) = (K-x)_+$. Denote by E_t the price of this option at time t, $0 \le t \le T$. Of

course $E_T = h(S(T))$. Assume for a moment – anticipating the proof in Section 9.3.9 below – that there exists a mapping

$$u : [0, T] \times \mathbb{R}_+ \to \mathbb{R}_+$$

such that, for all $0 \leq t \leq T$,

$$E_t = u(t, S_t).$$

Assume also – this again can be proved, but it is a little more difficult – that

$$u \in C^{1,2}((0, T) \times \mathbb{R}_+),$$

so that we can apply Itô's formula.

Note that

$$E_t = u(t, S_0 \exp(\lambda t + \sigma B_t)),$$

hence an application of Itô's formula from Theorem 3.4 yields

$$dE_t = \left(\frac{\partial u}{\partial t}(t, S_t) + \mu S_t \frac{\partial u}{\partial x}(t, S_t) + \frac{\sigma^2}{2} S_t^2 \frac{\partial^2 u}{x^2}(t, S_t) \right) dt + \sigma S_t \frac{\partial u}{\partial x}(t, S_t) dB_t$$

The absence of arbitrage implies that there exists an admissible strategy $\{(X_t, Y_t); 0 \leq t \leq T\}$ such that the associated value of the portfolio at the final time is

$$V_T(X, Y) = h(S_T).$$

Then necessarily

$$V_t(X, Y) = E_t, \quad 0 \leq t \leq T.$$

Moreover,

$$V_t(X, Y) = X_t R_t + Y_t S_t,$$

and the self-financing condition in continuous time reads

$$dV_t(X, Y) = X_t dR_t + Y_t dS_t.$$

But

$$dR_t = r R_t dt$$

and, using Itô's formula once more, we deduce that

$$dS_t = \mu S_t dt + \sigma S_t dB_t.$$

From the above identities there follows a second formula for the differential of E_t, namely

$$dE_t = (rX_tR_t + \mu Y_tS_t)\,dt + \sigma Y_tS_t\,dB_t.$$

We now use a well-known result in stochastic calculus, the fact that whenever two Itô processes are identical, the coefficient of dB_t and that of dt necessarily coincide.

Hence, we have first

$$\sigma Y_tS_t = \sigma S_t \frac{\partial u}{\partial x}(t, S_t),$$

or in other words

$$Y_t = \frac{\partial u}{\partial x}(t, S_t)$$

(we have just identified the hedging strategy), and second

$$rX_tR_t + \mu Y_tS_t = \frac{\partial u}{\partial t}(t, S_t) + \mu S_t \frac{\partial u}{\partial x}(t, S_t) + \frac{\sigma^2}{2} S_t^2 \frac{\partial^2 u}{\partial u^2}(t, S_t).$$

But we already know that

$$Y_t = \frac{\partial u}{\partial x}(t, S_t),$$

from which we deduce, thanks to

$$X_tR_t + Y_tS_t = u(t, S_t),$$

that

$$X_t = R_t^{-1}\left(u(t, S_t) - S_t \frac{\partial u}{\partial x}(t, S_t)\right).$$

Hence, the above relation becomes

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, S_t) + rS_t \dfrac{\partial u}{\partial x}(t, S_t) + \dfrac{\sigma^2}{2} S_t^2 \dfrac{\partial^2 u}{\partial x^2}(t, S_t) = ru(t, S_t), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u(T, S_T) = h(S_T). \end{cases}$$

A necessary and sufficient condition for these relations to hold almost surely is that u is a solution of the parabolic PDE

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, x) + rx \dfrac{\partial u}{\partial x}(t, x) + \dfrac{\sigma^2 x^2}{2} \dfrac{\partial^2 u}{\partial x^2}(t, x) = ru(t, x), \\ \qquad 0 < t < T,\ x \in \mathbb{R}_+; \qquad u(T, x) = h(x),\ x \in \mathbb{R}_+. \end{cases} \tag{9.4}$$

9.3.5 The Black–Scholes formula (2)

Recall that

$$S_t = S_0 + \mu \int_0^t S_s ds + \sigma \int_0^t S_s dB_s.$$

Define

$$B_t^* := B_t + \frac{\mu - r}{\sigma} t.$$

Then

$$S_t = S_0 + r \int_0^t S_s ds + \sigma \int_0^t S_s dB_s^*.$$

Now let $\mathbb{P}^*$ be a probability on $(\Omega, \mathcal{F})$ such that $\{B_t^*;\ t \geq 0\}$ is a Brownian motion. Such a probability exists, and we shall see below that it is equivalent to the probability $\mathbb{P}$ (under which $\{B_t\}$ is a Brownian motion).

Note that

$$d(R_t^{-1} S_t) = \sigma R_t^{-1} S_t dB_t^*,$$

hence the discounted price $\tilde{S}_t = R_t^{-1} S_t$ is a martingale under $\mathbb{P}^*$, which again is interpreted as the *risk-neutral probability*.

It then follows from the Feynman–Kac formula (Theorem 3.8) and equation (9.4) that

$$u(t, x) = \mathbb{E}^* \left[e^{-r(T-t)} h(S_T) | S_t = x \right]$$

or, in other words,

$$E_t = u(t, S_t) = \mathbb{E}^* \left[e^{-r(T-t)} h(S_T) | S_t \right],$$

and in particular

$$E_0 = u(0, S_0) = \mathbb{E}^* \left[e^{-rT} h(S_T) \right].$$

Since under $\mathbb{P}^*$ the law of $\log(S_T/S_0)$ is the probability distribution $N((r - \frac{\sigma^2}{2})T, \sigma^2 T)$, we obtain in particular the formulae for C_0 and P_0 which we previously obtained in Section 9.2.6.

9.3.6 Generalization of the Black–Scholes model

In the case of a call, we note that $h'(x) \geq 0$, and we expect that

$$\frac{\partial u}{\partial x}(t, x) \geq 0.$$

Clearly if the price of the underlying asset rises, the price of the option rises; hence we expect that $Y_t \geq 0$. Note that these inequalities are reversed in the case of a put !

On the other hand, there is no reason why the part X_t of the portfolio invested in the riskless asset must be positive or negative (i.e. a deposit or a loan), and the assumption that the interest rate is the same for the two cases is completely unrealistic.

Suppose that a deposit attracts an interest rate r^+, while the loans are subject to an interest rate r^-. Writing

$$R_t^+ = e^{r^+ t}, \quad R_t^- = e^{r^- t},$$

we have, in the case of a self-financing strategy, with the notation $X_t^+ = \max(0, X_t)$, $X_t^- = \max(0, -X_t)$,

$$dV_t = (X_t^+ r^+ R_t^+ - X_t^- r^- R_t^-)dt + Y_t dS_t.$$

If we follow the same argument which in Section 9.3.4 led us to the Black–Scholes PDE, we see that we again have

$$Y_t = \frac{\partial u}{\partial x}(t, S_t),$$

and this time

$$X_t^+ R_t^+ = (u(t, S_t) - S_t \frac{\partial u}{\partial x}(t, S_t))_+,$$
$$X_t^- R_t^- = (u(t, S_t) - S_t \frac{\partial u}{\partial x}(t, S_t))_-,$$

from which we deduce the *nonlinear* PDE

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, x) + \dfrac{\sigma^2 x^2}{2} \dfrac{\partial^2 u}{\partial x^2} = H\left(r^+, r^-, x, u(t, x), \dfrac{\partial u}{\partial x}(t, x)\right); \\ u(T, x) = h(x), \end{cases}$$

with $H(a, b, x, y, p) = a(y - xp)_+ - b(y - rp)_-$.

9.3.7 The Black–Scholes formula (3)

We now ask a question which is a little more general than the preceding one. What is the fair price of an option that pays its holder the amount $H(\geq 0)$ at time T? Assume that the random variable H is $\mathcal{F}_T$-measurable, where again, up to zero measure sets,

$$\mathcal{F}_t = \sigma\{B_s;\ 0 \leq s \leq t\} = \sigma\{S_s; 0 \leq s \leq t\}.$$

A particular example is the case where

$$H = h(S_T),$$

which notably covers the cases of a European call or put, but we shall also see below other types of options which are not of this particular form. We shall be guided in this section by what we have done in the case of the discrete model, and in the previous sections.

We phrase the question as follows: find V_0 and a self-financing strategy $\{(X_t, Y_t); 0 \le t \le T\}$ such that

$$\begin{aligned} V_T(X, Y) &= V_0 + \int_0^T X_t dR_t + \int_0^T Y_t dS_t \\ &= H. \end{aligned}$$

Again $R_t = e^{rt}$, and we define the discounted value of the portfolio at time t:

$$\widetilde{V}_t(X, Y) = R_t^{-1} V_t(X, Y) = X_t + \widetilde{Y}_t,$$

where $\widetilde{Y}_t \triangleq R_t^{-1} Y_t S_t$ is the discounted value of the part of the portfolio which is invested on the stock. Then

$$\widetilde{Y}_t = \widetilde{V}_t - X_t.$$

Moreover,

$$\begin{aligned} d\widetilde{V}_t(X, Y) &= -r\widetilde{V}_t dt + R_t^{-1} dV_t \\ &= -r\widetilde{V}_t dt + rX_t dt + R_t^{-1} Y_t dS_t \\ &= -r\widetilde{Y}_t dt + R_t^{-1} Y_t dS_t. \end{aligned}$$

But

$$dS_t = \mu S_t dt + \sigma S_t dB_t,$$

hence

$$d\widetilde{V}_t = (\mu - r)\widetilde{Y}_t dt + \sigma \widetilde{Y}_t dB_t.$$

Finally, define

$$B_t^* \triangleq \frac{\mu - r}{\sigma} t + B_t.$$

Then

$$d\widetilde{V}_t = \sigma \widetilde{Y}_t dB_t^*,$$

that is,

$$\widetilde{V}_t = e^{-rT} H - \sigma \int_t^T \widetilde{Y}_s dB_s^*.$$

Now define $\mathbb{P}^*$ to be the probability on $(\Omega, \mathcal{F}_T)$ such that

$$\{B_t^*;\ 0 \le t \le T\}$$

is a $\mathbb{P}^*$ Brownian motion. We suppose that

$$\mathbb{E}^*(H^2) < \infty.$$

Under this assumption, we can show easily that

$$\mathbb{E}^* \int_0^T |\widetilde{Y}_t|^2 dt < \infty,$$

hence in particular

$$\widetilde{V}_t = e^{-rT}\mathbb{E}^*(H|\mathcal{F}_t),$$

that is,

$$E_t = V_t = e^{-r(T-t)}\mathbb{E}^*(H|\mathcal{F}_t),$$

from which we again deduce the Black–Scholes formula for the prices of the European call and put.

What about the hedging strategy? Under $\mathbb{P}^*$, $\{\widetilde{V}_t\}$ is a square integrable martingale which is adapted to the filtration $\mathcal{F}_t^{B^*}$. A theorem due to Itô tells us that this implies the existence of a unique $Z \in M^2(0,T)$ such that

$$\widetilde{V}_t = V_0 + \int_0^t Z_s dB_s^*, \quad 0 \le t \le T.$$

Then

$$Y_t = \frac{R_t Z_t}{\sigma S_t}.$$

In the case where $H = h(S_T)$, we have the Black–Scholes PDE and Y_t can be computed in terms of the derivative of its solution. In more general cases, the computation can be done (but not in a very explicit way!) with other tools from stochastic calculus, in particular Malliavin's calculus.

Let us close this section with two classical examples of options which are not of the form $H = h(S_T)$.

Example 3.9 Barrier call option *This is a contingent claim which at exercise time T pays*

$$H = \mathbf{1}_{\{\sup_{0\le t\le T} S_t < \beta\}}(S_T - K)_+,$$

in other words it pays the same as a European call, except that the holder is allowed to exercise his option only if the stock price of the underlying asset has never reached the barrier β between time 0 and time T.

Example 3.10 Asian call option *This is an option which at exercise time T pays*

$$H = \left(\frac{1}{T}\int_0^T S_t dt - K\right)_+.$$

9.3.8 Girsanov's theorem

The risk-neutral probability $\mathbb{P}^*$ which we have introduced above may seem mysterious. In fact it can be deduced from the model probability $\mathbb{P}$ by Girsanov's theorem. We shall establish a simplified version of Girsanov's theorem, due to Cameron and Martin, which will be sufficient for our purpose.

We first prove the following lemma.

Lemma 3.11 *A continuous process* $\{B_t;\ 0 \le t \le T\}$ *is a Brownian motion if and only if, for all* n, *all* $0 = t_0 < t_1 < \ldots < t_n \le T$, $u_1, \ldots, u_n \in \mathbb{R}$,

$$\mathbb{E} \exp\left[\sum_1^n u_k \left(B_{t_k} - B_{t_{k-1}}\right)\right] = \exp\left[\frac{1}{2}\sum_1^n u_k^2 (t_k - t_{k-1})\right].$$

PROOF The necessity of the condition follows from the fact that if $\{B_t\}$ is a Brownian motion, then the law of the random vector $(B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}})$ is the Gaussian probability distribution $N(0, \Lambda_n)$, where Λ_n is an $n \times n$ diagonal matrix whose kth diagonal equals $t_k - t_{k-1}$.

Sufficiency follows from the fact that if that formula is true, then the law of $(B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}})$ is the Gaussian probability distribution $N(0, \Lambda_n)$, for all n, $0 < t_1 < \ldots < t_{n-1} \le t_n$, hence (i) and (ii) in Definition 3.1 hold. □

Theorem 3.12 *Let* $\{B_t;\ 0 \le t \le T\}$ *be a Brownian motion defined on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *If* $f \in L^2(0, T)$,

$$B_t^* = B_t - \int_0^t f(s)ds, \quad 0 \le t \le T,$$
$$Z_t = \exp\left(\int_0^t f(s)dB_s - \frac{1}{2}\int_0^t f^2(s)ds\right),$$

and $\mathbb{P}^*$ *is the probability on* $(\Omega, \mathcal{F}_T)$ *defined by*

$$\frac{d\mathbb{P}^*}{d\mathbb{P}} = Z_T,$$

then $\{B_t^*;\ 0 \le t \le T\}$ *is a Brownian motion under* $\mathbb{P}^*$.

PROOF In view of the preceding lemma, it suffices to show that, for all $g \in L^2(0, T)$,

$$\mathbb{E}^*\left[\exp\left(\int_0^T g(t)dB_t^*\right)\right] = \exp\left(\frac{1}{2}\int_0^T g^2(t)dt\right).$$

But

$$\begin{aligned}\mathbb{E}^*\left[\exp\left(\int_0^T g(t)dB_t^*\right)\right] &= \mathbb{E}^*\exp\left\{\int_0^T [f(t)+g(t)]dB_t - \frac{1}{2}\int_0^T [2f(t)g(t)+f^2(t)]dt\right\} \\ &= \exp\left(\frac{1}{2}\int_0^T g^2(t)dt\right).\end{aligned}$$

□

9.3.9 Markov property and partial differential equation

In order to deduce the so-called 'Black–Scholes PDE' in Section 9.3.4, we assumed that, at each time t, the price E_t of the option is a function of the price S_t of the underlying asset, that is, that E_t can be written as

$$E_t = u(t, S_t).$$

We have seen in Section 9.3.7 that

$$E_t = \frac{R_t}{R_T}\mathbb{E}^*(H|\mathcal{F}_t).$$

Why and under which conditions is this conditional expectation a function of t and S_t only?

Definition 3.13 *Let $\{X_t; t \geq 0\}$ be a stochastic process. $\{X_t\}$ is called a Markov process if, for all $0 < s < t$, all $f \in C_b(\mathbb{R})$,*

$$\mathbb{E}[f(X_t)|\mathcal{F}_s^X] = \mathbb{E}[f(X_t)|X_s],$$

where $\mathcal{F}_s^X \stackrel{\triangle}{=} \sigma\{X_r; 0 \leq r \leq s\}$ (up to zero measure sets).

Note that (again up to zero measure sets) $\mathcal{F}_t = \sigma\{S_s; 0 \leq s \leq t\}$. We have the following proposition.

Proposition 3.14 *Under $\mathbb{P}^*$, $\{S_t; 0 \leq t \leq T\}$ is a Markov process.*

PROOF If $0 < s < t$,

$$S_t = S_s \exp\left[\left(r - \frac{\sigma^2}{2}\right)(t-s) + \sigma(B_t^* - B_s^*)\right].$$

Hence

$$\begin{aligned}\mathbb{E}^*\left[f(S_t)|\mathcal{F}_s\right] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f\left(S_s \exp\left[\left(r - \frac{\sigma^2}{2}\right)(t-s) + \sigma x\sqrt{t-s}\right]\right)e^{-x^2/2}dx \\ &= \mathbb{E}^*[f(S_t)|S_s],\end{aligned}$$

since S_s is $\mathcal{F}_s$ measurable, and $B_t^* - B_s^*$ is independent of $\mathcal{F}_s$ under $\mathbb{P}^*$. □

Using the Feynman–Kac formula, we can deduce the PDE satisfied by the function $u(t,x)$ from the formula

$$E_t = \mathbb{E}^*\left[e^{-r(T-t)}h(S_T)|S_t\right].$$

We can now ask ourselves whether in the cases of the barrier option and the Asian option, the computation of the price of the option can still be done by solving a PDE.

PDE associated with the barrier option

Consider again the barrier call option from Example 9.3.9. Define

$$S_t^\beta = S_{t\wedge\tau_\beta},$$

where $\tau_\beta = \inf\{t \le T;\ \ S_t \ge \beta\}$, and

$$h(x) = \begin{cases} 0, & \text{if } x \le K, \\ x - K, & \text{if } K \le x < \beta, \\ 0, & \text{if } x \ge \beta. \end{cases}$$

Then, in the case of the barrier call option, the random variable H can be rewritten as

$$H = h(S_t^\beta).$$

By an argument analogous to that in the proof of Proposition 3.14, we show that $\{S_t^\beta;\ \ 0 \le t \le T\}$ is a Markov process. Under the *risk-neutral probability*, both $R_t^{-1}S_t$ and $R_{t\wedge\tau_\beta}^{-1}\ S_{t\wedge\tau_\beta}$ are martingales, and the arguments from Section 9.3.7 again lead to the formula

$$E_t = \mathbb{E}^*\left[\frac{R_t}{R_T}h(S_T^\beta)|S_t^\beta\right].$$

Note that, for $t \ge \tau_\beta$, $E_t = 0$. In other words, if $E_t = u(t, S_t^\beta)$, $u(t,\beta) = 0$. The PDE becomes

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) + rx\dfrac{\partial u}{\partial x}(t,x) + \dfrac{\sigma^2x^2}{2}\dfrac{\partial^2 u}{\partial x^2}(t,x) = ru(t,x), \quad 0 < t < T,\ 0 < x < \beta; \\ \qquad u(t,\beta) = 0,\ 0 < t < T;\ \ u(T,x) = h(x), \quad 0 < x < \beta. \end{cases}$$

PDE associated with the Asian option

Write $U_t = \int_0^t S_s ds$. Then in the case of the Asian option (Example 3.10),

$$H = h(U_T),$$

now with $h(x) = (T^{-1}x - K)_+$. It is not too hard to check that $\{U_t;\ 0 \le t \le T\}$ is not a Markov process. However, the two-dimensional process $\{(S_t, U_t);\ 0 \le t \le T\}$ is a Markov process, hence

$$E_t = \mathbb{E}^*\left(\frac{R_t}{R_T}H|\mathcal{F}_t\right) = \mathbb{E}^*\left(\frac{R_t}{R_T}H|S_t, U_t\right) = u(t, S_t, U_t),$$

where $\{u(t, x, y);\ 0 \le t \le T,\ x>0,\ y>0\}$ solves the PDE

$$\begin{cases} \dfrac{\partial u}{\partial t}(t, x, y) + \left(\dfrac{\sigma^2 x^2}{2}\dfrac{\partial^2 u}{\partial x^2} + rx\dfrac{\partial u}{\partial x} + x\dfrac{\partial u}{\partial y} - ru\right)(t, x, y) = 0, \\ \qquad\qquad 0 \le t \le T,\ x, y>0; \qquad u(T, x, y) = h(y), \\ \qquad\qquad\qquad\qquad\qquad\qquad x>0,\ y>0. \end{cases}$$

9.3.10 Contingent claim on several underlying stocks

So far we have only considered options based upon a single underlying risky asset (e.g. a stock). There do exist contingent claims based upon several underlying assets. A first example is the so-called *spread option*, which is written on the difference between two stocks. In other words, $H = (S_T^1 - S_T^2)_+$, where S^1 and S^2 are the prices of two different assets. A second example is the *basket option*, such as an option on an index (e.g. the Dow Jones). A put option based on a portfolio is one way to insure a portfolio. Given a portfolio consisting of a_i shares of stock i priced at S_t^i at time t, $i = 1, \ldots, d$, a put with payoff $(K - \sum_{i=1}^n a_i S_T^i)_+$ guarantees that the portfolio can be sold at least at the price K at the exercise time.

Suppose that, in addition to the riskless asset, whose price is $R_t = e^{rt}$ at time t, the market includes d risky assets, whose prices are S_t^i, $i = 1, \ldots, d$, and which fluctuate according to the model

$$dS_t^i = \mu_i S_t^i dt + S_t^i \sum_{j=1}^d \sigma_{ij} dB_t^j, \quad 1 \le i \le d,\ t \ge 0,$$

where $B_t^1, \ldots, B_t^d$ are d mutually independent Brownian motions. An application of Itô's formula allows us to check that the unique solution of the above multidimensional SDE is

$$S_t^i = S_0^i \exp\left[\lambda_i t + \sum_{j=1}^d \sigma_{ij} B_t^j\right], \quad 1 \le i \le d,\ t \ge 0,$$

with $\lambda_i = \mu_i - \frac{1}{2}\sum_{j=1}^d \sigma_{ij}^2$.

The first question to raise, in order to generalize the Black–Scholes theory to this situation, is that of the existence of a *risk-neutral probability* $\mathbb{P}^*$ equivalent to $\mathbb{P}$, under which the processes of the discounted prices $\{e^{-rt}S_t = e^{-rt}(S_t^1, \ldots, S_t^d);$

$t \geq 0\}$ are a d-dimensional martingale. This holds provided we have a *d-dimensional $\mathbb{P}^*$-Brownian motion* $\{B_t^*;\ t \geq 0\}$ such that

$$dS_t^i = rS_t^i dt + S_t^i \sum_{j=1}^{d} \sigma_{ij} dB_t^{*,j}, \quad 1 \leq i \leq d,\ t \geq 0.$$

Write $\mathbf{r}$ for the vector in $\mathbb{R}^d$ whose entries are all equal to r, and

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \vdots \\ \mu_d \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \cdots & \sigma_{1d} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \cdots & \sigma_{dd} \end{pmatrix}.$$

The second way of writing the S_t^i is equivalent to

$$(\mathbf{r} - \mu)t = \Sigma(B_t - B_t^*), \quad t \geq 0.$$

We are thus led to formulate the following crucial assumption:

$$\Sigma \quad \text{is invertible.} \tag{9.5}$$

Under this assumption, we deduce from the previous relation between B_t and B_t^* the formula

$$B_t^* = \Sigma^{-1}(\mu - \mathbf{r})t + B_t, \quad t \geq 0. \tag{9.6}$$

It follows from a natural generalization (whose proof is left as an exercise) of Girsanov's Theorem 3.12 that if

$$\frac{d\mathbb{P}^*}{d\mathbb{P}} = \exp\left[<\Sigma^{-1}(\mu - \mathbf{r})B_T> - \frac{1}{2}\|\Sigma^{-1}(\mu - \mathbf{r})\|^2 T \right],$$

then $\{B_t^*;\ 0 \leq t \leq T\}$ is a d-dimensional Brownian motion (i.e. $\{B_t^1\}, \ldots, \{B_t^d\}$ are scalar mutually independent Brownian motions) under $\mathbb{P}^*$.

Using the same arguments as in Section 9.3.7, we have that the price E_t of the option at time t is given by the formula (which is formally identical to that in the scalar case)

$$E_t = e^{-r(T-t)}\mathbb{E}^*(H|\mathcal{F}_t),$$

hence in particular

$$E_0 = e^{-rT}\mathbb{E}^* H. \tag{9.7}$$

If $H = h(S_T)$, then we have $E_0 = e^{-rT}\mathbb{E}^* h(S_T)$. For $t \geq 0$, denote by $\log S_t$ the vector $(\log S_t^1, \ldots, \log S_t^d)$. Under $\mathbb{P}^*$, the probability distribution of $\log(S_T)$ is the multidimensional Gaussian distribution $N(\log(S_0) + (\mathbf{r} - \frac{1}{2}s^2)T, \Sigma\Sigma^* T)$, where $s^2 = (s_1^2, \ldots, s_d^2)$, with $s_i^2 = \sum_{j=1}^d \sigma_{ij}^2$.

The $\mathbb{R}^d$-valued process $\{S_t;\ t \geq 0\}$ is a Markov process, hence in the case where $H = h(S_T)$, there exists a function $u : [0, T] \times \mathbb{R}^d \to \mathbb{R}_+$ such that

$$E_t = u(t, S_t), \quad 0 \leq t \leq T.$$

We show by an argument analogous to that in Section 9.3.4 that u solves the following parabolic PDE in $\mathbb{R}^d_+$, with $a = \Sigma\Sigma^*$:

$$\begin{cases} r\dfrac{\partial u}{\partial t}(t,x) + r\displaystyle\sum_{i=1}^{d} x_i \frac{\partial u}{\partial x_i}(t,x) + \frac{1}{2}\sum_{i,j=1}^{d} x_i x_j a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j}(t,x) = ru(t,x), \\ \qquad\qquad 0 \leq t \leq T,\ x \in \mathbb{R}^d_+; \qquad\qquad u(T,x) = h(x),\ x \in \mathbb{R}^d_+. \end{cases}$$

Moreover, the hedging portfolio is determined by the relation

$$Y_t^i = \frac{\partial u}{\partial x_i}(t, S_t), \quad 1 \leq i \leq d,\ 0 \leq t \leq T,$$

in the sense that Y_t^i is the number of shares of stock i that this portfolio should contain. The value of the portfolio is then

$$V_t(X, Y) = X_t R_t + \sum_{i=1}^{d} Y_t^i S_t^i, \quad 0 \leq t \leq T.$$

9.3.11 Viability and completeness

The notions of *viable* and *complete* market are defined as in the discrete model case, except that now a complete market is such that with each $\mathcal{F}_T$-measurable and square integrable random variable $H \geq 0$, we can associate an initial wealth V_0 and an admissible strategy (X, Y) such that

$$H = V_0 + \int_0^T X_t dR_t + \sum_{i=1}^{d} Y_t^i dS_t^i.$$

The restriction that H is square integrable is useless in the case of a discrete model, since in this model Ω is finite, hence all random variables are bounded.

We have the following fundamental result.

Theorem 3.15 *The market is viable if and only if there exists at least one risk-neutral probability $\mathbb{P}^*$ equivalent to $\mathbb{P}$. The market is complete if there exists exactly one risk-neutral probability $\mathbb{P}^*$ equivalent to $\mathbb{P}$.*

The fact that existence of a risk-neutral probability implies viability is proved as in the case of the discrete model. The rest of the proof of the theorem can be found, for example, in [24] (see Theorems 2.6 and 3.4).

9.3.12 Remarks on effective computation

We have associated a PDE with each of the options we have considered so far. (However, in the case of the Asian option, the price for this was a doubling of the dimension, which is not good news for the numerical solution.) Can we produce a probabilistic formula for the price of the option in case of the model considered in Section 9.3.6? The answer is yes, using the theory of 'backward SDEs'; but the probabilistic formula is not very explicit. Let us now discuss the effective computation of the price of the option.

Explicit formulae

A first approach, which can only be used in the simplest cases (these include almost all the examples we have considered so far) involves exploiting the knowledge of the law of the random variables which we consider. In the case of a European call option we use the formula

$$C_0 = S_0 F(d_1) - K e^{-rT} F(d_2)$$

(see the end of Section 9.2.6), where F denotes the standard normal distribution function $N(0, 1)$, which is available to very good precision in Matlab or whose numerical computation is very easy to program. Concerning the barrier option, we can use the knowledge of the law of the random vector $(\sup_{0\leq t\leq T} S_t, S_T)$; and concerning the Asian option, progress has recently been made towards identifying the law of the random variable U_T.

Monte Carlo method

A second method, with a much broader domain of application, is the Monte Carlo method, presented in Chapter 1. Recall that in order to reduce the variance (and hence accelerate the convergence) one should, provided one can use a call–put parity formula, compute the price of the put by Monte Carlo, rather than directly that of the call.

The Monte Carlo method is to all intents and purposes the universal method for pricing options in banks. It is particularly well suited for basket options with several tens of underlying assets.

Note that a variant of the Monte Carlo method involves simulating a binomial tree (possibly with large N; see Section 9.2.6).

Numerical solution of a PDE

A final method involves computing an approximate solution of the PDE by a numerical scheme, such as the finite difference method. Note that such a scheme implies the discretization of both time and space (and replacing $\mathbb{R}_+$ by a bounded set). These numerical methods have been developed by specialists in scientific computation for many other applications. Their main limitation is the dimension of the spatial variable, in the case of a model with several underlying assets.

9.3.13 Historical and implicit volatility

The use of the above models for pricing an option requires knowledge of very few parameters (it should be noticed that the drift parameter μ in the model for the fluctuations of S_t has disappeared). The interest rate r can be assumed to be known.

On the other hand, the parameter σ – the volatility – is not known a priori. One can try to estimate it from past fluctuations; this is the so-called *historical volatility*.

But since there is a market for options, where prices are regulated by the law of supply and demand, one can invert the Black–Scholes formula in order to deduce a value for σ; this is called the *implicit volatility*. It should be noted that inverting the formula for different options on the same underlying asset, with the same exercise time T, but with different strike prices, produces different implicit volatilities – this is the so-called *smile*. This fact contradicts the Black–Scholes model, which is too simple to reflect the full complexity of the markets. On the other hand, its simplicity leads to explicit and intuitive formulae which are very popular with practitioners.

9.4 American options in the discrete model

Unlike the European option, an American option can be exercised at any time between the date of signature of the contract (time 0) and the expiry of the contract (time T). Write $h(S_t)$ for the payoff of the option, if it is exercised at time t. In the case of an American call, $h(x) = (x - K)_+$; and in the case of an American put, $h(x) = (K - x)_+$. Let $Z_t = h(S_t)$, $0 \le t \le T$, and denote by A_t the price of the American option at time t, $0 \le t \le T$. In particular, A_0 is the *premium* which the buyer of the American option must pay upon signing the contract (at time 0).

We again assume that

$$R_t = (1 + r)^t, \quad 0 \le t \le T,$$

and we define the discounted values

$$\tilde{Z}_t = \frac{Z_t}{(1 + r)^t}, \quad \tilde{A}_t = \frac{A_t}{(1 + r)^t}.$$

Let us first try to compute the values A_t by means of a backward recurrence. It is first of all clear that

$$A_T = Z_T.$$

At time $T - 1$, the option holder has the choice between exercising immediately and retaining the option in the hope of making make more profit at time T. Hence,

$$A_{T-1} = Z_{T-1} \vee \frac{1}{1 + r}\mathbb{E}^*(Z_T | S_{T-1}).$$

By the same argument, for all $0 < t \leq T$,

$$A_{t-1} = Z_{t-1} \vee \frac{1}{1+r}\mathbb{E}^*(A_t|S_{t-1}).$$

In terms of discounted values, we get

$$\begin{cases} \tilde{A}_T = \tilde{Z}_T, \\ \tilde{A}_{t-1} = \tilde{Z}_{t-1} \vee \mathbb{E}^*(\tilde{A}_t|S_{t-1}), \quad 0 < t \leq T. \end{cases} \tag{9.8}$$

Definition 4.1 *A supermartingale (submartingale) is an adapted sequence* $\{M_t;\ 0 \leq t \leq T\}$ *such that, for* $0 < t \leq T$,

$$\mathbb{E}[M_t|\mathcal{F}_{t-1}] \leq (\geq) M_{t-1}.$$

Proposition 4.2 *The sequence* $\{\tilde{A}_t;\ 0 \leq t \leq T\}$ *is a* $\mathbb{P}^*$*-supermartingale. It is the smallest* $\mathbb{P}^*$*-supermartingale which is greater than or equal to the sequence* $\{\tilde{Z}_t;\ 0 \leq t \leq T\}$.

Proof The supermartingale property and the fact that the sequence $\{\tilde{A}_t\}$ is bigger than the sequence $\{\tilde{Z}_t\}$ follow directly from (9.8). Now let $\{M_t\}$ be another $\mathbb{P}^*$-supermartingale which is bigger than $\{\tilde{Z}_t\}$. Then $M_T \geq \tilde{Z}_T = \tilde{A}_T$ and, if $M_t \geq \tilde{A}_t$,

$$M_{t-1} \geq \mathbb{E}^*(M_t|\mathcal{F}_{t-1}) \geq \mathbb{E}^*(\tilde{A}_t|\mathcal{F}_{t-1}),$$

hence also $M_{t-1} \geq \tilde{Z}_{t-1}$, consequently

$$M_{t-1} \geq \tilde{Z}_{t-1} \vee \mathbb{E}^*(\tilde{A}_t|\mathcal{F}_{t-1}) = \tilde{A}_{t-1}.$$

□

We now need to describe the smallest supermartingale bigger than a given adapted sequence.

9.4.1 Snell envelope

Recall that a $\{0, 1, \ldots, T\}$-valued random variable ν is called a stopping time if for all $0 \leq t \leq T$,

$$\{\nu = t\} \in \mathcal{F}_t.$$

Given an adapted random sequence $\{X_t;\ 0 \leq t \leq T\}$ and a stopping time ν, the stopped sequence $\{X_{t\wedge\nu};\ 0 \leq t \leq T\}$ is again adapted. This follows from the fact that ν is a stopping time if and only if $\{\nu \leq t\} \in \mathcal{F}_t$ for all $0 \leq t \leq T$. We have moreover the optional stopping theorem:

Theorem 4.3 *If* $\{M_t;\ 0 \leq t \leq T\}$ *is a martingale (supermartingale), then* $\{M_{t\wedge\nu};\ 0 \leq t \leq T\}$ *is also a martingale (supermartingale).*

PROOF It suffices to observe (in the notation of Proposition 2.6) that

$$M_{t\wedge\nu} = M(Y)_t, \quad 0 \le t \le T,$$

if $Y_t = \mathbf{1}_{\{t\le\nu\}}$. Since $\{t \le \nu\} = \{\nu \le t-1\}^c$, the sequence Y is predictable, and the result for martingales follows from Proposition 2.6. The same argument, exploiting the positivity of Y, yields the result for supermartingales. □

Given an adapted sequence $\{Z_t;\ 0 \le t \le T\}$, we wish to study its *Snell envelope*, that is, the smallest supermartingale which is bigger than Z. This is the sequence $\{U_t;\ 0 \le t \le T\}$ defined by

$$\begin{cases} U_T = Z_T, \\ U_t = Z_t \vee \mathbb{E}(U_{t+1}|\mathcal{F}_t), \quad 0 \le t < T. \end{cases}$$

Note that when $U_t > Z_t$, $U_t = \mathbb{E}(U_{t+1}|\mathcal{F}_t)$. This remark can be formalized as follows.

Proposition 4.4 *The random variable defined by*

$$\nu = \inf\{0 \le t \le T | U_t = Z_t\}$$

is a stopping time, and the stopped sequence $\{U_{t\wedge\nu};\ 0 \le t \le T\}$ *is a martingale.*

PROOF Note first that

$$\{\nu = t\} = \{U_0 > Z_0\} \cap \cdots \cap \{U_{t-1} > Z_{t-1}\} \cap \{U_t = Z_t\} \in \mathcal{F}_t.$$

We again write $Y_t = \mathbf{1}_{\{t\le\nu\}}$, $U_{t\wedge\nu} = U(Y)_t$. Then

$$U(Y)_{t+1} - U(Y)_t = \mathbf{1}_{\{t+1\le\nu\}}(U_{t+1} - U_t),$$

and on the set $\{t+1 \le \nu\}$, $U_t = \mathbb{E}(U_{t+1}|\mathcal{F}_t)$, hence $\mathbb{E}(U(Y)_{t+1} - U(Y)_t|\mathcal{F}_t) = 0$. □

Denote by $\mathcal{T}_t$ the set of stopping times which take their values in the set $\{t, t+1, \ldots, T\}$.

Corollary 4.5 *We have*

$$U_0 = \mathbb{E}(Z_\nu|\mathcal{F}_0) = \sup_{\tau\in\mathcal{T}_0} \mathbb{E}(Z_\tau|\mathcal{F}_0).$$

PROOF From Proposition 4.4, $\{U_{\cdot\wedge\nu}\}$ is a martingale, hence

$$U_0 = \mathbb{E}(U_{T\wedge\nu}|\mathcal{F}_0) = \mathbb{E}(Z_\nu|\mathcal{F}_0).$$

If $\tau \in \mathcal{T}_0$, from Theorem 4.3, $\{U_{\cdot\wedge\tau}\}$ is a supermartingale, hence

$$U_0 \ge \mathbb{E}(U_{N\wedge\tau}|\mathcal{F}_0) \ge \mathbb{E}(Z_\tau|\mathcal{F}_0).$$

□

Corollary 4.5 generalizes as

$$U_t = \sup_{\tau \in \mathcal{T}_t} \mathbb{E}(Z_\tau | \mathcal{F}_t) = \mathbb{E}(Z_{\nu_t} | \mathcal{F}_t),$$

if $\nu_t = \inf\{s \geq t | U_s = Z_s\}$. An *optimal stopping time* is a stopping time which satisfies the same optimality property as established in the corollary. The next theorem says that ν is the smallest of the optimal stopping times.

Theorem 4.6 *The stopping time ν is optimal if and only if the following conditions are fulfilled:*

(i) $Z_\nu = U_\nu$;

(ii) $\{U_{t \wedge \nu};\ 0 \leq t \leq T\}$ *is a martingale.*

Proof Conditions (i) and (ii) imply $U_0 = \mathbb{E}(Z_\nu | \mathcal{F}_0)$, hence optimality of ν according to Corollary 4.5.

Suppose now that ν is optimal. Then

$$U_0 = \mathbb{E}(Z_\nu | \mathcal{F}_0) \leq \mathbb{E}(U_\nu | \mathcal{F}_0),$$

and since $\{U_{\cdot \wedge \nu}\}$ is a supermartingale, $\mathbb{E}(U_\nu | \mathcal{F}_0) \leq U_0$, hence $\mathbb{E}(U_\nu | \mathcal{F}_0) = \mathbb{E}(Z_\nu | \mathcal{F}_0) = U_0$, consequently since U is bigger than Z, $U_\nu = Z_\nu$, which is condition (i).

Again, $\{U_{\cdot \wedge \nu}\}$ is a supermartingale, hence

$$U_0 \geq \mathbb{E}(U_{t \wedge \nu} | \mathcal{F}_0) \geq \mathbb{E}(U_\nu | \mathcal{F}_0),$$

but since the two extreme terms coincide, we have

$$\mathbb{E}(U_{t \wedge \nu} | \mathcal{F}_0) = \mathbb{E}(U_\nu | \mathcal{F}_0) = \mathbb{E}(\mathbb{E}(U_\nu | \mathcal{F}_t) | \mathcal{F}_0).$$

Since, on the other hand, $U_{t \wedge \nu} \geq \mathbb{E}(U_\nu | \mathcal{F}_t)$, we have the identity $U_{t \wedge \nu} = \mathbb{E}(U_\nu | \mathcal{F}_t)$, from which (ii) follows. □

9.4.2 Doob's decomposition

Proposition 4.7 *Any supermartingale $\{U_t;\ 0 \leq t \leq T\}$ can be uniquely written in the form*

$$U_t = M_t - C_t,$$

where $\{M_t;\ 0 \leq t \leq T\}$ is a martingale and $\{C_t;\ 0 \leq t \leq T\}$ is an increasing predictable sequence such that $C_0 = 0$.

Proof We must have $M_0 = U_0$ and $C_0 = 0$. Moreover,

$$U_{t+1} - U_t = M_{t+1} - M_t - C_{t+1} + C_t,$$

hence conditioning upon $\mathcal{F}_t$,

$$\mathbb{E}(U_{t+1}|\mathcal{F}_t) - U_t = -C_{t+1} + C_t,$$

and $C_{t+1} \geq C_t$. Furthermore,

$$M_{t+1} - M_t = U_{t+1} - \mathbb{E}(U_{t+1}|\mathcal{F}_t),$$

hence $M_t = \mathbb{E}(M_{t+1}|\mathcal{F}_t)$. □

Proposition 4.8 *Let* $\{Z_t;\ 0 \leq t \leq T\}$ *be an adapted sequence whose Snell envelope is* $U_t = M_t - C_t$. *Then the largest optimal stopping time is given by*

$$\nu_{\max} = \inf\{t \leq T;\ C_{t+1} \neq 0\}.$$

PROOF The fact that $\nu_{\max}$ is a stopping time follows from the fact that $\{C_\cdot\}$ is predictable. Since $C_t = 0$ for $t \leq \nu_{\max}$, we have

$$U_{\cdot\wedge\nu_{\max}} = M_{\cdot\wedge\nu_{\max}},$$

hence $\{U_{\cdot\wedge\nu_{\max}}\}$ is a martingale. From Theorem 4.6, optimality of $\nu_{\max}$ will follow from the identity $U_{\nu_{\max}} = Z_{\nu_{\max}}$. Write $Y_t = \mathbf{1}_{\{\nu_{\max}=t\}}$.

$$\begin{aligned} U_{\nu_{\max}} &= \sum_{t=0}^{T-1} Y_t U_t + Y_T U_T \\ &= \sum_{t=0}^{T-1} Y_t \max(Z_t, \mathbb{E}(U_{t+1}|\mathcal{F}_t)) + Y_T Z_T \\ &= \sum_{t=0}^{T-1} Y_t Z_t + Y_T Z_T \\ &= Z_{\nu_{\max}}, \end{aligned}$$

where we have used the following facts: $\mathbb{E}(U_{t+1}|\mathcal{F}_t) = M_t - C_{t+1}$ and, on the set $\{Y_t = 1\}$, $C_t = 0$ and $C_{t+1}>0$, hence $\mathbb{E}(U_{t+1}|\mathcal{F}_t) = M_t - C_{t+1} < U_t$, which implies that $U_t = Z_t$.

It remains to show that there is no optimal stopping time ν such that $\nu \geq \nu_{\max}$ and $\mathbb{P}(\nu>\nu_{\max})>0$. Indeed, that would imply that

$$\mathbb{E}(U_\nu) = \mathbb{E}(M_\nu) - \mathbb{E}(C_\nu) = \mathbb{E}(U_0) - \mathbb{E}(C_\nu) < \mathbb{E}(U_0),$$

which would prevent $\{U_{\cdot\wedge\nu}\}$ from being a martingale. □

9.4.3 Snell envelope and Markov chain

We now specify the Snell envelope in the Markovian case. Let $\{X_t;\ 0 \le t \le T\}$ be a homogeneous E-valued Markov chain, where E is a finite set, with transition matrix $P = \{P_{xy};\ x, y \in E\}$. We assume that, for $0 \le t \le T$, $\mathcal{F}_t = \sigma\{X_0, \ldots, X_t\}$. The following proposition is a consequence of the above results.

Proposition 4.9 *Let $\{Z_t\}$ be a sequence defined by $Z_t = \psi(t, X_t)$, with $\psi : \{0, 1, \ldots, T\} \times E \to \mathbb{R}$. Then the Snell envelope $\{U_t\}$ of the sequence $\{Z_t\}$ is given by the formula $U_t = u(t, X_t)$, where the function u is defined by*

$$\begin{cases} u(T, x) = \psi(T, x), \quad x \in E, \\ u(t, x) = \psi(t, x) \vee \sum\limits_y P_{xy} u(t+1, y), \quad x \in E,\ 0 \le t < T. \end{cases}$$

In practice, to determine the stopping time

$$\nu = \inf\{t;\ U_t = Z_t\},$$

one might compute the solution u of the backward recurrence from Proposition 4.9, and stop at time $\nu = \inf\{t \le T;\ u(t, X_t) = \psi(t, X_t)\}$.

9.4.4 Back to American options

By Proposition 4.2, the discounted price $\{\tilde{A}_t\}$ of the American option is the $\mathbb{P}^*$-Snell envelope of the sequence $\{\tilde{Z}_t = (1+r)^{-t} h(S_t) = R_t^{-1} h(S_t)\}$. We know (generalization of Corollary 4.5) that

$$\tilde{A}_t = \sup_{\nu \in \mathcal{T}_t} \mathbb{E}^*(R_\nu^{-1} h(S_\nu) | \mathcal{F}_t),$$

hence

$$A_t = R_t \sup_{\nu \in \mathcal{T}_t} \mathbb{E}^*(R_\nu^{-1} h(S_\nu) | \mathcal{F}_t).$$

From Doob's decomposition, $\tilde{A}_t = \tilde{M}_t - \tilde{C}_t$, with $\{\tilde{M}_t\}$ a $\mathbb{P}^*$-martingale and $\tilde{C}_t$ increasing, predictable and null at 0.

Since the market is complete, there exist an initial wealth V_0 and a self-financing strategy $\{(X_t, Y_t)\}$ such that $V_T(X, Y) = R_T \tilde{M}_T$, hence $\tilde{V}_T(X, Y) = \tilde{M}_T$, and consequently, for all $0 \le t \le T$, $\tilde{V}_t(X, Y) = \tilde{M}_t$, since both are $\mathbb{P}^*$-martingales. It follows that $\tilde{A}_t = \tilde{V}_t(X, Y) - \tilde{C}_t$ and $A_t = V_t(X, Y) - C_t$, where $C_t = R_t \tilde{C}_t$. An optimal exercise time τ satisfies $A_\tau = h(S_\tau)$. It must also satisfy $\tau \le \nu_{\max} = \inf\{t;\ C_{t+1} \ne 0\}$, since by exercising the option at time $\nu_{\max}$, its holder earns $A_{\nu_{\max}} = V_{\nu_{\max}}(X, Y)$, and with the strategy $\{(X, Y)\}$ his portfolio is worth strictly more than the option at times $\nu_{\max} + 1, \nu_{\max} + 2, \ldots, T$.

One can also check that if the option holder exercises it at a non-optimal time, the seller makes a strictly positive profit.

9.4.5 American and European options

Proposition 4.10 *Let A_t denote the price at time t of an American option which pays Z_t if exercised at time t, and E_t the price at time t of a European option which pays Z_T at the exercise time. Then $A_t \geq E_t$, $0 \leq t \leq T$. If, moreover, $E_t \geq Z_t$ for all t, then $E_t = A_t$ for all t.*

PROOF The first identity follows clearly from the martingale (supermartingale) property of $\{E_t\}$ ($\{A_t\}$) under $\mathbb{P}^*$. If $E_t \geq Z_t$, $\{\tilde{E}_t\}$ is a $\mathbb{P}^*$-martingale (hence also a supermartingale) which is greater than $\{\tilde{Z}_t\}$, hence also greater than $\{\tilde{A}_t\}$, by Proposition 4.2. □

Corollary 4.11 *In the case of European and American calls with the same T and the same strike price K, on the same underlying asset with price S_t at time t, $A_t = E_t$, $0 \leq t \leq T$.*

PROOF We have $Z_t = (S_t - K)_+$ and

$$\tilde{E}_t = R_T^{-1}\mathbb{E}^*((S_T - K)_+|\mathcal{F}_t) \geq \mathbb{E}^*(\tilde{S}_T - R_T^{-1}K|\mathcal{F}_t) = \tilde{S}_t - R_T^{-1}K,$$

hence

$$E_t \geq S_t - \frac{R_t}{R_T}K \geq S_t - K,$$

but also $E_t \geq 0$, hence $E_t \geq Z_t$. It remains to apply Proposition 4.10. □

This property is not satisfied by a put or a call on an asset which distributes a dividend.

9.4.6 American options and Markov model

Consider again an American option which pays $Z_t = h(S_t)$ if exercised at time t, and suppose now that $\{S_t;\ 0 \leq t \leq T\}$ is a Markov chain. Then the price A_t can be written in the form $A_t = u(t, S_t)$. Write $\tilde{u}(t, x) = R_t^{-1}u(t, x)$ and $\tilde{h}(t, x) = R_t^{-1}h(x)$. It follows from Proposition 4.9 that

$$\tilde{u}(t, x) = \tilde{h}(t, x) \vee \sum_y P_{xy}\tilde{u}(t + 1, y),$$

from which we deduce the recurrence formula

$$u(t, x) = h(x) \vee \sum_y P_{xy}\frac{u(t + 1, y)}{1 + r}. \tag{9.9}$$

An optimal exercise time is defined by

$$\nu = \inf\{t \leq T;\ u(t, S_t) = h(S_t)\}.$$

9.5 American options in the Black–Scholes model

The study of American options in the Black–Scholes model requires sophisticated mathematical tools. We shall describe the associated PDE as a limit of the formulae in Section 4.6.

Let us rewrite formula (9.9) in the framework of Section 2.6, with the following changes of variable:

$$g(x) = h(e^x), \quad v(t,x) = u(t,e^x).$$

We get

$$v\left(t-\frac{1}{N},x\right) = g(x) \vee e^{-r/N}\left(p_+^N v\left(t,x+\frac{r}{N}+\frac{\sigma}{\sqrt{N}}\right) + p_-^N v\left(t,x+\frac{r}{N}-\frac{\sigma}{\sqrt{N}}\right)\right),$$

with

$$p_+^N = \mathbb{P}\left(\eta_k^N = \frac{\sigma}{\sqrt{N}}\right) = \frac{1}{2} - \frac{\sigma}{4\sqrt{N}} + O(N^{-3/2}),$$

$$p_-^N = \mathbb{P}\left(\eta_k^N = -\frac{\sigma}{\sqrt{N}}\right) = \frac{1}{2} + \frac{\sigma}{4\sqrt{N}} + O(N^{-3/2}).$$

Define

$$(A_N v)(t,x)e^{-r/N}\left(p_+^N v\left(t,x+\frac{r}{N}+\frac{\sigma}{\sqrt{N}}\right) + p_-^N v\left(t,x+\frac{r}{N}-\frac{\sigma}{\sqrt{N}}\right)\right).$$

Then (9.10) can be rewritten

$$v\left(t-\frac{1}{N},x\right) \geq g(x),$$

$$(A_N v)(t,x) - v\left(t-\frac{1}{N},x\right) \leq 0,$$

$$\left(v\left(t-\frac{1}{N},x\right) - g(x)\right)\left((A_N v)(t,x) - v\left(t-\frac{1}{N},x\right)\right) = 0.$$

If we assume that v is sufficiently regular, we get after multiplication by N, with the help of an expansion, that as $N \to \infty$,

$$N\left[(A_N v)(t,x) - v\left(t-\frac{1}{N},x\right)\right] \to$$

$$Av(t,x) := \frac{\partial v}{\partial t}(t,x) + \left(r - \frac{\sigma^2}{2}\right)\frac{\partial v}{\partial x}(t,x) + \frac{\sigma^2}{2}\frac{\partial^2 v}{\partial x^2}(t,x) - rv(t,x).$$

Then the price of the American option is

$$A_t = v(t, \log S_t),$$

where v is the solution of the variational inequality

$$\begin{aligned} v(t,x) &\geq g(x), \\ Av(t,x) &\leq 0, \\ (v(t,x) - g(x))Av(t,x) &= 0, \end{aligned}$$

and an optimal exercise time for the option is given by the stopping time

$$\nu = \inf\{t \leq T;\ v(t, \log S_t) = g(\log S_t) = h(S_t)\}.$$

9.6 Interest rate and bonds

So far we have assumed that the interest rate is a constant (both deterministic and constant in time). Such an assumption is reasonable when dealing with stocks and options on stocks, but not with bonds and options on bonds.

By a *zero-coupon bond* we mean a security which pays one dollar at the exercise time T. We denote by $O_{t,T}$ the value at time t of this security.

9.6.1 Future interest rate

The interest rate on a loan depends on both its issue date t and its maturity T. A person borrowing one dollar at time t must repay R_T^t dollars at time T. In the case of the constant interest rate of the Black–Scholes model, we would have

$$R_T^t = \exp[(T-t)r].$$

More generally, in a deterministic framework, where all quantities $\{R_T^t;\ 0 \leq t \leq T\}$ are known, the absence of arbitrage opportunity requires R to satisfy

$$R_T^t = R_u^t R_T^u, \quad \forall 0 \leq t < u < T.$$

From this relation, together with $R_t^t = 1$ and the continuity of R, we deduce that there must exist a function $t \to r(t)$ such that

$$R_T^t = \exp\left(\int_t^T r(s)ds\right), \quad \forall 0 \leq t \leq T.$$

Consequently, we must have

$$O_{t,T} = \exp\left(-\int_t^T r(s)ds\right).$$

9.6.2 Future interest rate and bonds

Writing R_t for R_t^0, we now assume that

$$R_t = \exp\left(\int_0^t r_s ds\right),$$

where $\{r_t;\ t \geq 0\}$ is a stochastic process which is adapted to the natural filtration $\{\mathcal{F}_t;\ t \geq 0\}$ of a Brownian motion $\{W_t;\ t \geq 0\}$ (i.e. up to zero measure sets, $\mathcal{F}_t = \sigma\{W_s;\ 0 \leq s \leq t\}$). r_t is called the *instantaneous rate*.

We shall assume that there exists a probability $\mathbb{P}^*$ equivalent to $\mathbb{P}$ under which $\tilde{O}_{t,u} := (R_t)^{-1} O_{t,u}$, $0 \leq t \leq u$, is a martingale, for all $0 < u \leq T$. Since $O_{u,u} = 1$, $\tilde{O}_{u,u} = (R_u)^{-1}$, this assumption implies that

$$\tilde{O}_{t,u} = \mathbb{E}^*\left(\exp\left[-\int_0^u r_s ds\right] | \mathcal{F}_t\right),$$

hence also

$$O_{t,u} = \mathbb{E}^*\left(\exp\left[-\int_t^u r_s ds\right] | \mathcal{F}_t\right).$$

If we wish to write the quantity $O_{t,u}$ explicitly, we need to compute the Radon–Nikodym derivative of the probability $\mathbb{P}^*$ with respect to $\mathbb{P}$. Denote this density by L_T; it is such that, for any bounded random variable X,

$$\mathbb{E}^*(X) = \mathbb{E}(XL_T).$$

If, moreover, X is $\mathcal{F}_t$-measurable, then writing $L_t = \mathbb{E}(L_T|\mathcal{F}_t)$, we have

$$\mathbb{E}^*(X) = \mathbb{E}(XL_t).$$

L_t is the density of the restriction to $\mathcal{F}_t$ of $\mathbb{P}^*$ with respect to $\mathbb{P}$. It follows from Girsanov's theorem (which was partly proved in Section 9.3.8 above).

Proposition 6.1 *There exists an adapted process* $\{q_t;\ 0 \leq t \leq T\}$ *satisfying*

$$\int_0^T q_t^2 dt < \infty \ a.s.,$$

such that, for all $0 \leq t \leq T$, *almost surely*

$$L_t = \exp\left(\int_0^t q_s dW_s - \frac{1}{2}\int_0^t q_s^2 ds\right).$$

Corollary 6.2 *The price at time* t *of a zero-coupon bond with maturity* $u \geq t$ *is of the form*

$$O_{t,u} = \mathbb{E}\left(\exp\left[-\int_t^u \left(r_s + \frac{q_s^2}{2}\right) ds + \int_t^u q_s dW_s\right] | \mathcal{F}_t\right).$$

PROOF Write $X = \exp(-\int_t^u r_s ds)$. We need to compute $\mathbb{E}^*(X|\mathcal{F}_t)$. Let Y be an $\mathcal{F}_t$-measurable and bounded random variable. We have

$$\begin{aligned}\mathbb{E}^*(XY) &= \mathbb{E}(XYL_T)\\ &= \mathbb{E}(\mathbb{E}[XL_T|\mathcal{F}_t]Y)\\ &= \mathbb{E}^*\left(\frac{\mathbb{E}[XL_T|\mathcal{F}_t]}{L_t}Y\right),\end{aligned}$$

hence from the characterization of $\mathbb{E}^*(X|\mathcal{F}_t)$,

$$\mathbb{E}^*(X|\mathcal{F}_t) = \frac{\mathbb{E}[XL_T|\mathcal{F}_t]}{L_t}.$$

The result follows. □

Proposition 6.3 *For any maturity u, there exists an adapted process* $\{\sigma_t^u;\ 0 \le t \le u\}$ *such that, on* $[0, u]$,

$$dO_{t,u} = (r_t - \sigma_t^u q_t)O_{t,u}dt + O_{t,u}\sigma_t^u dW_t.$$

PROOF Using first the formula established in the proof of Corollary 6.2, then the martingale property of $\{\tilde{O}_{t,u};\ 0 \le t \le u\}$ under $\mathbb{P}^*$, we deduce that, for $0 \le s \le t \le u$,

$$\mathbb{E}(\tilde{O}_{t,u}L_t|\mathcal{F}_s) = \mathbb{E}^*(\tilde{O}_{t,u}|\mathcal{F}_s)L_s = \tilde{O}_{s,u}L_s.$$

Hence, $\{\tilde{O}_{t,u}L_t;\ 0 \le t \le u\}$ is a martingale under $\mathbb{P}$, which is strictly positive, and its log is a semimartingale, whose martingale part takes the form

$$\int_0^t \theta_s^u dW_s,$$

and again since $\{\tilde{O}_{t,u}L_t;\ 0 \le t \le u\}$ is a $\mathbb{P}$-martingale,

$$\tilde{O}_{t,u}L_t = \tilde{O}_{0,u}\exp\left(\int_0^t \theta_s^u dW_s - \frac{1}{2}\int_0^t (\theta_s^u)^2 ds\right).$$

Multiplying by $R_t(L_t)^{-1}$, we get

$$O_{t,u} = O_{0,u}\exp\left(\int_0^t \left(r_s - \frac{(\theta_s^u)^2 - q_s^2}{2}\right)ds + \int_0^t (\theta_s^u - q_s)dW_s\right).$$

The result follows by using Itô's formula, if we write $\sigma_t^u = \theta_t^u - q_t$. □

Comparing the formula in Proposition 6.3 with that concerning the interest rate, namely

$$dR_t = r_t R_t dt,$$

we see that the bond is 'more risky' than a bank account. The term $r_t - \sigma_t^u q_t$ is a sort of mean relative return per unit time of the obligation; $-\sigma_t^u q_t$ is the difference between that return and the riskless rate. Hence the interpretation of $-q_t$ as a 'risk premium'. Note that under the risk-neutral probability $\mathbb{P}^*$, $W_t^* = W_t - \int_0^t q_s ds$ is a standard Brownian motion and

$$dO_{t,u} = r_t O_{t,u} dt + O_{t,u} \sigma_t^u dW_t^*.$$

9.6.3 Option based on a bond

Consider the case of a European option with exercise time T on a zero-coupon bond with maturity T', with $T \leq T'$. In case of a call with strike price K, the price of the option at time T is clearly $(O_{T,T'} - K)_+$, and we shall see that we can apply the same methodology here as in Section 3.7.

The evolution of the hedging portfolio is given in the case of a self-financing strategy by the formula

$$dV_t(X, Y) = X_t dR_t + Y_t dO_{t,T'}.$$

Definition 6.4 *A strategy* $\{(X_t, Y_t);\ 0 \leq t \leq T\}$ *is admissible if it is self-financing and if the discounted value*

$$\tilde{V}_t(X, Y) = X_t + Y_t \tilde{O}_{t,T'}$$

of the corresponding portfolio is positive and square integrable under $\mathbb{P}^*$ *for all* t.

Under reasonable assumptions, we can hedge a European option with exercise time $T < T'$.

Proposition 6.5 *Suppose that* $\sup_{0 \leq t \leq T} |r_t| < \infty$ *almost surely,* $\inf_{0 \leq t \leq T} |\sigma_t^{T'}| > 0$, *and* $T < T'$. *Let* H *be an* $\mathcal{F}_T$*-measurable random variable such that* $He^{-\int_0^T r_s ds}$ *is* $\mathbb{P}^*$ *square integrable. Then there exist* V_0 *and an admissible strategy* $\{(X, Y)\}$ *such that* $V_T(X, Y) = H$. *Moreover,*

$$V_t(X, Y) = \mathbb{E}^* \left(e^{-\int_t^T r_s ds} H | \mathcal{F}_t \right).$$

Proof We have

$$d\tilde{V}_t(X, Y) = Y_t d\tilde{O}_{t,T'} = Y_t \tilde{O}_{t,T'} \sigma_t^{T'} dW_t^*.$$

We then deduce that $\{\tilde{V}_t;\ 0 \leq t \leq T\}$ is a $\mathbb{P}^*$-martingale, hence

$$V_t(X, Y) = e^{\int_0^t r_s ds} \mathbb{E}^* \left(e^{-\int_0^T r_s ds} H | \mathcal{F}_t \right).$$

It remains to produce a corresponding strategy. It follows from Itô's representation theorem that

$$He^{-\int_0^T r_s ds} = \mathbb{E}^* \left(He^{-\int_0^T r_s ds} \right) + \int_0^T J_t dW_t^*,$$

with a given process $\{J_t;\ 0 \le t \le T\}$ such that $\int_0^T J_t^2 dt < \infty$ almost surely. It suffices to choose

$$Y_t = \frac{J_t}{\tilde{O}_{t,T'}\sigma_t^{T'}}, \quad X_t = \mathbb{E}^*\left(He^{-\int_0^T r_s ds}|\mathcal{F}_t\right) - \frac{J_t}{\sigma_t^{T'}}.$$

□

9.6.4 An interest rate model

We consider *Vasicek*'s model, which is the simplest interest rate model (though probably not the most satisfying). In this model, the process $\{r_t;\ 0 \le t \le T\}$ solves the SDE

$$dr_t = a(b - r_t)dt + \sigma dW_t,$$

where a, b and σ are positive constants, and the process q from Proposition 6.1 is also a constant, $q_t \equiv -\lambda$. Then if $W_t^* = W_t + \lambda t$, and $b^* = b - \lambda\sigma/a$, we have equivalently

$$dr_t = a(b^* - r_t)dt + \sigma dW_t^*.$$

We can easily show that r_t can be rewritten as

$$r_t = r_0 e^{-at} + b(1 - e^{-at}) + \sigma e^{-at}\int_0^t e^{as} dW_s,$$

and that the law of r_t under $\mathbb{P}$ is the Gaussian probability distribution

$$N\left(r_0 e^{-at} + b(1 - e^{-at}),\ \sigma^2 \frac{1 - e^{-2at}}{2a}\right).$$

The law of r_t under $\mathbb{P}^*$ is the same law, with b replaced by b^*. Hence, r_t takes a negative value with non-zero probability. However, this probability might be very small.

Consider now the price of the zero-coupon bond,

$$\begin{aligned} O_{t,T} &= \mathbb{E}^*\left(e^{-\int_t^T r_s ds}|\mathcal{F}_t\right) \\ &= e^{-b^*(T-t)}\mathbb{E}^*\left(e^{-\int_t^T X_s ds}|\mathcal{F}_t\right), \end{aligned}$$

with $\{X_s = r_s - b^*\}$ solution of the SDE

$$dX_s = -aX_s ds + \sigma dW_s^*. \tag{9.11}$$

Then

$$\mathbb{E}^*\left(e^{-\int_t^T X_s ds}|\mathcal{F}_t\right) = F(T - t, r_t - b^*),$$

where F is defined by

$$F(s, x) = \mathbb{E}^*\left(e^{-\int_0^s X_t^x dt}\right),$$

$\{X_t^x\}$denoting the solution of the SDE (9.11) satisfying $X_0 = x$. Since $\int_0^s X_t^x dt$ is a Gaussian random variable,

$$F(s,x) = \exp\left(-\mathbb{E}^* \int_0^s X_t^x dt + \frac{1}{2}\operatorname{var}\left[\int_0^s X_t^x dt\right]\right).$$

Now

$$\mathbb{E}^* \int_0^s X_t^x dt = x\frac{1-e^{-as}}{a}$$

and

$$\operatorname{var}\left[\int_0^s X_t^x dt\right] = \int_0^s \int_0^s \operatorname{cov}(X_t^x, X_r^x) dt dr,$$

but

$$\operatorname{cov}(X_t^x, X_r^x) = \sigma^2 e^{-a(t+r)}\mathbb{E}^*\left(\int_0^t e^{as} dW_s^* \int_0^r e^{as} dW_s^*\right)$$

$$= \sigma^2 e^{-a(t+r)}\frac{e^{2a(t\wedge r)}-1}{2a},$$

hence

$$Var\left[\int_0^s X_t^x dt\right] = \frac{\sigma^2 s}{a^2} - \frac{\sigma^2}{a^3}(1-e^{-as}) - \frac{\sigma^2}{2a^3}(1-e^{-as})^2.$$

Finally,

$$O_{t,T} = \exp\left(-(T-t)R(T-t, r_t)\right),$$

where $R(T-t, r_t)$, the mean interest rate over the period $[t, T]$, is given by the formula

$$R(s,r) = R - (as)^{-1}\left[(R-r)(1-e^{-as}) - \frac{\sigma^2}{4a^2}(1-e^{-as})^2\right],$$

with $R = b^* - \sigma^2/2a^2$. R is interpreted as a long-term rate, which here is independent of the spot instantaneous rate r. This last property is a defect of the model.

9.7 Exercises

Exercise 7.1 *From the corresponding inequalities at time T, show that if C_t denotes the price of a European call option at time t with exercise time T and strike price K on an underlying asset whose price $\{S_t\}$ evolves according to the Black–Scholes model, then*

$$S_t - Ke^{-r(T-t)} \le C_t \le S_t.$$

Show that the price P_t of the European put satisfies

$$P_t \le Ke^{-r(T-t)}.$$

Exercise 7.2 *Let* $\{S_t;\ 0 \le t \le T\}$ *denote the price of an underlying asset which follows the Black–Scholes model. Write* $U_t = \log S_t$. *Deduce from Itô's formula the differential of* U_t, *in terms of* dt *and* dB_t^*. *Define* $v(t, y) = u(t, e^y)$, *where* $\{u(t, x)\}$ *is the solution of the Black–Scholes PDE. Write a parabolic PDE satisfied by* $\{v(t, y);\ 0 \le t \le T,\ y \in \mathbb{R}\}$.

Exercise 7.3 *(A chooser option) Consider an option over an underlying asset* $\{S_t;\ 0 \le t \le T\}$ *(evolving according to the Black–Scholes model) which gives its holder the choice at a contracted time* $0 \le t \le T$ *between a call and a put option, both being exercised (if at all) at time* T, *with strike price* $K > 0$. *At time* 0, *the buyer of the option pays the premium* X_0; *at time* t *he chooses between call and put; at time* T *he has the right to exercise the option chosen at time* t, *the strike price being* K *in both cases.*

1. *Denote by* C_t (P_t) *the price of the call (put) at time* t. *Show that the best choice is to choose the call (put) if* $C_t > P_t$ $(P_t > C_t)$. *Check with the help of the call–put parity formula that* $C_t \neq P_t$ *almost surely, under* $\mathbb{P}$ *as well as under* $\mathbb{P}^*$.

2. *Deduce that at time* T *the option pays almost surely*

$$\begin{aligned} H &= (S_T - K)_+ \mathbf{1}_{\{C_t \ge P_t\}} + (K - S_T)_+ \mathbf{1}_{\{C_t < P_t\}} \\ &= (S_T - K)_+ + (K - S_T)\mathbf{1}_{\{C_t < P_t\}} \\ &= (K - S_T)_+ + (S_T - K)\mathbf{1}_{\{C_t > P_t\}}. \end{aligned}$$

3. *Recall that the general theory of options tells us that* $X_0 = e^{-rT}\mathbb{E}^*(H)$. *Write the events* $F_t = \{C_t < P_t\}$ *and* $G_t = \{C_t > P_t\}$ *in terms of* S_t, K *and* $r(T - t)$. *Show that* $e^{-rT}\mathbb{E}^*(S_T \mathbf{1}_{F_t}) = e^{-rt}\mathbb{E}^*(S_t \mathbf{1}_{F_t})$, *and similarly for* G_t.

4. *Deduce formulae for the quantities* $X_0 - C_0$ *and* $X_0 - P_0$, *which you should write explicitly in a form similar to that of the formulae for* C_0 *and* P_0 *at the end of Section 9.2.*

5. *Show that the price of the chooser option at time* t *is* $\max(C_t, P_t)$. *Write* $\{u(s, x)\}$ *for the solution of the Black–Scholes PDE for the call, and* $\{v(s, x)\}$ *for the solution of the same PDE for the put (these PDEs differ only in their final condition at time* T*). Denote, finally, by* $\{w(s, x);\ 0 \le s \le t, x > 0\}$ *the solution of the same PDE with the final condition* $\sup(u(t, x), v(t, x))$ *at time* t. *Describe a hedging strategy for the chooser option, in terms of these three functions.*

Exercise 7.4 *(Programming) Consider a European call option over an underlying asset whose price at time* 0 *is* 105 *dollars, with strike price* $K = 110$ *dollars, exercise time* 1 *year, interest rate equal to 5% (i.e.* $rT = 0.05$*), and a volatility such that* $\sigma\sqrt{T} = 0.3$.

1. *Compute the price of the call using the Monte Carlo method applied to the formula*

$$C_0 = \mathbb{E}^*\left[\left(S_T - Ke^{-rT}\right)_+\right],$$

with 1000 simulations. Find (approximately) the variance of the random variable whose expectation you are trying to evaluate, and give a confidence interval for the desired quantity.

2. *Do the same computation (including the confidence interval) combining the same method applied to the formula for the price of the put,*

$$P_0 = \mathbb{E}^*\left[\left(Ke^{-rT} - S_T\right)_+\right],$$

with the same number of simulations, and the call–put parity formula.

3. *Compute the price of the same call option a third time, using the formula*

$$C_0 = S_0 F(d_1) - Ke^{-rT} F(d_2),$$

with

$$d_1 = \frac{1}{\sigma\sqrt{T}} \log\left(\frac{S_0}{K}\right) + \frac{r\sqrt{T}}{\sigma} + \frac{\sigma\sqrt{T}}{2},$$
$$d_2 = \frac{1}{\sigma\sqrt{T}} \log\left(\frac{S_0}{K}\right) + \frac{r\sqrt{T}}{\sigma} - \frac{\sigma\sqrt{T}}{2},$$

and the distribution function F of the Gaussian law $N(0, 1)$ given by Matlab.

4. *Compare the results with those given by a pricer on the internet.*

5. *The market price for the above option being 15 dollars, deduce the implicit volatility (i.e. invert the Black–Scholes formula) using Newton's method.*

Exercise 7.5 *(Programming) In the model with several underlying assets from Section 3.10, denote by C the price of a call option with exercise time T, strike price K, over a basket containing a_i shares of stock i, $1 \leq i \leq d$, and P the price of the corresponding put; that is, $C = E_0$ with $H = h(S_t)$ and*

$$h(x) = \left(\sum_{i=1}^{d} a_i x_i - K\right)_+,$$

and let P be given by the same formula with

$$h(x) = \left(K - \sum_{i=1}^{d} a_i x_i\right)_+,$$

where E_0 is defined by formula (9.7). Apply the Monte Carlo method to compute the price of this option with $d = 5$,

$$a = \begin{pmatrix} 5 \\ 3 \\ 8 \\ 2 \\ 4 \end{pmatrix}, \quad S_0 = \begin{pmatrix} 80 \\ 95 \\ 105 \\ 75 \\ 35 \end{pmatrix}, \quad K = 2000, \quad rT = 0.05,$$

$$\sqrt{T}\Sigma = \begin{pmatrix} 0.3 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \end{pmatrix}.$$

1. *Compute the price of the call by applying the Monte Carlo method to the computation of the formula for C_0 with 1000 simulations. Find (approximately) the variance of the random variable whose expectation you are trying to estimate, and give a confidence interval for the searched quantity.*

2. *Write the call–put parity formula explicitly in this case. Repeat the same computation (including the confidence interval) by combining the same method applied to the formula for the price of the put P_0 with the same number of simulations, and the call–put parity formula. How do these two approaches compare?*

10

Solutions to selected exercises

10.1 Chapter 1

Exercise 1.5.1

1. We draw i.i.d. random variables $\{U_k;\ k \geq 1\}$, whose common probability distribution is uniform on the interval [0, 1], and we define

$$X = \inf\{k \geq 1;\ U_k \leq p\};$$

in other words, X is the index of the first of those U_k which are smaller than p.

2. Let U denote a random variable whose probability distribution is uniform on [0, 1]. Define

$$X = \left[\frac{\log U}{\log(1-p)}\right] + 1,$$

where $[x]$ denotes the integer part of the real number x. Then for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}(X > k) &= \mathbb{P}\left(\left[\frac{\log U}{\log(1-p)}\right] \geq k\right)\\ &= \mathbb{P}\left(\frac{\log U}{\log(1-p)} \geq k\right)\\ &= \mathbb{P}(U \leq (1-p)^k)\\ &= (1-p)^k,\end{aligned}$$

hence the probability distribution of X is the geometric distribution. This method is more efficient than the previous one, since one unique U produces one X.

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

Exercise 1.5.2

1. The classical method is the Box–Müller method, which is presented in Section 1.3

2. We need to find $k > 1$ such that, for all $x > 0$,

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \leq k \frac{\lambda}{2} \exp(-\lambda x),$$

which is equivalent to

$$\sup_{x>0} \exp\left(\lambda x - \frac{x^2}{2}\right) = \exp\left(\frac{\lambda^2}{2}\right) \leq k\lambda\sqrt{\frac{\pi}{2}},$$

or

$$k \geq \sqrt{\frac{2}{\pi}} \lambda^{-1} \exp\left(\frac{\lambda^2}{2}\right).$$

The choice of λ which minimizes the right-hand side above is $\lambda = 1$, so we should choose

$$k = \sqrt{\frac{2e}{\pi}}.$$

With the notation of Proposition 3.2, f denoting the density of the Gaussian probability distribution $N(0, 1)$, g the density of the double exponential with parameter 1,

$$\alpha(x) = \frac{f(x)}{kg(x)} = \exp\left(-\frac{(|x| - 1)^2}{2}\right).$$

The rejection algorithm works as follows. At each step, we draw three mutually independent random variables U, V and W, whose probability distribution is uniform on $[0, 1]$. We write

$$X = -\log(U) \times r(V),$$

where

$$r(v) = \begin{cases} 1, & \text{if } v < 1/2, \\ -1, & \text{if } v \geq 1/2, \end{cases}$$

hence the law of X is double exponential with parameter 1. If $W \leq \alpha(X)$, accept X; if not, start afresh.

3. (a) This law is that of $|V|$, where the law of V is $N(0, 1)$. Indeed, since $\mathbb{P}(Y > y) = \exp(-y)$ and X and Y are independent,

$$\mathbb{P}(Y > (1 - X)^2/2) = \int_0^\infty e^{-(1-x)^2/2} e^{-x} dx = \sqrt{\frac{\pi}{2e}}.$$

Moreover,

$$\mathbb{P}(X > x, Y > (1-X)^2/2) = \int_x^\infty e^{-(1-y)^2/2} e^{-y} dy$$
$$= \frac{1}{\sqrt{e}} \int_x^\infty e^{-y^2/2} dy,$$

hence

$$\mathbb{P}(X > x | Y > (1-X)^2/2) = \sqrt{\frac{2}{\pi}} \int_x^\infty e^{-y^2/2} dy = \mathbb{P}(|V| > x).$$

(b) If $Z = |V|$ is the absolute value of a standard normal random variable, and S is independent of Z and equals ± 1 with probability $1/2$, then the law of ZS is $N(0, 1)$ since, for $x < 0$,

$$\mathbb{P}(ZS < x) = \mathbb{P}(S = -1)\mathbb{P}(Z > -x)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-x}^\infty e^{-y^2/2} dy$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy,$$

and for $x > 0$,

$$\mathbb{P}(ZS < x) = \mathbb{P}(S = -1) + \mathbb{P}(S = 1)\mathbb{P}(Z < x)$$
$$= \frac{1}{2} + \frac{1}{2}\left(1 - \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-y^2/2} dy\right)$$
$$= 1 - \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy.$$

(c) We simulate three mutually independent random variables, U, V and W, the law of each being uniform on $[0, 1]$. We let $X = -\log(U)$, $Y = -\log(V)$, $S = r(W)$ (with the function r as defined in part 2). If $Y > (1-X)^2/2$, the simulated value is XS. If not, we reject and start afresh.

Exercise 1.5.5

1. We simulate a sequence $\{X_n;\ n \geq 1\}$ of mutually independent random variables, each having the same law as X, and we accept the first value which is greater than or equal to m (for the justification of this procedure, see the end of the proof of Proposition 3.2). The bigger m is, the smaller $\mathbb{P}(X > m)$ is, and the less efficient the method is.

2. If $x \leq m$, $\mathbb{P}(Z \leq x) = 0$. If $x > m$,

$$\begin{aligned}\mathbb{P}(Z \leq x) &= \mathbb{P}\left(U \leq \frac{F(x) - F(m)}{1 - F(m)}\right) \\ &= \frac{F(x) - F(m)}{1 - F(m)} \\ &= \frac{\mathbb{P}(m < X \leq x)}{\mathbb{P}(X > m)} \\ &= \mathbb{P}(X \leq x | X > m).\end{aligned}$$

Hence, provided that $F(m)$ is known or computable, and F^{-1} is also known or at least approximately computable, we deduce a method of simulation of the law of X, conditioned upon $\{X > m\}$, which only requires the simulation of one random variable whose law is uniform on $[0, 1]$.

3. More generally, if $a < b$, the law of the random variable

$$Z = F^{-1}(F(a) + (F(b) - F(a))U)$$

is the conditional law of X, given that $a < X < b$.

4. Note that $X = \mu + \sigma Y$, with Y a standard normal random variable, and that $\{X > m\} = \{Y > (m - \mu)/\sigma\}$.

5. Replacing m by m', we may assume that we wish to simulate a standard normal random variable X, conditioned upon $X > m$. We need to choose θ such that (with the notation $\bar{F}(m) = \mathbb{P}(X > m)$)

$$[\sqrt{2\pi}\,\bar{F}(m)]^{-1} e^{-x^2/2} \leq k\theta e^{-\theta(x-m)},$$

for all $x \geq m$, with k as small as possible. If $\theta \leq m$, the above inequality is equivalent, with $k' = k\sqrt{2\pi}\,\bar{F}(m)$, to

$$\exp(-m^2) \leq k'\theta,$$

which suggests that θ should be as large as possible in order to minimize k', hence the choice $\theta = m$. We then proceed as in Exercise 1.5.2.

Exercise 1.5.7

1. Note that

$$\begin{aligned}\mathbb{E}\left(e^{5X}\mathbf{1}_{\{X > 0\}}\right) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{5x - x^2/2} dx \\ &= \frac{e^{25/2}}{\sqrt{2\pi}} \int_0^\infty e^{-(x-5)^2/2} dx\end{aligned}$$

$$= \frac{e^{25/2}}{\sqrt{2\pi}} \int_{-5}^{\infty} e^{-y^2/2} dy$$
$$= e^{25/2} \mathbb{P}(X > -5).$$

In other words, if $N(0, 1)$ is the law of X, the random variables $e^{5X}\mathbf{1}_{\{X > 0\}}$ and $e^{25/2}\mathbf{1}_{\{X+5 > 0\}}$ have the same expectation. The variance of the former is close to that of e^{5X}, that is, of $e^{50} - e^{25/2}$ (in fact it is bigger than or equal to $e^{50} - e^{25/2} - 1$), while the variance of the latter is bounded above by $e^{25}/4$.

2. Since $\mathbb{E}[e^{5X}] = e^{25/2}$,

$$I = e^{25/2} - \mathbb{E}\left(e^{5X}\mathbf{1}_{\{X<0\}}\right).$$

The reduction of the variance here is enormous, since

$$e^{5X}\mathbf{1}_{\{X<0\}} \leq 1, \quad \text{hence } \operatorname{var}\left(e^{5X}\mathbf{1}_{\{X<0\}}\right) \leq 1.$$

3. The antithetic variable method here involves replacing the random variable

$$e^{5X}\mathbf{1}_{\{X<0\}} \quad \text{by} \quad \frac{e^{5X}\mathbf{1}_{\{X<0\}} + e^{-5X}\mathbf{1}_{\{X > 0\}}}{2},$$

which has the same expectation in the Monte Carlo computation. We may note that here we do strictly better than dividing the variance by 2, since

$$\operatorname{cov}\left(e^{5X}\mathbf{1}_{\{X<0\}}, e^{-5X}\mathbf{1}_{\{X > 0\}}\right) = -\left(\mathbb{E}\left[e^{5X}\mathbf{1}_{\{X<0\}}\right]\right)^2 < 0.$$

Exercise 1.5.8

1. In the two cases $X > Y$ and $X < Y$, we have

$$(f(X) - f(Y))(g(X) - g(Y)) \geq 0,$$

hence

$$f(X)g(X) + f(Y)g(Y) \geq f(X)g(Y) + f(Y)g(X),$$

from which the first result follows by taking the expectation. The second result follows if we choose X and Y to be independent with the same law.

2. The identity in question follows from independence of the random variables, with

$$\Phi(x) = \mathbb{E}[f(X_1, \ldots, X_{n-1}, x)g(X_1, \ldots, X_{n-1}, x)].$$

Suppose that the desired result holds for functions of $n-1$ random variables. Then

$$\Phi(x) \geq F_n(x)G_n(x),$$

if

$$F_n(x) = \mathbb{E}f(X_1, \ldots, X_{n-1}, x), \quad G_n(x) = \mathbb{E}g(X_1, \ldots, X_{n-1}, x),$$

and these two functions are increasing, hence from part 1,

$$\mathbb{E}\Phi(X_n) \geq \mathbb{E}[F_n(X_n)G_n(X_n)] \geq \mathbb{E}F_n(X_n)\mathbb{E}G_n(X_n),$$

which is what we wanted to prove for functions of n random variables.

3. The desired inequality follows from part 2 and the fact that if h is increasing (or decreasing) in each of its arguments, the same is true for the function g defined by $g(x_1, \ldots, x_n) = -h(1-x_1, \ldots, 1-x_n)$. But $h(U_1, \ldots, U_n)$ and $h(1-U_1, \ldots, 1-U_n)$ have the same law, hence the inequality just established shows that

$$\operatorname{cov}(h(U_1, \ldots, U_n), h(1-U_1, \ldots, 1-U_n)) \leq 0,$$

hence

$$\mathbb{E}\left[\frac{h(U_1, \ldots, U_n) + h(1-U_1, \ldots, 1-U_n)}{2}\right] = \mathbb{E}h(U_1, \ldots, U_n),$$

$$\operatorname{var}\left[\frac{h(U_1, \ldots, U_n) + h(1-U_1, \ldots, 1-U_n)}{2}\right] \leq \frac{1}{2}\operatorname{var}[h(U_1, \ldots, U_n)].$$

10.2 Chapter 2

Exercise 2.10.2

States 1 and 5 communicate, as well as states 2 and 4. Starting at 4, we can (in addition to going to 2) go to 5 or to 3. The recurrent classes are $\{1, 5\}$ and $\{3\}$ (3 is an absorbing state). The class $\{2, 4\}$ is transient. Starting from this class, we eventually end up in one of the two recurrent classes. The transition matrix of the chain restricted to recurrent class $\{1, 5\}$ is

$$\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix},$$

hence the invariant probability whose support is $\{1, 5\}$ is $(1/2, 0, 0, 0, 1/2)$. The invariant probability whose support is $\{3\}$ is $(0, 0, 1, 0, 0)$. Hence π is invariant by P if and only if there exists $0 \leq p \leq 1$ such that $\pi = (p/2, 0, 1-p, 0, p/2)$. We

would obtain the same result if we looked for probability distributions which are solutions of the equation $\pi P = \pi$.

Exercise 2.10.5

$x \not\to y$ if and only if $(P^n)_{xy} = 0$, for all n, hence if and only if $\sum_{n\geq 0}(P^n)_{xy} = 0$. On the other hand, $x \to y$ and the fact that x is recurrent together imply that $x \leftrightarrow y$. Consequently, there exists $m > 0$ such that $(P^m)_{xy} > 0$. But

$$(P^{m+n})_{xy} \geq (P^n)_{xx}(P^m)_{xy}.$$

The result now follows from Corollary 4.3.

Exercise 2.10.6

1. The greatest common divisor of n and $n+1$ divides their difference, hence it is 1.

2. If $n, n+1 \in N_x$, then $kn + \ell(n+1) \in N_x$, for all $k, \ell \in \mathbb{N}$, since

$$(P^{kn+\ell(n+1)})_{xx} \geq \left[(P^n)_{xx}\right]^k \left[(P^{n+1})_{xx}\right]^\ell.$$

   Any integer $m \geq n^2$ is of the form $m = n^2 + \ell n + k$, with $k < n$. Then from the above property $m = (n + \ell - k)n + k(n+1) \in N_x$.

3. Write $A_n = N_x \cap [0, n]$, and denote by α_n the greatest common divisor of A_n. Then $\alpha_{n+1} \leq \alpha_n$. If the greatest common divisor of N_x is 1, there must exist n_0 such that $\alpha_{n_0} = 1$. Let $a_1, a_2, \ldots, a_k$ denote the distinct elements of A_{n_0}. Iterating the Euclidean algorithm, we show that there exist k elements $u_1, u_2, \ldots, u_k$ of $\mathbb{Z}$ such that

$$a_1u_1 + a_2u_2 + \cdots + a_ku_k = 1,$$

   hence there exist $b_1, \ldots, b_\ell$, $b_{\ell+1}, \ldots, b_k$ in N_x, and positive integers $v_1, \ldots, v_k$ such that

$$b_1v_1 + \ldots + b_\ell v_\ell = b_{\ell+1}v_{\ell+1} + \ldots + b_kv_k + 1.$$

   Since the b_i belong to N_x and the v_i are integers, $b_1v_1 + \ldots + b_\ell v_\ell \in N_x$, and $b_{\ell+1}v_{\ell+1} + \ldots + b_kv_k \in N_x$, from which the result follows.

4. Clearly if N_x contains all integers greater than or equal to a certain z, the greatest common divisor of N_x is 1. We have already shown that, conversely, if the greatest common divisor is 1, then there exists $n \in \mathbb{N}$ such that $n, n+1 \in N_x$, hence N_x contains all integers greater than or equal to n^2.

Exercise 2.10.7

1. We deduce the diagonal entries of P from the fact that the sum over each row is 1. We find successively as diagonal entries 1/2, 2/3, 1/8, 0, 1/4, 2/5.

2. Clearly 1 and 2 communicate, 3 and 5 communicate, as well as 4 and 6. Hence the chain can move from 4 to 1, 2 or 5, and from 6 to 2 or 5. But the reverse is not true. Hence there are three classes, namely $\{1, 2\}$, $\{3, 5\}$ and $\{4, 6\}$.

3. As noted above, the chain can leave the class $\{4, 6\}$, while once it is in the class $\{1, 2\}$ or $\{3, 5\}$, it cannot exit from it. Hence if the chain starts from 4 or 6, its ends up in one of the other two classes sooner or later. The states 4 and 6 are transient, and the two classes $\{1, 2\}$ and $\{3, 5\}$ are recurrent, since once the chain reaches one of those classes, it stays there.

4. If $x \in \mathcal{C}$, then $X_0 \in \mathcal{C}$ almost surely under $\mathbb{P}_x$, hence $T = 0\, \mathbb{P}_x$ almost surely, and $\rho_x = 1$. If $x \in \mathcal{C}'$, under $\mathbb{P}_x$ the chain never reaches $\mathcal{C}$, since it stays in $\mathcal{C}'$, hence $T = \infty\, \mathbb{P}_x$ almost surely, and $\rho_x = 0$. If $x \in \mathcal{T}$, $\mathbb{P}_x(X_1 = 2) > 0$, and $\mathbb{P}_x(X_1 = 5) > 0$, hence, since $\{X_1 = 2\} \subset \{T < \infty\}$ and $\{X_1 = 5\} \subset \{T = \infty\}$, $0 < \rho_x < 1$.

5. Choose $x \in \mathcal{T}$.

$$\begin{aligned}
\mathbb{P}_x(T < \infty) &= \mathbb{P}_x(T = 0) + \mathbb{P}_x(T = 1) + \mathbb{P}_x(2 \leq T < \infty) \\
&= 0 + \mathbb{P}_x(X_1 \in \mathcal{C}) + \sum_{z \in \mathcal{T}} \mathbb{P}_x(X_1 = z, T < \infty) \\
&= \sum_{y \in \mathcal{C}} P_{xy} + \sum_{z \in \mathcal{T}} P_{xz} \rho_z \\
&= \sum_{y \in E} P_{xy} \rho_y,
\end{aligned}$$

   where we have used for the penultimate step

$$\sum_{z \in \mathcal{T}} \mathbb{P}_x(X_1 = z, T < \infty) = \sum_{z \in \mathcal{T}} P_{xz} \mathbb{P}_z(T < \infty),$$

   which follows from the Markov property, and in the last step the values of ρ_y for $y \in \mathcal{C} \cup \mathcal{C}'$.

6. The above equation leads to

$$\begin{aligned}
\rho_4 &= P_{41} + P_{42} + P_{44}\rho_4 + P_{46}\rho_6 \\
\rho_6 &= P_{61} + P_{62} + P_{64}\rho_4 + P_{66}\rho_6,
\end{aligned}$$

or

$$\rho_4 = 1/2 + \rho_6/4, \quad \rho_6 = 1/5 + \rho_4/5 + 2\rho_6/5,$$

from which we deduce $\rho_4 = 7/11$, $\rho_6 = 6/11$.

7. The chain ends up almost surely in one of the two recurrent classes (and does not visit the other one), hence for all $x \in E$, $\mathbb{P}_x(T < \infty) + \mathbb{P}_x(T_{\mathcal{C}'} < \infty) = 1$. Consequently, $\mathbb{P}_4(T_{\mathcal{C}'} < \infty) = 1 - \rho_4 = 4/11$ and $\mathbb{P}_6(T_{\mathcal{C}'} < \infty) = 1 - \rho_6 = 5/11$.

Exercise 2.10.8

1. The diagonal entries are $(5/12, 1/6, 1/2, 1/6, 0, 5/12)$.

2. It is easily verified that $\mathcal{T} = \{1, 2, 3\}$ and $\mathcal{R} = \{4, 5, 6\}$.

3. If $x \in \mathcal{R}$, $T = 0\ \mathbb{P}_x$ almost surely, hence $h_x = 0$. If $x \in \mathcal{T}$, $T \geq 1\ \mathbb{P}_x$ almost surely. Since, moreover, $\mathbb{P}_x(T > 1) > 0$,

$$\mathbb{E}_x(T) \geq \mathbb{P}_x(T = 1) + 2\mathbb{P}_x(T > 1) > 1.$$

To show that $h_x < \infty$, we modify the matrix P, changing the last row into $(5/12, 0, 0, 1/3, 1/4, 0)$, which makes the chain irreducible, without changing the law of T under $\mathbb{P}_x$, hence without changing h_x. We conclude with the help of Corollary 5.5.

4. The identity in question follows from writing that the time needed to reach $\mathcal{R}$, starting from x at time 0, equals 1 plus the time needed to reach $\mathcal{R}$, starting from X_1, hence

$$h_x = 1 + \mathbb{E}\left[\mathbb{E}_{X_1}(T)\right].$$

Taking into account that $h_4 = h_5 = h_6 = 0$, we deduce the linear system

$$h_1 = 1 + \frac{5}{12}h_1 + \frac{1}{4}h_2 + \frac{1}{3}h_3$$
$$h_2 = 1 + \frac{1}{4}h_1 + \frac{1}{6}h_2$$
$$h_3 = 1 + \frac{1}{2}h_1 + \frac{1}{2}h_3.$$

The solution is $h_1 = 236/21$, $h_2 = 32/7$, $h_3 = 278/21$.

Exercise 2.10.9

1. The chain can move from 1 to 2, from 2 to 4, from 4 to 3 and from 3 to 1, hence the chain is irreducible, hence positive recurrent since the state space is finite.

2. We solve the system of equations $\pi = \pi P$, $\sum_x \pi_x = 1$, and we find $\pi = (p^2, p(1-p), p(1-p), (1-p)^2)$.

3. We easily check that $(P^n)_{11} \geq (P_{11})^n = p^n > 0$, for all n. Hence, by Definition 6.1 and Lemma 6.2, the chain is aperiodic. It now follows from Theorem 6.4 that $(P^n)_{ij} \to \pi_j$ as $n \to \infty$, hence the result.

4. The identity $P^2 = \lim_{n\to\infty} P^n$ follows from an elementary computation. We also have $P^n = P^2$ for all $n \geq 2$. The law of X_n is μP^n, if μ denotes the law of X_0. For all $n \geq 2$, $\mu P^n = \pi$ for any probability distribution μ, since all entries of the xth column of P^n are identically equal to π_x. Hence, for all $n \geq 2$, the law of X_n is the invariant probability π. The asymptote is reached here for $n = 2$.

5. From Theorem 5.4, $\mathbb{E}_x(T_x) = 1/\pi_x$, hence $\mathbb{E}_4(T_4) = (1-p)^{-2}$.

Exercise 2.10.10

1. The states 1, 2, 3 and 4 communicate with 0. Consequently, the chain is irreducible, hence recurrent since the state space is finite.

2. Define $E' = \{1, 2, 3, 4\}$. For $k \geq 2$, we have

$$\begin{aligned}\{T = k\} &= \{X_1 \in E', \ldots, X_{k-1} \in E', X_k = 0\},\\ \mathbb{P}_0(T = k) &= \mathbb{P}_0(X_1 \in E', \ldots, X_{k-1} \in E', X_k = 0)\\ &= (1-p)^{k-2}p,\end{aligned}$$

since given that $X_j \in E'$, $X_{j+1} \in E'$ with probability $1-p$, and $X_{j+1} = 0$ with probability p, independently of the values of $X_0, X_1, \ldots, X_{j-1}$. Hence

$$\mathbb{E}_0(T) = \sum_{k=2}^{\infty} k(1-p)^{k-2}p = \frac{p+1}{p}.$$

3. Combining Theorems 5.7 and 5.4, we deduce that $n^{-1}N_n \to p/(p+1)$ as $n \to \infty$. Hence $n^{-1}M_n \to 1/(p+1)$, since $n^{-1}(M_n + N_n) = 1$.

4. The states 1, 2, 3, 4 play symmetric roles in this problem. From Theorem 5.4, $\pi_0 = p/(1+p)$. Then $\pi_1 = \pi_2 = \pi_3 = \pi_4 = [4(1+p)]^{-1}$.

5. Define $\pi' = \pi_\tau$.. We have

$$\pi_x = \sum_y \pi_y P_{y,x} = \sum_z \pi_{\tau z} P_{\tau z, x},$$

hence if $x = \tau u$,

$$\pi'_u = \sum_z \pi'_z P_{\tau z, \tau u} = \sum_z \pi'_z P_{z,u}.$$

Consequently, π' is invariant, and from uniqueness of the invariant probability distribution, $\pi' = \pi$.

Returning to our problem, we note that for any permutation τ of the points 1, 2, 3, 4, $P_{\tau x, \tau y} = P_{xy}$. The identity $\pi_1 = \pi_2 = \pi_3 = \pi_4$ follows.

Exercise 2.10.12

1. Clearly

$$\mathbb{E}(\exp[itY_1]) = \frac{1}{d}\sum_{j=1}^{d} \frac{e^{it_j} + e^{-it_j}}{2},$$

hence the formula for $\phi(t)$. Note that

$$(P^n)_{00} = \mathbb{P}(Y_1 + \cdots + Y_n = 0).$$

For X a $\mathbb{Z}$-valued random variable,

$$\mathbb{E}(e^{itX}) = \sum_{x \in \mathbb{Z}} (\cos(tx) + i\sin(tx))\mathbb{P}(X = x).$$

From the periodicity of the functions cos and sin, for all $x \in \mathbb{Z}\backslash\{0\}$,

$$\int_{-\pi}^{\pi} (\cos tx + i \sin tx)dt = 0,$$

hence

$$\mathbb{P}(X = 0) = (2\pi)^{-1} \int_{-\pi}^{\pi} \phi_X(t)dt.$$

If $t \to \phi^n(t)$ denotes the characteristic function of the $\mathbb{Z}^d$-valued random variable $Y_1 + \cdots + Y_n$, we have by a similar argument that

$$\mathbb{P}(Y_1 + \cdots + Y_n = 0) = (2\pi)^{-d} \int_{[-\pi,\pi]^d} \phi^n(t)dt.$$

2. The formula follows from the last result and the fact that for all $0 \le r < 1$, $|r\phi(t)| < 1$, and consequently

$$\sum_{n \ge 0} r^n \phi^n(t) = (1 - r\phi(t))^{-1}.$$

3. The first claim follows from the fact that $\|t\| \geq \alpha > 0 \Rightarrow \phi(t) \leq 1 - \rho(\alpha)$, with $\rho(\alpha) > 0$. For $\|t\|$ small enough, $-\pi/2 < t_j < \pi/2$, $1 \leq j \leq d$, and then $0 < \phi(t) \leq 1$, hence the second claim.

4. Clearly $\sum_{n\geq 0}(P^n)_{00} = \lim_{r\uparrow 1}\sum_{n\geq 0} r^n (P^n)_{00}$. But for $0 < r < 1$,

$$(2\pi)^d \sum_{n\geq 0} r^n (P^n)_{00} = \int_{[-\pi,\pi]^d \setminus C_\alpha} (1 - r\phi(t))^{-1} dt + \int_{C_\alpha} (1 - r\phi(t))^{-1} dt.$$

The first term is bounded for $0 < r \leq 1$. Hence the convergence or divergence of the series at $r = 1$ depends upon the value of the limit as $r \to 1$ of the second integral, which by the monotone convergence theorem equals

$$\begin{aligned} \int_{C_\alpha} \frac{dt}{1 - \phi(t)} &\simeq 2 \int_{C_\alpha} \frac{dt}{\|t\|^2} \\ &= c(d) \int_0^\alpha r^{d-3} dr, \end{aligned}$$

which is finite if and only if $d \geq 3$, hence the result by Corollary 4.3.

Exercise 2.10.14

1. If $X_n = 0$, $X_{n+1} = 1$, and if $X_n > 0$, $\mathbb{P}(X_{n+1} = X_n + 1) = p$, $\mathbb{P}(X_{n+1} = X_n - 1) = 1 - p$, hence $P_{01} = 1$, and for $x \in \mathbb{N}\setminus\{0\}$, $P_{x,x+1} = p$, $P_{x,x-1} = 1 - p$. Consequently, if $x \neq y$, $(P^{|x-y|})_{xy} > 0$, which implies irreducibility.

2. Since $Y_{n+1} \leq 1$, clearly

$$Y_{n+1} \leq \mathbf{1}_{\{X_n > 0\}} Y_{n+1} + \mathbf{1}_{\{X_n = 0\}},$$

from which we deduce that $X_n \geq X'_n$ by a recurrence on n. If $p > 1/2$, $X'_n \to \infty$ as $n \to \infty$, hence also $X_n \to \infty$, which establishes transience of the chain.

3. Provided $n < T$, for all $0 \leq k \leq n$, $X_{k+1} = X_k + Y_{k+1}$, whence $X_{n+1} = X'_{n+1}$, as claimed in the statement of the question. Suppose that $X_0 = 1$. Either $X_1 = 0$, in which case $X_2 = 1$, and the chain returns to 1. Otherwise $X_1 = 2$, and we have that $X_n = X'_n$ if $n \leq T$. It follows that $T = \inf\{n \geq 0;\, X'_n = 0\}$. Now either $\{X'_n\}$ is recurrent (if $p = 1/2$), or it tends to $-\infty$ as $n \to \infty$. In either case, starting at 2, $\{X'_n\}$ visits 1 in finite time almost surely, and since it must visit 1 before reaching 0, $\{X_n\}$ also visits 1 in finite time, hence the state 1 is recurrent, and the same is true for all states by irreducibility.

4. In the first case, the measure (with infinite mass) $\nu = (1/2, 1, 1, 1, \ldots)$ is invariant, since

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \ldots \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 & \ldots \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & \ldots \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix},$$

hence $\nu P = \nu$. In the case $p < 1/2$, we have that

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \ldots \\ 1-p & 0 & p & 0 & 0 & 0 & \ldots \\ 0 & 1-p & 0 & p & 0 & 0 & \ldots \\ 0 & 0 & 1-p & 0 & p & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix},$$

hence for μ as specified,

$$\begin{aligned} (1-p)\mu_1 &= \mu_0 \\ \mu_0 + (1-p)\mu_2 &= \mu_1 \\ p\mu_{x-1} + (1-p)\mu_{x+1} &= \mu_x, \quad x \geq 2, \end{aligned}$$

that is, $\mu P = \mu$. But $p < 1/2$ implies that

$$\frac{p}{1-p} < 1,$$

and we conclude that $\sum_{x \geq 0} \mu_x < \infty$.

Exercise 2.10.16

1. With $\{Y_n;\ n \geq 1\}$ and $\{Z_n;\ n \geq 1\}$ mutually independent random variables with

$$\mathbb{P}(Y_n = 1) = p = 1 - \mathbb{P}(Y_n = 0), \quad \mathbb{P}(Z_n = 1) = q = 1 - \mathbb{P}(Z_n = 0),$$

define the sequence $\{X_n;\ n \geq 0\}$ by

$$X_{n+1} = X_n + Y_{n+1} - Z_{n+1}\mathbf{1}_{\{X_n\ >\ 0\}}.$$

It follows from Lemma 1.2 that $\{X_n;\ n \geq 1\}$ is a Markov chain. Its transition matrix is given by

$$P = \begin{pmatrix} 1-p & p & 0 & 0 & 0 & 0 & \cdots \\ q(1-p) & \alpha_{p,q} & p(1-q) & 0 & 0 & 0 & \cdots \\ 0 & q(1-p) & \alpha_{p,q} & p(1-q) & 0 & 0 & \cdots \\ 0 & 0 & q(1-p) & \alpha_{p,q} & p(1-q) & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix},$$

with $\alpha_{p,q} = pq + (1-p)(1-q)$. Irreducibility follows from the fact that the entries $P_{x,x+1}$, $x \geq 0$, and $P_{x,x-1}$, $x \geq 1$, are strictly positive.

2. If $p > q$, $\mathbb{E}Y_n > \mathbb{E}Z_n$, and $X_n \to \infty$, as $n \to \infty$. If $p = q$, the measure $(1-p, 1, 1, 1, 1, \ldots)$ (with infinite mass) is invariant. $\{X_n;\ n \geq 1\}$ is positive recurrent if and only if $p < q$, since in that case the following probability distribution π is P-invariant:

$$\pi_0 = \frac{q-p}{q}, \quad \pi_x = \left(\frac{p(1-q)}{q(1-p)}\right)^x \frac{q-p}{q(1-q)},\ x \geq 1.$$

3. We have

$$\mathbb{E}_\pi(X_n) = \sum_{x \geq 0} x\pi_x = \frac{p(1-p)}{q-p}.$$

4. A customer who arrives at the arbitrary time n finds in front of him a random number of customers, whose probability distribution is π. The service times being i.i.d. geometric with expectation $1/q$, and independent of the number of customers which the newly arriving customer finds in front of him, $\mathbb{E}(T)$ equals

$$\frac{1 + \mathbb{E}_\pi(X_n)}{q} = \frac{q - p^2}{q(q-p)}.$$

Exercise 2.10.18

1. First compute $\mathbb{E}[u^Y | X = x]$, which equals the generating function of the binomial probability distribution $B(x, q)$, hence

$$\mathbb{E}[u^Y | X = x] = (qu + 1 - q)^x,$$

consequently

$$\begin{aligned}
\psi(u) &= \mathbb{E}[u^Y] \\
&= \sum_{x=0}^{\infty} \mathbb{E}[u^Y | X = x]\mathbb{P}(X = x) \\
&= \mathbb{E}\left[(qu + 1 - q)^X\right] \\
&= \phi(qu + 1 - q).
\end{aligned}$$

2. One way to construct the chain $\{X_n;\ n \in \mathbb{N}\}$ is as follows. Let $\{Z_{n,k};\ n, k \in \mathbb{N}^*\}$ be i.i.d. Bernoulli random variables such that $\mathbb{P}(Z_{n,k} = 1) =$

$1 - \mathbb{P}(Z_{n,k} = 0) = p$. Define

$$X'_n = \sum_{k=1}^{X_n} Z_{n+1,k}$$

and

$$X_{n+1} = X'_n + Y_{n+1}.$$

It is easy to see that one can apply Lemma 1.2 to show that $\{X_n;\ n \in \mathbb{N}\}$ is a Markov chain. Note, however, that the role of the Y_n in this lemma is played by the $(Y_n, Z_{n,\cdot})$, which take values in $\{0, 1\}^{\mathbb{N}}$. This set is not countable, but continuous (Lemma 1.2 is of course still valid in this case; see the analogous Lemma 1.3). This chain is irreducible, its transition matrix P having the property that $P_{x,y}>0$ for all $x, y \in \mathbb{N}$, since

$$P_{x,y} > \mathbb{P}(X'_n = 0 | X_n = x)\mathbb{P}(Y_{n+1} = y) > 0.$$

3. The second equality below follows from independence between Y_{n+1} and the pair (X_n, X'_n):

$$\begin{aligned}\mathbb{E}\left[u^{X_{n+1}} | X_n = x\right] &= \mathbb{E}\left[u^{X'_n + Y_{n+1}} | X_n = x\right] \\ &= \mathbb{E}\left[u^{Y_{n+1}}\right] \mathbb{E}\left[u^{X'_n} | X_n = x\right] \\ &= e^{\theta(u-1)}(pu + 1 - p)^x,\end{aligned}$$

4. It follows from the last computation that

$$\begin{aligned}\phi_{n+1}(u) &= \mathbb{E}\left[e^{\theta(u-1)}(pu + 1 - p)^{X_n}\right] \\ &= e^{\theta(u-1)}\phi_n(pu + 1 - p),\end{aligned}$$

from which the desired formula follows by recurrence.

5. As $n \to \infty$, $1 - p^n + p^n u \to 1$. Moreover, ϕ_0 is continuous and $\phi_0(1) = 1$, $\sum_0^{n-1} p^k \to (1 - p)^{-1}$,

$$\phi_n(u) \to \rho(u) = \exp[\theta(1 - p)^{-1}(u - 1)],$$

which is the generating function of the Poisson probability distribution with parameter $\theta(1 - p)^{-1}$.

6. It follows from the computations above that if ρ is the generating function of the distribution of X_n, then it is also that of the distribution of X_{n+1}. Hence the Poisson distribution with parameter $\theta(1 - p)^{-1}$ is an invariant probability distribution of the Markov chain $\{X_n;\ n \in \mathbb{N}\}$. Since this chain is irreducible, it follows from Theorem 5.4 that it is positive recurrent.

Exercise 2.10.19

1. If we define $Y_{n+1} = A_n - D_n$ and $f(x, y) = (x - y)^+$, the Markov property follows from Lemma 1.2. We have

$$\begin{cases} P_{0,\,0} &= \mathbb{P}(A_n - D_n \le 0) = r_0 + qr_1 \\ P_{x,\,x+k} &= qr_{k+1} + (1-q)r_k, \quad k \in \mathbb{N}, \\ P_{x,\,x-1} &= q\, r_0, \quad x \in \mathbb{N}^* \\ P_{x,\,x-\ell} &= 0, \quad \ell \ge 2, \ x \ge \ell. \end{cases}$$

   Since $r_0 < 1$, there exists $k > 0$ such that $r_k > 0$. Moreover, $q < 1$. Therefore $P_{x,x+k} > 0$ and, moreover, $P_{x,x-1} > 0$, hence starting from x, the chain can reach $x + nk - \ell$ in $n + \ell$ steps, $n,\ \ell \in \mathbb{N}^*$. But for all $x \ne y$, there exists $n,\ \ell \in \mathbb{N}$ such that $y = x + nk - \ell$.

2. We have

$$X_{n+1} \ge X_n + A_n - D_n.$$

   Hence if $X_n \ge S_n$, $X_{n+1} \ge S_{n+1}$. But $X_0 = S_0 = 0$. The inequality follows by recurrence. If $T > n$, $X_k > 0$, $1 \le k \le n$, then

$$X_{k+1} = X_k + A_k - D_k, \ \ 0 \le k < n,$$

   consequently $X_n = S_n > 0$ and $X_n \ge 1$. Since $A_n - D_n \ge -1$,

$$X_n + A_n - D_n \ge 0$$

$$X_{n+1} = S_{n+1}.$$

3. From the definition of S_n,

$$\begin{aligned} \frac{S_n}{n} &= \frac{1}{n}\sum_0^{n-1} A_k - \frac{1}{n}\sum_0^{n-1} D_k \\ &\to \mathbb{E}\, A_0 - \mathbb{E}\, D_0 = p - q \end{aligned}$$

   almost surely as $n \to \infty$, by the law of large numbers.

4. If $p < q$, $S_n/n \to p - q < 0$, hence $S_n \to -\infty$. Consequently, $T < \infty$ almost surely, since on the set $\{T = +\infty\}$, $X_n = S_n$, for all n, and $X_n \ge 0$.

5. If $p > q$, $S_n \to +\infty$. Since $X_n \ge S_n$, $X_n \to \infty$ almost surely, from which the result follows.

6. In the case $p > q$, we have just seen that the state 0 is transient (indeed, a recurrent state is visited infinitely often almost surely), hence the chain is transient. In the case $p < q$, $P(T < \infty) = 1$, which says exactly that 0 is recurrent, hence the chain is recurrent.

7. We have, writing $a = (1-q)(1-p)$,

$$P = \begin{pmatrix} 1-p+qp & (1-q)p & 0 & 0 & 0 \\ q(1-p) & qp+a & (1-q)p & 0 & 0 \\ 0 & q(1-p) & qp+a & (1-q)p & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

8. In the case $p = q$, as long as it stays in $\mathbb{N}^*$, the chain behaves like a symmetric random walk, except that $\mathbb{P}(X_{n+1} = X_n) > 0$. If we define $T_1 = \inf\{n \geq 0;\, X_n \neq X_0\}$, $T_2 = \inf\{n \geq T_1;\, X_n \neq X_{T_1}\}, \ldots$, etc., then $Z_n = X_{T_n}$, $n \in N$, is a Markov chain with transition matrix

$$P' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

On $\mathbb{N}^*$, $\{Z_n\}$ behaves like the symmetric random walk in Exercises 2.10.11 and 2.10.13. It is easy to deduce that the state 1 is recurrent. The same is true for the original chain $\{X_n\}$.

If $\pi = (1, 1, 1, \ldots)$, it is easily checked that $\pi P = \pi$, hence there exists an invariant measure with infinite total mass, and $\{X_n\}$ is null recurrent.

9. The chain $\{X_n\}$ being irreducible, it admits at most one invariant probability measure. We note that

$$\frac{q(1-p)}{p(1-q)}(1-q)p + qp + (1-p)(1-q) + q(1-p)\frac{p(1-q)}{q(1-p)} = 1,$$

hence

$$(\pi\ P)_x = \pi_x, \quad x \geq 1.$$

We note that we also have $(\pi\ P)_0 = \pi_0$. Hence $\pi = \{(1-a)a^x;\ x \in \mathbb{N}\}$ is an invariant probability distribution, and the chain is positive recurrent.

Exercise 2.10.22

1. It suffices to note that for many pairs $(k, \ell) \in E^2$, $Q_{k\ell} = 0 < Q_{\ell k}$.

2. If a chain is irreducible, recurrent, aperiodic and satisfies (ii), then summing over $k_1, k_2, \ldots, k_{m-1}$ and writing $\ell = k_m$, we obtain

$$(P^m)_{k\ell} P_{\ell k} = P_{k\ell} (P^m)_{\ell k}.$$

It remains to let $m \to \infty$ in order to deduce the detailed balance equation.

3. P satisfies (i). Let us check that it satisfies (ii) in the theorem. Consider an excursion. Each element $i \in (1, \ldots, n)$ experiences f_i moves ahead, and stays in front r_i times while no other point moves. The product of the transition probabilities along the excursion in the forward direction is

$$\prod_{1 \le i \le n} p_i^{f_i + r_i},$$

and the product of the transition probabilities in the backward direction has the same value.

4. We associate with each permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$ the quantities

$$I_{k\ell} = \begin{cases} 1, & \text{if } i_\ell < i_k, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\text{position of } k = 1 + \sum_{\ell \ne k} I_{k\ell}.$$

Hence

$$\begin{aligned} J_Q &= \sum_{1 \le k \le n} p_k \pi(\text{position of } k) \\ &= \sum_{1 \le k \le n} p_k \left[1 + \sum_{\ell \ne k} \pi(\ell \text{ is in front of } k) \right] \\ &= 1 + \sum_k \sum_{\ell \ne k} p_k \pi(\ell \text{ is in front of } k) \\ &= 1 + \sum_{k < \ell} \left[p_k \pi(\ell \text{ is in front of } k) + p_\ell \pi(k \text{ is in front of } \ell) \right] \\ &= 1 + \sum_{k < \ell} (p_k - p_\ell) \pi(\ell \text{ is in front of } k) + \sum_{k < \ell} p_\ell. \end{aligned}$$

Under π, the probability that ℓ is in front of k is the probability that ℓ has been called more recently than k. Hence

$$\pi(\ell \text{ is in front of } k) = \frac{p_\ell}{p_\ell + p_k}.$$

It remains to show that if $p_\ell > p_k$, then

$$\mu(\ell \text{ is in front of } k) > \frac{p_\ell}{p_\ell + p_k}.$$

We now exploit the reversibility of the chain with the transition matrix P. We have

$$p_{i_{j+1}} \mu(i_1, \ldots, i_j, i_{j+1}, \ldots, i_n) = p_{i_j} \mu(i_1, \ldots, i_{j+1}, i_j, \ldots, i_n).$$

Iterating this formula, we deduce that

$$\mu(\ldots, k, i_1, \ldots, i_j, \ell, \ldots) = \left(\frac{p_k}{p_\ell}\right)^{j+1} \mu(\ldots, \ell, i_1, \ldots, i_j, k, \ldots).$$

Hence if $p_\ell > p_k$, then

$$\mu(\ldots, k, i_1, \ldots, i_j, \ell, \ldots) < \frac{p_k}{p_\ell}\mu(\ldots, \ell, i_1, \ldots, i_j, k, \ldots).$$

Write $\alpha(k, \ell) \stackrel{\text{def}}{=} \mu(k \text{ is in front of } \ell)$. Summing over all permutations satisfying 'k is in front of ℓ', and using the above relation, one can check that $\alpha(k, \ell) < (p_k/p_\ell)\alpha(\ell, k)$. Since $\alpha(k, \ell) + \alpha(\ell, k) = 1$,

$$\alpha(\ell, k) > \frac{p_\ell}{p_\ell + p_k}.$$

10.3 Chapter 3

Exercise 3.5.1

1. We have $X_{n+1} = f(X_n, U_{n+1}, Y_{n+1})$ with

$$f(x, u, y) = \begin{cases} y, & \text{if } u \le \frac{p(y)}{cq(y)}, \\ x, & \text{otherwise.} \end{cases}$$

The fact that $\{X_n;\ n \ge 0\}$ is a Markov chain is easy to check, combining Lemmas 1.2 and 1.3 of Chapter 2.

2. We have

$$P_{xy} = c^{-1}p(y) + \delta_{xy}\left(1 - c^{-1}\right),$$

where δ_{xy} is the Kronecker delta, equal to 0 if $x \neq y$, and to 1 if $x = y$.

3. We have

$$(\mu P)_x = c^{-1}p(x) + (1 - c^{-1})\mu(x).$$

Hence p is P-invariant and

$$((\mu - p)P)_x = (1 - c^{-1})(\mu(x) - p(x)),$$
$$\left((\mu - p)P^n\right)_x = (1 - c^{-1})^n(\mu(x) - p(x)).$$

But $pP^n = p$, hence the above computations show that $\mu P^n \to p$ as $n \to \infty$, since $0 < (1 - c^{-1}) < 1$. Moreover, if $\bar{p}$ were another invariant probability, we would have $\bar{p} - p = (1 - c^{-1})(\bar{p} - p)$, hence $\bar{p} = p$.

4. This chain makes the same random drawings as the classical rejection method (see Proposition 3.2 of Chapter 1). If the indices k and ℓ correspond to two successive acceptances, the classical algorithm rejects the drawn values for $k < n < \ell$, and the chain $\{X_n\}$ is constant and equals X_k for $k \leq n < \ell$.

Exercise 3.5.2

1. We have $P_{xy} = \alpha c_y + Q_{xy}$, with $Q_{xy} \geq 0$, $\sum_y Q_{xy} = 1 - \alpha$. If ν is as defined in the exercise,

$$\sum_x \nu_x P_{xy} = \sum_x \nu_x Q_{xy},$$

hence

$$\begin{aligned} |\nu P| &= \sum_y |\sum_x \nu_x Q_{xy}| \\ &\leq \sum_{x,y} |\nu_x| Q_{xy} \\ &= (1-\alpha)|\nu|. \end{aligned}$$

The second claim follows if we choose $\nu = \mu - \mu'$.

2. The previous result implies that if $\mu P = \mu$ and $\mu' P = \mu'$, then $\mu = \mu'$ – hence uniqueness of the invariant probability distribution. Iterating the equalities of part 1, one can show that if μ and μ' are two probability distributions,

$$|\mu P^n - \mu' P^n| \leq (1-\alpha)^n |\mu - \mu'| \leq 2(1-\alpha)^n,$$

hence in particular, for $n, m \geq 1$,

$$|\mu P^{n+m} - \mu P^n| \leq 2(1-\alpha)^n,$$

hence this is a Cauchy sequence.

3. Since the sequence of probability distributions $\{\mu P^n;\ n \geq 1\}$ is Cauchy in the Banach space $\ell^1(E)$, it converges to a probability distribution π. It follows that as $n \to \infty$, the sequence μP^{n+1} converges both to π and to πP. Hence $\pi = \pi P$. Moreover, since $\pi P^n = \pi$, it follows from an inequality established in part 2 that

$$|\mu P^n - \pi| \leq 2(1-\alpha)^n.$$

4. In this case, we still have the Cauchy property, since for $k, n, m \geq 1$,

$$|\mu P^{k\ell+n} - \mu P^{k\ell+m}| \leq 2(1-\alpha)^k.$$

The limit π is again invariant, and this time

$$|\mu P^n - \pi| \leq 2(1-\alpha)^{[n/\ell]}.$$

5. This transition matrix is given by

$$\tilde{P}_{xy} = \begin{cases} P_{xy} \exp\left[-\beta(H(y) - H(x))_+\right], & \text{if } y \neq x, \\ 1 - \sum_{y \neq x} \tilde{P}_{xy}, & \text{if } y = x. \end{cases}$$

6. We have

$$\tilde{P}_{xy} \geq \alpha c_y e^{-\beta(H(y)-H(x))_+} \geq \alpha c_y e^{-\beta H(y)} \geq \beta c'_y,$$

where

$$c'_y = \frac{c_y e^{-\beta H(y)}}{\sum_z c_z e^{-\beta H(z)}}, \quad \beta = \alpha \sum_z c_z e^{-\beta H(z)} < \alpha.$$

Hence $\tilde{P}$ satisfies (3.2) with α replaced by β, c by c'. If we use Hastings' algorithm for simulating a Markov chain with the transition matrix $\tilde{P}$, we simulate a Markov chain whose law at time n converges at exponential rate towards the probability μ. Note that this exercise considers a case which is quite unlike those which are encountered in practice. The above assumption on P contradicts the fact that the cardinality of E is very large (the only case where the MCMC method is useful), and the transitions under P are easy to simulate! Indeed, these constraints imply that $P_{xy} = 0$ except for a small proportion of pairs (x, y).

10.4 Chapter 4

Exercise 4.8.1

Write $\mu = (0.3, 0.3, 0.1, 0.3)$ and $\pi = (0.2, 0.2, 0.4, 0.2)$. If $x_n \neq \mathrm{C}$,

$$\delta_y(n+1) = \frac{\delta_0(n) P_{0y} \vee \delta_1(n) P_{1y}}{4},$$

while if $x_n = \mathrm{C}$,

$$\delta_0(n+1) = (\delta_0(n) P_{00} \vee \delta_1(n) P_{10}) \mu_{x_{n+1}},$$
$$\delta_1(n+1) = (\delta_0(n) P_{01} \vee \delta_1(n) P_{11}) \pi_{x_{n+1}}.$$

If the ratio between the two coordinates of $\delta(n_1)$ is as given in the question, the successive maximizations lead to the choice of the diagonal entry of the matrix P, hence, given that $x_n \neq \mathrm{C}$ for $n_1 \leq n < n_2$,

$$\delta_0(n_2) = \delta_0(n_1)\left(\frac{0.9}{4}\right)^{n_2-n_1}, \quad \delta_1(n_2) = \delta_1(n_1)\left(\frac{0.9}{4}\right)^{n_2-n_1}.$$

10.5 Chapter 5

Exercise 5.7.2

1. Following step by step the proof of Theorem 5.1, one can check that the formulae for the Kalman filter are modified as follows:

$$\begin{aligned}
\hat{X}_{n+1} &= A_{n+1}\hat{X}_n + \Sigma_n H_{n+1}^{\star}(H_{n+1}\Sigma_n H_{n+1}^{\star} + R)^{-1}(Y_{n+1} - H_{n+1}A_{n+1}\hat{X}_n),\\
\Sigma_n &= A_{n+1}\Lambda_n A_{n+1}^{\star} + Q,\\
\Lambda_{n+1} &= \Sigma_n - \Sigma_n H_{n+1}^{\star}(H_{n+1}\Sigma_n H_{n+1}^{\star} + R)^{-1} H_{n+1}\Sigma_n,\\
\hat{X}_0 &= \bar{X}_0, \quad \Lambda_0 = P_0.
\end{aligned}$$

2. To simplify the notation, let us just show that the conditional law of X_2, given (Y_1, Y_2), is a Gaussian distribution, and for the sake of further simplification, we assume that $d = 1$ (the computation of the conditional law of X_1 given Y_1 poses no new difficulty).

   Note that given two random variables X and Y, if we denote the law of Y by μ_Y, the conditional probability of the event $X < x$ given that $Y = y$ is defined for μ_Y almost all y as $\varphi(y)$, where φ is the (class of μ_Y almost everywhere equal) function(s) from $\mathbb{R}$ into $[0, 1]$, which are such that $\varphi(Y) = \mathbb{P}(X < x|Y)$ almost surely. In the case where the pair (X, Y) is Gaussian, Proposition 5.2 allows us to define the conditional probability distribution of X, given that $Y = y$, which is the Gaussian law $N(\hat{X}_y, \hat{\Sigma})$ where in formula (i) of that Proposition we replace Y by y.

   Consider now the triple (X_2, Y_1, Y_2), which is not Gaussian. Note that

$$\begin{aligned}
X_1 &= AX_0 + \eta_1, \quad & X_2 &= A(Y_1)X_1 + \eta_2,\\
Y_1 &= HX_1 + \xi_1, \quad & Y_2 &= H(Y_1)X_2 + \xi_2,
\end{aligned}$$

   hence

$$\begin{aligned}
X_2 &= A(Y_1)AX_0 + A\eta_1 + \eta_2,\\
Y_1 &= H(AX_0 + \eta_1) + \xi_1,\\
Y_2 &= H(Y_1)(A(Y_1)[AX_0 + \eta_1] + \eta_2) + \xi_2.
\end{aligned}$$

Thus the conditional probability of the event $\{X_2 < x\}$, given that $(Y_1, Y_2) = (y_1, y_2)$, is the conditional probability of the event $\{A(y_1)AX_0 + A\eta_1 + \eta_2 < x\}$, given that

$$H(AX_0 + \eta_1) + \xi_1 = y_1, \quad H(y_1)(A(y_1)[AX_0 + \eta_1] + \eta_2) + \xi_2 = y_2.$$

But if we write

$$\begin{aligned} U &= A(y_1)AX_0 + A\eta_1 + \eta_2, \\ V &= H(AX_0 + \eta_1) + \xi_1, \\ W &= H(y_1)(A(y_1)[AX_0 + \eta_1] + \eta_2) + \xi_2, \end{aligned}$$

clearly the triple (U, V, W) is Gaussian, hence the conditional law of U, given that $(V, W) = (y_1, y_2)$, is a Gaussian law whose expectation and variance can be expressed in terms of (y_1, y_2). Finally, the conditional law of X_2, given (Y_1, Y_2), is a Gaussian law whose parameters are obtained by replacing (y_1, y_2) by (Y_1, Y_2) in the formulae. A rigorous justification of the above argument is not quite immediate. Since the joint law of the Y_n has a density, one can at least easily convince oneself that the argument can be justified when the functions A and H of the Y_n are constant on cubes of $\mathbb{R}^d$. The general result then follows by approximation.

It is now easy to check that the Kalman filter formulae in the 'conditionally Gaussian' case are given by the formulae of part 1, where A_n is replaced by $A(Y_1, \ldots, Y_{n-1})$ and H_n by $H(Y_1, \ldots, Y_{n-1})$.

10.6 Chapter 6

Exercise 6.5.2

We first identify the law of the pair $(X \wedge Y, X \vee Y - X \wedge Y)$. For all $s, t > 0$,

$$\begin{aligned} &\mathbb{P}(X \wedge Y > s, X \vee Y - X \wedge Y > t) \\ &\quad = \mathbb{P}(X \wedge Y > s, X \vee Y - X \wedge Y > t, X < Y) \\ &\quad\quad + \mathbb{P}(X \wedge Y > s, X \vee Y - X \wedge Y > t, X > Y) \\ &\quad = 2\mathbb{P}(X \wedge Y > s, X \vee Y - X \wedge Y > t, X < Y) \\ &\quad = 2\mathbb{P}(X > s, Y > X + t) \\ &\quad = 2\int_s^\infty \int_{u+t}^\infty \lambda e^{-\lambda u} \lambda e^{-\lambda v} dv du \end{aligned}$$

$$= 2\int_s^\infty \lambda e^{-\lambda u} e^{-\lambda(u+t)} du$$
$$= e^{-2\lambda s} e^{-\lambda t}.$$

We have just shown that the two random variables $U = X \wedge Y$ and $V = X \vee Y - X \wedge Y$ are independent, the law of U being exponential with parameter 2λ, and that of V exponential with parameter λ.

1. Note that

$$\{C \text{ exits first}\} = \{U + Z > U + V\} = \{Z > V\}.$$

But Z and V are independent (since Z and (X, Y) are), and they both have the same density. Consequently, $\mathbb{P}(Z > V) = \mathbb{P}(V > Z)$ and $\mathbb{P}(Z = V) = 0$. Hence $\mathbb{P}(Z > V) = 1/2$.

2. The total time spent by C at the post office equals $U + Z$, which has the same law as $U + V = X \vee Y$ (both are the sum of two independent random variables, one having the exponential law with parameter λ, the other the exponential law with parameter 2λ). We also have

$$\mathbb{P}(X \vee Y \leq t) = \mathbb{P}(X \leq t, Y \leq t) = (\mathbb{P}(X \leq t))^2,$$

hence the distribution function of the desired probability law is the function $(1 - e^{-\lambda t})^2$, and its density equals $2\lambda(1 - e^{-\lambda t})e^{-\lambda t}$.

3. The probability distribution of the time of the last departure is that of $U + (V \vee Z)$, where the three random variables U, V, and Z are globally independent (since U and V are independent, as well as (U, V) and Z). Its distribution function equals

$$\mathbb{P}(U + V \vee Z \leq t) = \int_0^t \mathbb{P}(V \vee Z \leq t - s)\lambda e^{-\lambda s} ds$$
$$= \lambda \int_0^t \left(1 - e^{-\lambda(t-s)}\right)^2 e^{-\lambda s} ds$$
$$= 1 - 2\lambda t e^{-\lambda t} - e^{-2\lambda t},$$

and its probability density is the function

$$t \to 2\lambda t e^{-\lambda t}(e^{-\lambda t} - 1 + \lambda t).$$

Exercise 6.5.4

1. Denote by π the random permutation of $\{1, \ldots, n\}$ which is such that $Y_k = X_{\pi(k)}$, $1 \leq k \leq n$. By symmetry, the probability of this permutation

is the uniform distribution on the set Π (whose cardinality equals $n!$) of all the permutations of the set $\{1, \ldots, n\}$. Hence, for any $f : [0, t]^n \to \mathbb{R}_+$, with the notation

$$\mathbb{E}[X; A] = \mathbb{E}[X\mathbf{1}_A],$$

we have

$$\begin{aligned}\mathbb{E}[f(Y_1, \ldots, Y_n)] &= \sum_{\mu \in \Pi} \mathbb{E}[f(X_{\mu(1)}, \ldots, X_{\mu(n)}); \pi = \mu] \\ &= n!\mathbb{E}[f(X_1, \ldots, X_n); X_1 < X_2 < \ldots < X_n] \\ &= \frac{n!}{t^n} \int_{\Delta_t^n} ds_1 \cdots ds_n,\end{aligned}$$

where Δ_t^n denotes the simplex

$$\{(s_1, s_2, \ldots, s_n);\ 0 < s_1 < s_2 < \ldots < s_n < t\}.$$

The law of the random vector $(Y_1, \ldots, Y_n)$ is uniform on Δ_t^n.

2. Compute, for $0 < t_1 < t_2 < \ldots < t_n < t$,

$$\begin{aligned}&\mathbb{P}(T_1 < t_1 < T_2 < t_2 < \ldots < T_n < t_n | N_t = n) \\ &= \frac{\mathbb{P}(T_1 < t_1 < T_2 < t_2 < \ldots < T_n < t_n < t < T_{n+1})}{\mathbb{P}(T_n \leq t < T_{n+1})} \\ &= \frac{\mathbb{P}(N_{t_1} = 1, N_{t_2} - N_{t_1} = 1, \ldots, N_{t_n} - N_{t_{n-1}} = 1, N_t - N_{t_n} = 0)}{\mathbb{P}(N_t = n)} \\ &= \frac{\lambda t_1 e^{-\lambda t_1} \lambda(t_2 - t_1) e^{-\lambda(t_2 - t_1)} \cdots \lambda(t_n - t_{n-1}) e^{-\lambda(t_n - t_{n-1})} e^{-\lambda(t - t_n)}}{(\lambda t)^n e^{-\lambda t}/n!} \\ &= \frac{n!}{t^n} t_1(t_2 - t_1) \cdots (t_n - t_{n-1}) \\ &= \frac{n!}{t^n} \int_{0 < s_1 < t_1 < s_2 < t_2 < \ldots < s_n < t_n} ds_1 \cdots ds_n,\end{aligned}$$

hence the result, by comparison with the law found in part 1.

Exercise 6.5.6

It follows from Exercise 6.5.4 that, conditionally upon $N_t = n$, the joint law of the infection times is the law of a set of n i.i.d. random variables, whose common law is uniform on $[0, t]$. An individual who is infected at time s has probability $G(t - s)$ of being sick at time t, and probability $\bar{G}(t) = 1 - G(t)$ of not yet being sick at time t. Hence each of the infected individuals has probability p_t of being

sick at time t, and q_t of not being sick at time t, with

$$p_t = \frac{1}{t}\int_0^t G(t-s)ds = \frac{1}{t}\int_0^t G(s)ds,\ \ q_t = \frac{1}{t}\int_0^t \bar{G}(s)ds.$$

By independence of the behaviours of the various individuals, if $n = k + \ell$,

$$\begin{aligned}
\mathbb{P}(N_t^1 = k, N_t^2 = \ell | N_t = n) &= \frac{n!}{k!\ell!} p_t^k q_t^\ell, \\
\mathbb{P}(N_t^1 = k, N_t^2) &= \mathbb{P}(N_t^1 = k, N_t^2 = \ell | N_t = n)\mathbb{P}(N_t = n) \\
&= \frac{n!}{k!\ell!} p_t^k q_t^\ell e^{-\lambda t}\frac{(\lambda t)^n}{n!} \\
&= e^{-\lambda t p_t}\frac{(\lambda t p_t)^k}{k!} e^{-\lambda t q_t}\frac{(\lambda t q_t)^\ell}{\ell!},
\end{aligned}$$

which proves exactly what was claimed.

10.7 Chapter 7

Exercise 7.11.1

Denote by $\{N_t;\ t \geq 0\}$ the counting process of the Poisson point process $\{T_n;\ n \geq 1\}$. It is not too hard to deduce from the assumptions that, with the same notation as in Definition 1.1,

$$\begin{aligned}
&\mathbb{P}(X_t = y | X_{t_0} = x_0, X_{t_1} = x_1, \ldots, X_{t_n} = x_n, X_s = x) \\
&= \sum_{n\geq 0} \mathbb{P}(X_t = y, N_t - N_s = n | X_{t_0} = x_0, X_{t_1} = x_1, \ldots, X_{t_n} = x_n, X_s = x) \\
&= \sum_{n\geq 0} (P^n)_{xy} e^{-\lambda(t-s)}\frac{[\lambda(t-s)]^n}{n!} \\
&= \mathbb{P}(X_t = y | X_s = x).
\end{aligned}$$

We deduce that the semi-group of transition matrices of the jump Markov process $\{X_t;\ t \geq 0\}$ is given by

$$P(t) = \exp[\lambda t(P - I)], \quad t \geq 0.$$

The infinitesimal generator is the matrix $Q = \lambda(P - I)$. Denote by S_1 the time of the first jump of $\{X_t\}$. Clearly $S_1 \geq T_1$, with $\mathbb{P}(S_1 > T_1) > 0$. More precisely, if we denote by N_x the first n such that $Z_n \neq Z_0 = x$, under $\mathbb{P}_x$ the law of N_x is

geometric with parameter P_{xx}, hence

$$\begin{aligned}\mathbb{P}_x(S_1 > t) &= \sum_{n\geq 1}(1-P_{xx})^{n-1}P_{xx}\mathbb{P}(T_n > t)\\ &= P_{xx}\sum_{n\geq 1}\frac{(1-P_{xx})^{n-1}}{(n-1)!}\int_{\lambda t}^{\infty} e^{-y}y^{n-1}dy\\ &= e^{-\lambda t P_{xx}}.\end{aligned}$$

Consequently, the law of S_1 is exponential with parameter λP_{xx}.

Exercise 7.11.3

1. We note that

$$\begin{aligned}&\mathbb{P}(X_t = y|X_{t_0} = x_0, X_{t_1} = x_1, \ldots, X_{t_n} = x_n, X_s = x)\\ &= \sum_{n\geq(y-x)^+}\mathbb{P}(N_t - N_s = n, P_t - P_s = k - y + x|X_{t_0} = x_0, \ldots, X_s = x)\\ &= \sum_{n\geq(y-x)^+}\mathbb{P}(N_t - N_s = n, P_t - P_s = k - y + x)\\ &= \sum_{n\geq(y-x)^+}\mathbb{P}(N_t - N_s = n, P_t - P_s = k - y + x|X_s = x)\\ &= \mathbb{P}(X_t = y|X_s = x),\end{aligned}$$

   from which the Markov property follows. The row indexed by x of the infinitesimal generator Q has only three non-zero entries: $Q_{x,x-1} = \mu$, $Q_{xx} = -(\lambda+\mu)$ and $Q_{x,x+1} = \lambda$. The process can jump from x to $x+1$ and to $x-1$, from which irreducibility follows.

2. $$\frac{X_t}{t} = \frac{N_t}{t} - \frac{P_t}{t} \to \lambda - \mu, \quad \text{as } t \to \infty.$$

   Hence, $X_t \to +\infty$ as $t \to \infty$, whenever $\lambda > \mu$, and $X_t \to -\infty$ as $t \to \infty$, whenever $\lambda < \mu$. In these two cases, the process $\{X_t;\ t \geq 0\}$ is transient.

3. In this case, the embedded chain is a symmetric nearest-neighbour $\mathbb{Z}$-valued random walk, which we know is recurrent. Moreover, the counting measure π of the points of $\mathbb{Z}$ satisfies $\pi Q = 0$, hence it is invariant. But that measure has infinite total mass, hence $\{X_t;\ t \geq 0\}$ is null recurrent.

Exercise 7.11.5

1. Approximating a general stopping time by stopping times which take their values in the set of multiples of 2^{-n} as in the proof of Proposition 3.2, we can restrict ourselves to the case where the stopping time S takes its values in the set $\{s_1, \ldots, s_N\}$, where $0 \leq s_1 < s_2 < \ldots < s_N \leq t$. Then

$$\begin{aligned} M_S &= \sum_{k=1}^{N} M_{s_k} \mathbf{1}_{\{S=s_k\}}, \\ \mathbb{E}M_S &= \sum_{k=1}^{N} \mathbb{E}\left(M_{s_k} \mathbf{1}_{\{S=s_k\}}\right) \\ &= \mathbb{E}M_t - \sum_{k=1}^{N} \mathbb{E}\left((M_t - M_{s_k})\mathbf{1}_{\{S=s_k\}}\right) \\ &= \mathbb{E}M_t. \end{aligned}$$

   The last equality follows from $\{S = s_k\} \in \mathcal{F}_{s_k}$ and the martingale property.

2. From the Markov property, if $0 < s < t$,

$$\begin{aligned} \mathbb{E}[M_t - M_s|\mathcal{F}_s^X] &= \mathbb{E}\left[f(X_t) - f(X_s) - \int_s^t Qf(X_r)dr|\mathcal{F}_s^X\right] \\ &= \mathbb{E}\left[f(X_t) - f(X_s) - \int_s^t Qf(X_r)dr|X_s\right] \\ &= \left[P_{t-s}f - f - \int_0^{t-s} P_r Qf dr\right](X_s) \end{aligned}$$

   If we admit the identity $P_r Q = Q P_r$, then, from the proof of Theorem 3.2, the right-hand side of the above identity vanishes.

3. If $f(x) = x$, then $Qf(x) = \beta - \alpha$, for all x. It then follows from the answers to parts 1 and 2 that

$$\mathbb{E}_x\left(X_{T_F \wedge t}\right) = x + (\beta - \alpha)\mathbb{E}_x(T_F \wedge t).$$

   We can pass to the limit as $t \to \infty$ in this equality, applying the dominated convergence theorem to the left-hand side, and the monotone convergence theorem to the right-hand side, whence if $\beta \neq \alpha$,

$$\mathbb{E}_x(T_F) = \frac{\mathbb{E}_x\left(X_{T_F}\right) - x}{\beta - \alpha}.$$

Now in the case $\alpha = \beta$ we choose $f(x) = x^2$, so that, for all x, $Qf(x) = 2\alpha$, hence by the same argument as above, we deduce

$$\mathbb{E}_x(T_F) = \frac{\mathbb{E}_x\left(X_{T_F}^2\right) - x^2}{2\alpha}.$$

Exercise 7.11.6

1. The transition matrix of the embedded chain is

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ \frac{\lambda}{\lambda+\mu} & 0 & \frac{\mu}{\lambda+\mu} & 0 & 0 & 0 & \dots \\ 0 & \frac{\lambda}{\lambda+\mu} & 0 & \frac{\mu}{\lambda+\mu} & 0 & 0 & \dots \\ 0 & 0 & \frac{\lambda}{\lambda+\mu} & 0 & \frac{\mu}{\lambda+\mu} & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

   All entries of the sub- and superdiagonals are non-zero, hence the embedded chain is irreducible, and the same is true for $\{X_t\}$.

2. Note that one way to construct the embedded chain $\{Z_n = X_{T_n}\}$ is to start with an i.i.d. sequence of $\{-1, 1\}$-valued random variables $\{Y_n;\ n \geq 1\}$ such that $\mathbb{P}(Y_1 = 1) = \mu/(\lambda + \mu) = 1 - \mathbb{P}(Y_1 = -1)$, and to define, for $n \geq 0$,

$$Z_{n+1} = \begin{cases} 1, & \text{if } X_n = 0, \\ Z_n + Y_n, & \text{if } X_n > 0. \end{cases}$$

   The transience or recurrence of $\{Z_n\}$, hence also of $\{X_t\}$, can be established as in Exercise 10.19 of Chapter 2.

3. The identity $(\pi\ Q)_0 = 0$ implies that $\pi_x = \pi_0(\mu/\lambda)^x$. With the additional condition $\sum_x \pi_x = 1$, we deduce

$$\pi_x = \frac{\lambda - \mu}{\lambda}\left(\frac{\mu}{\lambda}\right)^x.$$

   π is a geometric probability law on $\mathbb{N}$.

4. From what we have just seen, $\{X_t\}$ is positive recurrent if $\lambda > \mu$. If $\lambda = \mu$, the measure $\pi = (1/2,\ 1,\ 1,\ \dots)$ is invariant and has infinite total mass, hence $\{X_t\}$ is null recurrent.

Exercise 7.11.7

1. We have
$$Q = \begin{pmatrix} p-1 & 1-p & 0 & 0 \\ 0 & -1 & p & 1-p \\ p & 1-p & -1 & 0 \\ 0 & 0 & p & -p \end{pmatrix}, \quad P' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & p & 1-p \\ p & 1-p & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

2. Starting at state 1, $\{X_t\}$ waits an exponential time with parameter $1-p$, then jumps to state 2. Starting at state 2, $\{X_t\}$ waits an exponential time with parameter 1, then jumps to state 3 with probability p, and to state 4 with probability $1-p$. Starting from state 3, $\{X_t\}$ waits an exponential time with parameter 1, then jumps to state 1 with probability p, and to state 2 with probability $1-p$. Starting at state 4, $\{X_t\}$ waits an exponential time with parameter p, then jumps to state 3.

3. The embedded chain is clearly irreducible. Since the state space is finite, the chain is positive recurrent. The equation satisfied by the invariant probability is $\mu Q = 0$, which is equivalent to $\mu P = \mu$, hence the invariant probability is the same as that in Exercise 10.9 of Chapter 2.

4. The invariant probability of the embedded chain is the solution of $\mu P' = \mu$, which is $(p/3, 1/3, 1/3, (1-p)/3)$.

Exercise 7.11.8

1. The Markov chain $\{X_n\}$ is not the embedded chain associated with $\{X_t\}$. The first row of the transition matrix of the embedded chain is $(0, 1, \ldots)$, the rest of the transition matrix being the same for both chains.

2. Since $0 < p, q < 1$, both $\{X_n\}$ and the embedded chain $\{Y_n\}$ jump with positive probability from 0 to 1, and from x to both $x-1$ and $x+1$, for $x \geq 1$. Hence for $x, y \in \mathbb{N}$, each of the two chains has a non-zero probability, starting at x, of reaching y in $|y-x|$ steps. Both chains are then irreducible, as $\mathbb{N}$-valued processes.

3. It is easy to see that the equation $\pi P = \pi$ is equivalent to $\pi Q = 0$, since $Q = P - I$. Consequently, $\{X_n\}$ and $\{X_t\}$ have the same invariant measure.

   $\{X_t\}$ is transient or recurrent if and only if the embedded chain $\{Y_n\}$ has the same property, which in turn is equivalent to the fact that $\{X_n\}$ has the same property. Indeed, an irreducible $\mathbb{N}$-valued Markov chain $\{X_n\}$ is transient if and only if $X_n \to \infty$ as $n \to \infty$. This property depends only on the rows of the transition matrix after the mth, for arbitrary m, and these are the same in the transition matrices of both $\{X_n\}$ and $\{Y_n\}$.

Finally, one way to distinguish between positive recurrence and null recurrence is to check whether the invariant measures have finite or infinite total mass. We can conclude from the above arguments that if both $\{X_n\}$ and $\{X_t\}$ are recurrent, then either both are positive recurrent, or else both are null recurrent.

4. It suffices to consider $\{X_n\}$. But $X_{n+1} = X_n + \mathbf{1}_{\{X_n > 0\}} Y_{n+1} + \mathbf{1}_{\{X_n=0\}} Y^+_{n+1}$, with $X_0, Y_1, Y_2, \ldots$ independent, $\mathbb{P}(Y_k = 1) = p = 1 - \mathbb{P}(Y_k = -1)$. Hence $X_n \geq X_0 + \sum_{k=1}^n Y_k$, and since the Y_k are i.i.d., with $\mathbb{E}(Y_k) = p - q > 0$, $\sum_{k=1}^n Y_k \to +\infty$ as $n \to \infty$, and also $X_n \to +\infty$ as $n \to \infty$. $\{X_n\}$ is transient.

5. It suffices to show that $X_n \not\to \infty$ as $n \to \infty$. This will follow from the fact that, starting at any $x > 1$, the chain visits almost surely the state $x - 1$ (and then returns almost surely to x from irreducibility). Since between its visits to x and to $x - 1$ the trajectory of X_n does not visit 0, the probability distribution of the time it takes for X_n to reach $x - 1$, starting at x, is the same as for the random walk $\sum_1^n Y_k$, and we know from Exercise 10.11 of Chapter 2 that this chain is recurrent in the case $p = q$. In particular, the time taken to reach $x - 1$ starting from x is almost surely finite. One could then argue that the null recurrent behaviour of $\{X_n\}$ follows from that of $\sum_1^n Y_k$, but it is even simpler to note that the measure π defined by $\pi_x = 1$, for all $x \in \mathbb{N}$, is invariant for the chain $\{X_n\}$.

6. Solving the equation $\pi Q = 0$, we find that $\pi_1 = (p/q)\pi_0$, $\pi_2 = (\pi_1 - p\pi_0)/q = \pi_0(\lambda - p)/q = \lambda^2\pi_0$. More generally, for all $x \geq 1$, $\pi_{x+1} = (\pi_x - p\pi_{x-1})/q$. Hence if $\pi_x = \lambda\pi_{x-1}$, then $\pi_{x+1} = \pi_{x-1}(\lambda - p)/q = \lambda^2\pi_{x-1} = \lambda\pi_x$. We deduce easily that $\pi_x = \lambda^x\pi_0$, and π is a probability if $\pi_0 = 1 - \lambda$, from which we finally deduce that $\pi_x = (1 - \lambda)\lambda^x$, $x \in \mathbb{N}$. This is a geometric distribution; in fact the process which we consider here is a particular case of the $M/M/1$ queue (see Section 8.1). In the case $p < q$, both chains are positive recurrent.

7. It is easily checked that all the arguments above remain valid in the presence of the constant c, since it modifies neither the embedded chain, nor the equation for a possible invariant measure. The constant c changes only the law of the time spent by the continuous time process in each state, multiplying the parameter of the corresponding exponential distribution by c. If $c > 1$ those times are shortened, while if $c < 1$ they are lengthened.

8. One can easily check that the embedded chain associated with $\{Y_t\}$ is the same as that associated with $\{X_t\}$. But the equation for the invariant measure is now $\lambda q\pi_1 = p\pi_0$, and, for $x \geq 1$, $\lambda^{x+1}q\pi_{x+1} = \lambda^x\pi_x - \lambda^{x-1}p\pi_{x-1}$. Clearly the 'uniform' measure ($\pi_x = 1$, for all x) is invariant. Making the times spent in the states which are 'far from 0' (the larger x, the smaller the factor λ^x) longer and longer makes the expectation of the time taken to return to x infinite.

This last question gives one example of a jump Markov process which is null recurrent while its embedded chain is positive recurrent. For an example where the jump Markov process is positive recurrent while the embedded chain is null recurrent, it suffices to choose $p = q = 1/2$, and to multiply the xth row of the matrix Q by λ^{-x} ($\lambda < 1$ as above). Then the same geometric probability as in part 6 is invariant for the process in continuous time. Dividing the xth row by λ^x shortens the time spent at state x, the factor λ^x being smaller for larger x. The lengths of the excursions from 0 are shortened, in the sense that their expectations become finite, even though the number of states visited during an excursion is infinite (the embedded chain is null recurrent).

Exercise 7.11.9

1. Clearly Lemma 1.2 applies here, and we have a Markov chain. Its transition matrix is

$$P = \begin{pmatrix} 1-p/2 & p/2 & 0 & 0 & 0 & 0 \\ 1-p/2 & 0 & p/2 & 0 & 0 & 0 \\ 1-p & p/2 & 0 & p/2 & 0 & 0 \\ 1-p & 0 & p/2 & 0 & p/2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

and since all entries of the sub- and superdiagonals are non-zero, the chain can go in a finite number of steps from x to y, for all $x, y \in \mathbb{N}$, hence the irreducibility.

2. Let T denote the time of the first return to 0, starting from 0. $\mathbb{P}(T > k) \leq p^k$ for all $k \geq 1$. Hence,

$$\mathbb{E}(T) = \sum_{k=0}^{\infty} \mathbb{P}(T > k) < \infty.$$

Then the state 0 is positive recurrent, and from irreducibility the chain has that property.

3. It is easy to verify that the probability π from the statement of the exercise satisfies $\pi P = \pi$, hence it is an invariant probability distribution, which is unique from irreducibility.

4. The transition matrix of the embedded chain differs from the matrix P in part 1 only in its first row, which equals $(0\,1\,0\,0\,\ldots)$. Hence, the embedded chain is irreducible and recurrent (the fact that 0 is recurrent is not affected by a modification of the first row of the transition matrix), and $\{X_t\}$ is irreducible and recurrent. Since $Q = P - I$, and $\pi P = \pi$, $\pi Q = 0$, and by Theorem 6.2, $\{X_t\}$ is positive recurrent.

10.8 Chapter 8

Exercise 8.13.2

1. One can put the sequence $\{Y_n;\ n \geq 0\}$ in the form

$$Y_{n+1} = (Y_n + 1 - Z_n)^+,$$

where Z_n is the number of potential departures (effective number of departures if enough customers are present) between two arrivals. In other words, the law of Z_n is the law of N_T, where $\{N_t;\ t \geq 0\}$ and T are independent, the former being a Poisson process with intensity μ, and T is an exponential random variable with parameter λ. Then

$$\mathbb{P}(Z_n = k) = \frac{\lambda \mu^k}{(\lambda + \mu)^{k+1}}, \quad \mathbb{P}(Z_n \geq k) = \left(\frac{\mu}{\lambda + \mu}\right)^k.$$

Hence $\{Y_n;\ n \geq 0\}$ is a Markov chain with transition matrix

$$P = \begin{pmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} & 0 & 0 & 0 & \dots \\ \left(\frac{\mu}{\lambda+\mu}\right)^2 & \frac{\lambda\mu}{(\lambda+\mu)^2} & \frac{\lambda}{\lambda+\mu} & 0 & 0 & \dots \\ \left(\frac{\mu}{\lambda+\mu}\right)^3 & \frac{\lambda\mu^2}{(\lambda+\mu)^3} & \frac{\lambda\mu}{(\lambda+\mu)^2} & \frac{\lambda}{\lambda+\mu} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

Clearly the chain $\{Y_n;\ n \geq 0\}$ is irreducible as an $\mathbb{N}$-valued process, since all entries $P_{x,y}$ with $y \leq x + 1$ are non-zero. The chain is aperiodic since the diagonal entries of P are non-zero, hence the same is true for the diagonal entries of P^k, for all $k \geq 1$.

2. It is easy to check that the probability distribution π defined by

$$\pi_x = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^x, \quad x \in \mathbb{N},$$

is P-invariant. Hence the chain is positive recurrent, and by aperiodicity the law of Y_n converges to π as $n \to \infty$.

3. The mean time $\bar{D}$ spent by a customer in the system equals (1+ mean number of customers which he finds in front of him when he arrives)$\times$ mean service time; hence, from what we have just seen,

$$\bar{D} = \left(1 + \sum_{x \geq 0} x \pi_x\right) \frac{1}{\mu} = \frac{1}{\mu - \lambda}.$$

Then

$$\lambda \bar{D} = \frac{\lambda}{\mu - \lambda} = \mathbb{E}_\pi(X_t),$$

which is Little's formula.

4. For all $x > 0$, we have

$$\begin{aligned}
\mathbb{P}(X_{T_1^-} = x) &= \sum_{y \geq x} \mathbb{P}(Z_0 = y - x | X_0 = y) \mathbb{P}(X_0 = y) \\
&= \sum_{y \geq x} \mathbb{P}(Z_0 = y - x) \mathbb{P}(X_0 = y) \\
&= \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^{x+1},
\end{aligned}$$

while

$$\begin{aligned}
\mathbb{P}(X_{T_1^-} = 0) &= \sum_{y \geq x} \mathbb{P}(Z_0 = y | X_0 = y) \mathbb{P}(X_0 = y) \\
&= \sum_{y \geq x} \mathbb{P}(Z_0 \geq y) \mathbb{P}(X_0 = y) \\
&= \left(1 - \frac{\lambda}{\mu}\right) \frac{\lambda + \mu}{\mu}.
\end{aligned}$$

Denote by μ the law of $X_{T_1^-}$. We have that

$$\pi_0 + \pi_1 = \mu_0, \quad \pi_x = \mu_{x-1}, \ x \geq 2.$$

Hence,

$$\begin{aligned}
\sum_{x \geq 0} \mu_x f(x) &= (\pi_0 + \pi_1) f(0) + \sum_{x \geq 2} \pi_x f(x - 1) \\
&\leq \sum_{x \geq 0} \pi_x f(x),
\end{aligned}$$

provided that f is increasing, with strict inequality if f is not constant. This result is very intuitive, since $X_{T_1^-}$ equals X_0 minus the number of departures between 0 and T_1. But part 2 of the exercise tells us that $X_{T_n^-}$ converges in law towards π as $n \to \infty$. As in the case treated in Exercise 5.3 of Chapter 6, the law of $X_{T_1^-}$ would be its asymptotic distribution (here π), had the system been initialized at time $-\infty$.

10.9 Chapter 9

Exercise 9.7.1

We shall make use of the fact that $\tilde{S}_t$ is a martingale under $\mathbb{P}^*$ (see Section 9.3.5), hence

$$\mathbb{E}^*[e^{-r(T-t)}S_T|\mathcal{F}_t] = S_t.$$

But

$$S_T - K \leq (S_T - K)_+ \leq S_T,$$

hence, multiplying by $e^{-r(T-t)}$ and taking the conditional expectation given $\mathcal{F}_t$ under $\mathbb{P}^*$, we deduce that

$$S_t - e^{-r(T-t)}K \leq C_t \leq S_t.$$

The result for P_t can be deduced similarly from the inequality $(K - S_T)_+ \leq K$.

Exercise 9.7.3

1. Under $\mathbb{P}$, as under $\mathbb{P}^*$, the law of S_t is lognormal, in particular it has a density, hence the probability that $S_t = e^{-r(T-t)}K$, or equivalently that $C_t = P_t$, is zero. The inequality $C_t > P_t$ means that the conditional expectation, given the price S_t of the underlying asset at time t, of the gain from the call at time T is greater than the gain from the put, and vice versa if $P_t > C_t$.

2. If the buyer follows the strategy defined in part 1, the option yields at time T exactly what the statement of the question claims.

3. The call–put parity formula tells us that

$$C_t - P_t = S_t - e^{-r(T-t)}K,$$

hence

$$F_t = \{C_t < P_t\} = \{S_t < e^{-r(T-t)}K\},$$
$$G_t = \{C_t > P_t\} = \{S_t > e^{-r(T-t)}K\}.$$

In particular, these two events belong to the σ-algebra $\sigma(S_t)$, hence the identity of the statement follows from the fact that $\{e^{-rt}S_t;\ t \geq 0\}$ is a martingale under $\mathbb{P}^*$ (see Section 9.3.5).

4. From the formulae in part 2,

$$H - (S_T - K)_+ = (K - S_T)\mathbf{1}_{\{C_t < P_t\}},$$
$$H - (K - S_T)_+ = (S_T - K)\mathbf{1}_{\{C_t > P_t\}}.$$

Hence, multiplying by e^{-rT} and taking the expectation under $\mathbb{P}^*$, we deduce that

$$X_0 - C_0 = e^{-rT} K \mathbb{P}^*(F_t) - e^{-rt}\mathbb{E}^*[S_t \mathbf{1}_{F_t}],$$
$$X_0 - P_0 = e^{-rt}\mathbb{E}^*[S_t \mathbf{1}_{G_t}] - e^{-rT} K \mathbb{P}^*(G_t).$$

From now on let ξ denote a random variable whose probability distribution is $N(0, 1)$ under $\mathbb{P}^*$, and let F denote the distribution function of $N(0, 1)$. To compute $X_0 - C_0$ and $X_0 - P_0$ we first need to compute

$$\begin{aligned}\mathbb{P}^*(C_t < P_t) &= \mathbb{P}^*\left(e^{-rt}\frac{S_t}{S_0} < e^{-rT}\frac{K}{S_0}\right)\\ &= \mathbb{P}^*\left(\sigma\sqrt{t}\xi < -rT + \sigma^2\frac{t}{2} + \log\frac{K}{S_0}\right)\\ &= F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} + \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right),\end{aligned}$$

hence

$$\mathbb{P}^*(C_t > P_t) = 1 - F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} + \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right),$$

and

$$\begin{aligned}\mathbb{E}^*(S_t; C_t < P_t) &= S_0 e^{(r-\sigma^2/2)t}\mathbb{E}^*\left(e^{\sigma\sqrt{t}\xi}; \xi < \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}} + \frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0}\right)\\ &= S_0 e^{rt} F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} - \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right).\end{aligned}$$

Hence, since $\mathbb{E}^*(S_t) = S_0 e^{rt}$,

$$\mathbb{E}^*(S_t; C_t > P_t) = S_0 e^{rt}\left[1 - F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} - \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right)\right].$$

Finally,

$$\begin{aligned}X_0 - C_0 &= e^{-rT} K F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} + \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right)\\ &\quad - S_0 F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} - \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right),\\ X_0 - P_0 &= S_0\left[1 - F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} - \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right)\right]\\ &\quad - e^{-rT} K\left[1 - F\left(\frac{1}{\sigma\sqrt{t}}\log\frac{K}{S_0} + \frac{\sigma}{2}\sqrt{t} - \frac{rT}{\sigma\sqrt{t}}\right)\right].\end{aligned}$$

5. The equality $X_t = \max(C_t, P_t)$ follows from part 1. The hedging strategy is as follows. Between time $s = 0$ and time $s = t$ we invest

$$\frac{\partial w}{\partial x}(s, S_s)$$

on the risky asset, and between time $s = t$ and time $s = T$ we invest

$$\mathbf{1}_{\{u(t,S_t)>v(t,S_t)\}}\frac{\partial u}{\partial x}(s, S_s) + \mathbf{1}_{\{u(t,S_t)<v(t,S_t)\}}\frac{\partial v}{\partial x}(s, S_s).$$

References

[1] S. Audic, J.M. Claverie (1998) Self-identification of protein-coding regions in microbial genomes. *Proceedings of the National Academy of Sciences* **95**, 10026–10031.

[2] F. Baccelli, P. Brémaud (2002) *Elements of Queueing Theory*, Applications of Mathematics. Springer.

[3] P. Baldi (2000) On the convergence of a clustering algorithm for protein-coding regions in microbial genomes. *Bioinformatics* **16**, 367–371.

[4] F.E. Benth (2004) *Option Theory with Stochastic Analysis, Universitext*. Springer.

[5] R.N. Bhattacharya (1982) On the functional central limit theorem and the law of the iterated logarithm for Markov processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **60**, 185–201.

[6] F. Black, M. Scholes (1973) The pricing of options and corporate liabilities. *Journal of Political Economy* **81**, 637–654.

[7] N. Bouleau (1986) *Probabilités de l'ingénieur*. Hermann.

[8] P. Bratley, B.L. Fox, E.L. Schrage (1987) *A Guide to Simulation*, 2nd edition. Springer.

[9] L. Breiman (1992) *Probability*. SIAM. First published by Addison-Wesley in 1968.

[10] W.G. Cochran (1977) *Sampling Techniques*, 3rd edition. John Wiley & Sons, Inc.

[11] J.C. Cox, S.A. Ross, M. Rubinstein (1979) Option pricing: a simplified approach. *Journal of Financial Economics* **7**, 229–263.

[12] D. Dacunha-Castelle, M. Duflo (1983) *Probabilités et statistiques, 2. Problèmes á temps mobile*. Masson translation *Probability and statistics*, Vol II, Springer Verlag, New-York, 1986.

[13] L. Devroye (1986) *Non-Uniform Random Variate Generation*. Springer. English translation *Probability and statistics*, Vol II, Springer Verlag, New-York, 1986.

[14] P. Diaconis (1996) The cutoff phenomenon in finite Markov chains. *Proceedings of the National Academy of Sciences* **93**, 1659–1664.

[15] P. Diaconis, K. Khare, L. Saloff-Coste (2008) Gibbs sampling, exponential families, and orthogonal polynomials. *Statistical Science* **23**, 151–178.

[16] C.B. Do, M.S.P. Mahabhashyam, M. Brudno, S. Batzoglou (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research* **15**, 330–340.

[17] R. Durbin, S. Eddy, A. Krogh, G. Mitchison (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

[18] R. Edgar, S. Batzoglou (2006) Multiple sequence alignment. *Current Opinion in Structural Biology* **16**, 368–373.

[19] W. Ewens, G. Grant (2001) *Statistical Methods in Bioinformatics*. Springer.
[20] J. Felsenstein (2004) *Inferring Phylogenies*. Sinauer.
[21] F. Jelinek (1997) *Statistical Methods for Speech Recognition*. MIT Press.
[22] G.L. Jones (2004) On the Markov chain central limit theorem. *Probability Surveys* **1**, 299–320.
[23] D.E. Knuth (1981) *The Art of Computer Programming, Vol. 2. Seminumerical Algorithms*, 2nd edition. Addison-Wesley.
[24] D. Lamberton, B. Lapeyre (1991) *Introduction au calcul stochastique appliqué á la finance*, Mathématiques et Applications **9**, Ellipses. English translation, *Introduction to Stochastic Calculus Applied to Finance*, 2nd edition, published by Chapman & Hall/CRC in 2007.
[25] B. Lapeyre, É. Pardoux, R. Sentis (1998) *Méthodes de Monte-Carlo pour les équations de la physique*, Mathématiques et Applications **29**. Springer. English translation, *Introduction to Monte Carlo Methods for Transport and Diffusion Equations*, published by Oxford University Press in 2003.
[26] M. Matsumoto, T. Nishimura (1998) Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation* **8**, 3–30.
[27] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.
[28] F. Muri-Majoube, B. Prum (2001) Une approche statistique de l'analyse des génomes. *Gazette des Mathématiciens* **89**, 63–98.
[29] M. Musiela, M. Rutkowski (1997) *Martingale Methods in Financial Modeling: Theory and Applications*. Springer.
[30] C. Notredame, D. Higgins, J. Heringa (2000) T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology* **302**, 205–217.
[31] J.G. Propp, D.B. Wilson (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms*, Atlanta 1995, Vol. **9**, 223–252.
[32] B.D. Ripley (1987) *Stochastic simulation*. John Wiley & Sons, Inc.
[33] P. Robert (2000) *Réseaux et files d'attente: méthodes probabilistes*, Mathématiques et Applications **35**. Springer. English translation *Stochastic Networks and Queues*, Springer Verlag, Berlin, 2003.
[34] S. Robin, F. Rodolphe, S. Schbath (2003) *ADN, mots, modèles*. Belin. English translation, *DNA, Words and Models*, published by Cambridge University Press in 2005.
[35] S. Ross (1997) *Simulation*, 2nd edition. Academic Press.
[36] L. Saloff-Coste (1997) Lectures on finite Markov chains. In *École de Probabilité de Saint-Flour XXVI*, Lecture Notes in Mathematics **1665** 301–413. Springer.
[37] D.W. Stroock (2005) *An Introduction to Markov Processes*. Springer.
[38] Z. Yang (2006) *Computational Molecular Evolution*. Oxford University Press.

Index

Markov Processes and Applications: Algorithms, Networks, Genome and Finance E. Pardoux

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, Second Edition
ALTMAN, GILL and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, Third Edition
*ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, Third Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

BARNETT · Environmental Statistics: Methods & Applications
BARNETT and LEWIS · Outliers in Statistical Data, Third Edition
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, Third Edition
BERNARDO and SMITH · Bayesian Theory
BERRY, CHALONER and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BHAT and MILLER · Elements of Applied Stochastic Processes, Third Edition
BHATTACHARYA and JOHNSON · Statistical Concepts and Methods
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BIEMER, GROVES, LYBERG, MATHIOWETZ and SUDMAN · Measurement Errors in Surveys
BILLINGSLEY · Convergence of Probability Measures, Second Edition
BILLINGSLEY · Probability and Measure, Third Edition
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE and MURTHY · Reliability: Modeling, Prediction and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, Second Edition
BOLLEN · Structural Equations with Latent Variables
BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOSQ and BLANKE · Inference and Prediction in Large Dimensions
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building
BOX, HUNTER and HUNTER · Statistics for Experimenters: Design, Innovation and Discovery, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CASTILLO, HADI, BALAKRISHNAN and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Regression Analysis by Example, Fourth Edition
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, Third Edition
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÉS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, Second Edition
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, Second Edition
*COCHRAN and COX · Experimental Designs, Second Edition
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Models for Categorical Data
CONGDON · Bayesian Statistical Modelling
CONGDON · Bayesian Statistical Modelling, Second Edition
CONOVER · Practical Nonparametric Statistics, Second Edition
COOK · Regression Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
CORNELL · Experiments with Mixtures, Designs, Models and the Analysis of Mixture Data, Third Edition
COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods
*COX · Planning of Experiments
CRESSIE · Statistics for Spatial Data, Revised Edition
CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, Sixth Edition
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, Second Edition
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning

*Now available in a lower priced paperback edition in the Wiley Classics Library.

DAVID and NAGARAJA · Order Statistics, Third Edition
*DEGROOT, FIENBERG and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, Second Edition
*DOOB · Stochastic Processes
DOWDY, WEARDEN and CHILKO · Statistics for Research, Third Edition
DRAPER and SMITH · Applied Regression Analysis, Third Edition
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, Second Edition
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, Third Edition
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS (editors) · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS and PEACOCK · Statistical Distribution, Third Edition
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, Third Edition, Revised; Volume II, Second Edition
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD and WARE · Applied Longitudinal Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, Second Edition
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, Second Edition
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models.
GEISSER · Modes of Parametric Statistical Inference

*Now available in a lower priced paperback edition in the Wiley Classics Library.

GELMAN and MENG (editors) · Applied Bayesian Modeling and Casual Inference from Incomplete-data Perspectives
GEWEKE · Contemporary Bayesian Econometrics and Statistics
GHOSH, MUKHOPADHYAY and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GIVENS and HOETING · Computational Statistics
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, Second Edition
GOLDSTEIN and LEWIS · Assessment: Problems, Development and Statistical Issues
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS and HARRIS · Fundamentals of Queueing Theory, Third Edition
*HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HEDEKER and GIBBONS · Longitudinal Data Analysis
HELLER · MACSYMA for Statisticians
HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HINKELMANN and KEMPTHORNE · Design and analysis of experiments, Volume 2: Advanced Experimental Design
HOAGLIN, MOSTELLER and TUKEY · Exploratory Approach to Analysis of Variance
HOAGLIN, MOSTELLER and TUKEY · Exploring Data Tables, Trends and Shapes
*HOAGLIN, MOSTELLER and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, Second Edition
HOEL · Introduction to Mathematical Statistics, Fifth Edition
HOGG and KLUGMAN · Loss Distributions
HOLLANDER and WOLFE · Nonparametric Statistical Methods, Second Edition
HOSMER and LEMESHOW · Applied Logistic Regression, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
HUBER · Robust Statistics
HUBERTY · Applied Discriminant Analysis
HUNT and KENNEDY · Financial Derivatives in Theory and Practice, Revised Edition
HUSKOVA, BERAN and DUPAC · Collected Works of Jaroslav Hajek – with Commentary
HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
IMAN and CONOVER · A Modern Approach to Statistics
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, Fifth Edition
JUDGE, GRIFFITHS, HILL, LU TKEPOHL and LEE · The Theory and Practice of Econometrics, Second Edition
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, Second Edition
JOHNSON, KOTZ and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, Second Edition
JOHNSON, KOTZ and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ and KEMP · Univariate Discrete Distributions, Second Edition
JUREĔKOVÁ and SEN · Robust Statistical Procedures: Asymptotics and Interrelations
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE and SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, Second Edition
KARIYA and KURATA · Generalized Least Squares
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, Second Edition
KHURI, MATHEW and SINHA · Statistical Tests for Mixed Linear Models

*KISH · Statistical Design for Research
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLUGMAN, PANJER and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN and JOHNSON · Continuous Multivariate Distributions, Volume 1, Second Edition
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, Second Edition
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, Third Edition
LAWLESS · Statistical Models and Methods for Lifetime Data, Second Edition
LAWSON · Statistical Methods in Spatial Epidemiology, Second Edition
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, Third Edition
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, Second Edition
LLOYD · The Statistical Analysis of Categorical Data
LOWEN and TEICH · Fractal-Based Point Processes
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition

*Now available in a lower priced paperback edition in the Wiley - Interscience Paperback Series.

MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY and TOLLEY · Statistical Applications Using Fuzzy Sets
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and practice
MARDIA and JUPP · Directional Statistics
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
MASON, GUNST and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, Second Edition
MCCULLOCH and SERLE · Generalized, Linear and Mixed Models
MCFADDEN · Management of Data in Clinical Trials
MCLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
MCLACHLAN, DO and AMBROISE · Analyzing Microarray Gene Expression Data
MCLACHLAN and KRISHNAN · The EM Algorithm and Extensions
MCLACHLAN and PEEL · Finite Mixture Models
MCNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MICKEY, DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, Third Edition
*MILLER · Survival Analysis, Second Edition
MONTGOMERY, PECK and VINING · Introduction to Linear Regression Analysis, Fourth Edition
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MULLER and STEWART · Linear Model Theory: Univariate, Multivariate and Mixed Models
MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis and Nonlinear Optimization
MURTHY, XIE and JIANG · Weibull Models
MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Second Edition
MYERS, MONTGOMERY and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†NELSON · Accelerated Testing, Statistical Models, Test Plans and Data Analysis
† NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, SUGIHARA and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, Second Edition
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PALTA · Quantitative Methods in Population Health: Extentions of Ordinary Regression
PANJER · Operational Risks: Modeling Analytics
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO and TSAY · A Course in Time Series Analysis
PIANTADOSI · Clinical Trials: A Methodologic Perspective
PORT · Theoretical Probability for Applications
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
PRESS · Bayesian Statistics: Principles, Models and Applications
PRESS · Subjective and Objective Bayesian Statistics, Second Edition
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PUKELSHEIM · Optimal Experimental Design
PURI, VILAPLANA and WERTZ · New Perspectives in Theoretical and Applied Statistics
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
QIU · Image Processing and Jump Regression Analysis
RAO · Linear Statistical Inference and its Applications, Second Edition
RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods and Applications, Second Edition
RENCHER · Linear Models in Statistics
RENCHER · Methods of Multivariate Analysis, Second Edition
RENCHER · Multivariate Statistical Inference with Applications
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROBINSON · Practical Strategies for Experimenting
ROHATGI and SALEH · An Introduction to Probability and Statistics, Second Edition
ROLSKI, SCHMIDLI, SCHMIDT and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice

† Now available in a lower priced paperback edition in the Wiley - Interscience Paperback Series.

ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROSSI, ALLENBY and MCCULLOCH · Bayesian Statistics and Marketing
ROUSSEEUW and LEROY · Robust Regression and Outline Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, Second Edition
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN and SCOTT (editors) · Sensitivity Analysis
*SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation and Application
SCHOTT · Matrix Analysis for Statistics
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, Second Edition
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: Its Only a Game!
SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order and Shape Restrictions
SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL and MECKE · Stochastic Geometry and Its Applications, Second Edition
STOYAN and STOYAN · Fractals, Random and Point Fields: Methods of Geometrical Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON and SONG · Methods for Meta-Analysis in Medical Research

*Now available in a lower priced paperback edition in the Wiley Classics Library.

TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, Second Edition
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA and STIGLER (editors) · Box on Quality and Discovery: with Design, Control and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VAN BELLE, FISHER, HEAGERTY and LUMLEY · Biostatistics: A Methodology for the Health Sciences, Second Edition
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
WEISBERG · Applied Linear Regression, Second Edition
WELISH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, Second Edition
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, Second Edition
WU and HAMADA · Experiments: Planning, Analysis and Parameter Design Optimization
WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches
YANG · The Construction Theory of Denumerable Markov Processes
YOUNG, VALERO-MORA and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics

*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions: Applications in the Health Sciences
ZHOU, OBUCHOWSKI and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.