Universitext

Universitext

Series editors

Universitext is a series of textbooks that presents material from a wide variety of mathematical disciplines at master's level and beyond. The books, often well class-tested by their author, may have an informal, personal, even experimental approach to their subject matter. Some of the most successful and established books in the series have evolved through several editions, always following the evolution of teaching curricula, into very polished texts.

Thus as research topics trickle down into graduate-level teaching, first textbooks written for new, cutting-edge courses may make their way into *Universitext*.

More information about this series at http://www.springer.com/series/223

Jean-François Collet

Discrete Stochastic Processes and Applications

Springer

Jean-François Collet
Laboratoire J.A. Dieudonné
Université de Nice Sophia-Antipolis
Nice Cedex 02
France

ISSN 0172-5939 ISSN 2191-6675 (electronic)
Universitext
ISBN 978-3-319-74017-1 ISBN 978-3-319-74018-8 (eBook)
https://doi.org/10.1007/978-3-319-74018-8

Library of Congress Control Number: 2017964594

Mathematics Subject Classification (2010): Primary: 60J10, 60J27, 60J28, 60J75, 94A17; Secondary: 26B25, 60J80, 60J85, 94A24

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG
part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Pour vous quatre

Preface

Stochastic processes are quite prevalent in scientific modeling, and the fine analysis of their mathematical properties relies on tools coming from such diverse branches of mathematics as (aside from probability theory of course) linear algebra, convex analysis, and information theory.

The aim of this book is to give a self-contained introduction to this extensive toolbox at a rather elementary level and to discuss its use in the study of Markov processes as well as in some other applications such as coding theory, population dynamics, and the design of search engines.

The first part of the book focuses on the rigorous theory of Markov chains and could provide a basis for a one-semester introductory course on this topic.

The main goal is to make it possible for the reader to develop a good intuition for the probabilistic concepts relevant to Markov processes without having to digest measure theory first. This is why we restrict ourselves to Markov processes on countable spaces, a.k.a. Markov chains. In this case, the mathematical form of the idea of absence of long-term memory is rather intuitive and does not require the extra technical tools (such as stopping times and filtrations) needed in the continuous case. The approach will be very gradual, starting with the discrete-time case, then moving on to continuous time. Along the way, new technicalities arise but the underlying probabilistic concepts are the same.

One of the most important results from the theory of Markov chains is the convergence theorem for ergodic chains (Theorem 1.39 in the text). This will be proved twice, first with probabilistic techniques in Chapter 1 (the coupling method is fundamentally probabilistic in nature) and second (for finite spaces) by pure linear algebra in Chapter 2. The fact that such a general result may be proved by attacking it from such radically different directions should cause readers to marvel once they have reached the end of Chapter 2, and it says a lot about the richness of the field and the beauty of mathematics. Chapter 2 ends with a popular application of the convergence result for Markov chains, the PageRank algorithm used in search engines.

The next step is to move on to continuous-time processes. The study of the Poisson process in Chapter 3 will serve as a pretext to introduce some of the relevant concepts without getting too technical. This will begin with a handmade construction of the Poisson process, before a review of the classical definitions and proof of their equivalence. Chapter 4 then deals with the rigorous theory of general continuous-time Markov processes. Since this text is intended as only a first exposition to stochastic processes, in order to avoid technicalities (which would mean introducing stopping times and filtrations) we do not cover the strong Markov property. Admittedly, the price to pay is that the rigorous proofs based on the use of skeletons may get a bit technical in places, so the reader who is interested mostly in applications might skip the detailed proofs.

Chapter 5 should provide some necessary relief after so much theory. It explores two kinds of examples: random walks and birth–death processes. The approach taken is much more applied than in the previous chapters and focuses on actual computations rather than theoretical results.

The first chapters provide two approaches to the study of the asymptotic behavior of Markov chains; a third route (and probably the least represented in textbooks on stochastic processes) is the use of entropy.

The second part of the text provides an introduction to various uses of convexity in probability as well as to the concept of entropy. It is more applied in nature (although mathematically rigorous) and might serve as a one-semester gentle introduction to coding and information theory.

At the heart of the uses of entropy is a certain number of inequalities that are in fact more or less all convexity inequalities. Assuming no prior knowledge of convexity whatsoever, Chapter 6 is an introduction to convex sets and convex maps, and to the relevant inequalities. In particular, it includes a discussion of Bregman divergences, which although they are now used in data clustering, are not often found in basic probability textbooks.

Chapter 7 collects the main quantities used in information theory; according to the expositional philosophy of the first part we present everything in the framework of discrete distributions, but the extension to absolutely continuous distributions is trivial.

The main theme here is that many quantities (such as Bregman divergences and Φ-divergences and in particular the Kullback–Leibler divergence) defined on the space of probability vectors may be used as distances even though they are not distances in the true mathematical sense of the term. What is meant here by "may be used as distances" becomes clear once the main inequalities, Jensen's and Pinsker's, are understood. In Section 7.4 we show how to use entropy to give a third proof of the convergence theorem for ergodic chains (at which point the above comment on the beauty and unity of mathematics should be reiterated). This is a use in the setting of Markov chains of what is known as the "entropy dissipation method" in partial differential equations and dynamical systems. The chapter ends with a detailed

introduction to exponential families, which are parametric families of probability distributions that under some constraints maximize entropy. Finally, Chapter 8 explores the connection with binary coding; essentially, it is shown that the entropy of a random source provides an incompressible lower bound for the length of any uniquely decipherable code. To make a long story short, loss of memory in a Markov chain or complexity of a random source may be quantified using the same tool, entropy.

The prerequisites for this text are as follows.

- From linear algebra: multiplication of matrices, scalar product, and the concept of eigenvalue (no knowledge of diagonalization results is needed); the section on the Perron–Frobenius theorem is totally self-contained.
- From calculus: sequences and proofs by induction, manipulation of convergent series, the chain rule, computation of one-dimensional integrals by change of variables and integration by parts. From calculus in several variables, the notions of gradient and Hessian.
- From probability theory: discrete and absolutely continuous random variables; conditional probability, computations of expectations, moments, Gaussian variables.

Acknowledgments. Now comes the most pleasant and easiest paragraph of this book to write. It is indeed a pleasure to thank all the people without whom this book would not have been written: Didier Auroux, Florent Berthelin, Jaques Blum, Cédric Boulbe, David Chiron, Jean-Antoine Désidéri, Frédéric Gruy, Bernard Guy, David Hoff, Stéphane Junca, Frédéric Poupaud, Francesca Rapetti, Vitaly Volpert.

Touët-sur-Var
Jean-François Collet July 2017

Contents

Terminology, Abbreviations, and Typographical Conventions

Terminology and Abbreviations

Throughout the text, the words "positive" and "negative" are understood strictly; the components of a vector q are interchangeably denoted by q_i or $q(i)$, and similarly, the components of a matrix p are interchangeably denoted by $p_{i,j}$ or $p(i,j)$. We refer to the probability mass function of a discrete random variable X simply as its probability vector q, and then write $q(x)$ for $P(X = x)$, even in the case that X takes its values in an infinite countable set.

Unless otherwise stated, vectors appearing in products are understood as row vectors; therefore, in a product of a vector by a matrix, the vector will appear to the left of the matrix.

The end of a rigorous proof is indicated by the symbol $\square$. The notation $:=$ indicates that an equality should be understood as a definition of its left-hand-side (as opposed to equalities resulting from a computation).

Notation

- A^t: the transpose matrix of a matrix A, meaning $A^t(i,j) := A(j,i)$. This notation includes vectors, so for example, if x is a column vector, then x^t is the corresponding row vector.
- I_n: the $n \times n$ identity matrix; sometimes denoted by I when n is implicit.
- $\delta_{x,y}$ is the well-known Kronecker delta : it is 1 if $x = y$, and 0 otherwise.
- $\mathbb{N}^*$, $\mathbb{R}^*$: positive integers, positive real numbers.
- $[x,y]$: the segment joining two points $x, t \in \mathbb{R}^n$, which means the set of all points of the form $\lambda x + (1-\lambda)y$ for some $0 \leq \lambda \leq 1$. In particular, if $n = 1$, this is just the familiar *interval* $[x,y]$. If we wish to exclude (for instance) y, we write $[x,y)$.

- $\nabla f(x)$: the gradient of a scalar-valued map (in this text we never have to use the derivatives of a vector-valued map) $f : \mathbb{R}^n \to \mathbb{R}$:

$$\nabla f(x)(i) := \frac{\partial f}{\partial x_i}(x).$$

- $\nabla^2 f(x)$: the Hessian matrix of a map $f : \mathbb{R}^n \to \mathbb{R}$:

$$\nabla^2 f(x)(i,j) := \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

- $H(X), H(p)$: entropy of a random variable, of a probability vector
- $\Gamma(\alpha), B(\alpha, \beta)$: the Euler gamma and beta functions:

$$\Gamma(\alpha) := \int_0^\infty e^{-t} t^{\alpha-1}\, dt, \quad B(\alpha,\beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt.$$

- $I_0(z)$: the modified Bessel function of order zero :

$$I_0(z) := \frac{1}{\pi} \int_0^\pi \exp(x \cos t)\, dt.$$

- $D(f), R(f)$: the domain and range of a map f. The notation $f : X \to Y$ means that the function f is defined on some subset of X, the domain $D(f)$. This means that what we call a function is what some people call a *partial function*. The raison d'être for this choice is that we use X just as a way to specify the nature of the variable (e.g., scalar or vector), without having to make assumptions about $D(f)$.
- $\mathcal{P}_n$: the set of n-component probability vectors:

$$\mathcal{P}_n := \{p \in \mathbb{R}^n : p_i \geq 0, p_1 + \cdots + p_n = 1\}.$$

- $\mathcal{P}_n^*$: the set of n-component probability vectors with positive components:

$$\mathcal{P}_n^* := \{p \in \mathbb{R}^n : p_i > 0, p_1 + \cdots + p_n = 1\}.$$

- $\log_2 x$: the logarithm to the base two; if a computation or statement is valid independently of the base, then the notation log is used. In this case, we assume only that the base is strictly greater than 1 (so that the corresponding logarithm function is increasing).
- $E_p(\phi)$: the expectation of ϕ under the probability distribution p. In other words, $E_p(\phi) := E(\phi(X))$, where X is a random variable having probability distribution p.
- $\mathcal{A}^+$: the set of all finite strings over a finite set $\mathcal{A}$.
- c^+: the extension to $\mathcal{A}^+$ of a binary code c defined on $\mathcal{A}$.
- $l_c(x)$: the code length of the symbol x, that is, the length of the codeword $c(x)$.

- $L(c)$: the average code length of the code c.
- $d_H(x, y)$: the Hamming distance between two binary words x and y of the same length.
- $\stackrel{d}{=}$: equality in distribution.
- $\stackrel{d}{\rightarrow}$, $\stackrel{p}{\rightarrow}$, $\stackrel{a.s.}{\rightarrow}$: convergence in distribution, in probability, almost sure.
- $\sim$: distributed as:
  1. $X \sim \mathcal{P}(\lambda)$ means that X is a Poisson variable of parameter λ;
  2. $X \sim \mathcal{B}(n, p)$ means that X is a binomial variable of parameters n and p;
  3. $X \sim \mathcal{NB}(k, p)$ means that X is a negative binomial variable of parameters k and p;
  4. $X \sim \mathcal{G}(p)$ means that X is a geometric variable of parameter p;
  5. $X \sim \mathcal{E}(\lambda)$ means that X is an exponential variable of parameter λ;
  6. $X \sim \mathcal{TE}(\lambda)$ means that X is a truncated exponential variable of parameters λ and n;
  7. $X \sim \Gamma(k, \lambda)$ means that X is a Γ variable of parameters λ and k.