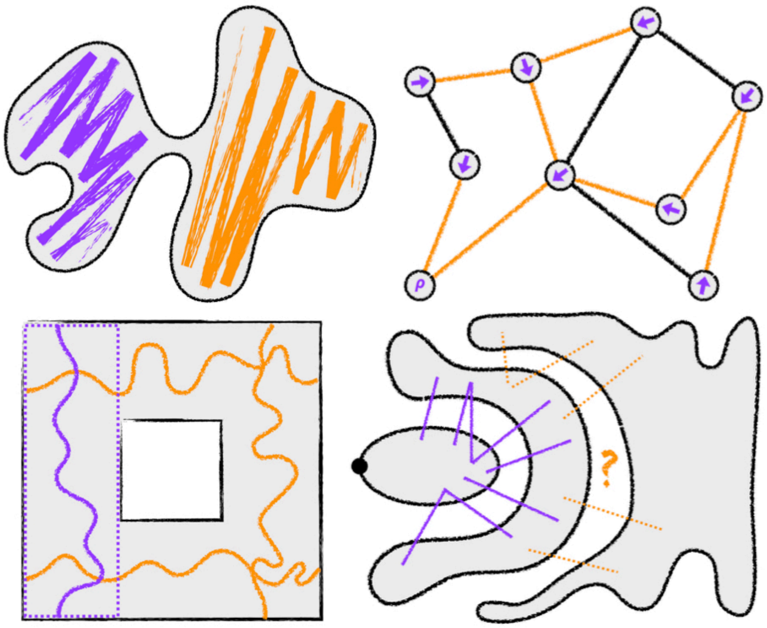
Modern Discrete Probability

An Essential Toolkit

Sébastien Roch



December 20, 2023

To Betsy

Contents

Preface

This book arose from a set of lecture notes prepared for a one-semester topics course I taught at the University of Wisconsin–Madison in 2014, 2017, 2020 and 2023 which attracted a wide spectrum of students in mathematics, computer sciences, engineering, and statistics.

What is it about?

The purpose of the book is to provide a graduate-level introduction to discrete probability. Topics covered are drawn primarily from stochastic processes on graphs: percolation, random graphs, Markov random fields, random walks on graphs, etc. No attempt is made at covering these broad areas in depth. Rather, the emphasis is on illustrating important techniques used to analyze such processes. Along the way, many standard results regarding discrete probability models are worked out.

The "modern" in the title refers to the (non-exclusive) focus on nonasymptotic methods and results, reflecting the impact of the theoretical computer science literature on the trajectory of this field. In particular several applications in randomized algorithms, probabilistic analysis of algorithms and theoretical machine learning are used throughout to motivate the techniques described (although, again, these areas are not covered exhaustively).

Of course the selection of topics is somewhat arbitrary and driven in part by personal interests. But the choice was guided by a desire to introduce techniques that are widely used across discrete probability and its applications. The material discussed here is developed in much greater depth in the following (incomplete list of) excellent textbooks and expository monographs, many of which influenced various sections of this book:

- Agarwal, Jiang, Kakade, Sun. *Reinforcement learning: Theory and algorithms.* [AJKS22]
- Aldous, Fill. *Reversible Markov chains and random walks on graphs.* [AF]
- Alon, Spencer. *The Probabilistic Method.* [AS11]

- Béla Bollobás. *Random graphs*. [Bol01]
- Boucheron, Lugosi, Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. [BLM13]
- Chung, Lu. *Complex graphs and networks*. [CL06]
- Durrett. *Random Graph Dynamics*. [Dur06]
- Frieze and Karoński. *Introduction to random graphs*. [FK16]
- Grimmett. *Percolation*. [Gri10b]
- Janson, Luczak, Rucinski. *Random Graphs*. [JLR11]
- Lattimore, Szepesvári. *Bandit Algorithms*. [LS20]
- Levin, Peres, Wilmer. *Markov chains and mixing times*. [LPW06]
- Lyons, Peres. *Probability on trees and networks.* [LP16]
- Mitzenmacher, Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. [MU05]
- Motwani, Raghavan. *Randomized algorithms*. [MR95]
- Rassoul-Agha, Seppäläinen. *A course on large deviations with an introduction to Gibbs measures.* [RAS15]
- S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. [SSBD14]
- van Handel. *Probability in high dimension.* [vH16]
- van der Hofstad. *Random graphs and complex networks. Vol. 1.* [vdH17]
- Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. [Ver18]

In fact the book is meant as a first foray into the basic results and/or toolkits detailed in these more specialized references. My hope is that, by the end, the reader will have picked up sufficient fundamental background to learn advanced material on their own with some ease. I should add that I used many additional helpful sources; they are acknowledged in the "Bibliographic remarks" at the end of each chapter. It is impossible to cover everything. Some notable omissions include, e.g., graph limits [Lov12], influence [KS05], and group-theoretic methods [Dia88], among others.

Much of the material covered here (and more) can also be found in [HMRAR98], [Gri10a], and [Bre17] with a different emphasis and scope.

Prerequisites

It is assumed throughout that the reader is fluent in undergraduate linear algebra, for example, at the level of [Axl15], and basic real analysis, for example, at the level of [Mor05].

In addition, it is recommended that the reader has taken at least one semester of graduate probability at the level of [Dur10]. I am also particularly fond of [Wil91], which heavily influenced the appendix where measure-theoretic background is reviewed. Some familiarity with countable Markov chain theory is necessary, as covered for instance in [Dur10, Chapter 6]. An advanced undergraduate or Masters' level treatment such as [Dur12], [Nor98], [GS20], [Law06] or [Bre20] will suffice however.

Organization

The book is organized around five major "tools." The reader will have likely encountered those tools in prior probability courses. The goal here is to develop them further, specifically with their application to discrete random structures in mind, and to illustrate them in this setting on a variety of major, classical results and applications.

In the interest of keeping the book relatively self-contained and serving the widest spectrum of readers, each chapter begins with a "background" section which reviews the basic material on which the rest of the chapter builds. The remaining sections then proceed to expand on two or three important specializations of the tools. While the chapters are meant to be somewhat modular, results from previous chapters do occasionally make an appearance.

The techniques are illustrated throughout with simple examples first, and then with more substantial ones in separate sections marked with the symbol ▷ . I have attempted to provide applications from many areas of discrete probability and theoretical computer science, although some techniques are better suited for certain types of models or questions. The examples and applications are important: many of the tools are quite straightforward (or even elementary), and it is only when seen in action that their full power can be appreciated. Moreover, the ▷ sections serve as an excuse to introduce the reader to classical results and important applications—beyond their reliance on specific tools.

Chapter 1 introduces some of the main models from probability on graphs that we come back to repeatedly throughout the book. It begins with a brief review of graph theory and Markov chain theory.

Chapter 2 starts out with the probabilistic method, including the first moment principle and second moment method, and then it moves on to concentration inequalities for sums of independent random variables, mostly sub-Gaussian and sub-exponential variables. It also discusses techniques to analyze the suprema of random processes.

Chapter 3 turns to martingales. The first main topic there is the Azuma-

Hoeffding inequality and the method of bounded differences with applications to random graphs and stochastic bandit problems. The second main topic is electrical network theory for random walks on graphs.

Chapter 4 introduces coupling. It covers stochastic domination and correlation inequalities as well as couplings of Markov chains with applications to mixing. It also discusses the Chen-Stein method for Poisson approximation.

Chapter 5 is concerned with spectral methods. A major topic there is the use of the spectral theorem and geometric bounds on the spectral gap to control the mixing time of a reversible Markov chain. The chapter also introduces spectral methods for community recovery in network analysis.

Chapter 6 ends the book with applications of branching processes. Among other applications, an introduction to the reconstruction problem on trees is provided. The final section gives a detailed analysis of the phase transition of the Erdös-Rényi graph, where techniques from all chapters of the book are brought to bear.

Acknowledgments

The lecture notes on which this book is based were influenced by graduate courses of David Aldous, Steve Evans, Elchanan Mossel, Yuval Peres, Alistair Sinclair, and Luca Trevisan at UC Berkeley, where I learned much of this material. In particular scribe notes for some of these courses helped shape early iterations of this book.

I have also learned a lot over the years from my collaborators and mentors as well as my former and current Ph.D. students and postdocs. I am particularly grateful to Elchanan Mossel and Allan Sly for encouragements to finish this project and to the UW-Madison students who have taken the various iterations of the course that inspired the book for their invaluable feedback.

Warm thanks to everyone in the departments of mathematics at UCLA and UW-Madison who have provided the stimulating environments that made this project possible. Beyond my current department, I am particularly indebted to my colleagues in the NSF-funded Institute for Foundations of Data Science (IFDS) who have significantly expanded my knowledge of applications of this material in machine learning and statistics.

Finally I thank my parents, my wife and my son for their love, patience and support.

Notation

Throughout the book, we will use the following notation.

- The real numbers are denoted by $\mathbb{R}$, the nonnegative reals are denoted by $\mathbb{R}_+$, the integers are denoted by $\mathbb{Z}$, the nonnegative integers are denoted by $\mathbb{Z}_+$, the natural numbers (i.e., positive integers) are denoted by $\mathbb{N}$ and the rational numbers are denoted by $\mathbb{Q}$. We will also use the notation $\overline{\mathbb{Z}}_+ := \{0, 1, \ldots, +\infty\}$.

- For two reals $a, b \in \mathbb{R}$,

  $$a \wedge b := \min\{a, b\}, \quad a \vee b := \max\{a, b\},$$

  and

  $$a^+ = 0 \vee a, \quad a^- = 0 \vee (-a).$$

- For a real a, $\lfloor a \rfloor$ is the largest integer that is smaller than or equal to a and $\lceil a \rceil$ is the smallest integer that is larger than or equal to a.

- For $x \in \mathbb{R}$, the natural (i.e., base e) logarithm of x is denoted by $\log x$. We also let $\exp(x) = e^x$.

  natural logarithm

- For a positive integer $n \in \mathbb{N}$, we let

  $$[n] := \{1, \ldots, n\}.$$

- Let A be a set. The cardinality of A is denoted by $|A|$. The power set of A, i.e., the collection of all of its subsets, is denoted by 2^A.

- For two sets A, B, their cartesian product is denoted by $A \times B$.

- We will use the following notation for standard vectors: $\mathbf{0}$ is the all-zero vector, $\mathbf{1}$ is the all-one vector, and $\mathbf{e}_i$ is the standard basis vector with a one in coordinate i and a zero elsewhere. In each case, the dimension is implicit, as well as whether it is a row or column vector.

- For a vector $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$ and real $p > 0$, its *p-norm* (or *ℓ^p-norm*) is *p-norm*
$$\|\mathbf{u}\|_p := \left(\sum_{i=1}^n |u_i|^p \right)^{1/p}.$$
When $p = +\infty$, we have
$$\|\mathbf{u}\|_\infty := \max_i |u_i|.$$
We also use the notation $\|\mathbf{u}\|_0$ to denote the number of nonzero coordinates of $\mathbf{u}$ (although it is not a norm; see Exercise 1.1). For two vectors $\mathbf{u} = (u_1, \ldots, u_n), \mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, their *inner product* is *inner product*
$$\langle \mathbf{u}, \mathbf{v} \rangle := \sum_{i=1}^n u_i v_i.$$
The same notations apply to row vectors.

- For a matrix A, we denote the entries of A by $A(i,j)$, $A_{i,j}$, or A_{ij}. The i-th row of A is denoted by $A(i,\cdot)$ or $A_{i,\cdot}$. The j-th column of A is denoted by $A(\cdot,j)$ or $A_{\cdot,j}$. The transpose of A is A^T.

- For a vector $\mathbf{z} = (z_1, \ldots, z_d)$, we let $\mathrm{diag}(\mathbf{z})$ be the diagonal matrix with diagonal entries $z_1, \ldots, z_d$.

- The *binomial coefficients* are defined as *binomial coefficients*
$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$
where $k, n \in \mathbb{N}$ with $k \leq n$ and $n! = 1 \times 2 \times \cdots \times n$ is the factorial of n. Some standard approximations for $\binom{n}{k}$ and $n!$ are listed in Appendix A. See also Exercises 1.2, 1.3, and 1.4.

- We use the abbreviation a.s. for "almost surely," that is, with probability 1. We use "w.p." for "with probability."

- Convergence in probability is denoted as $\to_{\mathrm{p}}$. Convergence in distribution is denoted as $\xrightarrow{\mathrm{d}}$.

- For a random variable X and a probability distribution μ, we write $X \sim \mu$ to indicate that X has distribution μ. We write $X \overset{\mathrm{d}}{=} Y$ if the random variables X and Y have the same distribution.

- For an event A, the random variable $\mathbf{1}_A$ is the indicator of A, that is, it is 1 if A occurs and 0 otherwise. We also use $\mathbf{1}\{A\}$.

- For probability measures μ, ν on a countable set S, their *total variation distance* is
$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \subseteq S} |\mu(A) - \nu(A)|.$$

total variation distance

- For nonnegative functions $f(n)$, $g(n)$ of $n \in \mathbb{Z}_+$ we write $f(n) = O(g(n))$ if there exists a positive constant $C > 0$ such that $f(n) \leq Cg(n)$ for all n large enough. Similarly, $f(n) = \Omega(g(n))$ means that $f(n) \geq cg(n)$ for some constant $c > 0$ for all n large enough. The notation $f(n) = \Theta(g(n))$ indicates that both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ hold. We also write $f(n) = o(g(n))$ or $g(n) = \omega(f(n))$ or $f(n) \ll g(n)$ or $g(n) \gg f(n)$ if $f(n)/g(n) \to 0$ as $n \to +\infty$. If $f(n)/g(n) \to 1$ we write $f(n) \sim g(n)$. The same notations are used for functions of a real variable x as $x \to +\infty$.