Djalil Chafaï, Florent Malrieu

A Collection of Stochastic Models

Post-publication version, Summer 2026, automatically translated
from the French original by using ChatGPT Codex



Typeset with the LaTeX 2$_\varepsilon$ class `svmono.cls`

Preface



«*. . . I prefer concrete things and I don't like to learn more about abstract stuff than I absolutely have to. . . .* ». Marc Kac (1914 – 1984), in *Enigmas of Chance: An Autobiography* (1985, chapter 5, page 112).

This collection has grown over time, and we are delighted to share it. It is designed to be consulted occasionally, at random, by browsing its pages, its table of contents or its index. We believe that this kind of document is useful in a highly specialised scientific world that leaves little room for breadth and eclecticism.

Our primary audience is probability researchers and lecturers, whether new to the field or experienced. This collection aims to broaden their knowledge of probability by arousing their curiosity. We hope it will inspire their master's courses and their supervision of introductory research projects. Many chapters may also directly benefit master's students and students preparing for the French agrégation examination.

The guiding principle of this book is to organise the material around models rather than tools, devoting one chapter to each model. Although sometimes

related, the chapters are largely self-contained. They contain little or no review of prerequisites. However, at the end of the book, just before the bibliography, we provide recommended references on limit theorems, discrete-time martingales, Markov chains with at most countable state spaces, and the basics of stochastic calculus. Each chapter opens with lists of keywords and tools used, together with an indication of difficulty (one to three stars). The introduction and, especially, the "Further reading" section of each chapter place the topic in context and suggest references.

This collection does not claim to be a reference work on probabilistic models in biology, computer science, engineering or physics, still less to provide solutions for practitioners in these fields. The models considered here are often merely inspired by applications, and the treatment generally concerns considerably simplified versions. That is already quite a lot!

This collection grew out of the applied mathematics master's courses and preparation courses for the modelling component of the mathematics agrégation that we taught at the universities of Toulouse, Rennes, Marne-la-Vallée, Paris-Dauphine and Tours. The quality of the text, in both content and form, owes much to our earliest readers at every level. Many thanks, then, to our former students! Many thanks also to Arnaud Guyader, Bernard Bercu and Olivier Zindy for their extensive reading and comments. We also thank Jürgen Angst, Amine Asselah, Jean-Baptiste Bardet, Charles Bordenave, Raphaël Butez, Marie-Line Chabanol, Bertrand Cloez, Yan Doumerc, Olivier Durieu, Lucas Gerin, Hélène Guérin, Bénédicte Haas, Igor Kortchemski, Jamal Najim, Julien Poisat, Justin Salez, Pierre Tarrès, Marie-Noémie Thai, Frédérique Watbled, and Pierre-André Zitt for their help. This collection also benefited from careful and constructive readings by several anonymous referees. Finally, we are grateful to Sylvie Méléard for encouraging us to turn lecture notes into a book and for her thorough editorial work. Nevertheless, like any human endeavour, this collection remains limited and imperfect. Errata will be available online, together with the programs used to produce the illustrations.

Enjoy your reading!

Djalil Chafaï and Florent Malrieu
Vincennes and Saint-Avertin, autumn 2015
Updated in winter 2021

$$(\Omega, \mathcal{F}, \mathbb{P})$$

All random variables in this collection are defined on this space.

Contents

1 Heads, tails, and coupons 1
- 1.1 Coin tossing 1
- 1.2 The binomial-Gaussian approximation 4
- 1.3 Total variation distance 4
- 1.4 The binomial-Poisson approximation 8
- 1.5 The coupon collector problem 9
- 1.6 Further reading 14

2 Random walks 17
- 2.1 Simple random walk on the line 17
- 2.2 Simple symmetric random walk in space 24
- 2.3 The Dirichlet problem and the Gaussian free field 27
- 2.4 Random walk on the symmetric group 31
- 2.5 Further reading 34

3 Branching and Galton–Watson processes 37
- 3.1 Extinction and the threshold phenomenon 38
- 3.2 The three possible cases 43
- 3.3 Tree size in the critical and subcritical regimes 47
- 3.4 Immigration 50
- 3.5 Further reading 52

4 Permutations, partitions, and graphs 53
- 4.1 A basic algorithm for discrete distributions 53
- 4.2 Random permutations 54
- 4.3 Random partitions 56
- 4.4 Random graphs 58
- 4.5 Random trees 60
- 4.6 Further reading 62

5 Gibbs measures 65
5.1 Gibbs measures 65
5.2 The Metropolis–Hastings algorithm 67
5.3 The simulated annealing algorithm 70
5.4 The Propp–Wilson algorithm 72
5.5 Modeling a chemical compound 74
5.6 Further reading 76

6 Internal diffusion limited aggregation 79
6.1 A diffusion aggregation model 79
6.2 The one-dimensional model 80
6.3 Almost sure convergence 81
6.4 Convergence in distribution 82
6.5 A concentration inequality 85
6.6 Higher dimensions 85
6.7 Further reading 87

7 Hidden Markov chains 89
7.1 The forward-backward algorithm 89
7.2 The Kalman filter 94
7.3 Further reading 98

8 The EM algorithm and mixtures 99
8.1 A model for gull nests 99
8.2 A favorable but unrealistic situation 101
8.3 Solving the actual problem 102
8.4 Further reading 105

9 Ehrenfest urns 107
9.1 The microscopic model 107
9.2 The Ehrenfest urn 115
9.3 Pressure over time 116
9.4 Fluctuations around the mean at equilibrium 118
9.5 When will the urn be empty again? 119
9.6 Further reading 121

10 Records, extremes and recruitment 123
10.1 Elitism 123
10.2 Above average 126
10.3 The exponential distribution case 127
10.4 The Pareto distribution case 130
10.5 Further reading 131

11 M/M/Infinity queue 135
11.1 Exponential distributions 137
11.2 Half-life of the initial customers 139
11.3 Distribution of the number of customers 141
11.4 Long-time behavior 145
11.5 Further reading 148

12 Wright–Fisher model 151
12.1 Hardy–Weinberg law 151
12.2 Moran model 152
12.3 Wright–Fisher model and fixation 155
12.4 Wright–Fisher model with mutation 159
12.5 Wright–Fisher model with selection 161
12.6 Cannings model 162
12.7 Further reading 165

13 Genealogies and coalescence 167
13.1 Continuous-time genealogical model 170
13.2 Length of the genealogical tree 172
13.3 Mutations 174
13.4 Kingman's coalescent 177
13.5 Further reading 179

14 Chinese restaurants 181
14.1 Ewens distributions 183
14.2 Number of tables 185
14.3 Extreme tables 189
14.4 Additional combinatorics 191
14.5 Further reading 192

15 Reinforcement 195
15.1 Pólya's urn 195
15.2 Barabási–Albert graph 199
15.3 Reinforced random walk 202
15.4 Rubin's theorem 206
15.5 Further reading 208

16 Percolation 211
16.1 Percolation on a graph 211
16.2 Bethe graph (regular tree) 213
16.3 Euclidean graph (lattice) 216
16.4 Complete graph and Erdős–Rényi model 220
16.5 Further reading 222

17 Growth and fragmentation . . . 225
17.1 The TCP window-size process in computing . . . 225
17.2 Position-dependent jump intensity . . . 232
17.3 Branching and growth-fragmentation . . . 233
17.4 Further reading . . . 236

18 Ruin of an insurance company . . . 239
18.1 Compound Poisson process . . . 241
18.2 Is ruin almost sure? . . . 242
18.3 Expression for the probability of ruin . . . 243
18.4 Exponential claims . . . 244
18.5 Further reading . . . 246

19 Directed polymers in a random environment . . . 247
19.1 Model and main results . . . 247
19.2 Normalized partition function . . . 251
19.3 A lower bound in high dimensions . . . 254
19.4 Mean free energy . . . 255
19.5 An upper bound on the critical parameter . . . 257
19.6 Further reading . . . 259

20 The traveling salesman problem . . . 261
20.1 Concentration in the uniform case . . . 263
20.2 Estimating the mean in the uniform case . . . 267
20.3 Proof in the uniform case . . . 271
20.4 Further reading . . . 271

21 Random matrices . . . 273
21.1 Wigner's theorem . . . 274
21.2 Reduction to a simpler case . . . 277
21.3 Convergence of moments and weak convergence . . . 281
21.4 Proof of the simplified Wigner theorem . . . 284
21.5 Further reading . . . 288

22 Births and assassinations . . . 291
22.1 A sufficient condition for stability . . . 293
22.2 A sufficient condition for instability . . . 294
22.3 Further reading . . . 297

23 The telegraph model . . . 299
23.1 Markov aspects of the full process . . . 300
23.2 Connecting the equation and the process . . . 303
23.3 An ergodic modification . . . 304
23.4 Further reading . . . 307

24 The Dirichlet problem 309
24.1 Probabilistic interpretation 311
24.2 The unit ball 313
24.3 Numerical approaches 314
24.4 The upper half-plane: from discrete to continuous 315
24.5 Recurrence, transience, and harmonic functions 317
24.6 Further reading 318

25 The Ornstein-Uhlenbeck process 319
25.1 Basic properties 319
25.2 Spectral decomposition and the Poincaré inequality 325
25.3 Entropy, coupling, and long-time behavior 329
25.4 The logarithmic Sobolev inequality 331
25.5 Further reading 335

26 Kinetic diffusion models 337
26.1 Langevin model 337
26.2 Confined Langevin process 339
26.3 Long-time behaviour of the confined process 343
26.4 Further reading 347

27 From Markov chains to diffusion processes 349
27.1 Laziness and sampling 349
27.2 Convergence of a chain to a diffusion 351
27.3 Diffusion on an interval 354
27.4 Application to Wright-Fisher processes 356
27.5 Empirical distribution function 359
27.6 Euler scheme 361
27.7 Further reading 363

Suggested reading 365

References 367

Index 381

Main notation and abbreviations 389

1

Heads, tails, and coupons

Keywords. Combinatorics; discrete distribution; total variation distance; law of small numbers; cutoff; Gumbel distribution. **Tools.** Law of large numbers; central limit theorem; Markov inequality. **Difficulty.** *

This chapter concerns *coin tossing* and the *coupon collector problem*, two important probabilistic objects that are worth knowing and recognizing. It also studies total variation distance for discrete probability measures.

1.1 Coin tossing

Coin tossing consists of successive tosses of a coin, each resulting in either heads (success, encoded as 1) or tails (failure, encoded as 0). We model this by a sequence $(X_n)_{n\geqslant 1}$ of independent identically distributed random variables with Bernoulli distribution $\mathrm{Ber}(p)$:

$$\mathbb{P}(X_n = 1) = 1 - \mathbb{P}(X_n = 0) = p \in [0, 1].$$

The number of successes in the first n tosses,

$$S_n = X_1 + \cdots + X_n$$

is a random variable taking values in $\{0, 1, \ldots, n\}$ with binomial distribution $\mathrm{Bin}(n, p)$ of size n and parameter p, given, for every $k = 0, 1, \ldots, n$, by

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

The mean and variance of S_n are

$$\mathbb{E}(S_n) = np \quad \text{et} \quad \mathrm{Var}(S_n) = np(1-p).$$

The time T of the first success, also the number of tosses needed for the first success, is

$$T := \inf\{n \geqslant 1 : X_n = 1\}$$

If $p > 0$, then T has the geometric distribution $\mathrm{Geo}(p)$ on $\mathbb{N}^*$ with parameter p given by

$$\mathbb{P}(T = k) = (1-p)^{k-1}p, \quad k \in \mathbb{N}^*.$$

We have $T \equiv \infty$ if $p = 0$ and $\mathbb{P}(T < \infty) = 1$ otherwise. Moreover,

$$\mathbb{E}(T) = \frac{1}{p} \quad \text{et} \quad \mathrm{Var}(T) = \frac{1-p}{p^2}.$$

The number of failures before the first success,

$$T' := \inf\{n \geqslant 0 : X_{n+1} = 1\} = T - 1$$

has the geometric distribution $\mathrm{Geo}_{\mathbb{N}}(p)$ on $\mathbb{N}$ with parameter p, given by

$$\mathbb{P}(T' = k) = \mathbb{P}(T - 1 = k) = (1-p)^k p, \quad k \in \mathbb{N},$$

and

$$\mathbb{E}(T') = \mathbb{E}(T) - 1 = \frac{1-p}{p} \quad \text{et} \quad \mathrm{Var}(T') = \mathrm{Var}(T) = \frac{1-p}{p^2}.$$

For every $r \in \mathbb{N}^*$, the number of tosses T_r needed to obtain r successes is recursively defined by

$$T_1 := T \quad \text{et} \quad T_{r+1} := \inf\{n > T_r : X_n = 1\}.$$

The random variables $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are independent and identically distributed with geometric distribution $\mathrm{Geo}(p)$. The random variable T_r has the Pascal, or negative binomial, distribution $\mathrm{Geo}(p)^{*r}$. For every $k \geqslant r$,

$$\mathbb{P}(T_r = k) = \sum_{\substack{k_1 \geqslant 1, \ldots, k_r \geqslant 1 \\ k_1 + \cdots + k_r = k}} (1-p)^{k_1 - 1}p \cdots (1-p)^{k_r - 1}p = (1-p)^{k-r}p^r \binom{k-1}{r-1}$$

and

$$\mathbb{E}(T_r) = r\mathbb{E}(T) = \frac{r}{p} \quad \text{et} \quad \mathrm{Var}(T_r) = r\mathrm{Var}(T) = r\frac{1-p}{p^2}.$$

The *Bernoulli process* $(S_n)_{n \geqslant 0}$ has piecewise constant trajectories with jumps of size $+1$ at times $(T_r)_{r \geqslant 1}$ (i.i.d. geometric interarrival times). It counts events separated by independent identically distributed geometric durations, and is the discrete analogue of the *Poisson process*. Since S_n is a sum of independent variables, $(S_n)_{n \geqslant 0}$ is a Markov chain on $\mathbb{N}$ with kernel $\mathbf{P}(x, y) = p\mathbb{1}_{y=x+1} + (1-p)\mathbb{1}_{y=x}$, and $(S_n - np)_{n \geqslant 0}$ is a *martingale*.

By the *law of large numbers* (LLN) and the *central limit theorem* (CLT),

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\text{p.s.}} p \quad \text{et} \quad \frac{\sqrt{n}}{\sqrt{p(1-p)}}\left(\frac{S_n}{n} - p\right) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

This allows us, in particular, to construct an asymptotic confidence interval for p called the *Wald interval*, which is nevertheless not very accurate.

Remark 1.1 (The Clopper–Pearson interval)**.** *Nonasymptotic confidence intervals for p can also be constructed, such as the Clopper–Pearson interval, based on the beta-binomial correspondence. Let $U_1, \dots, U_n$ be i.i.d. real random variables uniformly distributed on $[0,1]$, and let $U_{(1,n)} \leqslant \cdots \leqslant U_{(n,n)}$ be their increasing rearrangement. For every $k = 1, \dots, n$, the variable $U_{(k,n)}$ has distribution* $\mathrm{Beta}(k, n-k+1)$ *on* $[0,1]$ *with density*

$$t \in [0,1] \mapsto \frac{t^{k-1}(1-t)^{n-k}}{\mathrm{Beta}(k, n-k+1)}, \quad \text{où} \quad \mathrm{Beta}(a,b) := \int_0^1 s^{a-1}(1-s)^{b-1}\, ds,$$

and

$$\mathbb{P}(S_n \geqslant k) = \mathbb{P}(\mathbb{1}_{\{U_1 \leqslant p\}} + \cdots + \mathbb{1}_{\{U_n \leqslant p\}} \geqslant k) = \mathbb{P}(U_{(k,n)} \leqslant p).$$

This identity gives an exact expression for the probability that S_n lies in a given interval. Since S_n is discrete, its quantiles are also discrete, preventing an exact interval of arbitrary level $\alpha \in [0,1]$ and suggesting a smoothing procedure.

Remark 1.2 (Repeated patterns and zero-one laws)**.** *When $0 < p < 1$, the* Borel–Cantelli lemma *(independent case) implies that every finite string of* 0 *and* 1 *appears almost surely infinitely often in the sequence $X_1, X_2, \dots$ Independence is crucial. An immortal monkey typing at a keyboard will eventually write the complete works of William Shakespeare! Alternatively, this follows from the geometric distribution of the first success time in a coin-tossing process obtained by dividing the original sequence into successive blocks of the same length as the desired string.*

Remark 1.3 (Coin tossing and the uniform distribution on $[0,1]$)**.** *If U is a random variable on $[0,1]$ and $U = \sum_{n=1}^{\infty} b_n 2^{-n}$ is its base-2 expansion, then U is uniformly distributed on $[0,1]$ if and only if the bits $(b_n)_{n \geqslant 1}$ in its base-2 expansion are i.i.d. Bernoulli random variables with parameter $1/2$.*

Remark 1.4 (Von Neumann's debiasing algorithm)**.** *Let $(X_n)_{n \geqslant 1}$ be a sequence of independent Bernoulli random variables with unknown parameter $0 < p < 1$. Construct the sequence $(Y_n)_{n \geqslant 1}$ of independent identically distributed random variables on $\{0,1,2\}$ as follows:*

$$\underbrace{X_1 X_2}_{Y_1} \underbrace{X_3 X_4}_{Y_2} \cdots$$

setting $Y_n = 0$ if $(X_{2n-1}, X_{2n}) = (0,1)$, $Y_n = 1$ if $(X_{2n-1}, X_{2n}) = (1,0)$, and $Y_n = 2$ otherwise. The sequence $(Z_n)_{n \geqslant 1}$ obtained from $(Y_n)_{n \geqslant 1}$ by deleting the 2 consists of independent Bernoulli random variables with parameter $1/2$. Producing each term of $(Z_n)_{n \geqslant 1}$ requires a geometric random number of terms of $(X_n)_{n \geqslant 1}$.

1.2 The binomial-Gaussian approximation

Let $(X_n)_{n\geqslant 1}$ be i.i.d. real random variables with mean m and variance $0 < \sigma^2 < \infty$, and let $S_n = X_1 + \cdots + X_n$. The central limit theorem states that, for every $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - nm}{\sqrt{n}\sigma} \leqslant t\right) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

The Berry–Esseen central limit theorem refines this asymptotic result by providing a uniform quantitative bound: if $X_1, X_2, \ldots$ are Bernoulli with parameter $p \in]0,1[$, then $m := \mathbb{E}(X_1) = p$, $\sigma^2 := \mathbb{E}((X-m)^2) = p(1-p)$, $\tau^3 := \mathbb{E}(|X_1 - m|^3) = p(1-p)(1-2p(1-p))$, and

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\Big(S_n \leqslant \sqrt{np(1-p)}t + np\Big) - \int_{-\infty}^{t} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\, dx \right| \leqslant \frac{\tau^3}{\sqrt{n}\sigma^3} = \frac{1-2p(1-p)}{\sqrt{p(1-p)}\sqrt{n}}.$$

This Gaussian approximation to the binomial distribution improves as $(1-2p(1-p))/\sqrt{np(1-p)}$ decreases. For fixed n, the bound is minimal at $p = 1/2$ but diverges as p approaches 0 or 1. In these extreme regimes, a Poisson approximation, discussed in Section 1.4, is more appropriate.

When $X_1, X_2, \ldots$ are Bernoulli variables with parameter $p \in]0,1[$, the CLT suggests approximating the binomial distribution $\mathrm{Bin}(n,p)$ of S_n by the Gaussian distribution $\mathcal{N}(nm, n\sigma^2)$ for large n. The *de Moivre–Laplace central limit theorem* states more precisely that, for every $-\infty < a < b < +\infty$,

$$\lim_{n\to\infty} \sqrt{n} \sup_{k\in I_n(a,b)} \left| \mathbb{P}(S_n = k) - \frac{\exp\left(-\frac{(k-np)^2}{2np(1-p)}\right)}{\sqrt{2\pi np(1-p)}} \right| = 0$$

where

$$I_n(a,b) = \left\{ 0 \leqslant k \leqslant n : \frac{k-np}{\sqrt{np(1-p)}} \in [a,b] \right\},$$

which recovers the CLT upon integration:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \in [a,b] \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\, dx.$$

1.3 Total variation distance

Throughout this section, E is an at most countable set equipped with the discrete topology and sigma-algebra $\mathcal{P}(E)$. The set of probability distributions on E is a complete metric space for the *total variation distance*

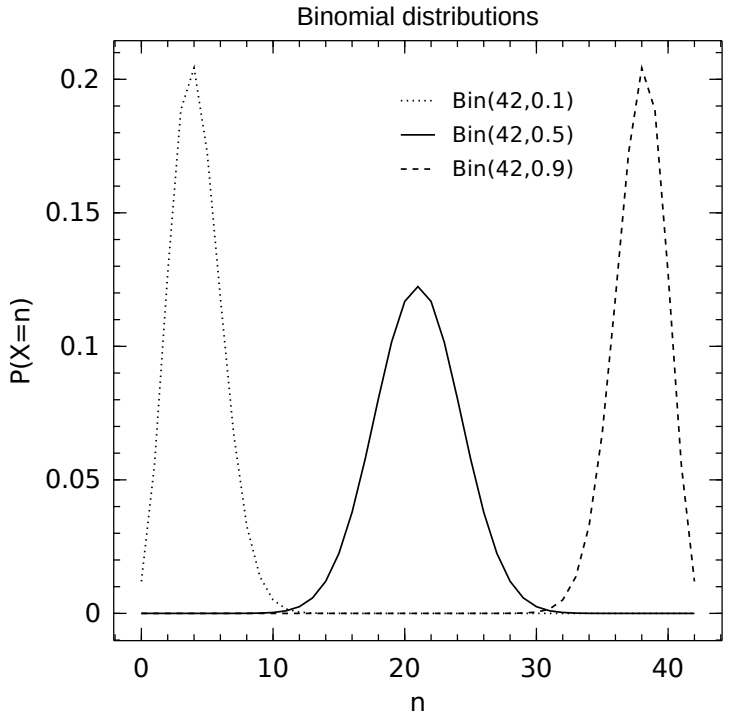


Fig. 1.1. Binomial distributions of the same size, for three values of the second parameter, illustrating when Gaussian and Poisson approximations to the binomial distribution are appropriate.

$$d_{\mathrm{VT}}(\mu,\nu) := \sup_{A\subset E} |\mu(A)-\nu(A)|.$$

Note that $d_{\mathrm{VT}}(\mu,\nu) = \|\mu-\nu\|_{\mathrm{VT}}$, where $\|\eta\|_{\mathrm{VT}} := \sup_{A\subset E}|\eta(A)|$ for every finite signed measure η on E. We have $0 \leqslant d_{\mathrm{VT}}(\mu,\nu) \leqslant 1$, since the diameter of $[0,1]$ is 1. Moreover, $d_{\mathrm{VT}}(\mu,\nu) = 1$ if μ and ν have disjoint supports.

Theorem 1.5 (Alternative expressions)**.** *If μ and ν are probability distributions on E, then*[1]

$$d_{\mathrm{VT}}(\mu,\nu) = \frac{1}{2}\sup_{f:E\to[-1,1]}\left|\int f\,d\mu - \int f\,d\nu\right| = \frac{1}{2}\sum_{x\in E}|\mu(x)-\nu(x)|.$$

Moreover, the supremum in the definition of $d_{\mathrm{VT}}(\cdot,\cdot)$ is attained at the set

$$A_* = \{x\in E : \mu(x) \geqslant \nu(x)\},$$

[1] In particular, $2\|\cdot\|_{\mathrm{VT}} = \|\cdot\|_{\ell^1(E,\mathbb{R})}$.

and the supremum in the functional variational expression for $d_{\mathrm{VT}}(\cdot,\cdot)$ *is attained at*

$$f = \mathbb{1}_{A_*} - \mathbb{1}_{A_*^c}.$$

Proof. The second equality follows from the inequality

$$\left|\int f\,d\mu - \int f\,d\nu\right| \leqslant \sum_{x\in E} |f(x)||\mu(x)-\nu(x)| \leqslant \sup_{x\in E}|f(x)| \sum_{x\in E}|\mu(x)-\nu(x)|$$

with equality for $f = \mathbb{1}_{A_*} - \mathbb{1}_{A_*^c}$. For the first equality, write

$$|\mu(A)-\nu(A)| = \frac{1}{2}\left|\int f_A\,d\mu - \int f_A\,d\nu\right|$$

where $f_A := \mathbb{1}_A - \mathbb{1}_{A^c}$, giving

$$|\mu(A)-\nu(A)| \leqslant \frac{1}{2}\sup_{f:E\to[-1,1]}\left|\int f\,d\mu - \int f\,d\nu\right| = \frac{1}{2}\sum_{x\in E}|\mu(x)-\nu(x)|$$

with equality for $A = A_*$, since by definition of A_*,

$$\begin{aligned} 2|\mu(A_*)-\nu(A_*)| &= \mu(A_*)-\nu(A_*)+\nu(A_*^c)-\mu(A_*^c) \\ &= \sum_{x\in A_*}|\mu(x)-\nu(x)| + \sum_{x\in A_*^c}|\mu(x)-\nu(x)|. \end{aligned}$$

□

Theorem 1.6 (Convergence in distribution)**.** *If* $(X_n)_{n\geqslant 1}$ *are random variables on* E *and* μ_n *denotes the distribution of* X_n*, then for every distribution* μ *on* E*, the following are equivalent:*

1. $\lim_{n\to\infty}\int f\,d\mu_n = \int f\,d\mu$ *for every bounded function* $f : E \to \mathbb{R}$*;*
2. $\lim_{n\to\infty}\mu_n(x) = \mu(x)$ *for every* $x \in E$*;*
3. $\lim_{n\to\infty} d_{\mathrm{VT}}(\mu_n,\mu) = 0$.

When these hold, (X_n) *is said to converge in distribution to* μ *as* $n\to\infty$.

Every function $E \to \mathbb{R}$ is continuous for the discrete topology on E.

Proof. To deduce *1.* from *3.*, use the functional variational expression for $d_{\mathrm{VT}}(\cdot,\cdot)$. To deduce *2.* from *1.*, take $f = \mathbb{1}_{\{x\}}$. To deduce *3.* from *2.*, observe that, for every $A \subset E$,

$$\sum_{x\in E}|\mu_n(x)-\mu(x)| = \sum_{x\in A}|\mu_n(x)-\mu(x)| + \sum_{x\in A^c}|\mu_n(x)-\mu(x)|.$$

By *2.*, if A is finite, then for every $\varepsilon > 0$ there exists an integer $N = N(A,\varepsilon)$ such that the first term on the right is bounded by ε for every $n \geqslant N$. The second term can be bounded as follows:

$$\sum_{x\in A^c} |\mu_n(x) - \mu(x)| \leqslant \sum_{x\in A^c} \mu_n(x) + \sum_{x\in A^c} \mu(x).$$

Since

$$\sum_{x\in A^c} \mu_n(x) = \sum_{x\in A} \mu(x) - \sum_{x\in A} \mu_n(x) + \sum_{x\in A^c} \mu(x),$$

we obtain

$$\sum_{x\in A^c} |\mu_n(x) - \mu(x)| \leqslant \sum_{x\in A} |\mu_n(x) - \mu(x)| + 2\sum_{x\in A^c} \mu(x).$$

Since $\mu \in \mathcal{P}$, for every $\varepsilon' > 0$ we can choose a finite A such that $\mu(A^c) \leqslant \varepsilon'$. Hence

$$\lim_{n\to\infty} \sum_{x\in E} |\mu_n(x) - \mu(x)| = 0,$$

which is precisely *3.*, by Theorem 1.5. □

Remark 1.7 (Escape of mass to infinity)**.** *If $(\mu_n)_{n\geqslant 1}$ are probability distributions on E and $\mu(x) := \lim_{n\to\infty} \mu_n(x)$, then μ need not be a probability distribution unless E is finite. When E is infinite, mass may escape to infinity. A counterexample is $E = \mathbb{N}$ with μ_n assigning mass $1/n$ to the singletons $\{1\}, \ldots, \{n\}$, yielding μ identically zero.*

Theorem 1.8 (Minimum)**.** *If μ and ν are probability distributions on E, then*

$$d_{\mathrm{VT}}(\mu, \nu) = 1 - \sum_{x\in E} (\mu(x) \wedge \nu(x)).$$

In particular, $d_{\mathrm{VT}}(\mu, \nu) = 1$ if and only if μ and ν have disjoint supports.

Proof. Recall that $a \wedge b = \min(a, b)$. Simply write

$$\sum_{x\in E} (\mu(x) \wedge \nu(x)) = \frac{1}{2} \sum_{x\in E} (\mu(x) + \nu(x) - |\mu(x) - \nu(x)|) = 1 - d_{\mathrm{VT}}(\mu, \nu).$$

□

Theorem 1.9 (Coupling)**.** *If μ and ν are probability distributions on E, then*

$$d_{\mathrm{VT}}(\mu, \nu) = \inf_{(X,Y)} \mathbb{P}(X \neq Y)$$

where the infimum is over pairs of random variables on $E \times E$ with marginal distributions μ and ν. Moreover, equality is attained for some such pair, so the infimum is a minimum.

Theorem 1.9 is often used as follows: exploit the setting to explicitly construct a pair (X, Y) of random variables with $X \sim \mu$ and $Y \sim \nu$, and deduce that $d_{\mathrm{VT}}(\mu, \nu) \leqslant \mathbb{P}(X \neq Y)$.

Proof. Let (X, Y) be a pair of random variables on $E \times E$ with marginal distributions μ and ν. Since $\mathbb{P}(X = x, Y = x) \leqslant \mu(x) \wedge \nu(x)$ for every $x \in E$,

$$1 - d_{\mathrm{VT}}(\mu, \nu) = \sum_{x \in E} (\mu(x) \wedge \nu(x)) \geqslant \sum_{x \in E} \mathbb{P}(X = x, Y = x) = \mathbb{P}(X = Y).$$

It therefore suffices to construct a pair (X, Y) attaining equality. Set $p = 1 - d_{\mathrm{VT}}(\mu, \nu) \in [0, 1]$ and distinguish three cases.

- *Case* $p = 0$. Take (X, Y) with X and Y independent, with respective distributions μ and ν. Since $d_{\mathrm{VT}}(\mu, \nu) = 1$, μ and ν have disjoint supports, hence $\mathbb{P}(X = Y) = \sum_{x \in E} \mu(x)\nu(x) = 0$;
- *Case* $p = 1$. Then $d_{\mathrm{VT}}(\mu, \nu) = 0$, hence $\mu = \nu$. Take (X, Y), where $X \sim \mu$ and $Y = X$;
- *Case* $0 < p < 1$. Let U, V, W be random variables with distributions

  $$p^{-1}(\mu \wedge \nu), \quad (1-p)^{-1}(\mu - (\mu \wedge \nu)), \quad (1-p)^{-1}(\nu - (\mu \wedge \nu)).$$

  Note that $p = \sum_{x \in E} (\mu(x) \wedge \nu(x))$ (Theorem 1.8). Let B be a Bernoulli random variable with distribution $\mathrm{Ber}(p)$, independent of the triple (U, V, W). Set $(X, Y) = (U, U)$ if $B = 1$ and $(X, Y) = (V, W)$ if $B = 0$. Then $X \sim \mu$ and $Y \sim \nu$. Since the distributions of V and W have disjoint supports, $\mathbb{P}(V = W) = 0$, and therefore $\mathbb{P}(X = Y) = \mathbb{P}(B = 1) = p$.

□

1.4 The binomial-Poisson approximation

If S_n has the binomial distribution $\mathrm{Bin}(n, p)$, then for every $k \in \mathbb{N}$,

$$\mathbb{P}(S_n = k) - e^{-np}\frac{(np)^k}{k!} = \left(\frac{n}{n(1-p)} \cdots \frac{n-k+1}{n(1-p)}(1-p)^n - e^{-np}\right)\frac{(np)^k}{k!}.$$

This shows that, if p depends on n with $\lim_{n\to\infty} np = \lambda$, then the distribution of S_n tends to $\mathrm{Poi}(\lambda)$. Total variation distance quantifies this convergence in distribution: *Le Cam's inequality* in Theorem 1.10 below gives

$$\sum_{k=0}^{\infty} \left| \mathbb{P}(S_n = k) - e^{-np}\frac{(np)^k}{k!} \right| \leqslant 2np^2,$$

which is useful when np^2 is small. For example, if $p = \lambda/n$, then $np = \lambda$ and $np^2 = \lambda^2/n$.

Theorem 1.10 (Le Cam's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with Bernoulli distributions $\mathrm{Ber}(p_1), \ldots, \mathrm{Ber}(p_n)$. Let μ_n be the distribution of $S_n = X_1 + \cdots + X_n$, and let $\nu_n = \mathrm{Poi}(p_1 + \cdots + p_n)$ be the Poisson distribution with the same mean as S_n. Then*

$$d_{\mathrm{VT}}(\mu_n, \nu_n) \leqslant p_1^2 + \cdots + p_n^2.$$

Proof. Induction on n shows that, if $\alpha_1, \dots, \alpha_n$ and $\beta_1, \dots, \beta_n$ are probability distributions on $\mathbb{N}$, then the subadditivity inequality holds:

$$d_{\mathrm{VT}}(\alpha_1 * \cdots * \alpha_n, \beta_1 * \cdots * \beta_n) \leqslant d_{\mathrm{VT}}(\alpha_1, \beta_1) + \cdots + d_{\mathrm{VT}}(\alpha_n, \beta_n).$$

Next establish $d_{\mathrm{VT}}(\mathrm{Ber}(p), \mathrm{Poi}(p)) \leqslant p^2$ and use the convolution semigroup property of Poisson distributions: $\mathrm{Poi}(a) * \mathrm{Poi}(b) = \mathrm{Poi}(a+b)$. □

The following $\ell^2 - \ell^1 - \ell^\infty$ inequality,

$$p_1^2 + \cdots + p_n^2 \leqslant (p_1 + \cdots + p_n) \max_{1 \leqslant k \leqslant n} p_k$$

recovers the *law of small numbers*: if $(X_{n,k})_{1 \leqslant k \leqslant n}$ is a triangular array of independent Bernoulli random variables with $X_{n,k} \sim \mathrm{Ber}(p_{n,k})$ for every $n \geqslant k \geqslant 1$, and if

$$\lim_{n \to \infty} p_{n,1} + \cdots + p_{n,n} = \lambda \quad \text{et} \quad \lim_{n \to \infty} \max_{1 \leqslant k \leqslant n} p_{n,k} = 0,$$

then $X_{n,1} + \cdots + X_{n,n}$ converges in distribution to $\mathrm{Poi}(\lambda)$ as $n \to \infty$.

1.5 The coupon collector problem

The *coupon collector problem* belongs in the same toolkit as coin tossing and dice rolling, to which it is closely related. Many concrete situations conceal a coupon collector or one of its variants. We consider only its simplest version.

A geometric random number of coin tosses is needed to see both sides of the coin. If we replace the coin with an $r \geqslant 2$-sided die, how many rolls are needed to see all r distinct faces? For a fixed integer $r \geqslant 2$, we model this by the random variable

$$\begin{aligned} T &:= \min\{n \geqslant 1 : \{X_1, \dots, X_n\} = \{1, \dots, r\}\} \\ &= \min\{n \geqslant 1 : \mathrm{card}\{X_1, \dots, X_n\} = r\} \end{aligned}$$

where $(X_n)_{n \geqslant 1}$ is an i.i.d. sequence uniformly distributed on $\{1, \dots, r\}$. The random variable T is the first time all r faces have appeared. This time of course depends on r, but we suppress this dependence in the notation for simplicity. The name "coupon collector" comes from collectible coupons included in some cereal packages.

Theorem 1.11 (Combinatorics)**.** *We have $T \geqslant r$ and, for every $n \geqslant r$,*

$$\mathbb{P}(T = n) = \frac{r!}{r^n} \left\{ {n-1 \atop r-1} \right\}$$

where the notation in braces denotes the Stirling number of the second kind*, the number of ways to partition a set of $n-1$ elements into $r-1$ nonempty subsets.*

Proof. We have $X_T \notin \{X_1, \dots, X_{T-1}\}$ because the coupon completing the collection cannot have appeared previously. Fixing $n \geqslant r$, the event $\{T = n\}$ corresponds to choosing the type of the last coupon and assigning the remaining $n-1$ coupons to the remaining $r-1$ types. The result follows because the types are uniformly distributed. □

Although explicit, Theorem 1.11 is not very illuminating. The following intuitive result greatly helps the study of T.

Lemma 1.12 (Decomposition)**.** *We have $T = G_1 + \cdots + G_r$, where $G_1, \dots, G_r$ are independent random variables with*

$$G_i \sim \mathrm{Geo}(\pi_i), \quad \pi_i := \frac{r-i+1}{r}, \quad 1 \leqslant i \leqslant r.$$

In particular, $\mathbb{P}(T < \infty) = 1$, and moreover

$$\mathbb{E}(T) = r(\log(r) + \gamma) + o_{r\to\infty}(r) \quad \textit{et} \quad \mathrm{Var}(T) = \frac{\pi^2}{6} r^2 + o_{r\to\infty}(r^2),$$

where $\gamma = \lim_{n\to\infty}(\sum_{i=1}^{n} 1/i - \log(n)) \approx 0.577$ is Euler's constant.

Proof. Set $G_1 \equiv 1$ and, for every $1 < i \leqslant r$,

$$G_i = \min\{n \geqslant 1 : X_{G_{i-1}+n} \notin \{X_1, \dots, X_{G_{i-1}}\}\}.$$

We have $\mathrm{card}(\{X_1, \dots, X_{G_i}\}) = i$ for every $1 \leqslant i \leqslant n$. The random variables $G_1, G_1 + G_2, \dots, G_1 + \cdots + G_r$ are the times of the first r successes in a special coin-tossing process whose success probability changes after each success, taking successive values

$$\pi_1 = 1, \ \pi_2 = \frac{r-1}{r}, \ \pi_3 = \frac{r-2}{r}, \ \dots, \ \pi_r = \frac{1}{r}.$$

This decreasing sequence reflects the growing difficulty of obtaining a new coupon type as the collection progresses.

Now compute the moments of T. Linearity of expectation gives

$$\begin{aligned} \mathbb{E}(T) &= \sum_{i=1}^{r} \mathbb{E}(G_i) = \sum_{i=1}^{r} \frac{1}{\pi_i} \\ &= \sum_{i=1}^{r} \frac{r}{r-i+1} = r \sum_{i=1}^{r} \frac{1}{i} = r(\log(r) + \gamma + o(1)). \end{aligned}$$

Independence of the random variables $G_1, \dots, G_r$ (exercise!) gives

$$\begin{aligned} \mathrm{Var}(T) &= \sum_{i=1}^{r} \mathrm{Var}(G_i) = \sum_{i=1}^{r} \frac{1-\pi_i}{\pi_i^2} \\ &= r \sum_{i=1}^{r-1} \frac{r-i}{i^2} = \frac{\pi^2}{6} r^2 - r(\log(r) + \gamma) + o(r^2). \end{aligned}$$

□

Theorem 1.13 (Distribution tail)**.** *For every $n \geqslant 1$,*

$$\mathbb{P}(T > n) = \sum_{k=1}^{r} (-1)^{k-1} \binom{r}{k} \left(1 - \frac{k}{r}\right)^n.$$

Proof. We have

$$\{T > n\} = E_{n,1} \cup \cdots \cup E_{n,r} \quad \text{où} \quad E_{n,i} := \{X_1 \neq i, \ldots, X_n \neq i\}.$$

If $i_1, \ldots, i_k \in \{1, \ldots, r\}$ are pairwise distinct, then, writing

$$R = \{1, \ldots, r\} \setminus \{i_1, \ldots, i_k\},$$

we have

$$\begin{aligned} \mathbb{P}(E_{n,i_1} \cap \cdots \cap E_{n,i_k}) &= \mathbb{P}(X_1 \in R) \cdots \mathbb{P}(X_n \in R) \\ &= \left(\frac{r-k}{r}\right)^n = \left(1 - \frac{k}{r}\right)^n. \end{aligned}$$

Finally use the *inclusion-exclusion principle*, or *Poincaré sieve*[2]. □

Because of the alternating signs, it is difficult to deduce the tail behavior of T as a function of n and r from Theorem 1.13.

Theorem 1.14 (Deviation)**.** *For every real $t > 0$,*

$$\mathbb{P}(T > rt + r\log(r)) = \mathbb{P}\left(\frac{T - r\log(r)}{r} > t\right) \leqslant e^{-t+\varepsilon_r}$$

where $\varepsilon_r = (rt + r\log(r) - \lfloor rt + r\log(r) \rfloor)/r \in [0, 1/r[$.

Proof. For every integer $n \geqslant 1$, write

$$\mathbb{P}(T > n) = \mathbb{P}(\cup_{i=1}^{r} E_{n,i}) \leqslant \sum_{i=1}^{r} \mathbb{P}(E_{n,i}) \quad \text{où} \quad E_{n,i} := \{X_1 \neq i, \ldots, X_n \neq i\}.$$

Since $\mathbb{P}(E_{n,i}) = (1 - 1/r)^n \leqslant e^{-n/r}$, choose $n = \lfloor rt + r\log(r) \rfloor$. □

For fixed $\alpha > 0$ and r, choose t large enough that $e^{-t+\varepsilon_r} \leqslant \alpha$, for example $t = -\log(\alpha) + 1/r$. This gives the nonasymptotic prediction interval $[r, r\log(r) - r\log(\alpha) + 1]$ of level $1 - \alpha$ for T.

Theorem 1.15 (Asymptotic behavior)**.**

$$\frac{T}{r\log(r)} \xrightarrow[r\to\infty]{\mathbb{P}} 1.$$

[2] $\mathbb{P}(\cup_{k=1}^{n} A_k) = \sum_{j=1}^{n} (-1)^{j+1} \sum_{k_1 < \ldots < k_j} \mathbb{P}(A_{k_1} \cap \cdots \cap A_{k_j})$.

Proof. Fix $\varepsilon > 0$. By Markov's inequality and Lemma 1.12,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{T}{r\log(r)} - 1\right| > \varepsilon\right) &\leqslant \frac{\mathbb{E}((T - r\log(r))^2)}{\varepsilon^2 r^2 \log(r)^2} \\ &= \frac{\mathrm{Var}(T) + (\mathbb{E}(T) - r\log(r))^2}{\varepsilon^2 r^2 \log(r)^2} \\ &= \mathcal{O}\left(\frac{1}{\log(r)^2}\right).\end{aligned}$$

□

This is a proof by the *second moment method*. In the same spirit, a fourth moment method is used for Theorem 11.6.

The bound established in the preceding proof is not summable in r and therefore cannot prove almost sure convergence through Borel–Cantelli, which would only make sense on a single probability space valid for every r. On the other hand, this bound yields a nonasymptotic prediction interval: for $\alpha = 0.05$, fixed r, and suitably chosen t, we have $\mathbb{P}(|T - r\log(r)|/r > t) = \mathcal{O}(1/t^2) = \alpha$. The prediction interval has width $2rt$ and deteriorates as t increases (α decreases).

The next theorem states that the asymptotic fluctuations in the preceding convergence follow a Gumbel distribution.

Theorem 1.16 (Asymptotic fluctuations)**.** *We have*

$$\frac{T - r\log(r)}{r} = \log(r)\left(\frac{T}{r\log(r)} - 1\right) \xrightarrow[r\to\infty]{\text{loi}} \text{Gumbel}$$

where the Gumbel distribution has distribution function $t \in \mathbb{R} \mapsto e^{-e^{-t}}$.

Figure 1.2 illustrates this result.

Proof. It suffices to show that, for every $t \in \mathbb{R}$,

$$\lim_{r\to\infty} \mathbb{P}(T > r\log(r) + tr) = S(t) := 1 - e^{-e^{-t}}.$$

Fix $t \in \mathbb{R}$ and suppose r is large enough that $r\log(r) + tr > r$. Introduce the integer $n_{t,r} = \lceil r\log(r) + tr \rceil$. Theorem 1.13 gives

$$\mathbb{P}(T > r\log(r) + tr) = \mathbb{P}(T > n_{t,r}) = \sum_{k=1}^{r} (-1)^{k-1} \binom{r}{k} \left(1 - \frac{k}{r}\right)^{n_{t,r}}.$$

Since $\binom{r}{k} \leqslant r^k/k!$ and $1 - u \leqslant e^{-u}$ for every $u \geqslant 0$,

$$\binom{r}{k} \left(1 - \frac{k}{r}\right)^{n_{t,r}} \xrightarrow[r\to\infty]{\leqslant} \frac{e^{-tk}}{k!}.$$

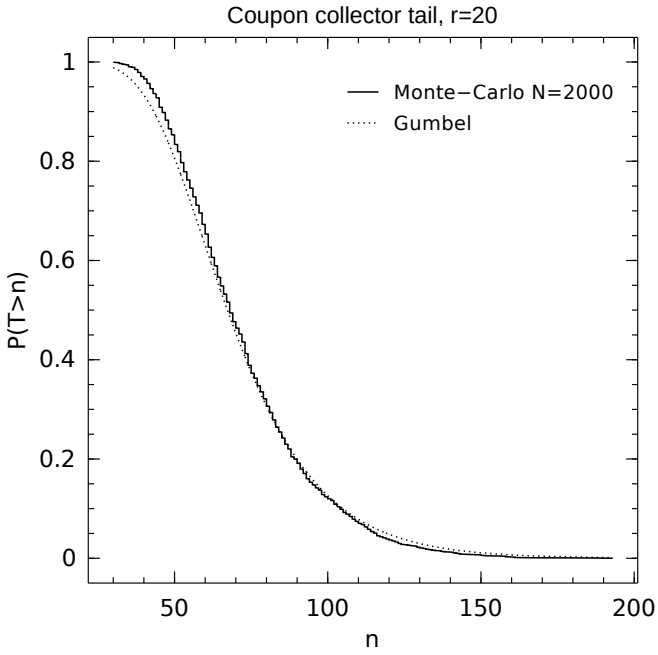


Fig. 1.2. Approximations of the tail of the random variable T by that of $r\log(r) + rG$, where G has the Gumbel distribution, by Theorem 1.16, and by a Monte Carlo method with N samples using the representation of T as a sum of independent geometric random variables.

Finally, dominated convergence gives

$$\lim_{r\to\infty} \sum_{k=1}^{r} (-1)^{k-1} \binom{r}{k} \left(1 - \frac{k}{r}\right)^{n_{t,r}} = \sum_{k=1}^{\infty} (-1)^{k-1} \frac{e^{-tk}}{k!} = S(t).$$

This completes the proof. Alternatively, the theorem can be proved using the exponential approximation of geometric distributions. Set $H_k := \frac{1}{r} G_{r-k+1}$. For every fixed $k \geqslant 1$, the sequence $(H_k)_{r \geqslant k}$ converges in distribution to the exponential distribution with parameter k. The following quantitative version is available for characteristic functions: for every $t \in \mathbb{R}$ and every $k \geqslant 1$, there is a constant $C > 0$ such that, for r large enough,

$$\left|\mathbb{E}(e^{itH_k}) - \mathbb{E}(e^{itE_k})\right| \leqslant \frac{C}{rk}$$

where E_k is an exponential real random variable with parameter k. Using the inequality

$$|a_1 \cdots a_r - b_1 \cdots b_r| \leqslant |a_1 - b_1| + \cdots + |a_r - b_r|, \quad a, b \in \{z \in \mathbb{C} : |z| \leqslant 1\}^r,$$

and considering independent exponential real random variables $E_1, \ldots, E_r$ with respective parameters $1, \ldots, r$ gives

$$\left| \mathbb{E}(e^{it\frac{T}{r}}) - \mathbb{E}(e^{it(E_1+\cdots+E_r)}) \right| \leqslant C\frac{\log(r)}{r}.$$

Rényi's lemma 11.3 states that $E_1 + \cdots + E_r$ has the same distribution as the random variable $\max(F_1, \ldots, F_r)$, where $F_1, \ldots, F_r$ are i.i.d. exponential random variables with parameter 1. Then use convergence in distribution to the Gumbel law of the sequence $(\max(F_1, \ldots, F_r) - \log(r))_{r \geqslant 1}$. □

Theorem 1.16 yields an asymptotic prediction interval for the random variable T: for all real $b \geqslant a$,

$$\lim_{r\to\infty} \mathbb{P}(T \in [r\log(r) + ra, r\log(r) + rb]) = e^{-e^{-b}} - e^{-e^{-a}}.$$

The Gumbel distribution is highly concentrated, and its distribution function rises sharply from 0 to 1. This yields a *threshold phenomenon*[3] for the random variable T. The quantity $\mathbb{P}(T > n)$ changes abruptly from ≈ 1 to ≈ 0 around $n = r\log(r)$ when $r \gg 1$:

Theorem 1.17 (Cutoff at $n = r\log(r)$)**.** *For every real $c > 0$,*

$$\mathbb{P}(T > r\log(r) + rc) \leqslant e^{-c+\varepsilon_r} \quad \text{où} \quad 0 \leqslant \varepsilon_r < 1/r.$$

If $(c_r)_{r\in\mathbb{N}^*}$ is a sequence tending to infinity, then

$$\lim_{r\to\infty} \mathbb{P}(T > r\log(r) - rc_r) = 1.$$

Proof. The first upper bound follows from Theorem 1.14. Theorem 1.16 states that $(T - r\log(r))/r$ converges in distribution to the Gumbel law as $r \to \infty$. Since $c_r \to \infty$, this gives

$$\mathbb{P}(T > r\log(r) - rc_r) = \mathbb{P}\left(\frac{T - r\log(r)}{r} > -c_r\right) \underset{r\to\infty}{\longrightarrow} 1.$$

□

1.6 Further reading

Our study of coin tossing leaves aside several remarkable properties, such as the arcsine law and invariance principles, discussed in William Feller's book [Fel68, Fel71]. It also contains a Fourier-analytic proof of the Berry–Esseen theorem. A beautiful *longest common subsequence problem* for two

[3] A sharp transition between two regimes.

coin-tossing sequences is studied in Michael Steele's book on randomized combinatorial optimization [Ste97]. Michael Steele's article [Ste94] on Poisson approximation is very illuminating. The book [BHJ92] by Andrew Barbour, Lars Holst, and Svante Janson contains a strengthened version of Le Cam's inequality, obtained by Barbour and Eagleson using Louis Chen and Charles Stein's integration-by-parts method:

$$d_{\mathrm{VT}}\left(\mathrm{Ber}(p_1) * \cdots * \mathrm{Ber}(p_n), \mathrm{Poi}(p_1 + \cdots + p_n)\right) \leqslant (1 - e^{-(p_1+\cdots+p_n)})\frac{p_1^2 + \cdots + p_n^2}{p_1 + \cdots + p_n}.$$

For comparison, an inhomogeneous version of the Berry–Esseen central limit theorem, from William Feller's book states that, if $(X_n)_{n\geqslant 1}$ are independent random variables, then writing $S_n = X_1 + \cdots + X_n$, $\sigma_k^2 = \mathbb{E}(|X_k - \mathbb{E}(X_k)|^2)$, and $\tau_k^3 = \mathbb{E}(|X_k - \mathbb{E}(X_k)|^3)$ gives

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \leqslant x\right) - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt\right|^2 \leqslant \frac{(\tau_1^3 + \cdots + \tau_n^3)^2}{(\sigma_1^2 + \cdots + \sigma_n^2)^3}.$$

Total variation distance $d_{\mathrm{VT}}(\cdot,\cdot)$ can also be viewed as a *Wasserstein distance* (coupling). First observe that $\mathbb{P}(X \neq Y) = \mathbb{E}(d(X,Y))$ for the discrete metric $d(x,y) = \delta_{x\neq y}$. Writing $\Pi(\mu,\nu)$ for the set of distributions on $E \times E$ with marginals μ and ν, we obtain

$$d_{\mathrm{VT}}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)}\int_{E\times E} d(x,y)\,d\pi(x,y).$$

The coupon collector problem is discussed, among other places, in William Feller's book [Fel68, Fel71], Rejeev Motwani and Prabhakar Raghavan's [MR95], and articles by Lars Holst [Hol01] and Aristides Doumas and Vassilis Papanicolaou [DP13]. Donald Newman and Lawrence Shepp showed in [NS60] that, if each type must be observed m times, the mean collection completion time is

$$r\log(r) + (m-1)r\log(\log(r)) + \mathcal{O}(r).$$

Other variants appear in the book by Claude Bouzitat, Gilles Pagès, Frédérique Petit, and Fabrice Carrance [BPPC99]. Theorem 1.16, showing asymptotic Gumbel fluctuations, was obtained by Paul Erdős and Alfréd Rényi [ER61a]. The nonuniform case associated with a discrete probability measure $(p_1,\ldots,p_r)$ is analyzed by Lars Holst [Hol01]. When r is unknown, one has the estimator

$$\widehat{r}_n := \mathrm{card}\{X_1,\ldots,X_n\}.$$

If $e_1,\ldots,e_r$ is the standard basis of $\mathbb{R}^r$, then the random vector

$$C_n := (C_{n,1},\ldots,C_{n,r}) := e_{X_1} + \cdots + e_{X_n}$$

in $\mathbb{N}^r$ has the multinomial distribution with size n and parameter $(p_1, \dots, p_r)$, and

$$\widehat{r}_n = \sum_{i=1}^{r} \mathbb{1}_{\{C_{n,i}>0\}} = r - \sum_{i=1}^{r} \mathbb{1}_{\{C_{n,i}=0\}}.$$

In particular,

$$\mathbb{E}(\widehat{r}_n) = r - \sum_{i=1}^{r} \mathbb{P}(C_{n,i} = 0) = r - \sum_{i=1}^{r} (1-p_i)^n.$$

Finally, since the r types are ordered, the random variable $\max(X_1, \dots, X_n)$ is an estimator of the right endpoint r of the support $\{1, \dots, r\}$.

The coupon collector is a special case of the covering problem discussed in David Aldous' article [Ald91], the book by David Aldous and James Fill [AF01], Amir Dembo's lecture notes [Dem05], and the book by David Levin, Yuval Peres, and Elizabeth Wilmer [LPW09]: if $(X_n)_{n\geqslant 1}$ is a sequence of random variables, not necessarily independent or identically distributed, taking values in the same finite set E, then the *cover time*[4] of E is $T := \inf\{n \geqslant 1 : \{X_1, \dots, X_n\} = E\}$.

[4] Also called the *covering time*.

2

Random walks

Keywords. Random walk; exit time; gambler's ruin; Dirichlet problem; harmonic function; Gaussian free field. **Tools.** Combinatorics; Markov chain; martingale; multinomial distribution; Fourier transform; symmetric group; coupon collector. **Difficulty.** *

Random walks are essential probabilistic objects of surprising richness and diversity. This chapter presents some elementary aspects, particularly in connection with their Markov property.

2.1 Simple random walk on the line

The simple random walk on the line $\mathbb{Z}$ is the sequence $(X_n)_{n\geqslant 0}$ of random variables defined by the recursive (or autoregressive) relation

$$X_{n+1} = X_n + \varepsilon_{n+1} = X_0 + \varepsilon_1 + \cdots + \varepsilon_{n+1}$$

for every $n \geqslant 0$, where $(\varepsilon_n)_{n\geqslant 1}$ is a sequence of i.i.d. real random variables with Rademacher distribution, independent of X_0. Set

$$p := \mathbb{P}(\varepsilon_n = 1) = 1 - \mathbb{P}(\varepsilon_n = -1) \in]0, 1[.$$

The sequence $(X_n)_{n\geqslant 0}$ is a Markov chain with state space $\mathbb{Z}$ and transition kernel given, for every $x, y \in \mathbb{Z}$, by

$$\mathbf{P}(x, y) = p\mathbb{1}_{\{y=x+1\}} + (1-p)\mathbb{1}_{\{y=x-1\}}$$

and the matrix $\mathbf{P}$ is tridiagonal. A random variable B has the Bernoulli distribution $(1-p)\delta_0 + p\delta_1$ if and only if the random variable $2B-1$ has the Rademacher distribution $(1-p)\delta_{-1} + p\delta_1$. Thus, for every $n \geqslant 0$,

$$\frac{X_n - X_0 + n}{2} \sim \mathrm{Bin}(n, p).$$

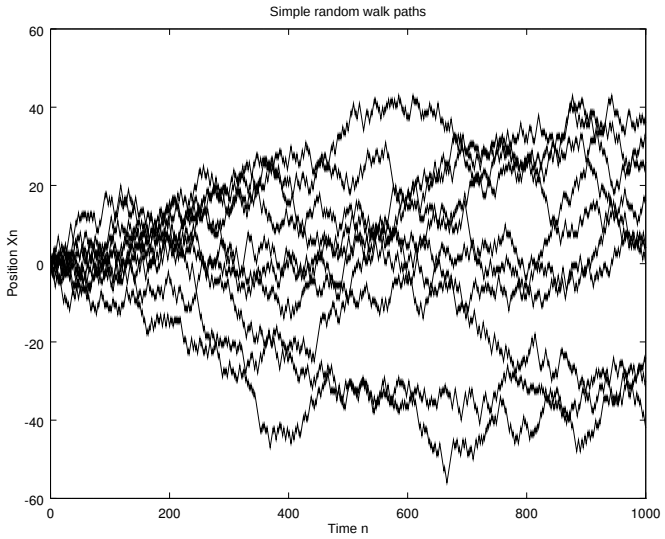


Fig. 2.1. Trajectories starting from 0 of the simple random walk on $\mathbb{Z}$.

Theorem 2.1 (Recurrence)**.** *The simple random walk on $\mathbb{Z}$ is null recurrent if $p = 1/2$ (simple symmetric random walk) and transient if $p \neq 1/2$ (simple asymmetric random walk).*

Proof. Recall that a state x is recurrent if the walk starting from x returns there almost surely, or equivalently if it visits x infinitely often. A useful criterion is that x is recurrent if and only if $\sum_n \mathbf{P}^n(x,x) = \infty$. The chain has period 2, so $\mathbf{P}^{2n+1}(x,x) = 0$ for every $n \geqslant 0$. Since it is irreducible, all states have the same classification, and we can restrict attention to state 0. The binomial formula gives

$$\mathbf{P}^{2n}(0,0) = \mathbb{P}(X_{2n} = 0 \,|\, X_0 = 0) = \binom{2n}{n} p^n (1-p)^n.$$

Stirling's formula $n! \sim \sqrt{2\pi n}(n/e)^n$ gives, writing $\rho = 4p(1-p)$,

$$\mathbf{P}^{2n}(0,0) \sim \frac{\rho^n}{\sqrt{\pi n}}.$$

If $p = 1/2$, then $\rho = 1$ and the chain is recurrent; if $p \neq 1/2$, then $\rho < 1$ and it is transient. When $p = 1/2$, counting measure is symmetric (since the matrix $\mathbf{P}$ is symmetric), hence invariant. Since this measure is not finite, the chain is null recurrent.

Alternatively, the case $p \neq 1/2$ can be handled with more probabilistic intuition. Indeed, the strong law of large numbers gives $X_n = X_0 + n(2p -$

$1 + o_{n\to\infty}(1)$) almost surely, so $X_n \to +\infty$ almost surely if $p > 1/2$ and $X_n \to -\infty$ almost surely if $p < 1/2$, ruling out recurrence. □

The next theorem studies the *gambler's ruin problem*: a gambler wins 1 euro with probability p and loses 1 euro with probability $1-p$. The initial fortune is x, and the gambler stops upon reaching $a < x$ euros (ruin) or $b > x$ euros (gain). We use the traditional conditional notation $\mathbb{P}_x(\cdot) := \mathbb{P}(\cdot \,|\, X_0 = x)$ and $\mathbb{E}_x(\cdot) = \mathbb{E}(\cdot \,|\, X_0 = x)$.

Theorem 2.2 (Exiting a box[1] or gambler's ruin)**.** *Let $a < b$ belong to $\mathbb{Z}$, and define the stopping times τ_a, τ_b, and τ by*

$$\tau_a = \inf\{n \geqslant 0 : X_n = a\}, \quad \tau_b = \inf\{n \geqslant 0 : X_n = b\}, \quad \text{et} \quad \tau = \min(\tau_a, \tau_b).$$

Then for every $x \in [a,b] \cap \mathbb{Z}$ there exists $c > 0$ such that $\mathbb{E}_x(e^{c\tau}) < \infty$. In particular, $\mathbb{E}_x(\tau) < \infty$, and hence $\mathbb{P}_x(\tau < \infty) = 1$. Moreover, setting $\rho = \dfrac{1-p}{p}$,

$$\mathbb{P}_x(X_\tau = a) = \begin{cases} \dfrac{\rho^b - \rho^x}{\rho^b - \rho^a} & \text{si } p \neq \frac{1}{2}, \\ \dfrac{b-x}{b-a} & \text{si } p = \frac{1}{2}, \end{cases}$$

and

$$\mathbb{E}_x(\tau) = \begin{cases} \dfrac{x-a}{1-2p} - \dfrac{(b-a)}{1-2p}\dfrac{\rho^x - \rho^a}{\rho^b - \rho^a} & \text{si } p \neq \frac{1}{2}, \\ (b-x)(x-a) & \text{si } p = \frac{1}{2}. \end{cases}$$

If $p = 1/2$, the chain is recurrent and almost surely visits every state infinitely often, so $\mathbb{P}_x(\tau_a < \infty) = 1$ and $\mathbb{P}_x(\tau_b < \infty) = 1$ for every $a \leqslant x \leqslant b$. If $p \neq 1/2$, however, it is transient, and the hitting times of a or b are no longer almost surely finite (depending on the probability p and the starting point x). This is visible in the formulas of Theorem 2.2 by letting a or b tend to infinity. The exit time τ from $[a,b]$ has the same distribution as the absorption time T into $\{a,b\}$ of the chain $Y = (Y_n)_{n\geqslant 0}$ on the finite state space $\{a,\dots,b\}$, with the same transitions as $(X_n)_{n\geqslant 0}$ but absorption at a and b. For Y, states a and b are recurrent and all the others (finitely many) are transient. Since the chain Y almost surely visits each transient state only finitely often, we deduce $\mathbb{P}_x(\tau < \infty) = \mathbb{P}_x(T < \infty) = 1$ for every $a \leqslant x \leqslant b$.

Proof. We show that $\mathbb{E}_x(\tau) < \infty$ for every $a \leqslant x \leqslant b$. It suffices to obtain a geometric upper bound on the tail of τ. One could use coupling, as for the Wright–Fisher model; see Remark 12.7. For every $a \leqslant x \leqslant b$, there is a path ℓ_x of length $|\ell_x| \leqslant (b-a)$ from x to a or b. We have

$$\mathbb{P}_x(\tau > (b-a)) \leqslant \mathbb{P}(X_{1:|\ell_x|} \neq \ell_x)$$

[1] Quite natural for a drunkard modeled by a random walk.

$$\begin{aligned} &= 1 - \mathbb{P}(X_{1:|\ell_x|} = \ell_x) \\ &\leqslant 1 - \min(p, 1-p)^{|\ell_x|}. \end{aligned}$$

If $\eta = \max_{a<x<b}(1 - \min(p, 1-p)^{|\ell_x|}) < 1$, induction and the conditional independence of past and future given the present give, for every $k \geqslant 1$,

$$\begin{aligned} &\mathbb{P}_x(\tau > k(b-a)) \\ &\quad = \sum_{a<y<b} \mathbb{P}_x(\tau > k(b-a), X_{(k-1)(b-a)} = y, \tau > (k-1)(b-a)) \\ &\quad = \sum_{a<y<b} \mathbb{P}_y(\tau > (b-a))\mathbb{P}_x(X_{(k-1)(b-a)} = y, \tau > (k-1)(b-a)) \\ &\quad \leqslant \eta \mathbb{P}_x(\tau > (k-1)(b-a)) \leqslant \eta\eta^{k-1} = \eta^k. \end{aligned}$$

Since $\eta < 1$, we obtain $\mathbb{E}_x(e^{c\tau}) < \infty$ for some real $c > 0$. In particular, all moments of τ are finite under $\mathbb{P}_x$, and $\mathbb{P}_x(\tau < \infty) = 1$. Compute

$$r(x) := \mathbb{P}_x(X_\tau = a).$$

For every $a < x < b$,

$$\begin{aligned} r(x) &= \mathbb{P}_x(X_\tau = a \,|\, X_1 = x+1)p + \mathbb{P}_x(X_\tau = a \,|\, X_1 = x-1)(1-p) \\ &= pr(x+1) + (1-p)r(x-1). \end{aligned}$$

The solution set of this second-order linear recurrence is a vector space containing the constant solution 1. If $p \neq 1/2$, then ρ^x is another solution, linearly independent of 1, so all solutions have the form $A + B\rho^x$ with constants A and B. The boundary conditions $r(a) = 1$, $r(b) = 0$ determine A and B, giving the unique solution

$$r(x) = \frac{\rho^b - \rho^x}{\rho^b - \rho^a}.$$

If $p = 1/2$, then $\rho = 1$ and the preceding two fundamental solutions coincide. In this case, however, x is also a solution, linearly independent of 1. Solutions thus have the form $A + Bx$ with constants A and B. The boundary conditions $r(a) = 1$ and $r(b) = 0$ determine A and B, giving the unique solution

$$r(x) = \frac{b-x}{b-a}.$$

Now compute

$$R(x) := \mathbb{E}_x(\tau).$$

Conditioning on X_1 gives, for every $a < x < b$, the linear recurrence (the method applies to any Markov chain, and likewise to $r(x)$)

$$R(x) = pR(x+1) + (1-p)R(x-1) + 1.$$

The nonzero right-hand side 1 suggests looking for particular solutions. If $p \neq 1/2$, then $x/(1-2p)$ is a particular solution, and all solutions have the form $R(x) = x/(1-2p) + A + B\rho^x$. The boundary conditions $R(a) = 0$ and $R(b) = 0$ finally give

$$R(x) = \frac{x-a}{1-2p} - \frac{(b-a)}{1-2p}\frac{\rho^b - \rho^x}{\rho^b - \rho^a}.$$

If $p = 1/2$, then $-x^2$ is a particular solution, and all solutions have the form $-x^2 + A + Bx$. The boundary conditions $R(a) = R(b) = 0$ finally give

$$R(x) = (b-x)(x-a).$$

Similarly, we can compute the following functions:

$$F(x,s) = \mathbb{E}_x(s^\tau), \quad G(x,s) = \mathbb{E}_x(s^\tau \mathbb{1}_{X_\tau = a}), \quad \text{et} \quad G(x,s) = \mathbb{E}_x(s^\tau \mathbb{1}_{X_\tau = b})$$

which, for $a \leqslant x \leqslant b$ and $s \in]-1,1[$, satisfy the recurrence

$$r(x) = psr(x+1) + (1-p)sr(x-1)$$

with respective boundary conditions

$$F(a,s) = F(b,s) = 1, \quad G(a,s) = 1 = 1 - G(b,s), \quad H(a,s) = 0 = 1 - H(b,s).$$

The expression for $\mathbb{E}_x(\tau)$ can then be recovered by differentiating the generating function. The explicit expressions for F, G, and H are rather cumbersome. □

Remark 2.3 (Limit theorems to the rescue)**.** *Here is another proof that $\mathbb{P}_x(\tau < \infty) = 1$. Set $m = 2p - 1$ and $\sigma^2 = 4p(1-p)$. If $m \neq 0$, the law of large numbers implies that almost surely $(X_n)_{n \geqslant 1}$ tends to $+\infty$ if $m > 0$ and to $-\infty$ if $m < 0$, so $\mathbb{P}_x(\tau < \infty) = 1$. If $m = 0$, then for every $n \geqslant 1$, setting $I_n = \frac{1}{\sqrt{n}}\,]a,b[$ gives*

$$\mathbb{P}_x(\tau = \infty) \leqslant \mathbb{P}(a < X_n < b) = \mathbb{P}\left(\frac{X_n}{\sqrt{n}} \in I_n\right).$$

Now $(n^{-1/2}X_n)_{n \geqslant 1}$ converges in distribution to $\mathcal{N}(0, \sigma^2)$ by the central limit theorem. Although I_n depends on n, since $(I_n)_{n \geqslant 1}$ is nonincreasing,

$$\limsup_{n \to \infty} \mathbb{P}\left(\frac{X_n}{\sqrt{n}} \in I_n\right) \leqslant \inf_{n \geqslant 1} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{I_n} e^{-\frac{t^2}{2\sigma^2}}\, dt = 0.$$

Remark 2.4 (Martingales)**.** *If $(Z_n)_{n \geqslant 0}$ is a Markov chain on a finite set E with kernel $\mathbf{P} = \mathbf{L} + \mathbf{I}$, then for every function $f : E \to \mathbb{R}$ the sequence $(M_n)_{n \geqslant 0}$ given by $M_0 = 0$ and, for every $n \geqslant 1$,*

$$M_n = f(Z_n) - f(Z_0) - \sum_{k=0}^{n-1} (\mathbf{L}f)(Z_k)$$

is a martingale for the natural filtration of $(Z_n)_{n\geqslant 0}$*. This formula can be viewed as a discrete Itô formula. When* f *is harmonic for* $\mathbf{L}$*, that is,* $\mathbf{L}f = 0$*, then* $(f(Z_n) - f(Z_0))_{n\geqslant 0}$ *is a martingale. The vector of hitting probabilities of a closed set is harmonic. The formulas in Theorem 2.2 can be recovered using suitable martingales and the optional stopping theorem. For example, the martingale* $(X_n - n(p-q))_{n\geqslant 0}$ *gives* $\mathbb{E}_0(X_\tau) = (p-q)\mathbb{E}_0(\tau)$*, while* $(\rho^{X_n})_{n\geqslant 0}$ *gives* $\mathbb{E}_0(\rho^{X_\tau}) = 1$*. This martingale method extends to processes with continuous time and state space, providing in particular exit-time formulas for a diffusion on* $\mathbb{R}^d$*; see Chapter 27. Moreover, the equations satisfied by* r *and* R *in the proof of Theorem 2.2 are special cases of the Dirichlet problem in Theorem 2.8.*

Theorem 2.5 (Catalan numbers)**.** *If* $\tau := \inf\{n \geqslant 1 : X_n = 0\}$*, then*

$$\mathbb{P}_0(\tau = 2n+2) = \frac{2}{n+1}\binom{2n}{n} p^{n+1}(1-p)^{n+1}, \quad n \geqslant 0.$$

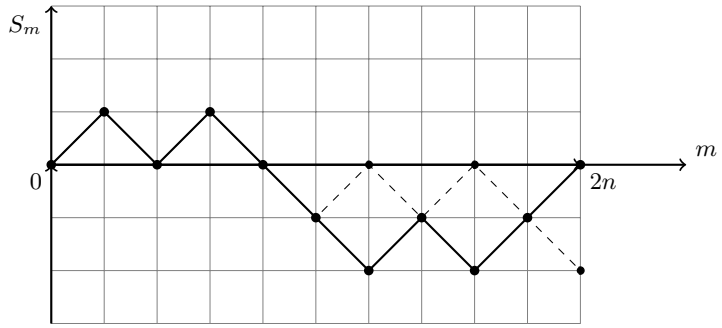


Fig. 2.2. A path from 0 to 0 that does not remain nonnegative, and the associated path from 0 to -2 in the proof of Theorem 2.5.

We recognize the n^{e} Catalan number $\frac{1}{n+1}\binom{2n}{n}$. Catalan numbers are also the even moments of the semicircle distribution; see Chapter 21.

Proof. Given $\{X_0 = 0\}$, the event $\{\tau = 2n+2\}$ corresponds to a trajectory of length $2n+2$ starting from 0 and returning to zero while staying strictly positive or strictly negative. These cases are equally likely, giving the factor 2. In either case, there must have been $n+1$ increments $+1$ and $n+1$ increments -1, hence

$$\mathbb{P}_0(\tau = 2n+2) = 2C_n p^{n+1}(1-p)^{n+1}.$$

where C_n is the number of paths of length $2n+2$ starting and ending at zero while staying strictly positive in between. The first increment must be $+1$ and the last -1. Thus C_n equals the number of nonnegative paths of length $2n$ starting and ending at zero. Such paths have n increments $+1$ and n increments -1. Consider all paths starting and ending at zero with n increments $+1$ and n increments -1. There are $\binom{2n}{n}$ of them. If such a path is not nonnegative, reverse all increment signs after the first negative position, interchanging $+1$ and -1. Figure 2.2 illustrates this. The resulting path has $n-1$ increments $+1$ and $n+1$ increments -1. Every path starting at zero with $n-1$ increments $+1$ and $n+1$ increments -1 arises this way, and there are $\binom{2n}{n-1}$ such paths. This bijection therefore gives $C_n = \binom{2n}{n} - \binom{2n}{n-1} = \frac{1}{n+1}\binom{2n}{n}$ (Désiré André's formula). □

Theorem 2.6 (The ballot theorem[2])**.** *If $n = a+b$ and $k = a-b$ with $0 \leqslant b \leqslant a$, then*

$$\mathbb{P}(S_1 > 0, \dots, S_n > 0 \,|\, S_0 = 0, S_n = k) = \frac{k}{n} = \frac{a-b}{a+b}.$$

Theorem 2.6 says that, in an election with candidates A and B and n voters, where A receives a votes and B receives $b \leqslant a$ votes, the probability that A stays ahead of B throughout the counting of the n ballots is $(a-b)/(a+b)$.

Proof. First note that $0 \leqslant k \leqslant n$ and $(n+k, n-k) = 2(a,b)$. Let $P_{n,k}$ be the set of simple random walk paths of length n starting at $(0,0)$ and ending at (n,k). All have exactly a increments $+1$ and b increments -1, so there are $\binom{a+b}{a}$ such paths. Let $P^+_{n,k}$ be the subset of paths strictly positive at times $1, \dots, n$. We have

$$\mathbb{P}(S_1 > 0, \dots, S_n > 0 \,|\, S_0 = 0, S_n = k) = \operatorname{card}(P^+_{n,k}) \frac{p^a (1-p)^b}{\mathbb{P}(S_n = k \,|\, S_0 = 0)}.$$

On the other hand, $\mathbb{P}(S_n = k \,|\, S_0 = 0) = \operatorname{card}(P_{n,k})\, p^a (1-p)^b$, hence

$$\mathbb{P}(S_1 > 0, \dots, S_n > 0 \,|\, S_0 = 0, S_n = k) = \frac{\operatorname{card}(P^+_{n,k})}{\operatorname{card}(P_{n,k})}.$$

Remarkably, this formula does not depend on p. The set $P_{n,k} \setminus P^+_{n,k}$ is invariant under reflection of the path segment before the return to 0. This is the trick used in the proof of Theorem 2.5. It follows that the elements of $P_{n,k} \setminus P^+_{n,k}$ beginning with an increment $+1$ are in bijection with those of $P_{n,k} \setminus P^+_{n,k}$ beginning with an increment -1. The latter are in bijection with the elements of $P_{n,k}$ beginning with an increment -1, themselves in bijection

[2] Named after its interpretation in vote counting.

with $P_{n-1,k+1}$. This gives $\mathrm{card}(P_{n,k}) - \mathrm{card}(P^+_{n,k}) = 2\mathrm{card}(P_{n-1,k+1})$. Since $\mathrm{card}(P_{n,k}) = \binom{n}{(n+k)/2}$, we finally obtain

$$\mathrm{card}(P^+_{n,k}) = \binom{n}{(n+k)/2} - 2\binom{n-1}{(n+k)/2} = \frac{k}{n}\binom{n}{(n+k)/2}.$$

In other words, $\mathrm{card}(P^+_{n,k}) = (k/n)\mathrm{card}(P_{n,k})$. □

The Catalan-number formula of Theorem 2.5 can be written as

$$2^{2n+2}\mathbb{P}(S_1 > 0, \dots, S_{2n+1} > 0 \,|\, S_0 = 0, S_{2n+2} = 0) = \mathrm{card}(P^+_{2n+1,1}) = \frac{1}{n+1}\binom{2n}{n}.$$

2.2 Simple symmetric random walk in space

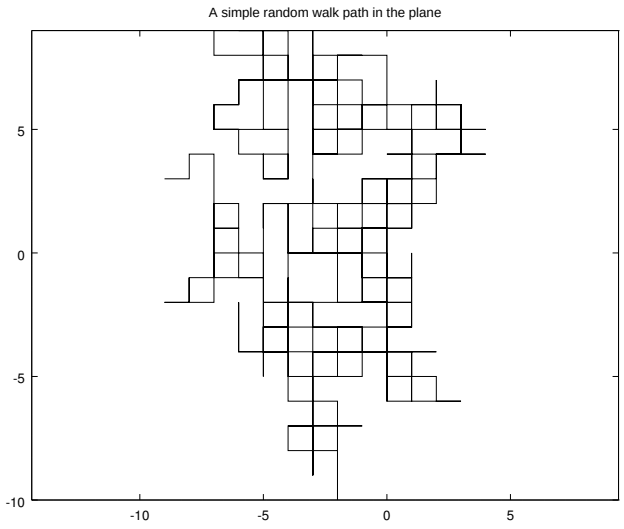


Fig. 2.3. Beginning of a simple random walk trajectory on $\mathbb{Z}^2$.

The simple symmetric random walk on $\mathbb{Z}^d$, $d \geqslant 1$, is defined by

$$X_{n+1} = X_n + \varepsilon_{n+1} = X_0 + \varepsilon_1 + \cdots + \varepsilon_{n+1}$$

where $(\varepsilon_n)_{n\geqslant 1}$ is an i.i.d. sequence of random variables independent of the initial position X_0 and uniformly distributed on $\{\pm e_1, \dots, \pm e_d\}$, where $e_1, \dots, e_d$

is the standard basis of $\mathbb{R}^d$. The sequence $(X_n)_{n\geqslant 0}$ is a Markov chain with state space $\mathbb{Z}^d$ and transition kernel

$$\mathbf{P}(x,y)=\frac{1}{2d}\mathbb{1}_{|x-y|_1=1}$$

where $|x|_1:=|x_1|+\cdots+|x_d|$. It is also called the symmetric nearest-neighbor random walk for the norm $|\cdot|_1$. The increments $(\varepsilon_n)_{n\geqslant 1}$ are uniformly distributed on the unit sphere for $|\cdot|_1$. Viewing the increments $\varepsilon_1,\ldots,\varepsilon_n$ as arising from n rolls of a fair $2d$-sided die shows that $\mathrm{Loi}(X_n-X_0)$ is the image of the multinomial distribution

$$\left(\frac{1}{2d}\delta_{e_1}+\frac{1}{2d}\delta_{-e_1}+\cdots+\frac{1}{2d}\delta_{-e_d}+\frac{1}{2d}\delta_{e_d}\right)^{*n}=\mathrm{Mul}\left(n,\left(\frac{1}{2d},\ldots,\frac{1}{2d}\right)\right)$$

under the map

$$(k_1,\ldots,k_{2d})\in\mathbb{N}^{2d}\mapsto(k_1-k_2,\ldots,k_{2d-1}-k_{2d})\in\mathbb{Z}^d.$$

The random time by which the walk has taken all $2d$ available directions has the coupon collector distribution studied in Chapter 1, with $r=2d$.

Theorem 2.7 (Simple symmetric random walk)**.** *The simple symmetric random walk on $\mathbb{Z}^d$ is null recurrent if $d\leqslant 2$ and transient if $d\geqslant 3$.*

Proof. The proof uses Fourier analysis. First recall that, for every random variable Z taking values in $\mathbb{Z}^d$,

$$\mathbb{P}(Z=z)=\int_{[0,1[^d}e^{2\pi i\langle z,x\rangle}\varphi_Z(x)\,dx\quad\text{où}\quad\varphi_Z(x)=\sum_{z\in\mathbb{Z}^d}e^{-2\pi i\langle z,x\rangle}\mathbb{P}(Z=z).$$

Introduce a parameter $\rho\in[0,1[$ to interchange the sum and integral. Using the preceding identity yields

$$\begin{aligned}\sum_{n=0}^{\infty}\mathbf{P}^n(0,0)&=\lim_{\rho\uparrow 1}\sum_{n=0}^{\infty}\rho^n\mathbf{P}^n(\varepsilon_1+\cdots+\varepsilon_n=0)\\&=\lim_{\rho\uparrow 1}\sum_{n=0}^{\infty}\int_{[0,1[^d}\rho^n e^{2\pi i\langle 0,x\rangle}\varphi_{\varepsilon_1+\cdots+\varepsilon_n}(x)\,dx\\&=\lim_{\rho\uparrow 1}\sum_{n=0}^{\infty}\int_{[0,1[^d}\rho^n(\varphi_{\varepsilon_1}(x))^n\,dx\\&=\lim_{\rho\uparrow 1}\int_{[0,1[^d}\sum_{n=0}^{\infty}\rho^n(\varphi_{\varepsilon_1}(x))^n\,dx\\&=\lim_{\rho\uparrow 1}\int_{[0,1[^d}\frac{1}{1-\rho\varphi_{\varepsilon_1}(x)}\,dx\end{aligned}$$

$$= \lim_{\rho \uparrow 1} \int_{[0,1[^d} \frac{d}{\sum_{k=1}^d (1 - \rho \cos(2\pi x_k))}\, dx$$
$$= \int_{[0,1[^d} \frac{d}{\sum_{k=1}^d (1 - \cos(2\pi x_k))}\, dx,$$

where monotone convergence justifies interchanging the limit as $\rho \uparrow 1$ and the integral. The integral's behavior depends only on the integrand near 0. Spherical coordinates and a Taylor expansion at 0 show that it is finite if and only if $r \geqslant 0 \mapsto r^{d-3}$ is integrable near 0, that is, $d > 2$, or equivalently $d \geqslant 3$ since d is an integer! □

Recurrence of the simple symmetric random walk on $\mathbb{Z}^d$ $(X_n)_{n \geqslant 0}$ can also be studied combinatorially. The chain is irreducible with period 2 for every $d \geqslant 1$, so it suffices to study $\sum_n \mathbf{P}^{2n}(0,0)$. The multinomial formula gives

$$\mathbf{P}^{2n}(0,0) = \frac{1}{(2d)^{2n}} \sum_{\substack{r_1 + \cdots + r_{2d} = n \\ r_1 = r_2, \ldots, r_{d-1} = r_d}} \binom{2n}{r_1, \ldots, r_{2d}}$$
$$= \frac{1}{(2d)^{2n}} \sum_{r_1 + \cdots + r_d = n} \frac{(2n)!}{r_1!^2 \cdots r_d!^2}.$$

The case $d = 1$ has already been treated. The cases $d = 2$ and $d \geqslant 3$ remain.

Case $d = 2$ (direct method). Vandermonde's identity[3] gives

$$\mathbf{P}^{2n}(0,0) = \frac{1}{4^{2n}} \sum_{n_1 + n_2 = n} \frac{(2n)!}{n_1!^2 n_2!^2}$$
$$= \frac{(2n)!}{4^{2n} n!^2} \sum_{k=0}^n \binom{n}{k}^2$$
$$= \frac{(2n)!}{4^{2n} n!^2} \frac{(2n)!}{n!^2} \underset{n \to \infty}{\sim} \frac{1}{\pi n}$$

which is not summable, so the chain is recurrent. It is null recurrent because counting measure, which is infinite since $\mathbb{Z}^2$ is infinite, is symmetric and hence invariant.

Case $d = 2$ (a clever method). Let U_n and V_n be the projections of X_n onto the two diagonals (slopes ± 1). One checks that $(U_n)_{n \geqslant 0}$ and $(V_n)_{n \geqslant 0}$ are independent simple symmetric random walks on $\frac{1}{\sqrt{2}}\mathbb{Z}$! Moreover, $X_n = 0$ if and only if $U_n = V_n = 0$, yielding $\mathbf{P}^{2n}(0,0) \sim \frac{1}{\pi n}$, so the chain is recurrent. Note that the components of $(X_n)_{n \geqslant 0}$ (the two coordinates) are not simple random walks on $\mathbb{Z}$, but Markov chains with kernel $\mathbf{Q}(x,y) = \frac{1}{4}\mathbb{1}_{|x-y|=1} + \frac{1}{2}\mathbb{1}_{x=y}$ (they can be shown to be recurrent, but this is not particularly useful here).

[3] The "convolution" formula $\binom{n+m}{r} = \sum_{k=0}^r \binom{n}{k}\binom{m}{r-k}$.

Case $d = 3$ (direct method). The multinomial formula gives

$$\begin{aligned}\mathbf{P}^{2n}(0,0) &= \frac{1}{6^{2n}} \sum_{r_1+r_2+r_3=n} \frac{(2n)!}{r_1!^2 r_2!^2 r_3!^2} \\ &= \frac{(2n)!}{2^{2n}3^n n!^2} \sum_{r_1+r_2+r_3=n} \binom{n}{r_1\ r_2\ r_3}^2 \left(\frac{1}{3}\right)^n.\end{aligned}$$

If $n = 3m$, a short analysis shows that

$$\binom{n}{r_1\ r_2\ r_3} \leqslant \binom{n}{m\ m\ m}$$

Hence, using the trinomial formula with size n and second parameter $(1/3, 1/3, 1/3)$, we obtain, for $n = 3m$,

$$\mathbf{P}^{2n}(0,0) \leqslant \frac{(2n)!}{2^{2n}3^n n!^2} \binom{n}{m\ m\ m} \sim \frac{1}{2}\left(\frac{3}{\pi n}\right)^{3/2} = \frac{c}{n^{3/2}}.$$

Thus $\sum_m \mathbf{P}^{6m}(0,0) < \infty$. Since, for $k = 1, 2$,

$$\frac{1}{6^{2k}}\mathbf{P}^{6m}(0,0) \leqslant \mathbf{P}^{6m+2k}(0,0),$$

we obtain

$$\sum_n \mathbf{P}^{2n}(0,0) = \sum_{k=0,1,2} \sum_m \mathbf{P}^{6m+2k}(0,0) < \infty,$$

so the chain is transient.

Case $d > 3$ (a clever reduction to $d = 3$). Let $X' = (X'_n)_{n\geqslant 0}$ be the chain projected onto $\mathbb{Z}^3$, consisting of the first three coordinates of $X = (X_n)_{n\geqslant 0}$. It is a Markov chain on $\mathbb{Z}^3$ with kernel

$$\mathbf{Q} := \Big(1 - \frac{d-3}{d}\Big)\mathbf{P} + \frac{d-3}{d}\mathbf{I}$$

where $\mathbf{P}$ is the kernel of the simple symmetric random walk X'' on $\mathbb{Z}^3$. A return to zero of X always entails a return to zero of X', so if X' is transient, X must also be transient. The trajectories of X' are those of X'' with geometric holding times at each site; in particular, their asymptotic visit frequencies agree. Thus transience of X'' implies transience of X', so X is transient.

2.3 The Dirichlet problem and the Gaussian free field

Let $(X_n)_{n\geqslant 0}$ be the simple random walk on $\mathbb{Z}^d$ with transition kernel $\mathbf{P}$. For every bounded function $f : \mathbb{Z}^d \to \mathbb{R}$ and every $x \in \mathbb{Z}^d$,

$$\mathbb{E}(f(X_{n+1}) \,|\, X_n = x) - f(x) = (\mathbf{P}f)(x) - f(x) = (\Delta f)(x),$$

where $\Delta := \mathbf{P} - \mathbf{I}$. The operator Δ is the generator of the simple symmetric random walk on $\mathbb{Z}^d$. It is a discrete Laplacian, computing the deviation from the average over neighboring sites:

$$(\Delta f)(x) = \left(\frac{1}{2d} \sum_{y:|y-x|_1=1} f(y) \right) - f(x) = \frac{1}{2d} \sum_{y:|y-x|_1=1} (f(y) - f(x)).$$

We call f harmonic on $A \subset \mathbb{Z}^d$ if $\Delta f = 0$ on A, meaning that at every point of A, the value of f equals its average over the $2d$ neighbors. The operator Δ is local: $(\Delta f)(x)$ depends only on the values of f at x and its nearest neighbors. Thus Δf on A depends only on the values of f on $\bar{A} := A \cup \partial A$, where

$$\partial A := \{y \notin A : \exists x \in A, |x - y|_1 = 1\}$$

is the outer boundary of A. The Dirichlet problem asks for a harmonic function on A with prescribed values on ∂A. This is a linear algebra problem. Theorem 2.8 below gives a probabilistic expression for its solution using the hitting time

$$\tau_{\partial A} := \inf\{n \geqslant 0 : X_n \in \partial A\}.$$

Chapter 24 considers the same question on an open subset of $\mathbb{R}^d$ for the classical Laplacian.

Theorem 2.8 (The Dirichlet problem)**.** *Let $A \subset \mathbb{Z}^d$ be a finite nonempty set. Then, for every $x \in A$, $\mathbb{P}_x(\tau_{\partial A} < \infty) = 1$. Moreover, for every function $g : \partial A \to \mathbb{R}$, the function $f : \bar{A} \to \mathbb{R}$ defined, for every $x \in \bar{A}$, by*

$$f(x) = \mathbb{E}_x(g(X_{\tau_{\partial A}}))$$

is the unique solution of the system

$$\begin{cases} f = g & \text{sur } \partial A, \\ \Delta f = 0 & \text{sur } A. \end{cases}$$

When $d = 1$, we recover the function r studied in the proof of the gambler's ruin theorem 2.2. Moreover, by Remark 2.4, the image of a Markov chain under a function harmonic for its generator is a martingale, which retrospectively explains the probabilistic solution formula through the optional stopping theorem.

The quantity $f(x) = \mathbb{E}_x(g(X_{\tau_{\partial A}})) = \sum_{y \in \partial A} g(y) \mathbb{P}_x(X_{\tau_{\partial A}} = y)$ is the mean of g under the distribution μ_x on ∂A, called harmonic measure, defined by

$$\mu_x(y) = \mathbb{P}_x(X_{\tau_{\partial A}} = y), \quad y \in \partial A.$$

One may also call this the discrete Poisson kernel, by analogy with the Dirichlet problem on an open subset of $\mathbb{R}^d$ (see Chapter 24).

Proof. The property $\mathbb{P}_x(\tau_{\partial A} < \infty) = 1$ for every $x \in A$ can be proved as in Remark 2.3, or as in the proof of Theorem 2.2, which also gives a geometric upper bound on the tail of τ.

Check that the proposed function f is a solution. For every $x \in \partial A$, we have $\tau_{\partial A} = 0$ on $\{X_0 = x\}$, hence $f = g$ on ∂A. We now show that $\Delta f = 0$ on A. By linearity, first reduce to the case $g = \mathbb{1}_{\{z\}}$ with $z \in \partial A$. For every $y \in \bar{A}$, write

$$\begin{aligned} f(y) = \mathbb{P}_y(X_{\tau_{\partial A}} = z) &= \sum_{n=0}^{\infty} \mathbb{P}_y(X_n = z, \tau_{\partial A} = n) \\ &= \mathbb{1}_{y=z} + \mathbb{1}_{y\in A} \sum_{n=1}^{\infty} \sum_{x_1,\dots,x_{n-1}\in A} \mathbf{P}(y, x_1)\mathbf{P}(x_1, x_2)\cdots\mathbf{P}(x_{n-1}, z). \end{aligned}$$

Moreover, since $f = 0$ on $\partial A \setminus \{z\}$ and Δ is local, for every $x \in A$,

$$(\mathbf{P}f)(x) = \sum_{y\in\mathbb{Z}^d} \mathbf{P}(x, y)f(y) = \mathbf{P}(x, z)f(z) + \sum_{y\in A} \mathbf{P}(x, y)f(y).$$

Consequently, for every $x \in A$,

$$\begin{aligned} (\mathbf{P}f)(x) &= \mathbf{P}(x, z)f(z) + \sum_{n=1}^{\infty} \sum_{y,x_1,\dots,x_{n-1}\in A} \mathbf{P}(x, y)\mathbf{P}(y, x_1)\cdots\mathbf{P}(x_{n-1}, z) \\ &= \mathbf{P}(x, z)f(z) + (f(x) - (\mathbb{1}_{x=z} + \mathbf{P}(x, z))) \\ &= f(x) \end{aligned}$$

where the last equality follows from $\mathbb{1}_{x=z} = 0$ and $f(z) = 1$. Thus $\mathbf{P}f = f$ on A, meaning $\Delta f = 0$ on A.

For uniqueness, linearity reduces the problem to showing that $f = 0$ is the unique solution when $g = 0$. If $f : \bar{A} \to \mathbb{R}$ is harmonic on A, interpreting $(\Delta f)(x)$ as the deviation from the nearest-neighbor average shows that both the minimum and maximum of f on $\bar{A}$ must be attained on the boundary ∂A (possibly elsewhere as well). Since f vanishes on ∂A, it follows that f vanishes on $\bar{A}$. □

Theorem 2.8 generalizes as follows:

Theorem 2.9 (The Dirichlet problem and Green's function)**.** *If $A \subset \mathbb{Z}^d$ is a finite nonempty set, then for every pair of functions $g : \partial A \to \mathbb{R}$ and $h : A \to \mathbb{R}$, the function $f : \bar{A} \to \mathbb{R}$ defined for every $x \in \bar{A}$ by*

$$f(x) = \mathbb{E}_x(g(X_{\tau_{\partial A}}) + \sum_{n=0}^{\tau_{\partial A}-1} h(X_n))$$

is the unique solution of

$$\begin{cases} f = g & \text{sur } \partial A, \\ \Delta f = -h & \text{sur } A. \end{cases}$$

When $d = 1$ and $h = 1$, we recover the function R from the proof of Theorem 2.2. Theorem 2.8 is the case $h = 0$. For every $x \in A$,

$$f(x) = \sum_{y \in \partial A} g(y) \mathbb{P}_x(\tau_{\partial A} = y) + \sum_{y \in A} h(y) G_A(x, y)$$

where $G_A(x, y)$ is the mean number of visits to y starting from x before exiting A, that is,

$$G_A(x, y) := \mathbb{E}_x \left(\sum_{n=0}^{\tau_{\partial A} - 1} \mathbb{1}_{X_n = y} \right) = \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = y, n < \tau_{\partial A}).$$

We call G_A the *Green function* of the simple symmetric random walk on A killed at the boundary ∂A. It is the inverse of the restriction $-\Delta_A$ of $-\Delta$ to functions on $\bar{A}$ vanishing on ∂A:

$$G_A = -\Delta_A^{-1}.$$

Indeed, if $g = 0$ and $h = \mathbb{1}_{\{y\}}$, then $f(x) = G_A(x, y)$, hence $\Delta_A G_A = -\mathbf{I}_A$.

Proof of Theorem 2.9. By Theorem 2.8 and linearity, it suffices to check that $f(x) = \mathbb{1}_{x \in A} G_A(x, z)$ is a solution when $g = 0$ and $h = \mathbb{1}_{\{z\}}$ with $z \in A$. For every $x \in A$, the Markov property gives

$$\begin{aligned} f(x) &= \mathbb{1}_{\{x=z\}} + \sum_{n=1}^{\infty} \mathbb{P}_x(X_n = z, n < \tau_{\partial A}) \\ &= \mathbb{1}_{\{x=z\}} + \sum_{y : |x-y|_1 = 1} \sum_{n=1}^{\infty} \mathbb{P}(X_n = z, n < \tau_{\partial A} \,|\, X_1 = y) \mathbf{P}(x, y) \\ &= \mathbb{1}_{\{x=z\}} + \sum_{u : |x-y|_1 = 1} f(y) \mathbf{P}(x, y). \end{aligned}$$

Thus $f = h + \mathbf{P}f$ on A, meaning $\Delta f = -h$ on A. □

The proofs of Theorems 2.8 and 2.9 remain valid for asymmetric random walks on $\mathbb{Z}^d$, provided the symmetric walk generator Δ is replaced by the generator $L := \mathbf{P} - \mathbf{I}$, a local operator: $|x - y| > 1 \Rightarrow L(x, y) = \mathbf{P}(x, y) = 0$. Beyond this setting, the notion of boundary must be adapted: $\{y \notin A : \exists x \in A, L(x, y) > 0\}$.

The *Gaussian free field*[4] is a random interface model related to the simple symmetric random walk and the Dirichlet problem. Let $A \subset \mathbb{Z}^d$ be a finite nonempty subset. An interface is a height function $f : \bar{A} \to \mathbb{R}$ assigning to every site $x \in \bar{A}$ a height $f(x)$, also called its *spin*. For simplicity, impose the boundary condition $f = 0$ on the outer boundary ∂A of A. Write $\mathcal{F}_A$ for the

[4] Abbreviated GFF.

set of interfaces f on $\bar{A}$ vanishing on the boundary ∂A, identified with $\mathbb{R}^A$. The energy $H_A(f)$ of interface $f \in \mathcal{F}_A$ is

$$H_A(f) = \frac{1}{4d} \sum_{\substack{\{x,y\} \subset \bar{A} \\ |x-y|_1 = 1}} (f(x) - f(y))^2,$$

where $f = 0$ on ∂A. The energy $H_A(f)$ is smaller when the interface f is "flatter." Writing

$$\langle u, v \rangle_A := \sum_{x \in A} u(x) v(x)$$

gives, for every $f \in \mathcal{F}_A$,

$$H_A(f) = \frac{1}{4d} \sum_{x \in A} \sum_{\substack{y \in \bar{A} \\ |x-y|_1 = 1}} (f(x) - f(y)) f(x) = \frac{1}{2} \langle -\Delta f, f \rangle_A.$$

Since $H_A(0) = 0$ and $H_A(f) = 0$ imply $f = 0$, the quadratic form H_A is nondegenerate. We can therefore define a Gaussian distribution Q_A on $\mathcal{F}_A$ favoring low energies:

$$Q_A(df) = \frac{1}{Z_A} e^{-H_A(f)}\, df \quad \text{où} \quad Z_A := \int_{\mathcal{F}_A} e^{-H_A(f)}\, df.$$

This Gaussian distribution, called the Gaussian free field, is characterized by its mean $m_A : A \to \mathbb{R}$ and covariance matrix $C_A : A \times A \to \mathbb{R}$, given, for every $x, y \in A$, by

$$m_A(x) := \int f_x \, Q_A(df) = 0$$

and

$$C_A(x,y) := \int f_x f_y \, Q_A(df) - m_A(x) m_A(y) = -(\Delta_A^{-1})(x,y) = G_A(x,y),$$

where f_x denotes the coordinate map $f_x : f \in \mathcal{F}_A \mapsto f(x) \in \mathbb{R}$.

2.4 Random walk on the symmetric group

Consider $r \geqslant 2$ playing cards stacked in a vertical deck and numbered from 1 to r. The top card is in position 1, and so on, with the bottom card in position r. We study a very simple card-shuffling procedure, called the *top-to-random shuffle*. More precisely, a k-insertion takes the top card and inserts it between positions k^{e} and $k+1^{\text{e}}$À 1-insertion has no effect. By convention, a r-insertion places the top card at the bottom. We perform successive k-insertions, using an i.i.d. uniform sequence on $\{1, \ldots, r\}$ to choose k.

A deck configuration is encoded by an element σ of the symmetric group Σ_r, so that $\sigma(k)$ denotes the position of card k. A k-insertion changes configuration σ into $(k, k-1, \dots, 1)\sigma$, where $(k, k-1, \dots, 1) \in \mathcal{S}_r$ denotes the k-cycle $k \to k-1 \to \cdots \to 1 \to k$.

Writing X_n for the deck configuration at time n gives a random sequence $(X_n)_{n \geqslant 0}$ in Σ_r satisfying, for every $n \geqslant 0$,

$$X_{n+1} = \varepsilon_{n+1} X_n = \varepsilon_{n+1} \cdots \varepsilon_1 X_0$$

where $(\varepsilon_n)_{n \geqslant 1}$ are i.i.d. uniformly distributed on $C := \{(k, k-1, \dots, 1) : 1 \leqslant k \leqslant r\} \subset \Sigma_r$. The sequence $(X_n)_{n \geqslant 0}$ is a left random walk on the nonabelian group Σ_r with increments C. It is also a Markov chain with finite state space Σ_r (of cardinality $r!$) and kernel

$$\mathbf{P}(\sigma, \sigma') = \frac{\mathbb{1}_C(\sigma' \sigma^{-1})}{|C|} = \frac{\mathbb{1}_C(\sigma' \sigma^{-1})}{r}.$$

1	$\sigma^{-1}(1)$
$\vdots$	$\vdots$
r	$\sigma^{-1}(r)$
Initial configuration	Configuration after shuffling

Fig. 2.4. The card in position k moves to position $\sigma(k)$.

Theorem 2.10 (Convergence in distribution)**.** *The chain $(X_n)_{n \geqslant 0}$ is recurrent, irreducible, and aperiodic. Its unique invariant distribution is uniform on Σ_r:*

$$\mu = \sum_{\sigma \in \Sigma_r} \frac{1}{|\Sigma_r|} \delta_\sigma = \frac{1}{r!} \sum_{\sigma \in \Sigma_r} \delta_\sigma.$$

The chain $(X_n)_{n \geqslant 0}$ converges in distribution to μ regardless of its initial distribution $\mathrm{Loi}(X_0)$.

Proof. The special set of transpositions

$$\mathcal{T} := \{(r, r-1), \dots, (2, 1), (1, r)\}$$

generates Σ_r. Since the k-insertions $(r, r-1, \dots, 1)$ and $(2, 1)$ corresponding to $k = r$ and $k = 2$ generate $\mathcal{T}$, they generate Σ_r. Hence C generates Σ_r and the chain is irreducible. Since the identity (1) is also a k-insertion $(k = 1)$, the diagonal of the transition matrix is > 0, so the chain is aperiodic. Since Σ_r is a group, exactly r states lead to each state. The columns of the transition matrix therefore have exactly r nonzero entries, all equal to $1/r$. Its transpose

is thus also a transition matrix[5], so the uniform distribution is invariant. Every finite recurrent aperiodic chain has a unique invariant distribution to which it converges in distribution from any initial law. □

We now study the rate at which the chain converges to its invariant distribution, starting from a fixed initial configuration X_0. By translation invariance, assume $X_0 = (1)$, meaning that all cards are initially in order. At time 0, card r is in position r (the bottom of the deck) and gradually rises to the top. Its ascent becomes faster and faster. For every $1 \leqslant k \leqslant r-1$, let T_k be the time this card spends in position $r-k$ (and $T_0 = 0$). For every $k \geqslant 1$, with the convention $T_1 + \cdots + T_{k-1} = 0$ if $k = 1$,

$$\begin{aligned} T_k &= \min\{n \geqslant 1 : X_{T_1+\cdots+T_{k-1}+n} = r-k\} \\ &= \min\{n \geqslant 1 : X_n(r) = r-k\} - (T_1 + \cdots + T_{k-1}). \end{aligned}$$

At time $T_1 + \cdots + T_{r-1}$, card r is in position 1 (the top). The random variable

$$T := 1 + T_1 + \cdots + T_{r-1}$$

has the coupon collector distribution with r uniformly drawn coupon types (Chapter 1). The random variables $T_1, \ldots, T_{r-1}$ are independent with $T_k \sim \mathrm{Geo}(k/r)$ for every $1 \leqslant k \leqslant r-1$. Note that $T \geqslant r$.

Theorem 2.11 (Well shuffled after the ascent)**.** *The time T is a* strong stationary time*: for every $n \geqslant 0$, the random variables X_{T+n} and T are independent, and X_{T+n} has the uniform distribution μ on Σ_r.*

Proof. The uniform distribution on Σ_r can be obtained by drawing the images of $1, \ldots, r$ uniformly without replacement. Moreover, the uniform distribution on Σ_r is translation invariant.

At time T_1, card r reaches position $r-1$ for the first time, because card 1 has been placed at the bottom. At time $T_1 + T_2$, card r first reaches position $r-2$ because card 2 has been inserted below or above card 1. The two bottom cards have numbers $1,2$ or $2,1$ with probability $1/2$. Inductively, at time $T_1 + \cdots + T_{r-1} = T-1$, card r first reaches the top, and the numbers of the $r-1$ cards below it are uniformly distributed without replacement in $\{1, \ldots, r-1\}$. At time T, card r is placed uniformly at random in the deck at a position between 1 and r, so X_T is uniform. Since the probability $\mathbb{P}(X_T = \sigma, T = k)$ does not depend on σ, T and X_T are independent. Since the distribution μ is invariant, we obtain $X_{T+n} \sim \mu$ for every $n \geqslant 0$. □

To shuffle the deck well, we could stop at time T. Unfortunately, T is unknown in practice! Alternatively, we could seek the smallest possible n

[5] The matrix is *doubly stochastic* (or *bistochastic*). The set of doubly stochastic $n \times n$ matrices is a compact convex polytope (an intersection of half-spaces) with $(n-1)^2$ degrees of freedom. Its extreme points are permutation matrices (Birkhoff and von Neumann). The symmetric group again!

for which $d_{\mathrm{VT}}\left(\mathrm{Loi}(X_n), \mu\right)$ is close to zero, where $d_{\mathrm{VT}}\left(\cdot, \cdot\right)$ is total variation distance (Section 1.3). For sufficiently large r, the quantity $d_{\mathrm{VT}}\left(\mathrm{Loi}(X_n), \mu\right)$ drops abruptly from 1 to 0[6] around $n = r\log(r)$. The next theorem quantifies this cutoff.

Theorem 2.12 (Cutoff at $n = r\log(r)$)**.** *For every real $c > 0$,*

$$d_{\mathrm{VT}}\left(\mathrm{Loi}(X_{\lfloor r\log(r)+cr\rfloor}), \mu\right) \leqslant e^{-c+\varepsilon_r} \quad \textit{avec} \quad 0 \leqslant \varepsilon_r < 1/r.$$

If $c_r \geqslant 0$ satisfies $\lim_{r\to\infty} c_r = \infty$ and $r\log(r) - rc_r > 0$ for every $r \geqslant 1$, then

$$\lim_{r\to\infty} d_{\mathrm{VT}}\left(\mathrm{Loi}(X_{r\log(r)-rc_r}), \mu\right) = 1.$$

Proof. By Theorem 2.11, for every $A \subset \Sigma_r$,

$$\begin{aligned}
\mathbb{P}(X_n \in A) &= \sum_{k=0}^{n} \mathbb{P}(X_n \in A, T = k) + \mathbb{P}(X_n = \sigma, T > n) \\
&\leqslant \sum_{k=0}^{n} \mathbb{P}(X_n \in A, T = k) + \mathbb{P}(T > n) \\
&= \sum_{k=0}^{n} \mathbb{P}(X_{T+n-k} \in A, T = k) + \mathbb{P}(T > n) \\
&= \sum_{k=0}^{n} \mu(A)\mathbb{P}(T = k) + \mathbb{P}(T > n) \\
&= \mu(A)\mathbb{P}(T \leqslant n) + \mathbb{P}(T > n) \\
&\leqslant \mu(A) + \mathbb{P}(T > n).
\end{aligned}$$

hence $\mathbb{P}(X_n \in A) - \mu(A) \leqslant \mathbb{P}(T > n)$. Applying this to A and A^c gives

$$d_{\mathrm{VT}}\left(\mathrm{Loi}(X_n), \mu\right) \leqslant \mathbb{P}(T > n), \quad n \geqslant 0.$$

The result follows from the coupon collector Theorem 1.17. □

Figure 1.2 illustrates the cutoff phenomenon.

2.5 Further reading

The simple random walk on $\mathbb{Z}^d$, studied notably by George Pólya and later Frank Spitzer, is one of probability's basic models. It receives particular attention in the books by William Feller [Fel68], Frank Spitzer [Spi70], Peter G. Doyle and Laurie Snell [DS84], and James Norris [Nor98a]. The proof of Theorem 2.7 is taken from an article [CF51] by Kai Lai Chung and Wolfgang

[6] This is called the *cutoff phenomenon*.

Heinrich Johannes Fuchs. This Fourier analysis can be developed further, as explained in the books by Gregory Lawler and Vlada Limic [Law13, LL10]. Désiré André's combinatorial trick appears in [And87].

Catalan numbers occur very frequently in combinatorics. Besides simple random walk paths, they count Dyck words, balanced parenthesis expressions, polygon triangulations, noncrossing partitions, subdiagonal paths in a square, plane trees, and many other objects.

Statistical physics has inspired many random walk models: walks in random environments, random scenery, and self-interacting walks (self-avoiding, reinforced, excited,...). This rich classical subject remains an active research area. Chapters 3, 6, and 15 involve random walks. Moreover, the central limit theorem allows Brownian motion to be viewed as a continuous-time and continuous-space analogue of the simple random walk, obtained as a scaling limit of discrete models ($\varepsilon\mathbb{Z}^d$ approximates $\mathbb{R}^d$). The discrete Dirichlet problem is studied in detail in Gregory Lawler's book [Law13]. Its continuous-time and continuous-space version, studied in Chapter 24, is a scaling limit of the discrete problem. At the process level, the simple symmetric random walk becomes Brownian motion through the central limit theorem, while at the generator level, the discrete Laplacian becomes the Laplace differential operator through Taylor's formula. The Gaussian free field is a fundamental object of mathematical physics, discussed for example in the book by James Glimm and Arthur Jaffe [GJ87]; its scaling limit is presented in Scott Sheffield's article [She07].

Although the card-shuffling model in this chapter is not very realistic, it gives a very simple route to the important cutoff phenomenon in convergence to equilibrium, shared by most studied shuffling procedures. One example is the *riffle shuffle*: split the deck in two and merge the subdecks by interleaving their cards. Cutoff in card shuffling belongs to the broader study of convergence to equilibrium for Markov chains, extensively developed by Persi Diaconis and his collaborators and beautifully presented in the books by David Aldous and James Fill [AF01] and David Levin, Yuval Peres, and Elizabeth Wilmer [LPW09].

The coupon collector also appears in the analysis of random walk on the hypercube (Section 9.1), a remarkable example of a random walk on a regular graph. Random walks can also be defined on more general directed or undirected graphs by specifying, for each vertex, a distribution on its neighbors. These walks are Markov chains. Conversely, every Markov chain with at most countable state space can be viewed as a random walk on its transition graph. Google's PageRank algorithm is based on a random walk on the directed graph of web pages. Another classical example is random walk on the Cayley graph of a finitely generated group. Finally, random walks on groups can be defined using i.i.d. "increments."

3

Branching and Galton–Watson processes

Keywords. Random tree; branching; genealogy; population dynamics; threshold phenomenon. **Tools.** Generating function; Laplace transform; martingale; Markov chain; random walk. **Difficulty.** *

Branching processes form an important class arising in many probabilistic situations. This chapter introduces the Galton–Watson branching process, whose role in probability is almost as important as that of random walks.

The Galton–Watson process describes the evolution of an asexual population over nonoverlapping generations. Let $P := p_0\delta_0 + p_1\delta_1 + \cdots$ be a distribution on $\mathbb{N}$ called the *offspring distribution*. Generation n produces generation $n+1$ as follows: every individual in generation n independently produces a random number of children with distribution P and then dies. If Z_n denotes the number of individuals in generation $n \in \mathbb{N}$, then Z_0 is the initial population size and Z_{n+1} satisfies the recurrence

$$Z_{n+1} = \sum_{k=1}^{Z_n} X_{n+1,k}$$

where $(X_{n,k})_{n\geqslant 1, k\geqslant 1}$ are i.i.d. random variables with distribution P, independent of Z_0. We adopt the convention $\sum_{\varnothing} = 0$, so that if $Z_n = 0$ then $Z_{n+1} = 0$. For every $n \in \mathbb{N}$ and $z_0, \ldots, z_n \in \mathbb{N}$,

$$\mathrm{Loi}(Z_{n+1} \,|\, Z_0 = z_0, \ldots, Z_n = z_n) = \mathrm{Loi}(Z_{n+1} \,|\, Z_n = z_n) = P^{*z_n}.$$

The sequence $(Z_n)_{n\geqslant 0}$ is therefore a Markov chain with state space $\mathbb{N}$ and transition kernel $\mathbf{P}$ given, for every $z \in \mathbb{N}$, by

$$\mathbf{P}(z, \cdot) = P^{*z}.$$

State 0 is absorbing (population extinction). The extinction time is

$$T := \inf\{n \geqslant 0 : Z_n = 0\} \in \mathbb{N} \cup \{\infty\}.$$

Remark 3.1 (Original motivations)**.** *One original motivation for this model was to estimate the survival probability of a family's male line, which probably explains why the model is asexual.*

Remark 3.2 (The Galton–Watson tree)**.** *The sequence* $(Z_n)_{n\geqslant 0}$ *starting from* $Z_0 = 1$ *can be represented as a random tree with offspring distribution* P*. The variable* Z_n *is the number of nodes at depth* n*. An* r*-ary tree corresponds to* $P = \delta_r$*, but it is not a* r*-regular graph: the root (the common ancestor) has* r *neighbors, whereas all other vertices (its descendants) have* $r+1$ *neighbors. The tree has* depth *or* height $T = \inf\{n \in \mathbb{N} : Z_n = 0\}$*,* width $\max_{n\in\mathbb{N}} Z_n$*, and* total size $N = Z_0 + \cdots + Z_{T-1}$*. The tree in Figure 3.2 has depth* 3*, width* 5*, and total size* 12*. When* $Z_0 > 1$*, one sometimes speaks of a* random forest*.*

We make the following observations:

- The behavior of $(Z_n)_{n\geqslant 0}$ reduces to the case $Z_0 = 1$, since conditionally on $\{Z_0 = z\}$, the sequence $(Z_n)_{n\geqslant 0}$ has the distribution of a sum of z i.i.d. copies starting from 1. This is the *branching property*:

$$Z^{(x+y)} \overset{\text{loi}}{=} Z^{(x)} + Z^{(y)}$$

  where $Z^{(x)}$ and $Z^{(y)}$ are independent;
- If $p_z = 1$ for some $z \in \mathbb{N}$, then $P = \delta_z$, hence for every $n \geqslant 0$,

$$Z_{n+1} = zZ_n = \cdots = z^{n+1}Z_0;$$

- If $p_0 = 0$ and $p_1 < 1$, then $\mathbb{P}(Z_n \nearrow \infty \,|\, Z_0 = 1) = 1$ by comparison with coin tossing;
- If $p_0 + p_1 = 1$ and $p_0 \neq 0$, then $\mathbb{P}(Z_n \searrow 0 \,|\, Z_0 = 1) = 1$ by comparison with coin tossing, and the extinction time T is geometric.

These observations lead us to assume throughout the sequel that

$$Z_0 = 1 \quad \text{et} \quad 0 < p_0 \leqslant p_0 + p_1 < 1.$$

This implies $p_z < 1$ for every $z \in \mathbb{N}$. It also implies $Z_1 \sim P$, since $Z_0 = 1$. Throughout, write $m := p_1 + 2p_2 + \cdots$ for the mean of P when it exists, and $\sigma^2 \in \overline{\mathbb{R}}_+$ for the variance of P when m exists.

3.1 Extinction and the threshold phenomenon

We begin with the first two moments of Z_n.

Theorem 3.3 (Mean and variance)**.**

- *if* m *exists, then* $\mathbb{E}(Z_n) = m^n$ *for every* $n \geqslant 0$*;*

- *if m exists and $\sigma^2 < \infty$, then for every $n \geqslant 0$,*

$$\mathrm{Var}(Z_n) = \begin{cases} \sigma^2 m^{n-1} \dfrac{(m^n - 1)}{m - 1} & \text{si } m \neq 1; \\ n\sigma^2 & \text{si } m = 1. \end{cases}$$

Proof. The expectation formula follows from $\mathbb{E}(Z_{n+1} \,|\, Z_n) = mZ_n$ and $Z_0 = 1$. For the variance, $\mathbb{E}(Z_{n+1}^2 \,|\, Z_n) = Z_n(\sigma^2 + m^2) + Z_n(Z_n - 1)m^2$ gives

$$\mathbb{E}(Z_{n+1}^2) = \mathbb{E}(Z_n)(\sigma^2 + m^2) + (\mathbb{E}(Z_n^2) - \mathbb{E}(Z_n))m^2,$$

and a calculation yields the recurrence $\mathrm{Var}(Z_{n+1}) = \mathrm{Var}(Z_n)m^2 + m^n\sigma^2$. □

The asymptotic behavior of the mean size $\mathbb{E}(Z_n) = m^n$ depends only on the parameter m and reveals a threshold phenomenon around the critical value $m = 1$. We naturally adopt the following terminology according to the parameter m:

- subcritical case: $m < 1$,

$$\mathbb{E}(Z_n) \searrow 0 \quad \text{avec} \quad \mathrm{Var}(Z_n) \to 0 \text{ si } \sigma^2 < \infty;$$

- critical case: $m = 1$,

$$\mathbb{E}(Z_n) = 1 \quad \text{avec} \quad \mathrm{Var}(Z_n) \to \infty \text{ si } \sigma^2 < \infty;$$

- supercritical case: $m > 1$,

$$\mathbb{E}(Z_n) \nearrow \infty \quad \text{avec} \quad \mathrm{Var}(Z_n) \to \infty \text{ si } \sigma^2 < \infty.$$

The expectation $\mathbb{E}(Z_n)$ and variance $\mathrm{Var}(Z_n)$ of Z_n can be obtained using a generating function. Recall that the generating function $g : [0,1] \mapsto [0,1]$ of the distribution P is the real power series

$$g(s) = \mathbb{E}(s^{X_{1,1}}) = \sum_{z=0}^{\infty} p_z s^z.$$

If P has a moment of order r, then

$$\lim_{s \to 1} g^{(r)}(s) = \mathbb{E}(X_{1,1}(X_{1,1} - 1) \cdots (X_{1,1} - r + 1))$$

(the *factorial moment of order r of P*); in particular,

$$g'(1^-) = m \quad \text{et} \quad g''(1^-) = \sigma^2 + m^2 - m.$$

Theorem 3.4 (Generating function of Z_n)**.** *For every $n \geqslant 1$, the generating function g_n of Z_n satisfies $g_n = g^{\circ n} := \underbrace{g \circ \cdots \circ g}_{n \text{ fois}}$.*

Proof. We have $g_1 = g$ since $Z_0 = 1$. For every $n \geqslant 1$ and every $s \in [0,1]$,

$$\begin{aligned} g_{n+1}(s) &= \mathbb{E}(s^{Z_{n+1}}) = \mathbb{E}(\mathbb{E}(s^{Z_{n+1}}|Z_n)) \\ &= \mathbb{E}((\mathbb{E}(s^{X_{1,1}}))^{Z_n}) = \mathbb{E}(g(s)^{Z_n}) = g_n(g(s)). \end{aligned}$$

□

Remark 3.5 (Connections between Galton–Watson and Wright–Fisher)**.** *Suppose the offspring distribution P is Poisson* $\mathrm{Poi}(\lambda)$*. For every fixed integer $N \geqslant 1$, the distribution of $(X_{1,1}, \ldots, X_{1,N})$ given $X_{1,1} + \cdots + X_{1,N} = N$ is the multinomial distribution* $\mathrm{Mul}(N, (1/N, \ldots, 1/N))$*. Indeed, for $(k_1, \ldots, k_N)$ in $\mathbb{N}^N$ such that $k_1 + \cdots + k_N = N$, the probability*

$$\mathbb{P}(X_{1,1} = k_1, \ldots, X_{1,N} = k_N | X_{1,1} + \cdots + X_{1,N} = N)$$

equals (since $X_{1,1} + \ldots + X_{1,N}$ is Poisson with parameter λN)

$$\frac{\mathbb{P}(X_{1,1} = k_1, \ldots, X_{1,N} = k_N)}{\mathbb{P}(X_{1,1} + \cdots + X_{1,N} = N)} = \frac{N!}{k_1! \ldots k_N!}.$$

Thus the Wright–Fisher model of Chapter 12 is related to the Galton–Watson model with Poisson offspring conditioned to have constant size (Remark 13.1). The Galton–Watson process describes population size evolution and says nothing at the individual level. Conversely, if $(Y_1, \ldots, Y_N)$ has the multinomial distribution $\mathrm{Mul}(N, (1/N, \ldots, 1/N))$*, then for every fixed $k \geqslant 1$, $(Y_1, \ldots, Y_k)$ converges in distribution as $N \to \infty$ to the product distribution* $\mathrm{Poi}(1)^{\otimes k}$*. This indicates that individuals in a large Wright–Fisher population tend to produce children independently with Poisson offspring counts.*

Theorem 3.6 (Dichotomy)**.** *Almost surely, the process either becomes extinct:*

$$\lim_{n\to\infty} Z_n = 0$$

or tends to infinity (explosion):

$$\lim_{n\to\infty} Z_n = \infty.$$

In other words, almost surely,

$$\lim_{n\to\infty} Z_n = \begin{cases} \infty & si\ T = \infty; \\ 0 & si\ T < \infty. \end{cases}$$

Proof. Since $p_0 > 0$, we have $\mathbf{P}(z,0) = p_0^z > 0$ for every $z \in \mathbb{N}^*$. Since 0 is absorbing, every $z \in \mathbb{N}^*$ is transient. Almost surely, the chain visits each transient state only finitely often. Thus, almost surely, it is either trapped in the absorbing state 0 or diverges to infinity. In other words, almost surely the process eventually leaves every finite interval of $\mathbb{N}^*$. □

Theorem 3.6 suggests studying the *extinction probability* $\mathbb{P}(T < \infty)$.

Theorem 3.7 (Extinction probability)**.** *The extinction probability* $\mathbb{P}(T < \infty)$ *is a fixed point of* g*. If* $m \leqslant 1$*, then* g *has a unique fixed point, equal to* 1*, and* $\mathbb{P}(T < \infty) = 1$*. If* $m > 1$*, then* g *has two fixed points,* $s \in]0, 1[$ *(attracting) and* 1 *(repelling), and* $\mathbb{P}(T < \infty) = s$*.*

Proof. Since Z_n is integer-valued,

$$\{T < \infty\} = \{\lim_{n\to\infty} Z_n = 0\} = \cup_{n\geqslant 0}\{Z_n = 0\}.$$

The sequence $(\{Z_n = 0\})_{n\geqslant 0}$ is nondecreasing and the extinction probability satisfies

$$\mathbb{P}(T < \infty) = \mathbb{P}(\lim_{n\to\infty} Z_n = 0) = \mathbb{P}(\cup_{n\geqslant 0}\{Z_n = 0\}) = \lim_{n\to\infty} \mathbb{P}(Z_n = 0).$$

Theorem 3.4 gives $\mathbb{P}(Z_n = 0) = g_n(0) = g^{\circ n}(0)$, hence

$$\mathbb{P}(T < \infty) = \lim_{n\to\infty} g_n(0) = \lim_{n\to\infty} g^{\circ n}(0).$$

The extinction probability $\rho := \mathbb{P}(T < \infty) = \lim_{n\to\infty} g^{\circ n}(0)$ satisfies $g(\rho) = \rho$ because $g : [0, 1] \to [0, 1]$ is continuous. The function g is convex since

$$g''(s) = \sum_{z=0}^{\infty}(z + 2)(z + 1)s^z p_{z+2}.$$

Since $p_0 + p_1 < 1$, the function g is actually strictly convex (it is affine when $p_0 + p_1 = 1$). Moreover, $g(0) = p_0 \in]0, 1[$ and $g(1) = 1$. Consequently, if $g'(1^-) = m \leqslant 1$, the graph of g lies above the diagonal, and 1 is the only fixed point. If $g'(1^-) = m > 1$, the graph of g crosses the diagonal exactly once in the interval $]0, 1[$, and g has a second fixed point $s \in]0, 1[$. Finally, if $m > 1$ then $g'(1^-) = m > 1$, so 1 is not an attracting fixed point. Figure 3.1 illustrates special cases from Examples 3.8 and 3.9. □

Example 3.8 (Poisson offspring)**.** *If* $P = \mathrm{Poi}(\lambda)$*, then* $m = \lambda$ *and* $g(s) = e^{\lambda(s-1)}$*. Thus, if* $\lambda \leqslant 1$ *the population becomes extinct almost surely, whereas if* $\lambda > 1$ *the extinction probability* s *is the unique solution in* $]0, 1[$ *of* $\lambda = \log(s)/(s - 1)$*.*

Example 3.9 (Geometric offspring)**.** *If* $P = \mathrm{Geo}_{\mathbb{N}}(p) = \sum_{n\geqslant 0} q^n p\delta_n$*, then* $m = q/p$ *and* $g(s) = p/(1 - sq)$*. Thus, if* $p \geqslant 1/2$ *the population becomes extinct almost surely, whereas if* $p < 1/2$ *the extinction probability* s *is the unique solution in* $]0, 1[$ *of* $qs^2 - s + p = 0$*. A pleasant little calculation gives* $s = p/q$*.*

Theorem 3.7 therefore gives the following classification:

- subcritical ($m < 1$): the population becomes extinct almost surely;

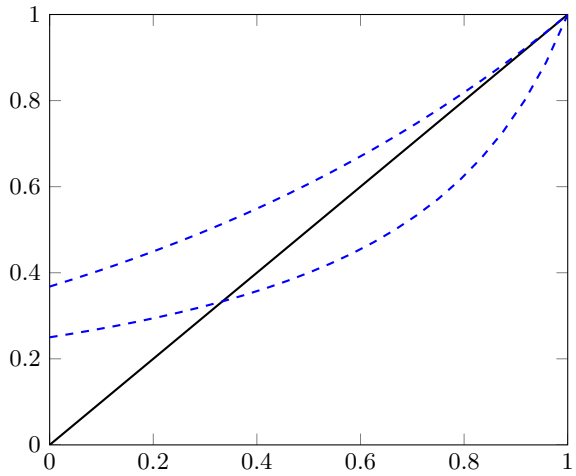


Fig. 3.1. Dashed curves: the generating function of the geometric distribution with parameter 1/4 (two fixed points: 1/3, attracting, and 1, repelling), and that of the Poisson distribution with parameter 1 (a unique fixed point at 1). The solid line is the diagonal.

- critical ($m = 1$): the population becomes extinct almost surely;
- supercritical ($m > 1$): the population becomes extinct with probability $p \in \,]0, 1[$.

To study $(Z_n)_{n\geqslant 0}$ more precisely, introduce $(Y_n)_{n\geqslant 0}$ defined by

$$Y_n = \frac{Z_n}{\mathbb{E}(Z_n)} = \frac{Z_n}{m^n}.$$

It is useful because it is a martingale for the natural filtration:

$$\mathbb{E}(Y_{n+1} \,|\, \sigma(Y_0, \dots, Y_n)) = \frac{\mathbb{E}(X_{n+1,1} + \cdots + X_{n+1,Z_n} \,|\, Z_n)}{m^{n+1}} = \frac{mZ_n}{m^{n+1}} = Y_n.$$

We have $\mathbb{E}(Y_n) = 1$. Since $(Y_n)_{n\geqslant 0}$ is a nonnegative martingale, it converges almost surely to a nonnegative integrable real random variable Y_∞; in particular, $Y_\infty < \infty$ almost surely. On the event $\{Y_\infty > 0\}$, we have $Z_n \sim_{n\to\infty} m^n Y_\infty$ almost surely.

Since $Y_n = 0$ if $n \geqslant T$, we obtain $\{T < \infty\} \subset \{Y_\infty = 0\}$ almost surely. By Theorem 3.7, in the subcritical ($m < 1$) and critical ($m = 1$) cases, $\mathbb{P}(T < \infty) = 1$, hence $\mathbb{P}(Y_\infty > 0) = 0$. Nevertheless, if $m \leqslant 1$, then $\sup_{n\geqslant 0} \frac{Z_n}{m^n}$ is not integrable: otherwise dominated convergence would give $0 = \mathbb{E}(Y_T) = \mathbb{E}(\lim_{n\to\infty} Y_n) = \lim_{n\to\infty} \mathbb{E}(Y_n) = 1$.

Remark 3.10 (Martingales)**.** *Part of Theorem 3.7 can be recovered using martingales. Indeed, if $m < 1$, then*

$$Y_n \geqslant Y_n \mathbb{1}_{\{T=\infty\}} \geqslant m^{-n} \mathbb{1}_{\{T=\infty\}}$$

since $Z_n \geqslant 1$ *if* $n < T$*, and consequently*

$$1 = \mathbb{E}(Y_0) = \mathbb{E}(Y_n) \geqslant m^{-n} \mathbb{P}(T = \infty),$$

hence $\mathbb{P}(T < \infty) = 1$*. When* $m = 1$*, we have* $Z_n = Y_n \to Y_\infty < \infty$ *almost surely, hence* $Z_n \not\to \infty$ *almost surely, and therefore* $\mathbb{P}(T < \infty) = 1$ *(Theorem 3.6). Now suppose* $m > 1$*, and let* s *be the smallest fixed point of* g*. The sequence*

$$(s^{Z_n})_{n \geqslant 0}$$

is a martingale taking values in $[0, 1]$*. Since* $Z_n \to \infty$ *almost surely on* $\{T = \infty\}$ *(Theorem 3.6),* $Z_n = 0$ *if* $n \geqslant T$*, and* $0 < s < 1$*, we obtain*

$$s^{Z_n} = s^{Z_n} \mathbb{1}_{\{T=\infty\}} + s^{Z_n} \mathbb{1}_{\{T<\infty\}} \xrightarrow[n\to\infty]{p.s.} \mathbb{1}_{\{T<\infty\}}.$$

Dominated convergence gives $s = \mathbb{E}(s^{Z_0}) = \mathbb{E}(s^{Z_n}) \to \mathbb{P}(T < \infty)$*.*

3.2 The three possible cases

The subcritical case

If $m < 1$, then $\mathbb{P}(T < \infty) = 1$ and $Z_n = Z_{n \wedge T} \to Z_T = 0$ almost surely (Theorem 3.7), while $\mathbb{E}(Z_n) = m^n \to 0$ and $\mathrm{Var}(Z_n) \to 0$ if $\sigma^2 < \infty$ (Theorem 3.3).

Theorem 3.11 (Yaglom)**.** *If* $m < 1$*, then* $\mathrm{Loi}(Z_n \,|\, Z_n > 0)$ *converges as* $n \to \infty$ *to a distribution* μ *on* $\mathbb{N}^*$ *whose generating function* h *satisfies*

$$(h - 1) \circ g = m(h - 1).$$

Proof. Let h_n be the generating function of $\mu_n := \mathrm{Loi}(Z_n \,|\, Z_n > 0)$. The sequence $(h_n(1))_{n \geqslant 1}$ is constant and equal to 1, and therefore converges to $h(1) := 1$. For every $s \in [0, 1[$,

$$1 - h_n(s) = 1 - \frac{\mathbb{E}(s^{Z_n} \mathbb{1}_{\{Z_n > 0\}})}{\mathbb{P}(Z_n > 0)} = 1 - \frac{\mathbb{E}(s^{Z_n}) - \mathbb{P}(Z_n = 0)}{\mathbb{P}(Z_n > 0)} = \frac{1 - g_n(s)}{1 - g_n(0)}$$

since $\mathbb{P}(Z_n > 0) = 1 - \mathbb{P}(Z_n = 0) = 1 - g_n(0)$. Recall that $g_{n+1} = g \circ g_n$, and that $g_n(s) \to 1$ and $g_n(0) \to 1$ as $n \to \infty$, because 1 is the unique fixed point of g since $m < 1$ (Theorem 3.7). Since the functions g_n and

$$s \mapsto r(s) := \frac{1 - g(s)}{1 - s}$$

are increasing on $[0, 1[$, we have

$$\frac{1-h_{n+1}(s)}{1-h_n(s)} = \frac{1-g_{n+1}(s)}{1-g_n(s)}\frac{1-g_n(0)}{1-g_{n+1}(0)} = \frac{r(g_n(s))}{r(g_n(0))} \geqslant 1.$$

Thus the nonnegative sequence $(1-h_n(s))_{n\geqslant 1}$ is nondecreasing and, being bounded above by 1, converges to a limit $1-h(s)$. We have

$$1-h_n(g(s)) = \frac{1-g_{n+1}(0)}{1-g_n(0)}\frac{1-g_{n+1}(s)}{1-g_{n+1}(0)}.$$

Since $g_n(0) \nearrow \mathbb{P}(T<\infty) = 1$, we deduce that

$$\frac{1-g_{n+1}(0)}{1-g_n(0)} = \frac{g(1)-g(g_n(0))}{g(1)-g_n(0)} \to g'(1) = m.$$

Hence $(1-h)\circ g = m(1-h)$ on $[0,1[$ and therefore on $[0,1]$. It remains to show that h is the generating function of a probability distribution μ on $\mathbb{N}^*$, and that $(\mu_n)_{n\geqslant 1}$ converges weakly to μ. Since $(h_n)_{n\geqslant 1}$ converges pointwise on $[0,1[$ to h, a standard argument shows that $\mu_n(z)$ converges to a limit $\mu(z) \in [0,1]$ for every $z \in \mathbb{N}^*$, whose generating function equals h on $[0,1[$. The fact that the subprobability μ is a probability measure follows from the functional equation satisfied by h by taking the limit at 1. □

The critical case

If $m = 1$, then $\mathbb{P}(T<\infty) = 1$ and $Z_n = Z_{n\wedge T} \to Z_T = 0$ almost surely (Theorem 3.7), while $\mathbb{E}(Z_n) = 1$ and $\mathrm{Var}(Z_n) \to \infty$ if $\sigma^2 < \infty$ (Theorem 3.3). The tree's "width" $\sup_{n\geqslant 0} Z_n$ is not integrable, since dominated convergence would otherwise imply $1 = \lim_{n\to\infty} \mathbb{E}(Z_n) = \mathbb{E}(\lim_{n\to\infty} Z_n) = 0$, a contradiction. Critical trees are therefore often very wide, but their depth is almost surely finite: a passing bout of grandeur inevitably overtaken by insufficient fertility. A critical case indeed! Moreover, if $m = 1$ then $(Z_n)_{n\geqslant 0} = (Y_n)_{n\geqslant 0}$, recovering $Y_\infty = 0$ almost surely.

Theorem 3.12 (The critical case)**.** *If $m = 1$ and $\sigma^2 < \infty$, then*

1. $\lim_{n\to\infty} n\mathbb{P}(Z_n > 0) = 2\sigma^{-2}$
2. $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}(Z_n \mid Z_n > 0) = \frac{1}{2}\sigma^2$
3. $\mathrm{Loi}(n^{-1}Z_n \mid Z_n > 0)$ *converges as $n \to \infty$ to the distribution* $\mathrm{Exp}(2\sigma^{-2})$.

Item 2 does not follow directly from item 3, because convergence in distribution does not imply convergence of moments, particularly expectations. A counterexample is $(Z_n)_{n\geqslant 1}$, which converges almost surely to 0, hence in distribution to δ_0, while $1 = \mathbb{E}(Z_n)$ does not converge to the mean of δ_0, which is 0. The theorem nevertheless asserts that the expectation of $n^{-1}Z_n$ converges to the expectation $2\sigma^{-2}$ of the limiting distribution of $n^{-1}Z_n$. Convergence in Wasserstein distance of order p is equivalent to convergence in distribution together with convergence of moments up to order p. Moment convergence plays an important role in the proof of Wigner's theorem in Chapter 21.

Proof. We have

$$n\mathbb{P}(Z_n > 0) = n(1 - g_n(0)) = \left(\frac{1}{n}\left(\frac{1}{1 - g_n(0)} - 1\right) + \frac{1}{n}\right)^{-1}.$$

Taylor's formula with integral remainder at 1 to order 2 gives

$$g(s) = s + \frac{1}{2}(1-s)^2\sigma^2 + (1-s)^2\alpha(s)$$

with α bounded on $[0,1]$ and $\lim_{s\to 1}\alpha(s) = 0$. It follows that

$$\begin{aligned}\frac{1}{1-g(s)} - \frac{1}{1-s} &= \frac{g(s)-s}{(1-g(s))(1-s)} \\ &= \frac{\frac{1}{2}\sigma^2 + \alpha(s)}{1 - \frac{1}{2}(1-s)\sigma^2 - (1-s)\alpha(s)} = \frac{\sigma^2}{2} + \beta(s)\end{aligned}$$

where β behaves like α. Since $(g_k)_{k\geqslant 1}$ converges uniformly to 1, we obtain

$$\lim_{n\to\infty}\frac{1}{n}\left(\frac{1}{1-g_n(s)} - \frac{1}{1-s}\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\left(\frac{1}{1-g(g_k(s))} - \frac{1}{1-g_k(s)}\right) = \frac{\sigma^2}{2}$$

uniformly for every $s \in [0,1[$. Similarly, since $Z_n\mathbb{1}_{\{Z_n>0\}} = Z_n$,

$$\mathbb{E}(n^{-1}Z_n \,|\, Z_n > 0) = \frac{\mathbb{E}(Z_n\mathbb{1}_{\{Z_n>0\}})}{n\mathbb{P}(Z_n>0)} = \frac{\mathbb{E}(Z_n)}{n(1-g_n(0))} \to \frac{\sigma^2}{2}.$$

In the same way, for every $t \in \mathbb{R}_+^*$, writing $s_n = e^{-t/n}$ gives

$$\begin{aligned}\mathbb{E}(e^{-t\frac{1}{n}Z_n} \,|\, Z_n > 0) &= \frac{\mathbb{E}(s_n^{Z_n}\mathbb{1}_{\{Z_n>0\}})}{\mathbb{P}(Z_n>0)} \\ &= \frac{g_n(s_n) - g_n(0)}{1 - g_n(0)} = 1 - \frac{1-g_n(s_n)}{1-g_n(0)}\end{aligned}$$

which can also be written as

$$1 - \frac{1}{n(1-g_n(0))}\left(\frac{1}{n}\left(\frac{1}{1-g_n(s_n)} - \frac{1}{1-s_n}\right) + \frac{1}{n(1-s_n)}\right)^{-1}$$

Thus

$$\begin{aligned}\mathbb{E}(e^{-t\frac{1}{n}Z_n} \,|\, Z_n > 0) &\to 1 - \frac{\sigma^2}{2}\left(\frac{\sigma^2}{2} + \frac{1}{t}\right)^{-1} \\ &= \frac{1}{1 + \frac{\sigma^2}{2}t} \\ &= \frac{2}{\sigma^2}\int_0^\infty e^{-tx}e^{-\frac{2}{\sigma^2}x}\,dx.\end{aligned}$$

□

The supercritical case

If $m > 1$, then $0 < \mathbb{P}(T < \infty) < 1$ (Theorem 3.7), with $Z_n \to 0$ on the event $\{T < \infty\}$ and $Z_n \to \infty$ on $\{T = \infty\}$ (Theorem 3.6). Moreover, $\mathbb{E}(Z_n) = m^n \to \infty$ and $\mathrm{Var}(Z_n) \to \infty$ if $\sigma^2 < \infty$ (Theorem 3.3). Finally, if $m > 1$, then on $\{T < \infty\}$ we have $Z_n \to 0$, while on $\{Y_\infty > 0\}$ we have $Z_n \sim_{n\to\infty} m^n Y_\infty$. We also have $\{T < \infty\} \subset \{Y_\infty = 0\}$ almost surely, and the following elegant theorem states that this almost sure inclusion is an almost sure equality.

Theorem 3.13 (Explosion in the supercritical case)**.** *If $\mathbb{E}(Y_\infty) = 1$, then $\{Y_\infty = 0\} = \{T < \infty\}$ almost surely, and hence almost surely*

- *either $Y_\infty = 0$ and $\lim_{n\to\infty} Z_n = 0$;*
- *or $Y_\infty > 0$ and $Z_n \sim_{n\to\infty} m^n Y_\infty \nearrow +\infty$.*

Proof. Since $\mathbb{E}(Y_\infty) = 1$ and $Y_\infty \geqslant 0$, we have $\mathbb{P}(Y_\infty > 0) > 0$. On $\{Y_\infty > 0\}$, we have $Z_n \sim_{n\to\infty} m^n Y_\infty \to \infty$, recovering $\mathbb{P}(T < \infty) < 1$. To show that the almost sure inclusion $\{Y_\infty > 0\} \subset \{T = \infty\}$ is an almost sure equality, we prove that

$$\mathbb{P}(Y_\infty = 0) = \mathbb{P}(T < \infty).$$

Indeed, for every $z \in \mathbb{N}^*$, by the Markov and branching properties,

$$\begin{aligned}\mathbb{P}(Y_\infty = 0 \,|\, Z_1 = z) &= \mathbb{P}(\lim_{n\to\infty} Y_n = 0 \,|\, Z_1 = z) \\ &= \mathbb{P}(\lim_{n\to\infty} Y_n = 0)^z = \mathbb{P}(Y_\infty = 0)^z,\end{aligned}$$

and since $Z_1 \sim P$, we have $\mathbb{P}(Y_\infty = 0) = g(\mathbb{P}(Y_\infty = 0))$, so $\mathbb{P}(Y_\infty = 0)$ is a fixed point of g. Since $\mathbb{P}(Y_\infty = 0) < 1$, Theorem 3.7 gives $\mathbb{P}(Y_\infty = 0) = \mathbb{P}(T < \infty)$, completing the proof. □

Fatou's lemma gives $\mathbb{E}(Y_\infty) \leqslant \liminf_{n\to\infty} \mathbb{E}(Y_n) = 1$, ensuring that Y_∞ is integrable, but the convergence of $(Y_n)_{n\geqslant 0}$ to Y_∞ need not hold in L^1. Finally, Scheffé's lemma says that convergence holds in L^1 if and only if $\mathbb{E}(Y_\infty) = 1$.

Theorem 3.14 (Distribution of Y_∞ in the supercritical case)**.** *If $m > 1$ and $\sigma^2 < \infty$, then the martingale $(Y_n)_{n\geqslant 0}$ converges almost surely and in L^2 to a random variable $Y_\infty \geqslant 0$ with mean $\mathbb{E}(Y_\infty) = 1$ and variance $\mathrm{Var}(Y_\infty) = \sigma^2/(m^2 - m)$. Moreover, the Laplace transform $t \in \mathbb{R}_+ \mapsto L_\infty(t) = \mathbb{E}(e^{-tY_\infty})$ of Y_∞ is characterized by the following properties:*

$$L'_\infty(0) = -1 \quad \textit{et} \quad L_\infty(mt) = g(L_\infty(t)) \quad \textit{pour tout } t \in \mathbb{R}_+.$$

Proof. The martingale $(Y_n)_{n\geqslant 0}$ is bounded in L^2 since

$$\mathbb{E}(Y_n^2) = \frac{\mathrm{Var}(Z_n) + \mathbb{E}(Z_n)^2}{m^{2n}} = \frac{\sigma^2}{m^2 - m} - \frac{\sigma^2}{m^n(m^2 - m)} + 1$$

(Theorem 3.3), which converges because $m > 1$. Thus $(Y_n)_{n\geqslant 0}$ converges almost surely and in L^2 to a real random variable Y_∞ whose first two moments

are the limits of those of Y_n. One can also argue directly without a martingale theorem, because the convergence L^2 is sufficiently fast here. Indeed, Theorem 3.3 gives, for every $n, k \geqslant 0$,

$$\mathbb{E}((Y_{n+k} - Y_n)^2) = \frac{\sigma^2}{m^n} \frac{1 - m^{-k}}{m^2 - m}.$$

Since $m > 1$, this shows that $(Y_n)_{n \geqslant 0}$ is a Cauchy sequence in L^2. Since L^2 is complete, it converges to a real random variable $Y_\infty \in L^2$. The series $\sum_{n=0}^{\infty} \mathbb{E}((Y_n - Y_\infty)^2)$ also converges by the geometric bound in m^{-n} on $\mathbb{E}((Y_\infty - Y_n)^2)$ obtained by taking $k \to \infty$. By monotone convergence, $\mathbb{E}(\sum_{n=0}^{\infty}(Y_n - Y_\infty)^2) < \infty$, so $(Y_n)_{n \geqslant 0}$ converges almost surely to Y_∞. The first two moments of Y_∞ are easily obtained.

We have $L'_\infty(0) = -\mathbb{E}(Y_\infty) = -1$. Moreover, the Laplace transform $t \in \mathbb{R}_+ \mapsto L_n(t) = \mathbb{E}(e^{-tY_n})$ of Y_n satisfies, for every $n \geqslant 0$,

$$L_{n+1}(mt) = g_{n+1}(e^{-\frac{mt}{m^{n+1}}}) = g(g_n(e^{-\frac{t}{m^n}})) = g(L_n(t)),$$

hence $\lim_{n\to\infty} L_\infty(mt) = g(L_\infty(t))$, since $L_\infty = \lim_{n\to\infty} L_n$ and g is continuous. □

3.3 Tree size in the critical and subcritical regimes

Theorem 3.3 gives the mean total population at time n:

$$\mathbb{E}(Z_0 + \cdots + Z_n) = \frac{1 - m^{n+1}}{1 - m} \mathbb{1}_{m \neq 1} + (n+1)\mathbb{1}_{m=1}.$$

A similar formula can be obtained for the variance (exercise!). In this section, we consider the critical or subcritical regime $m \leqslant 1$. We know that the population becomes extinct almost surely, so the total population (or tree size) is

$$N = \lim_{n\to\infty} Z_0 + \cdots + Z_n \quad (= Z_0 + \cdots + Z_{T-1}),$$

hence, by monotone convergence,

$$\mathbb{E}(N) = \begin{cases} \dfrac{1}{1-m} & \text{si } m < 1, \\ \infty & \text{si } m = 1. \end{cases}$$

Since $Z_n \geqslant 1$ if $n < T$, we have $T \leqslant Z_0 + \cdots + Z_{T-1} = N$, giving the inequality $\mathbb{E}(T) \leqslant \mathbb{E}(N) = 1/(m-1)$. In the critical case, the total population N is almost surely finite but not integrable.

We characterize the distribution of N in terms of the offspring distribution P. First, the branching property applies: since the root individual gives rise to $X_{1,1}$ i.i.d. Galton–Watson trees, the distribution of N satisfies

$$N \overset{\text{loi}}{=} 1 + \sum_{i=1}^{X_{1,1}} N^{(i)}$$

where $(N^{(i)})_{i\geqslant 1}$ are i.i.d. random variables with the same distribution as N. This distributional identity recovers the formula $\mathbb{E}(N) = (1-m)^{-1}\mathbb{1}_{m<1} + \infty\mathbb{1}_{m=1}$ and, more generally, gives the following lemma.

Lemma 3.15 (Generating function of N)**.** *If g and f are the respective generating functions of $X_{1,1}$ and N, then the function f solves*

$$f(s) = s(g\circ f)(s) \quad \textit{pour } s\in[0,1].$$

Example 3.16 (Geometric offspring)**.** *If $P = \mathrm{Geo}_{\mathbb{N}}(p) = \sum_{n\geqslant 0} q^n p\delta_n$ with $p \geqslant 1/2$, then $f(s)$ solves $qf(s)^2 - f(s) + sp = 0$, and*

$$f(s) = \frac{1-\sqrt{1-4pqs}}{2q} \quad \textit{pour } s\in[0,1].$$

This generating function equation is not always easy to use. The distribution of N can also be expressed through convolutions of P.

Theorem 3.17 (Tree size distribution and random walks)**.** *We have*

$$N \overset{\text{loi}}{=} T_{-1},$$

where $T_{-1} := \inf\{n \geqslant 1 : S_n = -1\}$ is the hitting time of -1 for a random walk $(S_n)_{n\geqslant 0}$ on $\mathbb{Z}$ starting from $S_0 = 0$ with increments $(U_i)_{i\geqslant 1}$ such that $1 + U_i \sim P$. Moreover, for every $n \geqslant 1$,

$$\mathbb{P}(N = n) = \frac{\mathbf{P}(n, n-1)}{n} = \frac{P^{*n}(\{n-1\})}{n}.$$

The distribution of N can of course be computed directly without the distributional identity. This beautiful identity nevertheless provides, in particular, tail bounds useful for studying the Erdős–Rényi random graph; see Theorem 16.14.

Proof. The idea is to associate a finite tree bijectively with a random walk trajectory segment. Given a finite tree of size n, number its vertices starting from the root (numbered 1) and from left to right within each generation. Let $X^{(i)}$ be the number of children of individual i, and $U_i = X^{(i)} - 1$ for $1 \leqslant i \leqslant n$, and set

$$S_0 = 0 \quad \text{et} \quad S_{i+1} = S_i + U_{i+1} \quad \text{pour } 0 \leqslant i \leqslant n-1.$$

Figure 3.2 gives an example. With the numbering convention fixed, the tree is easily reconstructed from the trajectory of S. The sequence $(S_i)_{0\leqslant i\leqslant n}$ has the distribution of a random walk with increments $(U_i)_{i\geqslant 1}$ starting from 0 and stopped at the hitting time T_{-1} of -1. Since the increments have nonpositive mean $m-1$ and downward jumps have size 1, this time is almost surely finite. The bijection yields the distributional identity $N \overset{\text{loi}}{=} T_{-1}$. Finally, the distribution of T_{-1} follows from Lemma 3.18.

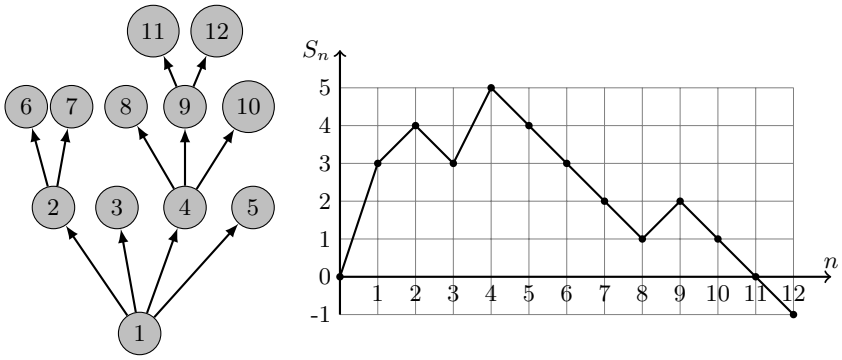


Fig. 3.2. A tree and its associated random walk.

Lemma 3.18 (The rotation principle)**.** *Let $(S_n)_{n\geqslant 0}$ be a random walk starting from 0 with i.i.d. integer-valued increments $(U_n)_{n\geqslant 1}$, and*

$$T_{-1} := \inf\{n \geqslant 1 : S_n = -1\}.$$

Then, for every $n \geqslant 1$,

$$\mathbb{P}(T_{-1} = n) = \frac{1}{n}\mathbb{P}(S_n = -1).$$

Proof of Lemma 3.18. Let $n \geqslant 1$ and $U_1, \ldots, U_n$ satisfy $S_n = -1$. For every integer k, let $\sigma(k)$ denote the integer between 1 and n congruent to k modulo n. For every $l = 1, \ldots, n$, define $S^{(l)}$ by

$$S_0^{(l)} = 0 \quad \text{et} \quad S_{k+1}^{(l)} = S_k^{(l)} + U_{\sigma(k+1+l)}.$$

Observe that, for every $l = 1, \ldots, n$, $S_n^{(l)} = S_n = -1$, and the trajectories $S^{(n)}$ and S coincide. Figure 3.3 gives an example with $n = 8$. The trajectory S attains its minimum at one or several times. Let l_0 be the earliest (in Figure 3.3, $l_0 = 4$). Then $S^{(l_0)}$ is the only trajectory among $S^{(1)}, \ldots, S^{(n)}$ that remains nonnegative until time $n - 1$. This proves the desired result. □

□

Example 3.19 (Poisson offspring)**.** *If $P = \mathrm{Poi}(\lambda)$ with $\lambda < 1$, then $U_1 \sim \mathrm{Poi}(\lambda) * \delta_{-1}$ and, for every $n \geqslant 1$,*

$$\mathbb{P}(N = n) = e^{-\lambda n}\frac{(\lambda n)^{n-1}}{n!}$$

since $S_n \sim \mathrm{Poi}(n\lambda) * \delta_{-n}$. Moreover, the generating series f of N solves the unwieldy functional equation $f(s) = se^{\lambda(f(s)-1)}$.

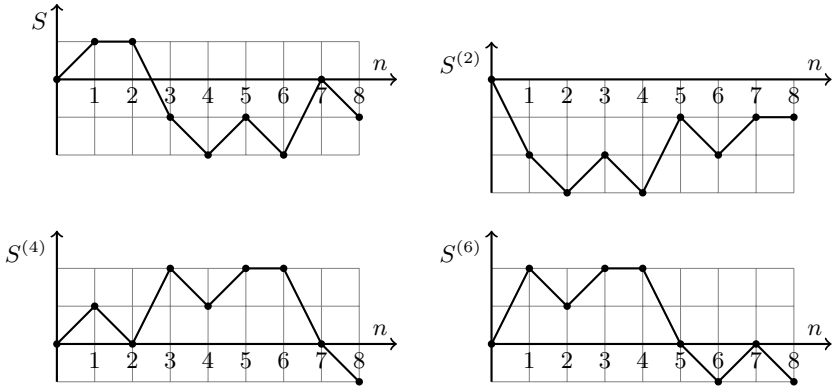


Fig. 3.3. One walk among n remains nonnegative before reaching -1 (here $S^{(4)}$).

3.4 Immigration

The Galton–Watson process with immigration $(Z_n)_{n\geqslant 0}$ starting from Z_0 is

$$Z_{n+1} := I_{n+1} + \sum_{k=1}^{Z_n} X_{n+1,k}$$

for every $n \geqslant 0$, where $(X_{n,k})_{n\geqslant 1,k\geqslant 1}$ are i.i.d. with distribution P on $\mathbb{N}$, and $(I_n)_{n\geqslant 1}$ are i.i.d. with distribution P_+ on $\mathbb{N}$; all these variables, together with Z_0, form an independent family. Suppose P and P_+ have respective means m and m_+. Write σ^2 and σ_+^2 for their variances when they exist. Assume $Z_0 = 1$. Let $\mathcal{F}_n$ be the sigma-algebra generated by $Z_0, (X_{i,j})_{1\leqslant i\leqslant n, j\geqslant 1}, I_1, \ldots, I_n$. For every $n \geqslant 0$,

$$\mathbb{E}(Z_{n+1} \,|\, \mathcal{F}_n) = mZ_n + m_+$$

from which a linear recurrence gives, for every $n \geqslant 1$,

$$\mathbb{E}(Z_n) = \left(m^n + \frac{m^n - 1}{m-1} m_+\right) \mathbb{1}_{m\neq 1} + (1 + nm_+)\mathbb{1}_{m=1}.$$

For the variance,

$$\mathbb{E}(Z_{n+1}^2 \,|\mathcal{F}_n) = Z_n(\sigma^2 + m^2) + Z_n(Z_n - 1)m^2 + 2Z_n m m_+ + \sigma_+^2 + m_+^2$$

which gives

$$\mathrm{Var}(Z_{n+1}) = \mathrm{Var}(Z_n)m^2 + \sigma^2 \mathbb{E}(Z_n) + \sigma_+^2.$$

If $m > 1$, the expectation of Z_n is of order m^n and its variance of order m^{2n}. If $m < 1$, then

$$\mathbb{E}(Z_n) \to \frac{m_+}{1-m} \quad \text{et} \quad \mathrm{Var}(Z_n) \to \frac{\sigma^2 m_+}{1-m} + \frac{\sigma_+^2}{1-m^2}.$$

Note that $Z_n = Z_{n,0} + \sum_{k=1}^{n} Z_{n,k}$, where $Z_{n,0}$ is the number of descendants at time n of the individual present at time 0 (a process without immigration!), while $Z_{n,k}$ with $1 \leqslant k \leqslant n$ is the number of descendants at time n of individuals who immigrated at time k. All these variables are independent. Observe that $Z_{n,k}$ has the distribution of a sum of I_k independent copies of $Z_{n-k,0}$. Although these observations can be used to study the case $m \leqslant 1$, for simplicity we restrict attention to $m > 1$.

Theorem 3.20 (The supercritical case with immigration)**.** *If $m > 1$, then $(Y_n)_{n\geqslant 0} = (Z_n/m^n)_{n\geqslant 0}$ converges almost surely to a real random variable $Y_\infty \geqslant 0$. If also $\sigma < \infty$ and $\sigma_+ < \infty$, then convergence holds in L^2 as well, and*[1]

$$\mathbb{E}(Y_\infty) = 1 + \frac{m_+}{m-1}.$$

Proof. The sequence $(Y_n)_{n\geqslant 0}$ is a submartingale for $(\mathcal{F}_n)_{n\geqslant 0}$, since

$$\mathbb{E}(Y_n \,|\, \mathcal{F}_{n-1}) = Y_{n-1} + m^{-n} m_+ \geqslant Y_{n-1}.$$

We have $\mathbb{E}(Y_0) = 1$ and $\mathbb{E}(Y_n) = \mathbb{E}(Y_{n-1}) + m^{-n} m_+$ for $n \geqslant 1$, hence

$$\mathbb{E}(Y_n) = 1 + (m^{-1} + \cdots + m^{-n}) m_+.$$

Since $m > 1$ and $Y_n \geqslant 0$, we obtain $\sup_n \mathbb{E}(|Y_n|) = \sup_n \mathbb{E}(Y_n) < \infty$. As a submartingale bounded in L^1, the sequence $(Y_n)_{n\geqslant 0}$ converges almost surely to an integrable real random variable $Y_\infty \geqslant 0$, and convergence holds in L^1 if the sequence is uniformly integrable. Assume $\sigma < \infty$ and $\sigma_+ < \infty$, and let us show that convergence holds in L^2. We have

$$\mathbb{E}(Y_n^2) = m^{-2n}\mathbb{E}(Z_n^2) = m^{-2n}\mathrm{Var}(Z_n) + m^{-2n}\mathbb{E}(Z_n)^2.$$

If $m > 1$, we know that the expectation of Z_n is of order m^n and its variance of order m^{2n}, hence $\ell := \lim_{n\to\infty} \mathbb{E}(Y_n^2) < \infty$. Finally, for every $n, k \geqslant 1$,

$$\begin{aligned}
\mathbb{E}((Y_{n+k} - Y_n)^2) &= \mathbb{E}(Y_{n+k}^2) + \mathbb{E}(Y_n^2) - 2\mathbb{E}(Y_{n+k} Y_n) \\
&= \mathbb{E}(Y_{n+k}^2) + \mathbb{E}(Y_n^2) - 2\mathbb{E}(\mathbb{E}(Y_{n+k} \,|\, \mathcal{F}_n) Y_n) \\
&= \mathbb{E}(Y_{n+k}^2) - \mathbb{E}(Y_n^2) - 2(m^{-n-k} + \cdots + m^{-n-1}) m_+ \mathbb{E}(Y_n) \\
&= \ell + o_{n\to\infty}(1) - \ell - o_{n\to\infty}(1) + o_{n\to\infty}(1).
\end{aligned}$$

Consequently, $(Y_n)_{n\geqslant 0}$ is Cauchy in L^2 and thus converges in L^2. Alternatively, one can directly invoke the convergence theorem for martingales bounded in L^2. □

[1] One can also compute $\mathrm{Var}(Y_\infty)$ (exercise!).

3.5 Further reading

Around 1875, Francis Galton and Henri William Watson wrote their article on the evolution of the number of English aristocratic family names. They were probably unaware of Irénée-Jules Bienaymé's earlier work on the same topic. The analysis of the Galton–Watson process can be greatly refined. The model itself can be enriched and modified to incorporate different situations: distinct sexes, individuals surviving across several generations, and so on. The study of branching processes through generating functions apparently goes back to Stanisław Ulam at Los Alamos during the Manhattan Project, where branching processes modeled the number of neutrons in a nuclear chain reaction. Many mathematical developments appear in the books by Theodore Harris [Har02], Khrishna Athreya and Peter Ney [AN04], Jean-François Delmas and Benjamin Jourdain [DJ06], and Patasy Haccou, Peter Jagers, and Vladimir Vatoutin [HJV07]. Branching processes have been extensively studied by the French school of probability, notably Jacques Neveu, Jean-François Le Gall, Jean Bertoin, and their mathematical descendants.

In Theorem 3.13 (explosion in the supercritical case), if $\mathbb{E}(Y_\infty) \neq 1$ then $\mathbb{E}(Y_n) < 1$ by Fatou's lemma. Moreover, Harry Kesten and Bernt P. Stigum proved in [KS66] that either $\mathbb{P}(Y_\infty = 0) = \mathbb{P}(T < \infty)$ or $\mathbb{P}(Y_\infty = 0) = 1$, and the following properties are equivalent:

1. $\mathbb{P}(Y_\infty = 0) = \mathbb{P}(T < \infty)$;
2. $\mathbb{E}(Y_\infty) = 1$;
3. $(Y_n)_{n \geqslant 1}$ converges in L^1;
4. $\mathbb{E}(X_{1,1} \log(X_{1,1})) < \infty$, that is, $z \in \mathbb{N} \mapsto z \log(z) \in L^1(P)$.

The process with immigration is studied in the collective volume [Rug01] and in Athreya and Ney's book [AN04]. Akiva Yaglom's Theorem 3.11 is a good starting point for the more general study of quasi-stationary distributions of population processes, presented in the survey by Sylvie Méléard and Denis Villemonais [MV12].

Lemma 3.18 appears in an article by Aryeh Dvoretzky and Theodore Motzkin [DM47]. The expression for the size of a Galton–Watson tree was established by Meyer Dwass [Dwa69] before the connection between the two situations was made. See Jean-François Marckert's habilitation thesis [Mar04] and the survey on the ballot theorem by Luigi Addario-Berry and Bruce Reed [ABR08]. Beyond Theorem 3.17, combinatorics abounds in bijections between collections of objects of different kinds.

4

Permutations, partitions, and graphs

Keywords. Uniform distribution; simulation; permutation; partition; graph; tree; Catalan numbers; Bell numbers. **Tools.** Symmetric group; random walk; discrete distribution. **Difficulty.** *

This chapter presents algorithms for sampling uniformly from several remarkable sets of permutations, partitions, and graphs. The concepts and techniques extend beyond the settings considered here, which have the advantage of being elementary.

4.1 A basic algorithm for discrete distributions

Consider the problem of sampling from a discrete distribution $\mu = \sum_{a\in E} \mu(a)\delta_a$, where E is finite or countable. We can enumerate the elements of E using a bijection $\varphi : E \to \{1, 2, \ldots\}$. Let $a_k := \varphi^{-1}(k)$. Partition the real interval $[0, 1]$ into blocks with respective Lebesgue measures $\mu(a_1), \mu(a_2)$, and so on, for example using the intervals

$$I_1 = [0, \mu(a_1)[\,,\ I_2 = [\mu(a_1), \mu(a_1) + \mu(a_2)[\,,\ \ldots,$$

If U is uniformly distributed on the interval $[0, 1]$, then $\mathbb{P}(U \in I_{\varphi(a)}) = \mu(a)$ for every $a \in E$. The basic algorithm for sampling μ is therefore as follows: generate a realization u of U; if $u \leqslant \mu(a_1)$, choose a_1; otherwise, if $u \leqslant \mu(a_1) + \mu(a_2)$, choose a_2, and so on. If F is the distribution function of the distribution $\mu \circ \varphi^{-1}$ on $\{1, 2, \ldots\} \subset \mathbb{R}$, with generalized inverse F^{-1}, then $(\varphi^{-1} \circ F^{-1})(U) \sim \mu$. This is a special case of inverse transform sampling. The cost of this algorithm is the number N of tests performed. This number is random, with distribution $\mu \circ \varphi^{-1}$. In particular, $\mathbb{P}(N < \infty) = 1$, and the average cost is

$$\mathbb{E}(N) = \sum_{a \in E} \varphi(a)\mu(a),$$

which may well be infinite if $\mu \circ \varphi^{-1}$ has no expectation (this can only happen if E is infinite)! If the enumeration φ minimizes the average cost $\mathbb{E}(N)$, then

$$\mu(a_1) \geqslant \mu(a_2) \geqslant \cdots .$$

For the geometric distribution (with any mean) and the Poisson distribution (with mean $\leqslant 1$), the natural enumeration has decreasing weights. Moreover, if $\mathrm{card}(E)$ is small, a sorting algorithm can determine the order of decreasing weights (at a cost).

Standard discrete distributions (binomial, geometric, Poisson, etc.) can be sampled using dedicated algorithms that exploit their special properties. In particular, one can sample a Poisson distribution with any mean λ using a generator for the Poisson distribution with mean 1. This simply requires thinning[1]. More precisely, generate $\lceil \lambda \rceil$ i.i.d. random variables $X_1, \ldots, X_{\lceil \lambda \rceil}$ with distribution $\mathrm{Poi}(1)$; then, conditionally on their sum $S = X_1 + \cdots + X_{\lceil \lambda \rceil}$, generate S i.i.d. Bernoulli random variables with mean $\lambda / \lceil \lambda \rceil$, and use the fact that $B_1 + \cdots + B_S \sim \mathrm{Poi}(\lambda)$ (a Poisson mixture of binomial distributions[2]).

Sampling uniformly from a finite set E can be very simple: $\varphi^{-1}(\lceil \mathrm{card}(E)U \rceil)$ has this distribution! However, this basic algorithm is impractical when E is difficult to enumerate and hence φ is difficult to access, or when $\mathrm{card}(E)$ is very large. We next study examples involving permutations, partitions, and graphs, for which we present efficient, exact algorithms.

4.2 Random permutations

Some situations require randomly permuting a finite list of objects: constructing experimental designs, anonymization, and so forth. This leads to the problem of sampling the uniform distribution $\sum_{\sigma \in \mathcal{S}_n} \mathrm{card}(\mathcal{S}_n)^{-1} \delta_\sigma$ on the finite set $\mathcal{S}_n$ of permutations of $\{1, \ldots, n\}$ (the symmetric group). But $\mathrm{card}(\mathcal{S}_n) = n! \sim \sqrt{2\pi n}(n/e)^n$ is a number with approximately $n \log(n)$ digits, and this combinatorial growth quickly rules out the basic algorithm for discrete distributions. One can sample uniformly from $\mathcal{S}_n$ by ordering a symmetric random sample: if $U_1, \ldots, U_n$ are i.i.d. random variables uniformly distributed on $[0, 1]$ and σ is a random permutation such that $U_{\sigma(1)} \leqslant \cdots \leqslant U_{\sigma(n)}$, then σ is uniformly distributed on $\mathcal{S}_n$. The complexity is that of the sorting algorithm used[3].

A naive algorithm for sampling a uniform permutation σ draws the values of $\sigma(1), \ldots, \sigma(n)$ uniformly without replacement from $\{1, \ldots, n\}$. This

[1] Randomly retaining some of the generated points.
[2] If $X \sim \mathrm{Poi}(\lambda)$ and $\mathrm{Loi}(Y|X = n) = \mathrm{Bin}(n, p)$ for every n, then $Y \sim \mathrm{Poi}(p\lambda)$.
[3] Of order $n \log(n)$ with high probability and n^2 in the worst case for quicksort.

seemingly attractive algorithm has higher complexity than sorting because its implementation must keep track of previously drawn elements. A good algorithm for sampling (exactly!) uniformly from $\mathcal{S}_n$ is known as the Fisher–Yates–Knuth algorithm[4]. It requires $n-1$ independent random variables uniformly distributed on $\{1,2\},\dots,\{1,2,\dots,n\}$, respectively, or of order $\sum_{k=2}^{n}\log(k)=\log(n!)\approx n\log(n)$ i.i.d. bits.

Theorem 4.1 (Permutations and the Fisher–Yates–Knuth algorithm)**.** *If $U_1,\dots,U_n$ are independent random variables, with U_i uniformly distributed on $\{1,\dots,i\}$ for every $1\leqslant i\leqslant n$, then the permutation formed by the product of random transpositions $(1,U_1)\cdots(n,U_n)$ is uniformly distributed on $\mathcal{S}_n$.*

Since inversion in $\mathcal{S}_n$ is bijective and transpositions are their own inverses, the reversed product $(n,U_n)\cdots(1,U_1)$ is also uniformly distributed on $\mathcal{S}_n$. The transposition $(1,U_1)$ is trivial (the identity of $\mathcal{S}_n$) and need not be included if $n\geqslant 2$ (although it makes the formula valid for $n=1$). The Fisher–Yates–Knuth algorithm for permuting a vector `v` has the following pseudocode:

```
for k from length(v) downto 2 do
swap(v[ceil(k*rand)],v[k])
```

Proof of Theorem 4.1. We proceed by induction on n. The assertion is trivial for $n=1$. Assume it holds for $n\geqslant 1$. For every $\sigma\in\mathcal{S}_n$, use the same notation σ for the element of $\mathcal{S}_{n+1}$ obtained from σ by adding the cycle $(n+1)$ (a fixed point). Suppose $\sigma_n=(1,U_1)\cdots(n,U_n)$ is uniformly distributed on $\mathcal{S}_n$. Let U_{n+1} be independent of σ_n and uniformly distributed on $\{1,\dots,n+1\}$. We show that $\sigma_{n+1}=\sigma_n(n+1,U_{n+1})$ is uniformly distributed on $\mathcal{S}_{n+1}$. For every $\sigma\in\mathcal{S}_{n+1}$,

$$\begin{aligned}\mathbb{P}(\sigma_{n+1}=\sigma)&=\sum_{i=1}^{n+1}\mathbb{P}(\sigma_n=\sigma(n+1,i))\mathbb{P}(U_{n+1}=i)\\&=\frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{P}(\sigma_n=\sigma(n+1,i)).\end{aligned}$$

Since $n+1$ is a fixed point of σ_n and is a fixed point of $\sigma(n+1,i)$ for exactly one value of i, denoted by i_σ, the preimage of $n+1$ under σ, we finally obtain

$$\mathbb{P}(\sigma_{n+1}=\sigma)=\frac{1}{n+1}\mathbb{P}(\sigma_n=\sigma(n+1,i_\sigma))=\frac{1}{n+1}\frac{1}{n!}=\frac{1}{(n+1)!}.$$

□

A permutation $\sigma\in\mathcal{S}_n$ is called a derangement if it has no fixed points: $\sigma(i)\neq i$ for every $1\leqslant i\leqslant n$. The fixed points of σ are the cycles of length 1 in its disjoint cycle decomposition. The uniform distribution on the set $\mathcal{D}_n\subset\mathcal{S}_n$ of derangements can be sampled by rejection, because the cardinality ratio $\mathrm{card}(\mathcal{D}_n)/\mathrm{card}(\mathcal{S}_n)$ is large.

[4] Also called the *Fisher–Yates shuffle* or *Knuth shuffle*.

Theorem 4.2 (Uniform distribution on derangements)**.** *If σ is a random permutation uniformly distributed on $\mathcal{S}_n$, then*

$$p_n := \mathbb{P}(\sigma \in \mathcal{D}_n) = \frac{\mathrm{card}(\mathcal{D}_n)}{\mathrm{card}(\mathcal{S}_n)} \underset{n\to\infty}{\longrightarrow} \frac{1}{e} \approx 0.37.$$

If $(\sigma_k)_{k\geqslant 1}$ is a sequence of independent identically distributed random permutations, each uniform on $\mathcal{S}_n$, and if

$$T := \inf\{k \geqslant 1 : \sigma_k \in \mathcal{D}_n\}$$

then the random permutation σ_T is uniformly distributed on $\mathcal{D}_n$, and the variable T has the geometric distribution with parameter p_n and expectation $1/p_n \underset{n\to\infty}{\longrightarrow} e \approx 2.72$.

One sometimes writes $!n = \mathrm{card}(\mathcal{D}_n)$, and we have $!(n+1) = (n+1)\times !n + (-1)^{n+1}$, analogous to $(n+1)! = (n+1)\times n!$.

Performance can be improved by stopping the Fisher–Yates–Knuth algorithm at each proposal step as soon as a fixed point appears.

Proof. If σ is uniformly distributed on $\mathcal{S}_n$, then $\{\sigma \notin \mathcal{D}_n\} = \cup_{i=1}^n A_i$, where $A_i = \{\sigma(i) = i\}$. Thus, by inclusion-exclusion,

$$\mathbb{P}(\sigma \in \mathcal{D}_n) = 1 - \mathbb{P}(\cup_{1\leqslant i\leqslant n} A_i) = 1 - \sum_{p=1}^{n} (-1)^{p+1} \sum_{1\leq i_1<\cdots<i_p\leq n} \mathbb{P}(A_{i_1}\cap\cdots\cap A_{i_p}).$$

For every $1 \leqslant p \leqslant n$,

$$\sum_{1\leqslant i_1<\cdots i_p\leqslant n} \mathbb{P}(A_{i_1}\cap\cdots\cap A_{i_p}) = \sum_{1\leqslant i_1<\cdots<i_p\leqslant n} \frac{(n-p)!}{n!} = \binom{n}{p}\frac{(n-p)!}{n!} = \frac{1}{p!}$$

hence $\mathbb{P}(\sigma \in \mathcal{D}_n) = 1 - \sum_{p=1}^n \frac{(-1)^{p+1}}{p!} \to e^{-1}$. □

4.3 Random partitions

A matching[5] of $2n$ points is a partition of $\{1,\ldots,2n\}$ into n subsets of cardinality 2, each forming a pair of matched points. Write $\mathcal{A}_{2n}$ for the set of matchings of $2n$ points. The set $\mathcal{A}_{2n}$ is in bijection with the set of involutive derangements of $\{1,\ldots,2n\}$, that is, elements of $\mathcal{S}_{2n}$ without fixed points that are their own inverses, or equivalently elements of $\mathcal{S}_{2n}$ obtained as products of n transpositions with pairwise disjoint supports. Isserlis' formula gives

$$\mathrm{card}(\mathcal{A}_{2n}) = \frac{(2n)!}{2^n n!}.$$

[5] A partition into pairs.

This is also the product of the odd numbers less than or equal to $2n-1$, denoted by $(2n-1)!!$ (double factorial). The ratio $\mathrm{card}(\mathcal{A}_{2n})/\mathrm{card}(\mathcal{S}_{2n})$ is small, which suggests seeking an alternative to rejection sampling. Here is an algorithm for sampling uniformly from $\mathcal{A}_{2n}$.

Theorem 4.3 (Uniform distribution on matchings)**.** *If σ is uniformly distributed on $\mathcal{S}_{2n}$, then the random matching $\{\{\sigma(1),\sigma(n+1)\},\dots,\{\sigma(n),\sigma(n+n)\}\}$ is uniformly distributed on $\mathcal{A}_{2n}$.*

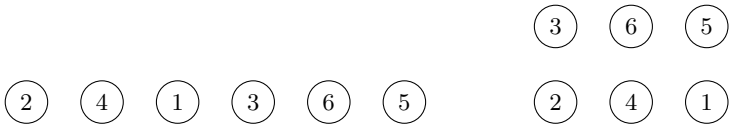


Fig. 4.1. Matching of $\{1,\dots,6\}$ obtained from an element of $\mathcal{S}_6$ (here $n=3$).

Proof. For every matching $\{\{a_1,a_{n+1}\},\dots,\{a_n,a_{n+n}\}\}$, there are $n!$ ways to permute the n blocks and 2^n ways to permute their contents, hence

$$\begin{aligned}\mathbb{P}(\{\{\sigma(1),\sigma(n+1)\},\dots,\{\sigma(n),\sigma(n+n)\}\} = \{\{a_1,a_{n+1}\},\dots,\{a_n,a_{n+n}\}\})\\ = 2^n n!\mathbb{P}(\sigma(1)=a_1,\dots,\sigma(2n)=a_{2n}) = \frac{2^n n!}{(2n)!}\end{aligned}$$

which does not depend on the matching and is precisely $1/\mathrm{card}(\mathcal{A}_{2n})$. □

The cycle decomposition of a uniformly random permutation of $\mathcal{S}_n$ provides a random partition of $\{1,\dots,n\}$. Its distribution is not uniform on the set of partitions Π_n of $\{1,\dots,n\}$ (Remark 14.6). Let us consider sampling uniformly from Π_n. This distribution assigns the same weight $1/B_n$ to every element of Π_n, where $B_n = \mathrm{card}(\Pi_n)$. In combinatorics, the sequence $(B_n)_{n\geqslant 1}$ consists of the Bell numbers. We have $B_1=1$, $B_2=2$, and more generally, using the convention $B_0=1$, the triangular recurrence

$$B_{n+1} = \sum_{k=0}^{n}\binom{n}{k}B_k,$$

where k is the number of elements outside the block containing $n+1$. It follows that the formal power series $G(X)=\sum_{n=0}^{\infty}\frac{B_n}{n!}X^n$ satisfies $G'(X)=\exp(X)G(X)$, yielding $G(X)=\exp(\exp(X)-1)$. We recognize the Laplace transform of the Poi(1) distribution. The Bell numbers are therefore the moments of this distribution, giving Dobinski's formula:

$$B_n = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}.$$

This formula appears in an algorithm for sampling uniformly from Π_n.

Theorem 4.4 (Stam's algorithm)**.** *Let $n \geqslant 1$. Let K be a random integer taking the value k with probability $k^n/(k!eB_n)$ for every $k \geqslant 0$. Given K, let $C_1, \dots, C_n$ be i.i.d. random variables uniformly distributed on $\{1, \dots, K\}$. Let P be the random partition of $\{1, \dots, n\}$ obtained by placing i, j in the same block if and only if $C_i = C_j$. Then P is uniformly distributed on Π_n.*

Dobinski's formula ensures that the distribution of K is well defined. It is convenient to interpret $C_1, \dots, C_n$ as colors: the blocks of P then group elements by color. The random integer K can be sampled using the basic algorithm for discrete distributions.

Proof. If $p \in \Pi_n$ has b blocks, then

$$\begin{aligned}\mathbb{P}(P = p) &= \sum_{k=b}^{\infty} \mathbb{P}(P = p \,|\, K = k)\mathbb{P}(K = k) \\ &= \sum_{k=b}^{\infty} \frac{k(k-1)\cdots(k-b+1)}{k^n} \frac{k^n}{k!eB_n} = \frac{1}{B_n}.\end{aligned}$$

□

4.4 Random graphs

Throughout this section, set $V = \{1, \dots, n\}$. A finite graph

$$G = (V, E)$$

is a pair with $E \subset \{\{i, j\} : i, j \in V, i \neq j\}$. The elements of V are the *vertices*[6] of the graph, while the elements of E are its *edges*[7]. There is at most one edge between two distinct vertices (no multiple edges), and none from a vertex to itself (no loops). Edges are undirected. We use the terminology for finite graphs introduced in Chapter 16: *paths*, *loops*, and so forth. The adjacency matrix A of G is the symmetric matrix $n \times n$ defined by $A_{j,k} = \mathbb{1}_{\{j,k\} \in E}$.

Theorem 4.5 (Uniform distribution on finite graphs)**.** *For every $n \geqslant 1$, the uniform distribution on the set $\mathcal{G}_n$ of finite graphs with vertices $\{1, \dots, n\}$ is obtained by making the $\binom{n}{2} = \frac{1}{2}n(n-1)$ edges independent and identically distributed with Bernoulli distribution* $\mathrm{Ber}(1/2)$.

A uniformly random graph in $\mathcal{G}_n$ is called an Erdős–Rényi random graph of size n and parameter $p = \frac{1}{2}$.

Proof. The set $\mathcal{G}_n$ is in bijection with the set of symmetric $n \times n$ matrices with entries in $\{0, 1\}$ and zero diagonal, itself in bijection with the product

[6] Also called *sites*, hence the notation V.
[7] Also called *links*; hence the notation E.

set $\{0,1\}^{n(n-1)/2}$. The uniform distribution on a product set is the product of the uniform distributions on its factors, and the uniform distribution on $\{0,1\}$ is the Bernoulli distribution $\mathrm{Ber}(1/2)$. □

In a finite graph $G = (V, E)$, the *degree* of a vertex $i \in V$ is the number, denoted by d_i, of vertices connected directly to i by an edge:

$$d_i := \mathrm{card}\{j \in V : \{i, j\} \in E\}.$$

We call G a d-regular graph, where $d \geqslant 0$ is a fixed integer, if $d_i = d$ for every $i \in V$. The 0-regular graphs consist of isolated vertices and have no edges. The 1-regular graphs consist of mutually disconnected edges. Finite 2-regular graphs consist of mutually disconnected cycles.

If G is a graph with vertices $\{1, \ldots, n\}$ such that $d_1 \geqslant \cdots \geqslant d_n$ then the following two properties hold:

1. the integer $d_1 + \cdots + d_n$ is even;
2. for every $1 \leqslant k \leqslant n$,

$$d_1 + \cdots + d_k \leqslant k(k-1) + \min(k, d_{k+1}) + \cdots + \min(k, d_n).$$

The sum is even because each edge is counted twice. The quantity $k(k-1) + \min(d_{k+1}, k) + \cdots + \min(d_n, k)$ is the maximal contribution to $d_1 + \cdots + d_k$ from edges incident to vertices 1 through k: there are at most $\binom{k}{2} = k(k-1)/2$ edges (counted twice) between vertices 1 through k, and at most $\min(k, d_{k+i})$ edges between vertex $i > k$ and vertices 1 through k.

A theorem of Erdős–Gallai states that, for every $d_1 \geqslant \cdots \geqslant d_n \geqslant 0$, a graph with $n \geqslant 1$ vertices of degrees $d_1, \ldots, d_n$ exists if and only if the two conditions above hold. We call $d_1, \ldots, d_n$ the graph's degree sequence[8]. For a d-regular graph with vertices $\{1, \ldots, n\}$, nd is even and $d \leqslant n-1$ (with equality for the complete graph).

Multigraphs are obtained by relaxing two constraints in the definition of graphs: multiple edges between vertices and loops are allowed. Let $d_1 \geqslant \cdots \geqslant d_n \geqslant 0$ be integers satisfying the Erdős–Gallai conditions, and let $\mathcal{M}_{d_1,\ldots,d_n}$ be the set of multigraphs with vertices $\{1, \ldots, n\}$ and prescribed degrees $d_1, \ldots, d_n$. Let $\sigma \in \mathcal{A}_{2n}$ be a matching of $2n$ points. Construct an element $M_\sigma \in \mathcal{M}_{d_1,\ldots,d_n}$ from σ as follows:

- for every $1 \leqslant k \leqslant n$, place d_k half-edges at vertex k, giving a total of $2(d_1 + \cdots + d_n)$ numbered half-edges;
- pair the $2n$ half-edges using the matching σ.

Let $\mathcal{G}_{d_1,\ldots,d_n} \subset \mathcal{M}_{d_1,\ldots,d_n}$ be the set of graphs in $\mathcal{M}_{d_1,\ldots,d_n}$. The configuration algorithm samples uniformly from $\mathcal{G}_{d_1,\ldots,d_n}$ using the multigraph M_σ associated with a random matching σ and rejection sampling.

[8] The sequence of vertex degrees.

Theorem 4.6 (Bollobás' configuration algorithm)**.** *If $(\sigma_k)_{k\geqslant 1}$ is a sequence of random matchings uniformly distributed on $\mathcal{A}_{2n}$, and T is the smallest integer $k \geqslant 1$ such that $M_{\sigma_k} \in \mathcal{G}_{d_1,\dots,d_n}$, meaning that the multigraph M_{σ_k} has neither multiple edges nor loops, then T is almost surely finite and geometrically distributed, and M_{σ_T} is uniformly distributed on $\mathcal{G}_{d_1,\dots,d_n}$.*

The random multigraph M_{σ_1} is called the *configuration model* with prescribed degrees. Uniform random *matchings* can be generated using Theorem 4.3.

Proof sketch. First, $\mathbb{P}(M_{\sigma_1} \in \mathcal{G}_{d_1,\dots,d_n}) > 0$ because the support of the distribution of M_{σ_1} is $\mathcal{M}_{d_1,\dots,d_n}$. The rejection algorithm stops at an almost surely finite geometric time T.

One can show that $M \mapsto \mathbb{P}(M_{\sigma_1} = M)$ is constant on $\mathcal{G}_{d_1,\dots,d_n}$. Thus the conditional distribution of M_{σ_1} given $\{M_{\sigma_1} \in \mathcal{G}_{d_1,\dots,d_n}\}$ is uniform on $\mathcal{G}_{d_1,\dots,d_n}$. This conditional distribution is precisely the distribution of M_{σ_T} (rejection sampling). □

Note that M_{σ_1} is not uniformly distributed on $\mathcal{M}_{d_1,\dots,d_n}$, because the map $M \mapsto \mathbb{P}(M_{\sigma_1} = M)$ is not constant on $\mathcal{M}_{d_1,\dots,d_n}$: two multigraphs differing only in the number of multiple edges between two particular vertices do not have the same probability, since those edges are indistinguishable and therefore interchangeable (the same applies to loops).

4.5 Random trees

A very common tree structure is that of a *binary tree*: each vertex has 0 or 2 children, that is, 1 or 3 neighbors when the tree is viewed as a graph. We consider *rooted trees*: the number of vertices is therefore odd. We consider *plane trees*: vertices in the same generation are numbered from left to right, starting from the root, numbered 1 and drawn at the bottom, as in Figure 4.2. Write $\mathcal{T}_n$ for the set of numbered trees of this type with $2n+1$ vertices.

Each element of $\mathcal{T}_n$ can be encoded by a simple random walk trajectory. More precisely, given an element of $\mathcal{T}_n$, let $x^{(i)}$ be the number of children of vertex i, and $u_i = x^{(i)} - 1$ for $1 \leqslant i \leqslant 2n+1$. Set $s_0 = 0$ and $s_{i+1} = s_i + u_{i+1}$ for every $0 \leqslant i \leqslant 2n$. Figure 4.2 gives an example of this correspondence. Induction shows that $s_{2n+1} = -1$ and $\min(s_1, \dots, s_{2n}) \geqslant 0$ always hold. Thanks to the numbering convention, the tree is easily reconstructed from the trajectory segment $(s_i)_{1\leqslant i\leqslant 2n}$. One can check that this construction defines a bijection between $\mathcal{T}_n$ and the set $\mathcal{P}_n$ of simple random walk trajectories of length $2n$ on $\mathbb{Z}$ that equal 0 at times 0 and $2n$ and remain nonnegative between these times. This is a special case of the bijection in the proof of Theorem 3.17. Since the bijection is explicit and computable, sampling uniformly from $\mathcal{T}_n$ reduces to sampling uniformly from $\mathcal{P}_n$, which can be done as follows.

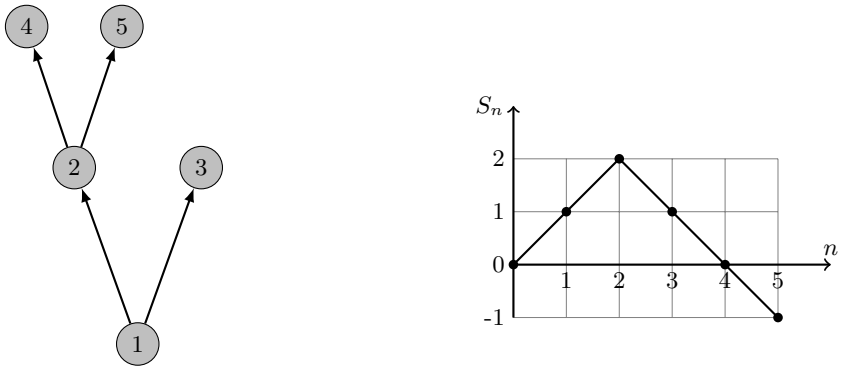


Fig. 4.2. An element of $\mathcal{T}_2$ and its associated random walk trajectory.

Theorem 4.7 (The Arnold–Sleep algorithm)**.** *Let $s = (s_i)_{0 \leqslant i \leqslant 2n}$ be the random path taking values in $\mathcal{P}_n$ whose distribution is given by $s_0 = s_{2n} = 0$ and, for every $0 \leqslant k \leqslant 2n-1$,*

$$\begin{aligned}\mathbb{P}(s_{k+1} - s_k = -1 \,|\, s_0, \dots, s_k) &= 1 - \mathbb{P}(s_{k+1} - s_k = 1 \,|\, s_0, \dots, s_k) \\ &= \frac{s_k(2n + k + s_k + 2)}{2(2n-k)(s_k+1)}.\end{aligned}$$

Then s is uniformly distributed on $\mathcal{P}_n$, and hence the associated binary tree is uniformly distributed on $\mathcal{T}_n$.

Proof. Construct a path $(s_i)_{1 \leqslant i \leqslant 2n}$ uniformly distributed on $\mathcal{P}_n$ step by step. Suppose $s_0, \dots, s_{2n-k}$ have already been constructed. The number of ways to extend this initial trajectory into an element of $\mathcal{P}_n$ depends only on k and $r = s_{2n-k}$; denote it by $N(r,k)$. The number of extensions beginning with an increment $+1$ is $N(r+1, k-1)$, while the number beginning with an increment -1 is $N(r-1, k-1)$. Thus $N(r,k) = N(r+1, k-1) + N(r-1, k-1)$, and

$$\mathbb{P}(s_{2n-k+1} = s_{2n-k} - 1 \,|\, s_0, \dots, s_{2n-k}) = \frac{N(r-1, k-1)}{N(r,k)} \quad \text{où} \quad r = s_{2n-k}.$$

The number $N(r,k)$ also counts nonnegative random walk paths connecting $(0,r)$ and $(k,0)$. Necessarily, $k \geqslant r$. Each such path has $t \geqslant r$ increments -1 and $k - t = t - r$ increments $+1$, hence $2t = k + r$, so $k + r$ is even. Using the notation and results from the proof of the ballot theorem 2.6 gives

$$N(r,k) = P^+_{k+1, r+1} = \frac{r+1}{k+1}\binom{k+1}{t+1} = \frac{r+1}{t+1}\binom{k}{t}.$$

It follows that

$$\mathbb{P}(s_{2n-k+1} = s_{2n-k} - 1 \,|\, s_0, \ldots, s_{2n-k}) = \frac{r(k+r+2)}{2k(r+1)} \quad \text{où} \quad r = s_{2n-k}.$$

Finally, the change of variable $2n - k \to k$ gives the stated result. Note that if $r = k$ then $\mathbb{P}(s_{2n-k+1} = s_{2n-k} - 1 \,|\, s_0, \ldots, s_{2n-k}) = 1$. □

By Theorem 2.5, the cardinality of $\mathcal{P}_n$ and $\mathcal{T}_n$ is the *Catalan number* C_{n-1}. The sets $\mathcal{T}_n$ and $\mathcal{P}_n$ are also in bijection with the set of balanced parenthesis expressions consisting of n pairs of parentheses. Simply encode the sequence of random walk increments using parentheses. For example, the increment sequence $+1, +1, -1, -1$ corresponding to the trajectory $(s_i)_{0 \leqslant i \leqslant 4}$ in Figure 4.2 becomes the parenthesis expression $(())$. This interpretation appears in the proof of Wigner's theorem 21.9.

4.6 Further reading

Sometimes the structure of the set of interest provides no attractive algorithm for uniform sampling, and rejection from a larger set is impractical. In that case, one can often use Markovian algorithms such as Metropolis–Hastings or Propp–Wilson, discussed in Chapter 5.

A probabilistic analysis of the complexity of randomized quicksort can be found, for example, in the book by Rajeev Motwani and Prabhakar Raghavan [MR95]. Donald Knuth's monumental work [Knu05] is an essential reference for algorithm analysis and uniform sampling from classical sets such as permutations and partitions. In the first edition, dating from the 1960s, Knuth popularized the algorithm for uniformly sampling permutations, following an earlier article by Richard Durstenfeld [Dur64]. The algorithm actually goes back to Ronald Fisher and Frank Yates [FY48]. It is implemented as standard in mathematical software. This algorithm corresponds exactly to the Chinese restaurant process of Chapter 14, and the uniform distribution on $\mathcal{S}_n$ coincides with the Ewens distribution with parameter $\theta = 1$.

For every fixed $n \geqslant 1$, the random walk on $\mathcal{S}_n$ with i.i.d. steps uniformly distributed over the transpositions, studied by Persi Diaconis and Mehrdad Shahshahani [DS81], converges abruptly to the uniform distribution on $\mathcal{S}_n$ after approximately $n \log(n)$ steps, as does the card-shuffling walk of Chapter 2. If $\sigma \in \mathcal{S}_n$ and $\tau = (i, j)$ is a transposition, then the cycle decomposition of $\sigma\tau$ is obtained from that of σ by merging the cycles of σ containing i and j if they differ, or by splitting the cycle of σ containing i and j otherwise. The Diaconis–Shahshahani chain, viewed on Π_n through the partition of cycle supports, is a fragmentation-coalescence chain. Its transition kernel can be described as follows: given that the chain is at $P \in \Pi_n$, draw i and j uniformly with replacement from $\{1, \ldots, n\}$, then merge the blocks of P containing i and j if they differ, or split uniformly the block of P containing i and j otherwise. This chain on Π_n is considered in an article by Persi Diaconis, Eddy Mayer-Wolf, Ofer Zeitouni, and Martin Zerner [DMWZZ04]. More generally,

beyond random transpositions, Nathanaël Berestycki, Oded Schramm, and Ofer Zeitouni proved in [BSZ11] that, for every $n \geqslant k \geqslant 2$, the random walk on $\mathcal{S}_n$ with i.i.d. uniform k-cycles as steps converges to the uniform distribution on $\mathcal{S}_n$ after approximately $(1/k)n\log(n)$ steps.

Aart Stam's algorithm for sampling uniformly from partitions can be found in [Sta83] and Knuth's book [Knu05, Volume 4A]. Partitions of a finite set should not be confused with integer partitions, which are related to Alfred Young's or Norman Ferrers' diagrams. The theorem of Paul Erdős and Tibor Gallai appears in [EG60] and has been reproved by several authors, including Claude Berge [Ber76]. A short, constructive proof can be found in an article by Amitabha Tripathi, Sushmita Venugopalan, and Douglas West [TVW10]. The configuration algorithm goes back to Béla Bollobás [Bol80] and was refined and extended notably by Brendan McKay and Nicholas Wormald [MW90]. It is discussed in the lecture notes of Charles Bordenave [Bor14a] and Remco van der Hofstad [vdH14]. Let M_n be the random multigraph of the configuration model with vertices $\{1,\dots,n\}$ and degree sequence $d_{n,1},\dots,d_{n,n}$, and let $p_{n,k} = (1/n)\sum_{i=1}^{n} \mathbb{1}_{\{d_{n,i}=k\}}$ be the proportion of vertices of degree k. Suppose there exists a distribution $(p_k)_{k\geqslant 1}$ such that $\lim_{n\to\infty} p_{n,k} = p_k$ and such that the first two moments converge:

$$\begin{aligned}\mu &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} d_{n,i} = \lim_{n\to\infty} \sum_{k=1}^{\infty} \frac{k}{n} \sum_{i=1}^{n} \mathbb{1}_{\{d_{n,i}=k\}} \\ &= \lim_{n\to\infty} \sum_{k=1}^{\infty} k p_{n,k} = \sum_{k=1}^{\infty} k p_k < \infty\end{aligned}$$

and

$$\nu = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{d_{n,i}(d_{n,i}-1)}{\sum_{j=1}^{n} d_{n,j}} = \frac{1}{\mu}\sum_{k=1}^{\infty} k(k-1)p_k < \infty,$$

Then one can show that the probability that M_n is a graph tends to $e^{-\frac{1}{2}\nu - \frac{1}{4}\nu^2}$ as $n \to \infty$. Thus the rejection sampling method of Theorem 4.6 for the configuration model remains reasonable if $n \gg 1$.

Although a special case of graphs, trees form a very rich class of structures. Surveys can be found in the books by Donald Knuth [Knu05, Volume 4A] and Michael Drmota [Drm09]. The algorithm of David Arnold and Ronan Sleep appears in [AS80], and in the survey by Jarmo Siltaneva and Erkki Mäkinen [SM02] on algorithms for sampling random binary trees. The uniform distribution on the set of plane binary trees $\mathcal{T}_n$ can also be sampled by an attractive recursive algorithm due to Jean-Luc Rémy [Ré85]: from an element of $\mathcal{T}_{n-1}$, construct an element of $\mathcal{T}_n$ by choosing a vertex uniformly at random. If it is a leaf, give it two children, which will therefore be leaves; otherwise, replace it with a new vertex whose children are a leaf and the original vertex. Other random tree algorithms include David Wilson's algorithm [Wil96] for generating a uniformly random *spanning tree*, and Luc Devroye's algorithm

[Dev12] for sampling a Galton–Watson tree conditioned to have a fixed size. The book by Russel Lyons and Yuval Peres [LP15] on probability on trees and networks is also a valuable reference.

5

Gibbs measures

Keywords. Gibbs measure; Metropolis–Hastings algorithm; Gibbs sampler; simulated annealing algorithm; Propp–Wilson algorithm. **Tools.** Markov chain; Laplace transform; Boltzmann entropy. **Difficulty.** *

This chapter concerns general algorithms for sampling probability measures using Markov chains. We work with discrete probability measures on finite sets. Throughout the chapter, E denotes a finite set, typically of very large cardinality, such as a symmetric group. The methods and concepts extend far beyond this restrictive setting, which has the advantage of requiring few technical tools.

5.1 Gibbs measures

Let $H : E \to \mathbb{R}$ be a function called the *energy* or *Hamiltonian*. For every $\beta \in \mathbb{R}$, the *Gibbs measure* μ_β is the probability measure on E given, for every $x \in E$, by

$$\mu_\beta(x) = \frac{1}{Z_\beta} e^{-\beta H(x)}, \quad \text{où} \quad Z_\beta = \sum_{y \in E} e^{-\beta H(y)}.$$

The normalizing constant Z_β is called the *partition function*. Allowing H to take the value $+\infty$, every probability measure μ on E is a Gibbs measure with $\beta = 1$ and $H(x) = -\log(\mu(x))$ for every $x \in E$, which gives[1] $H(x) = +\infty$ if and only if $\mu(x) = 0$.

When $\beta > 0$, the probability $\mu_\beta(x)$ is larger when the energy $H(x)$ is smaller (relative to the other values of H). The probability measure μ_β therefore favors configurations of low energy, increasingly so as β grows. When $\beta = 0$, we obtain the uniform probability measure on E.

Here are some examples of Gibbs measures:

[1] With the natural conventions $\log(0) = -\infty$ and $\exp(-\infty) = 0$.

- *The Ising model.* In this model, $E = \{-1, 1\}^{\Lambda}$, where $\varnothing \neq \Lambda \subset \mathbb{Z}^d$. Each $x \in E$ represents the magnetic configuration of the atoms in a piece of metal. For every lattice site $i \in \Lambda$, the value $x_i \in \{-1, 1\}$ is the magnetic orientation of the atom at site i, sometimes called its *spin*. The energy of configuration $x \in E$ has the form
$$H(x) = J \sum_{|i-j|_1=1} x_i x_j + h \sum_{i \in \Lambda} x_i$$
where $|i-j|_1 := |i_1 - j_1| + \cdots + |i_d - j_d|$, J is a coupling constant, and h models an external magnetic field;
- *The Ewens distribution.* Here $E = \mathcal{S}_n$ is the symmetric group, and the energy $H(\sigma) = |\sigma|$ of the permutation $\sigma \in E$ is given by the number of cycles $|\sigma|$ of σ (Theorem 14.3, $\beta = -\log(\theta)$);
- *Random graphs.* Here E is the set of graphs with n vertices, and the energy $H(g)$ of a graph $g \in E$ is its number of edges g. This gives the Erdős–Rényi random graph distribution (Theorem 16.14, $\beta = -\log(p) - \log(1-p)$). Other energies may be considered, such as the number of cycles in the graph;
- *The Gaussian free field.* Here $E = \mathbb{R}^{\Lambda}$ with $\varnothing \neq \Lambda \subset \mathbb{Z}^d$, and the energy $H(x)$ of the "field" or "interface" $x \in E$ is given by
$$H(x) = \sum_{|i-j|_1=1} (x_i - x_j)^2,$$
with $|i-j|_1 := |i_1 - j_1| + \cdots + |i_d - j_d|$ and, by convention, $x_i = 0$ if $i \notin \Lambda$. This is an example where E is infinite and continuous, studied in particular in Chapter 2 in connection with the simple random walk on $\mathbb{Z}^d$;
- *Directed polymers.* See Chapter 19.

Gibbs measures have the remarkable property of maximizing Boltzmann entropy at fixed mean energy. Recall that, if μ is a probability measure on E, its *Boltzmann entropy*, denoted by $\mathrm{Ent}(\mu)$, is defined, with the convention $0 \log(0) = 0$, by
$$\mathrm{Ent}(\mu) = -\sum_{x \in E} \mu(x) \log(\mu(x)).$$

Theorem 5.1 (Maximum entropy at fixed energy)**.** *Let $H : E \to \mathbb{R}$ be a function with minimum h_- and maximum h_+, and let $m \in [h_-, h_+]$, with $m \notin \{h_-, h_+\}$ if $h_- \neq h_+$. Let $\mathcal{M} = \{\mu : \mathbb{E}_\mu(H) = m\}$ be the set of probability measures on E with mean energy m. Then there exists $\beta \in \mathbb{R}$ such that $\mu_\beta \in \mathcal{M}$, and $\mathrm{Ent}(\mu) \leqslant \mathrm{Ent}(\mu_\beta)$ for every $\mu \in \mathcal{M}$, with equality if and only if $\mu = \mu_\beta$.*

Proof. Consider the function
$$f : \beta \in \mathbb{R} \mapsto \mathbb{E}_{\mu_\beta}(H) = \sum_{x \in E} H(x) \mu_\beta(x) = \sum_{x \in E} H(x) \frac{1}{Z_\beta} e^{-\beta H(x)}.$$

If $h_- = h_+$, then H is constant and equal to $m = h_- = h_+$, μ_β is the uniform measure on E for every $\beta \in \mathbb{R}$, and $\mathcal{M}$ is the set of all probability measures on E. If $h_- \neq h_+$, then f is strictly decreasing since[2]

$$\begin{aligned} f'(\beta) &= -\frac{1}{Z_\beta} \sum_{x \in E} H(x)^2 e^{-\beta H(x)} + \left(\frac{1}{Z_\beta} \sum_{x \in E} H(x) e^{-\beta H(x)} \right)^2 \\ &= -\mathrm{Var}_{\mu_\beta}(H) < 0, \end{aligned}$$

because H is not constant and μ_β has full support. Moreover, in this case, writing $M = \{x \in E : H(x) = h_-\}$ gives, for every $x \in E$,

$$\mu_\beta(x) = \frac{e^{-\beta(H(x)-h_-)}}{|M| + \sum_{y \notin M} e^{-\beta(H(y)-h_-)}} \underset{\beta \to +\infty}{\longrightarrow} \begin{cases} |M|^{-1} & \text{si } x \in M, \\ 0 & \text{sinon,} \end{cases}$$

Thus μ_β converges to the uniform measure on M as $\beta \to +\infty$, and consequently the image of $\mathbb{R}$ under f is the interval $]h_-, h_+[$.

If μ is a probability measure on E with the same mean energy as μ_β, then

$$\begin{aligned} \mathrm{Ent}(\mu_\beta) - \mathrm{Ent}(\mu) &= \sum_{x \in E} (\log(Z_\beta) + \beta H(x)) \mu_\beta(x) + \sum_{x \in E} \mu(x) \log(\mu(x)) \\ &= \sum_{x \in E} (\log(Z_\beta) + \beta H(x)) \mu(x) + \sum_{x \in E} \mu(x) \log(\mu(x)) \\ &= -\sum_{x \in E} \mu(x) \log(\mu_\beta(x)) + \sum_{x \in E} \mu(x) \log(\mu(x)) \\ &= \sum_{x \in E} \mu(x) \log\left(\frac{\mu(x)}{\mu_\beta(x)} \right) \\ &= \sum_{x \in E} \Phi\left(\frac{\mu(x)}{\mu_\beta(x)} \right) \mu_\beta(x) \\ &\geqslant \Phi\left(\sum_{x \in E} \frac{\mu(x)}{\mu_\beta(x)} \mu_\beta(x) \right) = \Phi(1) = 0, \end{aligned}$$

where we used Jensen's inequality for the distribution μ_β and the convex function

$$x \geqslant 0 \mapsto \Phi(x) := x \log(x).$$

Since Φ is strictly convex, equality holds if and only if $\mu = \mu_\beta$. □

5.2 The Metropolis–Hastings algorithm

In this section, we consider sampling a Gibbs measure on a finite space E, with energy $H : E \to \mathbb{R}$ and parameter $\beta > 0$, defined for every $x \in E$ by

[2] This calculation is typical of "exponential families" in statistics.

$$\mu_\beta(x) = \frac{1}{Z_\beta} e^{-\beta H(x)}, \quad \text{où} \quad Z_\beta = \sum_{x \in E} e^{-\beta H(x)}.$$

Even if β and H are known, evaluating the normalizing constant Z_β numerically is often difficult or impossible because E is too large or too complex to enumerate. In contrast, the ratio

$$\frac{\mu_\beta(x)}{\mu_\beta(y)} = e^{\beta H(y) - \beta H(x)}$$

depends only on the energy difference $H(y) - H(x)$, which is often even cheaper to evaluate in practice than $H(x)$ and $H(y)$ separately. The Metropolis–Hastings algorithm constructs and simulates a recurrent aperiodic Markov chain $(X_t)_{t \in \mathbb{N}}$ on E with invariant distribution μ_β, whose transition matrix involves μ_β only through the ratios $\mu_\beta(x)/\mu_\beta(y)$. Since E is finite, Doeblin's condition or the Perron–Frobenius theorem gives exponential convergence of the error[3]: there exist $c > 0$ and $\eta < 1$ such that, for every $t \in \mathbb{N}$,

$$d_{\mathrm{VT}}\left(\mathrm{Loi}(X_t), \mu_\beta\right) \leqslant c\eta^t.$$

This provides an approximate generator of μ_β: for $t \gg 1$,

$$\mu_\beta \approx \mathrm{Loi}(X_t).$$

Theorem 5.2 constructs the transition kernel $\mathbf{P}$ of $(X_t)_{t \in \mathbb{N}}$.

Theorem 5.2 (Metropolis–Hastings)**.** *Let $\mathbf{Q} : E \times E \to [0,1]$ be an irreducible transition kernel on E satisfying, for every $x, y \in E$,*

$$\mathbf{Q}(x,y) = 0 \quad \textit{ssi} \quad \mathbf{Q}(y,x) = 0.$$

Let $\alpha : \mathbb{R}_+ \to]0,1]$ be a function satisfying, for every $u \in \mathbb{R}_+$,

$$\alpha(u) = u\alpha\left(\frac{1}{u}\right).$$

Define $\mathbf{P} : E \times E \to \mathbb{R}$, for every $x, y \in E$, by

$$\mathbf{P}(x,y) = \begin{cases} \mathbf{Q}(x,y)\rho(x,y) & \textit{si } x \neq y, \\ 1 - \sum_{z \neq x} \mathbf{P}(x,z) & \textit{si } x = y, \end{cases}$$

where

$$\rho(x,y) = \alpha\left(\frac{\mu_\beta(y)\mathbf{Q}(y,x)}{\mu_\beta(x)\mathbf{Q}(x,y)}\right) \mathbb{1}_{\mathbf{Q}(x,y)>0}.$$

Then $\mathbf{P}$ is an irreducible positive recurrent transition kernel on E with reversible invariant distribution μ_β; it is aperiodic if $\alpha < 1$ or if $\mathbf{Q}$ is aperiodic.

[3] These bounds are rarely useful in practice, except in certain special cases.

The two most common choices of α are

$$\alpha(u) = \min(1, u) \quad \text{et} \quad \alpha(u) = \frac{u}{1+u}, \quad u \in \mathbb{R}_+.$$

The first has an intuitive interpretation. The second ensures that $\alpha < 1$.

If the measure μ_β is reversible for the kernel $\mathbf{Q}$, as happens for instance when μ_β is the uniform measure on E and $\mathbf{Q}(x,y) = \mathbf{Q}(y,x)$ for every $x, y \in E$, then the Metropolis–Hastings construction simply amounts to taking $\mathbf{P} = (1-\varepsilon)\mathbf{Q} + \varepsilon\mathbf{I}$, where $0 < \varepsilon < 1$ ensures aperiodicity.

Proof of Theorem 5.2. We have $0 \leqslant \mathbf{P}(x,y) \leqslant \mathbf{Q}(x,y)$ for every x, y in E with $x \neq y$, so $\mathbf{P}$ is a transition kernel. Since $\alpha > 0$, the kernel $\mathbf{P}$ inherits the transition structure of $\mathbf{Q}$, so it is irreducible and is aperiodic if $\mathbf{Q}$ is. If $\alpha < 1$, then $\mathbf{P}(x,x) > 0$ for every $x \in E$, so $\mathbf{P}$ is aperiodic. Since $\mathbf{P}$ is irreducible on the finite set E, it has a unique invariant distribution. The property of α gives, for every $x, y \in E$, $x \neq y$, $\mathbf{Q}(x,y) \neq 0$,

$$\begin{aligned}
\mu_\beta(x)\mathbf{P}(x,y) &= \mu_\beta(x)\mathbf{Q}(x,y)\alpha\left(\frac{\mu_\beta(y)\mathbf{Q}(y,x)}{\mu_\beta(x)\mathbf{Q}(x,y)}\right) \\
&= \mu_\beta(x)\mathbf{Q}(x,y)\frac{\mu_\beta(y)\mathbf{Q}(y,x)}{\mu_\beta(x)\mathbf{Q}(x,y)}\alpha\left(\frac{\mu_\beta(x)\mathbf{Q}(x,y)}{\mu_\beta(y)\mathbf{Q}(y,x)}\right) \\
&= \mu_\beta(y)\mathbf{P}(y,x).
\end{aligned}$$

Thus μ_β is reversible for $\mathbf{P}$ and hence is the invariant distribution of $\mathbf{P}$. □

The Metropolis–Hastings algorithm simulates trajectories of the chain $(X_t)_{t\in\mathbb{N}}$ with kernel $\mathbf{P}$. This reduces to sampling the discrete distribution $\mathbf{P}(x,\cdot)$ for an arbitrary x. To do so, let Y be a random variable on E with distribution $\mathbf{Q}(x,\cdot)$, and U a uniform random variable on $[0,1]$ independent of Y. Define the random variable Z by $Z = Y$ if $U < \rho(x,Y)$ and $Z = x$ otherwise. Then $Z \sim \mathbf{P}(x,\cdot)$, since for every $y \neq x$,

$$\mathbb{P}(Z = y) = \mathbb{P}(U < \rho(x,Y), Y = y) = \rho(x,y)\mathbf{Q}(x,y) = \mathbf{P}(x,y).$$

In other words, the proposal Y with distribution $\mathbf{Q}(x,\cdot)$ is accepted or rejected according to whether $U < \rho(x,Y)$ holds. Evaluating $\mathbf{P}(x,x)$ is unnecessary. We call ρ the *acceptance-rejection function*, and $\mathbf{Q}$ the *proposal or exploration kernel*. The kernel $\mathbf{Q}$ must be easy to sample. In practice, one often uses the simple nearest-neighbor random walk kernel on the graph associated with a natural distance on E:

$$\mathbf{Q}(x,y) = \frac{\mathbb{1}_{y\in V_x}}{|V_x|} \quad \text{où} \quad V_x := \{y \in E : \operatorname{dist}(x,y) = 1\}.$$

If the graph is regular, that is, $|V_x| = |V_y|$ for every $x, y \in E$, and if μ_β is the uniform measure on E, then $\mathbf{P} = (1-\varepsilon)\mathbf{Q} + \varepsilon\mathbf{I}$, where $0 < \varepsilon < 1$, and we can take $\varepsilon = 0$ if $\mathbf{Q}$ is aperiodic.

Example 5.3 (The Gibbs sampler)**.** *Let $F \subset \mathbb{Z}$ be a finite nonempty subset of $\mathbb{Z}$, such as $F = \{-1, 1\}$, as in the Ising model mentioned in the introduction. Let $\Lambda \subset \mathbb{Z}^d$ also be a finite nonempty subset of $\mathbb{Z}^d$, and finally let $E = F^\Lambda$. A natural exploration kernel $\mathbf{Q}$ on E starts from the current configuration $x \in E$, selects a site $i \in \Lambda$ at random*[4]*, then updates x_i using the conditional distribution*

$$\mu_\beta(x_i \,|\, x_{-i}) \quad \textit{où} \quad x_{-i} := (x_j)_{j \neq i}.$$

This conditional distribution is easy to compute and sample, for example in the Ising model. This amounts to taking, for every $x, y \in E$,

$$\mathbf{Q}(x, y) = \sum_{i \in \Lambda} q(i) \mu_{x,i}(y_i) \mathbb{1}_{\{x_{-i} = y_{-i}\}}, \quad \textit{où} \quad \mu_{x,i} := \mu_\beta(x_i \,|\, x_{-i})$$

where q is a distribution on Λ assigning positive mass to every state, such as the uniform distribution on Λ. The Gibbs sampler *is the name given to the Metropolis–Hastings algorithm when the exploration kernel $\mathbf{Q}$ is defined in this way.*

5.3 The simulated annealing algorithm

The *simulated annealing algorithm* aims to minimize a function $H : E \to \mathbb{R}$ on a finite set E. To this end, consider the Gibbs measure μ_β with energy H and parameter $\beta > 0$, defined for every $x \in E$ by

$$\mu_\beta(x) = \frac{1}{Z_\beta} e^{-\beta H(x)}, \quad \text{où} \quad Z_\beta = \sum_{x \in E} e^{-\beta H(x)}.$$

It is convenient to interpret the parameter $1/\beta$ as a temperature, or as a variance by analogy with Gaussian distributions. The distribution μ_β favors low-energy configurations, increasingly so as the temperature $1/\beta$ decreases. In the proof of Theorem 5.1, we established the following convergence[5]:

$$\mu_\infty(x) := \lim_{\beta \to \infty} \mu_\beta(x) = \frac{\mathbb{1}_{x \in M}}{|M|} \quad \text{où} \quad M := \Big\{ y \in E : H(y) = \inf_E H \Big\}.$$

The distribution μ_∞ is uniform on the set M of global minima of H. Do not confuse μ_∞ with the distribution μ_0, which is uniform on E. This suggests a stochastic algorithm for sampling μ_∞ called *simulated annealing*[6], which uses Metropolis–Hastings with a temperature $1/\beta = 1/\beta_t$ decreasing over time t. The resulting Metropolis–Hastings Markov chain $(X_t)_{t \in \mathbb{N}}$ is no longer time-homogeneous. Consider, for example, the case $\alpha(u) = \min(1, u)$ for every $u \in$

[4] Sometimes a systematic sweep through the sites is preferred.
[5] A very simple instance of the *Laplace principle*.
[6] Named after the annealing process in metallurgy.

$\mathbb{R}$, with $\mathbf{Q}$ satisfying $\mathbf{Q}(x,y) = \mathbf{Q}(y,x)$ for every $x, y \in E$. The acceptance-rejection function depends on time and is given by

$$\rho_t(x,y) = \alpha\left(\frac{\mu_{\beta_t}(y)}{\mu_{\beta_t}(x)}\right) = \alpha\Big(e^{-\beta_t(H(y)-H(x))}\Big) = e^{-\beta_t \max(0,H(y)-H(x))}.$$

Given that the chain (or algorithm) is in state x at time t, generate a proposal y with distribution $\mathbf{Q}(x,\cdot)$, and accept it if and only if $U < \rho_t(x,y)$, where U is an independent uniform random variable on $[0,1]$. Thus the acceptance probability is 1 if $H(y) \leqslant H(x)$ and $e^{-\beta_t(H(y)-H(x))}$ otherwise (equal to 0 if $1/\beta_t = 0$). The algorithm explores the graph of H, always accepts a transition that lowers the energy H, and accepts an energy-increasing transition with a probability tending to zero as the temperature decreases. Unlike gradient descent, simulated annealing can move "uphill" and thereby escape local minima at the beginning. The temperature still needs to be chosen: excessively rapid cooling may trap the trajectory near a local minimum, whereas excessively slow cooling may prevent concentration near M. A theoretical result ensures that, with a cooling schedule $1/\beta_t$ of the form $c/\log(t)$ as $t \to \infty$, the chain accumulates on points of M. This remarkable theorem is nevertheless of little practical use! The algorithm's performance depends strongly on the regularity of the graph of H relative to the exploration neighborhood structure.

Now consider the *traveling salesman problem* for n cities at positions $v_1, \ldots, v_n$ in space: find a tour of minimal length, that is, a permutation $x \in \mathcal{S}_n$ in the symmetric group $\mathcal{S}_n$ minimizing

$$x \in \mathcal{S}_n \mapsto H(x) := \sum_{i=1}^{n} \mathrm{dist}\big(v_{x(i)}, v_{x(i+1)}\big) \quad \text{où} \quad x(n+1) := x(1).$$

Here the tour length plays the role of energy. The cardinality of $E = \mathcal{S}_n$ is enormous: $n!$. With a fixed total computational budget, we can only explore a restricted subset $F \subset E$ and use the minimum of H over F to approximate the minimum of H over E. Such a subset $F \subset E$ can be generated deterministically[7], or stochastically, for example by simulated annealing. One possible exploration kernel $\mathbf{Q}$ is that of the simple random walk on $\mathcal{S}_n$ associated with transpositions, given by

$$\mathbf{Q}(x,y) = \begin{cases} \frac{2}{n(n-1)} & \text{si } y = x\tau \text{ pour une transposition } \tau, \\ 0 & \text{sinon.} \end{cases}$$

Since transpositions generate the symmetric group $\mathcal{S}_n$, the transition kernel $\mathbf{Q}$ is irreducible. A richer exploration kernel can of course be used, for example exchanging k randomly chosen cities.

[7] The deterministic Steinhaus–Johnson–Trotter algorithm lists the elements of $\mathcal{S}_n$ without comparing them to previously generated elements. Algebraically, it traverses a Cayley graph of $\mathcal{S}_n$; geometrically, it traverses adjacent vertices of the polytope called the permutohedron.

Chapter 20 studies the order of magnitude of $H(x)$ when the positions $v_1, \ldots, v_n$ of the n cities are i.i.d. random variables and n is large.

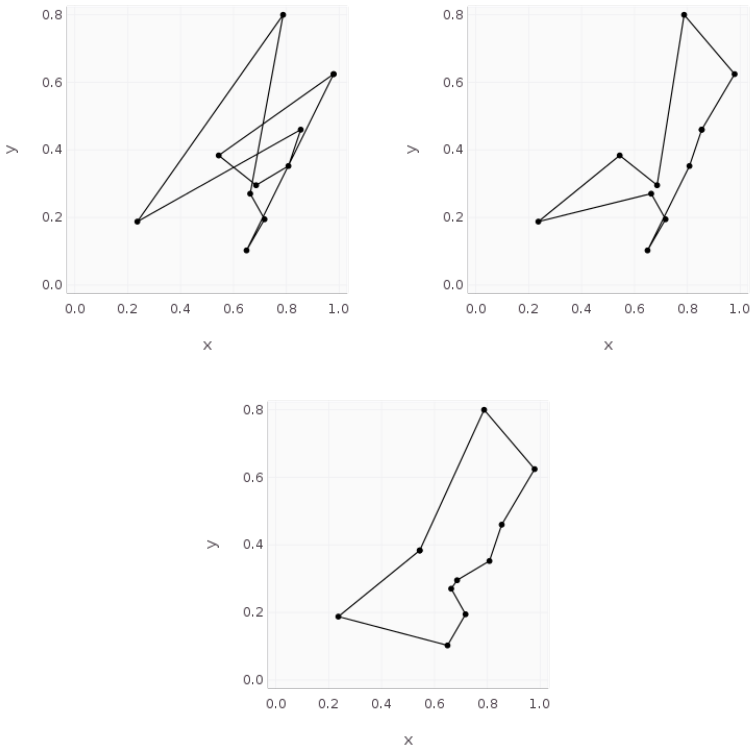


Fig. 5.1. Steps of simulated annealing for the traveling salesman problem ($n = 10$).

5.4 The Propp–Wilson algorithm

Metropolis–Hastings provides approximate sampling of a Gibbs measure μ_β on a finite set E through a Markov chain with kernel $\mathbf{P}$ admitting μ_β as a reversible invariant distribution (Theorem 5.2). We now present an algorithm for exact sampling of μ_β.

Let $\mathbf{P}$ be an irreducible Markov kernel on E with invariant probability measure μ_β. View Markov chains as random recurrence sequences. Thus let $g : E \times [0,1] \to E$ be a function such that $g(x,U) \sim \mathbf{P}(x,\cdot)$ for every $x \in E$, where U is uniformly distributed on $[0,1]$. Write $G : E \to E$ for the random function defined by $G(x) = g(x,U)$. Let $(G_n)_{n \geqslant 0}$ be an i.i.d. sequence of

random functions from E to E with the distribution of G, constructed like G using a sequence $(U_n)_{n\geqslant 0}$ of i.i.d. random variables uniformly distributed on $[0,1]$. For every n, construct the random maps $A_n : E \to E$ and $B_n : E \to E$ by composition as follows:

$$A_n = G_n \circ \cdots \circ G_1 \quad \text{et} \quad B_n = G_1 \circ \cdots \circ G_n.$$

By convention, A_0 and B_0 are the identity map on E. For every $x \in E$ and every n, the random variables $A_n(x)$ and $B_n(x)$ have distribution $\mathbf{P}^n(x,\cdot)$. The sequences $(A_n(x))_{n\geqslant 0}$ and $(B_n(x))_{n\geqslant 0}$ have the same 1-dimensional marginal distributions, but different laws as sequences. The sequence $(A_n(x))_{n\geqslant 0}$ is a Markov chain on E with kernel $\mathbf{P}$ and initial distribution δ_x. In contrast, $(B_n(x))_{n\geqslant 0}$ is not a Markov chain, because time is "reversed." Define the contraction times T_A and T_B with values in $\mathbb{N}\cup\{\infty\}$ by

$$T_A = \inf\{n \geqslant 0 : \operatorname{card}(A_n(E)) = 1\} \quad \text{où} \quad A_n(E) = \{A_n(x) : x \in E\}$$

and similarly,

$$T_B = \inf\{n \geqslant 0 : \operatorname{card}(B_n(E)) = 1\} \quad \text{où} \quad B_n(E) = \{B_n(x) : x \in E\}.$$

We call G *contracting* if $p = \mathbb{P}(\operatorname{card}(G(E)) = 1) > 0$.

Theorem 5.4 (Convergence at coalescence for the reversed sequence)**.** *If G is contracting, then $\mathbb{P}(T_B < \infty) = 1$; for every $x \in E$, $B_{T_B}(x) \sim \mu_\beta$, and $B_n(x) = B_{T_B}(x)$ for every $n \geqslant T_B$.*

Note that $(A_n(x))_{n\geqslant 0}$ converges in distribution to μ, whereas $(B_n(x))_{n\geqslant 0}$ converges almost surely to a random variable with distribution μ.

Proof. The events $C_n = \{\operatorname{card}(G_n(E)) = 1\}$ are independent and have the same probability $p = \mathbb{P}(\operatorname{card}(G(E)) = 1) > 0$. By the Borel–Cantelli lemma, almost surely, $\operatorname{card}(G_n(E)) = 1$ for infinitely many values of n, and in particular $\mathbb{P}(T_B < \infty) = 1$. The random time T_B almost surely defines a random singleton $\{x_B\}$ such that $B_{T_B}(x) = x_B$ for every $x \in E$. It follows that $B_n(x) = x_B$ for every $x \in E$ and every $n \geqslant T_B$, since

$$B_n(x) = B_{T_B}((G_{T_B+1} \circ \cdots \circ G_n)(x)) = x_B.$$

The trajectories of $(B_n(x))_{n\geqslant 0}$ *coalesce* regardless of the initial state x. Let μ_B be the distribution of x_B. By dominated convergence, for every $x \in E$ and every bounded function $f : E \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}(f(B_n(x)) = \lim_{n\to\infty} \mathbb{E}(f(B_n(x))\mathbb{1}_{\{T_B\leqslant n\}}) = \mathbb{E}(f(x_B)) = \mu_B f.$$

But $\mathbb{E}(f(B_n(x))) = \mathbf{P}^n(x,\cdot)f = \sum_{y\in E} \mathbf{P}^n(x,y)f(y)$, hence, for every $x, y \in E$,

$$\lim_{n\to\infty} \mathbf{P}^n(x,y) = \mu_B(y).$$

Thus $\mu_B = \mu_\beta$, completing the proof. □

If G is contracting, then $\mathbb{P}(T_A < \infty) = 1$ also holds, and T_A almost surely defines a random singleton $\{x_A\}$ satisfying $A_{T_A}(x) = x_A$ for every $x \in E$. However, nothing ensures that $A_n(x) = x_A$ for $n > T_A$ and every $x \in E$, or that x_A has distribution μ_β. For every $x, y \in E$, the sequence $((A_n(x), A_n(y)))_{n \geqslant 0}$ is a coalescent coupling: both components are chains with the same kernel and stay together after their first meeting. We call $((B_n(x), B_n(y)))_{n \geqslant 0}$ *coupling from the past*[8].

When G is not contracting, one can use the kernel $\mathbf{P}^r$, $r \geqslant 1$, which also has μ_β as its invariant distribution. In practice, determining the coupling (or coalescence) time from the past T_B is inconvenient because it involves all possible initial states. This time can nevertheless be controlled when the kernel $\mathbf{P}$ has a monotonicity property relative to an order on the state space E, by reducing to a minimal and a maximal state, as the detailed example in Section 5.5 shows.

5.5 Modeling a chemical compound

In this section, we implement a Metropolis–Hastings algorithm and an exact sampling algorithm for a Gibbs measure modeling the distribution of two strongly repelling chemical compounds, called types $+1$ and -1. The system is assumed to be open, so the number of particles of each compound is unconstrained. For simplicity, the compounds occupy a flat surface identified with a discrete square $\Lambda = \{1, \dots, L\}^2 \subset \mathbb{Z}^2$. Each site $i \in \Lambda$ is either empty or occupied by a compound of type $+1$ or -1. Repulsion between different types is modeled by prohibiting compounds of different types from being neighbors, that is, at distance 1 on the grid. The configuration space is therefore

$$E = \{x \in \{-1, 0, +1\}^\Lambda : \forall i, j \in \Lambda, |i - j|_1 = 1 \Rightarrow x_i x_j \neq -1\}.$$

Consider the energy

$$x \in E \mapsto H(x) = \mathrm{card}\{i \in \Lambda : |x_i| = 1\}$$

which counts the number of compounds present, and the associated Gibbs measure

$$\mu_\beta = \frac{1}{Z_\beta} e^{-\beta H(x)}, \quad x \in E, \quad \beta \in \mathbb{R}, \quad Z_\beta = \sum_{x \in E} e^{-\beta H(x)}.$$

The smaller (respectively larger) the real number β, the more the distribution μ_β favors dense (respectively sparse) configurations.

Remark 5.5 (Rejection sampling)**.** *Here is a first method for sampling μ_β exactly. Independently assign to each site of Λ a value -1, 0, or $+1$ with respective probabilities*

[8] A coalescent coupling constructed by running the dynamics from past times.

$$\frac{e^{-\beta}}{1+2e^{-\beta}}, \quad \frac{1}{1+2e^{-\beta}}, \quad \textit{et} \quad \frac{e^{-\beta}}{1+2e^{-\beta}}.$$

Accept the resulting configuration if it belongs to E; otherwise, repeat the procedure. This simple method is impractical because E is a tiny subset of $\{-1,0,+1\}^{\Lambda}$.

Let us now construct a Metropolis–Hastings algorithm. For every $u \in [-1,1], i \in \Lambda$, let

$$h_{i,u} : E \to \{-1,0,+1\}^{\Lambda}$$

be the function defined, for every $x \in E$ and $j \in \Lambda$, by

$$(h_{i,u}(x))_j = \begin{cases} x_j & \text{si } j \neq i; \\ -1 & \text{si } j = i \text{ et } u < -1/(1+2e^{-\beta}); \\ 0 & \text{si } j = i \text{ et } -1/(1+2e^{-\beta}) \leqslant u \leqslant 1/(1+2e^{-\beta}); \\ +1 & \text{si } j = i \text{ et } u > 1/(1+2e^{-\beta}). \end{cases}$$

Finally, set

$$g_{i,u}(x) = \begin{cases} h_{i,u}(x) & \text{si } h_{i,u}(x) \in E, \\ x & \text{sinon.} \end{cases}$$

Let $(U_n)_{n\geqslant 1}$ and $(V_n)_{n\geqslant 1}$ be independent sequences of i.i.d. random variables uniformly distributed on $[-1,1]$ and Λ, respectively, and let $G_n := g_{V_{n+1},U_{n+1}}$. Define the random recurrence sequence $(X_n)_{n\geqslant 0}$ by

$$X_{n+1} = G_n(X_n).$$

Theorem 5.6. *The sequence $(X_n)_{n\in\mathbb{N}}$ is an irreducible, recurrent, aperiodic Markov chain on E with reversible invariant distribution μ_β.*

Proof. The sequence $(X_n)_{n\in\mathbb{N}}$ is a Markov chain written as a random recurrence sequence on the finite set E. Every configuration can reach the zero configuration (no compound present), and conversely. The chain is therefore irreducible and recurrent. Let $\mathbf{P}$ be the transition kernel of $(X_n)_{n\in\mathbb{N}}$. If x and y belong to E, then $\mathbf{P}(x,y)$ is strictly positive if x and y agree at all but one site. Reversibility is checked case by case. For example, suppose x and y belong to E and, for $i_0 \in \Lambda$,

$$x_{i_0} = +1, \quad y_{i_0} = 0 \quad \text{et} \quad x_i = y_i \text{ pour } i \neq i_0.$$

Then $H(x) = H(y) + 1$, and hence

$$\begin{aligned} \mu_\beta(x)\mathbf{P}(x,y) &= \mu_\beta(y)e^{-\beta}\frac{2}{L^2(1+2e^{-\beta})} \\ &= \mu_\beta(y)\frac{1}{L^2}\left(1 - \frac{1}{1+2e^{-\beta}}\right) = \mu_\beta(y)\mathbf{P}(y,x). \end{aligned}$$

□

This random recurrence sequence also yields a perfect sampling algorithm. In Propp–Wilson, must we really follow all trajectories indexed by the points of E as initial conditions? The following fundamental property greatly simplifies the problem: two suffice!

Lemma 5.7 (Monotonicity)**.** *For every $u \in [-1,1]$ and $i \in \Lambda$, the function $x \in E \mapsto g_{i,u}(x)$ is increasing for the partial order on E defined by*

$$x \leqslant y \quad ssi \quad x_i \leqslant y_i \text{ pour tout } i \in \Lambda.$$

Moreover, the configuration denoted by $-\mathbb{1}$ (respectively $+\mathbb{1}$), consisting entirely of -1 (respectively $+1$), is a least (respectively greatest) element for this partial order.

Monotonicity now gives the bounds

$$G_n(-\mathbb{1}) \leqslant G_n(x) \leqslant G_n(+\mathbb{1}),$$

showing that coalescence from arbitrary x occurs before coalescence from $\pm\mathbb{1}$. In practice, it therefore suffices to iterate the algorithm until the trajectories starting from $\pm\mathbb{1}$ coalesce.

Finally, the kernel $\mathbf{P}$ is not contracting in the sense of Theorem 5.4, but $\mathbf{P}^{L^2}$ is, since for every $x \in E$, $\mathbf{P}^{L^2}(x,\cdot)$ assigns positive mass to the zero configuration.

5.6 Further reading

For an introduction to statistical physics, a major meeting point of physics and mathematics, and in particular explicit examples involving the Ising model, see Rodney Baxter's book [Bax82]. Models inspired by Ising are used in imaging to describe spatial noise.

The material on Metropolis–Hastings, simulated annealing, and Propp–Wilson is inspired by an earlier book [BC07, Chapter 4] and Thierry Bodineau's course at École Polytechnique [Bod14]. Metropolis–Hastings was introduced by Nicholas Metropolis, Arianna Rosenbluth, Marshall Rosenbluth, Augusta Teller, and Edward Teller [MRR$^+$53], then generalized by Keith Hastings [Has70]. It is now one of the MCMC (Markov chain Monte Carlo) methods, which use Markov chains to approximate expectations by Monte Carlo. MCMC methods are central to quantitative implementations of Bayesian statistics, discussed for instance in the book by Christian Robert and George Casella [RC04]. Particle versions can improve practical performance, particularly when the distribution μ is multimodal.

In metallurgy, annealing reheats a metal to escape local energy minima and obtain a low-energy metallic structure with greater strength. This inspired simulated annealing and explains its name. Simulated annealing converges

theoretically in infinite time under a temperature schedule decreasing in logarithmic stages, but this asymptotic result has little practical relevance. An analysis appears, for example, in Étienne Pardoux's book [Par07] and Olivier Catoni's survey [Cat99]. Improved versions, such as the Wang–Landau algorithm introduced by Fugao Wang and David Landau [WL01], are an active research topic. Metropolis–Hastings, simulated annealing, and Wang–Landau all have versions in continuous time and space. Propp–Wilson, proposed by James Propp and David Wilson [PW96, PW98], is presented in the book by David Levin, Yuval Peres, and Elizabeth Wilmer [LPW09] and in Olle Häggström's [Häg02]. The example in Section 5.5 is inspired by a text by Christophe Sabot. Other algorithms for exact sampling from the invariant distribution have been developed, such as James Fill's [Fil98].

6

Internal diffusion limited aggregation

Keywords. Growth model; random sets. **Tools.** Markov chain; martingale; coupling; concentration inequality. **Difficulty.** *

A barrel of radioactive waste is secretly buried. After a few years it becomes porous and starts leaking. To prevent excessive contamination, particle traps have been placed around it. Each captures the first particle that passes through it but then becomes inert (any new particle passes through unhindered). The medium is isotropic: a radioactive particle moves in the same way in every direction as long as it encounters traps that have already been activated, and is captured by the first free trap it encounters. We wish to determine the typical shape of the regions contaminated by the leak. To this end, we introduce a random growth model, called *internal diffusion limited aggregation*, and study in particular its asymptotic properties.

6.1 A diffusion aggregation model

To model the evolution of the polluted region, we introduce the following growth model. We identify space with $\mathbb{Z}^d$, where d is 1, 2, or 3, depending on the number of dimensions considered. Two elements x and y of $\mathbb{Z}^d$ are called neighbors, written $x \sim y$, if

$$|x - y|_1 = |x_1 - y_1| + \cdots + |x_d - y_d| = 1.$$

We place the barrel at the origin $0 \in \mathbb{Z}^d$ and a trap at every lattice point. Write $A(0)$ for the singleton $\{0\}$. The first activated trap will be one of the $2d$ sites neighboring the origin. We model the trajectory of a radioactive particle leaving the barrel by a simple symmetric random walk on $\mathbb{Z}^d$ starting at 0 and stopped when it leaves $A(0)$. Write $A(1)$ for the random set consisting

of 0 and the point reached by the particle on exit. For every neighbor x of 0, $A(1)$ equals $\{0, x\}$ with probability $1/2d$. We then iterate this procedure. Given a set $A(n) \subset \mathbb{Z}^d$, consider a symmetric random walk $(S_k)_{k\geqslant 0}$ starting at 0 and stopped when it exits $A(n)$. Define $A(n+1)$ as the union of $A(n)$ and the exit point of the walk S. This model is known as *internal diffusion limited aggregation*[1].

This produces a sequence $(A(n))_{n\geqslant 0}$ of random sets. The following result collects some immediate properties of this sequence.

Theorem 6.1 (Immediate properties)**.** *The sequence $(A(n))_{n\geqslant 0}$ is increasing with respect to inclusion. The cardinality of $A(n)$ is exactly $n+1$. For every $n \in \mathbb{N}$, if x and y are two elements of $A(n)$, then there exist $x_0, \dots, x_m$ elements of $A(n)$ such that $x_0 = x$, $x_m = y$, and $x_i \sim x_{i+1}$ for $i = 0, \dots, m-1$. We may say that the set $A(n)$ is path connected in $\mathbb{Z}^d$.*

We now wish to study the asymptotic behavior of the sequence $(A(n))_{n\geqslant 0}$. We examine dimension 1 in detail. The much more delicate study in higher dimensions is discussed in Sections 6.6 and 6.7.

6.2 The one-dimensional model

Assume that propagation takes place along a horizontal axis. We thus consider the case $d = 1$. The initial set A_0 is the singleton $\{0\}$; then A_1 equals $\{0, 1\}$ with probability $1/2$ and $\{-1, 0\}$ with probability $1/2$, and so on. Write $G_n = \min A_n$ and $D_n = \max A_n$. The set A_n has the form $A_n = \{G_n, G_n + 1, \dots, D_n - 1, D_n\}$. Since the cardinality of A_n is $n+1$, we have $D_n - G_n = n$. Thus A_n is characterized by $X_n = D_n + G_n$ and, in particular,

$$D_n = \frac{X_n + n}{2} \quad \text{et} \quad G_n = \frac{X_n - n}{2}.$$

The increments $(X_{n+1} - X_n)_{n\geqslant 0}$ can only take the values -1 or 1. More precisely, $X_{n+1} = X_n - 1$ if the walk starting at 0 reaches $G_n - 1$ before $D_n + 1$, and $X_{n+1} = X_n + 1$ if the walk starting at 0 reaches $D_n + 1$ before $G_n - 1$. The symmetry of the underlying random walk ensures that the distribution of X_n is symmetric for every $n \geqslant 0$. In particular, for every $n \geqslant 0$, we have $0 = \mathbb{E}(X_n) = \mathbb{E}(D_n) + \mathbb{E}(G_n)$ and hence $\mathbb{E}(D_n) = n/2$ and $\mathbb{E}(G_n) = -n/2$.

Theorem 6.2 (Markovian evolution)**.** *The sequence $(X_n)_{n\geqslant 0}$ is an inhomogeneous Markov chain on $\mathbb{Z}$ starting at 0, with transitions described by the following relations: for $-n \leqslant i \leqslant n$,*

$$\mathbb{P}(X_{n+1} = i - 1 \,|\, X_n = i) = \frac{n + 2 + i}{2(n+2)}$$

and

[1] Abbreviated IDLA.

$$\mathbb{P}(X_{n+1} = i+1 \,|\, X_n = i) = \frac{n+2-i}{2(n+2)}.$$

Proof. For every $n \geqslant 1$,

$$\{X_n = i\} = \left\{ D_n = \frac{i+n}{2},\ G_n = \frac{i-n}{2} \right\}.$$

The probability $\mathbb{P}(X_{n+1} = i-1 \,|\, X_n = i)$ therefore equals the probability that the simple random walk $(S_n)_{n\geqslant 0}$ starting at 0 reaches $(i-n)/2-1$ before $(i+n)/2+1$. By Theorem 2.2 (gambler's ruin), if $a \leqslant 0$ and $b \geqslant 0$ are distinct integers, and if $T_i := \inf\{n \geqslant 0,\ S_n = i\}$ for $i \in \{a, b\}$ and $T := T_a \wedge T_b$, then T is integrable and

$$\mathbb{P}(T = T_a) = 1 - \mathbb{P}(T = T_b) = \frac{b}{b-a} \quad \text{et} \quad \mathbb{E}(T) = -ab.$$

□

Remark 6.3 (Balancing)**.** *When X_n is strictly positive, $X_{n+1} - X_n$ is more likely to equal -1 than +1. This tendency is reversed when X_n is strictly negative. We therefore expect the process to have a stronger tendency to return to* 0 *than a simple symmetric random walk (whose increment is plus or minus 1 with probability* 1/2*).*

6.3 Almost sure convergence

We establish here the first long-time convergence result for the process $(X_n)_{n\geqslant 0}$ in the one-dimensional model. The proof makes the intuition of Remark 6.3 rigorous by comparing the absolute value of $(X_n)_{n\geqslant 0}$ with that of a simple random walk through coupling.

Theorem 6.4 (Almost sure convergence)**.** *The process $(|X_n|)_{n\geqslant 0}$ is an inhomogeneous Markov chain and*

$$\frac{|X_n|}{n} \xrightarrow[n\to\infty]{\text{p.s.}} 0 \quad \textit{d'où} \quad \frac{G_n}{n} \xrightarrow[n\to\infty]{\text{p.s.}} -\frac{1}{2} \quad \textit{et} \quad \frac{D_n}{n} \xrightarrow[n\to\infty]{\text{p.s.}} \frac{1}{2}.$$

Proof. Considering the possible values of X_n, we see that $|X_{n+1}| = 1$ if $|X_n| = 0$, and that if $|X_n| > 0$ then

$$|X_{n+1}| = \begin{cases} |X_n| - 1 & \text{avec probabilité } \dfrac{1}{2} + \dfrac{|X_n|}{2(n+2)}, \\ |X_n| + 1 & \text{avec probabilité } \dfrac{1}{2} - \dfrac{|X_n|}{2(n+2)}. \end{cases}$$

Thus $(|X_n|)_{n\geqslant 0}$ is an inhomogeneous Markov chain (a random recurrence sequence). To prove that $|X_n|/n \to 0$ almost surely, we use coupling. More

precisely, we construct two processes $(Y_n)_{n\geqslant 0}$ and $(Z_n)_{n\geqslant 0}$ as follows: $Y_0 = Z_0 = 0$ and, for $n \geqslant 0$,

$$Y_{n+1} = \begin{cases} Y_n + 2\mathbb{1}_{\{U_{n+1}<(n+2-Y_n)/2(n+2)\}} - 1 & \text{si } Y_n > 0, \\ 1 & \text{si } Y_n = 0, \end{cases}$$

and

$$Z_{n+1} = \begin{cases} Z_n + 2\mathbb{1}_{\{U_{n+1}<1/2\}} - 1 & \text{si } Z_n > 0, \\ 1 & \text{si } Z_n = 0, \end{cases}$$

where $(U_n)_{n\geqslant 1}$ is a sequence of i.i.d. real random variables uniformly distributed on $[0,1]$. The process $(Y_n)_{n\geqslant 0}$ has the distribution of $(|X_n|)_{n\geqslant 0}$, whereas the process $(Z_n)_{n\geqslant 0}$ has the same distribution as the absolute value of a simple random walk. This coupling satisfies, almost surely, for every $n \in \mathbb{N}$, $0 \leqslant Y_n \leqslant Z_n$. Indeed, if $0 < Y_n \leqslant Z_n$ then $Y_{n+1} \leqslant Z_{n+1}$ by construction. Moreover, if $0 = Y_n \leqslant Z_n$, then Z_n may be zero, in which case $Y_{n+1} \leqslant Z_{n+1}$ or $Z_n \geqslant 2$, since Z_n and Y_n have the same parity. Finally, $Z_n/n \to 0$ a.s. by the strong law of large numbers. □

Remark 6.5 (Expectation)**.** *Theorem 6.2 gives, with* $a_{n+1} := \frac{n+1}{n+2}$,

$$\mathbb{E}(|X_{n+1}|\,|\,|X_n|) = \mathbb{1}_{|X_n|=0} + a_{n+1}|X_n|\mathbb{1}_{|X_n|\neq 0}$$

hence, observing that $|X_n|\mathbb{1}_{|X_n|\neq 0} = |X_n|$,

$$\mathbb{E}(|X_{n+1}|) = \mathbb{P}(X_n = 0) + a_{n+1}\mathbb{E}(|X_n|).$$

6.4 Convergence in distribution

We now turn to the fluctuations of X_n/n around its limit. This section provides elements of a proof of the following result.

Theorem 6.6 (Convergence in distribution)**.** *The following convergence in distribution holds:*

$$\frac{X_n}{\sqrt{n}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}\left(0, \frac{1}{3}\right).$$

Remark 6.7 (Comparison with the central limit theorem)**.** *The rate of convergence in distribution of* $(X_n)_{n\geqslant 0}$ *is the same as for the simple random walk, and the limiting distribution is again Gaussian. Its asymptotic variance, however, is smaller than in the identically distributed case. This is due to the self-stabilizing tendency of the chain* $(X_n)_{n\geqslant 0}$.

Before proving Theorem 6.6, let us examine the behavior of the first moments of X_n/n as n tends to infinity. The distribution of X_n is symmetric, so all its odd moments vanish. For the moment of order 2, we have, for $n \geqslant 0$,

$$\mathbb{E}\big[X_{n+1}^2|\mathcal{F}_n\big] = \frac{n+2-X_n}{2(n+2)}(X_n+1)^2 + \frac{n+2+X_n}{2(n+2)}(X_n-1)^2$$
$$= 1 + \frac{n}{n+2}X_n^2.$$

Setting $x_n(2) = \mathbb{E}(X_n^2)$ and taking expectations in the relation above gives, for $n \geqslant 1$,

$$x_{n+1}(2) = 1 + \frac{n}{n+2}x_n(2) \quad \text{et} \quad x_1(2) = 1.$$

Induction yields

$$x_n(2) = 1 + \frac{1}{n(n+1)}\sum_{k=1}^{n-1} k(k+1),$$

and $(x_n(2)/n)_{n\geqslant 1}$ therefore converges to $1/3$. Similarly, for every $k \geqslant 0$, the sequence $(x_n(2k)/n^k)_{n\geqslant 1}$ has a limit $\mu(2k)$, and the sequence $(\mu(2k))_{k\in\mathbb{N}}$ satisfies

$$\mu(0) = 1 \quad \text{et} \quad \mu(2k+2) = \frac{2k+1}{3}\mu(2k).$$

By Theorem 21.7, the only probability measure whose even moments satisfy this recurrence and whose odd moments vanish is the distribution $\mathcal{N}(0, 1/3)$. Moreover, by Theorem 21.6, convergence of moments to those of a probability measure determined by its moments implies convergence in distribution. We thus obtain convergence in distribution of $X_n/\sqrt{n}$ to the centered Gaussian measure with variance $1/3$.

Another way to establish this convergence is to use the central limit theorem for martingales. Several sets of assumptions are available, with varying degrees of difficulty to check. Here is the theorem we shall use below.

Theorem 6.8 (CLT for martingales)**.** *Let $(M_n)_{n\geqslant 0}$ be a square-integrable martingale with respect to a filtration $(\mathcal{F}_n)_{n\geqslant 0}$, and let $(a_n)_{n\geqslant 0}$ be a deterministic sequence of strictly positive real numbers increasing to infinity. Set*

$$\langle M\rangle_0 = 0 \quad \textit{et} \quad \langle M\rangle_{n+1} = \langle M\rangle_n + \mathbb{E}((\Delta M_{n+1})^2 \,|\, \mathcal{F}_n),$$

where $\Delta M_n := M_n - M_{n-1}$. Assume that

1. *there exists a deterministic $\lambda \geqslant 0$ such that $a_n^{-1}\langle M\rangle_n \xrightarrow[n\to\infty]{\mathbb{P}} \lambda$,*
2. *(Lyapunov condition) there exists $\delta > 0$ such that*

$$\frac{1}{(a_n)^{1+\delta/2}}\sum_{k=1}^{n}\mathbb{E}\big(\Delta M_k^{2+\delta} \,|\, \mathcal{F}_{k-1}\big) \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Then

$$\frac{1}{\sqrt{a_n}}M_n \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,\lambda), \quad \textit{et} \quad \sqrt{a_n}\,\frac{M_n}{\langle M\rangle_n} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,\lambda^{-1}) \quad \textit{si } \lambda > 0.$$

We must slightly modify $(X_n)_{n\geqslant 0}$ to obtain a martingale.

Lemma 6.9 (A martingale in the model)**.** *The process $(M_n)_{n\geqslant 0}$ defined by $M_n = (n+1)X_n$ is a square-integrable martingale with respect to the natural filtration $(\mathcal{F}_n)_{n\geqslant 0}$ of $(X_n)_{n\geqslant 0}$, satisfying*

$$\mathbb{E}((\Delta M_{n+1})^2 \,|\, \mathcal{F}_n) = (n+2)^2 - X_n^2,$$

and

$$\mathbb{E}((\Delta M_{n+1})^4 \,|\, \mathcal{F}_n) = \big((n+2)^2 - X_n^2\big)\big((n+2)^2 + 3X_n^2\big),$$

and

$$\frac{\langle M\rangle_n}{n^3} \xrightarrow[n\to\infty]{\text{p.s.}} \frac{1}{3}.$$

Proof. Let $\alpha \in \{1,2,4\}$. The following quantity

$$\begin{aligned}\mathbb{E}((\Delta M_{n+1})^\alpha | \mathcal{F}_n) &= \mathbb{E}(((n+2)X_{n+1} - (n+1)X_n)^\alpha \,|\, X_n)\\ &= [X_n - (n+2)]^\alpha \frac{n+2+X_n}{2(n+2)} + [X_n + n + 2]^\alpha \frac{n+2-X_n}{2(n+2)},\end{aligned}$$

vanishes if $\alpha = 1$, and hence $(M_n)_{n\geqslant 0}$ is a martingale. If $\alpha \in \{2,4\}$ then

$$\mathbb{E}((\Delta M_{n+1})^\alpha \,|\, \mathcal{F}_n) = ((n+2)^2 - X_n^2)\frac{(n+2-X_n)^{\alpha-1} + (n+2+X_n)^{\alpha-1}}{2(n+2)},$$

which gives the first two desired relations. In particular,

$$\langle M\rangle_n = \sum_{k=0}^{n-1}((k+2)^2 - X_k^2) = \frac{n^3}{3} - \sum_{k=0}^{n-1} X_k^2 + o(n^3).$$

Moreover,

$$\frac{1}{n^3}\sum_{k=0}^{n-1} X_k^2 \leqslant \frac{1}{n}\sum_{k=1}^{n-1}\frac{X_k^2}{k^2}.$$

Since $X_n/n \to 0$ a.s., Cesàro's lemma completes the proof. □

To obtain convergence in distribution of $(M_n)_{n\geqslant 0}$ with the appropriate normalization, it remains to check the Lyapunov condition in Theorem 6.8 with $a_n = n^3$. For $\delta = 2$, we have

$$\sum_{k=1}^{n}\mathbb{E}((\Delta M_k)^4 \,|\, \mathcal{F}_{k-1}) \leqslant \sum_{k=2}^{n+1} k^4 + 2\sum_{k=1}^{n}(k+1)^2 X_{k-1}^2 = \mathcal{O}(n^5).$$

The sequence $(M_n/n^{3/2})_{n\geqslant 1}$ therefore converges in distribution to $\mathcal{N}(0,1/3)$. This immediately yields the convergence in Theorem 6.6.

6.5 A concentration inequality

We supplement the results on the behavior of X_n with a concentration inequality valid for every $n \geqslant 1$ (unlike the convergence in distribution theorem, which is only asymptotic).

Theorem 6.10 (Concentration)**.** *For every $n \geqslant 1$ and every $r > 0$,*

$$\mathbb{P}\left(\frac{|X_n|}{n} \geqslant r\right) \leqslant 2\exp\left(-\frac{3n^2r^2}{8(n+1)}\right).$$

Proof. By definition of the martingale $(M_n)_{n\geqslant 0}$ introduced in Lemma 6.9, for every $r > 0$,

$$\mathbb{P}\left(\frac{|X_n|}{n} \geqslant r\right) \leqslant \mathbb{P}(|M_n| \geqslant n(n+1)r).$$

The martingale $(M_n)_{n\geqslant 0}$ has bounded increments, since for every $n \geqslant 1$,

$$|M_n - M_{n-1}| = |(n+1)X_n - nX_{n-1}| \leqslant n|X_n - X_{n-1}| + |X_n| \leqslant 2n$$

because $|X_n - X_{n-1}| = 1$ and $|X_n| \leqslant n$. By the Azuma–Hoeffding concentration inequality in Lemma 20.3, for every $\lambda > 0$, we have

$$\mathbb{P}(|M_n| \geqslant \lambda) \leqslant 2\exp\left(-\frac{\lambda^2}{2\sum_{k=1}^n (2k)^2}\right).$$

Using the sum-integral comparison

$$\sum_{k=1}^{n}(2k)^2 \leqslant 4\int_1^{n+1} x^2\,dx = \frac{4}{3}(n+1)^3,$$

it remains to replace λ by $n(n+1)r$ to obtain the desired inequality. □

6.6 Higher dimensions

We now work on $\mathbb{Z}^d$ with $d > 1$. As in dimension 1, we expect the limiting shape of the random set to become relatively regular, since the random walk starting at the origin tends to reach the random set A_n^c at points with many neighbors in A_n. Figure 6.1 shows a realization of $A(1000)$ in dimension 2.

For every $r \geqslant 0$, write $\|x\|$ for the Euclidean norm of a point $x \in \mathbb{R}^d$, and

$$B(0,r) = \left\{y \in \mathbb{R}^d \,:\, \|y\| < r\right\} \quad \text{et} \quad \mathbb{B}(0,r) = B(0,r) \cap \mathbb{Z}^d.$$

The cardinality of a subset Λ of $\mathbb{Z}^d$ is denoted by $|\Lambda|$. Fix an integer n and consider the random set $A(|\mathbb{B}(0,n)|)$. The (random) radius of the largest ball contained in this set is

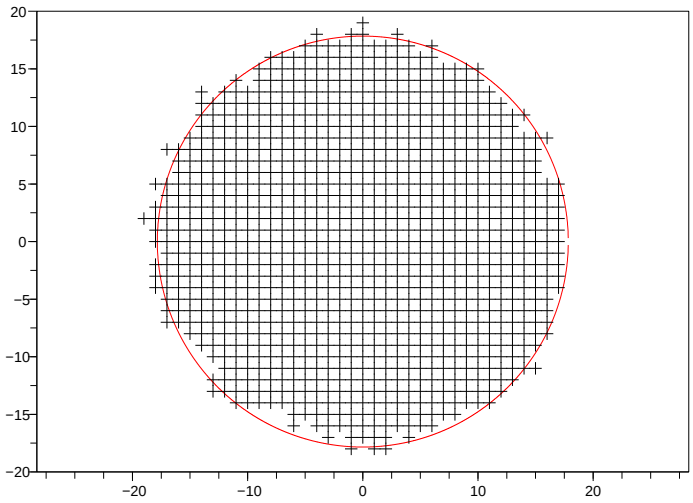


Fig. 6.1. A realization of $A(1000)$ in dimension 2.

$$\sup\{r \geqslant 0 \,:\, \mathbb{B}(0,r) \subset A(|\mathbb{B}(0,n)|)\}$$

while that of the smallest ball containing it is

$$\inf\{r \geqslant 0 \,:\, A(|\mathbb{B}(0,n)|) \subset \mathbb{B}(0,r)\}$$

We then define the inner error $\delta_I(n)$ by

$$n - \delta_I(n) = \sup\{r \geqslant 0 \,:\, \mathbb{B}(0,r) \subset A(|\mathbb{B}(0,n)|)\},$$

and the outer error $\delta_E(n)$ by

$$n + \delta_E(n) = \inf\{r \geqslant 0 \,:\, A(|\mathbb{B}(0,n)|) \subset \mathbb{B}(0,r)\}.$$

The following theorem shows that the asymptotic shape of $A(n)$ is spherical and also provides an upper bound on the extreme fluctuations of $A(n)$.

Theorem 6.11 (Shape and fluctuations)**.** *The limiting shape of* $(A_n)_{n\geqslant 0}$ *is the intersection of* $\mathbb{Z}^d$ *with a Euclidean ball. Moreover, there exists a constant* β_d *such that almost surely*

- *for* $d \geqslant 3$*:*

$$\limsup_{n\to\infty} \frac{\delta_I(n)}{\sqrt{\log(n)}} \leqslant \beta_d \quad et \quad \limsup_{n\to\infty} \frac{\delta_E(n)}{\sqrt{\log(n)}} \leqslant \beta_d,$$

- *for $d = 2$:*

$$\limsup_{n\to\infty} \frac{\delta_I(n)}{\log(n)} \leqslant \beta_2 \quad et \quad \limsup_{n\to\infty} \frac{\delta_E(n)}{\log(n)} \leqslant \beta_2.$$

6.7 Further reading

The internal diffusion limited aggregation[2] model is considered by Persi Diaconis and William Fulton in [DF91]. From a modeling perspective, it can be related to chemical polishing methods: particles are released into a tank and destroy an element of its boundary, making the tank smoother. References in this direction can be found in the articles [AG13a, AG13b] by Amine Asselah and Alexandre Gaudillière.

One of the earliest studies of $(A(n))_{n\geqslant 0}$ is due to Maury Bramson, David Griffeath, and Gregory Lawler in [LBG92]. In particular, they prove that the asymptotic shape of $A(n)$ is spherical. A complete study of the fluctuations around this limiting shape came much later. Theorem 6.11 can be found in the articles by David Jerison, Lionel Levin, and Scott Scheffied [JLS12, JLS13], and in [AG13a, AG13b].

The model in dimension 1 can be related to a ball-and-urn model named after Bernard Friedman and described in the article [Fre65] by David Freedman[3]. A drawn ball is returned to the urn together with α balls of the same color and β balls of the opposite color. If $\beta = 0$, we recover the Pólya urn model studied in Chapter 15. When $\beta > 0$, one can show that the asymptotic proportion of each color tends to $1/2$ for every α.

There is also a more difficult model, known as *external diffusion limited aggregation*, in which particles arrive from infinity and progressively accumulate.

The central limit theorem for martingales 6.8 can be found in the book by Peter Hall and Christopher Heyde [HH80, Cor. 3.1], and in [BC07, Th. 3.35].

[2] Abbreviated IDLA.
[3] *fried-man* $\neq$ *freed-man*!

7

Hidden Markov chains

Keywords. Parametric estimation; recursive estimation; hidden variable; conditional expectation; autoregressive process. **Tools.** Markov chain; Gaussian vector; martingales; maximum likelihood; Bayes' formula. **Difficulty.** *

Many real processes, although Markovian, are only partially observed. This chapter presents two models of *hidden Markov chains*, which are, by definition, projections of Markov chains. The mathematical challenge is to extract as much information as possible about the hidden part from partial observations. This reduces to computing conditional distributions. In general, these distributions cannot be computed explicitly. This is nevertheless possible in the two examples presented here: one uses Markov chains with a finite state space, and the other a Gaussian model.

7.1 The forward-backward algorithm

We model the structure of a DNA[1] strand by a character string over a finite alphabet $\mathcal{A}$. Some portions of the string correspond to sequences of codes for the amino acids needed to produce a protein, and are called coding regions, while others are noncoding. To account for this segmented structure, we model the string by a trajectory of a Markov chain with two regimes.

To a DNA strand of length l, denoted by $(X_1, \ldots, X_l)$, taking values in $\mathcal{A}$, we attach the l-tuple $(U_1, \ldots, U_l)$, taking values in $\mathcal{U} = \{0, 1\}$. If position n is coding (respectively noncoding), then $U_n = 1$ (respectively $U_n = 0$). We model the distribution of $(X, U) = ((X_n, U_n))_{n \geqslant 0}$ by a homogeneous Markov chain on $\mathcal{A} \times \mathcal{U}$ with transition matrix given by

$$P((x, u), (x', u')) = \mathbb{P}(X_{n+1} = x', U_{n+1} = u' \,|\, X_n = x, U_n = u)$$

[1] Deoxyribonucleic acid, the carrier of the genetic code of living organisms.

$$= \rho(u, u')\pi_u(x, x'),$$

where ρ is a transition matrix on $\{0, 1\}$, and π_0 and π_1 are transition matrices on $\mathcal{A}$. We assume that all entries of the matrices ρ, π_0, and π_1 are strictly positive. The component U acts as a kind of two-position (Markovian) switch.

The chain (X, U) is irreducible, recurrent, and aperiodic because all entries of its transition matrix are strictly positive. In general, the projection $X = (X_n)_{n \geqslant 0}$, which models the DNA strand, is not a Markov chain. The model is rigid: the length of a segment in category $u \in \mathcal{U}$ has a geometric distribution with parameter $1 - \rho(u, u)$.

We now assume that the transition matrix of (X, U), that is, the triple ρ, π_0, π_1, is known, for instance from a preliminary estimation procedure (to which we return below). The question is then whether we can determine the distribution of $(U_j)_{1 \leqslant j \leqslant l}$ conditionally on the observations $(X_i)_{1 \leqslant i \leqslant l}$. The Markov chain (X, U) is called hidden because the component U is not observed. Only the component X is observed.

Example 7.1. *Here is the example used in Figure 7.1:*

$$\rho = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}, \quad \pi_0 = \begin{pmatrix} 0.3 & 0.3 & 0.3 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.3 & 0.1 & 0.3 & 0.3 \\ 0.1 & 0.3 & 0.3 & 0.3 \end{pmatrix}, \quad \textit{et} \quad \pi_1 = \begin{pmatrix} 0.5 & 0.3 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.4 & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.1 \end{pmatrix}.$$

Description of the algorithm

We use the vector notation $X_{1:i}$ and $x_{1:i}$ for the i-tuples $(X_1, \ldots, X_i)$ and $(x_1, \ldots, x_i)$. For every $1 \leqslant i \leqslant l$, $x \in \mathcal{A}^l$, and $v \in \mathcal{U}$, define the following probabilities:

- *Prediction probability*:

$$P^i(v) := \mathbb{P}(U_i = v \,|\, X_{1:i-1} = x_{1:i-1}).$$

  This is the probability that the hidden state U_i is in position v given the observations $X_{1:i-1}$;
- *Filtering probability*:

$$F^i(v) := \mathbb{P}(U_i = v \,|\, X_{1:i} = x_{1:i}).$$

  This is the probability that the hidden state U_i is in position v given the observations $X_{1:i}$;
- *Smoothing probability*:

$$L^i(v) := \mathbb{P}(U_i = v \,|\, X_{1:l} = x_{1:l}).$$

  This is the probability that the hidden state U_i is in position v given the observations $X_{1:l}$.

To compute the smoothing probabilities, we use a *forward-backward*[2] algorithm: we recursively determine the prediction and filtering probabilities in increasing index order, then deduce the smoothing probabilities by a backward recursion. The recurrences required by this algorithm are collected in the following theorem.

Theorem 7.2 (Prediction, filtering, smoothing)**.** *For every* $1 \leqslant i \leqslant l$ *and* $v \in \mathcal{U}$*,*

$$\begin{aligned} P^i(v) &= \sum_{u \in \mathcal{U}} \rho(u,v) F^{i-1}(u), \\ F^i(v) &= \frac{\pi_v(x_{i-1}, x_i) P^i(v)}{\sum_{u \in \mathcal{U}} \pi_u(x_{i-1}, x_i) P^i(u)}, \\ L^{i-1}(u) &= F^{i-1}(u) \sum_{v \in \mathcal{U}} \rho(u,v) \frac{L^i(v)}{P^i(v)}. \end{aligned}$$

Proof. For the prediction equation, we introduce U_{i-1} and then use the fact that the distributions $\mathrm{Loi}(U_i \,|\, U_{i-1}, X_{1:i-1})$ and $\mathrm{Loi}(U_i \,|\, U_{i-1})$ coincide:

$$\begin{aligned} P^i(v) &= \sum_{u \in \mathcal{U}} \mathbb{P}(U_{i-1} = u, U_i = v \,|\, X_{i-} = x_{i-}) \\ &= \sum_{u \in \mathcal{U}} \mathbb{P}(U_i = v \,|\, U_{i-1} = u, X_{i-} = x_{i-}) \mathbb{P}(U_{i-1} = u \,|\, X_{i-} = x_{i-}) \\ &= \sum_{u \in \mathcal{U}} \rho(u,v) F^{i-1}(u) \end{aligned}$$

with the abbreviated notation $x_{i-} := x_{1:i-1}$. The basic relation $\mathbb{P}(A \,|\, B \cap C)\mathbb{P}(B \,|\, C) = \mathbb{P}(A \cap B \,|\, C)$ gives the filtering equation:

$$\begin{aligned} F^i(v) &= \frac{\mathbb{P}(U_i = v, X_i = x_i \,|\, X_{1:i-1} = x_{i-})}{\mathbb{P}(X_i = x_i \,|\, X_{i-} = x_{i-})} \\ &= \frac{\mathbb{P}(U_i = v, X_i = x_i \,|\, X_{i-} = x_{i-})}{\sum_{u \in \mathcal{U}} \mathbb{P}(U_i = u, X_i = x_i \,|\, X_{i-} = x_{i-})} \\ &= \frac{\mathbb{P}(X_i = x_i \,|\, U_i = v, X_{i-} = x_{i-}) \mathbb{P}(U_i = v, \,|\, X_{i-} = x_{i-})}{\sum_{u \in \mathcal{U}} \mathbb{P}(X_i = x_i \,|\, U_i = u, X_{i-} = x_{i-}) \mathbb{P}(U_i = u \,|\, X_{i-} = x_{i-})}. \end{aligned}$$

By definition of the transition matrix of (X, U),

$$\begin{aligned} \mathbb{P}(X_i = x_i \,|\, U_i = u, X_{1:i-1} = x_{1:i-1}) &= \mathbb{P}(X_i = x_i \,|\, U_i = u, X_{i-1} = x_{i-1}) \\ &= \pi_{u_{i-1}}(x_i - 1, x_i). \end{aligned}$$

For the smoothing equation, write

[2] The algorithm combines a forward pass with a backward pass.

$$\mathbb{P}(U_{i-1}=u,U_i=v\,|\,X_{1:l}=x_{1:l})=\mathbb{P}(U_{i-1}=u\,|\,U_i=v,X_{1:l}=x_{1:l})\\ \mathbb{P}(U_i=v\,|\,X_{1:l}=x_{1:l}),$$

and then

$$\mathbb{P}(U_{i-1}=u\,|\,U_i=v,X_{1:l}=x_{1:l})\\ =\frac{\mathbb{P}(U_{i-1}=u,X_{1:i-1}=x_{1:i-1},X_{i+1:l}=x_{i+1:l}\,|\,U_i=v,X_i=x_i)}{\mathbb{P}(X_{1:i-1}=x_{1:i-1},X_{i+1:l}=x_{i+1:l}\,|\,U_i=v,X_i=x_i)}.$$

The Markov property ensures that the past and future are conditionally independent given the present. Simplifying therefore gives

$$\mathbb{P}(U_{i-1}=u\,|\,U_i=v,X_{1:l}=x_{1:l})=\mathbb{P}(U_{i-1}=u\,|\,U_i=v,X_{1:i}=x_{1:i}).$$

Again, using the expression for the transition matrix of (X,U), we write

$$\mathbb{P}(U_{i-1}=u\,|\,U_i=v,X_{1:i}=x_{1:i})\\ =\frac{\mathbb{P}(U_{i-1}=u,U_i=v,X_i=x_i,\,|\,X_{1:i-1}=x_{1:i-1})}{\mathbb{P}(U_i=v,X_i=x_i\,|\,X_{1:i-1}=x_{1:i-1})}\\ =\frac{F^{i-1}(u)\rho(u,v)\pi_v(x_{i-1},x_i)}{P^i(v)\pi_v(x_{i-1},x_i)}.$$

We have thus obtained

$$\mathbb{P}(U_{i-1}=u,U_i=v\,|\,X_{1:l}=x_{1:l})=F^{i-1}(u)\rho(u,v)\frac{L^i(v)}{P^i(v)},$$

which yields the smoothing relation upon summing over v. □

We initialize the algorithm by choosing the initial distribution (for example, the stationary distribution) as $P^1(u)$. The prediction and filtering relations of Theorem 7.2 give all the probabilities P^i and F^i by forward recursion. We then deduce the probabilities L^i by backward recursion using the smoothing relation of Theorem 7.2, initialized with $L^l(v)=F^l(v)$. Figure 7.1 illustrates the effectiveness of the algorithm: most of the time, the probability $L^i(v)$ exceeds $1/2$ when $U_i=v$ (in exactly 95.2% of cases in this example). Nevertheless, in this example the algorithm misses a region where U equals 1 (around position 120), and always lags slightly behind regime changes. The algorithm works better when the matrices π_0 and π_1 differ substantially (the two regimes behave very differently) and when $\rho(0,0)$ and $\rho(1,1)$ are large (the intervals between regime changes are long).

Estimating a transition matrix

The preceding study assumes that the transition matrices are known. It is nevertheless important to be able to estimate the transition matrix of a Markov

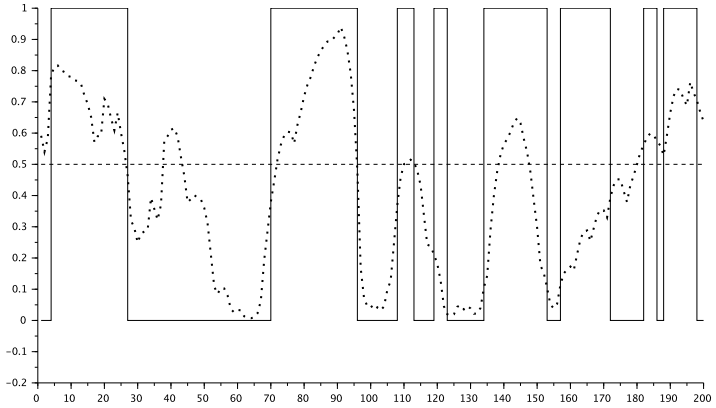


Fig. 7.1. Trajectory of U (solid curve) and smoothing probability for $v = 0$ (dashed curve), with $l = 200$ and the transitions of Example 7.1.

chain from one of its trajectories. Consider a Markov chain $(Z_n)_{n\geqslant 0}$ on the finite state space E with transition matrix $\theta_* = (\theta_*(i,j))_{(i,j)\in E^2}$. Assume that the chain is irreducible and recurrent, and write μ_* for its invariant distribution. We assume that the initial measure ν is known but the transition matrix θ_* is unknown. Write Θ for the set of transition matrices on E. Let λ be counting measure on E. The distribution of $(Z_0, \ldots, Z_n)$ has, with respect to $\lambda^{\otimes n+1}$, density $(z_0, \ldots, z_n) \in E^{n+1} \mapsto f_{\theta_*}(z_0, \ldots, z_n)$ given by

$$f_{\theta_*}(z_0, \ldots, z_n) = \mathbb{P}(Z_0 = z_0, \ldots, Z_n = z_n) = \nu(z_0)\theta_*(z_0, z_1) \cdots \theta_*(z_{n-1}, z_n).$$

For every $\theta \in \Theta$ and $z \in E^{n+1}$, write

$$f_\theta(z_0, \ldots, z_n) := \nu(z_0)\theta(z_0, z_1) \cdots \theta(z_{n-1}, z_n).$$

The random quantity $L_{Z_1,\ldots,Z_n}(\theta) := f_\theta(Z_0, \ldots, Z_n)$ is called the *likelihood*[3] of θ for the sample $(Z_0, \ldots, Z_n)$. The *maximum likelihood estimator* $\widehat{\theta}_n$ of θ_* is, by definition, the random element of Θ that maximizes the random function $\theta \in \Theta \mapsto L_{Z_0,\ldots,Z_n}(\theta)$.

Lemma 7.3 (Maximum likelihood)**.** *The maximum likelihood estimator $\widehat{\theta}_n$ of θ_* is given, for every $i, j \in E$, by*

$$\widehat{\theta}_n(i,j) = \begin{cases} \dfrac{N_n^{ij}}{N_n^i} & \text{si } N_n^i > 0, \\ 0 & \text{sinon,} \end{cases} \quad \text{où} \quad \begin{cases} N_n^i & := \sum_{p=0}^{n-1} \mathbb{1}_{\{Z_p = i\}}, \\ N_n^{ij} & := \sum_{p=0}^{n-1} \mathbb{1}_{\{Z_p = i, Z_{p+1} = j\}}. \end{cases}$$

[3] Hence the notation L.

Note that if $\theta(i,j)=0$ then $\widehat{\theta}_n(i,j)=0$, but the converse is false.

Proof. Number E so that $E=\{1,\dots,s\}$. Write $L:=L_{Z_1,\dots,Z_n}$. It is more convenient to maximize $\log(L)$ than L. We must maximize $\theta\in\Theta\mapsto\log(L(\theta))$ under the constraints $\theta(i,j)\geqslant 0$ and $\sum_{k=1}^{s}\theta(i,k)=1$ for every $1\leqslant i,j\leqslant s$. We have

$$\partial_{\theta(i,j)}(\theta\mapsto\log(L(\theta)))=\frac{N_n^{ij}}{\theta(i,j)},$$

and the s extremality conditions amount to saying that, for every $1\leqslant i\leqslant s$, the function $j\mapsto N_n^{ij}/\theta(i,j)$ must be constant. Taking $\theta\in\Theta$ into account gives the result. □

The estimator $\hat{\theta}$ is intuitive: it estimates $\theta(i,j)$ by the proportion of transitions from state i to state j among the n observed jumps. Its asymptotic properties follow from the behavior of the sequences $(N_n^{ij})_n$ and $(N_n^i)_n$ established below.

Theorem 7.4 (Consistency and asymptotic normality)**.** *For every $x\in E$, conditionally on $\{X_0=x\}$, the estimator $\widehat{\theta}_n$ is consistent and asymptotically normal: for every $i,j\in E$,*

$$\widehat{\theta}_n(i,j)\xrightarrow[n\to\infty]{\text{p.s.}}\theta(i,j)$$

and

$$\sqrt{n\mu(i)}(\widehat{\theta}_n(i,j)-\theta(i,j))\xrightarrow[n\to\infty]{\text{loi}}\mathcal{N}(0,\theta(i,j)(1-\theta(i,j))).$$

Proof. For every $x,i,j\in E$, conditionally on $\{X_0=x\}$, we have

$$\frac{1}{n}N_n^i\xrightarrow[n\to\infty]{\text{p.s.}}\mu(i)\quad\text{et}\quad\frac{1}{n}N_n^{ij}\xrightarrow[n\to\infty]{\text{p.s.}}\mu(i)\theta(i,j)$$

whereas

$$\frac{1}{\sqrt{n}}\big(N_n^{ij}-N_n^i\theta(i,j)\big)\xrightarrow[n\to\infty]{\text{loi}}\mathcal{N}(0,\mu(i)\theta(i,j)(1-\theta(i,j))).$$

These almost sure convergences follow from the law of large numbers for the Markov chains Z and Y, where Y is the irreducible recurrent Markov chain with state space E^2 defined by $Y_n=(Z_n,Z_{n+1})$, whose invariant distribution is $(i,j)\mapsto\mu(i)\theta(i,j)$. Convergence in distribution follows from the central limit theorem for Markov chains. □

7.2 The Kalman filter

An aircraft travels between Paris and London, attempting to follow a nominal trajectory defined by its flight plan. Ground-based air traffic controllers monitor the aircraft using a radar that receives an echo at regular intervals. The

aircraft's actual trajectory deviates from the nominal one for many reasons (weather, autopilot inaccuracies, turbulence,...). We therefore seek to locate the aircraft throughout its flight from successive radar observations.

Write X_n for the (unknown) deviation between the nominal trajectory and the aircraft's position at time n. Also write Y_n for the radar measurement at time n. This measurement is subject to errors due to radar inaccuracies. The controller's problem is to estimate the aircraft's position at time n as accurately as possible from the observations $Y_0, \dots, Y_n$. To simplify the study, we assume that the observed object moves in a space of dimension 1. We model the random sequences $(X_n)_{n\geqslant 0}$ and $(Y_n)_{n\geqslant 0}$ by setting

$$\begin{cases} X_0 = W_0, \\ X_n = aX_{n-1} + W_n & n \geqslant 1, \\ Y_n = X_n + V_n & n \geqslant 0, \end{cases}$$

where a is a deterministic real number, and $(V_n)_{n\geqslant 0}$ and $(W_n)_{n\geqslant 0}$ are independent random sequences, with $(V_n)_{n\geqslant 0}$ i.i.d. Gaussian with distribution $\mathcal{N}(0,\tau^2)$ and $(W_n)_{n\geqslant 0}$ i.i.d. Gaussian with distribution $\mathcal{N}(0,\sigma^2)$. The random variables $(W_n)_{n\geqslant 0}$ represent the instantaneous fluctuations in the deviation between nominal and actual positions. The random variables $(V_n)_{n\geqslant 0}$ model radar measurement errors. The parameter a models the pilot's actions.

In signal theory, $(X_n)_{n\geqslant 0}$ is called an autoregressive process of order 1, denoted AR(1), with Gaussian noise, and we have

$$X_n = \sum_{k=0}^{n} a^k W_{n-k} = \sum_{k=0}^{n} a^{n-k} W_k, \quad n \in \mathbb{N}.$$

The problem is now to estimate X_n given the observations $Y_0, \dots, Y_n$. The natural candidate is the conditional expectation $\mathbb{E}(X_n \,|\, Y_0, \dots, Y_n)$, since it is the function of the observations closest to X_n in L^2 (least squares). The Gaussian nature of the model and the properties of Gaussian vectors allow us to compute the conditional distribution $\mathrm{Loi}(X_n \,|\, Y_0, \dots, Y_n)$ explicitly, and in particular its mean $\mathbb{E}(X_n \,|\, Y_0, \dots, Y_n)$.

The first key observation is that, for every $n \geqslant 1$,

$$(X_0, \dots, X_n, Y_0, \dots, Y_n)$$

is a Gaussian random vector, since every linear combination of its coordinates is a linear combination of the independent Gaussian random variables $(W_i)_{0\leqslant i\leqslant n}$ and $(V_i)_{0\leqslant i\leqslant n}$. In particular, for every $n \geqslant 0$, $(X_n, Y_0, \dots, Y_n)$ is also a Gaussian vector. Moreover, the distribution of X_n given $(Y_0, \dots, Y_n)$ is Gaussian, with mean denoted by $\hat{X}_n$ and variance by P_n. We aim to compute these quantities recursively. As for the finite-state Markov chains in the first part of this chapter, the recursion has two steps. The prediction step expresses the distribution $\mathrm{Loi}(X_n \,|\, Y_0, \dots, Y_{n-1})$ in terms of

$\mathrm{Loi}(X_{n-1} \,|\, Y_0, \dots, Y_{n-1})$. Then the filtering step incorporates the observation Y_n to express $\mathrm{Loi}(X_n \,|\, Y_0, \dots, Y_n)$ in terms of $\mathrm{Loi}(X_n \,|\, Y_0, \dots, Y_{n-1})$.

Lemma 7.5 (Conditional distributions)**.** *We have*

$$\mathrm{Loi}(X_n \,|\, Y_0, \dots, Y_n) = \mathcal{N}(\hat{X}_n, P_n)$$

where

$$\hat{X}_n = a\hat{X}_{n-1} + \frac{P_n}{\tau^2}(Y_n - a\hat{X}_{n-1}) \quad \textit{et} \quad P_n = \frac{a^2\tau^2 P_{n-1} + \sigma^2\tau^2}{a^2 P_{n-1} + \sigma^2 + \tau^2}.$$

Proof. First recall the following result on Gaussian vectors, which we call Bayes' formula: if $(X, Y_0, \dots, Y_{n-1})$ is a Gaussian vector in $\mathbb{R}^{n+1}$ with

$$\mathrm{Loi}(X \,|\, Y_0, \dots, Y_{n-1}) = \mathcal{N}(\mu, \gamma^2) \quad \text{et} \quad \mathrm{Loi}(Y \,|\, Y_0, \dots, Y_{n-1}, X) = \mathcal{N}(X, \delta^2)$$

then

$$\mathrm{Loi}(X \,|\, Y_0, \dots, Y_{n-1}, Y) = \mathcal{N}\left(\rho^2\left(\frac{\mu}{\gamma^2} + \frac{Y}{\delta^2}\right), \rho^2\right) \quad \text{où} \quad \frac{1}{\rho^2} := \frac{1}{\gamma^2} + \frac{1}{\delta^2}.$$

As announced, we now proceed recursively in two steps.

- *Initialization.* Since $Y_0 = X_0 + V_0$, we have $\mathrm{Loi}(Y_0 \,|\, X_0) = \mathcal{N}(X_0, \tau^2)$. Bayes' formula ensures that $\mathrm{Loi}(X_0 \,|\, Y_0) = \mathcal{N}(\hat{X}_0, P_0)$, where
$$\hat{X}_0 = \frac{\sigma^2}{\sigma^2 + \tau^2} Y_0 \quad \text{et} \quad P_0 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$
- *Prediction.* We have $\mathrm{Loi}(X_{n-1} \,|\, Y_0, \dots, Y_{n-1}) = \mathcal{N}(\hat{X}_{n-1}, P_{n-1})$ and, by the model,
$$\mathrm{Loi}(X_n \,|\, Y_0, \dots, Y_{n-1}) = \mathcal{N}(a\hat{X}_{n-1}, a^2 P_{n-1} + \sigma^2).$$
- *Filtering.* Again, the model gives
$$\mathrm{Loi}(Y_n \,|\, Y_0, \dots, Y_{n-1}, X_n) = \mathcal{N}(X_n, \tau^2).$$
We then apply Bayes' formula to reverse the conditioning between Y_n and X_n: the distribution is $\mathrm{Loi}(X_n \,|\, Y_0, \dots, Y_n) = \mathcal{N}(\hat{X}_n, P_n)$ with
$$\frac{1}{P_n} = \frac{1}{a^2 P_{n-1} + \sigma^2} + \frac{1}{\tau^2} \quad \text{et} \quad \hat{X}_n = P_n\left(\frac{a\hat{X}_{n-1}}{a^2 P_{n-1} + \sigma^2} + \frac{Y_n}{\tau^2}\right).$$

□

Remark 7.6 (Gain from observations)**.** *The upper bound*

$$P_n = \mathbb{E}\Big[(X_n - \hat{X}_n)^2\Big] \leqslant \mathbb{E}\big[(X_n - Y_n)^2\big] = \tau^2,$$

is unsurprising in view of the properties of conditional expectation, and emphasizes the benefit of using all the observations $Y_0, \dots, Y_n$ instead of just the latest one Y_n.

Estimating some model parameters

In this section, we assume that the aircraft's positions are observed without error, so we have access to the sequence $(X_i)_{1\leqslant i\leqslant n}$. We wish to estimate the coefficients a and σ^2. This is not entirely straightforward, since the observations are not independent.

Lemma 7.7 (Maximum likelihood estimator)**.** *The maximum likelihood estimator $(\hat{a},\hat{\sigma})$ of (a,σ) is given by*

$$\hat{a}_n = \frac{\sum_{k=1}^n X_{k-1}X_k}{\sum_{k=1}^n X_{k-1}^2} \quad et \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^n (X_k - \hat{a}_n X_{k-1})^2.$$

Proof. It suffices to maximize the logarithm of the likelihood given by

$$L_{(X_1,\dots,X_n)}(a,\sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}}\prod_{k=1}^n \exp\left(-\frac{(X_k - aX_{k-1})^2}{2\sigma^2}\right).$$

□

Theorem 7.8 (Consistency and normality)**.** *If $-1 < a < 1$ then*

$$(\hat{a}_n,\hat{\sigma}_n) \xrightarrow[n\to\infty]{\text{p.s.}} (a,\sigma),$$

and

$$\sqrt{n}(\hat{a}-a) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1-a^2), \quad et \quad \sqrt{n}\big(\hat{\sigma}^2-\sigma^2\big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,2\sigma^4).$$

Proof. We only prove the results for $\hat{a}$, assuming σ known and fixed. Rewrite $\hat{a}_n$ as follows:

$$\hat{a}_n = a + \frac{\sum_{k=1}^n X_{k-1}W_k}{\sum_{k=1}^n X_{k-1}^2} = a + \frac{M_n}{\sum_{k=1}^n X_{k-1}}$$

where $M_n = \sum_{k=1}^n X_{k-1}W_k$, with $M_0 = 0$. The sequence $(M_n)_{n\geqslant 0}$ is a martingale with respect to the natural filtration $(\mathcal{F}_n)_n$ of W, with increasing process

$$\langle M\rangle_0 = 0 \quad \text{et, pour } n \geqslant 1, \quad \langle M\rangle_n = \sigma^2\sum_{k=1}^n X_{k-1}^2.$$

First, using the fact that X is an autoregressive process of order 1, AR(1), with $|a|<1$, we check that

$$\frac{\langle M\rangle_n}{n} \xrightarrow[n\to\infty]{\text{p.s.}} \frac{\sigma^4}{(1-a^2)},$$

(in particular, $\langle M\rangle_n \to +\infty$ a.s. as $n\to\infty$). We then use the law of large numbers and the central limit theorem for square-integrable martingales, which give

$$\frac{M_n}{\langle M\rangle_n} \xrightarrow[n\to\infty]{\text{p.s.}} 0 \quad \text{et} \quad \frac{M_n}{\sqrt{\langle M\rangle_n}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

□

7.3 Further reading

The forward-backward *segmentation algorithm* dates back at least to the work of Leonard Baum and Lloyd Welch in the 1960s, and can be viewed as an instance of the general concept of dynamic programming. Hidden Markov chains were used notably for speech recognition in the 1970s and for genomics from the 1980s onward. The book by Stéphane Robin, François Rodolphe, and Sophie Schbath [RRS05] offers an accessible introduction to their use in genomics. See also Étienne Pardoux's book [Par07]. Although the models used in practice are more sophisticated than the one presented here, especially concerning the state spaces $\mathcal{A}$ and $\mathcal{U}$, they rely on the same concepts and tools. In practice, however, neither the transition matrices nor even the number of hidden states needed to describe the sequence distribution are generally known. Algorithms must therefore also estimate these complexity parameters. The book by Jean-François Delmas and Benjamin Jourdain [DJ06] also discusses this question. Moreover, one can test whether a random sequence is Markovian by using, for example, N^i and N^{ij} and the χ^2 test, as explained in the book by Didier Dacunha-Castelle and Marie Duflo [DCD83].

The section on the Kalman filter is inspired by David Williams' book [Wil91]. Presented here in a simple form, the Kalman filter is a classic of signal theory, developed in the 1960s by Thorvald Thiele and Peter Swerling, and by Rudolf Kalman and Richard Bucy, notably for the National Aeronautics and Space Administration's Apollo program. An expression for the Kalman filter in dimensions greater than 1 can be found in [Par07]. The law of large numbers and central limit theorem for martingales used in the proof of Theorem 7.8 are taken from [BC07]. Maximum likelihood estimation of the parameters a, σ, τ of the Gaussian model from Y without knowing X is discussed, for example, in the book by Peter Brockwell and Richard Davis [BD02, Sec. 8.5]. Finally, [KDS+15] puts this in perspective in relation to state-space models.

8

The EM algorithm and mixtures

Keywords. Mixture distributions; parametric estimation. **Tools.** Bayes' formula; maximum likelihood; relative entropy; Gaussian distribution. **Difficulty.** *

The Expectation-Maximization (EM) algorithm is one of the most important algorithms in statistics. It provides a numerical approximation to the maximum likelihood estimator for partially observed models. Many variants are available. In this chapter, we illustrate the basic workings of the EM algorithm through a simple example.

Four different species of gulls coexist on an island. Ornithologists wish to estimate the proportion of gulls belonging to each species by observing nest sizes. Unlike birds, nests do not move, which makes counting easier. Unfortunately, the different species build rather similar nests: we do not know which species built them. We assume, however, that the distribution of nests is characteristic of each species. The overall distribution of nest sizes is a mixture of four probability distributions, each describing the distribution of nest sizes for one bird species.

8.1 A model for gull nests

Let us model the distribution of the sizes of nests built by species j by a distribution $\mu_j = \mathcal{N}(m(j), v(j))$ with density $\gamma_{m(j),v(j)}$. The distribution μ of gull nest sizes therefore has density

$$x \mapsto \sum_{j=1}^{J} \alpha(j) \gamma_{m(j),v(j)}(x) = \sum_{j=1}^{J} \frac{\alpha(j)}{\sqrt{2\pi v(j)}} \exp\left(-\frac{(x-m(j))^2}{2v(j)}\right),$$

with $\alpha_1 > 0, \ldots, \alpha_J > 0$ and $\alpha_1 + \cdots + \alpha_J = 1$. This finite convex combination of probability densities is called a finite mixture[1] of populations. If X denotes the nest size and Z the species of bird that built it, the distribution of the pair (X, Z) is given as follows:

- the variable Z has the discrete distribution on $\{1, \ldots, J\}$ with weights $\alpha(1), \ldots, \alpha(J)$:
$$\mathrm{Loi}(Z) = \sum_{j=1}^{J} \alpha(j)\delta_j$$
- for every $1 \leqslant j \leqslant J$, conditionally on $\{Z = j\}$, the variable X has distribution $\mathcal{N}(m(j), v(j))$:
$$\mathrm{Loi}(X|Z = j) = \mathcal{N}(m(j), v(j)).$$

The parameters $(\alpha_j, m(j), v(j))_{1 \leqslant j \leqslant J}$ are unknown. The vector α contains the mixture weights, while the vectors m and v contain the means and variances of the Gaussian distributions, respectively. We wish to estimate all these quantities from the sizes of n nests alone. To this end, we assume that there exists θ in the set

$$\Theta = \left\{ \theta = (\alpha_j, m_j, \sigma_j^2)_{1 \leqslant j \leqslant J} : \alpha_1 > 0, \ldots, \alpha_J > 0, \alpha_1 + \cdots + \alpha_J = 1 \right\}$$

such that the measured sizes $x_1, \ldots, x_n$ of n nests form a realization of an n-sample $X_1, \ldots, X_n$ whose density at x, for the distribution of X, with respect to Lebesgue measure $d\lambda$ on $\mathbb{R}$ is

$$f_\theta(x) = \sum_{j=1}^{J} \alpha(j)\gamma_{m(j),v(j)}(x).$$

We write $\overline{X}$ for the vector $(X_1, \ldots, X_n)$ and $\overline{Z}$ for the vector $(Z_1, \ldots, Z_n)$. We write dN for counting measure on $\mathbb{N}$. We introduce the following notation:

- density at (x, z) of the distribution of the pair (X, Z) with respect to $d\lambda \otimes dN$:
$$h_\theta(x, z) = \alpha(z)\gamma_{m(z),v(z)}(x), \quad x \in \mathbb{R},\ z \in \{1, \ldots, J\},$$
- density at x of the distribution of X with respect to $d\lambda$:
$$f_\theta(x) = \sum_{j=1}^{J} \alpha(j)\gamma_{m(j),v(j)}(x), \quad x \in \mathbb{R},$$
- density at z of the distribution of Z with respect to dN:
$$g_\theta(z) = \alpha(z), \quad z \in \{1, \ldots, J\},$$

[1] A *finite mixture* of populations.

- density at x of the distribution of X given Z with respect to $d\lambda$:

$$f_\theta(x|Z=j) = \gamma_{m(j),v(j)}(x), \quad x \in \mathbb{R},$$

- density at z of the distribution of Z given X with respect to dN:

$$g_\theta(z|X=x) = \frac{\alpha(z)\gamma_{m(z),v(z)}(x)}{\sum_{j=1}^J \alpha(j)\gamma_{m(j),v(j)}(x)}, \quad z \in \{1,\dots,J\}.$$

The last expression follows from the definition of conditional probability: indeed, for every Borel set A and every $1 \leqslant z \leqslant J$,

$$\mathbb{P}(X \in A,\ Z=z) = \mathbb{P}(Z=z)\mathbb{P}(X \in A|Z=z) = \alpha(z)\int_A \gamma_{m(z),v(z)}(x)\,dx$$

and

$$\begin{aligned}\mathbb{P}(X \in A,\ Z=z) &= \mathbb{P}(Z=z|X \in A)\mathbb{P}(X \in A)\\ &= \mathbb{P}(Z=z|X \in A)\sum_{j=1}^J \alpha(j)\int_A \gamma_{m(j),v(j)}(x)\,dx,\end{aligned}$$

so that, if A has the form $[x-\varepsilon, x+\varepsilon]$ with ε tending to 0, we obtain

$$\alpha(z)\gamma_{m(z),v(z)}(x) = \mathbb{P}(Z=z|X=x)\sum_{j=1}^J \alpha(j)\gamma_{m(j),v(j)}(x).$$

8.2 A favorable but unrealistic situation

In this section, we assume that both X and Z are observed, that is, both the nest size and the species that built the nest. The unknown parameters are then easily estimated by maximum likelihood. The log-likelihood of the complete model (the logarithm of the density of the sample $(\overline{X}, \overline{Z})$ with respect to the measure $(d\lambda \otimes dN)^{\otimes n}$) is

$$\begin{aligned}L(\theta, \overline{X}, \overline{Z}) &= \log \prod_{i=1}^n h_\theta(X_i, Z_i)\\ &= \sum_{i=1}^n \big[\log \alpha(Z_i) + \log \gamma_{m(Z_i),v(Z_i)}(X_i)\big].\end{aligned}$$

For a sample of size n, write, for every $1 \leqslant j \leqslant J$,

$$A_j = \{i = 1,\dots,n,\ Z_i = j\} \quad \text{et} \quad C_j = \operatorname{card}(A_j).$$

Grouping terms according to the possible values of $\overline{Z}$ gives

$$L(\theta, \overline{X}, \overline{Z}) = \sum_{j=1}^J C_j \log \alpha(j) + \sum_{j=1}^J \sum_{i \in A_j} \log \gamma_{m(j),v(j)}(X_i).$$

Theorem 8.1 (Estimation for the complete model)**.** *The complete likelihood attains its maximum at a unique point $\hat{\theta} \in \Theta$ given by:*

$$\hat{\alpha}(j) = \frac{C_j}{n}, \quad \hat{m}(j) = \frac{1}{C_j} \sum_{i \in A_j} X_i \quad et \quad \hat{v}(j) = \frac{1}{C_j} \sum_{i \in A_j} (X_i - \hat{m}(j))^2.$$

Proof. Under the constraint $\alpha(1)+\cdots+\alpha(J) = 1$, the function $\theta \mapsto L(\theta, \overline{X}, \overline{Z})$ has a unique critical point, which is a maximum. □

This result does not answer the original question, since in practice only the nest sizes $x_1, \ldots, x_n$ are observed. Instead, we should study the log-likelihood of the sample $X_1, \ldots, X_n$, denoted by L_{obs} for the log-likelihood of the observations, which is given by:

$$L_{\text{obs}}(\theta, \overline{X}) = \log \prod_{i=1}^{n} f_\theta(X_i) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{J} \alpha(j) \gamma_{m(j),v(j)}(X_i) \right].$$

Finding a set of parameters θ that maximizes this quantity is not easy.

8.3 Solving the actual problem

Since the complete log-likelihood L is unavailable, we replace it by its conditional expectation given the observations: we define the conditional log-likelihood of the observations under the distribution with parameter $\tilde{\theta}$, denoted by $L_c(\theta, \tilde{\theta}, \overline{X})$, as:

$$L_c(\theta, \tilde{\theta}, \overline{X}) = \mathbb{E}_{\tilde{\theta}}(L(\theta, \overline{X}, \overline{Z}) | \overline{X}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \log h_\theta(X_i, j) g_{\tilde{\theta}}(j | X = X_i).$$

Here is the expression for the conditional likelihood L_c in our Gaussian mixture setting.

Theorem 8.2 (Conditional log-likelihood)**.** *The function L_c has the following form:*

$$\begin{aligned} L_c(\theta, \tilde{\theta}, \overline{X}) = &- \frac{n}{2} \log(2\pi) + \sum_{j=1}^{J} \left(\sum_{i=1}^{n} g_{\tilde{\theta}}(j | X = X_i) \right) \log \alpha(j) \\ &- \frac{1}{2} \sum_{j=1}^{J} \left(\sum_{i=1}^{n} g_{\tilde{\theta}}(j | X = X_i) \right) \left(\log v(j) + \frac{(X_i - m(j))^2}{v(j)} \right). \end{aligned}$$

The maximum of the conditional likelihood is easy to determine.

Theorem 8.3 (Maximum of the conditional log-likelihood)**.** *The function* $\theta \mapsto L_c(\theta, \tilde{\theta}, \overline{X})$ *has a unique maximum* θ_M *given by*

$$\begin{aligned}
\alpha_M(j) &= \frac{1}{n}\sum_{i=1}^{n} g_{\tilde{\theta}}(j|X = X_i) \\
m_M(j) &= \frac{\sum_{i=1}^{n} X_i g_{\tilde{\theta}}(j|X = X_i)}{\sum_{i=1}^{n} g_{\tilde{\theta}}(j|X = X_i)} \\
v_M(j) &= \frac{\sum_{i=1}^{n} (X_i - m_M(j))^2 g_{\tilde{\theta}}(j|X = X_i)}{\sum_{i=1}^{n} g_{\tilde{\theta}}(j|X = X_i)}.
\end{aligned}$$

Proof. Under the constraint $\alpha(1) + \cdots + \alpha(J) = 1$, we obtain

$$\alpha_M(j) = \frac{\sum_{i=1}^{n} g_{\tilde{\theta}}(j|X = X_i)}{\sum_{l=1}^{J}\sum_{i=1}^{n} g_{\tilde{\theta}}(l|X = X_i)} = \frac{1}{n}\sum_{i=1}^{n} g_{\tilde{\theta}}(j|X = X_i).$$

Moreover, (m_M, v_M) is a critical point of the last term in the conditional log-likelihood. □

The EM algorithm repeatedly performs the following two consecutive steps:

- E(xpectation) step: given a parameter value θ_k, compute the conditional log-likelihood of the observations $L_c(\theta, \theta_k, \overline{X})$ using Theorem 8.2,
- M(aximization) step: using Theorem 8.3, choose θ_{k+1} so that the function $\theta \mapsto L_c(\theta, \theta_k, \overline{X})$ is maximal at θ_{k+1}.

In practice, given the observations $x_1, \ldots, x_n$ and initial parameter values, the algorithm repeats the following calculations:

- from the parameter θ_k, compute the following matrix $H^{(k)}$ of size $n \times J$:

$$H_{ij}^{(k)} = \frac{\alpha_k(j)\gamma_{m_k(j),v_k(j)}(X_i)}{\sum_{l=1}^{J} \alpha_k(l)\gamma_{m_k(l),v_k(l)}(X_i)},$$

- deduce θ_{k+1} from the following relations:

$$\alpha_{k+1}(j) = \frac{1}{n}\sum_{i=1}^{n} H_{ij}^{(k)},$$

$$m_{k+1}(j) = \frac{\sum_{i=1}^{n} X_i H_{ij}^{(k)}}{\sum_{i=1}^{n} H_{ij}^{(k)}},$$

$$v_{k+1}(j) = \frac{\sum_{i=1}^{n} (X_i - m_{k+1}(j))^2 H_{ij}^{(k)}}{\sum_{i=1}^{n} H_{ij}^{(k)}}.$$

Theorem 8.3 is useful only if we can estimate $\tilde{\theta}$. In the EM algorithm, this is done numerically. The following result shows that the log-likelihood L_{obs} increases along the algorithm.

Theorem 8.4 (Increasing likelihood)**.** *The sequence $(\theta_k)_{k\geqslant 0}$ constructed by the EM algorithm satisfies the following numerical stability property:*

$$L_{\text{obs}}(\theta_{k+1}, \overline{X}) \geqslant L_{\text{obs}}(\theta_k, \overline{X}),$$

where $L_{\text{obs}}(\theta, \overline{X})$ is the log-likelihood of the observations.

Proof. Set

$$H(\theta, \theta_k) = \mathbb{E}_{\theta_k}(\log g_\theta(\overline{Z} \,|\, \overline{X})) = \sum_{i=1}^{n} \mathbb{E}_{\theta_k}(\log g_\theta(Z_i \,|\, X_i)).$$

Since $h_\theta(x, z) = f_\theta(x) g_\theta(z|X = x)$, the conditional likelihood can be written as

$$L_c(\theta, \theta_k, \overline{X}) = L_{\text{obs}}(\theta, \overline{X}) + H(\theta, \theta_k).$$

During the M step, the function $\theta \mapsto L_c(\theta, \theta_k, \overline{X})$ is maximized over θ, hence

$$L_c(\theta_{k+1}, \theta_k, \overline{X}) \geqslant L_c(\theta_k, \theta_k, \overline{X}).$$

It follows that

$$L_{\text{obs}}(\theta_{k+1}, \overline{X}) - L_{\text{obs}}(\theta_k, \overline{X}) \geqslant H(\theta_k, \theta_k) - H(\theta_{k+1}, \theta_k).$$

By definition of H,

$$H(\theta_k, \theta_k) - H(\theta_{k+1}, \theta_k) = \mathbb{E}_{\theta_k} \log \frac{g_{\theta_k}(\overline{Z} \,|\, \overline{X})}{g_{\theta_{k+1}}(\overline{Z} \,|\, \overline{X})}.$$

Writing μ_θ for the distribution of $\overline{Z}$ given $\overline{X}$ under $\mathbb{P}_\theta$, we see that the quantity above is precisely the relative entropy of μ_{θ_k} with respect to $\mu_{\theta_{k+1}}$:

$$H(\theta_k, \theta_k) - H(\theta_{k+1}, \theta_k) = \int \log \frac{d\mu_{\theta_k}}{d\mu_{\theta_{k+1}}} d\mu_{\theta_k} = \int \frac{d\mu_{\theta_k}}{d\mu_{\theta_{k+1}}} \log \frac{d\mu_{\theta_k}}{d\mu_{\theta_{k+1}}} d\mu_{\theta_{k+1}}.$$

Jensen's inequality ensures that this quantity is nonnegative. □

Theorem 8.4 shows that the sequence $(L_{\text{obs}}(\theta_k, \overline{X}))_{k\geqslant 0}$ is nondecreasing. It therefore converges to a critical point (a local maximum or saddle point) of the likelihood $L_{\text{obs}}(\cdot, \overline{X})$. Unlike the case where both X and Z are observed, local maxima may trap the algorithm, making its outcome sensitive to the starting point. The behavior of the algorithm can be improved by introducing some randomness, as in the simulated annealing algorithm of Chapter 4, although this offers only limited hope. Moreover, we do not know how many iterations are required to approach a local maximum sufficiently closely. There is no effective stopping criterion.

Figure 8.1 shows the estimation of a mixture of three Gaussian distributions by the EM algorithm with 1000 iterations and parameters

$$\alpha = (0.4\ 0.4\ 0.2), \quad m = (-2\ 0\ 2), \quad v = (0.3\ 0.2\ 0.2), \quad n = 1000.$$

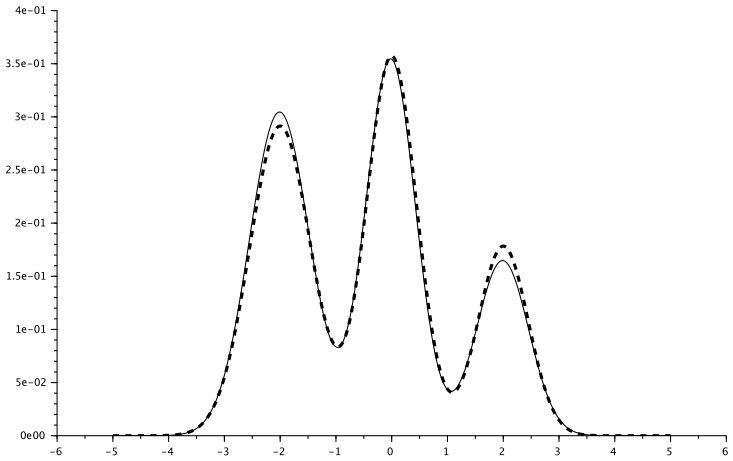


Fig. 8.1. Estimation of a mixture of three Gaussian distributions by the EM algorithm.

8.4 Further reading

Theorem 8.3 extends to exponential families, with complete likelihood

$$L(\theta, \overline{x}, \overline{z}) = \sum_{i=1}^{n} \log h_\theta(x_i, z_i) = \sum_{i=1}^{n} \log g_\theta(z_i) + \sum_{i=1}^{n} \log f_\theta(x_i \,|\, Z = z_i).$$

Integrating and recalling that $g_\theta(z) = \alpha(z)$ gives

$$L_c(\theta, \tilde{\theta}, \overline{x}) = \sum_{j=1}^{J} \left(\sum_{i=1}^{n} g_{\tilde{\theta}}(j \,|\, X = x_i) \right) \log \alpha(j) + \sum_{j=1}^{J} \sum_{i=1}^{n} g_{\tilde{\theta}}(j \,|\, X = x_i) \log f_\theta(x_i \,|\, Z = j).$$

If $\theta = (\alpha, \beta)$, where α again gives the mixture weights and $\beta(j)$ represents the parameters of the distribution $\mathrm{Loi}(X \,|\, Z = j)$, we determine α_M (whose expression remains unchanged) and β_M separately. The latter is obtained almost exactly as for the maximum of the complete likelihood: the contribution of observation X_i is simply weighted by $g_{\tilde{\theta}}(j \,|\, X = X_i)$. For example, if the mixture components $(\mathrm{Loi}(X \,|\, Z = j))_{1 \leqslant j \leqslant J}$ are exponential distributions with parameters $(\lambda(i))_{1 \leqslant i \leqslant J}$, the conditional log-likelihood is

$$L_c(\theta, \tilde{\theta}, \overline{X}) = \sum_{j=1}^{J} \left(\sum_{i=1}^{n} g_{\tilde{\theta}}(j \mid X = X_i) \right) \log \alpha(j) + \sum_{j=1}^{J} \sum_{i=1}^{n} g_{\tilde{\theta}}(j \mid X = X_i)(\log \lambda(j) - \lambda(j) x_i).$$

One step of the EM algorithm takes the following form:

$$H_{ij}^{(k)} = \frac{\alpha_k(j)\lambda_k(j)e^{-\lambda_k(j)x_i}}{\sum_{l=1}^{J} \alpha_k(l)\lambda_k(l)e^{-\lambda_k(l)x_i}},$$

$$\alpha_{k+1}(j) = \frac{1}{n} \sum_{i=1}^{n} H_{ij}^{(k)},$$

$$\lambda_{k+1}(j) = \frac{\sum_{i=1}^{n} H_{ij}^{(k)}}{\sum_{i=1}^{n} X_i H_{ij}^{(k)}}.$$

Although considerable work on this question already existed, it was the article [DLR77] by Arthur Dempster, Nan Laird, and Donald Rubin that first defined the EM algorithm in a general framework. References to earlier results can be found in its bibliography. Since then, the algorithm has been widely used in various settings and forms. To reduce dependence on initial conditions, many randomized versions of the EM algorithm have been developed, inspired in particular by the Kiefer–Wolfowitz stochastic approximation algorithm for maximizing a function expressed as an expectation. That algorithm was itself inspired by the Robbins–Monro stochastic approximation algorithm for finding a zero of a function expressed as an expectation. See, for example, the article [DLM99] by Bernard Delyon, Marc Lavielle, and Éric Moulines, and the book [Duf97] by Marie Duflo. When the conditional log-likelihood cannot be computed, a Monte Carlo method such as the Metropolis–Hastings algorithm of Chapter 5 can be used to approximate this function, whose maximum is then sought numerically. This approach is studied, for example, by Estelle Kuhn and Marc Lavielle in [KL04].

Christophe Biernacki and Stéphane Chrétien study the degeneracy of the likelihood of finite Gaussian mixtures and its connections with the EM algorithm in [BC03].

9

Ehrenfest urns

Keywords. Urn; invariant measure; reversible measure; return time; convergence to equilibrium; cutoff. **Tools.** Markov chain; continuous-time Markov chain; coupling; total variation distance; spectral decomposition; coupon collector; Chernoff inequality; Chebyshev inequality; central limit theorem. **Difficulty.** **

The second law of thermodynamics states that, for every isolated system, an extensive macroscopic quantity called *entropy* increases over time. In Boltzmann's kinetic theory of gases, the second law follows from a statistical approach to classical mechanics applied to microscopic constituents. Although appealing, this approach was criticized by opponents of atomism. How can one resolve the contradiction between the *microscopic reversibility* inherited from classical mechanics and the *macroscopic irreversibility* of thermodynamics? The *Ehrenfest model* offers an explanation of this paradox in an extremely simple setting. The key tool here is finite Markov chains.

9.1 The microscopic model

Consider two containers A and B connected by a small hole and containing a gas made up of a molecules (Figure 9.1). At each time, one of the a molecules, chosen at random, passes through the hole from one container to the other. We wish to describe the evolution of the system. We model these containers as two urns A and B containing a balls numbered from 1 to a. To a configuration, that is, a distribution of the a balls between the urns, we associate an element $x = (x(1), \ldots, x(a))$ of the discrete cube

$$F := \{0, 1\}^a$$

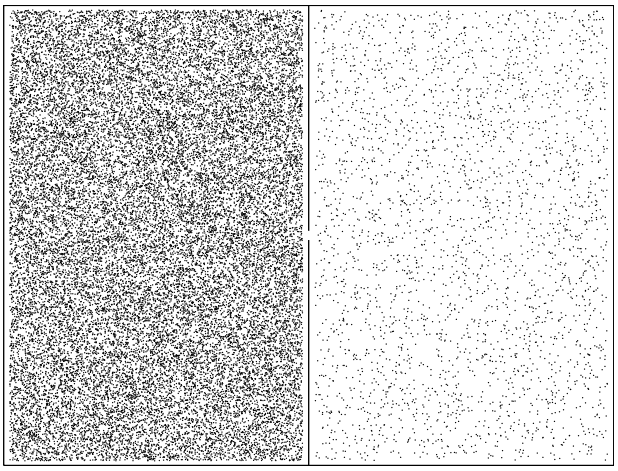


Fig. 9.1. Distribution of molecules between the two urns.

where $x(i) = 1$ if ball i is in urn A and $x(i) = 0$ otherwise. Two configurations x and y in F are called neighbors, written $x \sim y$ (or $x \sim_i y$), if they differ in exactly one coordinate (which is i). This relation gives F the structure of an undirected graph whose vertices are the points of F and whose edges are pairs of neighboring vertices. The random walk on the discrete cube $Y = (Y_n)_{n \geqslant 0}$ is the Markov chain with state space F and transition matrix given, for every $x, y \in F$, by

$$\mathbf{Q}(x,y) = \begin{cases} \dfrac{1}{a} & \text{si } x \sim y, \\ 0 & \text{sinon.} \end{cases}$$

This matrix is symmetric and, in particular, doubly stochastic. The *hitting time* of $y \in E$ is the random variable $\inf\{n \geqslant 0 : Y_n = y\}$. On the event $\{Y_0 = y\}$, the *return time* to y is $\inf\{n \geqslant 1 : Y_n = y\}$.

Lemma 9.1 (First properties)**.** *The Markov chain Y with state space F and transition matrix $\mathbf{Q}$ is irreducible, recurrent, and has period 2. The uniform distribution on F, denoted by π, is invariant and reversible*[1]*. Moreover, the mean return time to $x \in F$ does not depend on x and equals 2^a: for every $x \in F$,*

[1] That is, $\pi(x)\mathbf{Q}(x,y) = \pi(y)\mathbf{Q}(y,x)$ for every x, y. For the trajectories of Y, if $Y_0 \sim \pi$ then $(Y_0, \ldots, Y_n)$ and $(Y_n, \ldots, Y_0)$ have the same distribution, for every $n \in \mathbb{N}$.

$$\mathbb{E}(\inf\{n \geqslant 1 : Y_n = x\} \,|\, Y_0 = x) = \frac{1}{\pi(x)} = 2^a.$$

The distribution of the hitting time of y starting from x depends only on

$$d(x, y) = \sum_{i=1}^{a} |x_i - y_i| = \|x - y\|_1.$$

This quantity is a distance counting the coordinates in which x and y differ. Write m_d (temporarily suppressing the dependence on a) for the mean hitting time of y starting from x when $d(x, y) = d$, that is,

$$m_d := \mathbb{E}(\inf\{n \geqslant 0 : Y_n = y\} \,|\, Y_0 = x).$$

Theorem 9.2 (Mean times)**.** *For $1 \leqslant d \leqslant a$, we have*

$$m_d = \sum_{i=1}^{d} \mathbf{Q}^a_{a-i} \quad \textit{où} \quad \mathbf{Q}^a_i = \sum_{k=0}^{i} \frac{\binom{a}{k}}{\binom{a-1}{i}} \quad \textit{si } 1 \leqslant i \leqslant a-1.$$

Proof. By the Markov property, the sequence $(m_d)_{0 \leqslant d \leqslant a+1}$ satisfies

$$\begin{cases} m_d = 1 + \dfrac{d}{a} m_{d-1} + \dfrac{a-d}{a} m_{d+1} & \text{si } 1 \leqslant d \leqslant a, \\ m_0 = m_{a+1} = 0, \end{cases}$$

and this system of recurrence equations has a unique solution. □

Since Y has period 2, the distribution of Y_n does not converge to the invariant distribution π as $n \to \infty$. To obtain an aperiodic chain, we make it lazy[2]. This consists in modifying the transition matrix by setting, for every $x, y \in F$,

$$\mathbf{R}(x, y) = \frac{1}{2} \mathbb{1}_{\{y=x\}} + \frac{1}{2} \mathbf{Q}(x, y).$$

The chain $(Z_n)_{n \geqslant 0}$ with transition matrix $\mathbf{R}$ stays in place with probability $1/2$ or, with the same probability, moves to a uniformly chosen neighboring vertex. This chain is irreducible, recurrent, aperiodic, and reversible with respect to π. The distribution of Z_n therefore converges to π regardless of the distribution of Z_0. We present two methods to quantify this convergence. The first uses coupling and the coupon collector of Section 1.5. The second, sharper method uses the spectral decomposition of $\mathbf{R}$. Write $d_{\mathrm{VT}}(\cdot, \cdot)$ for the total variation distance studied in Section 1.3.

Theorem 9.3 (Long-time behavior and the coupon collector)**.** *For every $n \geqslant 0$,*

$$\max_{x \in F} d_{\mathrm{VT}}(\mathbf{R}^n(x, \cdot), \pi) \leqslant \mathbb{P}(T_a > n),$$

where T_a is the completion time of a collection of a coupons.

[2] A chain that may stay at its current state.

Proof. Fix $n \geqslant 0$. By Theorem 1.5, for every $x \in F$,

$$\begin{aligned} d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\pi\right) &= d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\pi\mathbf{R}^n\right) \\ &= \frac{1}{2}\sum_{y\in F}\Big|\mathbf{R}^n(x,y) - \sum_{x'\in F}\pi(x')\mathbf{R}^n(x',y)\Big| \\ &\leqslant \frac{1}{2}\sum_{y\in F}\sum_{x'\in F}\pi(x')\Big|\mathbf{R}^n(x,y) - \mathbf{R}^n(x',y)\Big| \\ &\leqslant \max_{x'\in F} d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\mathbf{R}^n(x',\cdot)\right). \end{aligned}$$

To bound $d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\mathbf{R}^n(x',\cdot)\right)$, construct a pair (Z,Z') of Markov chains $Z := (Z_n)_{n\geqslant 0}$ and $Z' := (Z'_n)_{n\geqslant 0}$ with the same transition matrix $\mathbf{R}$ and initial states $Z_0 = x$ and $Z'_0 = x'$, whose trajectories coincide after a random (coalescence) time that we can control. More precisely, take two independent sequences of independent random variables $(U_n)_{n\geqslant 0}$ and $(V_n)_{n\geqslant 0}$ with respective distributions Bernoulli $\mathrm{Ber}(1/2)$ and uniform on $\{1,\dots,a\}$. For every $n \geqslant 0$ and $1 \leqslant i \leqslant a$, set

$$\begin{cases} Z_{n+1}(V_n) = U_n, \\ Z_{n+1}(i) = Z_n(i) \text{ si } i \neq V_n, \end{cases} \qquad \begin{cases} Z'_{n+1}(V_n) = U_n, \\ Z'_{n+1}(i) = Z'_n(i) \text{ si } i \neq V_n. \end{cases}$$

It is easy to check that Z and Z' are Markov chains with transition matrix $\mathbf{R}$. At each time, a coordinate is chosen uniformly and assigned the same random value in both chains. If the coordinates were equal, they remain equal; otherwise, they become equal. The distribution of the global *coalescence time*

$$T_d := \inf\{n \geqslant 0 : Z_n = Z'_n\}$$

depends only on the number $d := d(x,x')$ of differing coordinates between x and x'. For every function $f : F \to \mathbb{R}$, writing $\Delta_n := f(Z_n) - f(Z'_n)$ gives

$$\mathbb{E}(\Delta_n) = \mathbb{E}(\underbrace{\Delta_n \mathbb{1}_{\{T_d \leqslant n\}}}_{=0}) + \mathbb{E}(\Delta_n \mathbb{1}_{\{T_d > n\}}) = \mathbb{E}(\Delta_n \mathbb{1}_{\{T_d > n\}}),$$

hence $\mathbb{E}(\Delta_n) \leqslant 2\|f\|_\infty \mathbb{P}(T_d > n)$, and therefore, by Theorem 1.5,

$$\begin{aligned} d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\mathbf{R}^n(x',\cdot)\right) &= d_{\mathrm{VT}}\left(\mathrm{Loi}(Z_n),\mathrm{Loi}(Z'_n)\right) \\ &= \frac{1}{2}\sup_{\|f\|_\infty \leqslant 1}\mathbb{E}(f(Z_n) - f(Z'_n)) \\ &\leqslant \mathbb{P}(T_d > n). \end{aligned}$$

Alternatively, use Theorem 1.9 and the definition of T_d:

$$d_{\mathrm{VT}}\left(\mathrm{Loi}(Z_n),\mathrm{Loi}(Z'_n)\right) \leqslant \mathbb{P}(Z_n \neq Z'_n) \leqslant \mathbb{P}(T_d > n).$$

To bound the tail of T_d, first observe that T_d is precisely the first time at which all coordinates in $I := \{i \in \{1, \dots, a\} : x_i \neq x'_i\}$ have been updated at least once. We have $\mathrm{card}(I) = d$. To simplify notation, suppose $I = \{1, \dots, d\}$. If the random variables $(\tau_i)_{1 \leqslant i \leqslant d}$ denote the successive times at which a new coordinate is modified, then $T_d = \tau_d$, and the random variables $(G_i)_{1 \leqslant i \leqslant d}$ defined by $G_1 = 1$ and $G_i = \tau_i - \tau_{i-1}$ for $2 \leqslant i \leqslant d$ are independent and geometrically distributed with respective parameters $(d+1-i)/d$. This is the coupon collector of Section 1.5. It only remains to observe that $\mathbb{P}(T_d > n)$ increases with d and attains its maximum at $d = a$, corresponding to $x = (0, \dots, 0)$ and $x' = (1, \dots, 1)$. □

Theorem 1.14 ensures that, for every real $c \geqslant 0$,

$$\mathbb{P}(T_a \geqslant a \log(a) + ca) \leqslant e^{-c+1/a}.$$

The next theorem refines this estimate of $\max_{x \in F} d_{\mathrm{VT}}\left(\mathbf{R}^n(x, \cdot), \pi\right)$.

Theorem 9.4 (Long-time behavior)**.** *If*

$$n = \frac{a \log(a)}{2} + ca \quad \textit{avec} \quad c \geqslant 0$$

then

$$\max_{x \in F} d_{\mathrm{VT}}\left(\mathbf{R}^n(x, \cdot), \pi\right) \leqslant \frac{e^{-c}}{\sqrt{2}}.$$

Proof. Equip $\mathbb{R}^F \equiv \mathbb{R}^{2^a}$ with its usual inner product, denoted by $\langle \cdot, \cdot \rangle$. The stochastic matrix $\mathbf{R}$ is symmetric[3]; its spectrum is real and contained in

$$[-1, 1] = \mathbb{R} \cap \{z \in \mathbb{C} : |z| \leqslant 1\}.$$

Moreover, there is an orthonormal basis of $\mathbb{R}^F$ consisting of eigenvectors $(f_j)_{1 \leqslant j \leqslant 2^a}$ of $\mathbf{R}$ associated with the eigenvalues $(\lambda_j)_{1 \leqslant j \leqslant 2^a}$ in decreasing order. In particular, λ_1 equals 1 and f_1 is the constant function equal to $2^{-a/2}$. Recall that π is the uniform distribution on F, which has cardinality 2^a, so $\pi(x) = 2^{-a}$ for every $x \in F$. For every function f from F to $\mathbb{R}$ and x in F,

$$\mathbf{R}^n(f)(x) = \sum_{j=1}^{2^a} \langle f, f_j \rangle f_j(x) \lambda_j^n = \pi(f) + \sum_{j=2}^{2^a} \langle f, f_j \rangle f_j(x) \lambda_j^n.$$

For the function $z \in F \mapsto f(z) = \mathbb{1}_y(z)$, we obtain the spectral decomposition

$$\mathbf{R}^n(x, y) = \frac{1}{2^a} + \sum_{j=2}^{2^a} f_j(x) f_j(y) \lambda_j^n.$$

[3] For any transition matrix on a finite state space, the uniform distribution is invariant (respectively reversible) if and only if the matrix is doubly stochastic (respectively symmetric).

By Theorem 1.5 and Schwarz's inequality,

$$4d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\pi(\cdot)\right)^2 = \Big(\sum_{y\in F}|\mathbf{R}^n(x,y)-\pi(y)|\Big)^2 \leqslant 2^a \sum_{y\in F}|\mathbf{R}^n(x,y)-\pi(y)|^2.$$

Since $(f_i)_i$ is an orthonormal basis, the spectral decomposition gives

$$\sum_{y\in F}|\mathbf{R}^n(x,y)-\pi(y)|^2 = \sum_{i,j=2}^{2^a} f_i(x)f_j(x)\langle f_i,f_j\rangle\lambda_i^n\lambda_j^n = \sum_{i=2}^{2^a} f_i(x)^2\lambda_i^{2n}.$$

By symmetry of F and $\mathbf{R}$, the quantity $\sum_{y\in F}|\mathbf{R}^n(x,y)-\pi(y)|^2$ does not depend on $x\in F$. Summing over $x\in F$ gives

$$2^a\sum_{y\in F}|\mathbf{R}^n(x,y)-\pi(y)|^2 = \sum_{i=2}^{2^a}\lambda_i^{2n},$$

which yields the upper bound

$$4d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\pi(\cdot)\right)^2 \leqslant \sum_{i=2}^{2^a}\lambda_i^{2n}.$$

One can show that the eigenvalues of $\mathbf{R}$ are the real numbers $(1-k/a)_{0\leqslant k\leqslant a}$ with respective multiplicities $\binom{a}{k}$ for $0\leqslant k\leqslant a$. Hence

$$\begin{aligned}4d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\pi\right)^2 &\leqslant \sum_{k=1}^{a}\left(1-\frac{k}{a}\right)^{2n}\binom{a}{k}\\ &\leqslant \sum_{k=1}^{a} e^{-2nk/a}\binom{a}{k} = \left(1+e^{-2n/a}\right)^a - 1.\end{aligned}$$

Choosing $n=(1/2)a\log(a)+ca$ gives

$$4d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\pi\right)^2 \leqslant \left(1+\frac{e^{-2c}}{a}\right)^a - 1 \leqslant e^{e^{-2c}}-1.$$

We have $e^{e^{-2c}}-1\leqslant 2e^{-2c}$ whenever $c\geqslant 0$, which proves the result. □

Another way to avoid the periodicity of the matrix $\mathbf{Q}$ is to introduce a continuous-time version $(Z_t)_{t\geqslant 0}$ of the chain $(Y_n)_{n\geqslant 0}$. Suppose each coordinate flips[4] from 0 to 1 or from 1 to 0 at random times. More precisely, let $(N^{(i)})_{1\leqslant i\leqslant a}$ be independent Poisson processes with the same intensity $1/a$. Set

$$Z_t(i) = Z_0(i) + N_t^{(i)} \bmod 2.$$

[4] Changes to the opposite value.

Each coordinate changes independently of the others at exponential times with parameter $1/a$. This dynamics can be summarized by the infinitesimal generator of the process $\mathbf{L} = \mathbf{Q} - \mathbf{I}$: for every function f from F to $\mathbb{R}$ and every $x \in F$,

$$\mathbf{L}f(x) = \frac{1}{a}\sum_{y\sim x}(f(y) - f(x)) = \frac{1}{a}\sum_{i=1}^{n}(f(x_{(i)}) - f(x))$$

where $x_{(i)} \in F$ is obtained from x by flipping coordinate i[e]The uniform distribution π solves $\pi\mathbf{L} = 0$ and is the invariant distribution of Z.

Theorem 9.5 (Long-time behavior of the continuous-time model)**.** *For every* $t \geqslant 0$*,*

$$\max_{x\in F} d_{\mathrm{VT}}\left(\mathrm{Loi}(Z_t \,|\, Z_0 = x), \pi\right) \leqslant \mathbb{P}(S_a > t)$$

where $S_a := \max_{1\leqslant i\leqslant a} E_i$ *and* $E_1, \ldots, E_a$ *are i.i.d. real random variables with distribution* $\mathrm{Exp}(2/a)$*.*

Proof. We use the coupling proof of Theorem 9.3. First, in the same way, for every $x \in F$,

$$d_{\mathrm{VT}}\left(\mathrm{Loi}(Z_t \,|\, Z_0 = x), \pi\right) \leqslant \max_{x'\in F} d_{\mathrm{VT}}\left(\mathrm{Loi}(Z_t \,|\, Z_0 = x), \mathrm{Loi}(Z_t \,|\, Z_0 = x')\right).$$

Then construct a coupling as follows: the coordinate pairs $(Z_\cdot(i), Z'_\cdot(i))_{1\leqslant i\leqslant a}$ evolve independently, and for each i, $Z_t(i)$ and $Z'_t(i)$ are independent until they meet and remain together thereafter. The infinitesimal generator $\mathbb{L}$ of the process $((Z_t, Z'_t))_{t\geqslant 0}$ is given, for every function f from $F \times F$ to $\mathbb{R}$ and every $(x, x') \in F \times F$, by

$$\begin{aligned}\mathbb{L}f(x,x') = \frac{1}{a}\sum_{i=1}^{a}\Big((f(x_{(i)}, x'_{(i)}) - f(x,x'))\mathbb{1}_{x_i = x'_i} \\ + (f(x, x'_{(i)}) - f(x,x'))\mathbb{1}_{x_i \neq x'_i} \\ + (f(x_{(i)}, x') - f(x,x'))\mathbb{1}_{x_i \neq x'_i}\Big).\end{aligned}$$

It coincides with $\mathbf{L}$ when f depends only on the first or only on the second variable: Z and Z' are chains with generator $\mathbf{L}$. Take the initial condition $(Z_0, Z'_0) = (x, x')$. As in the coupling proof of Theorem 9.3, the distribution of the global coalescence time

$$S_d := \inf\{t \geqslant 0 : Z_t = Z'_t\}$$

depends on (x, x') only through $d := d(x, x')$, and

$$d_{\mathrm{VT}}\left(\mathrm{Loi}(Z_t \,|\, Z_0 = x), \mathrm{Loi}(Z_t \,|\, Z_0 = x')\right) \leqslant \mathbb{P}(S_d > t).$$

To bound the tail of S_d, we seek a more convenient representation. Initially, there are $a-d+2d = a+d$ possible transitions, all with rate $1/a$, of which $2d$ cause a new coordinate coalescence. This gives a total rate of $\lambda = (a+d)/a$ and a coordinate coalescence probability of $p = 2d/(a+d)$. The construction of trajectories[5] and a property of exponential distributions[6] show that the first coordinate coalescence occurs after a random duration F_d exponentially distributed with parameter

$$p\lambda = \frac{2d}{a+d}\frac{a+d}{a} = \frac{2}{a}d.$$

More generally, for every $i \in \{1,\dots,d\}$, coordinate coalescence i^{e} occurs after a random duration F_{d-i+1} exponentially distributed with parameter

$$\frac{2(d-i+1)}{a+d-i+1}\frac{a+d-i+1}{a} = \frac{2}{a}(d-i+1).$$

These durations are independent, and the global coalescence time

$$S_d := \inf\{t \geqslant 0 : Z_t = Z'_t\}$$

has the same distribution as $F_d + \cdots + F_1$, where $F_d,\dots,F_1$ are independent real random variables with $F_k \sim \mathrm{Exp}(2k/a)$ for every $k \in \{1,\dots,d\}$. Rényi's lemma 11.3 states that $F_d + \cdots + F_1$ has the same distribution as $\max(E_1,\dots,E_d)$, where $E_1,\dots,E_d$ are i.i.d. real random variables exponentially distributed with parameter $2/a$.

$$S_d \overset{\text{loi}}{=} \max_{1\leqslant i\leqslant d} E_i.$$

For every $t \geqslant 0$, the quantity $\mathbb{P}(S_d > t)$ increases with d and attains its maximum $\mathbb{P}(S_a > t)$ at $d = a$, corresponding to $x = (0,\dots,0)$ and $x' = (1,\dots,1)$. □

The random time S_a is easier to study than the collection completion time T_a, although they have the same asymptotic behavior, as the following result shows.

Theorem 9.6 (Long-time behavior of the continuous-time chain)**.**

$$\frac{S_a - (a/2)\log(a)}{a/2} \xrightarrow[a\to\infty]{\text{loi}} \text{Gumbel}$$

where the Gumbel distribution has distribution function $t \in \mathbb{R} \mapsto e^{-e^{-t}}$.

[5] If Z is a continuous-time Markov chain on a discrete state space E with infinitesimal generator $\mathbf{G} = (\mathbf{G}(x,y))_{x,y\in E}$, then, given that the chain is at x at time t, the next jump occurs after an exponential time with parameter $-\mathbf{G}(x,x)$ and goes to $y \neq x$ with probability $-\mathbf{G}(x,y)/\mathbf{G}(x,x)$.

[6] Lemma 11.4 states that if $G, E_1, E_2, \dots$ are independent random variables, with G geometrically distributed with parameter p and $E_1, E_2, \dots$ exponentially distributed with parameter λ, then the random sum $E_1 + \cdots + E_G$ is exponentially distributed with parameter $p\lambda$.

As for T_a (Theorem 1.17), S_a exhibits cutoff around $(a/2)\log(a)$ with a window of width a.

Proof. For every $a \geqslant 1$ and $r \geqslant -\log(a)$,

$$\begin{aligned}\mathbb{P}\left(\frac{S_a - (a/2)\log(a)}{a/2} \leqslant r\right) &= \mathbb{P}\left(\max_{1\leqslant i\leqslant a}(E_i) \leqslant \frac{ar}{2} + \frac{a\log(a)}{2}\right)\\ &= \mathbb{P}\left(E_1 \leqslant \frac{a}{2}(r+\log(a))\right)^a\\ &= \left(1 - \frac{e^{-r}}{a}\right)^a.\end{aligned}$$

Thus, for every $r \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{S_a - (a/2)\log a}{a/2} \leqslant r\right) \underset{a\to\infty}{\longrightarrow} e^{-e^{-r}},$$

which is the stated convergence. □

9.2 The Ehrenfest urn

Observing the evolution of the chain Y would require determining the position of every molecule, which is physically impossible since a is of the order of Avogadro's number[7]. The macroscopic quantity of interest, which we can measure, is the proportion or number of molecules in urn A. To simplify some expressions, we shall sometimes assume that a is even. The letter b will then always denote the integer $a/2$. To the Markov chain Y describing the system microscopically, we associate the process $(X_n)_{n\geqslant 0}$ taking values in

$$E := \{0, 1, \ldots, a\}$$

where

$$X_n := \sum_{i=1}^{a} Y_n(i)$$

is the sum of the coordinates of Y_n, describing the system macroscopically.

Theorem 9.7 (The Ehrenfest chain)**.** *The process $(X_n)_{n\geqslant 0}$ is a Markov chain on E with transition matrix*

$$\mathbf{P}(x,x') = \begin{cases} x/a & \text{si } x' = x-1;\\ 1-x/a & \text{si } x' = x+1;\\ 0 & \text{sinon,}\end{cases} \quad \text{pour tous } x, x' \in E,$$

[7] This number, approximately 6.022×10^{23}, defines the mole and corresponds to the number of atoms in 12 grams of carbon 12. It is also approximately the number of molecules in 22.414 liters of an ideal gas under standard conditions of temperature (0°C) and pressure (1 atmosphere).

that is irreducible, recurrent, and has period 2*. The binomial distribution* $\mathrm{Bin}(a, 1/2)$ *is invariant and reversible.*

Proof. Let $\mathcal{F}_n$ and $\mathcal{G}_n$ be the sigma-algebras generated by the sequences $(X_k)_{0\leqslant k\leqslant n}$ and $(Y_k)_{0\leqslant k\leqslant n}$, respectively, and let $S(y) = y(1) + \cdots + y(a)$ for every $y \in F$. Since $X_n = S(Y_n)$, the sigma-algebra $\mathcal{F}_n$ is contained in $\mathcal{G}_n$. Therefore, for every function f from E to $\mathbb{R}$,

$$\mathbb{E}(f(X_{n+1}) \,|\, \mathcal{F}_n) = \mathbb{E}(\mathbb{E}(f(X_{n+1}) \,|\, \mathcal{G}_n) \,|\, \mathcal{F}_n).$$

By definition of X and since Y is a Markov chain,

$$\mathbb{E}(f(X_{n+1}) \,|\, \mathcal{G}_n) = \mathbb{E}(f(S(Y_{n+1})) \,|\, \mathcal{G}_n) = \mathbb{E}(f(S(Y_{n+1})) \,|\, Y_n).$$

Finally,

$$\mathbb{E}(f(S(Y_{n+1})) \,|\, Y_n) = \left(1 - \frac{S(Y_n)}{a}\right) f(S(Y_n) + 1) + \left(\frac{S(Y_n)}{a}\right) f(S(Y_n) - 1).$$

Hence

$$\begin{aligned}\mathbb{E}(f(X_{n+1}) \,|\, \mathcal{F}_n) &= \mathbb{E}\left(\left(1 - \frac{X_n}{a}\right) f(X_n + 1) + \left(\frac{X_n}{a}\right) f(X_n - 1) \,|\, \mathcal{F}_n\right) \\ &= \left(1 - \frac{X_n}{a}\right) f(X_n + 1) + \left(\frac{X_n}{a}\right) f(X_n - 1),\end{aligned}$$

and thus X is a Markov chain with transition matrix $\mathbf{P}$. It is irreducible because i and $i + 1$ communicate for every $0 \leqslant i \leqslant a$, with the convention $a+1 = 0$. Since the state space is finite, the matrix is (positive) recurrent. For every n, X_n and X_{n+1} have opposite parity, while $\mathbf{P}^2_{0,0} > 0$, so the chain has period 2. Note that, in general, the image of a Markov chain under a function is not a Markov chain. □

9.3 Pressure over time

The pressure at time n in urn A is of the order of

$$P_n := \frac{X_n}{a}.$$

We study the first two moments of this random variable.

Theorem 9.8 (Mean and variance of the pressure)**.** *For every* $n \in \mathbb{N}$*, writing* $\alpha = 1 - 2/a$ *and* $\beta = 1 - 4/a$*,*

$$\begin{aligned}\mathbb{E}(P_n) &= \frac{1}{2} + \left(\mathbb{E}(P_0) - \frac{1}{2}\right)\alpha^n, \\ \mathrm{Var}(P_n) &= \frac{1 - \beta^n}{4a} + \mathrm{Var}(P_0)\beta^n - \left(\mathbb{E}(P_0) - \frac{1}{2}\right)^2 (\alpha^{2n} - \beta^n).\end{aligned}$$

Proof. By conditioning,

$$\begin{aligned}\mathbb{E}(X_{n+1} \,|\, X_n) &= (X_n+1)\left(1-\frac{X_n}{a}\right) + (X_n-1)\frac{X_n}{a} \\ &= \alpha X_n + 1.\end{aligned}$$

Taking expectations yields a recurrence for the sequence $(\mathbb{E}(P_n))_n$, which immediately gives the formula for the mean. Similarly,

$$\begin{aligned}\mathbb{E}(X_{n+1}^2 \,|\, X_n) &= (X_n+1)^2\left(1-\frac{X_n}{a}\right) + (X_n-1)^2\frac{X_n}{a} \\ &= \left(1-\frac{4}{a}\right)X_n^2 + 2X_n + 1.\end{aligned}$$

This implies

$$\begin{aligned}\mathrm{Var}(X_{n+1}) &= \beta\mathrm{Var}(X_n) + \frac{4}{a^2}\mathbb{E}(X_n)(a-\mathbb{E}(X_n)) \\ &= \beta\mathrm{Var}(X_n) + 1 - \frac{4}{a^2}\left(\mathbb{E}(X_0) - \frac{a}{2}\right)^2\alpha^{2n}.\end{aligned}$$

A straightforward calculation then gives the variance formula. □

Remark 9.9 (Eigenvalues and eigenvectors)**.** *Let f be the function on E defined by $f(x) = x - b$, where $b = a/2$, identified with the column vector $(-b, -b+1, \ldots, b-1, b)^T$. A direct calculation gives $\mathbf{P}f = (1-2/a)f$. Thus, for every $n \geqslant 0$ and every $x \in E$,*

$$\mathbb{E}(X_n \,|\, X_0 = x) - b = \mathbb{E}(f(X_n) \,|\, X_0 = x) = \mathbf{P}^n f(x) = \alpha^n f(x) = \alpha^n(x-b).$$

Taking expectations with respect to the distribution of X_0 recovers the mean relation in Theorem 9.8. More generally, one can show that if $a = 2b$ is even, then the eigenvalues of $\mathbf{P}$ are the rational numbers $(i/b)_{-b\leqslant i\leqslant b}$.

By Theorem 9.8, if urn A is initially empty, then $X_0 = 0$, $P_0 = 0$, and the mean pressure $\mathbb{E}(P_n)$ at time n converges geometrically to $1/2$ as $n \to \infty$. Moreover, $\mathrm{Var}(P_n) \leqslant 1/(4a)$ since $\mathrm{Var}(P_0) = 0$ and $\alpha^2 \geqslant \beta$, and Chebyshev's inequality gives

$$\mathbb{P}(|P_n - \mathbb{E}(P_n)| \geqslant \varepsilon) \leqslant \frac{1}{4a\varepsilon^2}.$$

Since a is very large, the variable P_n is practically deterministic, in agreement with experiment. Figure 9.2 shows a trajectory of $(P_n)_{n\geqslant 0}$ for various initial conditions.

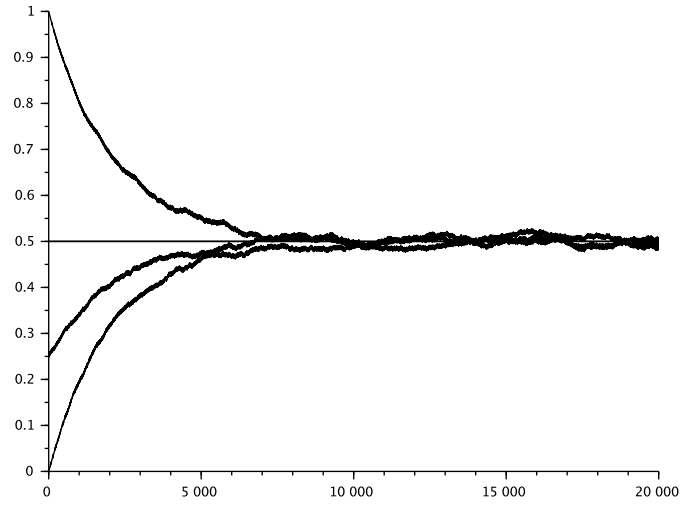


Fig. 9.2. Evolution of the proportion of molecules in urn A for initial values 0, 0.25, and 1, with $a = 4000$ molecules.

9.4 Fluctuations around the mean at equilibrium

Suppose the system is at equilibrium: X has the binomial distribution $\mathrm{Bin}(a, 1/2)$. By the central limit theorem, $P = X/a$ satisfies

$$\mathbb{P}\Big(|P - 1/2| \geqslant \frac{x}{2\sqrt{a}}\Big) \underset{a \text{ grand}}{\simeq} 2(1 - \Phi(x)) \leqslant \frac{2}{x\sqrt{2\pi}} e^{-x^2/2},$$

where Φ is the distribution function of the standard Gaussian $\mathcal{N}(0, 1)$.

A nonasymptotic bound on this probability can also be obtained using Chernoff's inequality. More precisely, if $(Z_i)_{1 \leqslant i \leqslant n}$ are independent identically distributed random variables with Rademacher distribution $(1/2)(\delta_{-1} + \delta_1)$, then, for every $r \geqslant 0$ and every $\lambda > 0$,

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} Z_i \geqslant r\Big) = \mathbb{P}\Big(\exp\Big(\lambda \sum_{i=1}^{n} Z_i\Big) \geqslant e^{\lambda n r}\Big) \leqslant e^{-\lambda n r} \mathrm{ch}(\lambda)^n.$$

Since $\mathrm{ch}(\lambda) \leqslant \exp\left(\lambda^2/2\right)$ (as follows, for example, from a power series expansion), optimizing over λ, which leads to choosing $\lambda = r$, gives

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} Z_i \geqslant r\Big) \leqslant e^{-n r^2/2}.$$

This ensures that

$$\mathbb{P}(|P - 1/2| \geqslant r) \leqslant 2e^{-ar^2/2}.$$

If $a = 6\times10^{23}$, then taking $r = 10^{-9}$ shows that the probability of a fluctuation exceeding one billionth is less than $2e^{-3\times10^5}$. This extremely small probability makes the event unobservable.

9.5 When will the urn be empty again?

The Ehrenfest chain $(X_n)_{n\geqslant0}$ is irreducible with a finite state space, so it visits every state, and the hitting time of each state is integrable (in particular, almost surely finite). This may seem paradoxical in view of macroscopic irreversibility. Let us therefore estimate these mean hitting times. It is simplest to start with return times. On the event $\{X_0 = i\}$, the return time T_{ii} to i is the random variable

$$T_{ii} := \inf\{n \geqslant 1 : X_n = i\}.$$

Theorem 9.10 (Mean return times)**.** *If a is even and $b := a/2$, then*

$$\mathbb{E}(T_{00} \,|\, X_0 = 0) = 2^a \quad \textit{et} \quad \mathbb{E}(T_{bb} \,|\, X_0 = b) \sim \sqrt{\pi b}.$$

Proof. The first result follows from Lemma 9.1. We know that $\mathbb{E}(T_{ii} \,|\, X_0 = i)$ equals $1/\pi(i)$, where π is the invariant distribution, namely $\mathrm{Bin}(a, 1/2)$. We have $\pi(i) = 2^{-a}\binom{a}{i}$, so the result is immediate for $i = 0$ and follows from Stirling's formula for $i = b$. □

We must therefore wait an immense time before the initially empty urn A becomes empty again! We now seek more precise estimates of the hitting times of 0 and b starting from 0 or b. Let

$$m_{i,j} := \mathbb{E}(\inf\{n \geqslant 0 : X_n = j\} \,|\, X_0 = i)$$

be the mean hitting time of j for the chain starting from i. Starting from $i+1$, the chain must pass through i to reach 0, so, for $0 \leqslant i < a$,

$$m_{i+1,0} = m_{i+1,i} + m_{i,0}.$$

By symmetry, for $0 \leqslant i < a$,

$$m_{i,i+1} = m_{a-i,a-i-1} = m_{a-i,0} - m_{a-i-1,0}.$$

Thus the numbers $(m_{i,j})_{ij}$ are determined by $(m_{i,0})_i$. Now observe that $X_n = 0$ if and only if $Y_n = (0,\dots,0)$. Hence the mean time taken by X to go from i to 0 equals the mean time taken by Y to go from any element of F at distance i from $(0,\dots,0)$ to $(0,\dots,0)$, namely m_i. Theorem 9.2 therefore gives, for $0 \leqslant i \leqslant a-1$,

$$m_{i,i+1} = m_{a-i,0} - m_{a-i-1,0} = \sum_{k=1}^{a-i} \mathbf{Q}^a_{a-k} - \sum_{k=1}^{a-i-1} \mathbf{Q}^a_{a-k} = \mathbf{Q}^a_i.$$

Corollary 9.11 (Mean hitting times)**.** *We have*

$$m_{i,j} = \begin{cases} 2^a / \binom{a}{i} & \text{si } i = j, \\ \sum_{k=i}^{j-1} \mathbf{Q}_k^a & \text{si } i < j, \\ \sum_{k=a-i}^{a-j-1} \mathbf{Q}_k^a & \text{si } i > j. \end{cases}$$

Theorem 9.12 (Mean hitting times)**.** *If a is even and $b := a/2$, then*

$$a\Big(2\log(2) - \frac{1}{2}\Big) - 2 \leqslant m_{0,b} \leqslant a\Big(\frac{1}{4} + \frac{1}{2}\log(a)\Big),$$

and

$$2^a\left(1 + \frac{1}{a-1}\right) \leqslant m_{0,a} \leqslant 2^a\left(1 + \frac{2}{a-1}\right).$$

Proof of the second pair of bounds in Theorem 9.12. For every $0 \leqslant i \leqslant a-1$,

$$m_{i,i+1} + m_{i+1,i} = \frac{1}{\binom{a-1}{i}}\left(\sum_{k=0}^{i}\binom{a}{k} + \sum_{k=i+1}^{a}\binom{a}{k}\right) = \frac{2^a}{\binom{a-1}{i}}.$$

Hence

$$m_{0,b} + m_{b,0} = 2^a \sum_{i=0}^{b-1} \frac{1}{\binom{a-1}{i}}.$$

Moreover, since $m_{b,0} = m_{b,a}$, we have

$$m_{0,a} = m_{0,b} + m_{b,a} = 2^a \sum_{i=0}^{b-1} \frac{1}{\binom{a-1}{i}}.$$

Separating the first two terms from the rest of the sum gives

$$1 + \frac{1}{a-1} \leqslant \sum_{i=0}^{b-1} \frac{1}{\binom{a-1}{i}} \leqslant 1 + \frac{1}{a-1} + \frac{b-1}{\binom{a-1}{2}}$$

which yields the desired bounds. □

Theorem 9.12 shows that the time needed to go from a state of complete imbalance to perfect equilibrium is utterly negligible compared with the time needed for the reverse evolution. With 100 molecules, $m_{0,50}$ is bounded above by 256, whereas $m_{50,0}$ is of the order of 10^{30}, or a few thousand billion billion billion.

9.6 Further reading

The mixing time of the random walk on the discrete cube exhibits cutoff around $(a/2)\log(a)$ with window a: within an interval of length of order a around $(a/2)\log a$, the total variation distance between $\mathbf{R}^n(x,\cdot)$ and π drops from near 1 to near 0. The upper bound is provided by Theorem 9.4, which can be supplemented by the following estimate from the book by David Levin, Yuval Peres, and Elizabeth Wilmer [LPW09]: if $n = (1/2)a\log(a) - ca$ with $c > 0$, then

$$\max_{x\in F} d_{\mathrm{VT}}\left(\mathbf{R}^n(x,\cdot),\pi\right) \geqslant 1 - 8e^{-(2c+1)}.$$

This result appears to originate in work by Persi Diaconis and Mehrdad Shahshahani [DS87]. The eigenvalues and eigenvectors of the transition matrix $\mathbf{P}$ of the Ehrenfest chain were obtained by Mark Kac in [Kac47]. They can also be obtained from those of the transition matrix $\mathbf{R}$ of the microscopic model (the walk on the discrete cube), described in Levin, Peres, and Wilmer [LPW09, Example 12.15].

The classical relation between expected return times and the invariant distribution used in the proof of Theorem 9.10 is established, for example, in James Norris' book [Nor98a]. Theorems 9.2 and 9.12 are taken from the book by John Kemeny and Laurie Snell [KS76].

Molecules move very rapidly between the urns. Microscopic time therefore differs from macroscopic time. Suppose one unit of macroscopic time corresponds to a urn changes. How does pressure evolve on this new time scale? Theorem 9.8 immediately implies that, for every $t > 0$,

$$\mathbb{E}\big(P_{\lfloor at\rfloor}\big) \underset{a\to\infty}{\longrightarrow} \frac{1}{2} + \left(\mathbb{E}(P_0) - \frac{1}{2}\right)e^{-2t} \quad \text{et} \quad \mathbb{E}\big(V_{\lfloor at\rfloor}\big) \underset{a\to\infty}{\longrightarrow} \mathrm{Var}(X_0)e^{-4t},$$

since $\lfloor at\rfloor/a$ tends to t. This convergence is only the tip of the iceberg: suitably rescaled in time and space, the Ehrenfest chain converges in distribution to the Ornstein–Uhlenbeck diffusion process, as presented in Chapter 27.

10

Records, extremes and recruitment

Keywords. Extreme; order statistic.

Tools. Piecewise deterministic Markov process; martingale; random permutation; centred series theorem; Lindeberg-Lévy theorem; exponential distribution; Pareto distribution; Laplace transform.

Difficulty. **

A headhunting firm wishes to assemble an outstanding team through sequential recruitment: each candidate receives a score after an interview, and the recruiters decide immediately whether to hire them. The aim is to recruit a team that is both excellent and large. In this chapter we compare two recruitment policies: under the first, a person is hired if their score exceeds all previous candidates' scores; under the second, if it exceeds the average of these scores. The more flexible second procedure should lead to more recruitment, but perhaps lower quality. We model successive candidates' scores by a sequence $(X_n)_{n\geqslant 1}$ of independent real random variables with common distribution μ and distribution function F. To avoid special cases, we assume F is continuous on $\mathbb{R}$. The right endpoint of the support of μ is denoted by x_F:

$$x_F = \sup\{x \in \mathbb{R} \,:\, F(x) < 1\} \in \mathbb{R} \cup \{+\infty\}.$$

10.1 Elitism

For every $n \geqslant 1$, the relative rank of X_n among $X_1, \ldots, X_n$ is given by

$$R_n = 1 + \sum_{k=1}^{n-1} \mathbb{1}_{\{X_n > X_k\}}.$$

Candidate n is hired if their rank is n, that is, if their score breaks the previous record. The number of records, or people hired, up to time n is given by

$$Z_n = \sum_{i=1}^{n} \mathbb{1}_{\{R_i = i\}}.$$

We also define the record times recursively:

$$T_1 = 1 \quad \text{et, pour } j \geqslant 1, \quad T_{j+1} = \inf\{n > T_j : X_n > X_i \text{ pour } i < n\}.$$

Lemma 10.1 (Rank distributions)**.** *The random variables $(R_n)_{n\geqslant 1}$ are independent and, for every $n \in \mathbb{N}^*$, R_n is uniformly distributed on $\{1, \ldots, n\}$,*

$$\mathbb{E}(Z_n) = \sum_{k=1}^{n} \frac{1}{k} = \log(n) + \gamma + \mathcal{O}(1/n),$$

and

$$\mathrm{Var}(Z_n) = \sum_{k=1}^{n} \frac{k-1}{k^2} = \log(n) + \gamma - \frac{\pi^2}{6} + \mathcal{O}(1/n),$$

where γ denotes Euler's constant.

Proof. The distribution function F of μ is continuous, so for any distinct integers i and j, $\mathbb{P}(X_i \neq X_j) = 1$. In particular, for $n \geqslant 1$, almost surely there exists a unique (random) permutation σ of $\{1, \ldots, n\}$ such that

$$X_{\sigma(1)} < \cdots < X_{\sigma(n)}.$$

Moreover, this permutation is uniformly distributed on $\mathcal{S}_n$. Finally, the correspondence between a vector $(R_1, \ldots, R_n)$ and the permutation σ is bijective, hence

$$\mathbb{P}(R_1 = r_1, \ldots, R_n = r_n) = \frac{1}{n!}$$

for every $(r_1, \ldots, r_n)$ satisfying $1 \leqslant r_i \leqslant i$ for $1 \leqslant i \leqslant n$. The random variable Z_n is therefore the sum of n independent Bernoulli random variables with parameters 1, $1/2$,..., $1/n$. Its expectation and variance follow. □

Note that the sequence $(Z_n)_{n\geqslant 1}$ has the same distribution as the sequence $(|B_n|)_{n\geqslant 1}$ of the Chinese restaurant process in Chapter 14; see also Theorem 13.5 on the sequence $(K_n)_{n\geqslant 1}$ counting alleles in a sample, obtained when studying the genealogy of the Wright-Fisher model.

Theorem 10.2 (Asymptotic behaviour of the number of records)**.**

$$\frac{Z_n}{\log(n)} \xrightarrow[n\to\infty]{\text{p.s.}} 1 \quad \textit{et} \quad \frac{Z_n - \log(n)}{\sqrt{\log(n)}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

Proof. Almost sure convergence relies on the centred series theorem recalled below.

Theorem 10.3 (Centred series)**.** *Let $(Q_n)_{n\geqslant 1}$ be a sequence of independent centred random variables and $(b_n)_{n\geqslant 1}$ a strictly increasing, unbounded (deterministic) sequence of strictly positive real numbers.*

$$Si \quad \sum_{n=1}^{\infty} \frac{\mathbb{E}(Q_n^2)}{b_n^2} < \infty, \quad alors \quad \frac{1}{b_n}\sum_{k=1}^{n} Q_k \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

We apply this theorem to the sequence $(Q_n)_{n\geqslant 1}$ defined by $Q_n = \mathbb{1}_{\{R_n=n\}} - 1/n$ for $n \geqslant 1$, with $b_n = \log(n)$, noting that

$$\frac{\mathrm{Var}(Q_n)}{b_n^2} = \frac{n-1}{n^2\log(n)^2}$$

is the general term of a convergent series.

The second part follows from the Lindeberg-Lévy central limit theorem recalled below. This extends the central limit theorem to sequences of independent square-integrable random variables that need not be identically distributed. The Lindeberg condition means that the total variance is asymptotically well distributed.

Theorem 10.4 (Lindeberg-Lévy)**.** *Let $(Y_i)_{i\geqslant 1}$ be a sequence of independent, square-integrable, centred real random variables. Write $S_n = Y_1 + \cdots + Y_n$, and denote its standard deviation by s_n. If the* Lindeberg condition *holds: for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{s_n^2}\sum_{i=1}^{n} \mathbb{E}\big[Y_i^2 \mathbb{1}_{\{|Y_i|>\varepsilon s_n\}}\big] = 0,$$

then

$$\frac{S_n}{s_n} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

Apply this theorem to the variables $Y_n = \mathbb{1}_{\{R_n=n\}} - 1/n$ for $n \geqslant 1$, noting that $\mathbb{1}_{\{|Y_i|>\varepsilon s_n\}} = 0$ for sufficiently large n. □

Remark 10.5 (Using martingales)**.** *Theorem 10.2 can also be proved using the law of large numbers and the central limit theorem for martingales. Set, for $n \geqslant 1$,*

$$M_n = Z_n - \sum_{k=1}^{n} \frac{1}{k} = \sum_{k=1}^{n} \left(\mathbb{1}_{\{R_k=k\}} - \frac{1}{k}\right).$$

Then $(M_n)_{n\geqslant 1}$ is a square-integrable martingale with increasing process

$$\langle M\rangle_n = \sum_{k=1}^{n} \frac{k-1}{k^2} = \mathcal{O}(\log(n)).$$

Thus $M_n = o(\log(n))$ almost surely. Convergence in distribution also follows from the central limit theorem for $(M_n)_{n\geqslant 1}$.

10.2 Above average

We now construct the sequences $(Y_n)_{n\geqslant 1}$, $(\overline{Y}_n)_{n\geqslant 1}$ and $(T_n)_{n\geqslant 1}$ recursively by:

$$Y_1 = X_1, \quad \overline{Y}_1 = Y_1 \quad \text{et} \quad T_1 = 1,$$

and then, for every $n \geqslant 1$, by

$$T_{n+1} = \inf\left\{i > T_n \,:\, X_i > \overline{Y}_n\right\}, \quad Y_{n+1} = X_{T_{n+1}} \quad \text{et} \quad \overline{Y}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} Y_i.$$

For every $n \geqslant 1$, Y_n is the performance of the n-th successful candidate, T_n is the number of candidates interviewed to hire n, and $\overline{Y}_n$ is the average performance of the first n successful candidates.

Theorem 10.6 (Towards maximum quality)**.**

$$\overline{Y}_n \xrightarrow[n\to\infty]{\text{p.s.}} x_F.$$

Proof. The sequence $(\overline{Y}_n)_{n\geqslant 1}$ is increasing by construction. It therefore converges to a random variable Y taking values in $\overline{\mathbb{R}}$. We use the following formula: $\overline{Y}_k = \sum_{j=1}^{k}(Y_j - \overline{Y}_{j-1})/j$. Now let $x < x_F$ and $\varepsilon > 0$ be such that $x + \varepsilon < x_F$. If $x_F = +\infty$, we may take $\varepsilon = 1$. On the event $\{Y \leqslant x\}$, since $\overline{Y}_j \leqslant Y \leqslant x$, we have

$$Y = \sum_{j=1}^{\infty} \frac{Y_j - \overline{Y}_{j-1}}{j} \geqslant \sum_{j=1}^{\infty} \frac{(Y_j - x)^+}{j} \geqslant \sum_{j=1}^{\infty} \frac{\varepsilon \mathbb{1}_{\{Y_j > x+\varepsilon\}}}{j}.$$

For $k \geqslant 1$, set $M_k = \#\{j \leqslant k \,:\, Y_j > x+\varepsilon\}$ and $N_k = \#\{j \leqslant k \,:\, X_j > x+\varepsilon\}$. Then $M_k \geqslant N_k$ for every $k \geqslant 1$ and $\mathbb{1}_{\{Y_j > x+\varepsilon\}} = M_j - M_{j-1}$. Thus

$$\begin{aligned}\sum_{j=1}^{\infty} \frac{\mathbb{1}_{\{Y_j > x+\varepsilon\}}}{j} &= \sum_{j=1}^{\infty} \frac{M_j - M_{j-1}}{j} = \sum_{j=1}^{\infty} \frac{M_j}{j(j+1)} \\ &\geqslant \sum_{j=1}^{\infty} \frac{N_j}{j(j+1)} \geqslant \sum_{j=1}^{\infty} \frac{N_j - N_{j-1}}{j} = \sum_{j=1}^{\infty} \frac{\mathbb{1}_{\{X_j > x+\varepsilon\}}}{j}.\end{aligned}$$

It remains to note that almost surely the series with general term $\mathbb{1}_{\{X_j > x+\varepsilon\}}/j$ diverges, since the associated centred series converges but the series of expectations does not. In particular, it diverges on $\{Y \leqslant x\}$, implying that this set has probability zero. □

We now consider the number of candidates that must be interviewed to form a team of a given size. The following result provides an answer under an assumption on the convergence of the sequence $(\overline{Y}_n)_{n\geqslant 1}$ that is fairly often satisfied in practice.

Theorem 10.7 (Asymptotic behaviour of the number interviewed)**.** *Let the sequence of random variables* $(P_n)_{n\geqslant 1}$ *be defined by* $P_1 = 1$ *and, for* $n \geqslant 2$, $P_n = 1 - F(\overline{Y}_{n-1})$. *If there exist a real* $\alpha > 0$ *and a random variable* W *such that* $\mathbb{P}(0 < W < \infty) = 1$ *and* $n^\alpha P_n \xrightarrow[n\to\infty]{\text{p.s.}} W$, *then*

$$\frac{T_n}{n^{\alpha+1}} \xrightarrow[n\to\infty]{\text{p.s.}} \frac{1}{(\alpha+1)W}.$$

Proof. The idea is to apply Theorem 10.3 conditionally on the sigma-field $\overline{\mathcal{F}}$ generated by the sequence $(\overline{Y}_n)_{n\geqslant 1}$, setting, for $n \geqslant 1$,

$$b_n = \sum_{i=1}^{n} \frac{1}{P_i} \quad \text{et} \quad Q_n = T_n - T_{n-1} - \frac{1}{P_n} \quad \text{avec} \quad T_0 = 0.$$

The sequence $(b_n)_{n\geqslant 1}$ is strictly increasing and tends to $+\infty$, since $1/P_i$ is of order i^α/W. More precisely, a sum-integral comparison gives

$$b_n \sim \frac{n^{\alpha+1}}{(\alpha+1)W}.$$

Moreover, $T_n - T_{n-1}$ has a geometric distribution with parameter P_n, so

$$\mathbb{E}\left(\frac{Q_n^2}{b_n^2} \,\middle|\, \overline{\mathcal{F}}\right) = \frac{1-P_n}{b_n^2 P_n^2} \underset{n\to\infty}{\sim} \frac{(\alpha+1)^2}{n^2}$$

which is the general term of a convergent series. The centred series theorem (Theorem 10.3) therefore ensures that

$$\frac{1}{b_n}\sum_{k=1}^{n} Q_k = \frac{T_n}{b_n} - 1 \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

The asymptotic equivalent of b_n gives the claimed result. □

10.3 The exponential distribution case

Theorem 10.8 (Evolution of the average and team size)**.** *Suppose the random variables* $(X_n)_{n\geqslant 1}$ *have an exponential distribution with parameter* λ. *Then there exists a random variable* G *with a Gumbel distribution, with distribution function* $x \mapsto \exp(-e^{-x})$, *such that*

$$\lambda\overline{Y}_n - \log(n) \xrightarrow[n\to\infty]{L^2} G, \quad \lambda\overline{Y}_n - \log(n) \xrightarrow[n\to\infty]{\text{p.s.}} G, \quad \textit{et} \quad \frac{T_n}{n^2} \xrightarrow[n\to\infty]{\text{p.s.}} \frac{e^G}{2}.$$

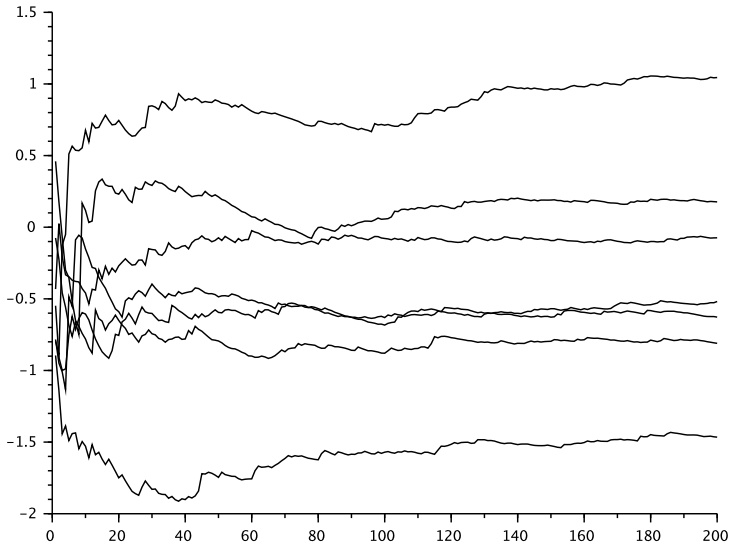


Fig. 10.1. Almost sure convergence of the sequence $(\overline{Y}_n - \log(n))_{n\geqslant 2}$.

Proof. By considering the sequence $(\lambda Y_n)_{n\geqslant 1}$, we may assume $\lambda = 1$. With the convention $\overline{Y}_0 = 0$, set, for every $n \geqslant 0$,

$$\eta_n = Y_{n+1} - \overline{Y}_n = (n+1)(\overline{Y}_{n+1} - \overline{Y}_n).$$

Thus $\overline{Y}_n = \sum_{j=1}^{n} \eta_{j-1}/j$. Write $\mathcal{F}_n$ for the sigma-field generated by $(Y_1, \ldots, Y_n)$. For every $n \geqslant 1$, the conditional distribution of η_n given $\mathcal{F}_n$ equals that of $Y_{n+1} - \overline{Y}_n$ given $\overline{Y}_n$ and $\{Y_{n+1} > \overline{Y}_n\}$; hence, by the memoryless property of exponential distributions,

$$\mathbb{P}(\eta_n > x \,|\, \mathcal{F}_n) = \mathbb{P}(Y_{n+1} > \overline{Y}_n + x \,|\, \overline{Y}_n, Y_{n+1} > \overline{Y}_n) = e^{-x}.$$

The random variables $(\eta_n)_{n\geqslant 0}$ are therefore independent and identically distributed, with exponential distribution of parameter 1. With $h_n = \sum_{k=1}^{n} \frac{1}{k}$, we have

$$\overline{Y}_n - h_n = \sum_{k=1}^{n} \frac{\eta_{k-1} - 1}{k}.$$

The sequence $(\overline{Y}_n - h_n)_{n\geqslant 1}$ is a martingale (partial sums of independent integrable centred random variables). It is bounded in L^2 since

$$\mathbb{E}((\overline{Y}_n - h_n)^2) = \sum_{k=1}^{n} \frac{\mathbb{E}((\eta_{k-1} - 1)^2)}{k^2} \leqslant \mathbb{E}((\eta_0 - 1)^2)\frac{\pi^2}{6} < \infty.$$

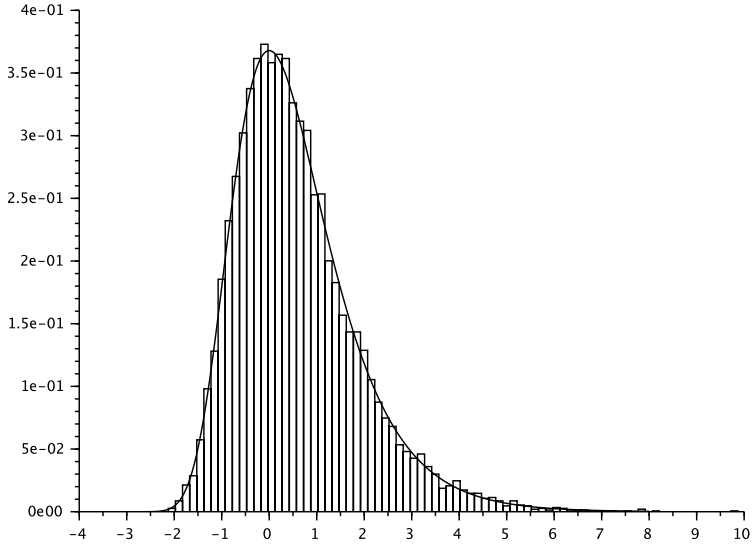


Fig. 10.2. Histogram of the almost sure limit of the random sequence $(\overline{Y}_n - \log n)_n$ and density of the Gumbel distribution.

There therefore exists a random variable $\widetilde{G}$ such that $\overline{Y}_n - h_n$ converges to $\widetilde{G}$ in L^2 and almost surely. We now show that $(\overline{Y}_n - \log(n))_{n\geqslant 1}$ converges in distribution to the Gumbel distribution. It will follow that $\widetilde{G} = G - \gamma$, where G has the Gumbel distribution and $\gamma = \lim_{n\to\infty}(h_n - \log(n))$ is Euler's constant.

The Laplace transform of $\overline{Y}_n$, denoted by H_n, is, for every $\theta \geqslant 0$,

$$H_n(\theta) = \mathbb{E}\Big(e^{-\theta\overline{Y}_n}\Big) = \prod_{k=1}^{n} \frac{k}{k+\theta} = n!\left(\prod_{k=1}^{n}(k+\theta)\right)^{-1}.$$

Now let $(Z_n)_{n\geqslant 1}$ be a sequence of i.i.d. real random variables with exponential distribution of parameter 1. For every $n \geqslant 1$, let $Z_{(n)} = \max(Z_1, \ldots, Z_n)$. We compute the Laplace transform L_n of Z_n. The density of $Z_{(n)}$ is $ne^{-z}(1 - e^{-z})^{n-1}\mathbb{1}_{[0,+\infty)}(z)$, since $\mathbb{P}(Z_{(n)} \leqslant z) = \mathbb{P}(Z_1 \leqslant z)\cdots\mathbb{P}(Z_n \leqslant z) = (1 - e^{-z})^n\mathbb{1}_{[0,+\infty)}(z)$. Integration by parts gives

$$\begin{aligned} L_n(\theta) &:= \mathbb{E}\big(e^{-\theta Z_{(n)}}\big) \\ &= \int_0^\infty e^{-\theta z} n e^{-z}\big(1 - e^{-z}\big)^{n-1}\, dz \end{aligned}$$

$$= \frac{n}{1+\theta} \int_0^\infty e^{-\theta z}(n-1)e^{-2z}\big(1-e^{-z}\big)^{n-2}\,dz$$
$$= \frac{n}{1+\theta} L_{n-1}(\theta+1),$$

and a straightforward induction gives

$$L_n(\theta) = \prod_{k=1}^{n} \frac{n+1-k}{k+\theta} = n!\left(\prod_{k=1}^{n}(k+\theta)\right)^{-1}.$$

Hence $\overline{Y}_n$ and $Z_{(n)}$ have the same distribution. In fact, this is only part of Rényi's lemma 11.3. Thus, for every real x,

$$\mathbb{P}(\overline{Y}_n - \log(n) \leqslant x) = \mathbb{P}(Z_{(n)} - \log(n) \leqslant x) = \left(1 - \frac{e^{-x}}{n}\right)^n \underset{n\to\infty}{\longrightarrow} e^{-e^{-x}},$$

which completes the proof of the first two statements. For the last one, we have

$$nP_n = n(1 - F(\overline{Y}_{n-1})) = e^{-(\overline{Y}_n - \log(n))} \underset{n\to\infty}{\xrightarrow{\text{p.s.}}} e^{-G},$$

which gives $T_n/n^2 \underset{n\to\infty}{\xrightarrow{\text{p.s.}}} e^G/2$ by Theorem 10.7. □

10.4 The Pareto distribution case

Theorem 10.9 (Evolution of the average)**.** *Suppose the random variables $(X_n)_{n\geqslant 1}$ have the Pareto distribution with parameter (θ, β), where $\theta > 0$ and $\beta > 1$, that is, they have the following distribution function:*

$$F(x) = \left[1 - \left(\frac{\theta}{x}\right)^\beta\right] \mathbb{1}_{[\theta,+\infty[}(x).$$

Then there exists a random variable Y such that $\mathbb{P}(0 < Y < \infty) = 1$ and

$$\frac{\overline{Y}_n}{n^{1/(\beta-1)}} \underset{n\to\infty}{\xrightarrow{\text{p.s.}}} Y.$$

Proof. If X has the Pareto distribution with parameter (θ, β), then X/θ has the Pareto distribution with parameter $(1, \beta)$. We therefore assume $\theta = 1$ in the remainder of the proof. Moreover, if X has the Pareto distribution with parameter $(1, \beta)$, then the distribution of X given $X > c$ is that of cX. Writing $\mathcal{F}_n$ for the sigma-field generated by $(Y_1, \ldots, Y_n)$, we have, for every $x > 1$,

$$\mathbb{P}(Y_n > x\overline{Y}_{n-1} \,|\, \mathcal{F}_{n-1}) = \mathbb{P}(X > x\overline{Y}_{n-1} \,|\, X > \overline{Y}_{n-1}, \overline{Y}_{n-1}) = x^{-\beta}.$$

where X is a random variable with Pareto distribution of parameter β, independent of $\mathcal{F}_{n-1}$. In other words, the distribution of $Y_n/\overline{Y}_{n-1}$ given $\mathcal{F}_{n-1}$ is Pareto with parameter β. Consequently,

$$\mathbb{E}(\overline{Y}_n \,|\, \mathcal{F}_{n-1}) = \frac{(n-1)\overline{Y}_{n-1} + \mathbb{E}(Y_n \,|\, \overline{Y}_{n-1})}{n} = \frac{n-1+\mathbb{E}(X)}{n}\overline{Y}_{n-1}.$$

Set, for $n \geqslant 1$,

$$a_n = \frac{n-1+\mathbb{E}(X)}{n} = 1 + \frac{1}{n(\beta-1)} \quad \text{et} \quad b_n = \frac{1}{\prod_{j=1}^n a_j}.$$

The sequence $(M_n)_{n\geqslant 1}$ defined by $M_n = b_n\overline{Y}_n$ is a nonnegative martingale with expectation 1. It therefore converges almost surely to a nonnegative random variable M in L^1. We omit the fairly technical proof that $\mathbb{P}(0 < M < \infty) = 1$. Finally, note that

$$b_n = \exp\Big(-\sum_{j=1}^n \log\Big(1 + \frac{1}{(\beta-1)j}\Big)\Big) = \exp\Big(-\frac{1}{\beta-1}\sum_{j=1}^n \frac{1}{j} + o(1)\Big).$$

In other words, $n^{1/(\beta-1)}b_n$ converges to a strictly positive real number. □

Corollary 10.10 (Recruitment time)**.** *There exists a random variable T taking values in $(0,\infty)$ such that*

$$\frac{T_n}{n^{(2\beta-1)/(\beta-1)}} \xrightarrow[n\to\infty]{\text{p.s.}} T.$$

Proof. For every $n \geqslant 1$, we have $P_n = (\overline{Y}_{n-1})^{-\beta}$. The preceding result ensures that $n^{\beta/(\beta-1)}P_n$ converges almost surely to a random variable W satisfying $\mathbb{P}(0 < W < \infty) = 1$. We conclude using Theorem 10.7. □

10.5 Further reading

One may also study the number M_n of candidates hired after n interviews and the average A_n performance of the candidates hired after n interviews.

This chapter is mainly based on an article by John Preater [Pre00] for the exponential distribution case, and an article by Abba Krieger, Moshe Pollak and Ester Samuel-Cahn [KPSC08] for the general case. Proofs of the centred series theorem (Theorem 10.3) and the Lindeberg-Lévy theorem (Theorem 10.4) can be found, for example, in William Feller's indispensable book [Fel68].

Theorem 10.2 can be supplemented with non-asymptotic deviation inequalities, in contrast to the convergence-in-distribution theorem. Since the random variables $(R_n)_{n\geqslant 1}$ take values in $[0,1]$, (Wassily) Hoeffding's inequality ensures that, for every $r > 0$,

$$\mathbb{P}\left(Z_n - \sum_{k=1}^n \frac{1}{k} \geqslant r\right) \leqslant \exp\left(-\frac{2r^2}{n}\right).$$

This inequality is not very sharp here, since it does not adequately account for the fact that the variances of the increments of $(Z_n)_{n\geqslant 1}$ tend to 0. A finer lower-tail result follows from (Andreas) Maurer's inequality for sums of nonnegative random variables [Mau03], which gives, for every $r > 0$,

$$\mathbb{P}\left(Z_n - \sum_{k=1}^{n} \frac{1}{k} \leqslant -r\right) \leqslant \exp\left(-\frac{r^2}{2\log(n)}\right).$$

Finally, we mention the famous *secretary problem*[1] which is delightfully described in an article by Thomas Ferguson [Fer89]. The setting is the same as in this chapter, but only one person is hired from n candidates. The question is to find the strategy that selects the best candidate with maximal probability. The first step is to show that the optimal strategy necessarily has the following form: fix r between 1 and n, reject the first $r-1$ candidates, then select the first candidate whose score exceeds all their predecessors'. Under this strategy, the best candidate is hired at rank j (between r and n) if and only if the best candidate arrives in position j and the candidates numbered r to $j-1$ are worse than one of the first $r-1$. Thus,

$$\mathbb{P}(j \text{ est le meilleur et il est sélectionné}) = \frac{1}{n}\frac{r-1}{j-1}$$

and the probability of hiring the best candidate with this strategy is

$$p_n(r) = \sum_{j=r}^{n} \frac{1}{n}\frac{r-1}{j-1} = \frac{r-1}{n}\sum_{j=r}^{n}\frac{1}{j-1}.$$

For $x \in]0,1[$,

$$p_n(\lfloor xn \rfloor) = \frac{\lfloor xn \rfloor}{n}\sum_{j=\lfloor xn \rfloor}^{n}\frac{1}{j-1} \underset{n\to\infty}{\longrightarrow} -x\log(x).$$

The maximal probability is therefore $1/e$ for an optimal choice of $r \approx n/e$.

[1] This optimal stopping problem is a classic of stochastic control and decision theory. It is also known, among other names, as the marriage problem.

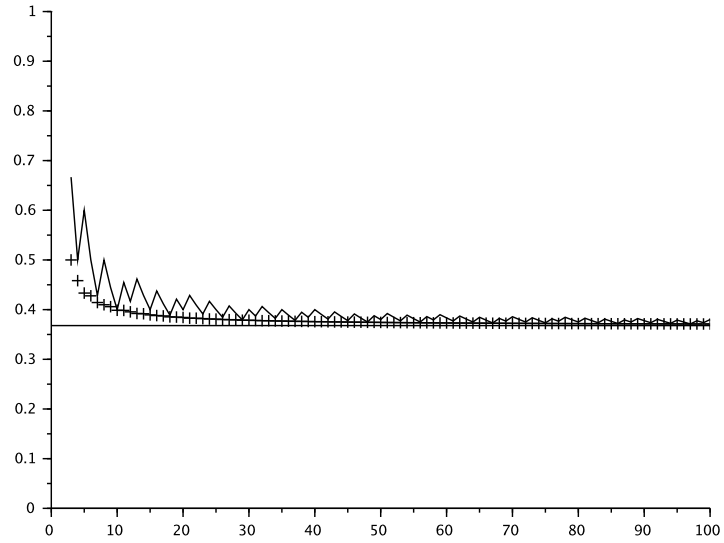


Fig. 10.3. Convergence of the optimal probability of choosing the best candidate (crosses) and the ideal proportion (solid line) to $1/e$.

11

M/M/Infinity queue

Keywords. Exponential distribution; binomial distribution; Poisson distribution; queue; invariant distribution.

Tools. Continuous-time Markov chain; Poisson process; coupling; Wasserstein distance; semigroup; infinitesimal generator; generating function; Markov's inequality.

Difficulty. **

Queues[1] are among the most widespread and useful random models. Perhaps the simplest example is the following: customers queue in front of a counter called a server. The times separating successive customer arrivals are modeled by i.i.d. real-valued random variables with exponential distribution of parameter λ, while successive service times are modeled by i.i.d. real-valued random variables with exponential distribution of parameter μ. The exponential distribution is justified by its memoryless property, which suits many practical situations. We study the number X_t of customers in the queue at time t. The process $(X_t)_{t\geqslant 0}$ is a continuous-time Markov chain with state space $\mathbb{N}$. In Kendall's notation, this is an M/M/1 queue with rates λ and μ: the first M indicates memorylessness[2] of interarrival times, the second M memorylessness of service times, and the final 1 a single server. More generally, M/M/s queues can be defined for any integer s. More generally still, M/M/s/K includes a maximum queue size K, modeling a waiting room with limited capacity. An M/M/K/K queue models, for example, a parking lot with K spaces. Additional phenomena such as impatience, priority, and service discipline[3] can be included. Queues can also be arranged in tandem or in networks. Finally, it is sometimes more realistic to let arrival rates λ and service rates μ vary over time, naturally leading to point processes.

[1] Called "files d'attente" in French.
[2] M stands for "*Memoryless*".
[3] The standard service discipline is FIFO, for "*First In First Out*".

This chapter treats the M/M/∞ queue, arguably the simplest queue to study and a discrete-state analogue of the Ornstein–Uhlenbeck process. In the M/M/∞ model, customer interarrival times are i.i.d. exponential with parameter λ. Each customer has a dedicated server and begins service immediately upon arrival. Service times are i.i.d. exponential with parameter μ, independent of the arrival process. A special feature of the M/M/∞ queue is that the number of customers in the queue always equals the number currently being served.

The M/M/∞ queue is a simplified version of M/M/K/K when λ/μ is negligible compared with K. It models, for example, the number of cars in a very large parking lot: cars are customers, parking spaces are servers, and a car's service time is its parking duration. Many practical situations can be modeled by an M/M/∞ queue.

To construct the process $(X_t)_{t\geqslant 0}$ rigorously, first assume that the queue is initially empty: $X_0 = 0$. Customer interarrival times are modeled by a sequence $(E_n)_{n\geqslant 0}$ of i.i.d. real-valued random variables exponential with parameter λ. Customer number n^{e} thus arrives at time $T_n := E_1 + \cdots + E_n$, which has distribution Gamma(n, λ). For every real $t \geqslant 0$, the number of customers arriving in the time interval $[0, t]$ is

$$N_t := \sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \leqslant t\}}.$$

The counting process $(N_t)_{t\geqslant 0}$ is a Poisson process starting from 0 with intensity λ. For every integer $n \geqslant 0$, denote by S_n the service time of customer n^{e}, who leaves the queue at time $E_n + S_n$. The sequence $(S_n)_{n\geqslant 0}$ consists of i.i.d. real-valued random variables exponential with parameter μ. Assume that the random variables X_0, $(S_n)_{n\geqslant 0}$, and $(E_n)_{n\geqslant 0}$ are independent. For every $t \in \mathbb{R}_+$, the number of customers in the queue at time t is

$$\sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \leqslant t < T_n + S_n\}}.$$

In the general case where X_0 need not vanish, introduce a sequence $(S'_n)_{n\geqslant 0}$ of i.i.d. real-valued random variables exponential with parameter μ, independent of X_0, and model the service times of the initial customers by $S'_0, \ldots, S'_{X_0}$. Assume that X_0, $(S_n)_{n\geqslant 0}$, $(S'_n)_{n\geqslant 0}$, and $(E_n)_{n\geqslant 0}$ are independent. For every real $t \geqslant 0$, the number of customers in the queue at time t is therefore

$$X_t := \sum_{n=0}^{X_0} \mathbb{1}_{\{S'_n > t\}} + \sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \leqslant t < T_n + S_n\}}.$$

11.1 Exponential distributions

The following lemmas are useful far beyond queueing theory. In particular, they play a role in the structure of Markov processes.

Lemma 11.1 (Memorylessness)**.** *If E_1 and E_2 are independent exponential random variables with respective parameters λ_1 and λ_2, then*

$$\mathrm{Loi}(E_2 - E_1 \,|\, E_2 > E_1) = \mathrm{Exp}(\lambda_2) = \mathrm{Loi}(E_2)$$

that is, $\mathbb{P}(E_2 - E_1 > t \,|\, E_2 > E_1) = e^{-\lambda_2 t}$ for every $t \geqslant 0$.

Proof. For every $t \geqslant 0$, we have $\mathbb{P}(E_2 > E_1 + t \,|\, E_2 > E_1) = e^{-\lambda_2 t}$ because

$$\mathbb{P}(E_2 > E_1) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

while $\mathbb{P}(E_2 > E_1 + t, E_2 > E_1) = \mathbb{P}(E_2 > E_1 + t)$ equals

$$\iint_{\substack{v > u+t \\ u > 0}} \lambda_1 e^{-\lambda_1 u} \lambda_2 e^{-\lambda_2 v}\, du dv = \frac{\lambda_1}{\lambda_1 + \lambda_2} e^{-\lambda_2 t}.$$

□

Lemma 11.2 (Competing exponential clocks)**.** *If $(E_i)_{i \in \mathcal{I}}$ is a finite or countably infinite family of independent real-valued random variables with exponential distributions of respective parameters $(\lambda_i)_{i \in \mathcal{I}}$, with $\lambda := \sum_{i \in \mathcal{I}} \lambda_i < \infty$, then the random variable $M := \inf_{i \in \mathcal{I}} E_i$ has exponential distribution $\mathrm{Exp}(\lambda)$. Moreover, almost surely the infimum is attained at a unique random integer I, independent of M, whose distribution is given by $\mathbb{P}(I = i) = \lambda_i / \lambda$ for every $i \in \mathcal{I}$.*

Proof of Lemma 11.2. For every $i \in \mathcal{I}$ and every $x \in \mathbb{R}_+$,

$$\mathbb{P}(\min_{k \neq i}(E_k) > E_i \geqslant x) = \lambda_i \int_x^\infty \prod_{k \neq i} \mathbb{P}(E_k > y)\, e^{-\lambda_i y}\, dy = \frac{\lambda_i}{\lambda} e^{-\lambda x}.$$

This is the formula for $\mathbb{P}(I = i, M \geqslant x)$, giving the distribution of the pair (I, M). □

Lemma 11.3 (Rényi's lemma on exponential order statistics)**.** *If $E_1, \ldots, E_n$ are i.i.d. real-valued random variables with exponential distribution $\mathrm{Exp}(\lambda)$ and*

$$\min(E_1, \ldots, E_n) = E_{(1)} < \cdots < E_{(n)} = \max(E_1, \ldots, E_n)$$

is the increasing rearrangement (order statistics) of $E_1, \ldots, E_n$, then

$$(E_{(1)}, \ldots, E_{(n)}) \overset{\text{loi}}{=} \left(\frac{E_1}{n}, \cdots, \frac{E_1}{n} + \cdots + \frac{E_n}{1} \right).$$

In particular, for every $1 \leqslant k \leqslant n$, with the convention $E_0 := 0$,

$$E_{(k)} \overset{\text{loi}}{=} \frac{E_1}{n} + \cdots + \frac{E_k}{n-k+1} = E_{(k-1)} + \frac{E_k}{n-k+1}.$$

Proof. To see this, for every bounded measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}$, changes of variables give

$$\begin{aligned}\mathbb{E}(\varphi(E_{(1)},\dots,E_{(n)})) &= \int_{[0,\infty)^n} \varphi(x_{(1)},\dots,x_{(n)})\,\lambda^n e^{-\lambda(x_1+\cdots+x_n)}\,dx_1\cdots dx_n \\ &= n!\int_{0\leqslant y_1<\cdots<y_n} \varphi(y_1,\dots,y_n)\,\lambda^n e^{-\lambda(y_1+\cdots+y_n)}\,dy_1\cdots dy_n \\ &= \int_{[0,\infty)^n} \varphi(z_1,\dots,z_1+\cdots+z_n)\,n!\lambda^n e^{-\lambda(nz_1+\cdots+1z_n)}\,dz,\end{aligned}$$

where $dz := dz_1\cdots dz_n$. This means that the random variables

$$E_{(1)}, E_{(2)}-E_{(1)}, \dots, E_{(n)}-E_{(n-1)}$$

are independent with $E_{(k)}-E_{(k-1)} \sim \mathrm{Exp}((n-k+1)\lambda)$, $1 \leqslant k \leqslant n$. □

Note that the distributional identity for the minimum

$$\min(E_1,\dots,E_n) = E_{(1)} \stackrel{\text{loi}}{=} \frac{1}{n}E_1$$

is a special case of Lemma 11.2 with $I=\{1,\dots,n\}$ and $\lambda_1=\cdots=\lambda_n$. The identity for the maximum

$$\max(E_1,\dots,E_n) = E_{(n)} \stackrel{\text{loi}}{=} \frac{E_1}{n}+\cdots+\frac{E_n}{1}$$

is proved differently in Theorem 10.8 (using Laplace transforms) and Theorem 13.4 (through the density, by induction on n). It also appears in Chapters 1, 9, and 17 (Yule process). We use it in the next section for the half-life.

Lemma 11.4 (Stability under geometric summation)**.** *Let $G, E_1, E_2, \dots$ be independent random variables, with G geometric of parameter $p \in]0,1]$ and $E_1, E_2, \dots$ exponential of parameter λ. Then the random sum $E_1+\cdots+E_G$ is exponentially distributed with parameter $p\lambda$.*

Since $E_1+\cdots+E_n \sim \mathrm{Gamma}(n,\lambda)$, a geometric mixture of gamma distributions with the same scale parameter is therefore exponential.

Proof. For every $t \in \mathbb{R}$ and every $k \geqslant 1$, we have $\mathbb{E}(e^{itE_k}) = \lambda/(\lambda-it)$ and

$$\begin{aligned}\mathbb{E}(e^{it(E_1+\cdots+E_G)}) &= \sum_{n=1}^{\infty}\mathbb{E}(e^{it(E_1+\cdots+E_n)})\mathbb{P}(G=n) \\ &= \sum_{n=1}^{\infty}\mathbb{E}(e^{itE_1})^n(1-p)^{n-1}p \\ &= \frac{p\mathbb{E}(e^{itE_1})}{1-\mathbb{E}(e^{itE_1})(1-p)}\end{aligned}$$

$$= \frac{\lambda p}{\lambda p - it}.$$

□

Lemma 11.4 says that accepting each customer of the Poisson arrival process (with parameter λ)) with probability p produces a "thinned" Poisson process of intensity $p\lambda$.

The following lemma completes our short catalogue of fundamental exponential-distribution properties. For example, it implies that the sample paths of the Poisson process $(N_t)_{t\geqslant 0}$ are well defined (they do not explode).

Lemma 11.5 (Zero-one law)**.** *If $(E_n)_{n\geqslant 1}$ are independent real-valued random variables with exponential distributions of respective parameters $(\lambda_n)_{n\geqslant 1}$, with $0 < \lambda_n < \infty$ for every $n \geqslant 1$, then the event $\{\sum_{n=1}^{\infty} E_n = \infty\}$ has probability 0 or 1 and*

$$\mathbb{P}\Big(\sum_{n=1}^{\infty} E_n = \infty\Big) = 1 \quad \textit{si et seulement si} \quad \sum_{n=1}^{\infty} \frac{1}{\lambda_n} = \infty.$$

The series $\sum_{n=1}^{\infty} E_n$ diverges a.s. if and only if it diverges in mean.

Proof. The variable $T_\infty = \sum_{n=1}^{\infty} E_n$ takes values in $\mathbb{R}_+ \cup \{\infty\}$. Monotone convergence gives $\mathbb{E}(T_\infty) = \sum_{n=1}^{\infty} 1/\lambda_n$ in $\mathbb{R}_+ \cup \{\infty\}$. Consequently, if $\sum_{n=1}^{\infty} 1/\lambda_n < \infty$, then $\mathbb{E}(T_\infty) < \infty$ and hence $\mathbb{P}(T_\infty = \infty) = 0$. Independence has not been used. Conversely, monotone convergence and independence give

$$\mathbb{E}\big[e^{-T_\infty}\big] = \prod_{n=1}^{\infty} \mathbb{E}\big[e^{-E_n}\big] = \prod_{n=1}^{\infty} \left(1 + \frac{1}{\lambda_n}\right)^{-1}.$$

Now $\sum_{n=1}^{\infty} 1/\lambda_n = \infty$ if and only if $\prod_{n=1}^{\infty}(1 + 1/\lambda_n)^{-1} = 0$, since for every sequence $(a_n)_{n\geqslant 1}$ of nonnegative real numbers, the product $\prod_{n=1}^{\infty}(1 + a_n)$ converges if and only if the series $\sum_{n=1}^{\infty} a_n$ converges. Thus, if $\sum_{n=1}^{\infty} 1/\lambda_n = \infty$, then $\mathbb{E}(\exp(-T_\infty)) = 0$, hence $\mathbb{P}(T_\infty = \infty) = 1$. □

11.2 Half-life of the initial customers

If $\mu = 0$, the process $(X_t)_{t\geqslant 0}$ coincides with the Poisson process $(X_0 + N_t)_{t\geqslant 0}$ starting from X_0 with intensity λ, where $(N_t)_{t\geqslant 0}$ is the Poisson arrival process, hence $X_t \to +\infty$ a.s. as $t \to \infty$ if $X_0 > 0$ and $\lambda > 0$.

In this section, consider instead the case $\lambda = 0$ and $X_0 = N$. The queue contains N customers at time $t = 0$, and no new customers arrive. The process $(X_t)_{t\geqslant 0}$ is then a pure death process, and X_t has distribution $\mathrm{Bin}(N, e^{-\mu t})$, hence $X_t \to 0$ a.s. as $t \to \infty$ if $\mu > 0$ or $X_0 = 0$. In the introductory notation, the service times of the N initial customers are $S'_1, \ldots, S'_N$. Successive service completion times are denoted by $(T_n^{(N)})_{1\leqslant n\leqslant N}$:

$$\min(S'_1,\dots,S'_n) = T_1^{(N)} \leqslant \cdots \leqslant T_N^{(N)} = \max(S'_1,\dots,S'_n).$$

We have $X_t = 0$ if $t > T_N^{(N)}$. Rényi's lemma 11.3 states that the real-valued random variables

$$(\tau_n^{(N)})_{1\leqslant n\leqslant N} \quad \text{où} \quad \tau_n^{(N)} := T_n^{(N)} - T_{n-1}^{(N)} \quad \text{et} \quad T_0^{(N)} := 0$$

are independent and exponential with respective parameters

$$((N-n+1)\mu)_{1\leqslant n\leqslant N}.$$

We study the half-life $T_{\lfloor N/2\rfloor}^{(N)}$, the time by which half of the N initial customers have left the queue.

Theorem 11.6 (Convergence of the half-life)**.** *We have*

$$T_{\lfloor N/2\rfloor}^{(N)} \xrightarrow[N\to\infty]{\text{p.s.}} \frac{\log(2)}{\mu}.$$

Proof. For simplicity, omit the superscript "(N)", assume that N is even, and set $M := N/2$. First observe that

$$\begin{aligned}\mathbb{E}\big(T_{N/2}\big) &= \sum_{n=1}^{N/2}\mathbb{E}(\tau_n) = \frac{1}{\mu}\sum_{n=1}^{N/2}\frac{1}{N-n+1}\\ &= \frac{1}{\mu}(H(N)-H(N/2)) = \frac{\log(2)}{\mu}+o(1),\end{aligned}$$

where $H(n)$ is the harmonic series. Moreover,

$$T_M - \mathbb{E}(T_M) = \sum_{n=1}^{M}(\tau_n - \mathbb{E}(\tau_n)),$$

where the variables $(\tau_n - \mathbb{E}(\tau_n))_{n\geqslant 1}$ are independent and centered, and $\tau_n - \mathbb{E}(\tau_n)$ has the same distribution as $(E-1)/\mu(N+1-n)$, where E is exponential of parameter 1. To establish convergence, use Markov's inequality with a sufficiently high-order moment. The moment of order 4 is controlled as follows:

$$\begin{aligned}\mathbb{E}\Big[(T_M-\mathbb{E}(T_M))^4\Big] &= \sum_{n=1}^{M}\mathbb{E}\big[(\tau_n-\mathbb{E}(\tau_n))^4\big]\\ &\quad + 6\sum_{1\leqslant m\neq n\leqslant M}\mathbb{E}\big[(\tau_n-\mathbb{E}(\tau_n))^2\big]\mathbb{E}\big[(\tau_m-\mathbb{E}(\tau_m))^2\big]\\ &= \frac{\mathbb{E}\big[(E-1)^4\big]}{\mu^4}\sum_{n=M+1}^{N}\frac{1}{n^4}\end{aligned}$$

$$+ 6\frac{\left(\mathbb{E}\left[(E-1)^2\right]\right)^2}{\mu^4} \sum_{1+M\leqslant m\neq n\leqslant N} \frac{1}{n^2}\frac{1}{m^2}$$
$$\leqslant \frac{9}{\mu^4}\sum_{n=M+1}^{N}\frac{1}{n^4} + \frac{6}{\mu^4}\left(\sum_{n=M+1}^{N}\frac{1}{n^2}\right)^2,$$

since $\mathbb{E}((E-1)^2) = 1$ and $\mathbb{E}((E-1)^4) = 9$. Thus

$$\mathbb{E}\Big[\big(T_M - \mathbb{E}(T_M)\big)^4\Big] = \mathcal{O}\big(N^{-2}\big).$$

Markov's inequality then gives, for every $r \geqslant 0$,

$$\mathbb{P}(|T_M - \mathbb{E}(T_M)| \geqslant r) \leqslant \frac{C}{r^4 N^2}.$$

Finally, the Borel–Cantelli lemma gives $T_M \xrightarrow[N\to\infty]{\text{p.s.}} \log(2)/\mu$. □

Remark 11.7 (Half-life and radioactivity)**.** *The lifetime of carbon-*14 *atoms is well modeled by an exponential random variable with half-life* 5734 *years. This allows archaeologists to date organic material samples.*

11.3 Distribution of the number of customers

This section computes the distribution of X_t using two different methods.

Theorem 11.8 (Distribution of X_t)**.** *For every $k \in \mathbb{N}$ and every $t \geqslant 0$,*

$$\mathrm{Loi}(X_t \,|\, X_0 = k) = \mathrm{Bin}(k, e^{-\mu t}) * \mathrm{Poi}\left(\frac{\lambda}{\mu}\big(1 - e^{-\mu t}\big)\right)$$

with the convention $\mathrm{Bin}(0,p) = \delta_0$*. Moreover, if* $X_0 \sim \mathrm{Poi}(\alpha)$*, then*

$$X_t \sim \mathrm{Poi}\left(\frac{\lambda}{\mu}\big(1 - e^{-\mu t}\big) + \alpha e^{-\mu t}\right).$$

Proof. Use the introductory notation. By independence,

$$\mathrm{Loi}(X_t \,|\, X_0 = k) = \mathrm{Loi}\left(\sum_{n=0}^{X_0} \mathbb{1}_{\{S'_n > t\}} \,\Big|\, X_0 = k\right) * \mathrm{Loi}\left(\sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \leqslant t < T_n + S_n\}}\right).$$

Again by independence (this time of X_0 and $(S'_n)_{n\geqslant 0}$),

$$\mathrm{Loi}\left(\sum_{n=0}^{X_0} \mathbb{1}_{\{S'_n > t\}} \,\Big|\, X_0 = k\right) = \mathrm{Loi}\left(\sum_{n=0}^{k} \mathbb{1}_{\{S'_n > t\}}\right) = \mathrm{Bin}(k, e^{-\mu t})$$

since $\mathbb{1}_{\{S'_1>t\}}, \dots, \mathbb{1}_{\{S'_k>t\}}$ are i.i.d. Bernoulli random variables with distribution $\mathrm{Ber}(e^{-\mu t})$. It remains to show that

$$\mathrm{Loi}(Y_t) = \mathrm{Poi}\left(\frac{\lambda}{\mu}\big(1-e^{-\mu t}\big)\right) \quad \text{où} \quad Y_t := \sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \leqslant t < T_n + S_n\}}.$$

Recall that $(N_t)_{t\geqslant 0}$ is the Poisson process counting the times $(T_n)_{n\geqslant 1}$. For $t > 0$, the random variable $N_t = \mathrm{card}\{n \geqslant 1 : T_n \in [0,t]\}$ has Poisson distribution of parameter λt. Moreover, conditional on $N_t = n$, the n jump times have the distribution of a sorted sample of n i.i.d. real-valued random variables uniform on $[0,t]$. A customer arriving at a random time uniform on $[0,t]$ is still in the queue at time t with probability

$$q(t) := \frac{1}{t}\int_0^t (1-F_\mu(s))\,ds = \frac{1}{\mu t}(1-e^{-\mu t}),$$

where F_μ is the distribution function of $\mathrm{Exp}(\mu)$. Independence thus gives, for every $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{P}(Y_t = k) &= \sum_{n=k}^{\infty} \mathbb{P}(Y_t = k \,|\, N_t = n)\mathbb{P}(N_t = n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} q(t)^k (1-q(t))^{n-k} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\
&= e^{-\lambda t q(t)} \frac{(\lambda t q(t))^k}{k!},
\end{aligned}$$

showing that Y_t has Poisson distribution of parameter $\lambda(1-e^{-\mu t})/\mu$. □

Remark 11.9 (Arbitrary service distribution and the M/G/∞ queue)**.** *If service times are i.i.d. with an arbitrary distribution function F, the proof of Theorem* (11.8) *shows that the conditional distribution of X_t given $X_0 = 0$ is Poisson with parameter*

$$\lambda \int_0^t (1-F(s))\,ds.$$

*This is called an M/G/∞ queue, where G stands for "*General*".*

Theorem 11.8 can also be obtained using the Chapman–Kolmogorov equations. This approach applies to continuous-time Markov chains far beyond the particular case studied here.

Theorem 11.10 (Chapman–Kolmogorov equations)**.** *For all $t \in \mathbb{R}_+$ and $n \in \mathbb{N}$, let $p_n(t) := \mathbb{P}(X_t = n)$. The functions $(p_n)_{n\in\mathbb{N}}$ are of class $\mathcal{C}^1$ and satisfy the infinite system of linear differential equations*

$$\forall n \in \mathbb{N}, \quad p'_n(t) = -(\lambda + n\mu)p_n(t) + \lambda p_{n-1}(t) + (n+1)\mu p_{n+1}(t),$$

with $p_{-1}(t) := 0$ for every $t \geqslant 0$, for a given initial distribution $(p_n(0))_{n\in\mathbb{N}}$.

Proof. To compute $p_n(t+h)$, observe that if $X_{t+h} = n$, one of the following mutually exclusive conditions holds:

1. $X_u = n$ for every $u \in [t, t+h]$;
2. $X_t = n-1$ and exactly one transition $(n-1 \to n)$ occurs in the time interval $[t, t+h]$;
3. $X_t = n+1$ and exactly one transition $(n+1 \to n)$ occurs in the time interval $[t, t+h]$;
4. at least two transitions occur in the time interval $[t, t+h]$.

By Lemmas 11.1 and 11.2, the process remains in state n for an exponential time of parameter $\lambda + n\mu$, then jumps from n to $n-1$ or $n+1$ with respective probabilities $n\mu/(\lambda+n\mu)$ and $\lambda/(\lambda+n\mu)$. Conditioning on the event $\{X_t = n\}$ gives

$$p_n(t+h) = p_n(t)\{1 - \lambda h - n\mu h\} + \lambda h p_{n-1}(t) + (n+1)\mu h p_{n+1}(t) + o(h).$$

Letting h tend to 0 yields the desired result. □

The following corollary of Theorem 11.10 states that the generating function of X_t solves a partial differential equation. It therefore provides a new, analytic proof of Theorem 11.8, based on generating functions (or Laplace transforms). For all $t \in \mathbb{R}_+$ and $s \in [-1, 1]$, let

$$G(s,t) := \mathbb{E}\big(s^{X_t}\big) = \sum_{n=0}^{\infty} p_n(t) s^n.$$

Corollary 11.11 (Generating function of X_t)**.** *If G_0 is the generating function of X_0, then for every $t \geqslant 0$ and every $s \in [-1, 1]$,*

$$G(s,t) = G_0(1 - e^{-\mu t} + s e^{-\mu t}) \exp\left(\frac{\lambda}{\mu}(1 - e^{-\mu t})(s-1)\right).$$

Proof. By the Chapman–Kolmogorov equations (Theorem 11.10),

$$\begin{aligned}
\partial_t G(s,t) &= \sum_{n=0}^{\infty} p_n'(t) s^n \\
&= \sum_{n=0}^{\infty} (-(\lambda + n\mu) p_n(t) + \lambda p_{n-1}(t) + (n+1)\mu p_{n+1}(t)) s^n \\
&= -\lambda(1-s) \sum_{n=0}^{\infty} p_n(t) s^n + \mu(1-s) \sum_{n=0}^{\infty} n p_n(t) s^{n-1}.
\end{aligned}$$

The function G therefore solves the partial differential equation

$$\partial_t G(s,t) = (1-s)[\mu \partial_s G(s,t) - \lambda G(s,t)].$$

Set $H = \log(G)$, then change variables to $(s, \tau) = (s, (1-s)e^{-\mu t})$. Writing J for the function such that $H(s,t) = J(s,\tau)$, the function J takes the form

$$J(s,\tau) = \frac{\lambda}{\mu} s + h(\tau),$$

where h remains to be determined. Equivalently, H has the form

$$H(s,t) = \frac{\lambda}{\mu} s + h((1-s)e^{-\mu t}).$$

Conclude by observing that h is characterized by $H(s,0)$ equaling $\log(G_0(s))$ for every $s \in [-1,1]$. □

The process $(X_t)_{t \geqslant 0}$ is a continuous-time Markov chain. Its Markov semigroup is the family of operators $(P_t)_{t \geqslant 0}$ defined as follows: for all $t \geqslant 0$ and $f : \mathbb{N} \to \mathbb{R}$, $P_t f$ is the function $\mathbb{N} \to \mathbb{R}$ given by

$$P_t f(n) := \mathbb{E}(f(X_t) \,|\, X_0 = n).$$

If ν is a probability measure on $\mathbb{N}$, denote by νP_t the distribution of X_t when X_0 has distribution ν. For every bounded function f,

$$\nu P_t f = \sum_{n \in \mathbb{N}} \nu(n) P_t f(n).$$

By Theorem 11.8, the probability measure $P_t(\cdot)(n)$ is the distribution

$$\mathrm{Bin}(n, e^{-\mu t}) * \mathrm{Poi}\left(\frac{\lambda}{\mu}(1 - e^{-\mu t})\right).$$

Remark 11.12 (Fundamental properties of the Markov semigroup)**.** *The semigroup $(P_t)_{t \geqslant 0}$ satisfies the following fundamental properties:*

1. *Identity: $P_0 = I$, meaning $P_0 f = f$ for every f;*
2. *Positivity and normalization: if $f \geqslant 0$, then $P_t f \geqslant 0$, and $P_t 1 = 1$;*
3. *Semigroup property: for every function f and all $s, t \geqslant 0$,*

$$P_{t+s} f = (P_t \circ P_s) f;$$

4. *Chapman–Kolmogorov equation: for every function f,*

$$\partial_t P_t f = L P_t f = P_t L f;$$

where L is called the infinitesimal generator of X and is defined by

$$Lf(n) := \lambda(f(n+1) - f(n)) + n\mu(f(n-1) - f(n)).$$

Formally, everything behaves as if $P_t = e^{tL}$. The first two properties are immediate. The third follows from the explicit formula for P_t and restates the Markov property. The last follows from 3 and an argument analogous to the proof of Theorem 11.10.

Remark 11.13 (Reading the dynamics from the infinitesimal generator)**.** *Define the transition kernel*

$$Q(n,\cdot) := \frac{\lambda}{\lambda+n\mu}\delta_{n+1} + \frac{n\mu}{\lambda+n\mu}\delta_{n-1}$$

and the holding rate $D(n) := \lambda+n\mu$. The infinitesimal generator then becomes

$$Lf(n) = D(n)\int (f(m)-f(n))\,Q(n,dm).$$

The generator reveals an algorithm for simulating sample paths: conditional on being at n, the process X jumps after an exponential time of parameter $D(n)$ to $n+1$ with probability $\lambda/(\lambda+n\mu)$ and to $n-1$ with probability $n\mu/(\lambda+n\mu)$. Another example appears in Chapter 9, page 114. Note that L can be viewed as a matrix with infinitely many rows and columns by setting $L(n,m) := L\mathbb{1}_n(m)$, so that $Lf(n) = \sum_{m\in\mathbb{N}} L(n,m)f(m)$.

Any reasonable Markov process has an associated semigroup and infinitesimal generator. While the latter is often explicit, the measures $P_t(\cdot)(x)$ rarely are. In this respect, the process studied here is remarkable.

11.4 Long-time behavior

Use the notation

$$\rho := \frac{\lambda}{\mu} \quad \text{et} \quad \pi := \mathrm{Poi}(\rho).$$

If $X_0 \sim \pi$, Theorem 11.8 ensures that $X_t \sim \pi$ for every $t \geqslant 0$. Thus the distribution π is invariant under P_t. Equivalently, $\pi L = 0$. As for all birth-and-death processes, the invariant distribution π, if it exists, is also reversible for L (or P_t): L and P_t are self-adjoint in $L^2(\pi)$. Finally, for any initial distribution (of X_0),

$$X_t \xrightarrow[t\to\infty]{\text{loi}} \pi.$$

We now aim to quantify this convergence in distribution.

Let $\mathcal{P}_1(\mathbb{N})$ be the set of probability measures ν on $\mathbb{N}$ having a moment of order 1, namely those $\nu : \mathbb{N} \to [0,1]$ satisfying $\sum_{x\in\mathbb{N}} \nu(x) = 1$ and $\sum_{x\in\mathbb{N}} x\nu(x) < \infty$. The Wasserstein distance on $\mathcal{P}_1(\mathbb{N})$ is defined, for all $\nu,\tilde{\nu} \in \mathcal{P}_1(\mathbb{N})$, by the variational coupling formula

$$W_1(\nu,\tilde{\nu}) := \inf\left\{\mathbb{E}|X-\tilde{X}| : X\sim\nu,\ \tilde{X}\sim\tilde{\nu}\right\}.$$

Relate this distance to the total variation distance introduced in Chapter 1. For integer-valued random variables, $X \neq \tilde{X}$ implies $|X - \tilde{X}| \geqslant 1$. Thus

$$d_{\mathrm{VT}}(\nu, \tilde{\nu}) \leqslant W_1(\nu, \tilde{\nu}).$$

Theorem 11.14 (Wasserstein contraction)**.** *If $\nu, \tilde{\nu} \in \mathcal{P}_1(\mathbb{N})$, then*

$$W_1(\nu P_t, \tilde{\nu} P_t) \leqslant e^{-\mu t} W_1(\nu, \tilde{\nu}),$$

where, for every $\eta \in \mathcal{P}(\mathbb{N})$, ηP_t is the distribution of X_t when X_0 has distribution η.

Proof. Use coupling. First suppose that $\nu = \delta_m$ and $\tilde{\nu} = \delta_n$ with $m < n$. Let X_t have distribution $\delta_m P_t$ and let B_t have distribution $\mathrm{Bin}(n-m, e^{-\mu t})$, independently of X_t. By Theorem 11.8, the random variable $\tilde{X}_t = X_t + B_t$ has distribution $\delta_n P_t$. Thus

$$\mathbb{E}|X_t - \tilde{X}_t| = \mathbb{E}(B_t) = (n-m)e^{-\mu t}.$$

Since $(X_t, \tilde{X}_t)$ is a coupling of $(\delta_m P_t, \delta_n P_t)$,

$$W_1(\delta_m P_t, \delta_n P_t) \leqslant e^{-\mu t}(n-m) = e^{-\mu t} W_1(\delta_m, \delta_n).$$

Now conclude in the general case. Let ν and $\tilde{\nu}$ belong to $\mathcal{P}_1(\mathbb{N})$, and let $(X_0, \tilde{X}_0)$ be a coupling of ν and $\tilde{\nu}$. As above, construct (X_t) and $(\tilde{X}_t)$ such that

$$\mathrm{Loi}(|X_t - \tilde{X}_t| | X_0, \tilde{X}_0) = \mathrm{Bin}(|X_0 - \tilde{X}_0|, e^{-\mu t}).$$

We deduce that

$$W_1(\nu P_t, \tilde{\nu} P_t) \leqslant e^{-\mu t} \mathbb{E}|X_0 - \tilde{X}_0|.$$

Taking the infimum over all couplings of X_0 and $\tilde{X}_0$ gives the desired property. □

This contraction result bounds convergence in Wasserstein distance of the distribution of X_t to the invariant distribution π of X. The resulting convergence rate does not depend on the birth rate λ!

Corollary 11.15 (Convergence to equilibrium)**.** *If $\nu \in \mathcal{P}_1(\mathbb{N})$, then*

$$W_1(\nu P_t, \pi) \leqslant e^{-\mu t} W_1(\nu, \pi).$$

Remark 11.16 (Optimal convergence rate)**.** *The constant μ in the exponential convergence to equilibrium at rate $e^{-\mu t}$ is optimal. Indeed, suppose convergence held with a constant $\mu' > \mu$. Taking initial distribution $\nu = \delta_n$ with n greater than the mean ρ of the invariant distribution π, and denoting by X a random variable with distribution π, we would have*

$$(n-\rho)e^{-\mu t} = \mathbb{E}_n(X_t) - \rho \leqslant W_1(\delta_n P_t, \pi) \leqslant e^{-\mu' t} W_1(\nu, \pi) = \mathbb{E}|n - X| e^{-\mu' t},$$

a contradiction, since $n - \rho > 0$ and $e^{-\mu' t}/e^{-\mu t} \underset{t \to \infty}{\longrightarrow} 0$.

Similarly, one can show that the function $P_t f$ converges to the constant equal to the integral of f under π by proving that its (discrete) gradient converges to 0.

Theorem 11.17 (Commutation)**.** *If $f : \mathbb{N} \to \mathbb{R}$, then for every $t \geqslant 0$,*

$$DP_t f(n) = e^{-\mu t} P_t(Df)(n),$$

where Dg is the function defined on $\mathbb{N}$ by $Dg(n) = g(n+1) - g(n)$.

Proof. By Theorem 11.8,

$$P_t f(n+1) = \mathbb{E} f(Z_1 + \cdots + Z_{n+1} + Y),$$

where $Z_1, \ldots, Z_{n+1}$ are independent random variables with distribution $\mathrm{Ber}(e^{-\mu t})$, independent of Y with Poisson distribution of parameter $\rho(1 - e^{-\mu t})$. Thus

$$\begin{aligned} P_t f(n+1) &= \mathbb{E}(f(Z_1 + \cdots + Z_{n+1} + Y)) \\ &= e^{-\mu t} \mathbb{E}(f(Z_1 + \cdots + Z_n + 1 + Y)) \\ &\quad + (1 - e^{-\mu t}) \mathbb{E}(f(Z_1 + \cdots + Z_n + Y)) \\ &= e^{-\mu t} P_t(Df)(n) + P_t f(n), \end{aligned}$$

which proves the result. □

If f is 1-Lipschitz, meaning that $|Df(n)| \leqslant 1$ for every $n \in \mathbb{N}$, the preceding theorem ensures that $P_t f$ is $e^{-\mu t}$-Lipschitz. This recovers convergence in Wasserstein distance through the Kantorovich–Rubinstein dual formulation:

$$W_1(\nu, \tilde{\nu}) = \sup \left\{ \int f \, d\nu - \int f \, d\tilde{\nu} \ : \ f \text{ 1-lipschitzienne} \right\}.$$

Remark 11.18 (Queues, births, and deaths)**.** *Figure 11.1 compares sample paths of M/M/1 and M/M/∞ queues. These queues are birth-and-death processes*[4]*, namely continuous-time Markov processes with state space $\mathbb{N}$ and nearest-neighbor transitions. Their infinitesimal generator has the form (with the convention $\mu_0 = 0$)*

$$Lf(n) = \lambda_n(f(n+1) - f(n)) + \mu_n(f(n-1) - f(n)).$$

The M/M/1 case corresponds to constant $n \in \mathbb{N} \mapsto \lambda_n$ and $n \in \mathbb{N}^* \mapsto \mu_n$, whereas M/M/$\infty$ corresponds to constant $n \in \mathbb{N} \mapsto \lambda_n$ and $\mu_n = n\mu$ for every $n \in \mathbb{N}$. The case of linear $n \mapsto \lambda_n$ and $\mu \equiv 0$ is the Yule process (a continuous-time Galton–Watson process with offspring distribution δ_2).

[4] Also called "life-and-death processes".

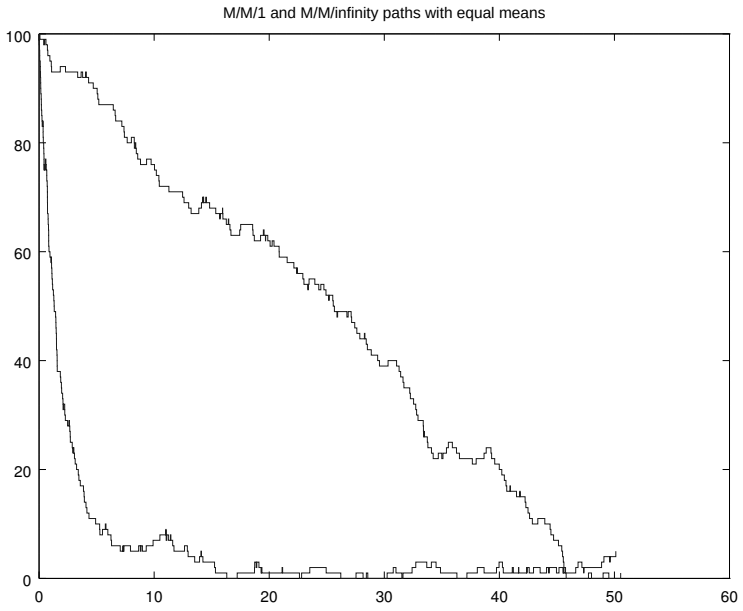


Fig. 11.1. Initial sample-path segments of an M/M/1 queue with parameters 1 and 3 and an M/M/∞ queue with parameters 1 and .5. Both start from 100. The first process has invariant distribution geometric on $\mathbb{N}$ with parameter 1/3; the second has Poisson invariant distribution of parameter $1/.5 = 2$. In both cases, the invariant distribution has mean 2. The first path exhibits linear behavior and the second exponential behavior.

11.5 Further reading

Classical results about exponential distributions can be found, for example, in James Norris' book [Nor98a]. Geoffrey Grimmett and David Stirzaker's book [GS01] studies a discrete-time version of the process X. Amusingly, it can model the number of typos remaining in a manuscript after n proofreading passes: in each pass, the author independently corrects each typo with probability p and inadvertently introduces a Poisson-distributed number of new ones (the *law of rare events* or *law of small numbers*).

Suitably rescaled, this discrete-state process converges to the Ornstein–Uhlenbeck process studied in Chapter 25. See, for example, Philippe Robert's book [Rob00]. More generally, that book gives an accessible and fairly general study of queues using martingales and point processes. See also the book by François Baccelli and Pierre Brémaud [BB03].

Many other results can be established for the process X. For example, [Cha06] gives functional inequalities related to its long-time behavior. The

commutation relation in Theorem 11.17 is a remarkable semigroup property. The Ornstein–Uhlenbeck process satisfies the same relation with the discrete gradient D replaced by the usual gradient on $\mathbb{R}$. This connection is an instance of a more general phenomenon treated in Chapter 27. Slightly weaker versions for more general processes provide estimates of convergence to the invariant distribution; see, for example, the article by Pietro Caputo, Paolo Dai Pra, and Gustavo Posta [CDPP09].

12

Wright–Fisher model

Keywords. Population dynamics; population genetics.

Tools. Markov chain; martingale; optional stopping theorem; spectral decomposition.

Difficulty. **

This chapter considers models of genotype evolution across generations in a population of constant size. The starting point is the Hardy–Weinberg law, related to Mendelian theory in an infinite population. We then present classical finite-population models: Moran's model, followed by Wright–Fisher models. The final section treats Cannings' model, which generalizes them.

12.1 Hardy–Weinberg law

Mendelian theory ensures the preservation of genotype variability.

Theorem 12.1 (Hardy–Weinberg law for Mendelian theory)**.** *Consider allele frequencies for a gene with two possible alleles A and B in an ideal hermaphroditic*[1] *diploid*[2] *population:*

1. *generations are separate (they do not coexist);*
2. *the population is infinite;*
3. *individuals mate randomly (no mate choice);*
4. *there is no migration, mutation, or selection.*

If $p \in [0,1]$ is the proportion of A alleles in the initial population and $1-p$ the proportion of B alleles, these proportions remain the same in each subsequent generation. Moreover, if r, s, and t are the respective proportions of genotypes

[1] Hermaphroditic: each individual can be male or female (bisexuality).
[2] Diploid: possessing two sets of similar chromosomes.

AA, AB, and BB in the initial population, with $r + s + t = 1$, then the frequencies of AA, AB, and BB in every subsequent generation equal

$$\Big(r + \frac{s}{2}\Big)^2, \quad 2\Big(r + \frac{s}{2}\Big)\Big(t + \frac{s}{2}\Big), \quad \Big(t + \frac{s}{2}\Big)^2.$$

Note that $p = r + s/2$ (contributions from AA and AB normalized to 1).

Proof. An individual of type AA arises

- with certainty from a pair AA and AA, which forms with probability r^2;
- with probability $1/2$ from a pair AA and AB, which forms with probability $2rs$;
- with probability $1/4$ from a pair AB and AB, which forms with probability s^2.

The frequency of genotype AA in generation 1 is therefore

$$r^2 + rs + \frac{s^2}{4} = \Big(r + \frac{s}{2}\Big)^2.$$

By symmetry, the frequency of genotype BB is $(t + s/2)^2$. The remaining frequency follows by subtraction. Applying these formulas again shows that the proportions are constant from the first generation onward. The frequency of allele A in the initial population is $2r + s$ (normalized to 2 because each individual has two alleles). In the next generation it is

$$2\Big(r + \frac{s}{2}\Big)^2 + 2\Big(r + \frac{s}{2}\Big)\Big(t + \frac{s}{2}\Big) = 2r + s.$$

and is therefore constant over time. □

The Hardy–Weinberg law generalizes to a locus[3] with several alleles $A_1, \ldots, A_k$. Dominance relationships between alleles do not affect allele-frequency evolution. In practice, Hardy–Weinberg behavior is observed in many populations beyond the assumptions made here. Substantial departures from these assumptions are needed to break this equilibrium and produce evolution across generations. This motivates the models studied in the rest of the chapter, describing finite-population evolution.

12.2 Moran model

The Moran model is an extremely simple model of changes in the frequencies of different alleles of a gene over time. In this section, consider two alleles A and B in a hermaphroditic population of constant finite size N, with separate generations and no migration, mutation, or selection. To move from generation

[3] Locus: the precise location of a particular gene on a chromosome.

n to generation $n+1$, sample two individuals with replacement from generation n: the first has two offspring, both inheriting its allele, the second has no offspring, and every other individual has exactly one offspring. Denote by X_n the number of A alleles in generation n. The number of A alleles

- increases by one if the sample with replacement gives (A, B);
- decreases by one if the sample is (B, A);
- otherwise remains constant.

We can realize $(X_n)_{n\geqslant 0}$ as a random recurrent sequence as follows. Let $(U_n)_{n\geqslant 1}$ and $(V_n)_{n\geqslant 1}$ be two independent sequences of i.i.d. real-valued random variables uniform on $[0,1]$. Let X_0 be a random variable taking values in $\{0,1,\dots,N\}$, independent of those introduced above. Set

$$X_{n+1} = X_n + \mathbb{1}_{\{U_{n+1}<X_n/N\}} - \mathbb{1}_{\{V_{n+1}<X_n/N\}}.$$

The sequence $(X_n)_{n\geqslant 0}$ is a Markov chain on $\{0,1,\dots,N\}$ with transition matrix $\mathbf{P}$ given, for all $x, y \in \{0,1,\dots,N\}$, by

$$\mathbf{P}(x,y) = \frac{1}{N^2}\begin{cases} x(N-x) & \text{si } |x-y|=1; \\ x^2+(N-x)^2 & \text{si } y=x; \\ 0 & \text{sinon.} \end{cases}$$

States 0 and N are absorbing. The others lead to $\{0,N\}$ and are therefore transient. The first time at which only one allele is present is denoted by

$$T := \inf\{n \geqslant 0 \,:\, X_n \in \{0,N\}\}.$$

We call T the *fixation time*. Since the state space is finite, the hitting time T of $\{0,N\}$ is almost surely finite and integrable:

$$\mathbb{P}(T<\infty) = 1 \quad \text{et} \quad \mathbb{E}(T) < \infty.$$

Moreover, X_T has a Bernoulli distribution supported on 0 and N, and

$$X_n \xrightarrow[n\to\infty]{\text{p.s.}} X_T.$$

We have $\mathbb{P}(X_T=0)+\mathbb{P}(X_T=N)=1$. The event $\{X_T=N\}$ (respectively $\{X_T=0\}$) means that allele A (respectively B) becomes fixed.

Use the standard notation $\mathbb{P}_x = \mathbb{P}(\cdot \,|\, X_0 = x)$ and $\mathbb{E}_x = \mathbb{E}(\cdot \,|\, X_0 = x)$.

Theorem 12.2 (Fixation state and time)**.** *For every $x \in \{0,1,\dots,N\}$,*

$$\mathbb{P}_x(X_T = N) = 1 - \mathbb{P}_x(X_T = 0) = \frac{x}{N}.$$

Moreover, for every $p \in]0,1[$,

$$\mathbb{E}_{\lfloor pN \rfloor}(T) \underset{N\to\infty}{\sim} -N^2(p\log p + (1-p)\log(1-p)).$$

Proof. Consider the case $X_0 \sim \delta_x$. We have $\mathbb{P}_x(T < \infty) = 1$, and since T is a fixation time, $(X_n)_{n\geqslant 0}$ converges a.s. to X_T. The bounded state space allows dominated convergence, so convergence holds in L^1, hence

$$\lim_{n\to\infty} \mathbb{E}_x(X_n) = \mathbb{E}(X_T) = 0\mathbb{P}_x(X_T = 0) + N\mathbb{P}_x(X_T = N) = N\mathbb{P}_x(X_T = N).$$

The formula for $\mathbf{P}$ gives, for every $n \geqslant 1$,

$$\mathbb{E}(X_n \,|\, X_{n-1} = x) = \frac{(x(N-x)(x-1+x+1) + (x^2 + (N-x)^2)x)}{N^2} = x,$$

hence

$$\mathbb{E}(X_n) = \mathbb{E}(\mathbb{E}(X_n \,|\, X_{n-1})) = \mathbb{E}(X_{n-1}) = \cdots = \mathbb{E}(X_0) = x,$$

which finally yields $\mathbb{P}_x(X_T = N) = x/N = 1 - \mathbb{P}_x(X_T = 0)$. These formulas can also be obtained from the linear system satisfied by the vector of hitting probabilities for the Markov chain $(X_n)_{n\geqslant 0}$.

Denote by $m(x) = \mathbb{E}_x(T)$ the mean fixation time of the chain starting from x. We have $m(0) = m(N) = 0$ and $m(x) = 1 + \sum_{y=0}^{N} \mathbf{P}(x,y)m(y)$ for every $x \in \{1, \ldots, N-1\}$. This gives the three-term recurrence

$$m(x+1) - 2m(x) + m(x-1) = -\frac{N^2}{x(N-x)}.$$

We deduce that

$$m(x) = N\left(\sum_{y=1}^{x} \frac{N-x}{N-y} + \sum_{y=x+1}^{N-1} \frac{x}{y}\right).$$

Let $0 < p < 1$. For every $N \geqslant 1$, set $x_N = \lfloor pN \rfloor$. Then

$$\begin{aligned}
\frac{m(x_N)}{N^2} &= \frac{1}{N}\sum_{y=1}^{x_N} \frac{N-x_N}{N-y} + \frac{1}{N}\sum_{y=x_N+1}^{N-1} \frac{x_N}{y} \\
&= \frac{1-N^{-1}x_N}{N}\sum_{y=1}^{x_N} \frac{1}{1-N^{-1}y} + \frac{N^{-1}x_N}{N}\sum_{y=x_N+1}^{N-1} \frac{1}{N^{-1}y} \\
&\underset{N\to\infty}{\longrightarrow} (1-p)\underbrace{\int_0^p \frac{1}{1-t}\,dt + p\int_p^1 \frac{1}{t}\,dt}_{=-((1-p)\log(1-p)+p\log(p))}.
\end{aligned}$$

Thus, in a population of size N with large N, the mean absorption time starting from $x = pN$ with $0 < p < 1$ is of order

$$-N^2(p\log p + (1-p)\log(1-p)).$$

□

The function $p \in [0,1] \mapsto -(p\log(p) + (1-p)\log(1-p))$ is the Boltzmann or Shannon entropy of the Bernoulli distribution $\mathrm{Ber}(p)$. It is continuous, strictly concave, nonnegative, symmetric about $1/2$, and attains its maximum $\log(2)$ at $p = 1/2$ and its minimum 0 at $p = 0$ and $p = 1$.

Remark 12.3 (Martingale)**.** *Let $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ for every $n \geqslant 0$. Returning to the proof of Theorem 12.2, for every $n \geqslant 1$ we have*

$$\mathbb{E}(X_n \,|\, \mathcal{F}_{n-1}) = \mathbb{E}(X_n \,|\, X_{n-1}) = X_{n-1}.$$

Thus, besides being a Markov chain, the sequence $(X_n)_{n\geqslant 0}$ is a martingale for its natural filtration $(\mathcal{F}_n)_{n\geqslant 0}$. In particular, $\mathbb{E}(X_0) = \mathbb{E}(X_n)$ for every $n \geqslant 1$. The optional stopping theorem states that $(X_{n\wedge T})_{n\geqslant 0}$ is also a martingale, so

$$\mathbb{E}(X_0) = \mathbb{E}(X_{n\wedge T}),$$

which converges to $\mathbb{E}(X_T)$ by dominated convergence as $n \to \infty$, since T is finite a.s. and X is uniformly bounded. Since T is a fixation time, in fact $X_n = X_{n\wedge T}$. One can also apply the theorem on almost sure and L^p convergence of martingales bounded in L^p with $p > 1$.

12.3 Wright–Fisher model and fixation

The Wright–Fisher model is a simple model of allele-frequency evolution across generations. Assume a hermaphroditic population of constant finite size, with separate generations, no migration, and mating independent of the trait under study.

Initially, ignore mutation and selection. Denote the population size by $N \in \mathbb{N}^*$ and successive generations by $n = 0, 1, 2, \ldots$. At the locus studied, two alleles, A and B, are possible. The random variable X_n counts A alleles in generation n. The population in generation $n+1$ is obtained from generation n by sampling with replacement N individuals from generation n, where the probability of obtaining A is X_n/N. We can realize $(X_n)_{n\geqslant 0}$ as the random recurrent sequence defined by

$$X_{n+1} = \sum_{k=1}^{N} \mathbb{1}_{\{U_{n+1,k} \leqslant \psi_{X_n}\}} \quad \text{où} \quad \psi_x := \frac{x}{N}$$

where $(U_{n,k})_{n\geqslant 1, 1\leqslant k\leqslant N}$ are i.i.d. random variables uniform on $[0,1]$, independent of X_0. For all $x_0, \ldots, x_n$, we have

$$\mathrm{Loi}(X_{n+1} \,|\, X_0 = x_0, \ldots, X_n = x_n) = \mathrm{Loi}(X_{n+1} \,|\, X_n = x_n) = \mathrm{Bin}(N, \psi_{x_n}).$$

Remark 12.4 (More than two alleles)**.** *The Wright–Fisher model with $\ell \geqslant 2$ alleles $A_1, \ldots, A_\ell$ involves the multinomial distribution. More precisely, X_n takes values in $\{x \in \mathbb{N}^\ell : x_1 + \cdots + x_\ell = N\}$ and*

$$\mathrm{Loi}(X_{n+1} \,|\, X_0, \ldots, X_n) = \mathrm{Loi}(X_{n+1} \,|\, X_n) = \mathrm{Mul}(N, \frac{X_n}{N}).$$

The sequence $(X_n)_{n\geqslant 0}$ is a Markov chain with state space $\{0,1,\dots,N\}$ and transition matrix $\mathbf{P}$ given, for all $x,y\in\{0,1,\dots,N\}$, by

$$\mathbf{P}(x,y) := \mathbb{P}(X_{n+1}=y \,|\, X_n=x) = \binom{N}{y}\psi_x^y(1-\psi_x)^{N-y}.$$

States 0 and N are absorbing (two singleton recurrent classes), whereas all other states lead to $\{0,N\}$ and are transient. As in the Moran model, when only one allele remains in the population, it is said to be fixed. The *fixation time* is

$$T := \inf\{n\geqslant 0 : X_n\in\{0,N\}\}.$$

Since the state space is finite, the hitting time T of $\{0,N\}$ satisfies

$$\mathbb{P}(T<\infty)=1 \quad \text{et} \quad \mathbb{E}(T)<\infty.$$

Moreover, $(X_n)_{n\geqslant 0}$ converges a.s. to the random variable X_T, whose Bernoulli distribution is supported on 0 and N. We have $\mathbb{P}(X_T=0)+\mathbb{P}(X_T=N)=1$. The event $\{X_T=N\}$ means that allele A becomes fixed, while $\{X_T=0\}$ means that allele B becomes fixed.

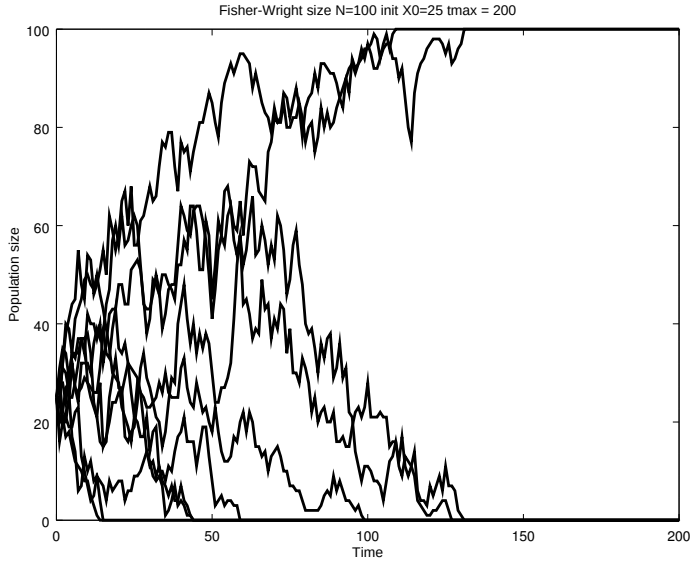


Fig. 12.1. Some sample paths, up to time $t=200$, of the Wright–Fisher chain for a population of size $N=100$, all starting from $X_0=N/4=25$.

Use the notation $\mathbb{P}_x=\mathbb{P}(\,\cdot\,|\,X_0=x)$ and $\mathbb{E}_x=\mathbb{E}(\,\cdot\,|\,X_0=x)$.

Theorem 12.5 (Fixation probability)**.** *For every $x \in \{0, 1, \dots, N\}$,*

$$\mathbb{P}_x(X_T = N) = 1 - \mathbb{P}_x(X_T = 0) = \frac{x}{N}.$$

Proof. The proof is identical to that for the Moran model (Theorem 12.2), since $\mathrm{Loi}(X_n \,|\, X_{n-1}) = \mathrm{Bin}(N, \frac{X_{n-1}}{N})$, hence

$$\mathbb{E}(X_n) = \mathbb{E}(\mathbb{E}(X_n \,|\, X_{n-1})) = \mathbb{E}(N\frac{X_{n-1}}{N}) = \mathbb{E}(X_{n-1}) = \cdots = \mathbb{E}(X_0) = x,$$

□

Remark 12.6 (Martingale)**.** *All of Remark 12.3 (Moran model) remains valid for the Wright–Fisher model, since for every $n \geqslant 1$,*

$$\mathbb{E}(X_n \,|\, \mathcal{F}_{n-1}) = \mathbb{E}(X_n \,|\, X_{n-1}) = N\frac{X_{n-1}}{N} = X_{n-1},$$

where $\mathcal{F}_n := \sigma(X_0, \dots, X_n)$ for every $n \geqslant 0$.

Remark 12.7 (Coupling with coin tossing)**.** *A geometric random variable can bound T through coupling. First observe that for every x and every n,*

$$\mathbb{P}(X_{n+1} \in \{0, N\} \,|\, X_n = x) = \mathbf{P}(x, \{0, N\}) \geqslant p_*$$

where

$$p_* = \min_{0 \leqslant x \leqslant N} \mathbf{P}(x, \{0, N\}) = \min_{0 \leqslant x \leqslant N} (1 - \psi_x)^N + \psi_x^N \geqslant 2^{-N+1}$$

and the minimum is attained at $\psi_x = 1/2$. Thus the chain reaches an absorbing state 0 or N before a coin tosser obtains the first heads *using a coin with heads probability 2^{-N+1}. To make this argument rigorous, construct the pair of processes (coin tossing and the Wright–Fisher chain) on the same probability space. Represent the Markov chain as a random recurrent sequence: for every $x \in \{0, \dots, N\}$, let $(I_{x,y})_{0 \leqslant y \leqslant N}$ be a partition of the interval $[0, 1]$, with $I_{x,y}$ of length*

$$|I_{x,y}| = \mathbf{P}(x, y),$$

and let $(U_n)_{n \geqslant 1}$ be a sequence of i.i.d. random variables uniform on $[0, 1]$, independent of X_0. Realize $(X_n)_{n \geqslant 0}$ using the recurrence

$$X_{n+1} := y \quad \text{où } y = y(X_n, U_{n+1}) \text{ est tel que } U_{n+1} \in I_{X_n, y}.$$

One can draw a tower of $N + 1$ copies of the interval $[0, 1]$ with its $N + 1$ partitions. In this construction, the intervals can always be ordered so that $I_{x,0}$ and $I_{x,N}$ are the first two from the left. By construction, $p_* = \min_{0 \leqslant x \leqslant N} |I_{x,0} \cup I_{x,N}|$ and

$$T := \inf\{n \geqslant 0 : X_n \in \{0, N\}\} \leqslant T_* := \inf\{n \geqslant 1 : \mathbb{1}_{\{U_n \leqslant p_*\}} = 1\}$$

and T_* has the geometric distribution $\mathrm{Geo}(p_*)$. In particular, T is integrable and hence finite a.s. This coupling method more generally provides geometric bounds for hitting times of sets by Markov chains.

The speed of fixation can be quantified by *heterozygosity*, the probability that two genes sampled randomly without replacement from the whole population in generation n have different alleles. It is given by

$$H_n = \frac{2X_n(N-X_n)}{N(N-1)} = \frac{2}{N-1}\mathrm{Var}(X_{n+1} \,|\, X_n).$$

We have $H_n = 0$ if $n \geqslant T$, hence $H_n \xrightarrow[n\to\infty]{\text{p.s.}} 0$. The following theorem shows that mean heterozygosity $\mathbb{E}(H_n)$ decreases exponentially over time.

Theorem 12.8 (Mean heterozygosity)**.** *For every $n \geqslant 1$,*

$$\mathbb{E}(H_n) = h_0\lambda^n \quad \textit{où} \quad h_0 := \mathbb{E}(H_0), \quad \lambda := 1 - \frac{1}{N} \in]0,1[,$$

and

$$\mathrm{Var}(X_n) = \mathbb{E}(X_0)(N - \mathbb{E}(X_0))(1-\lambda^n) + \lambda^n \mathrm{Var}(X_0).$$

Proof. It suffices to show that for every $n \geqslant 1$,

$$\mathbb{E}(X_n(N-X_n)) = \left(1 - \frac{1}{N}\right)^n \mathbb{E}(X_0(N-X_0)).$$

We have

$$\mathbb{E}(X_n(N-X_n)) = N\mathbb{E}(X_n) - \mathbb{E}(X_n^2) = N\mathbb{E}(X_{n-1}) - \mathbb{E}\big(\mathbb{E}(X_n^2 \,|\, X_{n-1})\big).$$

Then write

$$\begin{aligned}\mathbb{E}(X_n^2 \,|\, X_{n-1}) &= \mathrm{Var}(X_n \,|\, X_{n-1}) + (\mathbb{E}(X_n \,|\, X_{n-1}))^2 \\ &= X_{n-1}\left(1 - \frac{X_{n-1}}{N}\right) + X_{n-1}^2.\end{aligned}$$

Collecting terms gives

$$\mathbb{E}(X_n(N-X_n)) = \left(1 - \frac{1}{N}\right)\mathbb{E}(X_{n-1}(N-X_{n-1})),$$

which yields the formula for $\mathbb{E}(H_n)$. For the variance, use the formula[4]

$$\mathrm{Var}(X_n) = \mathbb{E}(\mathrm{Var}(X_n \,|\, X_{n-1})) + \mathrm{Var}(\mathbb{E}(X_n \,|\, X_{n-1})).$$

□

[4] More generally, if Φ is convex, then $\mathrm{Var}^\Phi(X) := \mathbb{E}(\Phi(X)) - \Phi(\mathbb{E}(X)) \geqslant 0$ by Jensen's inequality, and $\mathrm{Var}^\Phi(X) = \mathbb{E}(\mathrm{Var}^\Phi(X \,|\, Y)) + \mathrm{Var}^\Phi(\mathbb{E}(X \,|\, Y))$ where $\mathrm{Var}^\Phi(X \,|\, Y) := \mathbb{E}(\Phi(X) \,|\, Y) - \Phi(\mathbb{E}(X \,|\, Y))$. If $\Phi(t) = t^2$, we recover the variance.

As in the gambler's ruin problem (Theorem 2.2) and the Moran model (Theorem 12.2), we seek the mean exit time as a function of the starting point. For every $0 \leqslant x \leqslant N$, write $m(x) = \mathbb{E}_x(T)$. Then $m(0) = m(N) = 0$ and, by the (weak) Markov property, we obtain a system of linear equations: for every $0 < x < N$,

$$m(x) = \sum_{y=0}^{N} \mathbb{E}_x(T\mathbb{1}_{\{X_1=y\}}) = \sum_{y=0}^{N} (1 + \mathbb{E}_y(T))\mathbf{P}(x,y) = 1 + \sum_{y=0}^{N} m(y)\mathbf{P}(x,y),$$

whose unique minimal nonnegative solution is m. Although m can be computed numerically, there is no simple explicit expression for m as for the exit time from an interval of the simple random walk (Theorem 2.2) or for the Moran model (Theorem 12.2). Nevertheless, an asymptotic equivalent of m can be found as the population size N tends to infinity: for every $0 < p < 1$,

$$m(\lfloor Np \rfloor) \underset{N\to\infty}{\sim} -2N(p\log(p) + (1-p)\log(1-p)),$$

reminiscent of the Moran model (Theorem 12.2). The key is to show that the sequence of processes $Y^{(N)}$, defined by

$$Y^{(N)} = \left(\frac{1}{N} X_{\lfloor Nt \rfloor}\right)_{t\in\mathbb{R}_+}$$

and taking values in $[0,1]$, converges in distribution as $N \to \infty$ to a diffusion process Y on $[0,1]$ solving the stochastic differential equation

$$dY_t = \sqrt{Y_t(1-Y_t)}\,dB_t.$$

Details are given in Chapter 27.

Remark 12.9 (Connection with the Moran model)**.** *For the Wright–Fisher model, the expected absorption time is of order N, whereas for the Moran model it is of order N^2 (see Theorem 12.2). This is because a Moran transition changes only one allele, while a Wright–Fisher transition affects all of them.*

12.4 Wright–Fisher model with mutation

Now suppose that before the sampling with replacement that produces the next generation, each individual can mutate independently: allele A mutates to B with probability $u \in [0,1]$, and B mutates to A with probability $v \in [0,1]$. The transition between generations thus amounts to first applying a mutation operator, then an inheritance operator (reversing this order would produce a different process). This mechanism defines a new process $(X_n)_{n\geqslant 0}$ taking values in $\{0,1,\ldots,N\}$ and satisfying

$$\mathrm{Loi}(X_{n+1} \,|\, X_0,\ldots,X_n) = \mathrm{Bin}(N, \psi_{X_n})$$

where now

$$\psi_x = \frac{x(1-u) + (N-x)v}{N}.$$

The process X is a Markov chain with state space $\{0, 1, \dots, N\}$.

- If $u = v = 0$, we recover the model without mutation. States 0 and N are absorbing, while $\{1, \dots, N-1\}$ is a transient class. The set of invariant distributions is $\{(1-p)\delta_0 + p\delta_N : p \in [0,1]\}$, and starting from x the chain converges in distribution to $(1-p_x)\delta_0 + p_x\delta_N$, where $p_x = x/N$;
- if $u = 0$ and $0 < v \leqslant 1$, state N is absorbing, while the state set $\{0, \dots, N-1\}$ is a transient class, and the chain converges in distribution to its unique invariant distribution δ_N;
- if $0 < u \leqslant 1$ and $v = 0$, state 0 is absorbing, while the state set $\{1, \dots, N\}$ is a transient class, and the chain converges in distribution to its unique invariant distribution δ_0;
- if $u = v = 1$, then $\{0, N\}$ is a recurrent class of period 2, while the state set $\{1, \dots, N-1\}$ is a transient class. The chain is almost surely absorbed by $\{0, N\}$ and then oscillates periodically between states 0 and N;
- if $0 < u \leqslant 1$ and $0 < v < 1$, or if $0 < u < 1$ and $0 < v \leqslant 1$, the chain is irreducible. Since the state space is finite, it is positive recurrent and has a unique invariant distribution μ, which assigns positive mass to every state. The law of large numbers for Markov chains states that, regardless of the initial distribution, a.s. for every $x \in \{0, 1, \dots, N\}$,

$$\lim_{n\to\infty} \frac{\operatorname{card}\{0 \leqslant k \leqslant n : X_k = x\}}{n+1} = \mu(x).$$

Moreover, the transition matrix has a nonzero diagonal entry, so the chain is aperiodic and converges in distribution to μ from any initial distribution: for every $x \in \{0, 1, \dots, N\}$,

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \mu(x).$$

The distribution μ can be computed numerically but has no simple formula. Its mean m, for example, can nevertheless be determined explicitly. Indeed, since μ is invariant,

$$m = \sum_{x=0}^{N} y\mu(y) = \sum_{y=0}^{N} y \sum_{x=0}^{N} \mu(x)\mathbf{P}(x,y) = \sum_{x=0}^{N} \mu(x) \sum_{y=0}^{N} y\mathbf{P}(x,y).$$

Since $\mathbf{P}(x,\cdot) = \operatorname{Bin}(N, \frac{(1-u)x + v(N-x)}{N})$, we have

$$\sum_{y=0}^{N} y\mathbf{P}(x,y) = (1-u)x + v(N-x).$$

This formula, affine in x, gives a closed equation for m:

$$m = \sum_{x=0}^{N} ((1-u)x + v(N-x))\mu(x) = (1-u)m + v(N-m),$$

hence

$$m = \frac{Nv}{u+v}.$$

For the variance σ^2 of μ, calculation gives

$$\sigma^2 = \frac{N^2 uv}{(u+v)^2(2N(u+v)+1)} + o_{N\to\infty}(N).$$

Note that if $u = v$, then $m = N/2$ no longer depends on u and v. Beyond the first two moments, one can show, as in Chapter 27, that the process

$$Y^{(N)} = \left(\frac{1}{N} X_{\lfloor Nt \rfloor}\right)_{t\in\mathbb{R}_+}$$

converges in distribution as $(N, Nu, Nv) \to (\infty, \alpha, \beta)$ to a diffusion process Y on $[0,1]$ solving the stochastic differential equation

$$dY_t = \sqrt{Y_t(1-Y_t)}dB_t - \alpha Y_t dt + \beta(1-Y_t)dt,$$

whose invariant distribution is a $\mathrm{Beta}(2\beta, 2\alpha)$ distribution, with mean and variance

$$\frac{\beta}{\alpha+\beta} \quad \text{et} \quad \frac{\alpha\beta}{(\alpha+\beta)^2(2(\alpha+\beta)+1)}.$$

12.5 Wright–Fisher model with selection

Biologically, alleles A and B may confer different abilities on their carriers: viability, attractiveness, and so forth. This asymmetry drives natural selection. For simplicity, consider the model without mutation. A selective advantage of allele A can be incorporated by transforming generation n before constructing generation $n+1$: replace each individual of type A by $(1+s)$ individuals of type A, where $s > 0$ is fixed (think of an urn). This amounts to taking

$$\mathrm{Loi}(X_{n+1} \,|\, X_0, \ldots, X_n) = \mathrm{Loi}(X_{n+1} \,|\, X_n) = \mathrm{Bin}(N, \psi_{X_n})$$

with, this time,

$$\psi_x = \frac{x(1+s)}{x(1+s) + N - x}.$$

The process $(X_n)_{n\geqslant 0}$ is a Markov chain. Its expectation does not satisfy a linear recurrence, since ψ_x is not linear in x (unlike the no-selection case $s = 0$). States 0 and N are absorbing, while $\{1, \ldots, N-1\}$ is a transient class. Since ψ_x is no longer linear in x, the method used without selection ($s = 0$)

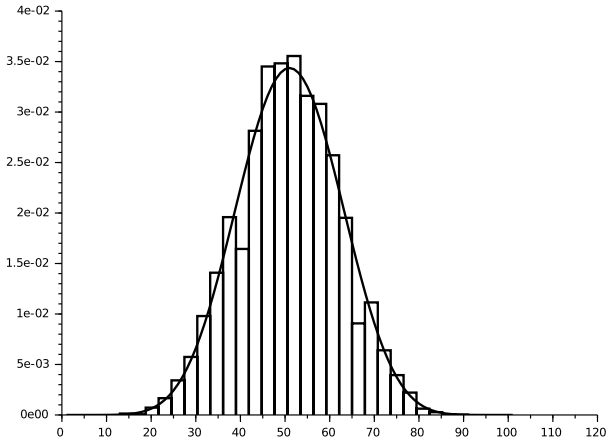


Fig. 12.2. Invariant distribution and histogram of a sample of size 10000 from X_{100} with $X_0 = 50$, for a population of size $N = 100$ and mutation rates $u = v = 0.05$.

no longer computes the fixation probability. Let $\mathbf{P}$ be the transition matrix of X and

$$\pi(x) := \mathbb{P}_x(X_T = N)$$

the fixation probability of allele A. We have

$$\pi(x) = \sum_{y=0}^{N} \mathbf{P}(x, y)\pi(y).$$

An asymptotic equivalent of π can be found as the population size N tends to infinity and s is close to 0. More precisely, Chapter 27 shows that if $Ns \to \alpha > 0$, then for every $0 < p < 1$,

$$\pi(\lfloor pN \rfloor) \underset{N\to\infty}{\longrightarrow} \frac{1 - e^{-2\alpha p}}{1 - e^{-2\alpha}}.$$

Thus, changing $s = 0$ to $s > 0$ substantially alters fixation probabilities.

12.6 Cannings model

In the Wright–Fisher model, generation n produces generation $n + 1$ by sampling with replacement N individuals from generation n. This assigns a number of offspring to each individual in generation n, with offspring inheriting the parent's allele. If $y_1, \ldots, y_N$ denote the offspring counts of individuals

in generation n (numbered from 1 to N), then $y_1 + \cdots + y_N = N$ and $(y_1, \ldots, y_N)$ has the symmetric multinomial distribution of size N and parameter $(1/N, \ldots, 1/N)$. This corresponds to rolling N times a fair die with N faces. This mechanism associating each individual in generation $n+1$ with an individual in generation n can study gene transmission with any number of alleles: offspring inherit the parent's allele. Cannings' model goes beyond this symmetric multinomial structure by allowing dependence among individuals in the same generation. It assumes only that the vector $(y_1, \ldots, y_N)$ satisfies $y_1 + \cdots + y_N = N$ and has an exchangeable distribution: for every permutation $\tau \in \mathcal{S}_N$,

$$(y_1, \ldots, y_N) \overset{\text{loi}}{=} (y_{\tau(1)}, \ldots, y_{\tau(N)}).$$

In particular, the $y_1, \ldots, y_N$ have the same distribution. Since $y_1 + \cdots + y_N = N$, we deduce that $\mathbb{E}(y_1) = 1$ and write $\sigma^2 = \mathrm{Var}(y_1)$. We also deduce $\mathrm{Cov}(y_1, y_2) = -\sigma^2/(N-1)$ because

$$0 = \mathrm{Var}(y_1 + \cdots + y_N) = N\mathrm{Var}(y_1) + N(N-1)\mathrm{Cov}(y_1, y_2).$$

Remark 12.10 (Moran model)**.** *Like the Wright–Fisher model, the Moran model is a special case of Cannings' model. It corresponds to the following distribution of $(y_1, \ldots, y_N)$: to pass from generation n to generation $n+1$, sample two individuals with replacement from generation n; the first has two offspring, the second none, and all others exactly one.*

With two alleles, this transition mechanism yields, as for the Wright–Fisher model, a Markov chain $(X_n)_{n \geqslant 0}$ with state space $\{0, 1, \ldots, N\}$ given, for every $x \in \{0, 1, \ldots, N\}$, by

$$\mathrm{Loi}(X_{n+1} \,|\, X_n = x) = \mathrm{Loi}(y_1 + \cdots + y_x).$$

States 0 and N are absorbing, whereas all other states lead to $\{0, N\}$ and are therefore transient. Note that $\mathbb{E}(X_{n+1} \,|\, X_n = x) = x$: hence $(X_n)_{n \geqslant 0}$ is a martingale. Moreover,

$$\begin{aligned}
\mathrm{Var}(X_{n+1} \,|\, X_n = x) &= \mathrm{Var}(y_1 + \cdots + y_x) \\
&= x\sigma^2 + x(x-1)\mathrm{Cov}(y_1, y_2) \\
&= x\sigma^2 - \frac{x(x-1)\sigma^2}{N-1} \\
&= \sigma^2 \frac{x(N-x)}{N-1}.
\end{aligned}$$

Remark 12.11 (Multinomial distribution)**.** *Let $n, d \in \mathbb{N}^*$ and $(p_1, \ldots, p_d) \in [0,1]^d$ with $p_1 + \cdots + p_d = 1$. The multinomial distribution of size n and parameter $(p_1, \ldots, p_d)$ is the distribution on the finite set*

$$\{(n_1, \ldots, n_d) \in \mathbb{N}^d : n_1 + \cdots + n_d = n\}$$

given by ($e_1, \dots, e_d$ is the canonical basis of $\mathbb{R}^d$)

$$(p_1\delta_{e_1} + \cdots + p_d\delta_{e_d})^{*n} = \sum_{\substack{(n_1,\dots,n_d)\in\mathbb{N}^d \\ n_1+\cdots+n_d=n}} \frac{n!}{n_1!\cdots n_d!} p_1^{n_1}\cdots p_d^{n_d}\delta_{(n_1,\dots,n_d)}.$$

It encodes n rolls of a die with d faces and outcome probabilities $p_1, \dots, p_d$. It is "symmetric" when $p_1 = \cdots = p_d = 1/d$. If $(X_1, \dots, X_d)$ has distribution $\mathrm{Mul}(n, (p_1, \dots, p_d))$ and $I_1, \dots, I_r$ is a partition of $\{1, \dots, n\}$, then

$$(X_{I_1}, \dots, X_{I_r}) \sim \mathrm{Mul}(n, (p_{I_1}, \dots, p_{I_r}))$$

where

$$X_I := \sum_{i\in I} X_i \quad \textit{et} \quad p_I := \sum_{i\in I} p_i.$$

If $I \subset \{1, \dots, n\}$, then $X_I \sim \mathrm{Bin}(n, p_I)$. Thus the marginal distributions of a multinomial distribution are binomial. We identify $\mathrm{Bin}(n, p)$ with $\mathrm{Mul}(n, (p, 1-p))$. If $(X_1, \dots, X_d) \sim \mathrm{Mul}(n, (p_1, \dots, p_d))$, then $X_d = n - X_{\{1,\dots,d-1\}}$.

Theorem 12.12 (Spectrum of the transition matrix)**.** *The eigenvalues of the transition matrix of Cannings' model are*

$$\lambda_0 = 1 \quad \textit{et} \quad \lambda_j = \mathbb{E}(y_1 y_2 \cdots y_j) \quad \textit{pour} \quad 1 \leqslant j \leqslant N.$$

In particular, for the Wright–Fisher model, $\lambda_0 = 1$ and, for $1 \leqslant j \leqslant N$,

$$\lambda_j = \frac{N(N-1)\cdots(N-j+1)}{N^j} = \left(1 - \frac{1}{N}\right)\cdots\left(1 - \frac{j-1}{N}\right).$$

Proof. Consider the Vandermonde matrix

$$V = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1^2 & 1^3 & \cdots & 1^N \\ 1 & 2 & 2^2 & 2^3 & \cdots & 2^N \\ 1 & 3 & 3^2 & 3^3 & \cdots & 3^N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & N & N^2 & N^3 & \cdots & N^N \end{pmatrix} \in \mathcal{M}_{N+1,N+1}(\mathbb{R}).$$

Let $\mathbf{P} \in \mathcal{M}_{N+1,N+1}(\mathbb{R})$ be the transition matrix of Cannings' model. For all $i, j \in \{0, 1, \dots, N\}$, we have

$$(\mathbf{P}V)_{i,j} = \sum_{k=0}^{N} \mathbf{P}_{i,k} V_{k,j} = \sum_{k=0}^{N} \mathbf{P}_{i,k} k^j = \mathbb{E}(X_1^j \,|\, X_0 = i).$$

Moreover, since

$$\mathrm{Loi}(X_1 \,|\, X_0 = i) = \mathrm{Loi}(y_1 + \cdots + y_i),$$

there exist real numbers $b_{i,k}$ such that

$$\mathbb{E}(X_1^j \,|\, X_0 = i) = \mathbb{E}((y_1 + \ldots + y_i)^j) = i^{[j]}\mathbb{E}(y_1 y_2 \cdots y_j) + \sum_{k=0}^{j-1} b_{i,k} i^k,$$

where

$$i^{[j]} := i(i-1)(i-2)\cdots(i-j+1).$$

To see this, use induction on j, observing that the distribution $\mathrm{Loi}(y_1, \ldots, y_i \,|\, y_k)$ is exchangeable. There is therefore an upper triangular matrix $T \in \mathcal{M}_{N+1,N+1}(\mathbb{R})$ such that $T_{0,0} = 1$, $T_{j,j} = \mathbb{E}(y_1 \cdots y_j)$ for every $j \in \{1, \ldots, N\}$, and, for all $i, j \in \{0, 1, \ldots, N\}$,

$$\mathbb{E}(X_1^j \,|\, X_0 = i) = \sum_{k=0}^{j} T_{k,j} i^k = (VT)_{i,j}.$$

Thus

$$\mathbf{P}V = VT,$$

and since V is invertible, the matrices $\mathbf{P}$ and T have the same spectrum. □

We have $\mathbb{E}(y_1) = 1$, so 1 is an eigenvalue of multiplicity 2, corresponding to the two recurrent classes $\{0\}$ and $\{N\}$. The second largest eigenvalue is

$$\lambda_2 = \mathbb{E}(y_1 y_2) = \mathrm{Cov}(y_1, y_2) + \mathbb{E}(y_1)^2 = 1 - \frac{\sigma^2}{N-1}.$$

12.7 Further reading

Genetic mechanisms have been widely discussed since their discovery by Gregor Mendel in the nineteenth century. Godfrey Harold Hardy and Wilhelm Weinberg independently discovered their law at the beginning of the twentieth century. Ronald Aylmer Fisher and Sewall Wright's model dates from the 1930s. John Moran's model dates from the 1950s, while Chris Cannings' model dates from the 1970s. These models and their extensions are studied, for example, in the books by Jean-François Delmas and Benjamin Jourdain [DJ06], Rick Durrett [Dur08], Warren Ewens [Ewe04], and Simon Tavaré [Tav04], and in Sylvie Méléard's lecture notes [Mé13]. They continue to inspire research combining biology and mathematics. Diffusion approximations of Markov chains are treated in Chapter 27. The genealogy of the Wright–Fisher process is studied in Chapter 13, leading to the Chinese restaurant process of Chapter 14.

13

Genealogies and coalescence

Keywords. Tree; branching; genealogy; Kingman's coalescent.

Tools. Geometric distribution; exponential distribution; multinomial distribution; Poisson distribution; Ewens distribution; Galton–Watson process; Wright–Fisher model.

Difficulty. **

In this chapter, we study the genealogy of the Wright–Fisher process from Chapter 12, relating it in particular to the Chinese restaurant process of Chapter 14. This study leads to Kingman's coalescent process.

Consider the evolution of a population of finite size N, constant across generations. Assume that generations do not overlap and that, for every $n \geqslant 0$, generation n dies while giving birth to generation $n+1$. The transition mechanism is Markovian: generation $n+1$ depends on generation n and an independent source of randomness. In each generation, number the individuals from 1 to N. For every $n \geqslant 0$, let a_i^{n+1} be the number of the parent of individual i in generation $n+1$. This parent belongs to generation n. Assume that the random variables $a_1^{n+1}, \dots, a_N^{n+1}$ are i.i.d. uniform on $\{1, \dots, N\}$. Denote by ν_i^n the number of descendants in generation $n+1$ of individual i alive in generation n. The random variables $(\nu_i^n)_{1 \leqslant i \leqslant N}$ are not independent, since they satisfy $\nu_1^n + \cdots + \nu_N^n = N$. We have

$$\nu_i^n = \sum_{j=1}^{N} \mathbb{1}_{\{a_j^{n+1} = i\}}.$$

Consequently, the random vector $\nu^n = (\nu_1^n, \dots, \nu_N^n)$ has the multinomial distribution of size N and parameter $(1/N, \dots, 1/N)$, that is,

$$\mathbb{P}(\nu_1^n = n_1, \dots, \nu_N^n = n_N) = \frac{N!}{n_1! \dots n_N!} \left(\frac{1}{N}\right)^N \mathbb{1}_{\{n_1 + \cdots + n_N = N\}}.$$

If each individual is represented by one allele, A or B, of a two-allele gene, and the number of A alleles in generation n equals x, then the number of A alleles in generation $n+1$ has binomial distribution $\mathrm{Bin}(N, x/N)$. We thus recover the Wright–Fisher model of Chapter 12. The connection with the multinomial distribution was already observed when studying Cannings' model in Chapter 12. Here we constructed the Wright–Fisher model backward through generations, as illustrated in Figure 13.1. Each individual of generation $n+1$ has an ancestor in generation n chosen independently of the others and uniformly on $\{1, \dots, N\}$. Some members of generation n therefore have no descendants in generation $n+1$, causing the phenomenon of genetic drift (fixation of an allele).

Remark 13.1 (Poisson and multinomial distributions)**.** *If $Y_1, \dots, Y_N$ are independent Poisson random variables with respective parameters $\theta_1, \dots, \theta_N$, then*

$$\mathrm{Loi}((Y_1, \dots, Y_N) \,|\, Y_1 + \cdots + Y_N = N) = \mathrm{Mul}(N, (p_1, \dots, p_N))$$

where $p_k = \theta_k/(\theta_1 + \cdots + \theta_N)$ for every $1 \leqslant k \leqslant N$. In particular, if $\theta_1 = \cdots = \theta_N$, then $p_1 = \cdots = p_N = 1/N$. Thus the Wright–Fisher model is a Galton–Watson process with Poisson offspring distribution, conditioned to have a fixed population size; see Chapter 3.

The Wright–Fisher model without mutation exhibits *genetic drift*: one allele prevails over the other, and the population becomes homozygous after a random but almost surely finite number of generations. Likewise, after some number of generations, all individuals descend from a single ancestor. We now aim to describe and quantify this phenomenon.

Since individuals choose their parents independently and uniformly, the probability that two fixed individuals have distinct parents (in the preceding generation) equals $N(N-1)/N^2 = 1-1/N$. These parents in turn choose their own parents independently, so the probability that the first common ancestor of two given individuals lies more than r generations in the past is $(1-1/N)^r$. The time T_2 to the *most recent common ancestor* (MRCA) therefore has a geometric distribution on $\mathbb{N}^*$ with parameter $p_2 = 1/N$. Now consider three individuals and denote by T_3' the time to an MRCA for at least two of them (the first coalescence time among three individuals). We have

$$\mathbb{P}(T_3' > 1) = \frac{N(N-1)(N-2)}{N^3} = 1 - \frac{3}{N} + \frac{2}{N^2}.$$

Parental choices are independent in each generation, hence, for $k \geqslant 1$,

$$\mathbb{P}(T_3' > k) = \left(1 - \frac{3}{N} + \frac{2}{N^2}\right)^k.$$

Thus T_3' has geometric distribution of parameter $p_3 = 3/N - 2/N^2$. Denote by $T_3 = T_3' + T_2'$ the time to an MRCA for all three individuals (here $T_2' :=$

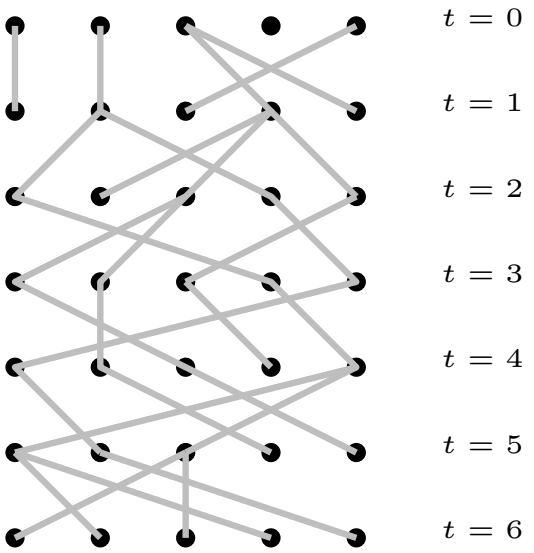


Fig. 13.1. In the Wright–Fisher model, everything behaves as if, in each generation, each individual independently chose a parent uniformly from the preceding generation and inherited from it. The diagram illustrates these parent-offspring relationships over several generations for a population of size $N = 5$. This genealogical viewpoint is central to the chapter.

$T_3 - T_3'$). If the MRCA of two individuals is the MRCA of all three, then $T_2' = 0$. Let us determine the distribution of T_2'. The probability that three distinct individuals have the same parent is $1/N^2$. The probability that T_2' is zero conditional on $T_3' = k$ therefore equals

$$\mathbb{P}(T_2' = 0 \,|\, T_3' = k) = \frac{\mathbb{P}(T_2' = 0\,,\, T_3' = k)}{\mathbb{P}(T_3' = k)} = \frac{(1-p_3)^{k-1}(1/N^2)}{p_3(1-p_3)^{k-1}} = \frac{1}{3N-2}.$$

Since parental choices are independent across generations, the first part implies that the conditional distribution of T_2' given $T_3' = k$ and $T_2' > 0$ is geometric with parameter p_2. Thus, conditional on $T_3' = k$, the variable T_2' equals in distribution UV, where U and V are independent with respective distributions $\mathrm{Ber}(1 - 1/(3N-2))$ and $\mathrm{Geo}_{\mathbb{N}^*}(p_2)$.

More generally, the number of distinct parents of a group of k individuals can be viewed as the number of occupied urns after k balls are thrown independently and uniformly into N urns. For every $j \in \{1, \ldots, k\}$, the probability that this number equals j is

$$g_{k,j}^{(N)} = \mathbb{P}(k \,\text{individus ont}\, j \,\text{parents distincts}) = \frac{N(N-1)\cdots(N-j+1)S_{k,j}}{N^k},$$

where $S_{k,j}$ is a Stirling number of the second kind[1]. Indeed, there are exactly $N(N-1)\cdots(N-j+1)$ ways to choose j distinct parents among N, and $S_{k,j}$ ways to associate with these j parents k children; finally, N^k counts the ways to assign k children to their parents. Now define the ancestral process $(A_n^N(r))_{r\in\mathbb{N}}$ by letting $A_n^N(r)$ be the number of distinct ancestors in generation $-r$ of a group of size n at time 0 (time runs backward here). The sequence $(A_n^N(r))_{r\in\mathbb{N}}$ is a Markov chain on $\{1,\dots,n\}$ with transition matrix

$$G_N = \left(g_{j,k}^{(N)} \mathbb{1}_{\{j\leqslant k\}}\right)_{1\leqslant j,k\leqslant n}.$$

For this chain, state 1 is absorbing, while states $2,\dots,n$ are transient since they all lead to 1. Fine properties of this chain, such as the hitting time of state 1, are difficult to study. We therefore replace this discrete-time model by a simpler continuous-time model.

13.1 Continuous-time genealogical model

Consider the continuous-time Markov chain $(A(t))_{t\geqslant 0}$ on $\mathbb{N}^*$ whose infinitesimal generator is given, for all $j,k\in\mathbb{N}^*$, by

$$L(j,k) = \begin{cases} \binom{j}{2} & \text{si } j\geqslant 2 \text{ et } k=j-1,\\ -\binom{j}{2} & \text{si } j\geqslant 2 \text{ et } k=j,\\ 0 & \text{sinon.}\end{cases}$$

It is a pure death process on $\mathbb{N}^*$ (only jumps from n to $n-1$), with absorbing state 1. The holding time in n has exponential distribution of parameter $\binom{n}{2}$. This admits the following interpretation: each pair of individuals seeks a parent independently of all other pairs and finds one after an exponential time of parameter 1. For n individuals, there are $\binom{n}{2}$ distinct pairs. The minimum of $\binom{n}{2}$ independent random variables with common exponential distribution Exp(1) has distribution Exp($\binom{n}{2}$) (see also Lemma 11.2). Denote by $(A_n(t))_{t\geqslant 0}$ the process starting from n, taking values in $\{1,\dots,n\}$.

We now show that the rescaled discrete-time process converges to the continuous-time process. Consider the limiting case of large N, with time measured in units of N generations. The time to the MRCA of two given individuals is then T_2'/N, where T_2' has a geometric distribution on $\mathbb{N}^*$ with parameter $p_2 = 1/N$.

Lemma 13.2 (Exponential distribution as a limit of rescaled geometric distributions)**.** *If $(V_n)_{n\geqslant 1}$ is a sequence of geometric random variables with respective parameters $(\mu_n)_{n\geqslant 1}$ such that $\lim_{n\to\infty} n\mu_n = \mu > 0$, then $(n^{-1}V_n)_{n\geqslant 1}$ converges in distribution to the exponential distribution of parameter μ.*

[1] The number of ways to partition a set of k elements into j nonempty sets. These numbers also appear in Chapter 1 in the study of the coupon collector!

Proof. Use characteristic functions. If $V \sim \mathrm{Exp}(\mu)$, then

$$\varphi_{V_n/n}(t) = \frac{\mu_n e^{it/n}}{1-(1-\mu_n)e^{it/n}} = \frac{n\mu_n}{n\mu_n - n(1-e^{it/n})} \underset{n\to\infty}{\longrightarrow} \frac{\mu}{\mu - it} = \varphi_V(t),$$

Alternatively, for every $x \geqslant 0$,

$$\mathbb{P}(n^{-1}V_n \geqslant x) = \mathbb{P}(V_n \geqslant nx) = e^{-\mu_n \lfloor nx \rfloor} \to e^{-\mu x}.$$

□

For a population of N individuals and a time scale with one unit equal to N generations, the time T_2 to the MRCA of two given individuals is approximately exponentially distributed with parameter 1. In particular, $\mathbb{E}(T_2) = 1$. We have therefore shown that

$$(A_2^N(\lfloor Nt \rfloor))_{t\geqslant 0} \underset{N\to\infty}{\overset{\text{loi}}{\longrightarrow}} (A_2(t))_{t\geqslant 0}.$$

Consider the genealogical tree of three individuals. As N tends to infinity, T_2'/N converges in distribution to $\mathrm{Exp}(1)$. Similarly, $(T_2'/N, T_3'/N)$ converges to (T_2, T_3), where T_2 and T_3 are independent with respective exponential distributions $\mathrm{Exp}(1)$ and $\mathrm{Exp}(3)$:

$$\begin{aligned}
\varphi_{(T_2'/N, T_3'/N)}(s,t) &= \mathbb{E}(\exp(isT_2'/N + itT_3'/N)) \\
&= \mathbb{E}(\mathbb{E}(\exp(isT_2'/N)|T_3')\exp(itT_3'/N)) \\
&= \frac{1}{3N-2}\mathbb{E}\Big(e^{itT_3'/N}\Big) + \frac{3N-3}{3N-2}\mathbb{E}\Big(e^{itV/N}\Big)\mathbb{E}\Big(e^{itT_3'/N}\Big)
\end{aligned}$$

and Lemma 13.2 shows that $(A_3^N(\lfloor Nt \rfloor))_{t\geqslant 0}$ converges in distribution to the process $(A_3(t))_{t\geqslant 0}$ as N tends to infinity. In other words, as N tends to infinity with time measured in units of N generations, the mechanism producing the MRCA of three distinct individuals in the limiting process is as follows:

- after an exponential time $T_3 \sim \mathrm{Exp}(3)$, a common ancestor of 2 individuals appears;
- the common ancestor then appears after a time $T_2 \sim \mathrm{Exp}(1)$;
- the random variables T_2 and T_3 are independent.

In particular, in the limit, the probability that the first common ancestor is common to all three individuals is zero. The result generalizes as follows.

Theorem 13.3 (From discrete to continuous time)**.** *For every $n \geqslant 2$, the following convergence of processes holds in the sense of marginal distributions:*

$$(A_n^N(\lfloor Nt \rfloor))_{t\geqslant 0} \underset{N\to\infty}{\overset{\text{loi}}{\longrightarrow}} (A_n(t))_{t\geqslant 0}.$$

Proof. We give only an outline. First consider the behavior of the transition matrix of G_N for large N. Since $S_{k,k-1} = \binom{k}{2}$, for every $2 \leqslant k \leqslant n$,

$$g_{k,k-1}^{(N)} = S_{k,k-1} \frac{N(N-1)\cdots(N-k+1)}{N^k} = \binom{k}{2}\frac{1}{N} + \mathcal{O}(N^{-2})$$

while for $j < k-1$, we have

$$g_{k,j}^{(N)} = S_{k,j} \frac{N(N-1)\cdots(N-j+1)}{N^k} = \mathcal{O}(N^{-2}).$$

Finally,

$$g_{k,k}^{(N)} = N^{-k} N(N-1)\cdots(N-k+1) = 1 - \binom{k}{2}\frac{1}{N} + \mathcal{O}(N^{-2}).$$

Thus $G_N = I + N^{-1}Q + \mathcal{O}(N^{-2})$. The rest of the proof, omitted here, identifies the limit of $(G_N)^{\lfloor Nt \rfloor}$. □

13.2 Length of the genealogical tree

The continuous-time process $(A_n(t))_{t \geqslant 0}$ with state space $\{1, \dots, n\}$ is a pure death process: its piecewise continuous sample paths are nonincreasing. It starts from $A_n(0) = n$ and decreases only through jumps of size -1. For every $k \geqslant 2$, conditional on the process being in state k, the waiting time T_k before entering state $k-1$ (the only possible transition) has distribution $\mathrm{Exp}(\binom{k}{2})$. Moreover, the random variables $(T_k)_{2 \leqslant k \leqslant n}$ are independent. The time to the MRCA W_n can therefore be written as

$$W_n = T_n + \cdots + T_2,$$

where the $(T_k)_{2 \leqslant k \leqslant n}$ are the holding times in states $2, \dots, n$. In particular,

$$\mathbb{E}(W_n) = 2\sum_{k=2}^{n} \frac{1}{k(k-1)} = 2\sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right) \to 2$$

and, since the $T_2, \dots, T_n$ are independent, also

$$\begin{aligned} \mathrm{Var}(W_n) &= \sum_{k=2}^{n} \mathrm{Var}(T_k) \\ &= \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} \\ &= 4\sum_{k=2}^{n} \left(\frac{1}{k^2} + \frac{1}{(k-1)^2} - \frac{2}{k(k-1)}\right) \end{aligned}$$

$$= 8 \sum_{k=1}^{n-1} \frac{1}{k^2} + \frac{4}{n^2} - 4 - 8\left(1 - \frac{1}{n}\right).$$

We have $\mathrm{Var}(W_n) \nearrow \frac{4\pi^2}{3} - 12 \simeq 1.16$ (note that $\mathrm{Var}(T_2) = 1$). An explicit expression for the density of W_n can be obtained by observing that $\mathbb{P}(W_n \leqslant t) = \mathbb{P}(A_n(t) = 1)$ and computing the transition matrix of the ancestral process (a rather unexciting task).

The genealogical tree of a group of n individuals consists of all their ancestors across all generations, including themselves, up to the first common ancestor of every individual. Denote by L_n the *length of the genealogical tree*, the random variable equal to the sum of the lifetimes of all individuals in the tree. The tree length L_n can be expressed through the times at which common ancestors appear:

$$L_n = 2T_2 + \cdots + nT_n.$$

In particular, since $T_2, \ldots, T_n$ are independent with $T_k \sim \mathrm{Exp}(\binom{k}{2})$,

$$\mathbb{E}(L_n) \underset{n\to\infty}{\sim} 2\log(n) \quad \text{et} \quad \mathrm{Var}(L_n) \underset{n\to\infty}{\sim} \frac{2\pi^2}{3}.$$

Theorem 13.4 (Length of the ancestral tree)**.** *The random variable L_n has the distribution of the maximum of $n-1$ i.i.d. real-valued random variables with exponential distribution of parameter $1/2$. Moreover,*

$$\frac{L_n}{\log(n)} \xrightarrow[n\to\infty]{\text{p.s.}} 2 \quad \textit{et} \quad \frac{L_n}{2} - \log(n) \xrightarrow[n\to\infty]{\text{loi}} \text{Gumbel}$$

where the Gumbel distribution has distribution function $t \in \mathbb{R} \mapsto e^{-e^{-t}}$.

Proof. The random variable L_n is the sum of independent random variables $2T_2, \ldots, nT_n$ with distributions $\mathrm{Exp}(1/2), \ldots, \mathrm{Exp}((n-1)/2)$. The first result follows from the following property: if $E_1, \ldots, E_n$ are independent random variables with distributions $\mathrm{Exp}(\lambda), \mathrm{Exp}(2\lambda), \ldots, \mathrm{Exp}(n\lambda)$, then their sum $S_n := E_1 + \cdots + E_n$ has the same distribution as $M_n := \max(F_1, \ldots, F_n)$, where $F_1, \ldots, F_n$ are i.i.d. random variables with distribution $\mathrm{Exp}(\lambda)$. The density of M_n is

$$\begin{aligned} f_n(x) &:= (\mathbb{P}(M_n \leqslant x))' \\ &= ((1 - e^{-\lambda x})^n \mathbb{1}_{\mathbb{R}_+}(x))' \\ &= n\lambda(1 - e^{-\lambda x})^{n-1} e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x). \end{aligned}$$

Let us show that S_n has density f_n. This holds for $n = 1$. By induction, if it holds for n, then the density of S_{n+1}, writing g_λ for that of $\mathrm{Exp}(\lambda)$, is

$$f_n * g_{(n+1)\lambda}(y) = (n+1)\lambda \int_{-\infty}^{y} f_n(x) e^{-\lambda(n+1)(y-x)}\, dx$$

$$
\begin{aligned}
&= \lambda(n+1)n\lambda e^{-\lambda(n+1)y} \int_0^y (e^{\lambda x}-1)^{n-1} e^{\lambda x}\, dx \\
&= \lambda(n+1)e^{-\lambda(n+1)y}(e^{\lambda y}-1)^n \\
&= \lambda(n+1)e^{-\lambda y}(1-e^{-n\lambda y})^n = f_{n+1}(y).
\end{aligned}
$$

In fact, this property is only part of Rényi's lemma 11.3.

To establish the theorem's second result (almost sure convergence), use the following fact[2]: if $(X_n)_n$ are independent centered real-valued random variables satisfying $\sum_n \mathbb{E}(X_n^2) < \infty$, then $\sum_n X_n$ converges almost surely and in L^2. With $X_n = nT_n - \mathbb{E}(nT_n) = nT_n - 2/(n-1)$, we have $\mathbb{E}(X_n) = 0$ and

$$
\sum_{n\geqslant 2} \mathbb{E}(X_n^2) = \sum_{n\geqslant 2} \mathrm{Var}(X_n) = \sum_{n\geqslant 2} \frac{4}{(n-1)^2} < \infty
$$

so $\sum_n X_n$ converges a.s., hence

$$
\begin{aligned}
\frac{L_n}{\log(n)} &= \frac{1}{\log(n)} \sum_{k=2}^{n} X_k + \frac{1}{\log(n)} \sum_{k=2}^{n} \mathbb{E}(kT_k) \\
&= \frac{1}{\log(n)} \sum_{k=2}^{n} X_k + \frac{2}{\log(n)} \sum_{k=1}^{n-1} \frac{1}{k} \xrightarrow[n\to\infty]{\text{p.s.}} 2.
\end{aligned}
$$

For the third result (Gumbel fluctuations), use the fact that L_n and M_n have the same distribution and write, for every $t \in \mathbb{R}$,

$$
\begin{aligned}
\mathbb{P}(L_n - 2\log(n) \leqslant 2t) &= \left(1 - e^{-\frac{1}{2}(2t+2\log(n))}\right)^{n-1} \\
&= \left(1 - \frac{e^{-t}}{n}\right)^n \underset{n\to\infty}{\longrightarrow} e^{-e^{-t}}.
\end{aligned}
$$

□

13.3 Mutations

To account for diversity within a population, we must consider possible DNA mutations, particularly during replication. Mutations generate genetic diversity among the offspring of a single individual. The mutation rate of a base[3] is very small. For simplicity, assume that this rate depends neither on the base nor on its position in the DNA strand and remains constant over time. Thus, in each generation, an individual has probability μ of a mutation distinguishing it from its parent. For a sequence of m bases, the probability that two mutations occur at the same position is $1/m$. Since mutation probabilities are

[2] A special case of the convergence theorem for martingales bounded in L^2.
[3] A DNA strand is a sequence of letters from the alphabet $\{A, C, G, T\}$.

small, this is very rare. We therefore assume infinitely many possible alleles, with each new mutation affecting a different site. A mutation thus always produces a new allele. The waiting time for a mutation in an individual's ancestral lineage has a geometric distribution of parameter μ. Set $\theta = 2\mu N$ and assume that θ is of order 1. In the limit as $N \to \infty$, the mutation process becomes a Poisson process. More precisely, equip each branch of the tree with a Poisson process of parameter $\theta/2$ counting mutations. The coalescence and mutation processes arise from different sources of randomness, and we take them to be independent.

Theorem 13.5 (Distribution of the number of alleles in a sample)**.** *If K_n denotes the number of distinct alleles in a group of n people, then*

$$K_n \overset{\text{loi}}{=} \eta_1 + \cdots + \eta_n$$

where $(\eta_k)_{k\geqslant 1}$ are independent random variables with $\eta_k \sim \mathrm{Ber}(\theta/(k-1+\theta))$, $k \geqslant 1$.

Note that $\eta_1 = 1$ almost surely. The random variable K_n corresponds to the number of tables in the Chinese restaurant process of Chapter 14. In particular, its first two moments can be computed, and its asymptotic behavior (convergence and fluctuations) is known.

Proof. For a group of n individuals, consider the first time (backward into the past) at which a coalescence or mutation occurs. Everything behaves as if U_n were the minimum of T_n, the first coalescence time (with distribution $\mathrm{Exp}(\binom{n}{2})$), and $R_1, \ldots, R_n$, the first mutation times of individuals $1, \ldots, n$, with common distribution $\mathrm{Exp}(\theta/2)$. Since these real-valued random variables are independent, their minimum has exponential distribution with parameter (Lemma 11.2)

$$\binom{n}{2} + n\frac{\theta}{2} = \frac{n(n-1+\theta)}{2},$$

and the probability that this first event is a coalescence is

$$\frac{\binom{n}{2}}{\frac{n(n-1+\theta)}{2}} = \frac{n-1}{n-1+\theta}.$$

Similarly, the probability that it is a mutation is

$$1 - \frac{n-1}{n-1+\theta} = \frac{\theta}{n-1+\theta}.$$

Set $\eta_n = 1$ if this first event is a mutation and $\eta_n = 0$ if it is a coalescence. The random variable η_n thus has Bernoulli distribution of parameter $\theta/(n-1+\theta)$. Consider what happens after time U_n:

- If $\eta_n = 0$, the number of ancestors at time U_n is $n-1$, and the number of distinct alleles in this group of $n-1$ individuals is $K_{n-1} = K_n$.

- If $\eta_n = 1$, the individual that mutated at time U_n gives rise to a different allele. In the initial population of n individuals, this corresponds to an allele distinct from all others. Moreover, this allele occurs only once in the population of n ancestors. We therefore obtain a population of $n-1$ individuals with $K_{n-1} = K_n - 1$ distinct alleles.

In either case, we have obtained $K_n = K_{n-1} + \eta_n$ and, from time U_n, consider a population of $n-1$ individuals with K_{n-1} distinct alleles. Induction gives the decomposition of K_n as a sum of Bernoulli random variables. The Markov property of the coalescence process and memorylessness of the mutation process ensure independence of the random variables $(\eta_k)_k$. □

We can describe the allele structure of a population of n individuals more precisely. In the infinite-alleles model, a sample of size n can be represented by a configuration $\boldsymbol{c} = (c_1, \ldots, c_n)$, where c_i is the number of alleles represented i times and $|\boldsymbol{c}| = c_1 + 2c_2 + \cdots + nc_n = n$. Denote by $\boldsymbol{e}_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ unit vector number i[e]. We now seek an equation satisfied by the probabilities $(q(\boldsymbol{c}))$ where $q(\boldsymbol{c})$ is the probability that a sample of size $|\boldsymbol{c}|$ drawn under the stationary distribution has configuration $\boldsymbol{c}$. Set $q(\boldsymbol{e}_1) = 1$. Suppose that the configuration is $\boldsymbol{c}$. Tracing the sample genealogy backward to the first change reveals either a mutation or a coalescence (a common ancestor of two individuals). Analyze these two events:

- The first event (mutation) has probability (Lemma 11.2)

$$\frac{n\theta/2}{n\theta/2 + n(n-1)/2} = \frac{\theta}{\theta + n - 1}.$$

  The configuration $\boldsymbol{b}$ that led to configuration $\boldsymbol{c}$ is
  - $\boldsymbol{b} = \boldsymbol{c}$ if the mutation occurred in one of the c_1 singletons (with probability c_1/n);
  - $\boldsymbol{b} = \boldsymbol{c} - 2\boldsymbol{e}_1 + \boldsymbol{e}_2$ if the mutation occurred in an allele represented twice (with probability $2b_2/n = 2(c_2+1)/n$);
  - $\boldsymbol{b} = \boldsymbol{c} - \boldsymbol{e}_1 - \boldsymbol{e}_{j-1} + \boldsymbol{e}_j$ if the mutation occurred in an allele represented j times (with probability $jb_j/n = j(c_j+1)/n$).
- The second event (coalescence) occurs with probability $(n-1)/(\theta+n-1)$. In that case, the preceding configuration had the form $\boldsymbol{b} = \boldsymbol{c} + \boldsymbol{e}_j - \boldsymbol{e}_{j+1}$: an allele present j times (among the c_j+1) was duplicated, changing from $c_j + 1$ to c_j the number of alleles present j times, and from $c_{j+1} - 1$ to c_{j+1} the number of alleles present $j+1$ times. This event has probability $jb_j/(n-1) = j(c_j+1)/(n-1)$.

We recover the Chinese restaurant process of Chapter 14, which shows that $\boldsymbol{c}$ has the Ewens distribution. In genetics this is called the *Ewens sampling formula*[4]: for all integers $a_1, \ldots, a_n$ satisfying $a_1 + 2a_2 + \cdots + na_n = n$,

[4] Abbreviated as *ESF*.

$$q(\boldsymbol{c}) = \mathbb{P}(\boldsymbol{c}_1(n) = a_1, \ldots, \boldsymbol{c}_n(n) = a_n)$$
$$= \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j}.$$

13.4 Kingman's coalescent

Kingman's coalescent is a continuous-time process taking values in the set $\mathcal{E}_n$ of equivalence relations on $[n] = \{1, \ldots, n\}$, or equivalently partitions of $[n]$. It describes the genealogy of a population of size n. So far, we have modeled the evolution of the number of ancestors of a given initial population. We now refine the description by keeping track of relationships among all individuals. Number the individuals from 1 to n. At time t, define the equivalence relation $\sim$ on $[n]$ by $i \sim j$ if and only if individuals i and j have the same ancestor at time t. Each equivalence class corresponds to an ancestor of the initial population and consists of all its descendants. Let $C(t)$ be this random partition.

Suppose that $C(0) = \alpha \in \mathcal{E}_n$, and denote by k its number of classes (writing $|\alpha| = k$). As t increases, we move into the past and the process remains constant until a common ancestor of the ancestors of two classes appears. When this happens, the two ancestors, and therefore all their descendants, share that common ancestor. The two equivalence classes merge (coalesce). This event occurs at rate 1. The process $(C(t))_{t \geqslant 0}$ is the continuous-time Markov process on $\mathcal{E}_n$ with initial state $C(0) = \Delta = \{\{1\}, \{2\}, \ldots, \{n\}\}$ (no one is related to anyone else) and generator Q

$$Q(\alpha, \beta) = \begin{cases} -\binom{k}{2} & \text{si } \alpha = \beta \text{ et } |\alpha| = k, \\ 1 & \text{si } \alpha \sim \beta, \\ 0 & \text{sinon,} \end{cases}$$

where $\alpha \sim \beta$ means that partition β can be obtained from partition α by merging two equivalence classes. The process C is called the coalescent or n-coalescent. To determine its distribution, first study the *embedded chain* $(C_k^{\text{inc}})_{k=n,n-1,\ldots,1}$. This chain starts from Δ and has transitions

$$\mathbb{P}(C_{k-1}^{\text{inc}} = \beta \,|\, C_k^{\text{inc}} = \alpha) = \begin{cases} \frac{1}{\binom{k}{2}} & \text{si } \alpha \sim \beta \text{ et } |\alpha| = k, \\ 0 & \text{sinon.} \end{cases}$$

The process C therefore evolves along a sequence

$$\Delta = C_n^{\text{inc}} \sim C_{n-1}^{\text{inc}} \sim \cdots \sim C_1^{\text{inc}} = \Theta,$$

where Θ is the trivial partition, and moves to partition C_k^{inc} after an exponential time of parameter $\binom{k}{2}$. Transition rates depend on a partition only through its cardinality, and

$$|C(t)| = A_n(t),$$

, since the classes of $C(t)$ correspond bijectively to the ancestors at time t. Thus the processes $(A_n(t))_{t\geqslant 0}$ and $(C_k^{\mathrm{inc}})_k$ are independent and, for every $t \geqslant 0$,

$$C(t) = C^{\mathrm{inc}}_{A_n(t)}.$$

More precisely,

$$\mathbb{P}(C(t) = \alpha) = \mathbb{P}(A_n(t) = |\alpha|)\mathbb{P}(C^{\mathrm{inc}}_{|\alpha|} = \alpha).$$

Theorem 13.6 (Distribution of Kingman's coalescent)**.** *Let $1 \leqslant j \leqslant n$ and let α be a partition of $[n]$ whose equivalence classes have cardinalities $\lambda_1, \dots, \lambda_j$. Then*

$$\mathbb{P}(C^{\mathrm{inc}}_j = \alpha) = \frac{(n-j)!j!(j-1)!}{n!(n-1)!}\lambda_1! \cdots \lambda_j!.$$

Proof. Proceed by downward induction. The result is clear for $j = n$ (if α is not the partition Δ, its probability is zero, and Δ is the only partition with n classes, all singletons). Suppose the result holds for $j \geqslant 2$. Then

$$\begin{aligned} p_{j-1}(\beta) &:= \mathbb{P}(C^{\mathrm{inc}}_{j-1} = \beta) \\ &= \sum_{\alpha\in\mathcal{E}_n} p_j(\alpha)\mathbb{P}(C^{\mathrm{inc}}_{j-1} = \beta \,|\, C^{\mathrm{inc}}_j = \alpha) \\ &= \sum_{\alpha\sim\beta} p_j(\alpha)\frac{2}{j(j-1)}. \end{aligned}$$

Denote by $\lambda_1, \dots, \lambda_{j-1}$ the sizes of the equivalence classes of β. Those of α are

$$\lambda_1, \dots, \lambda_{l-1}, m, \lambda_l - m, \lambda_{l+1}, \dots, \lambda_{j-1}$$

for some $1 \leqslant l \leqslant j-1$ and $1 \leqslant m \leqslant \lambda_l - 1$. Set

$$\pi_{l,m} := \lambda_1! \cdots \lambda_{l-1}!m!(\lambda_l - m)!\lambda_{l+1}! \cdots \lambda_{j-1}!.$$

By the induction hypothesis[5],

$$\begin{aligned} p_{j-1}(\beta) &= \frac{1}{2}\sum_{l=1}^{j-1}\sum_{m=1}^{\lambda_l-1} \frac{2(n-j)!j!(j-1)!}{j(j-1)n!(n-1)!}\pi_{l,m}\binom{\lambda_l}{m} \\ &= \frac{(n-j)!(j-1)!(j-2)!}{n!(n-1)!}\lambda_1! \cdots \lambda_{j-1}!\sum_{l=1}^{j-1}\sum_{m=1}^{\lambda_l-1} 1 \\ &= \frac{(n-j+1)!(j-1)!(j-2)!}{n!(n-1)!}\lambda_1! \cdots \lambda_{j-1}! \end{aligned}$$

because $\sum_{l=1}^{j-1}\sum_{m=1}^{\lambda_l-1} 1 = \lambda_1 + \cdots + \lambda_{j-1} - j + 1 = n - j + 1$. □

[5] The initial factor $1/2$ comes from interchanging m and $\lambda_l - m$. To understand it, consider the examples $\lambda_l = 2$ and $\lambda_l = 3$.

13.5 Further reading

In biology, *phylogeny* is the study of evolutionary relationships among living organisms. Stochastic modeling in phylogeny is discussed, for example, in the books by Warren Ewens [Ewe04], Rick Durrett [Dur08], and Jean-François Delmas and Benjamin Jourdain [DJ06]. Statistical aspects are treated in Simon Tavaré's course [Tav04]. John Kingman studies the process bearing his name in [Kin93]. His work led to many developments, including a theory of fragmentation and coalescence connected to branching processes, Lévy processes, and random partitions. See the books by Nathanaël Berestycki [Ber09] and Jean Bertoin [Ber06], Jim Pitman's Saint-Flour course [Pit06], and Sylvie Méléard's lecture notes [Mé13].

14

Chinese restaurants

Keywords. Random partition; random permutation; Ewens distribution; Fisher–Yates–Knuth algorithm; Bell numbers; Stirling numbers.

Tools. Combinatorics; symmetric group; Markov chain; martingale; Markov's inequality; Poisson distribution.

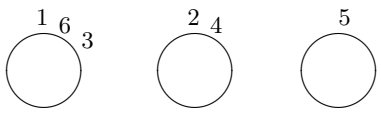
Difficulty. **

This chapter studies remarkable properties of the Ewens distribution, which appeared in Chapter 13 as the distribution of the partition of a population according to the alleles of a single gene in an infinite-alleles model. These properties are presented through the playful description of the so-called *Chinese restaurant process*, but the biological connections mentioned above provide its actual original motivation.

For every integer $n \geqslant 1$, denote by $\mathcal{S}_n$ the set of permutations of $\{1, \dots, n\}$ (the symmetric group). The *Chinese restaurant process* is a nonhomogeneous Markov chain $(\sigma_n)_{n\geqslant 1}$ with state space $\cup_{n\geqslant 1}\mathcal{S}_n$, where σ_n takes values in $\mathcal{S}_n$ for every $n \geqslant 1$, with initial value $\sigma_1 = (1)$ and transition kernel given, for all $(\sigma, \sigma') \in \mathcal{S}_n \times \mathcal{S}_{n+1}$, by

$$\mathbb{P}(\sigma_{n+1} = \sigma' \,|\, \sigma_n = \sigma) = \begin{cases} \dfrac{1}{\theta + n} & \text{si } \sigma' \text{ s'obtient en insérant } n+1 \\ & \text{dans l'un des cycles de } \sigma; \\ \dfrac{\theta}{\theta + n} & \text{si } \sigma' \text{ s'obtient en ajoutant} \\ & \text{le cycle } (n+1) \text{ à } \sigma; \\ 0 & \text{sinon;} \end{cases}$$

where $\theta \geqslant 0$ is a fixed parameter independent of n. This is indeed a transition kernel because the sum of the cycle lengths of σ is n. The process takes its name from interpreting the cycles of σ_n as circular tables occupied by the n customers of a Chinese restaurant, where strangers share tables. Initially, at time 1, the restaurant has a single table occupied by one customer numbered 1. Recursively in n, at time $n+1$ and conditional on the entire past $\sigma_1, \ldots, \sigma_n$, a new customer numbered $n+1$ enters the restaurant. They either join an occupied table (a cycle of σ_n), with probability proportional to its size (the cycle length), at a uniformly chosen seat, or sit at an empty table (creating a new cycle of length 1 containing $n+1$). This gastronomic interpretation assumes that customers never change tables, that tables can accommodate arbitrarily many customers, and that the restaurant can contain arbitrarily many tables.



$$\sigma_6 = (1,6,3)(2,4)(5) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 4 & 1 & 2 & 5 & 3 \end{pmatrix} \text{ and } \pi_6 = \{\{1,6,3\},\{2,4\},\{5\}\}$$

Fig. 14.1. A configuration with $n = 6$ customers at 3 tables in the Chinese restaurant.

Associate with the random permutation σ_n the random partition π_n of $\{1, \ldots, n\}$ given by its cycle supports. Each block of π_n consists of the customers at one table. We have $\pi_1 = \{1\}$, and for every $n \geqslant 1$, π_n takes values in the set Π_n of partitions of $\{1, \ldots, n\}$. In general, a function of a Markov chain is not a Markov chain[1]. Here, however, $(\pi_n)_{n \geqslant 1}$ is a nonhomogeneous Markov chain on $\cup_{n \geqslant 1} \Pi_n$ with transition kernel given, for all $(\pi, \pi') \in \Pi_n \times \Pi_{n+1}$, by

[1] A criterion due to Dynkin gives a sufficient condition on the function.

$$\mathbb{P}(\pi_{n+1} = \pi' \,|\, \pi_n = \pi) = \begin{cases} \dfrac{|b|}{\theta + n} & \text{si } \pi' \text{ s'obtient en insérant } n+1 \\ & \text{dans le bloc } b \text{ de } \pi; \\ \dfrac{\theta}{\theta + n} & \text{si } \pi' \text{ s'obtient en ajoutant} \\ & \text{le bloc singleton } \{n+1\} \text{ à } \pi; \\ 0 & \text{sinon;} \end{cases}$$

where $|b|$ denotes the cardinality of b. We have $\sum_{b \in \pi_n} |b| = n$, where $b \in \pi_n$ means that b is a block of π_n.

Remark 14.1 (Extreme cases)**.** *When $\theta = 0$, we have $\pi_n = \{\{1, \dots, n\}\}$, whereas if $\theta = \infty$, then $\pi_n = \{\{1\}, \dots, \{n\}\}$. These probabilities are monotone in θ. The larger θ is, the more customers tend to sit at empty tables instead of joining occupied ones.*

Remark 14.2 (Going backward in time)**.** *From this initial segment of a sample path of $(\pi_n)_{n \geqslant 1}$,*

$$\begin{array}{rc} \pi_1 = & \{\{1\}\} \\ \pi_2 = & \{\{1\}, \{2\}\} \\ \pi_3 = & \{\{1, 3\}, \{2\}\} \\ \pi_4 = & \{\{1, 3\}, \{2\}, \{4\}\} \\ \vdots & \vdots \end{array},$$

we immediately see that time can be reversed. For example, from π_4 we recover π_3 by locating the block containing 4, then π_2 by locating the block containing 3, and then π_1 by locating the block containing 2. More generally, for every $n \geqslant 1$ and every $\pi' \in \Pi_{n+1}$, there exists a unique $\pi \in \Pi_n$ such that $\mathbb{P}(\pi_{n+1} = \pi' \,|\, \pi_n = \pi) > 0$. All information about π_n is contained in π_{n+1} without any loss. The process carries its entire past with it.

14.1 Ewens distributions

Denote by $|\pi|$ the number of blocks in the partition $\pi \in \Pi_n$ and by $|\sigma|$ the number of cycles in the permutation $\sigma \in \mathcal{S}_n$. Thus $|\pi_n| = |\sigma_n|$ for every $n \geqslant 1$.

Theorem 14.3 (Ewens distribution on $\mathcal{S}_n$)**.** *For every $\sigma \in \mathcal{S}_n$, we have*

$$\mathbb{P}(\sigma_n = \sigma) = \frac{\theta^{|\sigma|}}{\theta(\theta+1)\cdots(\theta+n-1)}.$$

Proof. The formula holds for $n = 1$. Proceed by induction on n and assume it holds for $n \geqslant 1$. Let $\sigma' \in \mathcal{S}_{n+1}$. First observe that there is a unique $\sigma \in \mathcal{S}_n$

such that $\mathbb{P}(\sigma_{n+1} = \sigma' \,|\, \sigma_n = \sigma) > 0$. If $|\sigma'| = |\sigma|+1$, then $\mathbb{P}(\sigma_{n+1} = \sigma' \,|\, \sigma_n = \sigma) = \theta/(\theta+n)$, and if $|\sigma'| = |\sigma|$, then $\mathbb{P}(\sigma_{n+1} = \sigma' \,|\, \sigma_n = \sigma) = 1/(\theta+n)$. In either case,

$$\mathbb{P}(\sigma_{n+1} = \sigma') = \mathbb{P}(\sigma_n = \sigma)\,\mathbb{P}(\sigma_{n+1} = \sigma' \,|\, \sigma_n = \sigma) = \frac{\theta^{|\sigma'|}}{\theta(\theta+1)\cdots(\theta+n)}.$$

□

Remark 14.4 (Random walk, transpositions, and uniform distribution on $\mathcal{S}_n$)**.** *Conditional on $\{\sigma_n = \sigma\}$, we can construct σ_{n+1} from σ by first choosing a random element K of $\{1,\dots,n+1\}$, equal to $n+1$ with probability $\theta/(\theta+n)$ and to $1,\dots,n$ with probability $1/(\theta+n)$, then adding $n+1$ to the cycle of σ containing K if $K \leqslant n$, or creating a new cycle $(n+1)$ if $K = n+1$. This amounts to constructing from σ an element σ' of $\mathcal{S}_{n+1}$ by adjoining to σ the cycle $(n+1)$, then computing the product $\sigma'\tau$ in $\mathcal{S}_{n+1}$, where τ is the random transposition $(n+1, K)$. Thus*

$$(\sigma_n)_{n\geqslant 1} \overset{\text{loi}}{=} ((1,K_1)\cdots(n,K_n))_{n\geqslant 1}$$

where $(K_n)_{n\geqslant 1}$ is a sequence of independent random variables with $K_1 = 1$ and, for every $n \geqslant 1$, K_{n+1} having distribution $\frac{1}{\theta+n}\delta_1 + \cdots + \frac{1}{\theta+n}\delta_n + \frac{\theta}{\theta+n}\delta_{n+1}$. The process $(\sigma_n)_{n\geqslant 0}$ is a time-inhomogeneous random walk on $\mathcal{S}_n$, on the Cayley graph *of $\mathcal{S}_n$ generated by transpositions. If $\theta = 1$, then for every $n \geqslant 1$, K_n is uniformly distributed on $\{1,\dots,n\}$, while σ_n is uniformly distributed on $\mathcal{S}_n$, and we recover the Fisher–Yates–Knuth algorithm of Theorem 4.1.*

Theorem 14.5 (Ewens distribution on Π_n)**.** *For every $\pi \in \Pi_n$, we have*

$$\mathbb{P}(\pi_n = \pi) = \frac{\theta^{|\pi|}}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{b\in\pi} (|b|-1)!.$$

Proof. Theorem 14.3 gives

$$\mathbb{P}(\pi_n = \pi) = \sum_{\sigma\in E_\pi} \mathbb{P}(\sigma_n = \sigma) = \frac{\theta^{|\pi|}}{\theta(\theta+1)\cdots(\theta+n-1)} |E_\pi|$$

where E_π is the set of $\sigma \in \mathcal{S}_n$ whose cycle decomposition induces the partition π. There are $(k-1)!$ cycles of length k on a set of k elements, so the cardinality of E_π is $\prod_{b\in\pi}(|b|-1)!$. □

Remark 14.6 (Ewens and uniform distributions on Π_n)**.** *For $n \geqslant 3$, regardless of θ, the Ewens distribution on Π_n is never the uniform distribution studied in Chapter 4. Indeed, write $\pi = \{\{1\},\{2\},\dots,\{n\}\}$, $\pi' = \{\{1,2\},\{3\},\dots,\{n\}\}$, $\pi'' = \{\{1,2,\dots,n\}\}$. Then $\mathbb{P}(\pi_n = \pi) = \mathbb{P}(\pi_n = \pi')$ only for $\theta = 1$, and in that case $\mathbb{P}(\pi_n = \pi) \neq \mathbb{P}(\pi_n = \pi'')$.*

For every time $n \geqslant 1$ and every $1 \leqslant k \leqslant n$, let $A_{n,k}$ be the number of tables of size k at time n, that is, the number of blocks of size k in the random partition π_n. We have

$$n = A_{n,1} + 2A_{n,2} + \cdots + nA_{n,n} \quad \text{et} \quad |\pi_n| = A_{n,1} + \cdots + A_{n,n}.$$

Theorem 14.7 (Ewens distribution on blocks)**.** *For every $n \in \mathbb{N}^*$ and every $(a_1, \ldots, a_n) \in \mathbb{N}$ satisfying $a_1 + 2a_2 + \cdots + na_n = n$, we have*

$$\mathbb{P}(A_{n,1} = a_1, \ldots, A_{n,n} = a_n) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \frac{1}{a_j!} \left(\frac{\theta}{j}\right)^{a_j}.$$

Proof. Theorem 14.5 gives

$$\mathbb{P}(A_{n,1} = a_1, \ldots, A_{n,n} = a_n) = \frac{\theta^{a_1+\cdots+a_n}}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} ((j-1)!)^{a_j} |\Pi_n(a)|$$

where $\Pi_n(a)$ denotes the set of $\pi \in \Pi_n$ containing a_k blocks of size k for every $1 \leqslant k \leqslant n$. The set $\Pi_n(a)$ has cardinality

$$\frac{n!}{\prod_{j=1}^{n} (j!)^{a_j} a_j!},$$

which yields the formula. □

Remark 14.8 (Appearance and persistence of small tables)**.** *The Chinese restaurant process allows small tables to appear and persist over time because tables are chosen in proportion to their size!*

14.2 Number of tables

At time $n \geqslant 1$, the restaurant contains $|\pi_n|$ occupied tables.

Theorem 14.9 (Number of tables)**.** *For every $n \geqslant 1$, we have*

$$\mathbb{E}(|\pi_n|) = \sum_{k=0}^{n-1} \frac{\theta}{\theta+k} = \theta \log(n) + \mathcal{O}_{n\to\infty}(1) \underset{n\to\infty}{\sim} \theta \log(n),$$

and

$$\mathrm{Var}(|\pi_n|) = \sum_{k=1}^{n-1} \frac{\theta k}{(\theta+k)^2} = \theta \log(n) + \mathcal{O}_{n\to\infty}(1) \underset{n\to\infty}{\sim} \theta \log(n).$$

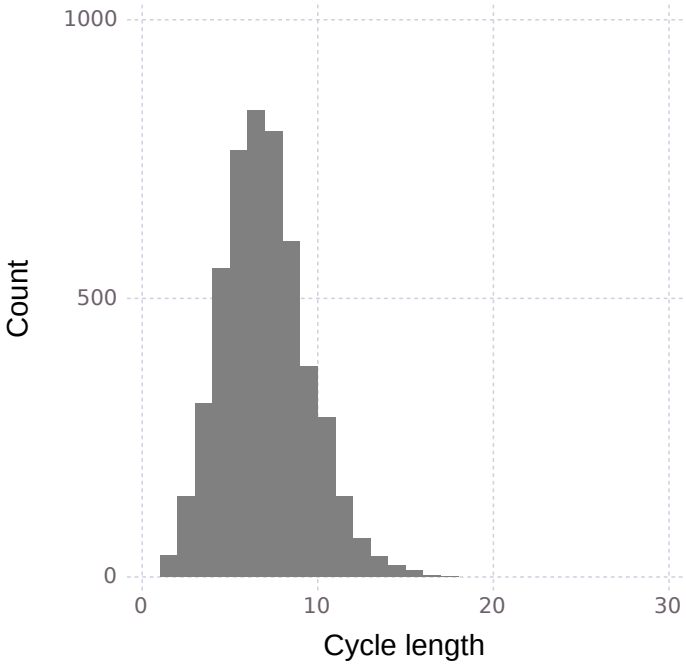


Fig. 14.2. Histogram of a sample of size 5000 from the Ewens distribution of size $n = 1000$ and parameter $\theta = 1$.

Proof. The Markov sequence $(\pi_n)_{n\geqslant 1}$ has the same distribution as the random recurrent sequence defined by $\pi_1 = \{\{1\}\}$ and, for every $n \geqslant 1$, by

$$\pi_{n+1} = f_n(\pi_n, \varepsilon_{n+1})$$

where $(\varepsilon_n)_{n\geqslant 1}$ are i.i.d. random variables uniform on $[0,1]$, and $f_n(\pi,\varepsilon)$ is the element of Π_{n+1} obtained from $\pi \in \Pi_n$ either by adding $n+1$ to block b_k of π if $\varepsilon \in [(|b_1|+\cdots+|b_{k-1}|)/(\theta+n), (|b_1|+\cdots+|b_k|)/(\theta+n)]$ with $1 \leqslant k \leqslant |\pi|$, where $b_1, \ldots, b_{|\pi|}$ are the blocks of π, or by adding the block $\{n+1\}$ to the partition π if $\varepsilon \in [n/(\theta+n), (\theta+n)/(\theta+n)]$. Recall that $n = |b_1|+\cdots+|b_{|\pi|}|$. Then $|\pi_n| = \xi_1 + \cdots + \xi_n$ for every $k \geqslant 1$, where

$$\xi_k := \mathbb{1}_{\{\varepsilon_k \in [\frac{k-1}{\theta+k-1}, \frac{\theta+k-1}{\theta+k-1}]\}}.$$

The sequence $(\xi_k)_{k\geqslant 1}$ consists of independent Bernoulli random variables, and for every $k \geqslant 1$,

$$\mathbb{P}(\xi_k = 1) = 1 - \mathbb{P}(\xi_k = 0) = \frac{\theta}{\theta + k - 1}.$$

The event $\{\xi_n = 1\}$ means that customer n starts a new table, while $\{\xi_n = 0\}$ means they join an existing table. In particular, $\xi_1 = 1$ a.s. This gives the

sum formulas for $\mathbb{E}(|\pi_n|)$ and $\mathrm{Var}(|\pi_n|)$. Recall the following sum-integral comparison: if $f : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and nonincreasing, then

$$\int_0^{n+1} f(t)\,dt \leqslant \sum_{k=0}^{n} f(k) \leqslant f(0) + \int_0^n f(t)\,dt.$$

Thus $\mathbb{E}(|\pi_n|)$ and $\mathrm{Var}(|\pi_n|)$ equal $\theta\log(n) + \mathcal{O}_{n\to\infty}(1) \sim_{n\to\infty} \theta\log(n)$. □

Remark 14.10 (Inhomogeneous coin tossing)**.** *By construction,* $(|\pi_n|)_{n\geqslant 1}$ *is a.s. nondecreasing and has only jumps of size* $+1$*. It is the counting process starting from* 1 *of events separated by independent random waiting times. However, these times are neither identically distributed nor geometrically distributed. It is instead a coin-tossing game whose winning probability changes at each toss. Although* $(\pi_n)_{n\geqslant 1}$ *is nondecreasing, its growth slows in a sense, since the quantity*

$$p_n = \mathbb{P}(|\pi_{n+1}| = |\pi_n| + 1) = \mathbb{P}(\xi_{n+1} = 1) = \theta(\theta + n)^{-1}$$

decreases as n *increases. Nevertheless, the mean and variance of* $|\pi_n|$ *are asymptotic to* $\theta\log(n)$ *as* n *tends to* ∞*. This differs from the coupon collector in Chapter 1, where the winning probability changes after each success.*

Theorem 14.11 (Asymptotics of the number of tables)**.** *We have*

$$\frac{|\pi_n|}{\log(n)} \xrightarrow[n\to\infty]{L^2} \theta \quad \textit{et en particulier} \quad \frac{|\pi_n|}{\log(n)} \xrightarrow[n\to\infty]{\mathbb{P}} \theta.$$

Proof. By Theorem 14.9, as $n \to \infty$,

$$\begin{aligned}\mathbb{E}\left(\left(\frac{|\pi_n|}{\log(n)} - \theta\right)^2\right) &= \frac{\mathrm{Var}(|\pi_n|) + (\mathbb{E}(|\pi_n|) - \theta\log(n))^2}{(\log(n))^2} \\ &= \mathcal{O}\left(\frac{1}{\log(n)}\right) = o(1).\end{aligned}$$

Convergence in probability follows from Markov's inequality. □

The following result states that convergence holds almost surely. The bound $\mathcal{O}(1/\log(n))$ obtained above by the second moment method is not summable, so the Borel–Cantelli lemma cannot directly yield almost sure convergence. This nevertheless suggests considering a higher-order moment. Alternatively, one can try to reduce the problem to a result about martingales.

Theorem 14.12 (Asymptotics of the number of tables)**.** *We have*

$$\frac{|\pi_n|}{\log(n)} \xrightarrow[n\to\infty]{\text{p.s.}} \theta.$$

Proof. Let $(\xi_n)_{n\geqslant 1}$ be as in the proof of Theorem 14.9. Then

$$\left(\frac{|\pi_n|}{\log(n)}\right)_{n\geqslant 2} \overset{\text{loi}}{=} \left(\frac{1}{\log(n)}\sum_{k=1}^{n}\xi_k\right)_{n\geqslant 2}.$$

Note that $\xi_1 = 1$ a.s. and therefore $\xi_1 - \mathbb{E}(\xi_1) = 0$. For every $n \geqslant 2$, we have

$$\frac{1}{\log(n)}\sum_{k=1}^{n}\xi_k = \frac{1}{\log(n)}\sum_{k=2}^{n}(\xi_k - \mathbb{E}(\xi_k)) + \frac{1}{\log(n)}\sum_{k=1}^{n}\mathbb{E}(\xi_k).$$

The second term on the right converges to θ as $n \to \infty$ by Theorem 14.9. It therefore suffices to show that the first term converges almost surely to 0. Now, if

$$S_n := \sum_{k=2}^{n} Y_k \quad \text{avec} \quad Y_k := \frac{\xi_k - \mathbb{E}(\xi_k)}{\log(k)}$$

then $(S_n)_{n\geqslant 2}$ is a martingale bounded in L^2 because (Theorem 14.9)

$$\mathbb{E}(S_n^2) = \sum_{k=2}^{n}\frac{\mathrm{Var}(\xi_k)}{(\log(k))^2} \leqslant \sum_{k=2}^{\infty}\frac{\theta}{(\theta+k-1)(\log(k))^2} < \infty.$$

Recall that the Bertrand series

$$\sum_{k=2}^{\infty}\frac{1}{k(\log(k))^{\beta}}$$

converges if and only if $\beta > 1$. Being bounded in L^2, the martingale $(S_n)_{n\geqslant 2}$ converges a.s. (and in L^2) to a random variable in L^2 (hence finite a.s.). Kronecker's lemma[2] then ensures almost sure convergence to 0 as $n \to \infty$ of

$$\frac{1}{\log(n)}\sum_{k=2}^{n}\log(k)Y_k = \frac{1}{\log(n)}\sum_{k=2}^{n}(\xi_k - \mathbb{E}(\xi_k)).$$

Alternatively, instead of the martingale theorem, one could use the centered series theorem (Theorem 10.3). □

Theorem 14.13 (Distributional asymptotics of the number of tables)**.** *We have*

$$\frac{|\pi_n| - \mathbb{E}(|\pi_n|)}{\sqrt{\mathrm{Var}(|\pi_n|)}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

Moreover, $\mathbb{E}(|\pi_n|)$ *and* $\mathrm{Var}(|\pi_n|)$ *can be replaced by* $\theta\log(n)$*.*

[2] If $\sum_n x_n$ converges, then $\lim_{n\to\infty} b_n^{-1}\sum_{k=1}^{n} b_k x_k = 0$ as $0 < b_n \nearrow \infty$.

Proof. First note that $\mathbb{E}(|\pi_n|)$ and $\mathrm{Var}(|\pi_n|)$ can be replaced by $\theta\log(n)$ using Slutsky's lemma and Theorem 14.9.

The result follows from the Berry–Esseen theorem of Section 1.6, applied with $X_k = \xi_k$ Bernoulli of mean $\theta/(\theta+k-1)$. This even gives a uniform rate of convergence for the distribution functions! The result can also be obtained using the Lindeberg–Lévy central limit theorem for sums of independent random variables that need not be identically distributed, or the central limit theorem for square-integrable martingales normalized by the increasing process.

An alternative proof uses the strengthened Le Cam inequality from Section 1.6, with $p_k = \theta/(\theta+k-1)$. Indeed, this gives $d_{\mathrm{VT}}(\mu_n, \nu_n) = o_{n\to\infty}(1)$, and consequently, for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$, with $\sigma_n^2 = \mathrm{Var}(|\pi_n|)$, we have

$$\left|\mathbb{E}\left(f\left(\frac{|\pi_n|-\lambda_n}{\sigma_n}\right)\right) - \mathbb{E}\left(f\left(\frac{P_n-\lambda_n}{\sigma_n}\right)\right)\right| \leqslant 2\|f\|_\infty\, d_{\mathrm{VT}}(\mu_n,\nu_n) \underset{n\to\infty}{\longrightarrow} 0.$$

where P_n denotes a random variable with Poisson distribution of mean $\lambda_n = p_1 + \cdots + p_n$. On the other hand, the central limit theorem, Theorem 14.9, which gives $\sqrt{\lambda_n}/\sigma_n \to 1$, and Slutsky's lemma yield

$$\frac{P_n-\lambda_n}{\sigma_n} = \frac{P_n-\lambda_n}{\sqrt{\lambda_n}}\frac{\sqrt{\lambda_n}}{\sigma_n} \underset{n\to\infty}{\overset{\text{loi}}{\longrightarrow}} \mathcal{N}(0,1).$$

□

The preceding proof shows that Poissonization is useful even when the mean diverges: $\lambda_n = p_1 + \cdots + p_n \to \infty$. On the other hand, we have $p_1^2 + \cdots + p_n^2 \not\to 0$, so Le Cam's inequality $d_{\mathrm{VT}}(\mu_n, , \nu_n) = \mathcal{O}(p_1^2 + \cdots + p_n^2)$ does not suffice.

14.3 Extreme tables

At time n, the restaurant has $A_{n,1}$ tables with a single customer. Combinatorics provides the distribution of $A_{n,1}$. Here we consider only the first two moments.

Theorem 14.14 (Solitary customers)**.** *At every time $n \geqslant 1$, the number $A_{n,1}$ of tables containing a single customer satisfies*

$$\mathbb{E}(A_{n,1}) = \frac{n\theta}{n+\theta-1} \qquad \textit{et} \qquad \mathrm{Var}(A_{n,1}) = \frac{n(n-1)(n-2+2\theta)\theta}{(n+\theta-2)(n+\theta-1)^2}.$$

In particular, if $\theta = 1$, then $\mathbb{E}(A_{n,1}) = \mathrm{Var}(A_{n,1}) = 1$ for every $n \geqslant 1$, while for every θ,

$$\lim_{n\to\infty} \mathbb{E}(A_{n,1}) = \theta \qquad \textit{et} \qquad \lim_{n\to\infty} \mathrm{Var}(A_{n,1}) = \theta.$$

Note that for $\theta = 1$, the random permutation σ_n is uniformly distributed on $\mathcal{S}_n$ and has $|\sigma_n| = |\pi_n| \sim \log(n)$ cycles and on average $\mathbb{E}(A_{n,1}) = 1$ fixed point for every $n \geqslant 1$.

Proof. We seek the distribution of the first component $A_{n,1}$ of a random vector A_n in $\mathbb{N}^n$ having the Ewens distribution. However, it is more convenient to view $A_{n,1}$ as the number of blocks of size 1 in π_n. We have $A_{1,1} = 1$ and $0 \leqslant A_{n,1} \leqslant n$ for every $n \in \mathbb{N}^*$. Moreover, for all integers $1 \leqslant a_1, \ldots, a_n \leqslant n$ and $0 \leqslant a_{n+1} \leqslant n+1$,

$$\mathbb{P}(A_{n+1,1} = a_{n+1} \,|\, A_{1,1} = a_1, \ldots, A_{n,1} = a_n) = \mathbb{P}(A_{n+1,1} = a_{n+1} \,|\, A_{n,1} = a_n)$$

which equals

$$\begin{cases} \dfrac{\theta}{\theta+n} & \text{si } a_{n+1} = a_n + 1 \text{ (s'attabler seul à une table vide);} \\ \dfrac{a_n}{\theta+n} & \text{si } a_{n+1} = a_n - 1 \text{ (rejoindre la table d'un solitaire);} \\ \dfrac{n-a_n}{\theta+n} & \text{si } a_{n+1} = a_n \text{ (rejoindre une table comptant 2 clients ou +);} \\ 0 & \text{sinon.} \end{cases}$$

In particular, $(A_{n,1})_{n\geqslant 1}$ is a nonhomogeneous Markov chain with state space $\mathbb{N}$. We obtain the remarkable affine formula[3]

$$\mathbb{E}(A_{n+1,1} \,|\, A_{n,1} = a) = \frac{a(\theta+n-1)+\theta}{\theta+n}.$$

This gives $(n+\theta)m_{n+1} = (\theta+n-1)m_n + \theta$, where $m_n := \mathbb{E}(A_{n,1})$. The stated formula for m_n follows. The variance calculation is similar but more involved, using

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X \,|\, Y)) + \mathrm{Var}(\mathbb{E}(X \,|\, Y))$$

where $\mathrm{Var}(X \,|\, Y) := \mathbb{E}(X^2 \,|\, Y) - \mathbb{E}(X \,|\, Y)^2$ is the conditional variance. □

The integer $A_{n,n}$ represents the number of tables containing n customers, or equivalently the number of blocks of size n in π_n. When such a table exists, it contains every customer and is therefore the unique table. We have $A_{n,n} = 1$ if and only if $|\pi_n| = 1$. The following theorem makes this precise and shows, in particular, that asymptotically the Chinese restaurant model does not have a unique table.

[3] A martingale is hidden in the Markov chain!

Theorem 14.15 (Unique table)**.** *We have* $A_{1,1} = 1$ *and* $A_{n,n} \in \{0, 1\}$ *for every* $n \geqslant 1$*. For every* $n \geqslant 2$*, the probability of having only one table is*

$$\mathbb{P}(A_{n,n} = 1) = \prod_{k=1}^{n-1} \frac{k}{\theta + k}.$$

Finally, the sequence $(A_{n,n})_{n\geqslant 1}$ *decreases almost surely to* 0*.*

Proof. The first assertions follow directly from the definition of π_n. The expression for the probability $\mathbb{P}(A_{n,n} = 1)$ follows by observing that $\mathbb{P}(A_{n,n} = 1) = \mathbb{P}(\xi_2 = 0, \ldots, \xi_n = 0) = \mathbb{P}(\xi_2 = 0) \cdots \mathbb{P}(\xi_n = 0)$. For almost sure convergence, first note that, since $\theta > 0$, the infinite product $\prod_{k=1}^{\infty}(1+\theta k^{-1})$ diverges because the harmonic series diverges, and consequently $\lim_{n\to\infty} \mathbb{P}(A_{n,n} = 1) = 0$. On the other hand, the sequence of events $(E_n)_{n\geqslant 1}$ defined by $E_n = \{A_{n,n} = 0\}$ is nondecreasing, so

$$\mathbb{P}(\lim_{n\to\infty} A_{n,n} = 0) = \mathbb{P}(\cup_{n=1}^{\infty} E_n) = \lim_{n\to\infty} \mathbb{P}(E_n) = 1 - \lim_{n\to\infty} \mathbb{P}(A_{n,n} = 1) = 1.$$

In fact, $A_{n,n} = \mathbb{1}_{\{n<T\}}$, where $T := \inf\{n \geqslant 1 : \xi_n = 1\}$ is the time of the first success in a coin-tossing game whose winning probability changes at each toss and decreases to 0 over time. However, the decrease is slow enough to ensure success with certainty. Indeed, $\{T < \infty\} = \cup_{n=1}^{\infty} E_n$, hence $\mathbb{P}(T < \infty) = 1$. Alternatively, apply the Borel–Cantelli lemma to the independent events $F_n = \{\xi_n = 1\}$, which satisfy

$$\sum_{n\geqslant 1} \mathbb{P}(F_n = 1) = \sum_{n\geqslant 1} \frac{\theta}{\theta + n} = \infty$$

and

$$\overline{\lim}\, F_n = \{\sum_{n\geqslant 1} \mathbb{1}_{\{\xi_n=1\}} = \infty\} \subset \{T < \infty\}.$$

□

14.4 Additional combinatorics

The cardinality of B_n is Bell number n, denoted by B_n (Chapter 4). We have

$$B_n = \sum_{k=1}^{n} \left\{ {n \atop k} \right\}$$

where the notation in braces denotes a Stirling number of the second kind, counting partitions with k blocks of $\{1, \ldots, n\}$. The recurrence formula is

$$\left\{ {n \atop k} \right\} = \left\{ {n-1 \atop k-1} \right\} + k \left\{ {n-1 \atop k} \right\}$$

with boundary conditions

$$\left\{ {n \atop 1} \right\} = 1 \quad \text{et} \quad \left\{ {n \atop n} \right\} = 1$$

, since choosing a partition of $\{1, \ldots, n\}$ with k blocks amounts either to choosing a partition of $\{1, \ldots, n-1\}$ with $k-1$ blocks and adding the singleton block $\{n\}$, or to adding the element n to one of the k blocks of a partition of $\{1, \ldots, n-1\}$ with k blocks. If X is a random variable with Poisson distribution of parameter λ, then

$$\mathbb{E}(X^n) = \sum_{k=1}^{n} \left\{ {n \atop k} \right\} \lambda^k \quad \text{en particulier} \quad \mathbb{E}(X^n) = B_n \quad \text{si } \lambda = 1.$$

We also have the explicit formula

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^n$$

which follows from the inclusion-exclusion principle, since a Stirling number of the second kind equals the number of surjections from $\{1, \ldots, n\}$ to $\{1, \ldots, k\}$ divided by $k!$.

A partition of $\{1, \ldots, n\}$ specifies cycle supports but not the internal structure of each cycle. Several permutations can therefore correspond to a given partition. The number of permutations of $\{1, \ldots, n\}$ with exactly k cycles is the unsigned Stirling number of the first kind, denoted by

$$\left[{n \atop k} \right],$$

, equal to the coefficient of x^k in the Pochhammer polynomial

$$x_{(n)} = x(x+1) \cdots (x+n-1).$$

14.5 Further reading

Most of this chapter is drawn from the expository article [CDM13]. Warren Ewens is a professor of mathematical biology born in Australia in 1937. In the late 1960s, he discovered the distribution now bearing his name while studying a sampling problem in population genetics related to the Wright–Fisher model discussed in Chapter 12. The parameter θ appears as an allele mutation rate, as explained in Chapter 13. Overviews can be found in his book [Ewe04], those of John Kingman [Kin80] and Rick Durrett [Dur08], and

Sylvie Méléard's lecture notes [Mé13]. Many statistical aspects are discussed in Simon Tavaré's course [Tav04]. Ewens' work generated a vast literature in quantitative biology and probability. The Ewens distribution appears in a broad range of so-called logarithmic discrete random structures, from combinatorics to number theory; see the book by Richard Arratia, Andrew Barbour, and Simon Tavaré [ABT03]. The name Chinese restaurant process seems to be due to Jim Pitman. It appears under this name in a course by David Aldous [Ald85]. Fred Hoppe showed in [Hop84] that it can be related to a Pólya-type urn model. It can also be related to Dirichlet processes and random partitions of $[0,1]$ (see below). Today, the Chinese restaurant process and Ewens distribution are standard examples in a more general theory of (fragmentation and) coalescence. See the books by Kingman [Kin93], Jean Bertoin [Ber06], Pitman [Pit06], and Nathanaël Berestycki [Ber09].

Here are two further remarkable representations of the Ewens distribution:

1. If $Z_1, \dots, Z_n$ are independent Poisson random variables with means $\theta/1, \dots, \theta/n$, then

$$\mathrm{Loi}(Z_1, \dots, Z_n \,|\, Z_1 + 2Z_2 + \cdots + nZ_n = n) \sim \mathrm{Ewens}(n, \theta)\,;$$

2. Let $(P_r)_{r\geqslant 1}$ be the random distribution on $\mathbb{N}^*$ generated from the following random partition of the interval $[0,1]$, viewed as a stick of chalk[4]:

$$P_1 = W_1, \quad P_2 = (1-W_1)W_2, \quad P_3 = (1-W_1)(1-W_2)W_3, \dots$$

where $(W_r)_{r\geqslant 1}$ are i.i.d. random variables with distribution $\mathrm{Beta}(1,\theta)$ and density $w \mapsto \theta(1-w)^{\theta-1}\mathbb{1}_{[0,1]}(w)$. Conditional on $(P_r)_{r\geqslant 1}$, let $X_1, \dots, X_n$ be i.i.d. random variables on $\mathbb{N}^*$ with distribution $(P_r)_{r\geqslant 1}$. The sequence $X_1, \dots, X_n$ contains at most n distinct integers. Their counts turn out to have distribution $\mathrm{Ewens}(n,\theta)$!

Finally, one can show that for all $0 \leqslant k \leqslant n$,

$$\mathbb{P}(|\pi_n| = k) = \begin{bmatrix} n \\ k \end{bmatrix} \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

and obtain a similar formula for the distribution of the pair $(A_n, |\pi_n|)$; see the article by Ewens and Tavaré in [JKB97].

[4] This construction is known as *stick breaking*.

15

Reinforcement

Keywords. Reinforcement; urn; random graph; preferential attachment; power law.

Tools. Markov chain; martingale; random walk; exchangeability; exponential distribution.

Difficulty. **

The phenomenon of *reinforcement* is widespread in nature, particularly in genetics, statistical physics, sociology, psychology, and neuroscience. This chapter presents three emblematic models: Pólya's urn, the Barabási–Albert preferential attachment random graph, and a reinforced random walk. It concludes with Rubin's theorem on generalized Pólya urns. Other instances include the Wright–Fisher and Moran models (Chapter 12), Ewens' model (Chapter 14), certain growth models (Chapter 6), and the Ehrenfest model (Chapter 9). In Markovian models, reinforcement often occurs in connection with a partial monotonicity property in a random recurrence.

15.1 Pólya's urn

Pólya's urn is a simple reinforcement model describing, for example, how success or wealth amplifies itself over time.

At time $n = 0$, prepare an urn containing $a > 0$ silver balls and $b > 0$ white balls. To obtain the configuration at time $n = 1$, draw a ball at random, then return it to the urn along with a new ball of the same color. Repeat this mechanism independently to obtain the urn's configuration at every time $n \in \mathbb{N}$. Denote by $M_n \in [0, 1]$ the proportion of silver balls at time $n \in \mathbb{N}$. At time $n \in \mathbb{N}$, the urn contains $a+b+n$ balls, of which $(a+b+n)M_n$ are silver. Conditional on M_n, a silver ball is added at time $n + 1$ with probability M_n. Thus $M_0 = a/(a + b)$, and

$$M_{n+1} = \frac{(a+b+n)M_n + \mathbb{1}_{U_{n+1} \leqslant M_n}}{a+b+n+1}$$

where $(U_n)_{n \geqslant 1}$ is a sequence of independent identically distributed random variables uniform on $[0,1]$, the event $\{U_{n+1} \leqslant M_n\}$ corresponding to the addition of a silver ball. The random recurrent sequence $(M_n)_{n \geqslant 0}$ is both a nonhomogeneous Markov chain with state space $[0,1]$ and a martingale. Encode the result of draw number n by a random variable X_n taking values in $\{\alpha, \beta\}$, where α and β indicate respectively that the drawn ball is silver or white. Finally, denote by Y_n and Z_n the respective numbers of silver and white balls added after the first n draws. At time n, the urn contains $a+Y_n+b+Z_n = a+b+n$ balls, including $a + Y_n = a + \sum_{k=1}^{n} \mathbb{1}_{\{X_k=\alpha\}} = (a+b+n)M_n$ silver balls and $b + Z_n = b + \sum_{k=1}^{n} \mathbb{1}_{\{X_k=\beta\}}$ white balls. We have $Y_n + Z_n = n$. The sequence $(X_n, Y_n, Z_n)_{n \geqslant 1}$ satisfies the random recurrence

$$(X_{n+1}, Y_{n+1}, Z_{n+1}) = \begin{cases} (\alpha, Y_n + 1, Z_n) & \text{si } U_{n+1} \leqslant \frac{a+Y_n}{a+b+n}, \\ (\beta, Y_n, Z_n + 1) & \text{si } U_{n+1} > \frac{a+Y_n}{a+b+n}, \end{cases}$$

with the conventions $Y_0 = 0$ and $Z_0 = 0$ (there is no need to define X_0).

Theorem 15.1 (Martingale)**.** *The sequence $(M_n)_{n \geqslant 0}$ is a martingale taking values in $[0,1]$ for the filtration $(\mathcal{F}_n)_{n \geqslant 0}$ defined by $\mathcal{F}_n = \sigma(U_1, \ldots, U_n)$. In particular, the mean proportion of silver balls is preserved over time: for every $n \in \mathbb{N}$,*

$$\mathbb{E}(M_n) = \mathbb{E}(M_0) = \frac{a}{a+b}.$$

Moreover, there exists a random variable M_∞ on $[0,1]$ such that

$$\lim_{n \to \infty} M_n = M_\infty$$

a.s. and in L^p for every $p \geqslant 1$. In particular, $\mathbb{E}(M_\infty) = \mathbb{E}(M_0) = \frac{a}{a+b}$.

Proof. For every $n \in \mathbb{N}$, M_n is $\mathcal{F}_n$-measurable, takes values in $[0,1]$, and is therefore bounded and integrable, and

$$\mathbb{E}(M_{n+1} \,|\, \mathcal{F}_n) = \mathbb{E}(M_{n+1} \,|\, M_n) = \frac{(a+b+n)M_n + M_n}{a+b+n+1} = M_n.$$

Thus $(M_n)_{n \geqslant 0}$ is a martingale for $(\mathcal{F}_n)_{n \geqslant 0}$. The convergence theorem for uniformly bounded martingales gives $M_n \to M_\infty$ almost surely and in L^p, where M_∞ is a random variable taking values in $[0,1]$[1]. □

[1] The martingale is uniformly integrable. Alternatively, use the convergence theorem for nonnegative martingales to obtain almost sure convergence, then the dominated convergence theorem for convergence in L^p.

Theorem 15.2 (Equilibrium)**.** *The random variable M_∞ appearing in Theorem 15.1 has the beta distribution on $[0,1]$ with parameter (a,b) and density*

$$u \in [0,1] \mapsto \frac{u^{a-1}(1-u)^{b-1}}{\mathrm{Beta}(a,b)} \quad \text{où} \quad \mathrm{Beta}(a,b) := \int_0^1 p^{a-1}(1-p)^{b-1}\,dp.$$

In particular, if $a = b = 1$, then M_∞ is uniformly distributed on $[0,1]$.

Proof. First recall that

$$\mathrm{Beta}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \text{où} \quad \Gamma(x) := \int_0^\infty t^{x-1}e^{-x}\,dx.$$

Now, for all c and k, write

$$c^{(k)} = c(c+1)\cdots(c+k-1) = \frac{(c+k-1)!}{(c-1)!} = \frac{\Gamma(c+k)}{\Gamma(c)}.$$

For all $x_1, \ldots, x_n$ in $\{0,1\}$, writing $k = x_1 + \cdots + x_n$, we have

$$\mathbb{P}(\mathbb{1}_{\{X_1=\alpha\}} = x_1, \ldots, \mathbb{1}_{\{X_n=\alpha\}} = x_n) = \frac{a^{(k)}b^{(n-k)}}{(a+b)^{(n)}}.$$

This probability is invariant under permutations of the $x_1, \ldots, x_n$: the distribution of the random vector $(\mathbb{1}_{\{X_1=\alpha\}}, \ldots, \mathbb{1}_{\{X_n=\alpha\}})$ is exchangeable. Thus the (random) number $Y_n = \sum_{k=1}^n \mathbf{1}_{\{X_k=\alpha\}}$ of silver balls drawn in the first n draws satisfies, for every $k \in \{0,1,\ldots,n\}$,

$$\begin{aligned}
\mathbb{P}(Y_n = k) &= \binom{n}{k}\frac{a^{(k)}b^{(n-k)}}{(a+b)^{(n)}} \\
&= \binom{n}{k}\frac{\Gamma(a+k)\Gamma(b+n-k)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)} \\
&= \binom{n}{k}\frac{\mathrm{Beta}(a+k,b+n-k)}{\mathrm{Beta}(a,b)} \\
&= \int_0^1 \binom{n}{k} p^k(1-p)^{n-k}\frac{p^{a-1}(1-p)^{b-1}}{\mathrm{Beta}(a,b)}\,dp.
\end{aligned}$$

We say that Y_n has the beta-binomial distribution, a mixture of binomial distributions of size n whose parameter p has the distribution Beta with parameter (a,b). We have $M_n = (a+Y_n)/(a+b+n)$ with $Y_0 = 0$. When $a = b = 1$, the formula for the distribution of Y_n shows that Y_n is uniform on $\{0,1,\ldots,n\}$, so M_n is uniform on $\{1/(n+2), \ldots, (n+1)/(n+2)\}$, implying that M_∞ is uniform on $[0,1]$. In the general case, the beta-binomial correspondence (Chapter 1) shows that, for every $t \in [0,1]$,

$$\mathbb{P}(M_\infty \leqslant t) = \lim_{n\to\infty} \mathbb{P}(Y_n \leqslant (a+b+n)t - a)$$

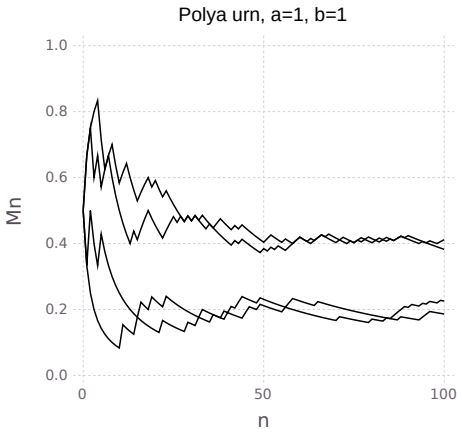


Fig. 15.1. Some sample paths of Pólya's urn.

$$= \cdots = \frac{1}{\mathrm{Beta}(a,b)} \int_0^t u^{a-1}(1-u)^{b-1}\, du.$$

□

Theorems 15.1 and 15.2 extend to the case where a and b are real numbers > 0, which in particular motivates the use of gamma and beta functions instead of factorials. In this case, the urn can be replaced by a partition of an interval.

Remark 15.3 (Generalized Pólya urn)**.** *Generalize the reinforcement mechanism as follows: fix an integer $r \geqslant -1$, and at each draw return $1 + r$ balls of the drawn color to the urn.*

- *If $r = 1$, we recover the Pólya urn studied above, for which Y_n has the beta-binomial distribution;*
- *if $r = 0$, we have sampling with replacement and Y_n is binomial;*
- *if $r = -1$, we have sampling without replacement and Y_n is hypergeometric;*
- *in general, for every $r \geqslant -1$ and every $k \in \{0, 1, \ldots, n\}$, we have*

$$\mathbb{P}(Y_n = k) = \frac{a^{(r,k)} b^{(r,k)}}{(a+b)^{(r,k)}} \quad \text{où} \quad c^{(r,k)} := c(c+r)\cdots(c+(k-1)r).$$

In this case, M_∞ has the beta distribution with parameter $(a/r, b/r)$.

An arbitrary number of colors can be considered, giving a model that includes sampling from the multitype hypergeometric distribution. Many reinforcement models can be obtained as generalized versions of Pólya's urn.

Remark 15.4 (Replacement matrix)**.** *Let $A = (a_{i,j})_{1 \leqslant i,j \leqslant k}$ be a $k \times k$ matrix of integers. Consider a generalized Pólya urn with k colors evolving as follows: draw a ball at random, note its color i, then return $a_{i,j}$ balls of color j to the urn for every $1 \leqslant j \leqslant k$. We call A the urn's replacement matrix. For the standard Pólya urn studied above, $k = 2$ and $A = 2I_2$.*

15.2 Barabási–Albert graph

Random graphs model many natural phenomena, such as friendship structures in social networks, links between pages on the World Wide Web, collaboration structures in artistic and scientific production, regulatory structures among proteins, connections between machines on the Internet, and so on. The Galton–Watson trees of Chapter 3 provide a random graph model suited to genealogical structures. The best-known and simplest random graph model is probably the Erdős–Rényi model discussed in Chapter 16: it is built recursively by adding a new vertex and then independently tossing a coin to decide its connection to each existing vertex. This model does not match real social random graphs, where new vertices preferentially connect to the most important existing vertices in terms of connectivity (degree). This is an instance of reinforcement that must be specifically accounted for.

The Barabási–Albert *preferential attachment random graph* is defined as follows: at time $n \geqslant 1$, the graph contains n sites (vertices) and a certain number of undirected links (edges) between them (see Figure 15.2). The degree of a vertex is the number of edges incident to it. At time $n = 1$, vertex 1 is connected to itself. This initialization yields elegant formulas but is not canonical; other initializations are possible. The degree of vertex 1 at time 1 is therefore 2. To evolve the graph recursively from time n to time $n + 1$, consider the degrees $d_{n,1}, \ldots, d_{n,n}$ of its n vertices at time n and the associated probability distribution

$$p_{n,k} = \frac{d_{n,k}}{d_{n,1} + \cdots + d_{n,n}},$$

Then connect the new vertex $n + 1$ to a vertex chosen randomly and independently among the n existing vertices, with probability distribution $p_{n,\cdot}$. This mechanism gives $d_{1,1} = 2$, $d_{2,1} = 3$, $d_{2,2} = 1$, and, for every $n \geqslant 1$, $d_{n,1} + \cdots + d_{n,n} = 2n$ (that is, n edges).

Remark 15.5 (Urn)**.** *This construction of the sequence $(d_{n,\cdot})_{n \geqslant 1}$ can be realized (discarding the graph's geometry) as a generalized Pólya urn model: at time $n \geqslant 1$ the urn contains $2n$ balls whose colors range from 1 to n. Draw a ball at random; if its color is k, return 2 balls of color k together with a new ball of color $n + 1$. This reinforces the vertex of color k and introduces a new vertex of color $n + 1$; in other words, the graph gains one vertex $(n + 1)$ and one edge $(k \leftrightarrow n + 1)$.*

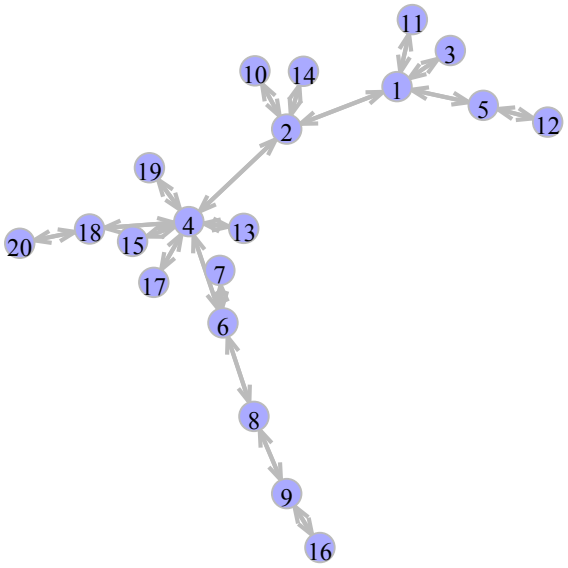


Fig. 15.2. A realization of a Barabási–Albert preferential attachment graph.

If the edge connecting 1 to itself is omitted, the Barabási–Albert preferential attachment graph has no cycles: it is a tree.

Theorem 15.6 (Power law)**.** *Fix $k > 1$. Let $d_{n,k}$ be the degree of vertex k in the Barabási–Albert preferential attachment random graph with n vertices, with $n \geqslant k$. Then the proportion $d_{n,k}/(d_{n,1} + \cdots + d_{n,k})$ converges almost surely as $n \to \infty$ to a random variable with beta distribution of parameter $(1, 2k-1)$ on $[0,1]$ and density $u \in [0,1] \mapsto (2k-1)(1-u)^{2(k-1)}$.*

The power law appearing in this preferential attachment model agrees well with real social networks and differs from the subexponential behavior of the Poisson distribution in Erdős–Rényi random graph models with nonpreferential attachment[2].

Proof. As in Remark 15.5, encode the evolution of the degree sequence (ignoring the underlying graph structure) by an urn model. At time k, introduce an additional coloring: the $2k-1$ balls whose color is less than k are white, while ball number $2k$ is silver. At time $n > k$, track only the evolution of balls

[2] In an Erdős–Rényi graph with parameter (n, p), each vertex has a random number of neighbors with binomial distribution $\mathrm{Bin}(n-1, p)$, converging to the Poisson distribution $\mathrm{Poi}(\lambda)$ if $np \to \lambda$ as $n \to \infty$. Now, if $X \sim \mathrm{Poi}(\lambda)$, then $\mathbb{P}(X \geqslant r) \leqslant C \exp(-cr \log(r))$ for $r \gg 1$, where $c > 0$ and $C > 0$ are constants.

with these colors. This corresponds to a Pólya urn with initial composition $a = 1$ and $b = 2k - 1$. We can then apply Theorem 15.2. □

Theorem 15.7 (Power law at a fixed degree)**.** *For all $n, d \geqslant 1$, if $N(n, d)$ denotes the number of vertices of degree d at time n in the Barabási–Albert preferential attachment random graph with n vertices, then*

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{N(n,d)}{n}\right) = \frac{4}{d(d+1)(d+2)}.$$

Thus, in a very large preferential attachment graph ($n \gg 1$), the probability that an edge has degree d decays polynomially in d^{-3} as $d \to \infty$. The heavy tail of the mean degree distribution is linked to the presence of highly connected vertices caused by reinforcement.

Proof. The sequence $(N(n,\cdot))_{n\geqslant 1}$ is a Markov chain. Conditional on $N(n,\cdot)$, creating a new vertex of degree d requires adding an edge to a vertex of degree $d-1$, which occurs with probability $(d-1)N(n,d-1)/(2n)$, whereas a vertex of degree d disappears from this category if an edge is added to it, which occurs with probability $dN(n,d)/(2n)$. Finally, a vertex of degree 1 is created at each step by construction. Thus the mean number $m_n(d) := \mathbb{E}(N(n,d))$ of vertices of degree d at time n satisfies a linear recurrence, called a master equation by Dorogovstev, Mendes, and Samukhin: for every $n, d \geqslant 1$,

$$m_{n+1}(d) - m_n(d) = -\frac{d}{2n} m_n(d) + \frac{d-1}{2n} m_n(d-1) + \mathbb{1}_{d=1},$$

with initial condition $m_1 = \mathbb{1}_{d=2}$. For $d = 1$, the equation becomes

$$m_{n+1}(1) = c + \left(1 - \frac{b}{n}\right) m_n(1) \quad \text{avec } c = 1 \text{ et } b = 1/2,$$

which gives

$$\begin{aligned} m_{n+1}(1) &= c + \left(1 - \frac{b}{n}\right)c + \left(1 - \frac{b}{n}\right)\left(1 - \frac{b}{n-1}\right) m_{n-1}(1) \\ &= c \sum_{k=1}^{n} \prod_{j=k+1}^{n} \left(1 - \frac{b}{j}\right) + \underbrace{m_1(1)}_{=0} \prod_{k=1}^{n} \left(1 - \frac{b}{k}\right). \end{aligned}$$

Now

$$\prod_{j=k+1}^{n} \left(1 - \frac{b}{j}\right) \approx \exp\Big(- \sum_{k=k+1}^{n} \frac{b}{j}\Big) \approx \exp\Big(- b(\log(n) - \log(k))\Big) = \Big(\frac{k}{n}\Big)^b,$$

hence

$$m_n(1) \approx cn^{-b} \int_0^n s^b \, ds = cn^{-b} \frac{n^{b+1}}{b+1} = \frac{cn}{b+1},$$

so that

$$\lim_{n\to\infty} \frac{m_n(1)}{n} = \frac{c}{b+1} = \frac{2}{3}.$$

More generally, for every $d > 1$, we have

$$m_{n+1}(d) = c_n(d) + \left(1 - \frac{b(d)}{n}\right) m_n(d)$$

where

$$b(d) = d/2 \quad \text{et} \quad c_n(d) = \frac{d-1}{2} \frac{m_n(d-1)}{n},$$

and induction on d shows that $c(d) = \lim_{n\to\infty} c_n(d)$ exists and

$$\frac{m_n(d)}{n} \underset{n\to\infty}{\longrightarrow} \frac{c(d)}{b(d)+1}.$$

Next, setting

$$\ell(d) := \lim_{n\to\infty} \frac{m_n(d)}{n}$$

gives $\ell(1) = 2/3$, while for every $d > 1$,

$$\ell(d) = \frac{\frac{(d-1)}{2}\ell(d-1)}{\frac{d}{2}+1} = \frac{d-1}{d+2}\ell(d-1).$$

Finally, the solution of this recurrence in d is (easy to check!)

$$\ell(d) = \frac{4}{d(d+1)(d+2)}.$$

□

15.3 Reinforced random walk

Reinforcement is at work in the formation of paths used by hikers in the mountains or ants on the ground. Consider two distinct paths α and β connecting the same starting and ending points, used by successive passersby. For every $n \geqslant 1$, encode by a random variable X_n taking values in $\{\alpha, \beta\}$ the path used on passage number n. The random variables A_n and B_n encode the attractiveness of paths α and β at passage $(n+1)^{\text{e}}$: for humans, this could reflect grass being worn away; for ants, the amount of pheromone. Fix (A_0, B_0) and a function $r :]0, \infty[\to]0, \infty[$, called the *reinforcement function*, such that $r(x) \geqslant x$ for every $x > 0$, and model $(X_n)_{n\geqslant 1}$ by

$$(X_{n+1}, A_{n+1}, B_{n+1}) = \begin{cases} (\alpha, r(A_n), B_n) & \text{si } U_{n+1} \leqslant \frac{A_n}{A_n+B_n}; \\ (\beta, A_n, r(B_n)) & \text{si } U_{n+1} > \frac{A_n}{A_n+B_n}, \end{cases}$$

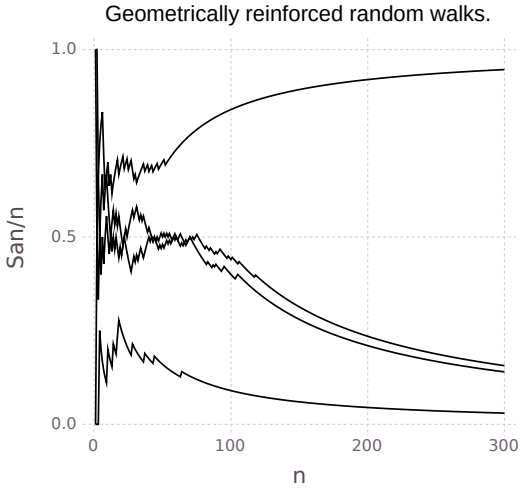


Fig. 15.3. Occupation of a path for the geometrically reinforced walk.

where $(U_n)_{n\geqslant 1}$ is a sequence of independent identically distributed random variables uniform on $[0,1]$. The random recurrent sequence $((X_n, A_n, B_n))_{n\geqslant 0}$ is a Markov chain (the value of X_0 plays no role in the recurrence). When A_0, B_0, and r take integer values, everything behaves as if an urn contained, at every time n, A_n silver balls and B_n white balls: draw a ball at random and then place $r(A_n)$ balls of the same color in the urn. This gives sampling without replacement if $r(x) = x$, Pólya's urn if $r(x) = x + 1$, and a kind of nonlinear Pólya urn in general.

The sequence $(X_n)_{n\geqslant 1}$ is a non-Markovian random walk on the two-point set $\{\alpha, \beta\}$. Its transitions depend on one another through the reinforcement mechanism associated with the time spent at each site. It is called a *vertex-reinforced random walk*. The reinforcement mechanism can also be viewed as a kind of stochastic algorithm.

For simplicity, assume that $A_0 = B_0 = 1$ (X_0 is unnecessary). The numbers of passages along path α and path β at time n are given by

$$Y_n = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=\alpha\}} \quad \text{et} \quad Z_n = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=\beta\}}.$$

For convenience, set $Y_0 = 0$ and $Z_0 = 0$, so that $Y_n + Z_n = n$ for every $n \in \mathbb{N}$. The following theorem describes the behavior of these two sequences for three reinforcement functions r.

Theorem 15.8 (Almost sure asymptotic behavior as $n \to \infty$)**.**

- No reinforcement. *If* $r(x) = x$ *for every* $x > 0$*, then almost surely* $\liminf |Y_n - Z_n| = 0$*, and*

$$Y_n \sim \frac{n}{2}, \quad Z_n \sim \frac{n}{2}, \quad \limsup \frac{|Y_n - Z_n|}{\sqrt{2n \log(\log(n))}} = 1.$$

- Linear reinforcement. *If* $r(x) = x+1$ *for every* $x > 0$*, there exists a random variable* U *uniform on* $[0,1]$ *such that almost surely,*

$$\frac{Y_n}{n} \to U \quad et \quad \frac{Z_n}{n} \to 1 - U.$$

- Geometric reinforcement. *If* $r(x) = \rho x$ *for every* $x > 0$*, where* $\rho > 1$ *is a constant, then almost surely the random sequence* $(X_n)_{n \geqslant 0}$ *is eventually constant on* n*, and its limiting value has the symmetric Bernoulli distribution on* $\{\alpha, \beta\}$*.*

Without reinforcement or with linear reinforcement, the frequency of use of each path converges over time to $1/2$ in the first case and to a uniform random number on $[0,1]$ in the second. Intuitively, superlinear reinforcement might force these frequencies to converge to the extreme values 0 or 1. This is confirmed for geometric reinforcement, where after some time one of the two paths is always used.

The linear reinforcement model coincides with the Pólya urn studied above. Geometric reinforcement is particularly appealing because a preferred path emerges, chosen randomly through reinforcement. It is a kind of generalized nonlinear, specifically superlinear, Pólya urn. Rubin's Theorem 15.9 below gives a criterion on the reinforcement for this phenomenon to occur.

Proof. We treat the three cases separately.

- *No reinforcement.* In this case, $(A_n)_{n \geqslant 0}$ and $(B_n)_{n \geqslant 0}$ are constant sequences equal to 1. We recover sampling with replacement, or the Bernoulli process (coin tossing). The sequence $(Y_n)_{n \geqslant 0}$ is a Bernoulli process on $\mathbb{N}$ starting from 0, whose increments have Bernoulli distribution on $\{0,1\}$ with parameter $A_0/(A_0 + B_0) = 1/2$. By the strong law of large numbers, almost surely $Y_n/n \to 1/2$ and $Z_n/n = 1 - Y_n/n \to 1/2$ as $n \to \infty$. The sequence $(Y_n - Z_n)_{n \geqslant 0}$ is a random walk on $\mathbb{Z}$ starting from 0, whose increments have Rademacher distribution on $\{-1,1\}$ with parameter $1/2$. It is an irreducible recurrent Markov chain: almost surely every state is visited infinitely often, particularly state 0, hence $\underline{\lim}\, |Y_n - Z_n| = 0$ a.s. The result concerning $\overline{\lim}$ follows from Strassen's law of the iterated logarithm;

- *Linear reinforcement.* In this case, $A_n = 1 + Y_n$ and $B_n = 1 + Z_n$ for every $n \in \mathbb{N}$. We recover Pólya's urn, and the desired result follows directly from the uniform case in Theorem 15.2. Nevertheless, let us give the simplified argument in the present notation. The relation $Y_n + Z_n = n$ reduces the

problem to studying Y_n. Let $(\mathcal{F}_n)_{n\geqslant 0}$ be the filtration defined by $\mathcal{F}_n = \sigma(U_1, \dots, U_n)$. We have

$$\begin{aligned}\mathbb{E}(1 + Y_{n+1} \,|\, \mathcal{F}_n) &= 1 + Y_n + \mathbb{E}(\mathbb{1}_{\{X_{n+1}=\alpha\}} \,|\, \mathcal{F}_n) \\ &= 1 + Y_n + \mathbb{E}(\mathbb{1}_{\{U_{n+1}\leqslant \frac{1+Y_n}{n+2}\}} \,|\, \mathcal{F}_n) \\ &= 1 + Y_n + \frac{1+Y_n}{n+2} = (2 + (n+1))\frac{1+Y_n}{2+n},\end{aligned}$$

so $((1 + Y_n)/(n + 2))_{n\geqslant 0}$ is a martingale for $(\mathcal{F}_n)_{n\geqslant 0}$. Taking values in $[0, 1]$, it is uniformly bounded and therefore converges almost surely (and in mean) to a random variable U taking values in $[0, 1]$. Finally, induction on n shows that $(1 + Y_n)/(n + 2)$ is uniformly distributed on $\{1/(n + 2), \dots, (n + 1)/(n + 2)\}$;

- *Geometric reinforcement.* In this case, $A_n = \rho^{Y_n}$ and $B_n = \rho^{Z_n}$ for every $n \in \mathbb{N}$. The random variable $\Delta_n := Y_n - Z_n$ satisfies $\Delta_0 = 0$ and

$$\Delta_{n+1} = \Delta_n + \mathbb{1}_{\{U_{n+1}\leqslant 1/(1+\rho^{-\Delta_n})\}} - \mathbb{1}_{\{U_{n+1}>1/(1+\rho^{-\Delta_n})\}}.$$

Set $\mathcal{F}_n = \sigma(U_1, \dots, U_n)$. It follows that

$$\begin{aligned}\mathbb{P}(|\Delta_{n+1}| = |\Delta_n| + 1 \,|\, \mathcal{F}_n) &= \frac{\rho^{|\Delta_n|}}{1 + \rho^{|\Delta_n|}}\mathbb{1}_{\{\Delta_n\neq 0\}} + \mathbb{1}_{\{\Delta_n=0\}} \\ \mathbb{P}(|\Delta_{n+1}| = |\Delta_n| - 1 \,|\, \mathcal{F}_n) &= \frac{1}{1 + \rho^{|\Delta_n|}}\mathbb{1}_{\{\Delta_n\neq 0\}}.\end{aligned}$$

If $f : \mathbb{N} \to \mathbb{R}$ satisfies $f(1) \geqslant f(0)$ and, for every $n \geqslant 1$,

$$\rho^n(f(n + 1) - f(n)) = f(n) - f(n - 1),$$

then

$$\mathbb{E}(f(|\Delta_{n+1}|) \,|\, \mathcal{F}_n) = f(|\Delta_n|) + \mathbb{1}_{\{\Delta_n=0\}} \geqslant f(|\Delta_n|),$$

and $(f(|\Delta_n|))_{n\geqslant 0}$ is a submartingale. Now consider the special case $f(0) = 0$ and, for every $n \geqslant 1$,

$$f(n) = \sum_{k=0}^{n-1} \rho^{-\frac{k(k+1)}{2}}.$$

Since f is bounded, the submartingale $f(|\Delta|)$ is uniformly bounded, so it converges a.s. (and in mean). Since f is injective and $|\Delta|$ takes values in $\mathbb{N}$, the values taken by $f(|\Delta|)$ are discrete. Since $f(|\Delta|)$ converges a.s., it can converge only to the sole accumulation point of $f(\mathbb{N})$, namely $\lim_{n\to\infty} f(n) = \sum_{k=0}^{\infty} \rho^{-\frac{k(k+1)}{2}}$. Thus a.s. $\lim_{n\to\infty} |\Delta_n| = \infty$. Now, for all $N, d \geqslant 0$, the probability that the sequence $|\Delta|$ is nondecreasing from index N onward, conditional on $|\Delta_N| = d$, is

$$\mathbb{P}(\forall n \geqslant N : |\Delta_{n+1}| = |\Delta_n| + 1 \,|\, |\Delta_N| = d) = \prod_{k=d}^{\infty} \frac{1}{1+\rho^{-k}} =: p_d.$$

The probability p_d does not depend on N and increases to 1 as $d \to \infty$. Thus the probability that $|\Delta|$ is eventually nondecreasing (from a random index onward) can be written as

$$\mathbb{P}(\exists N \geqslant 0, \forall n \geqslant N : |\Delta_{n+1}| = |\Delta_n| + 1)$$

It is bounded below by

$$\sup_{N \geqslant 0} \mathbb{P}(\forall n \geqslant N, |\Delta_{n+1}| = |\Delta_n| + 1) = \sup_{N \geqslant 0} \sum_{d=0}^{\infty} p_d \mathbb{P}(|\Delta_N| = d).$$

For every $d \geqslant 0$, we have $\lim_{N\to\infty} \mathbb{P}(|\Delta_N| = d) = 0$, since we know that $\lim_{N\to\infty} |\Delta_N| = \infty$ a.s. Consequently, for every $D \geqslant 0$,

$$\sup_{N \geqslant 0} \sum_{d=0}^{\infty} p_d \mathbb{P}(|\Delta_N| = d) \geqslant \sup_{N \geqslant 0} \sum_{d=D}^{\infty} p_d \mathbb{P}(|\Delta_N| = d) \geqslant p_D.$$

Since $P_D \to 1$ as $D \to \infty$, it follows that almost surely $|\Delta|$ is eventually nondecreasing, and hence the sequence X is absorbed at α or β. The distribution of the limit of X is symmetric by the model's symmetry in α and β, since $A_0 = B_0$.

□

15.4 Rubin's theorem

This section is devoted to Rubin's theorem on generalized Pólya urns. Fix two nondecreasing functions $S_\alpha, S_\beta : \mathbb{N} \to \mathbb{R}_+$ such that $S_\alpha(0) > 0$ and $S_\beta(0) > 0$. Construct a random recurrent sequence $(X_n)_{n\geqslant 1}$ taking values in $\{\alpha, \beta\}$ as follows: for every $n \in \mathbb{N}$, conditional on $X_1, \ldots, X_n$, writing $Y_n = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=\alpha\}}$ for the number of α and $Z_n = n - Y_n$ for the number of β,

$$(X_{n+1}, Y_{n+1}, Z_{n+1}) = \begin{cases} (\alpha, Y_n + 1, Z_n) & \text{si } U_{n+1} \leqslant \frac{S_\alpha(Y_n)}{S_\alpha(Y_n)+S_\beta(Z_n)}, \\ (\beta, Y_n, Z_n + 1) & \text{si } U_{n+1} > \frac{S_\alpha(Y_n)}{S_\alpha(Y_n)+S_\beta(Z_n)}, \end{cases}$$

where $(U_n)_{n\geqslant 1}$ is a sequence of i.i.d. random variables uniform on $[0,1]$. The sequence $(X_n)_{n\geqslant 1}$ encodes successive draws from a generalized Pólya urn. The model of Theorem 15.8, when (A_0, B_0) is deterministic, is obtained with $(S_\alpha(0), S_\beta(0)) = (A_0, B_0)$ and $S_\alpha(n) = S_\beta(n) = r(n)$ for every $n > 0$. If r is linear, we recover the absence of reinforcement (sampling with replacement).

If r is affine, we recover linear reinforcement (the standard Pólya urn). Finally, if r is a power function, we recover geometric reinforcement.

To study the general case, introduce the probability p_α (respectively p_β) that the sequence $(X_n)_{n\geqslant 1}$ contains only finitely many β (respectively α), that is, only α (respectively only β) from some index onward on n. These probabilities are given by

$$p_\alpha = \mathbb{P}(\cup_n \cap_{k\geqslant n} \{X_k = \alpha\}) \quad \text{et} \quad p_\beta = \mathbb{P}(\cup_n \cap_{k\geqslant n} \{X_k = \beta\}).$$

We have $p_\alpha + p_\beta \leqslant 1$. The following numbers in $[0,\infty]$ play a crucial role:

$$\varphi_\alpha = \sum_{n=0}^{\infty} \frac{1}{S_\alpha(n)} \quad \text{et} \quad \varphi_\beta = \sum_{n=0}^{\infty} \frac{1}{S_\beta(n)}.$$

Theorem 15.9 (Rubin)**.** *The values of p_α, p_β are related to $\varphi_\alpha, \varphi_\beta$:*

	$\boldsymbol{\varphi_\beta < \infty}$	$\boldsymbol{\varphi_\beta = \infty}$
$\boldsymbol{\varphi_\alpha < \infty}$	$p_\alpha > 0$, $p_\beta > 0$, $p_\alpha + p_\beta = 1$	$p_\alpha = 1$ *(hence* $p_\beta = 0$*)*
$\boldsymbol{\varphi_\alpha = \infty}$	$p_\beta = 1$ *(hence* $p_\alpha = 0$*)*	$p_\alpha = 0$, $p_\beta = 0$

Thus, whether the probabilities p_α and p_β are trivial depends only on the asymptotic behavior of the reinforcement functions S_α and S_β, and the critical cases are linked to the criterion for convergence of a Riemann series. Without reinforcement (S_α and S_β linear) and with linear reinforcement (S_α and S_β affine), we have $\varphi_\alpha = \varphi_\beta = \infty$ because the harmonic series diverges. As soon as reinforcement is superlinear, $\varphi_\alpha < \infty$ and $\varphi_\beta < \infty$, in particular for quadratic reinforcement and, of course, geometric reinforcement.

Proof. Let $(E_n^\alpha)_{n\geqslant 0}$ and $(E_n^\beta)_{n\geqslant 0}$ be two independent sequences of independent exponentially distributed random variables with, for every $n \geqslant 0$,

$$\mathbb{E}(E_n^\alpha) = \frac{1}{S_\alpha(n)} \quad \text{et} \quad \mathbb{E}(E_n^\beta) = \frac{1}{S_\beta(n)}$$

Now define the random sets

$$\mathcal{A} = \left\{ \sum_{k=0}^{n} E_k^\alpha : n \geqslant 0 \right\} \quad \text{et} \quad \mathcal{B} = \left\{ \sum_{k=0}^{n} E_k^\beta : n \geqslant 0 \right\} \quad \text{et} \quad \mathcal{G} = \mathcal{A} \cup \mathcal{B}.$$

Let $\xi_0 < \xi_1 < \cdots$ be the elements of $\mathcal{G}$ arranged in increasing order. Consider the random sequence $(X'_n)_{n\geqslant 1}$ taking values in $\{\alpha, \beta\}$ defined by $X'_n = \alpha$ if $\xi_{n-1} \in \mathcal{A}$ and $X'_n = \beta$ if $\xi_{n-1} \in \mathcal{B}$. The sequences $(X'_n)_{n\geqslant 1}$ and $(X_n)_{n\geqslant 1}$ have the same distribution, by properties of exponential distributions including memorylessness. Examine the equality in distribution of X_1 and X'_1. If U and V are independent exponentially distributed random variables with means $1/u$ and $1/v$, then $\mathbb{P}(U < V) = u/(u+v)$ and $\mathbb{P}(V < U) = v/(u+v)$, so

$$\mathbb{P}(X'_1 = \alpha) = \mathbb{P}(\xi_0 \in \mathcal{A}) = \mathbb{P}(E_0^\alpha \leqslant E_0^\beta) = \frac{S_\alpha(0)}{S_\alpha(0) + S_\beta(0)} = \mathbb{P}(X_1 = \alpha).$$

The same idea gives (with some work!) equality in distribution of $(X_n)_{n\geqslant 1}$ and $(X'_n)_{n\geqslant 1}$.

Now a zero-one law for exponential distributions (Theorem 11.5) states that $p := \mathbb{P}(\sum_{n=0}^{\infty} E_n^\alpha < \infty) \in \{0, 1\}$, with $p = 1$ if and only if $\varphi_\alpha < \infty$. The same property holds for the sequence $(E_n^\beta)_{n\geqslant 0}$ with φ_β. Moreover, we have the formulas $p_\beta = \mathbb{P}(\sum_{n=0}^{\infty} E_n^\alpha < \sum_{n=0}^{\infty} E_n^\beta)$ and $p_\alpha = \mathbb{P}(\sum_{n=0}^{\infty} E_n^\beta < \sum_{n=0}^{\infty} E_n^\alpha)$. □

15.5 Further reading

Pólya urns are studied in particular in the books by Hosam Mahmoud [Mah09] and Norman Johnson and Samuel Kotz [JK77], and in the surveys by Samuel Kotz and Narayanaswamy Balakrishnan [KB97] and Robin Pemantle [Pem07]. They are named after the Hungarian mathematician George Pólya. The Hungarian school greatly developed random discrete mathematics, notably through Paul Erdős and Alfréd Rényi's celebrated work on random graphs in the 1950s. The preferential attachment random graph model was introduced by the Hungarian physicists Albert-László Barabási and Réka Albert [AB99, AB02]. Here is their heuristic for the behavior in d^{-3}: if degrees and time were continuous, we would have the evolution equation

$$\partial_t d_{t,k} = \frac{d_{t,k}}{\sum_k d_{t,k}} = \frac{d_{t,k}}{2t},$$

giving the formula $d_{t,k} = (t/t_k)^{1/2}$, where t_k is the arrival time of vertex k. Since vertices are added uniformly on $[0, t]$, we obtain $\mathbb{P}(d_{t,k} > d) = \mathbb{P}(t_k < t/d^2) = 1/d^2$, and differentiation yields the simple formula $\mathbb{P}(d_{t,k} = d) = 2/d^3$, the polynomial (power-law) behavior suggested by simulations. This approach lacks rigor; in particular, the constant 2 is incorrect. The Barabási–Albert model was subsequently studied by Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády [BRST01]. It can be viewed as an instance of self-organization. The topic remains active at the time of writing, as illustrated by the article [BMR14] by Sébastien Bubeck, Elchanan Mossel, and Miklós Rácz, or [CDKM14] by Nicolas Curien, Thomas Duquesne, Igor Kortchemski, and Ioan Manolescu. This chapter's treatment of preferential attachment random graphs is inspired by Thierry Bodineau's course [Bod14] at the École Polytechnique and the books by Remco van der Hofstad [vdH14] and Rick Durrett [Dur10b]. The Azuma–Hoeffding inequality from Chapter 20 can be used to show that $N(n, d)$ (without expectation) behaves polynomially in d as $n \to \infty$. The preferential attachment rule can also be modified to obtain polynomial behavior with degree other than 3, corresponding to a modified (nonlinear) Pólya urn. Preferential attachment also appears in the branching process of Herbert Simon and Udny Yule. The discrete distribution with polynomial decay, the discrete version of the (Vilfredo) Pareto distribution, is known as the (George Kingsley) Zipf distribution.

The treatment of reinforced random walks is inspired by a text by Thierry Lévy. More generally, we refer to Robin Pemantle's survey [Pem07] on reinforced random processes and Pierre Tarrès' survey [Tar11] on reinforced random walks. A recent result on reinforcement can be found in the article by Margherita Disertori, Christophe Sabot, and Pierre Tarrès [DST14]. A continuous reinforcement model is studied by Michel Benaïm and Olivier Raimond in [BR11]. Herman Rubin's Theorem 15.9 appears in an appendix to an article by Burgess Davis [Dav90].

Beyond these models, reinforcement plays an important role in learning theory. It is also related to a range of stochastic algorithms, such as the Robbins–Monro stochastic approximation algorithm, linked to dynamical systems and their attractors. Martingales are a central tool in the stochastic approach to game theory and strategies, briefly presented in the book by Michel Benaïm and Nicole El Karoui [BEK05].

16

Percolation

Keywords. Percolation; threshold phenomenon; random graph; random tree; Bethe tree; Euclidean lattice; complete graph; Erdős–Rényi random graph.

Tools. Coupling; Fekete's subadditivity lemma; Kolmogorov's zero-one law; random walk; Galton–Watson process; martingale; optional stopping theorem; Markov's inequality; Laplace transform.

Difficulty. **

Percolation originally refers to the passage of a fluid through a permeable medium, such as water through coffee or oil through rock. One motivation for the mathematical study of this phenomenon was the modeling of gas-mask filters. Percolation can also model the spread of information through a social network, for example: each individual is directly connected to several others, and we seek to understand whom they are connected to in the overall population through these local connections.

We begin by introducing percolation on a general graph. We then discuss the relatively simple case of the Bethe graph (a tree), followed by the more physical case of the Euclidean graph. We conclude with a few words about a complete-graph model. These three settings can model the transmission of information in a strongly hierarchical company (the first), in a social network (the third), or the flow of a fluid through permeable rock (the second).

16.1 Percolation on a graph

An (undirected) graph is a pair $G = (V, E)$, where V is a nonempty finite or countably infinite set and $E \subset \mathcal{P}_2(V)$, where $\mathcal{P}_2(V)$ is the set of two-element subsets of V. The elements of V and E are respectively called[1] the *vertices*

[1] The notation comes from the words "*vertices*" and "*edges*".

and *edges* of the graph G. If the edge $\{x, y\}$ belongs to E, we say that x and y are neighbors and write $x \sim y$. The most immediate example is the Euclidean graph $\mathcal{E}_d = (V, E)$, for which $V = \mathbb{Z}^d$ and

$$E = \{\{x, y\} : x, y \in \mathbb{Z}^d, |x - y|_1 := |x_1 - y_1| + \cdots + |x_d - y_d| = 1\}.$$

A *path* of length n in the graph (V, E) connecting a vertex x to a vertex y is a finite sequence $x = x_0 \sim \cdots \sim x_n = y$. If there is a path connecting x to y, we say that x and y *communicate* (or are connected) and write $x \leftrightarrow y$. We never have $x \sim x$, but adopt the convention $x \leftrightarrow x$ for every $x \in V$. The binary relation $\leftrightarrow$ on V is an equivalence relation, and the equivalence class (or connected component) of the point x is denoted by $C(x)$. A path is *self-avoiding* if all its vertices are distinct. If $x \leftrightarrow y$ and $x \neq y$, there exists a self-avoiding path from x to y, obtained from any path by erasing loops. A path connecting a vertex to itself is called a *cycle*. We say that (V, E) is a *tree* when it is connected and has no cycles, or equivalently when any two distinct vertices are connected by a unique self-avoiding path.

In what follows, V is countably infinite unless stated otherwise. Let $F \subset E$. We say that *percolation* occurs at $x \in V$ (for F) if the class $C(x)$ of x in the graph (V, F) has infinite size ($|C(x)| = \infty$), and write $x \leftrightarrow \infty$. We identify the set of subsets of E with the set

$$\Omega := \{0, 1\}^E.$$

We shall choose F randomly in Ω, so that (V, F) is a *random graph*. We construct F from E by tossing a coin to decide whether to retain or delete each edge of E. This amounts to introducing the probability distribution

$$\mathbb{P}_p = \bigotimes_{e \in E} \mathrm{Ber}(p)$$

on Ω equipped with its cylinder sigma-algebra $\mathcal{F}$, where $p \in [0, 1]$ is a fixed parameter. Write $\mathbb{E}_p$ for expectation under $\mathbb{P}_p$. To study how the parameter p affects percolation under $\mathbb{P}_p$, we distinguish a particular vertex $\varnothing \in V$, called the *root*—we say that (V, E) is rooted at $\varnothing$—and define the *percolation probability* under $\mathbb{P}_p$ for $\varnothing$ by

$$\theta(p) := \mathbb{P}_p(\varnothing \leftrightarrow \infty) = \mathbb{P}_p(|C(\varnothing)| = \infty).$$

Lemma 16.1 (Monotonicity)**.** *The function $p \in [0, 1] \mapsto \theta(p)$ is nondecreasing, and*

$$\theta(0) = 0 \quad et \quad \theta(1) = 1.$$

The monotonicity of θ leads to the notion of *critical probability*:

$$p_c = \sup\{p \in [0, 1] \ : \ \theta(p) = 0\} \in [0, 1].$$

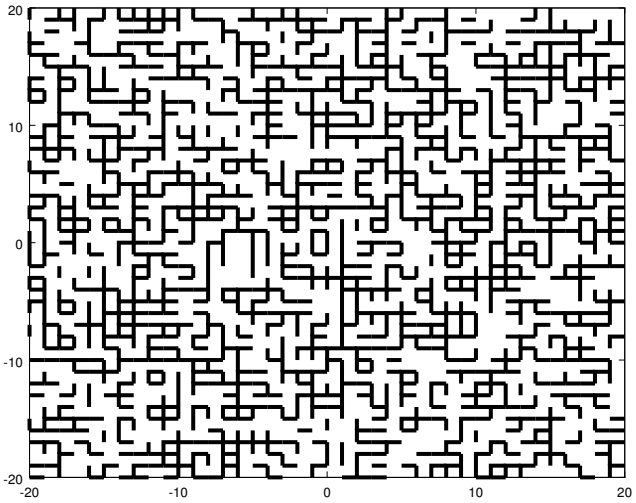


Fig. 16.1. A realization of the random graph (V, F) of the Euclidean graph $\mathcal{E}_2$ under $\mathbb{P}_{1/2}$.

Proof. If $p = 0$, then F is empty and hence $\theta(0) = 0$, whereas if $p = 1$, then $F = E$ and $\theta(1) = 1$. To establish monotonicity in p, we couple all the distributions $(\mathbb{P}_p)_{p\in[0,1]}$. Let $(U_e)_{e\in E}$ be i.i.d. real-valued random variables uniformly distributed on $[0, 1]$. For every $p \in [0, 1]$, realize $\mathbb{P}_p$ as the distribution of $F_p := (\mathbb{1}_{\{U_e\leqslant p\}})_{e\in E}$. If $p \leqslant p'$, then $F_p \subset F_{p'}$. Every edge of (V, F_p) is also an edge of $(V, F_{p'})$. If $x \leftrightarrow y$ for (V, F_p), this also holds for $(V, F_{p'})$. Consequently,

$$\theta(p) = \mathbb{P}_p(\varnothing \leftrightarrow \infty) \leqslant \mathbb{P}_{p'}(\varnothing \leftrightarrow \infty) = \theta(p').$$

□

Remark 16.2 (Coupling and monotonicity)**.** *Equip Ω with the partial order induced by inclusion in $\mathcal{P}_2(V)$: $F \leqslant F'$ if and only if every edge of (V, F) is also an edge of (V, F'). For every bounded $f : \Omega \to \mathbb{R}$, the coupling argument in the proof of Lemma 16.1 ensures that if f is nondecreasing, then so is $p \in [0, 1] \mapsto \mathbb{E}_p(f)$.*

16.2 Bethe graph (regular tree)

In this section, (V, E) is the *Bethe graph* $\mathcal{B}_r$ of degree $r \geqslant 2$, namely the infinite tree in which every vertex has r neighbors except one, which has only

$r-1$, denoted by 0 and called the *root* of the tree (see Figure 16.2). For every $p \in [0,1]$, under $\mathbb{P}_p$, the random graph (V,F) is a collection of trees, sometimes called a *forest*. The percolation probability and critical probability for the root vertex $\varnothing = 0$ in $\mathcal{B}_r$ under $\mathbb{P}_p$ are respectively denoted by

$$\theta_r(p) := \mathbb{P}_p(0 \leftrightarrow \infty) = \mathbb{P}_p(|C(0)| = \infty),$$
$$p_c(r) := \sup\{p \in [0,1] \,:\, \theta_r(p) = 0\}.$$

Theorem 16.3 (Threshold phenomenon for percolation on a Bethe graph)**.** *For every $r \geqslant 2$ and every $p \in [0,1]$, the quantity $\theta_r(p)$ is the largest root in $\theta \in [0,1]$ of the equation*

$$1 - \theta = (1 - p\theta)^{r-1},$$

while

$$p_c(r) = \frac{1}{r-1}.$$

Moreover, the function θ_r is continuous, vanishes on $[0, p_c(r)]$, is strictly increasing on $[p_c(r), 1]$, and for $r \geqslant 3$ we have

$$\lim_{p \to p_c(r)^+} \frac{\theta_r(p)}{p - p_c(r)} = \frac{2}{p_c(r)(1 - p_c(r))} = \frac{2(r-1)^2}{r-2}.$$

If $r \geqslant 3$, then $0 < p_c(r) < 1$, indicating a threshold phenomenon for percolation on the Bethe graph $\mathcal{B}_r$.

Proof. The Bethe graph $\mathcal{B}_r$ coincides with the Galton–Watson tree with offspring distribution δ_{r-1}. Under $\mathbb{P}_p$, we obtain a random tree in which each vertex is connected to each of its $r-1$ children with probability p. The number of these connections has offspring distribution $\mathrm{Bin}(r-1,p)$ with mean $m = (r-1)p$. According to Chapter 3, this tree is subcritical if $m < 1$, critical if $m = 1$, and supercritical if $m > 1$. By Theorem 3.7, the population extinc-

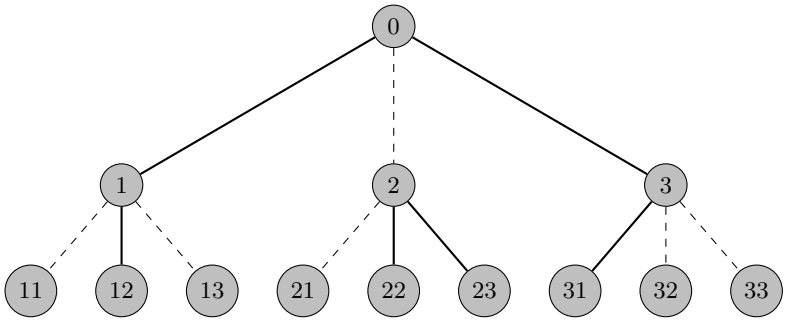


Fig. 16.2. First branches of $\mathcal{B}_4$ with edges deleted by $\mathbb{P}_p$ (dashed).

tion probability is the smallest root s_p of the equation $g_p(s) = s$, where g_p is the generating function of the offspring distribution: $g_p(s) = (ps + 1 - p)^{r-1}$ for $s \in [0, 1]$. Extinction of the population is equivalent to finiteness of the root's connected component. Thus $\theta_r(p) = 1 - s_p$, and $\theta_r(p)$ is the largest root in $[0, 1]$ of the equation $1 - \theta = (1 - p\theta)^{r-1}$.

Let h_p be the function defined on $[0, 1]$ by $h_p(\theta) = (1 - p\theta)^{r-1} - 1 + \theta$. It vanishes at 0, is strictly positive at 1, is strictly convex if $r \geqslant 3$, and affine if $r = 2$. Moreover, $h'_p(0)$ has the sign of $1 - p(r - 1)$. Thus, for $p(r - 1) \leqslant 1$, h_p vanishes only at 0 and $\theta_r(p) = 0$. Conversely, h_p vanishes exactly once on $]0, 1[$ whenever $p(r - 1) > 0$. In conclusion, the probability $\theta_r(p)$ is zero for $p \leqslant (r - 1)^{-1}$ and strictly positive for $p > (r - 1)^{-1}$. The function $p \mapsto h_p(\theta)$ is strictly decreasing and continuous. This ensures that the function θ_r is continuous on $[0, 1]$ and strictly increasing on $[p_c(r), 1]$.

Finally, $1-\theta = 1-(r-1)p\theta+\frac{1}{2}(r-1)(r-2)p^2\theta^2+o(\theta^2)$. Since $\theta_r(p_c(r)) = 0$ and $\theta_r(p)$ is strictly positive for $p > p_c = (r - 1)^{-1}$, simplification gives

$$\theta_r(p) = 2\frac{(r-1)p-1}{(r-1)(r-2)p^2} + o(\theta_r(p)) = \frac{2}{(r-2)p^2}(p - p_c) + o(\theta_r(p)).$$

We conclude by observing that $(r - 2)p_c = 1 - p_c$. □

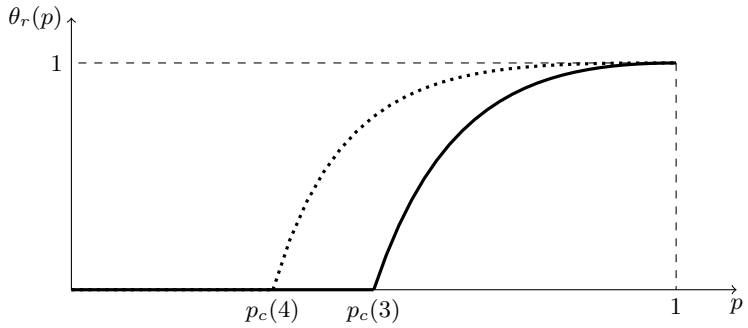


Fig. 16.3. Percolation for $\mathcal{B}_r$ with $r = 3$ (solid line) and $r = 4$ (dashed line).

Remark 16.4 (Explicit formulas in the cases $r \leqslant 4$)**.** *For $r = 2$, the Bethe graph $\mathcal{B}_2$ can be identified with the graph of $\mathbb{N}$, and under $\mathbb{P}_p$, the random variable $|C(0)|$ has a geometric distribution with parameter $1-p$, while $\theta_2 = \mathbb{1}_1$ and $p_c(2) = 1$. For $r = 3$ and $r = 4$, the equation satisfied by $\theta_r(p)$ in Theorem 16.3 can be solved explicitly (see Figure 16.3):*

$$\theta_3(p) = \frac{2p-1}{p^2}\mathbb{1}_{\{p>1/2\}} \quad et \quad \theta_4(p) = \frac{3p^2 - \sqrt{4p^3 - 3p^4}}{2p^3}\mathbb{1}_{\{p>1/3\}}.$$

Remark 16.5 (Mean size of the root component)**.** *Return to the proof of Theorem 16.3 and suppose that $p \leqslant p_c(r)$. Then $|C(0)| < \infty$ a.s. because the root's lineage dies out a.s. Moreover, by Section 3.3, the generating function h of $|C(0)|$ solves the equation*

$$h(s) = s(ph(s) + 1 - p)^{r-1}.$$

In particular,

$$\mathbb{E}_p(|C(0)|) = \begin{cases} \dfrac{1}{1-p(r-1)} & \text{si } p < p_c(r), \\ \infty & \text{si } p = p_c(r). \end{cases}$$

On the other hand, if $p \in]p_c, 1[$, there are infinitely many infinite classes, since individuals not connected to their parent are the roots of disjoint i.i.d. random trees. This situation, due to the tree structure, differs greatly from that of the Euclidean graph studied below.

16.3 Euclidean graph (lattice)

In this section, (V, E) is the Euclidean graph $\mathcal{E}_d$:

$$V = \mathbb{Z}^d \quad \text{et} \quad E = \left\{ \{x, y\} \in \mathcal{P}_2(\mathbb{Z}^d) \ : \ |x - y|_1 = \sum_{i=1}^{d} |x_i - y_i| = 1 \right\}.$$

The percolation probability and critical probability for the root vertex $\varnothing = 0$ in $\mathcal{E}_d$ under $\mathbb{P}_p$ are respectively denoted by

$$\begin{aligned} \theta_d(p) &:= \mathbb{P}_p(0 \leftrightarrow \infty) = \mathbb{P}_p(|C(0)| = \infty) \\ p_c(d) &:= \sup\{p \in [0,1] \ : \ \theta_d(p) = 0\}. \end{aligned}$$

For $d = 1$, the graph $\mathcal{E}_1$ can be identified with the graph $\mathbb{Z}$. For every $x \in \mathbb{Z}$, under $\mathbb{P}_p$, the random variable $|C(x)|$ has the distribution of $G_- + G_+ - 1$, where G_- and G_+ are independent random variables with geometric distribution of parameter $1 - p$. Thus, for $d = 1$, we obtain $\theta_1 = \mathbb{1}_1$ and $p_c(1) = 1$.

As soon as $d > 1$, the probability of the event $\{x \leftrightarrow y\}$ for $\mathcal{E}_d$ under $\mathbb{P}_p$ is not explicit even for $x \sim y$, because every edge of the graph may play a role, unlike the case of Bethe graphs. In terms of connectivity, the lattice is much more complex than a tree because cycles allow detours.

Lemma 16.6 (Small class)**.** *For all $d \geqslant 1$, $x \in \mathbb{Z}^d$, $p \in [0, 1]$,*

$$\mathbb{P}_p(C(x) = \{x\}) = (1-p)^{2d} \quad \textit{et} \quad \mathbb{P}_p(|C(x)| = 2) = p(1-p)^{2(2d-1)}.$$

Proof. We have $|C(x)| = 1$ if and only if x is connected to none of its $2d$ neighbors, whereas $|C(x)| = 2$ if and only if x is connected to just one of its $2d$ neighbors, which itself is connected only to x. □

Theorem 16.7 (Threshold phenomenon for percolation on a Euclidean graph)**.** *For every $2 \leqslant d \leqslant d'$, we have $\theta_d \leqslant \theta_{d'}$ and $p_c(d) \leqslant p_c(d')$. Moreover,*

$$0 < \frac{1}{2d-1} \leqslant \frac{1}{\kappa(d)} \leqslant p_c(d) \leqslant 1 - \frac{1}{\kappa(2)} \leqslant \frac{2}{3} < 1,$$

where the constant $\kappa(d)$ is defined in Lemma 16.8.

The constant $\kappa(d)$ is called the *connective constant* of $\mathcal{E}_d$. For clarity, the proof of Theorem 16.7 is split into four lemmas.

Lemma 16.8 (Connective constant)**.** *For all $d \geqslant 2$ and $n \geqslant 1$, let $\kappa_n(d)$ be the cardinality of the set $\mathcal{C}_n^d$ of self-avoiding paths in $\mathcal{E}_d$ of length n starting at the origin 0. We have*

$$d^n \leqslant \kappa_n(d) \leqslant 2d(2d-1)^{n-1}.$$

Moreover, the sequence $(\kappa_n(d)^{1/n})_{n\geqslant 1}$ converges, and its limit $\kappa(d)$ satisfies

$$d \leqslant \kappa(d) \leqslant 2d-1.$$

Proof of Lemma 16.8. The first edge of a self-avoiding path starting at 0 can be chosen among the $2d$ edges leaving 0. Subsequent steps cannot backtrack, hence $k_n(d) \leqslant 2d(2d-1)^{n-1}$. The lower bound $k_n(d) \geqslant d^n$ follows by observing that every path starting at 0 whose successive vertex coordinates are nondecreasing is self-avoiding. The sequence $(\kappa_n(d))_{n\geqslant 1}$ is submultiplicative: for all $n, m \geqslant 1$,

$$\kappa_{n+m}(d) \leqslant \kappa_n(d)\kappa_m(d),$$

because if $(0, x_1, \ldots, x_{n+m}) \in \mathcal{C}_{n+m}^d$, then

$$(0, x_1, \ldots, x_n) \in \mathcal{C}_n^d \quad \text{et} \quad (0, x_{n+1} - x_n, \ldots, x_{n+m} - x_n) \in \mathcal{C}_m^d.$$

The sequence $(\log(\kappa_n(d)))_{n\geqslant 1}$ is therefore subadditive, and Fekete's lemma[2] gives

$$\lim_{n\to\infty} \frac{\log(\kappa_n(d))}{n} = \lambda \quad \text{où} \quad \lambda := \inf_{n\geqslant 1} \frac{\log(\kappa_n(d))}{n} \in [-\infty, \infty[.$$

This gives $\kappa(d) = e^\lambda$, and the bounds on $\kappa(d)$ follow from those on $\kappa_n(d)$. □

Lemma 16.9 (Lower bound)**.** *For every $d \geqslant 2$, we have $p_c(d) \geqslant 1/\kappa(d)$.*

Proof of Lemma 16.9. Let $n \geqslant 1$ and let N_n be the random variable counting the elements of $\mathcal{C}_n^d$ contained in the random edge set $C(0)$ under $\mathbb{P}_p$. If $|C(0)| = \infty$, then $N_n \geqslant 1$. On the other hand, an element of $\mathcal{C}_n^d$ starts at the origin and has exactly n distinct edges. By independence, it has probability p^n of having all its edges in $C(0)$. Thus

[2] If $u_{n+m} \leqslant u_n + u_m$ for all $m, n \geqslant 1$, then $\frac{u_n}{n} \underset{n\to\infty}{\longrightarrow} \inf_{n\geqslant 1} \frac{u_n}{n} \in [-\infty, \infty[$.

$$\theta_d(p) \leqslant \mathbb{P}_p(N_n \geqslant 1) = \mathbb{E}_p(\mathbb{1}_{N_n \geqslant 1}) \leqslant \mathbb{E}_p(N_n) \leqslant p^n \kappa_n(d).$$

Lemma 16.8 gives $\kappa_n(d)^{1/n} = \kappa(d) + o_n(1)$, hence $\theta_d(p) \leqslant (p\kappa(d) + o_n(1))^n$. Consequently, if $p\kappa(d) < 1$, we obtain $\theta_d(p) = 0$ as $n \to \infty$. □

Lemma 16.10 (Monotonicity in dimension)**.** *If $2 \leqslant d \leqslant d'$, then $\theta_d \leqslant \theta_{d'}$ and $p_c(d') \leqslant p_c(d)$.*

Proof of Lemma 16.10. The idea is to use coupling, as in the proof of Lemma 16.1. Since the graph $\mathcal{E}_d$ can be identified with the subgraph of $\mathcal{E}_{d'}$ corresponding, for example, to the first d coordinates in $\mathbb{Z}^{d'}$, for every fixed $p \in [0,1]$ we can couple the random graphs $(\mathbb{Z}^d, F)$ and $(\mathbb{Z}^{d'}, F')$ of $\mathcal{E}_d$ and $\mathcal{E}_{d'}$ under $\mathbb{P}_p$ so that

$$\theta_d(p) = \mathbb{P}_p(0 \leftrightarrow +\infty \text{ dans } \mathcal{E}_d) \leqslant \mathbb{P}_p(0 \leftrightarrow +\infty \text{ dans } \mathcal{E}_{d'}) = \theta_{d'}(p).$$

In particular, $p_c(d) \leqslant p_c(d')$ follows. □

Lemma 16.11 (Upper bound)**.** *If $d = 2$, then $p_c(2) \leqslant 1 - 1/\kappa(2)$.*

Proof of Lemma 16.11. The dual graph $\mathcal{E}_2^* = (V^*, E^*)$ of the graph $\mathcal{E}_2 = (V, E)$ is the translate of $\mathcal{E}_2$ by the vector $(1/2, 1/2)$ (see Figure 16.3). Each edge e of the original graph intersects exactly one edge e_* of the dual graph. Construct the random graph (V^*, F^*) of $\mathcal{E}_2^*$ from the random graph (V, F) of $\mathcal{E}_2$ under $\mathbb{P}_p$ by declaring that $e_* \in F^*$ if and only if $e \notin F$.

A self-avoiding loop (or Jordan curve!) of length n is a path $(x_0, x_1, \ldots, x_n)$ such that the path $(x_0, x_1, \ldots, x_{n-1})$ is self-avoiding and $x_n = x_0$. Its complement consists of two connected components, one bounded (called the *interior*) and the other unbounded.

Let $\Lambda(m) := [-m, m]^2$. Introduce the following events:

$$F_m := \{\Lambda(m) \text{ est à l'intérieur d'une boucle auto-évitante de } \mathcal{E}_2^* \text{ à arêtes dans } F^*\},$$

and

$$G_m := \{\text{les arêtes reliant les sommets de } \Lambda(m) \text{ sont dans } F\}.$$

Percolation occurs at 0 on the event $F_m^c \cap G_m$, and since F_m and G_m involve disjoint edge sets, they are independent under $\mathbb{P}_p$. Now consider a loop γ whose interior contains $\Lambda(m)$. Its length n must exceed $4m$. Moreover, it must contain at least one vertex of the form $(1/2+k, 1/2)$ with $0 \leqslant k \leqslant n$. Let x_0 be the rightmost such point. Number the elements of γ by traversing the curve counterclockwise so that $\gamma = (x_0, x_1, \ldots, x_{n-1})$. This path is self-avoiding. There are at most $n\kappa_n(2)$ self-avoiding loops of length n surrounding $\Lambda(m)$. Thus, if M_n is the number of self-avoiding loops of length n surrounding $\Lambda(m)$ with all edges in F, we have

$$\mathbb{P}_p(F_m) \leqslant \mathbb{P}_p\left(\sum_{n\geqslant 4m} M_n \geqslant 1\right) \leqslant \sum_{n\geqslant 4m} \mathbb{E}_p(M_n) \leqslant \sum_{n\geqslant 4m} n\kappa_n(2)(1-p)^n.$$

Since $\kappa_n(2)^{1/n} = \kappa(2) + o_n(1)$, the series above (on the right-hand side) converges whenever $\kappa(2)(1-p) < 1$, and therefore, for sufficiently large m, $\mathbb{P}_p(F_m) < 1/2$. For such m, we have

$$\theta_2(p) \geqslant \mathbb{P}_p(F_m^c \cap G_m) = \mathbb{P}_p(F_m^c)\mathbb{P}_p(G_m) \geqslant \frac{1}{2}p^{4m^2} > 0.$$

Thus, if $(1-p)\kappa(2) < 1$, then $\theta_2(p) > 0$, hence $p_c(2) \leqslant 1 - 1/\kappa(2)$.

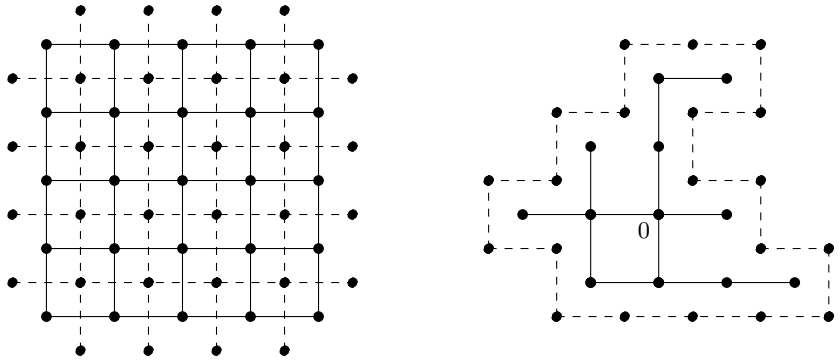


Fig. 16.4. Left: the Euclidean graph $\mathcal{E}_2$ (solid lines) and its dual $\mathcal{E}_2^*$ (dashed lines). Right: a finite component and its surrounding loop.

□

In the supercritical regime $p > p_c(d)$, the class $C(0)$ of the origin 0 may be infinite. The following theorem complements this result.

Theorem 16.12 (Existence of an infinite class)**.** *If $p > p_c(d)$, then $\mathbb{P}_p$-a.s. there exists at least one infinite class.*

Proof. The event $A :=$"There exists at least one infinite class" belongs to the tail sigma-algebra of the events $(A_e)_{e\in E}$, where $A_e := \{e \in F\}$. Under $\mathbb{P}_p$ these events are independent, so $\mathbb{P}_p(A) = 0$ or $\mathbb{P}_p(A) = 1$ by Kolmogorov's zero-one law. Now, if $p > p_c(d)$, then $\mathbb{P}_p(|C(0)| = \infty) > 0$ and $\{|C(0)| = \infty\} \subset A$, hence $\mathbb{P}_p(A) = 1$. □

However, the situation is very different from that of the Bethe graph: one can show that the infinite class is unique! The proof lies beyond the scope of this chapter. We give only a proof on $\mathcal{E}_2$, assuming the known fact that $\theta_2(1/2) = 0$.

Theorem 16.13 (Uniqueness of the infinite class)**.** *Suppose that $d = 2$ and that $\theta_2(1/2) = 0$ on the graph $\mathcal{E}_2$. Then, for every $p > p_c(2)$, there exists $\mathbb{P}_p$-a.s. a unique infinite class.*

Proof. Use the notation from the proof of Lemma 16.11.

For $p = 1/2$, percolation on the graph $\mathcal{E}_2$ and its dual graph $\mathcal{E}_2^*$ behaves in the same way: $\mathbb{P}_p$-a.s. there is no infinite component, since $\theta_2(1/2) = 0$. Let $\Lambda(m) = [-m, m]^2$. There exists a self-avoiding loop γ_* in $\mathcal{E}_2^*$ surrounding $\Lambda(m)$ whose edges belong to F^* (so the dual edges in $\mathcal{E}_2$ do not belong to F). By the same argument, there exists a loop γ in $\mathcal{E}_2$ containing γ_* and hence $\Lambda(m)$. By coupling, this remains true for every $p > 1/2$. If two infinite equivalence classes intersect $\Lambda(m)$, they also intersect γ and must therefore coincide! Since this holds for every m, there can be only one infinite class. □

16.4 Complete graph and Erdős–Rényi model

In this section, (V, E) is the *infinite complete graph* $\mathcal{K}_\infty$, meaning that V is countably infinite and $E = \mathcal{P}_2(V)$. Thus $x \sim y$ for all $x \neq y$ in V. All vertices play the same role. Since every pair of vertices are neighbors in $\mathcal{K}_\infty$, there is no geometry as in $\mathcal{B}_r$ or $\mathcal{E}_d$. Under $\mathbb{P}_p$, the random graph (V, F) is called the infinite Erdős–Rényi model. For all $p \in]0, 1]$ and $x \in V$, the vertex x has infinitely many neighbors in (V, F) and $\mathbb{P}_p(|C(x)| = \infty) = 1$, hence $\theta = \mathbb{1}_{]0,1]}$ and $p_c = 0$.

To make the model more interesting, we can consider percolation on the finite complete graph $\mathcal{K}_n$ with n vertices $V = \{1, \dots, n\}$ and, depending on n, choose the parameter p of $\mathbb{P}_p$. Denote by $\mathcal{G}(n, p)$ the distribution of the random graph (V, F) under $\mathbb{P}_p$. In this finite Erdős–Rényi model $\mathcal{G}(n, p)$, each vertex $x \in V$ has a random number of neighbors, with binomial distribution $\mathrm{Bin}(n-1, p)$ and mean $(n-1)p \sim np$. When $n \to \infty$ with $np \to \lambda > 0$, we have $p_n \sim \lambda/n \to 0$ and the number of neighbors of each vertex converges in distribution to the Poisson distribution $\mathrm{Poi}(\lambda)$ (the law of small numbers).

Theorem 16.14 (Threshold phenomenon and giant connected component)**.** *Let $\lambda > 0$ be a fixed real parameter and $\alpha := \lambda - 1 - \log(\lambda) > 0$. Let $(G_n)_{n \geqslant 1}$ be a sequence of Erdős–Rényi graphs defined on the same probability space, with $G_n = (V_n, F_n)$ having Erdős–Rényi distribution $\mathcal{G}(n, p)$ with $p = \lambda/n$. For every $v \in V_n$, let $G_n(v)$ be the connected component of vertex v.*

1. *If $\lambda < 1$, then for every $c > 1/\alpha$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{v \in V_n} |G_n(v)| \geqslant c \log(n) \right) = 0.$$

2. *If $\lambda > 1$, then*

$$\frac{\max_{v \in V_n} |G(v)|}{n} \xrightarrow[n \to \infty]{\text{p.s.}} 1 - \rho$$

where $\rho \in]0,1[$ is the extinction probability of a Galton–Watson branching process with offspring distribution $\mathrm{Poi}(\lambda)$*. Moreover, there exists a real number $c > 0$ such that almost surely the second largest connected component has size at most $c\log(n)$.*

In essence, if $\lambda < 1$, the random graph is so sparsely connected that it has no connected component larger than $\mathcal{O}(\log(n))$. The graph is fragmented. If $\lambda > 1$, the random graph is sufficiently connected to contain a unique giant connected component comprising a strictly positive fraction of the vertices, while the other connected components have size at most $\mathcal{O}(\log(n))$.

Proof. For simplicity, we prove only the first assertion. To do so, explore $G(v)$ using a *breadth-first search algorithm*[3], which explores the neighbors of v, forming the first generation, then their as yet unvisited neighbors, forming the second generation, and so on. This produces a finite spanning tree[4] of $G(v)$, similar to that in Figure 3.2. By forcing the absence of cycles, this approach also provides a coupling with a Galton–Watson random tree with offspring distribution $\mathrm{Bin}(n-1,p)$, which is subcritical because $(n-1)p \leqslant \lambda < 1$, whose total size N satisfies $N \geqslant |G(v)|$ (the tree is larger because there are no cycles). Now Theorem 3.17 states that

$$N \overset{\text{loi}}{=} T \quad \text{où} \quad T := \inf\{k \geqslant 1 : S_k = -1\}$$

is the hitting time of -1 for a random walk $(S_k)_{k\geqslant 0}$ on $\mathbb{Z}$ starting at $S_0 = 0$ with i.i.d. increments $(U_i)_{i\geqslant 1}$ such that $1 + U_i \sim \mathrm{Bin}(n-1,p)$. Fix any $\theta > 0$ and set

$$M_k = e^{\theta S_k}\varphi(\theta)^{-k} \quad \text{où} \quad \varphi(\theta) = \mathbb{E}(e^{\theta U_1})$$

which is the Laplace transform of the increment distribution of $(S_k)_{k\geqslant 0}$. The sequence $(M_k)_{k\geqslant 0}$ is a nonnegative martingale

for the natural filtration of $(S_k)_{k\geqslant 0}$, and T is an almost surely finite stopping time. We have $\lim_{n\to\infty} M_{T\wedge n} = M_T$ a.s. so, by Fatou's lemma and the optional stopping theorem,

$$e^{-\theta}\mathbb{E}(\varphi(\theta)^{-T}) = \mathbb{E}(M_T) = \mathbb{E}(\varliminf_{n\to\infty} M_{T\wedge n}) \leqslant \varliminf_{n\to\infty} \mathbb{E}(M_{T\wedge n}) = \mathbb{E}(M_0) = 1.$$

Let $\theta > 0$ be such that $\varphi(\theta) < 1$. For every $r > 0$, Markov's inequality gives

$$\mathbb{P}(T \geqslant r) = \mathbb{P}(\varphi(\theta)^{-T} \geqslant \varphi(\theta)^{-r}) \leqslant \varphi(\theta)^r\mathbb{E}(\varphi(\theta)^{-T}) \leqslant \varphi(\theta)^r e^{\theta}.$$

The convexity inequality $1 + x \leqslant e^x$, valid for every $x \in \mathbb{R}$, gives

$$\varphi(\theta) = e^{-\theta}(1 - p + pe^{\theta})^{n-1} = e^{-\theta}\left(1 + \frac{\lambda(e^{\theta}-1)}{n}\right)^{n-1} \leqslant e^{\lambda(e^{\theta}-1)-\theta}$$

[3] Abbreviated as *BFS*.
[4] One also speaks of a *Minimal Spanning Tree* (MST).

(note that $e^{\lambda(e^\theta-1)}$ is the Laplace transform of $\mathrm{Poi}(\lambda)$). The function $\theta \in \mathbb{R} \mapsto \lambda(e^\theta - 1) - \theta$ attains its minimum at $\theta_* = -\log(\lambda) > 0$, and $\varphi(\theta_*) = e^{-\alpha}$, where $\alpha := 1 - \lambda + \log(\lambda) > 0$, hence

$$\mathbb{P}(T \geqslant r) \leqslant \varphi(\theta_*)^r e^{\theta_*} = e^{-r\alpha+\theta_*} = \lambda^{-1} e^{-\alpha r}.$$

Consequently, for every $c > 0$,

$$\mathbb{P}(|G(v)| \geqslant c\log(n)) \leqslant \mathbb{P}(N \geqslant c\log(n)) = \mathbb{P}(T \geqslant c\log(n)) \leqslant \lambda^{-1} n^{-\alpha c}.$$

Now, since the random variables $(G(v))_{v\in V}$ are identically distributed, we have

$$\begin{aligned}\mathbb{P}\left(\max_{v\in V} |G(v)| \geqslant c\log(n)\right) &\leqslant \sum_{v\in V} \mathbb{P}(|G(v)| \geqslant c\log(n)) \\ &= n\mathbb{P}(|G(1)| \geqslant c\log(n)) \\ &\leqslant \lambda^{-1} n^{1-\alpha c},\end{aligned}$$

which gives the result upon taking $c > 1/\alpha$. □

16.5 Further reading

The geometry of regular trees (Bethe graphs) and complete graphs is simpler than that of the Euclidean graph (lattice), which in principle makes their probabilistic properties easier to study. In computing, telecommunications, and electrical engineering, a *network*[5] is a graph with a mark or weight on each edge, representing a conductance, resistance, length, cost, etc.; in our case, 0 (unavailable edge) or 1 (available edge). In this sense, our random graphs are random networks[6]. In the Euclidean case, the graph is also a lattice in the geometric sense[7], namely a discrete subgroup of Euclidean vector space.

Percolation is a classical phenomenon in statistical physics that can be studied on any random graph. The percolation mechanism also has many variants: oriented percolation (edges are oriented), first-passage and last-passage percolation, site percolation (vertices rather than edges are declared open or closed), etc. Much of this chapter is inspired by Thierry Lévy's lecture notes [Lé08]. The mathematical theory of percolation, developed since the mid-twentieth century by John Hammersley and Harry Kesten in particular, has remained a very active area of probability. Overviews can be found in the books [Gri99, Gri10] by Geoffrey Grimmett, whose 1974 thesis, supervised by Hammersley and Dominic Welsh (also a former student of Hammersley), already concerned random fields and graphs! Theorem 16.14 dates back to the mid-twentieth-century work of Paul Erdős and Alfréd Rényi [ER59, ER61b] and Edgar Gilbert [Gil59]. The proof of Theorem 16.14 is inspired by Charles Bordenave's lecture notes [Bor14a]. More generally, one can establish the following:

[5] The French term is "réseau".
[6] Networks chosen at random.
[7] Also called "réseau" in French.

Condition on (n,p)	Almost sure behavior of the random graph $G(n,p)$ as $n \to \infty$
$np < (1-\varepsilon)\log(n)$	isolated vertices exist (hence the graph is disconnected)
$np > (1+\varepsilon)\log(n)$	connected graph (hence $np \sim \log(n)$ is a connectivity threshold)
$np < 1$	no connected component of size $> \mathcal{O}(\log(n))$
$np = 1$	largest connected component of size $\approx n^{2/3}$
$np \to \lambda > 1$	unique giant connected component (fraction > 0 of the vertices)
	no other connected component is $> \mathcal{O}(\log(n))$.

For this topic, see [Bor14a] and the book by Remco van der Hofstad [vdH14], which provide a proof of the second assertion of Theorem 16.14.

Lemma 16.8 on the connective constant of the Euclidean graph is one of the keys to the proof of Theorem 16.7. However, the exact value of this constant $\kappa(d)$ is unknown for every $d \geqslant 2$. Hugo Duminil-Copin and Stanislav Smirnov showed in [DCS12] that it equals $\sqrt{2+\sqrt{2}}$ for the hexagonal graph on $\mathbb{Z}^2$. The proof of Theorem 16.13, taken from the survey article [HJ06], is due to Theodore Harris [Har60]. In that work, Harris also establishes that $\theta_2(1/2) = 0$, meaning that $p_c(2) \geqslant 1/2$. The equality $p_c(2) = 1/2$ was not proved until Kesten's work [Kes80]. For other values of d, the critical probability is unknown. Nevertheless, one can show that θ_d is of class $\mathcal{C}^\infty$ on the interval $]p_c(d), 1[$.

17

Growth and fragmentation

Keywords. Growth and fragmentation model; piecewise deterministic Markov process; infinitesimal generator; Yule process.

Tools. Poisson process; coupling; Wasserstein distance; autoregressive process; exponential distribution; geometric distribution; generating function; Laplace transform.

Difficulty. **

17.1 The TCP window-size process in computing

The maximum instantaneous outgoing throughput of a computer connected to a TCP/IP[1] network such as the Internet is regulated by the following algorithm: the maximum throughput increases deterministically by one unit at each time step, and upon receiving a congestion signal, it is multiplied by a factor between $[0, 1[$, typically $1/2$. This simple mechanism allows both efficient use of the network and an effective response to congestion.

We shall study an idealized Markovian model of this mechanism. If we assume that, over a given duration, each other computer on the network causes congestion with a small probability, independently of the others, then the law of small numbers suggests that congestion signals follow a Poisson process. Let $N = (N_t)_{t \geqslant 0}$ be a Poisson process starting from 0 with intensity λ, counting congestion signals, with successive jump times $(T_n)_{n \geqslant 0}$, where $T_0 := 0$. If $E_n := T_n - T_{n-1}$ for every $n \geqslant 1$, then

$$T_n = E_1 + \cdots + E_n,$$

[1] Standing for "*Transmission Control Protocol*" and "*Internet Protocol*".

and the sequence $E = (E_n)_{n\geqslant 1}$ consists of i.i.d. random variables with exponential distribution of mean $1/\lambda$. For every $t \in \mathbb{R}_+$, we have

$$N_t = \sum_{n\geqslant 1} \mathbb{1}_{[0,t]}(T_n).$$

Let $Q = (Q_n)_{n\geqslant 1}$ be a sequence of independent random variables taking values in $[0,1[$ with common distribution $\mathcal{Q}$. Let X_0 be a nonnegative random variable. We assume that X_0, E, and Q are independent. Define the continuous-time process $X = (X_t)_{t\geqslant 0}$ with state space $\mathbb{R}_+$ as follows:

$$X_t = \begin{cases} X_{T_n} + t - T_n & \text{si } T_n \leqslant t < T_{n+1}, \\ Q_{n+1}(X_{T_n} + T_{n+1} - T_n) & \text{si } t = T_{n+1}, \end{cases}$$

where $n = N_t \in \mathbb{N}$ is the number of jumps before time t. The sample paths of X are piecewise affine with slope 1, right-continuous with left limits, and each jump corresponds to multiplication by a number in $[0,1[$. This is called a piecewise deterministic Markov process[2] and, more specifically, an additive increase, multiplicative decrease process[3]. The process X is known as the TCP window-size process[4], since in computing, the maximum outgoing throughput is controlled by the size of a "TCP window".

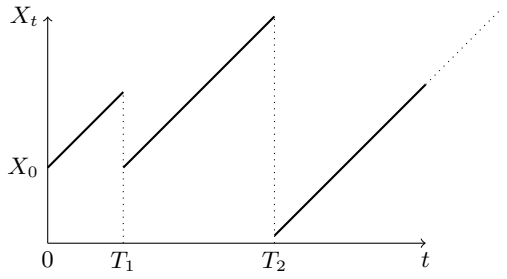


Fig. 17.1. Shape of a typical sample path of the process X.

Theorem 17.1 (Infinitesimal generator)**.** *The infinitesimal generator of the process X is given, for every function $f \in \mathcal{C}^1(\mathbb{R}_+, \mathbb{R})$ and every $x \geqslant 0$, by*

$$\begin{aligned} L(f)(x) &:= \lim_{t\to 0^+} \frac{\mathbb{E}(f(X_t) \,|\, X_0 = x) - f(x)}{t} \\ &= f'(x) + \lambda \int_0^1 (f(qx) - f(x))\, \mathcal{Q}(dq). \end{aligned}$$

[2] Abbreviated as *PDMP*.
[3] Abbreviated as *AIMD*.
[4] The window controls the permitted volume of outgoing data.

The process X does not explode because N does not explode. The conditional expectation $\mathbb{E}(f(X_t) \,|\, X_0 = x)$ is well defined because on $\{X_0 = x\}$ we have $X_t \leqslant x + t$, and since f is $\mathcal{C}^1$, the real-valued random variable $f(X_t)$ is bounded.

Proof. Let $f \in \mathcal{C}^1(\mathbb{R}_+, \mathbb{R})$ and $t \leqslant 1$. By the definitions of X and N,

$$\begin{aligned}\mathbb{E}\big(f(X_t)\mathbb{1}_{\{N_t=0\}} \,|\, X_0 = x\big) &= f(x+t)\mathbb{P}(N_t = 0) = f(x+t)e^{-\lambda t} \\ &= (f(x) + f'(x)t + o(t))(1 - \lambda t + o(t)) \\ &= f(x) + t(f'(x) - \lambda f(x)) + o(t).\end{aligned}$$

Conditional on $\{N_t = 1\}$, T_1 is uniformly distributed on $[0, t]$, so

$$\begin{aligned}\mathbb{E}(f(X_t) \,|\, X_0 = x, N_t = 1) &= \mathbb{E}(f(Q_1(x + T_1) + t - T_1) \,|\, N_t = 1) \\ &= \int_0^t \frac{ds}{t} \int_0^1 \mathcal{Q}(dq) f(q(x+s) + t - s) \\ &= \int_0^1 \mathcal{Q}(dq) f(qx) + o(1)\end{aligned}$$

where the $o(1)$ (as $t \to 0$) follows from the continuity of f, and hence

$$\begin{aligned}\mathbb{E}\big(f(X_t)\mathbb{1}_{\{N_t=1\}} | X_0 = x\big) &= \mathbb{P}(N_t = 1)\bigg(\int_0^1 \mathcal{Q}(dq) f(xq) + o(1)\bigg) \\ &= \lambda t \int_0^1 \mathcal{Q}(dq) f(xq) + o(t).\end{aligned}$$

Thirdly, on $\{X_0 = x\}$ we have $X_t \leqslant x + t$, and therefore, since $t \leqslant 1$, we have

$$\big|\mathbb{E}\big(f(X_t)\mathbb{1}_{\{N_t \geqslant 2\}} \,|\, X_0 = x\big)\big| \leqslant \left(\sup_{u \in [0, x+1]} |f(u)|\right) \mathbb{P}(N_t \geqslant 2) = \mathcal{O}(t^2).$$

Finally, collecting the three terms gives the desired result. □

For every $t \geqslant 0$, consider the linear operator P_t mapping each bounded measurable function $f : \mathbb{R}_+ \to \mathbb{R}$ to the bounded measurable function

$$x \in \mathbb{R}_+ \mapsto P_t f(x) = \mathbb{E}(f(X_t) \,|\, X_0 = x).$$

This quantity is also meaningful when f is measurable and locally bounded (for example, continuous), because conditional on $X_0 = x$, X_t takes its values in the bounded interval $[0, x + t]$. We have $P_0 = I$, and the Markov property implies that $(P_t)_{t \geqslant 0}$ is a semigroup:

$$P_{t+s} = P_{s+t} = P_t \circ P_s.$$

This semigroup is called Markovian because it preserves positivity and the unit function: $P_t(f) \geqslant 0$ if $f \geqslant 0$, and $P_t(1) = 1$. A bounded measurable

function $f : \mathbb{R}_+ \to \mathbb{R}$ is said to belong to the domain of the generator L of the process when the following limit exists for every $x \geqslant 0$:

$$L(f)(x) = \lim_{t \to 0^+} \frac{P_t(f)(x) - f(x)}{t} = \partial_{t=0} P_t(f)(x).$$

Everything behaves as if "$P_t = e^{tL}$". One can establish that the semigroup property gives, for every $t \geqslant 0$ and every function f of class $\mathcal{C}^1$,

$$\partial_t P_t(f) = \lim_{\varepsilon \to 0} P_t\left(\frac{P_\varepsilon f - f}{\varepsilon}\right) = P_t(Lf)$$
$$\partial_t P_t(f) = \lim_{\varepsilon \to 0} \frac{P_\varepsilon(P_t f) - P_t f}{\varepsilon} = L(P_t f).$$

These are called the *forward* and *backward* Chapman–Kolmogorov equations, respectively.

Theorem 17.1 states that the domain of the generator L contains functions of class $\mathcal{C}^1$. Moreover, the generator preserves polynomials without increasing their degree. This allows moments to be computed recursively.

Theorem 17.2 (Moments and invariant distribution)**.** *For every $n \in \mathbb{N}$, every $x \geqslant 0$, and every $t \geqslant 0$, the moment of order n of the distribution of X_t conditional on $\{X_0 = x\}$ satisfies*

$$\mathbb{E}(X_t^n \mid X_0 = x) = \frac{n!}{\prod_{k=1}^n \theta_k} + n! \sum_{m=1}^n \left(\sum_{k=0}^m \frac{x^k}{k!} \prod_{\substack{j=k \\ j \neq m}}^n \frac{1}{\theta_j - \theta_m} \right) e^{-\theta_m t}$$

where $\theta_p := \lambda(1 - \mathbb{E}(Q_1^p))$. Moreover, conditional on $\{X_0 = x\}$, the variable X_t converges in distribution as $t \to \infty$ to the distribution μ on $\mathbb{R}_+$ whose moments are

$$\int x^n \, \mu(dx) = \frac{n!}{\prod_{k=1}^n \theta_k}$$

for every $n \geqslant 0$. Finally, all moments of X_t converge to those of μ.

Proof. Set $\alpha_n(t) = \mathbb{E}(X_t^n \mid X_0 = x) = P_t(f_n)(x)$, where $f_n(x) = x^n$. For every $n \geqslant 1$, Theorem 17.1 gives

$$Lf_n(x) = nx^{n-1} + \lambda x^n \int_0^1 (q^n - 1) \, \mathcal{Q}(dq) = nf_{n-1}(x) - \theta_n f_n(x)$$

and therefore, using the forward Chapman–Kolmogorov equation,

$$\begin{aligned} \alpha_n'(t) &= \partial_t P_t(f_n)(x) = P_t(Lf_n)(x) = P_t(nf_{n-1} - \theta_n f_n)(x) \\ &= n\alpha_{n-1}(t) - \theta_n \alpha_n(t). \end{aligned}$$

Since α_0 is constant and equal to 1, we obtain the ordinary differential equation $\alpha_1'(t) = 1 - \theta_1 \alpha_1(t)$, which gives α_1. Induction on n (a triangular system!) yields the stated formula for $(\alpha_n)_{n \geqslant 0}$.

Since $\theta_p \leqslant \theta_1$ for $p \geqslant 1$, the moments of μ are bounded above by those of an exponential distribution with parameter θ_1. This implies that its Laplace transform is finite on $]-\infty, \theta_1[$. It is therefore holomorphic on a neighborhood (in $\mathcal{C}$) of the origin. In particular, the characteristic function of μ is analytic on a neighborhood (in $\mathbb{R}$) of the origin. Theorem 21.7 then ensures that the distribution μ is characterized by its moments.

Thus, conditional on $\{X_0 = x\}$, each moment of X_t converges as $t \to \infty$ to a finite limit. The fact that this sequence of limits is the moment sequence of a distribution, and that X_t converges in distribution to it, follows from an argument similar to the proof of Wigner's theorem by the method of moments in Chapter 21. Care is needed because here we work on $\mathbb{R}_+$ and not on the whole of $\mathbb{R}$. □

Remark 17.3 (Spectral decomposition)**.** *For all $n \geqslant 0$ and $t \geqslant 0$, the operators L and P_t, restricted to the space $\mathbb{R}_n[X]$ of polynomials of degree at most n, are endomorphisms whose eigenvalues are respectively $(-\theta_k)_{0 \leqslant k \leqslant n}$ and $(e^{-t\theta_k})_{0 \leqslant k \leqslant n}$.*

Remark 17.4 (Growth-collapse[5])**.** *If $\mathcal{Q} = \delta_0$, then the moment of order n of the invariant distribution μ of X equals $n!/\lambda^n$, which ensures that μ is the exponential distribution with parameter λ.*

Remark 17.5 (Distribution of the process at a fixed time)**.** *Consider the case $\mathcal{Q} = \delta_q$ with $0 < q < 1$, conditional on $\{X_0 = x\}$ with $x \geqslant 0$. Then for every $t > 0$, there exists an absolutely continuous distribution μ_t supported on the interval $[0, qx + t]$ such that*

$$\mathrm{Loi}(X_t) = e^{-\lambda t}\delta_{x+t} + (1 - e^{-\lambda t})\mu_t$$

and $\mathrm{dist}(x + t, \mathrm{supp}(\mu_t)) = (1 - q)x$*. Indeed, conditional on $\{N_t = n\}$,*

$$X_t = \begin{cases} x + t & \text{si } n = 0, \\ q^n x + \sum_{k=1}^{n+1} q^{n-k+1}(U_k - U_{k-1}) & \text{si } n > 0, \end{cases}$$

where $0 = U_0 < U_1 < \cdots < U_n < U_{n+1} = t$ is a uniform order statistic on $[0, t]$. Thus $\mathrm{Loi}(X_t \,|\, N_t = 0) = \delta_{x+t}$, whereas $\mathrm{Loi}(X_t \,|\, N_t = n)$ is absolutely continuous if $n > 0$, supported on $[q^n(x + t), q^n x + t]$. This gives the result with $\mu_t = \mathrm{Loi}(X_t \,|\, N_t > 0)$. It shows that X is asymptotically smoothing, since the mass of the atomic part of μ_t tends to 0 as $t \to \infty$. This is caused by the jump mechanism, the sole source of randomness.

Remark 17.6 (Irreducibility)**.** *For every $x \geqslant 0$, conditional on the event $\{X_0 = x\}$, the process X enters any interval in finite time with positive probability. In other words, its sample path intersects every horizontal strip in finite time with positive probability. This is immediate if the strip lies above x. If the strip lies below x, it suffices to consider a sawtooth sample path.*

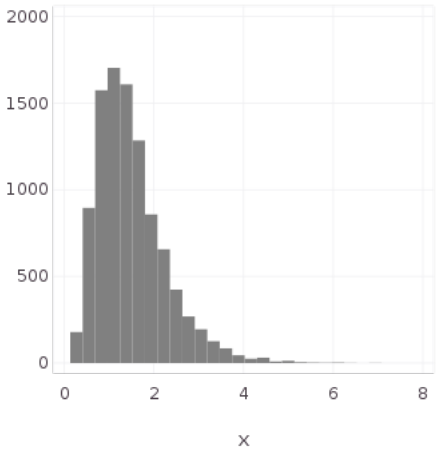


Fig. 17.2. Histogram of a sample from X_t for $t = 100$, $\lambda = 1$, $\mathcal{Q} = \delta_{1/2}$.

Let $p \geqslant 1$ be a fixed real number. If (E, d) is a metric space and μ a distribution on E, then the quantity $\int d(x,y)^p \, \mu(dy)$ is either infinite for every $x \in E$ or finite for every $x \in E$; in the latter case, we say that μ has a finite moment of order p. The Wasserstein distance W_p on the set of distributions on E having a finite moment of order p is defined by

$$W_p(\mu_1, \mu_2) = \inf_{\substack{(X,Y) \\ X_1 \sim \mu_1 \\ X_2 \sim \mu_2}} \mathbb{E}(d(X,Y)^p)^{1/p} = \inf_{\pi \in \Pi(\mu_1, \mu_2)} \left[\int_{E \times E} d(x,y)^p \, \pi(dx, dy) \right]^{1/p}$$

where $\Pi(\mu_1, \mu_2)$ is the convex set of probability distributions on $E \times E$ with marginal distributions μ_1 and μ_2. It contains the product distribution $\mu_1 \otimes \mu_2$. An element π of this set is a *coupling* of μ_1 and μ_2. Here we take $E = \mathbb{R}$ and $d(x,y) = |x - y|$. One can show that a sequence $(\mu_n)_n$ converges to μ for W_p if and only if it converges weakly for $\mathcal{C}_b$ and all moments of order at most p of μ_n converge to those of μ.

Theorem 17.7 (Long-time behavior by coupling)**.** *Let νP_t be the distribution at time t of the TCP process with initial distribution ν and parameters λ and $\mathcal{Q}$. Then, for any probability measures ν_1 and ν_2 on $\mathbb{R}_+$ with a finite moment of order $p \geqslant 1$, for every $t \geqslant 0$, setting $\theta_p := \lambda(1 - \mathbb{E}(Q_1^p))$, we have*

$$W_p(\nu_1 P_t, \nu_2 P_t) \leqslant W_p(\nu_1, \nu_2) \, \exp\left(-\frac{\theta_p t}{p} \right).$$

[5] Growth followed by complete collapse.

Proof. Consider a pair (X, Y) where X and Y start from x and y but use the same N and the same Q (the same jump times and multiplicative coefficients). Remarkably, the quantity $|X_t - Y_t|$ remains constant between successive jumps, and at the k^{e} jump it is multiplied by Q_k. Consequently, for every $t \geqslant 0$, we have

$$\begin{aligned}
\mathbb{E}(|X_t - Y_t|^p) &= \sum_{k=0}^{\infty} \mathbb{E}\big(|X_t - Y_t|^p \mathbb{1}_{\{N_t = k\}}\big) \\
&= \sum_{k=0}^{\infty} \mathbb{E}\big(|x - y|^p Q_1^p \dots Q_k^p \mathbb{1}_{\{N_t = k\}}\big) \\
&= |x - y|^p \sum_{k=0}^{\infty} \mathbb{E}(Q_1^p)^k \mathbb{P}(N_t = k) \\
&= |x - y|^p e^{-\lambda t(1 - \mathbb{E}(Q_1^p))}.
\end{aligned}$$

Thus, if X and Y are two TCP processes with parameters λ and $\mathcal{Q}$ and respective initial distributions ν_1 and ν_2, then for every coupling π of ν_1 and ν_2, every $t \geqslant 0$, and every $p \geqslant 1$,

$$W_p(\nu_1 P_t, \nu_2 P_t)^p \leqslant e^{-\theta_p t} \int_{\mathbb{R}_+ \times \mathbb{R}_+} |x - y|^p \, \pi(dx, dy).$$

Taking the infimum over couplings π of ν_1 and ν_2 completes the proof. □

The embedded chain $\widehat{X} := (\widehat{X}_n)_{n \geqslant 0} = (X_{T_n})_{n \geqslant 0}$ of X is the process sampled at its jump times. It is a homogeneous discrete-time Markov chain with continuous state space $\mathbb{R}_+$. It satisfies $\widehat{X}_0 = X_0$ and is a linear autoregressive process:

$$\widehat{X}_{n+1} = Q_{n+1}(\widehat{X}_n + E_{n+1}).$$

The transition kernel $\widehat{K}$ of $\widehat{X}$ is, for every $f \in \mathcal{C}_b(\mathbb{R}_+, \mathbb{R})$ and $x \geqslant 0$,

$$\widehat{K}(x, f) = \mathbb{E}(f(\widehat{X}_{n+1}) \,|\, \widehat{X}_n = x) = \lambda \int_{[0,1] \times \mathbb{R}_+} e^{-\lambda y} f(q(x + y)) \, \mathcal{Q}(dq) \, dy.$$

Theorem 17.8 (Convergence in distribution of the embedded chain)**.** *For any X_0, the embedded chain $(\widehat{X}_n)_{n \geqslant 0}$ converges in distribution to $\widehat{\mu} = \mathrm{Loi}(\sum_{n=1}^{\infty} Q_1 \cdots Q_n E_n)$.*

Proof. The autoregressive formula and the independence and stationarity assumptions give, by induction on n,

$$\begin{aligned}
\widehat{X}_n &= Q_n \cdots Q_1 X_0 + \sum_{k=1}^{n} Q_n \cdots Q_{n-k+1} E_{n-k+1} \\
&\overset{\mathrm{loi}}{=} Q_1 \cdots Q_n X_0 + \sum_{k=1}^{n} Q_1 \cdots Q_k E_k.
\end{aligned}$$

We have $\mathbb{E}(Q_1 \cdots Q_n) = \mathbb{E}(Q)^n$. Now $\mathbb{E}(Q) < 1$ since $\mathbb{P}(0 \leqslant Q < 1) = 1$, and hence by monotone convergence $\mathbb{E}(\sum_n Q_1 \cdots Q_n) = \sum_n \mathbb{E}(Q)^n < \infty$, so $\mathbb{P}(\sum_n Q_n < \infty) = 1$ and therefore $\mathbb{P}(Q_1 \cdots Q_n \to 0) = 1$. Consequently, $Q_1 \cdots Q_n X_0 \to 0$ a.s. On the other hand, by monotone convergence,

$$\sum_{k=1}^{n} Q_1 \ldots Q_k E_k \xrightarrow[n\to\infty]{\text{p.s.}} \sum_{n=1}^{\infty} Q_1 \cdots Q_n E_n \sim \widehat{\mu}.$$

Consequently, the embedded chain converges in distribution to $\widehat{\mu}$ for any X_0. □

The distribution $\widehat{\mu}$ has mean $\lambda^{-1}\mathbb{E}(Q_1)/\mathbb{E}(1 - Q_1)$. For the embedded chain, the loss of dependence on the initial condition is clear from the following formula:

$$\mathbb{E}(\widehat{X}_n) = \mathbb{E}(Q)^n \mathbb{E}(X_0) + \mathbb{E}(Q_1) \frac{1 - \mathbb{E}(Q_1)^n}{\lambda(1 - \mathbb{E}(Q_1))}.$$

The distribution $\widehat{\mu}$ satisfies the distributional equation

$$F \overset{\text{loi}}{=} Q_1(F + E_1)$$

where $F \sim \widehat{\mu}$ is independent of Q_1, E_1.

17.2 Position-dependent jump intensity

The jump intensity can depend on the path followed by the process. Consider a locally integrable function $\lambda : x \in \mathbb{R}_+ \mapsto \lambda(x) \in \mathbb{R}_+$ and a sequence $(E_n)_{n\geqslant 1}$ of i.i.d. real-valued random variables with exponential distribution of parameter 1. Construct the sample path of the process $(X_t)_{t\geqslant 0}$ starting from $X_0 = x_0$ by setting $X_s = x_0 + s$ for every $0 \leqslant s < T_1$, where the first jump time T_1 is defined by the equation (the time at which $\int$ reaches E_1)

$$\int_0^{T_1} \lambda(X_s)\, ds = \int_0^{T_1} \lambda(x_0 + s)\, ds = E_1.$$

Then set $X_{T_1} = Q_1 X_{T_1^-} = Q_1(x_0 + T_1)$ and, more generally, set

$$X_t = \begin{cases} X_{T_n} + t - T_n & \text{si } T_n \leqslant t < T_{n+1}, \\ Q_{n+1}(X_{T_n} + T_{n+1} - T_n) & \text{si } t = T_{n+1}, \end{cases}$$

where the increasing sequence of jump times $(T_n)_{n\geqslant 0}$ is defined by $T_0 = 0$ and

$$\int_{T_n}^{T_{n+1}} \lambda(X_s)\, ds = \int_0^{T_{n+1}-T_n} \lambda(X_{T_n} + s)\, ds = E_{n+1}.$$

For example, the special case $\lambda(x) = x$ gives

$$T_{n+1} - T_n = \sqrt{2E_{n+1} + (X_{T_n})^2} - X_{T_n}.$$

This construction yields the sample path $t \mapsto X_t$ on the interval $[0, T_\infty[$, where $T_\infty := \lim_{n\to\infty} T_n$ is the explosion time of the process. If λ is bounded, comparison with the constant case gives $\mathbb{P}(T_\infty = \infty) = 1$, and the process is said to be nonexplosive. When λ is unbounded, for example $\lambda(x) = x$, properties of $\mathcal{Q}$ may allow nonexplosion to be established. One can show that the generator of the process $(X_t)_{t\geqslant 0}$ is

$$L(f)(x) = f'(x) + \lambda(x) \int_0^1 (f(qx) - f(x))\, \mathcal{Q}(dq).$$

More generally, the same procedure constructs sample paths of a process with infinitesimal generator

$$L(f)(x) = G(f)(x) + \lambda(x) \int (f(y) - f(x))\, K(x, dy)$$

where K is a Markov kernel, λ is a nonnegative locally integrable function, and G is the generator of an arbitrary process that "acts" between jump times.

17.3 Branching and growth-fragmentation

The TCP process can model much more than the TCP window size in the TCP/IP protocol. For example, one can think of the size or mass of a cell that grows linearly and then divides. Another example is a polymer[6] whose length increases linearly over time before it breaks apart. In terms of mass, division or dislocation is a mass-preserving fragmentation. For simplicity, suppose that there are always two resulting fragments, evolving independently according to the same process. Consider the population size N_t at time $t \geqslant 0$. This amounts to studying a binary tree whose branch segment lengths are i.i.d. with exponential distribution of parameter λ. At time $t \geqslant 0$, the tree has N_t branches, and $N_0 = 1$. The process $(N_t)_{t\geqslant 0}$ is called the Yule process with parameter λ. It is a continuous-time Galton–Watson process with offspring distribution δ_2. It is also a continuous-time pure birth process.

Theorem 17.9 (Population size)**.** *For every $t \geqslant 0$, the population size N_t has a geometric distribution on $\mathbb{N}^*$ with parameter $e^{-\lambda t}$. In particular,*

$$\mathbb{E}(N_t) = e^{\lambda t} \quad et \quad \mathrm{Var}(N_t) = (1 - e^{-\lambda t})e^{2\lambda t}.$$

Proof. The memoryless property of exponential distributions shows that $(N_t)_{t\geqslant 0}$ is the counting process starting from $N_0 = 1$ associated with the sequence of real-valued random variables $(S_n)_{n\geqslant 1}$, where $S_n = F_1 + \cdots + F_n$,

[6] A molecule consisting of a chain of identical molecules (monomers).

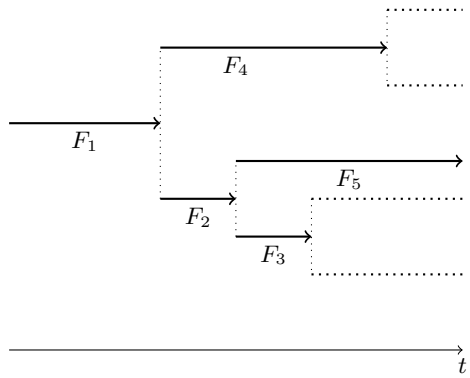


Fig. 17.3. Beginning of the binary tree defining the process N.

with $(F_n)_{n\geqslant 1}$ a sequence of independent real-valued random variables and F_n exponentially distributed with parameter $n\lambda$ for every $n \geqslant 1$. We have $N_t - 1 = \sum_{n=1}^{\infty} \mathbb{1}_{[0,t]}(S_n)$. Now, by Rényi's lemma 11.3, the real-valued random variable S_n has the distribution of $\max(E_1, \dots, E_n)$, where $E_1, \dots, E_n$ are i.i.d. with exponential distribution of parameter λ. Thus, for every integer $n \geqslant 0$,

$$\mathbb{P}(N_t - 1 \geqslant n) = \mathbb{P}(S_n \leqslant t) = \mathbb{P}(E_1 \leqslant t) \cdots \mathbb{P}(E_n \leqslant t) = (1 - e^{-\lambda t})^n.$$

□

Consider the case where each fragmentation produces $k + 1$ fragments, with $k \geqslant 1$ fixed. Denote the associated counting process by $(N_t^{(k)})_{t\geqslant 0}$. For every $n \geqslant 2$, the process starting from n is the sum of n i.i.d. copies of the process starting from 1. This is the superposition or branching property. We call $(N_t^{(k)})_{t\geqslant 0}$ a Yule process with offspring distribution δ_{k+1}. It is a homogeneous continuous-time Markov process with state space $\mathbb{N}$ and infinitesimal generator $L^{(k)}$ given, for all $f : \mathbb{N} \to \mathbb{R}$ and $n \in \mathbb{N}$, by

$$L^{(k)}(f)(n) = \lim_{t\to 0^+} \frac{\mathbb{E}(f(N_t^{(k)}) \,|\, N_0^{(k)} = n) - f(n)}{t} = n\lambda(f(n + k) - f(n)).$$

Compare this with the generator $\lambda(f(n + 1) - f(n))$ of the Poisson process. If $f(n) = n$ and $\alpha_n(t) = \mathbb{E}(N_t^{(k)} \,|\, N_0^{(k)} = n) = P_t^{(k)}(f)(n)$, then $(L^{(k)}f)(n) = k\lambda n = k\lambda f(n)$, and the forward Chapman–Kolmogorov equation gives the ordinary differential equation

$$\partial_t \alpha_n(t) = \partial_t P_t^{(k)}(f) = P_t(L^{(k)}f) = k\lambda P_t^{(k)}(f) = k\lambda \alpha_n(t).$$

The initial condition $\alpha_n(0) = n$ then gives $\mathbb{E}(N_t^{(k)} \,|\, N_0^{(k)} = n) = ne^{k\lambda t}$. Applying this method systematically leads to the following result.

Theorem 17.10 (Yule process)**.** *Suppose that* $N_0^{(k)} \sim \delta_1$*. For every real number* $t \geqslant 0$ *and every integer* $k \geqslant 1$*, if* $Z_1, \dots, Z_k$ *are i.i.d. copies of* $N_t^{(k)}$*, then*

$$\frac{Z_1 + \cdots + Z_k}{k} \sim \mathrm{Geo}(e^{-k\lambda t}).$$

In particular, the mean and variance of $N_t^{(k)}$ *are given by*

$$\mathbb{E}(N_t^{(k)}) = e^{k\lambda t} \quad et \quad \mathrm{Var}(N_t^{(k)}) = k(1 - e^{-k\lambda t})e^{2k\lambda t}.$$

Moreover, there exists a random variable $M_\infty \in L^2$ *with distribution* $\mathrm{Gamma}(1/k, 1/k)$ *and density*

$$x \mapsto \frac{(1/k)^{1/k}}{\Gamma(1/k)} x^{1/k-1} e^{-x/k} \mathbb{1}_{\mathbb{R}_+}(x)$$

such that

$$e^{-k\lambda t} N_t^{(k)} \underset{t\to\infty}{\longrightarrow} M_\infty \quad p.s. \text{ et dans } L^2.$$

Proof. A time change reduces the problem to $\lambda = 1$. Now fix $s \in [0,1]$ and consider the function $f(n) = s^n$. We have $P_t^{(k)}(f)(n) = \mathbb{E}(s^{N_t^{(k)}} \mid N_0^{(k)} = n)$. We study the generating function

$$g_t(s) := \mathbb{E}(s^{N_t^{(k)}} \mid N_0^{(k)} = 1) = P_t^{(k)}(f)(1).$$

By the branching property, $P_t^{(k)}(f)(n) = (P_t^{(k)}(f)(1))^n = g_t(s)^n$, hence

$$L^{(k)}(P_t^{(k)}(f))(n) = n(P_t^{(k)}(f)(n+k) - P_t^{(k)}(f)(n)) = n(g_t(s)^{n+k} - g_t(s)^n).$$

The backward Chapman–Kolmogorov equation now gives the ODE

$$\partial_t g_t(s) = \partial_t P_t^{(k)}(f)(1) = L^{(k)}(P_t^{(k)}(f))(1) = g_t(s)^{k+1} - g_t(s).$$

The initial condition $g_0(s) = s$ finally gives

$$g_t(s) = se^{-t}(1 - (1 - e^{-kt})s^k)^{-1/k}.$$

In particular, g_t^k is the generating function of kG, where G is a real-valued random variable with geometric distribution on $\mathbb{N}^*$ of parameter e^{-kt}. For convergence in distribution, for $\theta < 0$, the Laplace transform is

$$\begin{aligned}\mathbb{E}\Big(\exp\Big(\theta e^{-kt} N_t^{(k)}\Big)\Big) &= \exp\big(\theta e^{-kt}\big) e^{-t} \big(1 - \big(1 - e^{-kt}\big) \exp\big(\theta k e^{-kt}\big)\big)^{-1/k} \\ &= \big(e^{kt} \exp\big(-\theta k e^{-kt}\big) - e^{kt} + 1\big)^{-1/k}\end{aligned}$$

which converges as $t \to \infty$ to

$$(1-k\theta)^{-1/k} = \left(\frac{1/k}{1/k-\theta}\right)^{1/k},$$

the Laplace transform at θ of the distribution Gamma$(1/k, 1/k)$. Finally, $(e^{-k\lambda t}N_t^{(k)})_{t\geqslant 0}$ is a martingale bounded in L^2, and therefore converges a.s. and in L^2, as $t \to \infty$, to a random variable with distribution Gamma$(1/k, 1/k)$. □

Of course, fragmentation could be modeled more generally using a distribution on partitions of $[0, 1]$ with a possibly infinite number of pieces. This leads to the probabilistic theory of fragmentation.

Suppose that fragmentation into $k+1$ fragments divides the mass equally. At time $t \geqslant 0$, the fragment population consists of $N_t^{(k)}$ individuals positioned on a regular $k+1$ tree (each individual produces $k+1$ children and then dies). To number the individuals, we proceed as in Chapter 22, introducing the labels $\mathcal{U} = \cup_{r=0}^{\infty}(\mathbb{N}^*)^r$ with the convention $(\mathbb{N}^*)^0 = \varnothing$, where $(\mathbb{N}^*)^r$ numbers the individuals of generation r. Individuals alive at a given time t may belong to different generations. The population at time t is a random subset G_t of $\mathcal{U}$ and satisfies $N_t^{(k)} = |G_t|$. If $X_{u,t}$ represents the size of individual $u \in G_t$ alive at time t, then the distribution of masses in the population at time t is captured by the empirical measure

$$\mu_t = \frac{1}{N_t^{(k)}} \sum_{u\in G_t} \delta_{X_{u,t}}.$$

This random distribution gives the mass of an individual chosen uniformly among those alive at time t. For a test function $f : \mathbb{R}_+ \to \mathbb{R}$, we have

$$\mathbb{E}\left(\int f\, d\mu_t\right) = \mathbb{E}\left(\frac{1}{N_t^{(k)}} \sum_{u\in G_t} f(X_{u,t})\right) = \mathbb{E}(f(X_{U_t,t}))$$

where U_t is uniformly distributed on G_t. Intuitively, the evolution of mass along an arbitrary branch of the tree follows the TCP process. In fact, one must specify what is meant by an "arbitrary branch", since size bias may arise[7].

17.4 Further reading

The actual TCP window-size model is more complex and incorporates, in particular, a maximum window size. It is studied by Teunis Ott and Johannes Kamperman in [OK08], and by Vincent Dumas, Fabrice Guillemin, and Philippe Robert in [DGR02]. A connection with exponential functionals of Lévy processes is developed by Fabrice Guillemin, Philippe Robert,

[7] This is a classical sampling bias in statistics: a randomly chosen individual is more likely to come from a large family.

and Bert Zwart in [GRZ04]. The process is also discussed in the expository works of Philippe Robert [Rob05, Rob10]. Its long-time behavior is studied in [CMP10]. Graphical aspects are explored in the article [GR10] by Carl Graham and Philippe Robert. The branching version is studied by Vincent Bansaye, Jean-François Delmas, Laurence Marsalle, and Viet Chi Tran in [BDMT11], and by Bertrand Cloez in [Clo13]. The Yule process is named after Udny Yule and is studied in the books [Har02, AN04]. A connection with transport-fragmentation evolution equations can be found in the book by Benoît Perthame [Per07] and in the work of Bertrand Cloez [Clo13]. A theory of fragmentation is presented in the book by Jean Bertoin [Ber06].

Wasserstein distances are also known as coupling, transport, Kantorovich, Mallows, or Fréchet distances. These many names reflect their usefulness in various fields. These distances lie at the heart of optimal transport theory, developed in particular by Cédric Villani [Vil03]. If $E = \mathbb{R}$ and $d(x, y) = |x - y|$, then, writing F_μ^{-1} for the generalized inverse of the distribution function of μ, one can show that W_p is an L^p interquantile distance in the sense that

$$W_p(\mu_1, \mu_2) = \left(\int_0^1 |F_{\mu_1}^{-1}(x) - F_{\mu_2}^{-1}(x)|^p \, dx \right)^{1/p} = \|F_{\mu_1} - F_{\mu_2}\|_p.$$

18

Ruin of an insurance company

Keywords. Actuarial science; ruin time; hitting time; continuous-time Markov process.

Tools. Compound Poisson process; Laplace transform; characteristic function; martingale; random walk; law of large numbers; law of the iterated logarithm.

Difficulty. *

This chapter is mainly devoted to a simple, classical model of insurance risk, known as the Cramér–Lundberg model. A new insurance company wants to enter the market. It wishes to assess its probability of bankruptcy as a function of its initial capital and the price of its insurance policies. We assume that the company has initial capital c and denote by R_t its reserves at time t. The insurance company

- collects premiums from its customers, assumed to be paid monthly, spread uniformly over the year, and constant over time: the company's income over a time interval t is therefore pt, where p is the premium rate per unit time;
- pays compensation to policyholders who suffer losses, according to the damage incurred.

We model the occurrence and cost of claims as follows:

- the claim costs $(X_k)_{k\geqslant 1}$ are independent real-valued random variables with common distribution ν, common expectation λ, and finite variance $a-\lambda^2$;
- the claim arrival times $(T_n)_{n\geqslant 1}$ are distributed according to a Poisson process $(N_t)_{t\geqslant 0}$ with intensity μ. The time intervals between successive claims, denoted by $(\Delta_n)_{n\geqslant 1}$ with $\Delta_n = T_n - T_{n-1}$, are i.i.d. real-valued random variables with exponential distribution $\mathrm{Exp}(\mu)$.
- the sequence $(X_k)_{k\geqslant 1}$ and the process $(N_t)_{t\geqslant 0}$ are independent.

The insurance company's reserves at time t are therefore:

$$R_t = c + pt - \sum_{k=1}^{N_t} X_k.$$

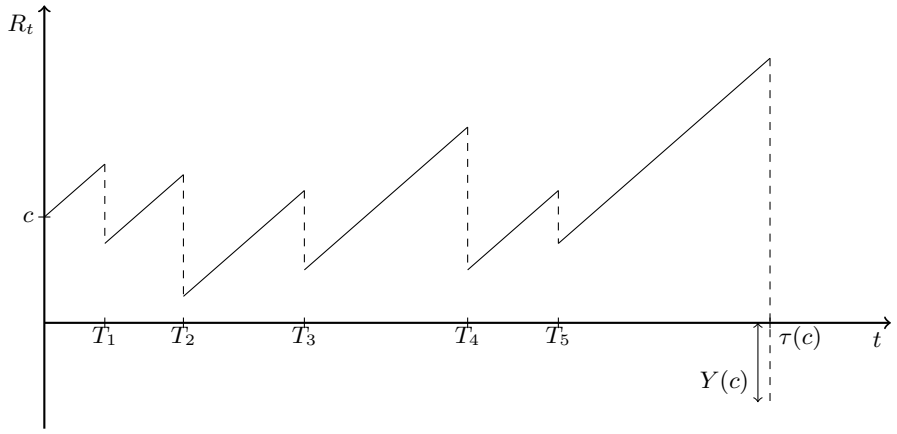


Fig. 18.1. A sample path leading to ruin at time $\tau(c)$ with a deficit $Y(c)$.

To simplify the calculations slightly, we instead consider the quantity

$$\tilde{R}_t = c - R_t = \sum_{k=1}^{N_t} X_k - pt.$$

Ruin occurs when the reserves R fall below 0 or when $\tilde{R}$ exceeds c. The company's ruin time, or default time, is defined by

$$\tau(c) = \inf\{t \geqslant 0 : R_t < 0\} = \inf\{t \geqslant 0 : \tilde{R}_t > c\}.$$

The company's probability of ruin is given by

$$\psi(c) = \mathbb{P}\left(\sup_{t \geqslant 0} \tilde{R}_t > c\right) = \mathbb{P}(\tau(c) < \infty).$$

Finally, we denote the deficit in the event of ruin by $Y(c)$ (see Figure 18.1):

$$Y(c) = -R_{\tau(c)} = \tilde{R}_{\tau(c)} - c.$$

Remark 18.1 (Markov process and infinitesimal generator)**.** *One can show that $\tilde{R}$ is a Markov process on $\mathbb{R}$ with infinitesimal generator*

$$Lf(x) = -pf'(x) + \mu \int_0^\infty (f(x+y) - f(x))\, \nu(dy).$$

This process evolves deterministically between its jump times. It is called a piecewise deterministic Markov process.

18.1 Compound Poisson process

Recall that the intensity of a Poisson process can easily be estimated from one of its sample paths, as the following result shows.

Lemma 18.2 (Long-time behavior)**.** *The following law of large numbers and central limit theorem hold:*

$$\frac{N_t}{t} \xrightarrow[t\to\infty]{\text{p.s.}} \mu \quad et \quad \sqrt{t}\left(\frac{N_t}{t} - \mu\right) \xrightarrow[t\to\infty]{\text{loi}} \mathcal{N}(0,\mu),$$

Proof. For every $i \in \mathbb{N}^*$, set $Y_i = N_i - N_{i-1}$. The variable Y_i represents the number of jumps of the Poisson process in the time interval $]i-1, i]$. The Poisson process has stationary independent increments. In particular, the random variables $(Y_i)_{i\geqslant 1}$ are independent with common Poisson distribution of parameter μ. Thus, by the law of large numbers, $N_{\lfloor t\rfloor}/\lfloor t\rfloor \to \mu$ a.s. Next, observe the bounds

$$\frac{N_{\lfloor t\rfloor}}{\lfloor t\rfloor}\frac{\lfloor t\rfloor}{t} \leqslant \frac{N_t}{t} \leqslant \frac{N_{\lfloor t\rfloor+1}}{\lfloor t\rfloor+1}\frac{\lfloor t\rfloor+1}{t}$$

which imply that $N_t/t \xrightarrow[t\to\infty]{\text{p.s.}} \mu$. Convergence in distribution follows by explicitly computing the characteristic function of $\sqrt{t}(N_t/t - \mu)$, since $N_t \sim \text{Poi}(\mu t)$. □

The process $C = (C_t)_{t\geqslant 0}$ defined by

$$C_t = \sum_{k=1}^{N_t} X_k,$$

is called a compound Poisson process with parameter μ and jump distribution ν.

Lemma 18.3 (Characteristic function)**.** *For all $t \geqslant 0$ and $u \in \mathbb{R}$,*

$$\varphi_{C_t}(u) := \mathbb{E}(e^{iuC_t}) = \exp(\mu t(\varphi_\nu(u) - 1)),$$

where φ_ν is the characteristic function of ν. If the first two moments λ and a of ν exist, then the first two moments of C_t exist and satisfy

$$\mathbb{E}(C_t) = \mu\lambda t \quad et \quad \text{Var}(C_t) = \mu a t,$$

and furthermore

$$\frac{C_t}{t} \xrightarrow[t\to\infty]{\text{p.s.}} \mu\lambda \quad et \quad \sqrt{t}\left(\frac{C_t}{t} - \mu\lambda\right) \xrightarrow[t\to\infty]{\text{loi}} \mathcal{N}(0,\mu a).$$

The moments of C_t can also be computed using Wald's lemma.

Proof. The expressions for the moments of C_t and the convergence in distribution follow from the expression for the characteristic function of C_t. Almost sure convergence follows from Lemma 18.2 and the law of large numbers. □

18.2 Is ruin almost sure?

Since the sample paths of $\tilde{R}$ decrease between jumps, the level c can only be crossed at a jump time. Define the sequence $(S_n)_{n\geqslant 0}$ by $S_n = \tilde{R}_{T_n}$. We call $(S_n)_{n\in\mathbb{N}}$ the embedded chain of $(\tilde{R}_t)_{t\in\mathbb{R}_+}$. By construction, since $\Delta_k = T_k - T_{k-1}$ for $k \geqslant 1$,

$$S_0 = 0 \quad \text{et} \quad \forall n \geqslant 1, \quad S_n = \sum_{k=1}^{n} (X_k - p\Delta_k).$$

In particular, S_n is the sum of n i.i.d. random variables with mean $\lambda - p/\mu$. The quantity $\lambda\mu$ represents the average claim amount per unit time, and its position relative to p determines, as the following result shows, the probability of ruin, which can also be written as

$$\psi(c) = \mathbb{P}\left(\sup_{n\geqslant 0} S_n > c\right).$$

Write $\tilde{\tau}(c) = \inf\{n \in \mathbb{N} : S_n > c\}$. By construction, we have

$$\tilde{\tau}(c) = n \Leftrightarrow \tau(c) = T_n \quad \text{et} \quad \tilde{\tau}(c) < +\infty \Leftrightarrow \tau(c) < +\infty.$$

Theorem 18.4 (Characterization of almost sure ruin)**.**

$$p \leqslant \lambda\mu \quad \Longleftrightarrow \quad \forall c \geqslant 0,\ \psi(c) = 1.$$

Proof. If $p < \lambda\mu$, the result follows from the law of large numbers applied to the sequence $(X_k - p\Delta_k)_{k\geqslant 1}$. If $p = \lambda\mu$, Strassen's law of the iterated logarithm applied to the centered i.i.d. random variables $(X_k - p\Delta_k)_{1\leqslant k\leqslant n}$ with variance $\sigma^2 = a - \lambda^2 + p^2/\mu^2$ gives

$$\liminf_{n\to\infty} \frac{S_n}{\sigma\sqrt{2n\log(\log(n))}} = -1 \quad \text{p.s.}$$

and

$$\limsup_{n\to\infty} \frac{S_n}{\sigma\sqrt{2n\log(\log(n))}} = +1 \quad \text{p.s.}$$

Finally, if $p > \lambda\mu$, the law of large numbers ensures that $S_n \to -\infty$ a.s. Suppose that the stopping time $\tau_1 = \inf\{n > 0 : S_n > c\}$ is finite a.s. The random walk $(S_{n+\tau_1} - S_{\tau_1})_{n\geqslant 0}$ has the same distribution as $(S_n)_{n\geqslant 0}$. Consequently, the stopping time $\tau_2 = \inf\{n > 0 : S_{n+\tau_1} - S_{\tau_1} > c\}$ is also finite a.s. Iterating this procedure shows that $\limsup S_n = \infty$, a contradiction. □

Remark 18.5 (Net profit condition)**.** *As intuition suggests, the company must ensure that $p > \lambda\mu$ to avoid almost sure ruin. Lemmas 18.2 and 18.3 allow the parameters λ and μ to be estimated from past claims, which can then help in choosing p.*

18.3 Expression for the probability of ruin

In what follows, we assume that the so-called net profit condition $p > \lambda\mu$ holds and seek an expression for the probability of ruin $\psi(c)$. Introduce the Laplace transform H of the claim distribution ν:

$$\forall x \in \mathbb{R}, \quad H(x) = \mathbb{E}\big(e^{xX_1}\big) \in [0, +\infty].$$

The function H is finite on an interval I containing $]-\infty, 0]$, since X_1 is a nonnegative random variable. Moreover, H is of class $\mathcal{C}^\infty$ and strictly convex on the interior of I.

Lemma 18.6 (Laplace transform of claims)**.** *Suppose that the interval I on which the Laplace transform H is finite is open and contains 0. Then the function*

$$x \mapsto H(x) - \frac{px}{\mu} - 1$$

vanishes at 0 and at a unique real number $u > 0$.

The real number u is called the *adjustment coefficient*.

Proof. The function $x \mapsto H(x) - px/\mu - 1$ is strictly convex on I and vanishes at 0 with a strictly negative derivative, since $H'(0) = \mathbb{E}(X_1) = \lambda < p/\mu$. On the other hand, since I is open, the function H tends to $+\infty$ at the right endpoint of I, and, for every $x_0 > 0$ such that $\mathbb{P}(X_1 \geqslant x_0) > 0$ and every $x \in I$,

$$H(x) \geqslant e^{xx_0}\mathbb{P}(X_1 \geqslant x_0).$$

Thus the function $x \mapsto H(x) - px/\mu - 1$ tends to $+\infty$ at the right endpoint of I. □

Theorem 18.7 (Probability of ruin)**.** *Under the assumptions of Lemma 18.6, the probability of ruin is given by*

$$\psi(c) = \frac{e^{-uc}}{\mathbb{E}\big(e^{uY(c)} | \tau(c) < +\infty\big)}$$

where $Y(c) = S_{\tilde{\tau}(c)} - c$ is the company's deficit at the time of ruin.

Proof. The sequence $(e^{uS_n})_{n\geqslant 0}$ is a nonnegative martingale because

$$\mathbb{E}\big(e^{uS_{n+1}} \mid \sigma(S_0, \ldots, S_n)\big) = e^{uS_n}\mathbb{E}\Big(e^{u(X_{n+1} - p\Delta_{n+1})}\Big).$$

Since X_{n+1} and Δ_{n+1} are independent and Δ_{n+1} has distribution $\mathrm{Exp}(\mu)$,

$$\mathbb{E}\Big(e^{u(X_{n+1} - p\Delta_{n+1})}\Big) = H(u)\frac{\mu}{\mu + pu} = 1.$$

Thus $E[e^{uS_n}] = 1$ for every $n \geqslant 0$. If $n > 0$, then $\tilde{\tau}(c) \wedge n$ is a bounded stopping time and

$$1 = \mathbb{E}\big(e^{uS_{\tilde{\tau}(c)\wedge n}}\big) = \mathbb{E}\big(e^{uS_{\tilde{\tau}(c)}}\mathbb{1}_{\{\tilde{\tau}(c)\leqslant n\}}\big) + \mathbb{E}\big(e^{uS_n}\mathbb{1}_{\{\tilde{\tau}(c)>n\}}\big).$$

Under the assumption $p > \lambda\mu$, $S_n \xrightarrow[n\to\infty]{\text{p.s.}} -\infty$. Moreover, on $\{\tilde{\tau}(c) > n\}$, we have $S_n \leqslant c$. Thus, by dominated convergence, the second expectation above converges to 0 as $n \to +\infty$. The first expectation converges by monotone convergence. Consequently, we obtain

$$1 = \mathbb{E}\big(e^{uS_{\tilde{\tau}(c)}}\mathbb{1}_{\{\tilde{\tau}(c)<+\infty\}}\big) = e^{uc}\mathbb{E}\Big(e^{uY(c)}|\tilde{\tau}(c) < +\infty\Big)\mathbb{P}(\tilde{\tau}(c) < +\infty),$$

which yields the stated result. □

Remark 18.8 (Lundberg bound)**.** *The preceding theorem gives* $\psi(c) \leqslant e^{-uc}$.

The Laplace transform of the measure with density

$$x \mapsto \frac{e^{-x}}{1+x^2}\mathbb{1}_{[0,\infty[}(x)$$

is finite at x if and only if $x \leqslant 1$. If $H(1) < 1 + p/\mu$, the function $x \mapsto H(x) - px/\mu - 1$ vanishes only at 0: this function, infinite on $]1, +\infty[$, is strictly negative on $]0, 1]$. In this case, Theorem 18.7 does not apply. However, the sequence $(e^{S_n})_{n\geqslant 0}$ is a supermartingale:

$$\mathbb{E}\big(e^{S_{n+1}} \,|\, \sigma(S_0, \ldots, S_n)\big) = e^{S_n}\frac{H(1)}{1+p/\mu} < e^{S_n}.$$

Proceeding as in the proof of Theorem 18.7, we deduce that

$$\psi(c) \leqslant \frac{e^{-c}}{\mathbb{E}\big(e^{Y(c)} \,|\, \tilde{\tau}(c) < +\infty\big)}.$$

18.4 Exponential claims

The probability of ruin is explicit when claim costs are exponentially distributed.

Theorem 18.9 (Probability of ruin)**.** *If claim costs have an exponential distribution with parameter* $1/\lambda$*, or mean* λ*, with* $p > \lambda\mu$*, then, for every* $c \geqslant 0$*, we have*

$$\psi(c) = \frac{\lambda\mu}{p}\exp\left(-\frac{p-\lambda\mu}{p\lambda}c\right).$$

Proof. For $x < 1/\lambda$, we have

$$H(x) = \frac{1}{1-\lambda x} \quad \text{et} \quad u = \frac{p-\mu\lambda}{p\lambda} < \frac{1}{\lambda}.$$

It suffices to show that, conditional on $\tilde{\tau}(c) < +\infty$, the variable $Y(c) = S_{\tilde{\tau}(c)} - c$ has an exponential distribution with expectation λ. If $\tilde{\tau}(c) = n$, then

$$Y(c) = S_n - c = X_n - p\Delta_n + S_{n-1} - c,$$

hence

$$\begin{aligned}
&\mathbb{P}(Y(c) > y, \tilde{\tau}(c) = n) \\
&\qquad = \mathbb{P}\Big(Y(c) > y, \max_{1\leqslant i\leqslant n-1} S_i \leqslant c, S_n > c\Big) \\
&\qquad\quad = \mathbb{P}\Big(X_n > y + p\Delta_n + c - S_{n-1}, \max_{1\leqslant i\leqslant n-1} S_i \leqslant c\Big) \\
&\qquad\qquad = \mathbb{P}\Big(X_n > y, X_n > y + p\Delta_n + c - S_{n-1}, \max_{1\leqslant i\leqslant n-1} S_i \leqslant c\Big).
\end{aligned}$$

The last equality follows from the fact that $X_n = Y(c) + c - S_{n-1} > y$ when $Y(c) > y$ and $\max_{1\leqslant i\leqslant n-1} S_i \leqslant c$. Now, since the random variable X_n has distribution $\mathrm{Exp}(1/\lambda)$ and is independent of $(\Delta_n, S_1, \ldots, S_{n-1})$, we have

$$\begin{aligned}
&\mathbb{P}(Y(c) > y, \tilde{\tau}(c) = n) \\
&\quad = \int_y^\infty \mathbb{P}\Big(x > y + p\Delta_n + c - S_{n-1}, \max_{1\leqslant i\leqslant n-1} S_i \leqslant c\Big)\, \lambda e^{-\lambda x}\, dx \\
&\qquad = e^{-\lambda y} \int_0^\infty \mathbb{P}\Big(x > p\Delta_n + c - S_{n-1}, \max_{1\leqslant i\leqslant n-1} S_i \leqslant c\Big)\, \lambda e^{-\lambda x}\, dx \\
&\qquad\quad = e^{-y/\lambda}\mathbb{P}\Big(S_n > c, \max_{1\leqslant i\leqslant n-1} S_i \leqslant c\Big) = e^{-y/\lambda}\mathbb{P}(\tilde{\tau}(c) = n) \qquad .
\end{aligned}$$

Thus $\mathbb{P}(Y(c) > y \,|\, \tilde{\tau}(c) < +\infty) = e^{-y/\lambda}$ for every $y > 0$. This ensures that

$$\mathbb{E}\Big(e^{uY(c)} \,|\, \tilde{\tau}(c) < +\infty\Big) = H(u).$$

We conclude by applying Theorem 18.7. □

Consider two companies insuring the same types of risk with the same pricing and compensation policies. Suppose that the customer base of the second differs from that of the first by a multiplicative factor ρ. If (λ, μ, p) denotes the parameters of the first company, those of the second are given by $(\lambda, \rho\mu, \rho p)$. We again assume that claims are exponentially distributed. By homogeneity, the probability of ruin of company $i = 1, 2$ is given by Theorem 18.9:

$$\psi(c_i) = \frac{\lambda\mu}{p} \exp\left(-\frac{p - \lambda\mu}{p\lambda} c_i\right).$$

Now suppose that they merge into a single company. Again, assume that the claims insured by the two companies are independent. The parameters of the new entity are $(\lambda, (1+\rho)\mu, (1+\rho)p)$ and its initial capital is $c_1 + c_2$. Once again, its probability of ruin is therefore given by $\psi(c_1 + c_2)$, which is smaller than the probability of ruin of either company. Insurers thus have an incentive to merge in order to pool their risks.

18.5 Further reading

The definition and essential properties of Poisson processes can be found, for example, in the books by Dominique Foata and Aimé Fuchs [FF04], James Norris [Nor98a], and Christiane Cocozza-Thivent [CT97]. Probability is one of the essential tools of *actuarial science*, the science of insurance. The model presented here, called the Cramér–Lundberg model, was introduced by Filip Lundberg around 1903 and popularized by Harald Cramér in the 1930s. Relevant references include the books by Søren Asmussen [Asm00], Thomas Mikosch [Mik09], and Arthur Charpentier and Michel Denuit [CD04]. From the perspective of stochastic process theory, the Cramér–Lundberg process is a piecewise deterministic Markov process (another example is treated in Chapter 17). A more sophisticated model accounts for insurers investing premium income as it is received. The infinitesimal generator of the process describing the company's wealth then takes the form

$$Lf(x) = (\gamma x + p) f'(x) + \mu \int_0^\infty (f(x-y) - f(x))\, \nu(dy).$$

The distributions of level crossings are then more difficult to describe.

19

Directed polymers in a random environment

Keywords. Threshold phenomenon; phase transition; FKG inequality; expectation depending on a parameter; directed polymer; random environment; Boltzmann-Gibbs measure.

Tools. Markov chain; Laplace transform; martingale; zero-one law.

Difficulty. ***

A polymer is a long molecule made of identical elementary building blocks called monomers. We are interested in modeling the structure of a polymer made of hydrophilic monomers, placed in an emulsion of oil and water. The oil is present as microscopic droplets. We assume that the monomers are regularly spaced and that the polymer extends in a preferred direction. The polymer is then said to be directed. By analogy with stochastic processes, this dimension is called time, and the polymer is viewed as a path in $\mathbb{Z}^d$, that is, a map from $\{0, 1, \ldots, n\}$ to $\mathbb{Z}^d$, where n is the polymer's length. At the points of $\mathbb{N} \times \mathbb{Z}^d$ we place i.i.d. random variables quantifying the sensitivity of each monomer of the polymer to the presence of oil or water at each site. Throughout the chapter, the integer d is fixed and at least 1.

The polymer's sensitivity to its environment is modeled by a real parameter denoted by β, introduced below. The question is to determine for which values of this parameter the polymer's behavior is noticeably different from that of a simple random walk, which models the structure of a polymer insensitive to its environment.

19.1 Model and main results

We model a polymer insensitive to its environment by a simple random walk on $\mathbb{Z}^d$, that is, a sequence $(S_n)_{n\geqslant 0}$ with $S_0 = 0$ and

$$S_n = \sum_{i=1}^{n} U_i, \quad n \geqslant 1,$$

where $(U_n)_{n\geqslant 1}$ is a sequence of i.i.d. random variables uniformly distributed on $\{\pm e_i : 1 \leqslant i \leqslant d\}$, where $e_1, \dots, e_d$ is the standard basis of $\mathbb{Z}^d$. We denote by $(\mathcal{F}_n)_{n\geqslant 0}$ the natural filtration of $(S_n)_{n\geqslant 0}$, and by $\mathcal{F}$ the sigma-algebra generated by the random variables $(U_n)_{n\geqslant 1}$.

It is convenient to represent the first $n+1$ terms of $(S_n)_{n\geqslant 0}$ by the random variable $S_{0:n} = (S_0, S_1, \dots, S_n)$, taking values in Γ_n, the set of paths starting from 0 and of length n in $\mathbb{Z}^d$, that is, the set of $(n+1)$-tuples $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_n)$ with $\gamma_0, \dots, \gamma_n$ in $\mathbb{Z}^d$, such that $\gamma_0 = 0$ and $\gamma_{i+1} - \gamma_i \in \{\pm e_j : 1 \leqslant j \leqslant d\}$ for every $0 \leqslant i \leqslant n-1$.

We model the environment by nonconstant i.i.d. real random variables

$$(\eta(n,x))_{n\geqslant 1,\, x\in\mathbb{Z}^d}$$

We assume that $(U_n)_{n\geqslant 1}$ and $(\eta(n,x))_{n\geqslant 0,\, x\in\mathbb{Z}^d}$ are independent. We denote by $\mathcal{G}$ the sigma-algebra generated by $(\eta(i,x))_{i\geqslant 1,\, x\in\mathbb{Z}^d}$ and by $\mathcal{G}_n$ the sigma-algebra generated by $(\eta(i,x))_{(i,x)\in E_n}$, where $E_n = \{(i,\gamma_i) : 1 \leqslant i \leqslant n,\ \gamma \in \Gamma_n\}$ is the environment seen by paths of length n.

For every $\beta \geqslant 0$, define the random probability measure μ_n on Γ_n by setting, for every $\gamma \in \Gamma_n$,

$$\mu_n(\gamma) = \frac{1}{Z_n(\beta)} e^{\beta H_n(\gamma)}\, \mathbb{P}(S_{0:n} = \gamma),$$

with

$$H_n(\gamma) = \sum_{i=1}^{n} \eta(i,\gamma_i) \quad \text{et} \quad Z_n(\beta) = \mathbb{E}\Big(e^{\beta H_n(S_{0:n})} \,|\, \mathcal{G}\Big).$$

The probability measure μ_n is a Gibbs measure (see Chapter 5). The quantity $H_n(S_{0:n})$ represents the affinity of the polymer's configuration for its environment. The higher it is, the more likely the polymer's configuration. The parameter β quantifies the influence of the environment on the polymer's structure. It is interpreted as the inverse of temperature. If $\beta = 0$ (infinite temperature), then all paths of length n have the same weight: the polymer is insensitive to the environment and behaves like the simple random walk. If $\beta = +\infty$ (zero temperature), the polymer assigns positive weight only to paths of length n maximizing H_n.

The quantity $Z_n(\beta)$ is a normalization constant called the *partition function* of the system. Set $\eta = \eta(0,0)$ and assume that the following function is well defined:

$$\lambda : t \in \mathbb{R} \mapsto \lambda(t) = \log \mathbb{E}\big(e^{t\eta}\big) \in \mathbb{R}_+.$$

The normalized partition function is

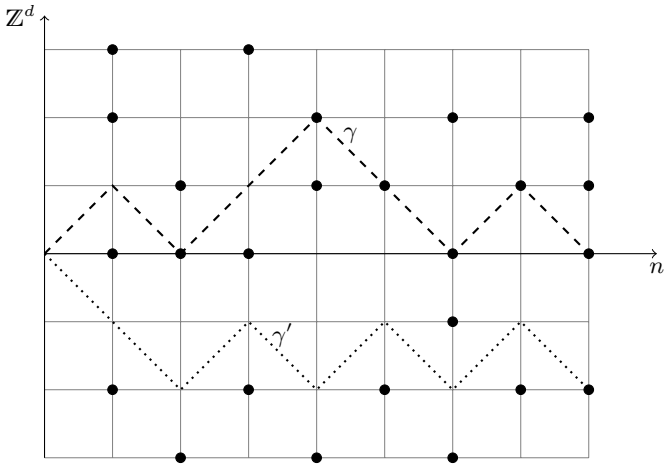


Fig. 19.1. Bernoulli environment: the black dots represent the sites where the environment equals +1). For paths γ and γ', $H_8(\gamma) = 6$ and $H_8(\gamma') = 1$. The first is favored over the second whenever $\beta > 0$.

$$W_n(\beta) = \mathbb{E}\Big(e^{\beta H_n(S_{0:n}) - n\lambda(\beta)} \,|\, \mathcal{G}\Big).$$

The system's *free energy* and its expectation are defined by

$$q_n(\beta) = \frac{1}{n} \log Z_n(\beta) \quad \text{et} \quad u_n(\beta) = \mathbb{E}(q_n(\beta)).$$

The quantities $Z_n(\beta)$, $q_n(\beta)$, $W_n(\beta)$, and μ_n are random, and we shall study their asymptotic behavior as n tends to infinity. There are several ways of measuring the influence of the environment on the polymer's structure. We present two, based on the asymptotic behavior of $q_n(\beta)$ and $W_n(\beta)$. The results are stated in this section and proved in the following ones.

We begin with the first significant result concerning a possible dichotomy in the polymer's behavior. It is based on the asymptotic behavior of the normalized partition function $W_n(\beta)$.

Theorem 19.1 (Threshold phenomenon)**.** *There exists a random variable $W_\infty(\beta)$ such that*

$$W_n(\beta) \xrightarrow[n\to\infty]{\text{p.s.}} W_\infty(\beta).$$

Moreover, there exists a constant $\bar{\beta}_c \in [0, +\infty]$ such that, almost surely,

$$\begin{cases} W_\infty(\beta) > 0 & \textit{si } \beta < \bar{\beta}_c, \textit{ et on dit que le polymère est en «faible désordre»}, \\ W_\infty(\beta) = 0 & \textit{si } \beta > \bar{\beta}_c, \textit{ et on dit que le polymère est en «fort désordre»}. \end{cases}$$

This result will be proved in Section 19.2. It yields the first notion of disorder. We now introduce the other notion of disorder. It relies on the following result.

Theorem 19.2 (Threshold phenomenon)**.** *There exists a function $u \leqslant \lambda$ such that*

$$u_n(\beta) \underset{n\to\infty}{\longrightarrow} u(\beta).$$

Moreover, there exists a constant $\beta_c \in [0, +\infty]$ such that

$$\begin{cases} u(\beta) = \lambda(\beta) & \text{si } \beta < \beta_c, \text{ (le polymère est en «très faible désordre»)}, \\ u(\beta) < \lambda(\beta) & \text{si } \beta > \beta_c, \text{ (le polymère est en «très fort désordre»)}. \end{cases}$$

This theorem is proved in Section 19.4. The two notions of disorder can be compared.

Lemma 19.3 (Very strong disorder implies strong disorder)**.** *We always have*

$$0 \leqslant \bar{\beta}_c \leqslant \beta_c \leqslant +\infty.$$

Proof. If $\beta < \bar{\beta}_c$ then $\mathbb{P}(\{W_\infty(\beta) > 0\}) = 1$ and consequently

$$(u - \lambda)(\beta) = \lim_{n\to\infty} \frac{1}{n} \log W_n(\beta) = 0.$$

In other words, $\beta \leqslant \beta_c$. □

The terminology requires care, since the model can simultaneously exhibit very weak disorder and strong disorder. The following table summarizes the situation.

β	$(0, \bar{\beta}_c)$	$(\bar{\beta}_c, \beta_c)$	$(\beta_c, +\infty)$
Fort Désordre	*non*	*oui*	*oui*
Très Fort Désordre	*non*	*non*	*oui*

The two essential questions are the following:

- do these threshold phenomena – sometimes called phase transitions in physics – actually occur? In other words, are the critical parameters finite and nonzero?
- are the two critical parameters different?

The answers may depend on the distribution of the environment and on the polymer's dimension d. In Section 19.3 we obtain an upper bound for β_c which may or may not be finite, depending on the environment's distribution and the dimension d. In Section 19.5, we show that if $d \geqslant 3$ then, for every environment, $\bar{\beta}_c > 0$. See Section 19.6 for the case of small dimensions, which turns out to be much more delicate.

Some proofs below rely on the following simple observation: if f and g are increasing, then $f(X)$ and $g(X)$ are positively correlated. This property,

called the FKG inequality[1], holds for functions of several independent random variables, as stated in Lemma 19.4.

Lemma 19.4 (FKG inequality)**.** *Let $X = (X_i)_{1 \leqslant i \leqslant k}$ be independent real random variables and $f, g : \mathbb{R}^k \to \mathbb{R}$ two functions such that $f(X)$ and $g(X)$ are square integrable, and increasing in the sense that $f(x) \leqslant f(y)$ if $x_i \leqslant y_i$ for every $1 \leqslant i \leqslant k$. Then $f(X)$ and $g(X)$ are positively correlated:*

$$\mathbb{E}[f(X)g(X)] \geqslant \mathbb{E}[f(X)]\,\mathbb{E}[g(X)].$$

Proof. For $k = 1$, if X' is an independent copy of X, we have

$$\mathbb{E}(f(X)g(X)) - \mathbb{E}f(X)\mathbb{E}g(X) = \frac{1}{2}\mathbb{E}[(f(X) - f(X'))(g(X) - g(X'))] \geqslant 0.$$

For $k > 1$, proceed by induction using conditioning:

$$\begin{aligned}
&\mathbb{E}(f(X)g(X)) - \mathbb{E}f(X)\mathbb{E}g(X) \\
&\qquad = \mathbb{E}(\mathbb{E}(f(X)g(X)|X_1) - \mathbb{E}(f(X)|X_1)\mathbb{E}(g(X)|X_1)) \\
&\qquad\quad + \mathbb{E}(\mathbb{E}(f(X)|X_1)\mathbb{E}(g(X)|X_1)) - \mathbb{E}f(X)\mathbb{E}g(X) \\
&\qquad = \mathbb{E}\mathrm{Cov}(f(X), g(X)|X_1) + \mathrm{Cov}(\mathbb{E}(f(X)|X_1), \mathbb{E}(g(X)|X_1)).
\end{aligned}$$

The first term is nonnegative by the induction hypothesis, since for fixed x_1 the functions f and g are increasing as functions of the $k-1$ variables $(x_2, \dots, x_k)$. The second is also nonnegative, since the functions $\mathbb{E}(f(X)|X_1 = \cdot)$ and $\mathbb{E}(g(X)|X_1 = \cdot)$ are increasing and the property holds for 1. □

19.2 Normalized partition function

This section is devoted to proving Theorem 19.1.

Lemma 19.5 (Convergence)**.** *The sequence $(W_n(\beta))_{n \geqslant 1}$ converges almost surely to a nonnegative random variable $W_\infty(\beta)$ whose expectation is at most 1.*

Proof. Jensen's inequality ensures that, for $p \geqslant 1$,

$$\mathbb{E}(W_n(\beta)^p) = \mathbb{E}\Big(\mathbb{E}\Big(e^{\beta H_n(S_{0:n}) - n\lambda(\beta)} \,|\, \mathcal{G}\Big)^p\Big) \leqslant \mathbb{E}\Big(e^{p\beta H_n(S_{0:n})}\Big) e^{-pn\lambda(\beta)}.$$

For every $\gamma \in \Gamma_n$, we have

$$\mathbb{E}\Big(e^{p\beta H_n(\gamma)}\Big) = \big[\mathbb{E}\big(e^{p\beta\eta}\big)\big]^n = e^{n\lambda(p\beta)}.$$

The random variable W_n^p is therefore integrable and $\mathbb{E}(W_n(\beta)^p) \leqslant e^{n(\lambda(p\beta) - p\lambda(\beta))}$. For $p = 1$, the upper bound is an equality: $\mathbb{E}(W_n(\beta)) = 1$. For every $n \geqslant 1$,

[1] For Fortuin, Kasteleyn, and Ginibre.

$$\begin{aligned}\mathbb{E}(W_{n+1}(\beta)\,|\,\mathcal{G}_n) &= \frac{e^{-(n+1)\lambda(\beta)}}{(2d)^{n+1}} \sum_{\gamma\in\Gamma_{n+1}} \mathbb{E}\Big(e^{\beta\sum_{i=1}^{n+1}\eta(i,\gamma_i)}\,|\,\mathcal{G}_n\Big)\\ &= \frac{e^{-(n+1)\lambda(\beta)}}{(2d)^{n+1}} \sum_{\gamma\in\Gamma_{n+1}} e^{\beta\sum_{i=1}^{n}\eta(i,\gamma_i)} e^{\lambda(\beta)}\\ &= \frac{1}{(2d)^n}\sum_{\gamma\in\Gamma_n} e^{\beta H_n(\gamma)-n\lambda(\beta)} = W_n(\beta).\end{aligned}$$

The sequence $(W_n(\beta))_{n\geqslant 1}$ is therefore a nonnegative martingale with respect to the filtration $(\mathcal{G}_n)_n$. It therefore converges almost surely to a nonnegative random variable $W_\infty(\beta)$. Moreover, by Fatou's lemma,

$$\mathbb{E}(W_\infty(\beta)) = \mathbb{E}\Big(\liminf_{n\to\infty} W_n(\beta)\Big) \leqslant \liminf_{n\to\infty}\mathbb{E}(W_n(\beta)) = 1,$$

which is the claimed result. □

So far, nothing ensures that the martingale converges in L^1 or that $W_\infty(\beta)$ has expectation 1. The next result shows that $W_\infty(\beta)$ is either zero or strictly positive almost surely.

Lemma 19.6 (Zero-one law)**.** *For every $\beta\geqslant 0$ and every $\theta\in(0,1)$,*

$$\mathbb{P}(W_\infty(\beta)=0)\in\{0,1\} \quad \textit{et} \quad \lim_{n\to\infty}\mathbb{E}(W_n(\beta)^\theta) = \mathbb{E}(W_\infty(\beta)^\theta).$$

Proof. The Markov property of the walk ensures that

$$W_{n+m}(\beta) = \mathbb{E}\Big(e^{\beta H_n(S_{0:n})-n\lambda(\beta)} Z_{n,m}^{S_n} e^{-m\lambda(\beta)}\,|\,\mathcal{G}\Big)$$

where, for $x\in\mathbb{Z}^d$ and positive integers n and m,

$$Z_{n,m}^x(\beta) = \mathbb{E}\left[\exp\left(\beta\sum_{i=1}^m \eta(n+i,x+S_i)\right)\,|\,\mathcal{G}\right].$$

Thus,

$$W_\infty(\beta) = \lim_{m\to\infty} W_{n+m}(\beta) = \mathbb{E}\Big(e^{\beta H_n(S_{0:n})-n\lambda(\beta)} \lim_{m\to\infty}\Big(Z_{n,m}^{S_n}e^{-m\lambda(\beta)}\Big)\,|\,\mathcal{G}\Big).$$

Since $e^{\beta H_n(S_{0:n})-n\lambda(\beta)}$ is strictly positive almost surely, the event

$$\{W_\infty(\beta)=0\} = \Big\{\lim_{m\to\infty}\Big(Z_{n,m}^x e^{-m\lambda(\beta)}\Big)=0 \ :\ \forall x,\ \mathbb{P}(S_n=x)>0\Big\}$$

belongs to the sigma-algebra $\sigma\big(\eta(j,x),\ j\geqslant n,\ x\in\mathbb{Z}^d\big)$. This holds for every n. Hence it belongs to the tail sigma-field $\bigcap_{n\geqslant 1}\sigma\big(\eta(j,x): j\geqslant n,\ x\in\mathbb{Z}^d\big)$. By Kolmogorov's zero-one law, every event in the tail sigma-algebra has probability 0 or 1. Since $W_n(\beta)$ has expectation 1, the sequence $(W_n(\beta)^\theta)_{n\geqslant 0}$ is bounded in $L^{1/\theta}$. Moreover, it converges to $W_\infty(\beta)^\theta$ almost surely. Hence

$$\lim_{n\to\infty} \mathbb{E}(W_n(\beta)^\theta) = \mathbb{E}(W_\infty(\beta)^\theta).$$

This limit is zero if and only if $W_\infty(\beta)$ is zero almost surely. □

It remains to establish a monotonicity property for the dichotomy above.

Lemma 19.7 (Monotonicity)**.** *For every $\theta \in]0,1[$, $\beta \mapsto \mathbb{E}(W_\infty(\beta)^\theta)$ is decreasing.*

Proof. First observe that

$$\partial_\beta W_n(\beta) = \mathbb{E}\Big((H_n(S_{0:n}) - n\lambda'(\beta))e^{\beta H_n(S_{0:n})-n\lambda(\beta)} \,|\, \mathcal{G}\Big).$$

Differentiating under the expectation therefore gives

$$\partial_\beta \mathbb{E}(W_n(\beta)^\theta) = \theta\mathbb{E}\Big(W_n(\beta)^{\theta-1}(H_n(S_{0:n}) - n\lambda'(\beta))e^{\beta H_n(S_{0:n})-n\lambda(\beta)}\Big).$$

The FKG inequality of Lemma 19.4, applied to the probability measure with density $e^{\beta H_n(\gamma)-n\lambda(\beta)}$ with respect to $\mathbb{P}$ and the functions $W_n(\beta)^{\theta-1}$ and $H_n(\gamma) - n\lambda'(\beta)$ – respectively decreasing and increasing functions of the variables $(\eta(k,x))_{(k,x)\in E_n}$ – ensures that

$$\mathbb{E}\Big(W_n(\beta)^{\theta-1}(H_n(\gamma) - n\lambda'(\beta))e^{\beta H_n(\gamma)-n\lambda(\beta)}\Big)$$

is at most

$$\mathbb{E}\Big(W_n(\beta)^{\theta-1}e^{\beta H_n(\gamma)-n\lambda(\beta)}\Big)\mathbb{E}\Big((H_n(\gamma) - n\lambda'(\beta))e^{\beta H_n(\gamma)-n\lambda(\beta)}\Big).$$

To conclude, we use

$$\mathbb{E}\Big((H_n(\gamma) - n\lambda'(\beta))e^{\beta H_n(\gamma)-n\lambda(\beta)}\Big) = \partial_\beta \mathbb{E}(W_n(\beta)) = 0.$$

□

We can now conclude. Passing to the limit $\theta \to 1$ shows that

$$\beta \mapsto \mathbb{E}(W_\infty(\beta))$$

is decreasing. Thus there exists $\bar{\beta}_c \in [0, +\infty]$ such that, almost surely,

$$\begin{cases} W_\infty(\beta) > 0 & \text{si } \beta < \bar{\beta}_c, \\ W_\infty(\beta) = 0 & \text{si } \beta > \bar{\beta}_c. \end{cases}$$

This completes the proof of Theorem 19.1.

19.3 A lower bound in high dimensions

Here we work in dimension $d \geqslant 3$. In this case, the simple random walk $(S_n)_{n\geqslant 0}$ is transient (see Theorem 2.7). Its probability of returning to 0 is strictly less than 1. Denote this probability by π_d.

Theorem 19.8 (A criterion for nonvanishing)**.** *If $d \geqslant 3$ and*

$$\tau(\beta) = \lambda(2\beta) - 2\lambda(\beta) < \log(1/\pi_d)$$

then $W_\infty > 0$ almost surely.

Proof. For every $\beta \geqslant 0$, we have $\tau'(\beta) = 2(\lambda'(2\beta) - \lambda'(\beta)) > 0$ since λ is strictly convex, being the logarithm of a Laplace transform. It follows that τ is strictly increasing. Now let $(U'_n)_{n\geqslant 1}$ have the same distribution as $(U_n)_{n\geqslant 1}$, with

$$(U_n)_{n\geqslant 1}, \quad (U'_n)_{n\geqslant 1}, \quad (\eta(n,x))_{n\geqslant 0,\, x\in\mathbb{Z}^d}$$

independent. Let $(S'_n)_{n\geqslant 1}$ be the sequence constructed from $(U'_n)_{n\geqslant 1}$ in the same way as $(S_n)_{n\geqslant 1}$ from $(U_n)_{n\geqslant 1}$. By independence of (S_n) and (S'_n),

$$\begin{aligned} W_n(\beta)^2 &= \mathbb{E}\left(\prod_{k=1}^n e^{\beta\eta(k,S_k)-\lambda(\beta)} \mid \mathcal{G}\right)\mathbb{E}\left(\prod_{k=1}^n e^{\beta\eta(k,S'_k)-\lambda(\beta)} \mid \mathcal{G}\right) \\ &= \mathbb{E}\left(\prod_{k=1}^n e^{\beta(\eta(k,S_k)+\eta(k,S'_k))-2\lambda(\beta)} \mid \mathcal{G}\right). \end{aligned}$$

Taking expectations gives

$$\mathbb{E}\big(W_n^2\big) = \mathbb{E}\left[\mathbb{E}\left(\prod_{k=1}^n e^{\beta(\eta(k,S_k)+\eta(k,S'_k))-2\lambda(\beta)} \mid \mathcal{F}'\vee\mathcal{F}\right)\right].$$

Now, for every $x, x' \in \mathbb{Z}^d$,

$$\mathbb{E}\Big(e^{\beta(\eta(k,x)+\eta(k,x'))-2\lambda(\beta)}\Big) = \begin{cases} e^{\tau(\beta)} & \text{si } x = x', \\ 1 & \text{si } x \neq x'. \end{cases}$$

Hence

$$\mathbb{E}\big(W_n(\beta)^2\big) = \mathbb{E}\Big(e^{\tau(\beta)I_n}\Big) \quad \text{où} \quad I_n = \sum_{k=0}^n \mathbb{1}_{\{S_k=S'_k\}}.$$

Observe that the sequences $(S_k - S'_k)_{k\geqslant 0}$ and $(S_{2k})_{k\geqslant 0}$ have the same distribution, since $-U'_n$ and U_n are independent and identically distributed. Thus the random variable I_n has the same distribution as the number N_{2n} of returns to 0 of S before time $2n$. By monotone convergence,

$$\mathbb{E}\Big(e^{\tau(\beta)I_n}\Big) = \mathbb{E}\Big(e^{\tau(\beta)N_{2n}}\Big) \underset{n\to\infty}{\longrightarrow} \mathbb{E}\Big(e^{\tau(\beta)N_\infty}\Big).$$

The strong Markov property of S ensures that N_∞ has the geometric distribution with parameter $1-\pi_d$ and that, for every $t \in \mathbb{R}$,

$$\mathbb{E}\big(e^{tN_\infty}\big) = \sum_{k=1}^{\infty}(1-\pi_d)\pi_d^{k-1}e^{tk} = \begin{cases} \dfrac{(1-\pi_d)e^t}{1-\pi_d e^t} & \text{si } t < -\log(\pi_d), \\ +\infty & \text{sinon.} \end{cases}$$

We thereby obtain a necessary and sufficient condition for the sequence $(W_n(\beta))_{n\geqslant 1}$ to be bounded in L^2:

$$\sup_{n\geqslant 1}\mathbb{E}\big(W_n(\beta)^2\big) < \infty \Longleftrightarrow \tau(\beta) < \log(1/\pi_d).$$

If $\tau(\beta) < \log(1/\pi_d)$, then $(W_n(\beta))_{n\geqslant 1}$ is a martingale bounded in L^2, and therefore converges almost surely and in L^2 to a square-integrable random variable $W_\infty(\beta)$. In particular, $\mathbb{E}(W_\infty(\beta)) = \lim_n \mathbb{E}(W_n(\beta)) = 1$ and therefore $W_\infty(\beta) > 0$ almost surely, by the zero-one law of Lemma 19.6. □

Corollary 19.9 (A criterion for nonvanishing)**.** *If $d \geqslant 3$, then for every environment, $\bar\beta_c > 0$.*

Proof. The function τ vanishes at 0 and is smooth on $\mathbb{R}$. It is also strictly increasing, since

$$\tau'(\beta) = 2(\lambda'(2\beta) - \lambda'(\beta)) > 0,$$

by strict convexity of λ. The condition of Theorem 19.8 is therefore always satisfied for sufficiently small β, regardless of the environment. □

Example 19.10 (Gaussian environment)**.** *If $\eta \sim \mathcal{N}(0,1)$, then $\tau(\beta) = \beta^2$ and $\tau(\beta) < \log(1/\pi_d)$ if and only if $\beta < \sqrt{\log(1/\pi_d)}$. Hence $\bar\beta_c > \sqrt{\log(1/\pi_d)}$.*

Example 19.11 (Bernoulli environment)**.** *If $\eta \sim \mathrm{Ber}(p)$, then*

$$\lambda(\beta) = \log\big(pe^\beta + 1 - p\big) \quad et \quad \lim_{\beta\to\infty}\tau(\beta) = -\log(p).$$

Thus, whenever $p > \pi_d$, the condition of Theorem 19.8 is satisfied for every $\beta \geqslant 0$, and $\bar\beta_c$ is infinite.

19.4 Mean free energy

The main purpose of this section is to prove Theorem 19.2. We first show that the mean free energy converges.

Lemma 19.12 (Convergence of the mean free energy)**.** *For every $\beta \geqslant 0$, there exists a constant $u(\beta)$ at most $\lambda(\beta)$ such that*

$$u_n(\beta) \underset{n\to\infty}{\longrightarrow} u(\beta).$$

Proof. Jensen's inequality ensures that

$$u_n(\beta) = \frac{1}{n}\mathbb{E}(\log(Z_n(\beta))) \leqslant \frac{1}{n}\log \mathbb{E}(Z_n(\beta)).$$

Since $\mathbb{E}(W_n(\beta)) = 1$, we have $\mathbb{E}(Z_n(\beta)) = e^{n\lambda(\beta)}$ and $u_n(\beta) \leqslant \lambda(\beta)$.

Conditionally on $\mathcal{G}$, the Markov property of the walk S ensures that

$$\mathbb{E}\Big(e^{\beta H_{n+m}(S_{0:n+m})} \,|\, \mathcal{G}\Big) = \mathbb{E}\Big(e^{\beta H_n(S_{0:n})} Z_{n,m}^{S_n} \,|\, \mathcal{G}\Big).$$

By definition of $Z_n(\beta)$ and μ_n, we simply have

$$\begin{aligned}\mathbb{E}\Big(e^{\beta H_n(S_{0:n})} Z_{n,m}^{S_n} \,|\, \mathcal{G}\Big) &= \frac{\mathbb{E}\big(e^{\beta H_n(S_{0:n})} Z_{n,m}^{S_n} \,|\, \mathcal{G}\big)}{\mathbb{E}\big(e^{\beta H_n(S_{0:n})} \,|\, \mathcal{G}\big)} Z_n(\beta) \\ &= \mu_n\big(Z_{n,m}^{S_n}\big) Z_n(\beta).\end{aligned}$$

In other words,

$$\log(Z_{n+m}(\beta)) = \log(Z_n(\beta)) + \log \sum_{x\in\mathbb{Z}^d} \mu_n(S_n = x) Z_{n,m}^x.$$

Jensen's inequality for the measure μ_n ensures that

$$\log(Z_{n+m}(\beta)) \geqslant \log(Z_n(\beta)) + \sum_{x\in\mathbb{Z}^d} \mu_n(S_n = x) \log(Z_{n,m}^x).$$

The random measure μ_n is $\mathcal{G}_n$-measurable and $Z_{n,m}^x$ is independent of $\mathcal{G}_n$. Hence

$$\mathbb{E}\left(\sum_{x\in\mathbb{Z}^d} \mu_n(S_n = x) \log(Z_{n,m}^x) \,|\, \mathcal{G}_n\right) = \sum_{x\in\mathbb{Z}^d} \mu_n(S_n = x)\mathbb{E}\big(\log(Z_{n,m}^x) \,|\, \mathcal{G}_n\big).$$

The random variables $Z_{n,m}^x$ and Z_m have the same distribution because the random variables $(\eta(n,x))_{n,x}$ are i.i.d. Taking expectations shows that the sequence $(\mathbb{E}(\log(Z_m)))_{n\geqslant 1}$ is *superadditive*, and therefore, by Fekete's lemma (see page 217), the sequence $(u_n(\beta))_{n\geqslant 1}$ converges in $\mathbb{R}\cup\{+\infty\}$. Since it is bounded by $\lambda(\beta)$, its limit $u(\beta)$ is finite. □

The threshold phenomenon will follow from a monotonicity property.

Lemma 19.13 (Monotonicity)**.** *The function $\beta \mapsto u(\beta) - \lambda(\beta)$ is decreasing on $\mathbb{R}_+$.*

Proof. We study the variation of $\beta \mapsto u_n(\beta) - \lambda(\beta)$. We have

$$\frac{\partial}{\partial\beta}\mathbb{E}(\log(Z_n(\beta))) = \mathbb{E}\left(\frac{\mathbb{E}\big(H_n(S_{0:n})e^{\beta H_n(S_{0:n})} \,|\, \mathcal{G}\big)}{Z_n(\beta)}\right)$$

$$= \frac{1}{(2d)^n} \sum_{\gamma \in \Gamma_n} \mathbb{E}\Big(\frac{H_n(\gamma) e^{\beta H_n(\gamma)}}{Z_n(\beta)} \Big).$$

Observe that $H_n(\gamma)$ (respectively $Z_n^{-1}(\beta)$) is an increasing (respectively decreasing) function of the variables $(\eta(i,x))_{(i,x)\in E_n}$. Moreover, the measure with density $e^{\beta H_n(\gamma) - n\lambda(\beta)}$ with respect to the measure $\mathbb{P}$ is a probability measure. The FKG inequality of Lemma 19.4 therefore gives

$$\mathbb{E}\Big(\frac{H_n(\gamma)}{Z_n(\beta)} e^{\beta H_n(\gamma) - n\lambda(\beta)} \Big) \leqslant \mathbb{E}\Big(H_n(\gamma) e^{\beta H_n(\gamma) - n\lambda(\beta)} \Big) \mathbb{E}\Big(Z_n(\beta)^{-1} e^{\beta H_n(\gamma) - n\lambda(\beta)} \Big).$$

We thus obtain

$$\begin{aligned} \frac{\partial}{\partial \beta} \mathbb{E}(\log(Z_n(\beta))) &\leqslant \frac{1}{(2d)^n} \sum_{\gamma \in \Gamma_n} e^{-n\lambda(\beta)} \mathbb{E}\Big(\frac{e^{\beta H_n(\gamma)}}{Z_n(\beta)} \Big) \mathbb{E}\Big(H_n(\gamma) e^{\beta H_n(\gamma)} \Big) \\ &\leqslant n\lambda'(\beta) \mathbb{E}\left(\frac{1}{(2d)^n} \sum_{\gamma \in \Gamma_n} \frac{e^{\beta H_n(\gamma)}}{Z_n(\beta)} \right) \\ &\leqslant n\lambda'(\beta). \end{aligned}$$

In other words, $(u_n - \lambda)'(\beta) \leqslant 0$: for every $n \geqslant 1$, the function $u_n - \lambda$ is decreasing. This remains true on passing to the limit as $n \to \infty$: the function $u - \lambda$ is decreasing on $\mathbb{R}_+$. □

The functions u and λ vanish at 0. Thus there exists $\beta_c \in [0, +\infty]$ such that

$$\begin{cases} u(\beta) = \lambda(\beta) & \text{si } \beta < \beta_c, \\ u(\beta) < \lambda(\beta) & \text{si } \beta > \beta_c. \end{cases}$$

19.5 An upper bound on the critical parameter

The aim of this section is to establish an upper bound on β_c.

Lemma 19.14. *The functions $(q_n)_{n\geqslant 1}$ and u satisfy the following properties:*

1. *q_n is strictly convex, $q_n(0) = 0$; u is convex and vanishes at 0.*
2. *$\beta \mapsto \beta^{-1} q_n(\beta)$ is increasing.*
3. *$\beta \mapsto \beta^{-1}(q_n(\beta) + \log(2d))$ is decreasing.*

Proof. The function $\beta \mapsto n q_n(\beta)$ is the logarithm of the Laplace transform of the real random variable $H_n(S_{0:n})$ (the variables $\eta(\cdot,\cdot)$ being fixed). Thus q_n is of class $\mathcal{C}^2$, convex, and zero at 0. These properties are preserved by taking expectations and then the limit as n tends to infinity: u is convex on $\mathbb{R}_+$ and

zero at 0. Since q_n is convex, the function (the slope of the chord between $(0,0)$ and $(\beta, q_n(\beta))$) $\beta \mapsto q_n(\beta)/\beta = (q_n(\beta) - q_n(0))/\beta$ is increasing.

For assertion 3, differentiation yields

$$\left(\frac{q_n(\beta) + \log(2d)}{\beta}\right)' = -\frac{1}{\beta^2}(q_n(\beta) + \log(2d)) + \frac{1}{n\beta}\mu_n(H_n).$$

Moreover, the entropy of μ_n defined by

$$\begin{aligned} h_n(\mu_n) &= \sum_{\gamma \in \Gamma_n} \frac{e^{\beta H(\gamma)}}{Z_n(2d)^n} \log\left(\frac{e^{\beta H(\gamma)}}{Z_n(2d)^n}\right) \\ &= \beta\, \mu_n(H_n) - (\log(Z_n) + n\log(2d)) \end{aligned}$$

is nonpositive. Thus,

$$\left(\frac{q_n(\beta) + \log(2d)}{\beta}\right)' = \frac{1}{n\beta^2} h_n(\mu_n) \leqslant 0.$$

The function $\beta \mapsto \beta^{-1}(q_n(\beta) + \log(2d))$ is decreasing on $\mathbb{R}_+^*$. □

Lemma 19.14 provides an upper bound for u and β_c.

Theorem 19.15 (Upper bounds)**.** *If the following condition holds:*

$$\lim_{\beta \to +\infty} \beta\lambda'(\beta) - \lambda(\beta) > \log(2d),$$

then the following equation in β

$$\beta\lambda'(\beta) = \lambda(\beta) + \log(2d)$$

has a unique root β_1 *and, for every* $\beta > \beta_1$*,*

$$u(\beta) \leqslant \frac{\beta}{\beta_1}(\lambda(\beta_1) + \log(2d)) - \log(2d),$$

Finally, $\beta_c \leqslant \beta_1$*.*

Proof. Since λ is strictly convex, $\beta \mapsto \beta\lambda'(\beta) - \lambda(\beta)$ is strictly increasing on $\mathbb{R}_+$ (its derivative equals $\beta \mapsto \beta\lambda''(\beta)$). It also vanishes at 0. The equation $\beta\lambda'(\beta) = \lambda(\beta) + \log(2d)$ therefore has at most one solution. A solution exists if and only if the condition of the theorem is satisfied.

Lemma 19.14 ensures that the function $\beta \mapsto (u_n(\beta) + \log(2d))/\beta$ is decreasing. Since also $u_n \leqslant \lambda$, we obtain

$$\begin{aligned} \frac{u_n(\beta) + \log(2d)}{\beta} &= \inf_{\alpha \in]0,\beta]} \frac{u_n(\alpha) + \log(2d)}{\alpha} \\ &\leqslant \inf_{\alpha \in]0,\beta]} \frac{\lambda(\alpha) + \log(2d)}{\alpha}. \end{aligned}$$

The function $\alpha \mapsto (\lambda(\alpha) + \log(2d))/\alpha$ attains its minimum on $[0, \beta]$ at β or at the point where its derivative vanishes, namely β_1. Thus, for every $\beta > \beta_1$,

$$u(\beta) \leqslant \frac{(\lambda(\beta_1) + \log(2d))}{\beta_1}\beta - \log(2d).$$

It remains to observe that for $\beta > \beta_1$,

$$\frac{(\lambda(\beta_1) + \log(2d))}{\beta_1}\beta - \log(2d) < \lambda(\beta)$$

since λ is strictly convex and equality holds at β_1. □

Example 19.16 (Gaussian and Bernoulli environments)**.** *If $\eta \sim \mathcal{N}(0,1)$ then $\beta\lambda'(\beta) - \lambda(\beta) = \beta^2/2$, and in this case $\beta_1 = \sqrt{2\log(2d)}$. In contrast, if $\eta \sim \mathrm{Ber}(p)$ then*

$$\beta\lambda'(\beta) - \lambda(\beta) = \frac{p\beta e^\beta}{pe^\beta + 1 - p} - \log\left(pe^\beta + 1 - p\right) \underset{\beta\to\infty}{\sim} -\log(p),$$

and β_1 is finite if and only if $2dp < 1$.

19.6 Further reading

This model of a directed polymer in a random environment was introduced by John Imbrie and Thomas Spencer [IS88]. The notion of strong disorder was studied in detail by Erwin Bolthausen in [Bol89], who proves, in particular, Theorems 19.1 and 19.8. Later, very strong disorder was studied, in particular, by Francis Comets and Nobuo Yoshida [CY06], who prove Theorem 19.2. Using a concentration inequality for q_n around its mean, one can (see [CH02, LW09]) strengthen the conclusions of Theorem 19.2 by proving that, for every $\beta \geqslant 0$,

$$q_n(\beta) \xrightarrow[n\to\infty]{\text{p.s.}} u(\beta) \quad \text{et} \quad q_n(\beta) \xrightarrow[n\to\infty]{L^1} u(\beta).$$

In dimensions 1 and 2, it is known that $\beta_c = 0$. Philippe Carmona and Yueyun Hu established in [CH02] that $\bar{\beta}_c = 0$ for $d = 1$ and $d = 2$ when the environment is Gaussian. The general case is treated by Francis Comets, Tokuzo Shiga, and Nobuo Yoshida in [CSY03]. Subsequently, Francis Comets and Vincent Vargas established in [CV06] that β_c vanishes for $d = 1$. Finally, Hubert Lacoin proved in [Lac10] that this also holds for $d = 2$. Whether the two critical parameters are equal when $d \geqslant 3$ appears to remain an open question.

The FKG inequality, due to Cees Fortuin, Pieter Kasteleyn, and Jean Ginibre [FKG71], is an important mathematical tool in statistical physics. See, for example, Thomas Liggett's book [Lig85] on interacting particle systems.

To conclude, we point out that, for clarity, we have slightly modified the standard notation used in the articles cited above. Indeed, the reference measure is usually written as the product of two measures, one for the environment

and the other for the distribution of the walk. One must then introduce two expectation symbols, corresponding to conditional expectations with respect to the sigma-algebras $\mathcal{F}$ and $\mathcal{G}$ in this chapter.

20

The traveling salesman problem

Keywords. Randomized combinatorial optimization.

Tools. Azuma-Hoeffding inequality; subadditivity; Poissonization; Poisson point process.

Difficulty. ***

Here are three flagship problems in randomized combinatorial optimization:

- traveling salesman problem[1];
- minimum spanning tree[2];
- minimum Euclidean matching[3].

The methods used include subadditivity, concentration of measure, Poissonization, and the objective method. We shall study the first problem and use some of these techniques. The traveling salesman problem consists in finding a tour, that is, a circular path (a *Hamiltonian cycle*) of minimal length through prescribed points $X_1, \dots, X_n \in \mathbb{R}^d$, $n, d \geqslant 2$. This amounts to solving, over the symmetric group,

$$\min_{\sigma \in \mathcal{S}_n} \sum_{k=1}^{n} \left| X_{\sigma(k)} - X_{\sigma(k+1)} \right|$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$, $\mathcal{S}_n$ is the symmetric group on $\{1, \dots, n\}$, and we use the convention $\sigma(n+1) = \sigma(1)$. Clearly, the choice of distance influences the solution. The combinatorial explosion of the symmetric group means that once n exceeds a few dozen, it is no longer possible to test

[1] Abbreviated TSP.
[2] In English: "*Minimum Spanning Tree*".
[3] In English: "*Minimum Euclidean Matching*".

all permutations to find the best one. It is, however, possible to use a global optimization algorithm such as *simulated annealing*, discussed in Section 5.3, to produce an approximate solution in reasonable time: a permutation for which the minimum is (almost) attained.

Rather than seeking an optimal permutation, in this chapter we study the minimum value and its behavior when n is large and the points $X_1, \ldots, X_n$ are independent random variables with common distribution μ on $\mathbb{R}^d$. We denote by $L_n = L_n(X_1, \ldots, X_n)$ the minimum tour length, which is a function of $X_1, \ldots, X_n$.

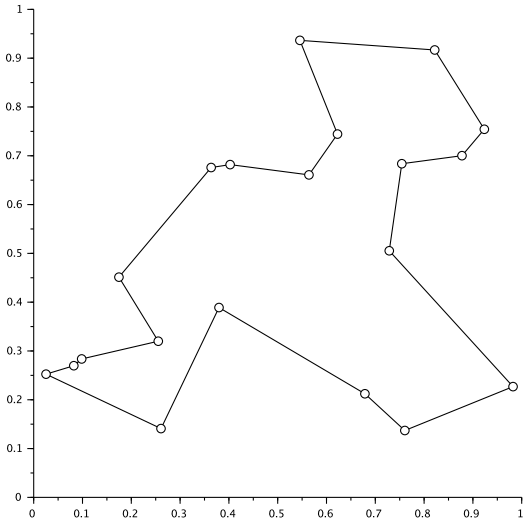


Fig. 20.1. Shortest route for $n = 20$ points uniformly distributed in the unit square, obtained by the stochastic simulated annealing algorithm (Chapter 4).

Theorem 20.1 (Bearwood-Halton-Hammersley)**.** *There exists a constant* $0 < \gamma_d < \infty$ *depending on* $d \geqslant 2$ *such that if* μ *has compact support and density* $f : \mathbb{R}^d \to \mathbb{R}_+$ *with respect to Lebesgue measure, then*

$$\frac{L_n(X_1, \ldots, X_n)}{n^{(d-1)/d}} \xrightarrow[n\to\infty]{\text{p.s.}} \gamma_d \int_{\mathbb{R}^d} f(x)^{(d-1)/d}\, dx.$$

In particular, $L_n(X_1, \ldots, X_n)$ is of order $\sqrt{n}$ in dimension $d = 2$. We shall prove this theorem when μ is uniform on the cube $[0, 1]^d$. We shall show that the random variable $L_n(X_1, \ldots, X_n)$ becomes increasingly concentrated

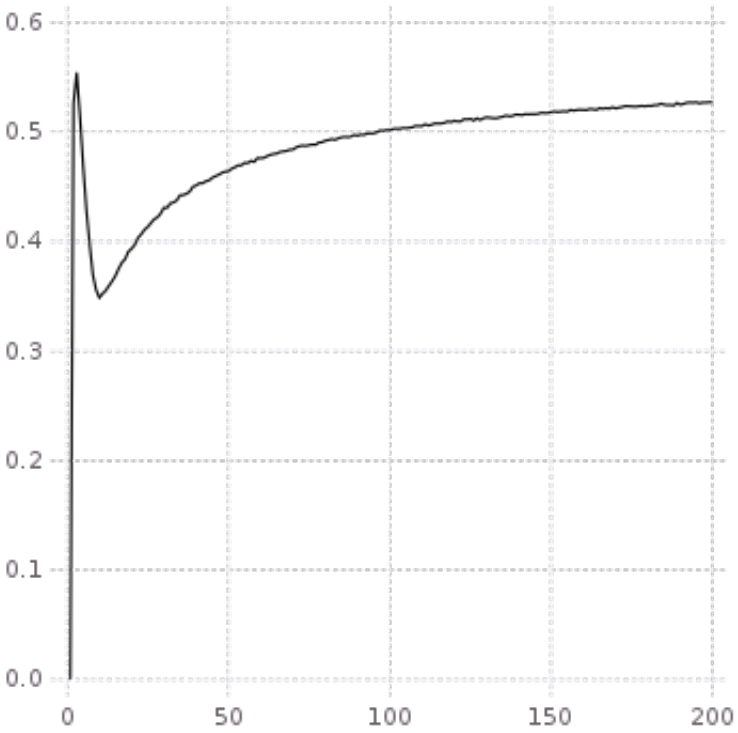


Fig. 20.2. Plot of an approximation of $n \mapsto n^{-1/2}\mathbb{E}(L_n)$ in the uniform case, using a Monte Carlo method and the simulated annealing algorithm (Chapter 5).

around its expectation $\mathbb{E}(L_n(X_1, \ldots, X_n))$ as n grows, and then that this expectation is of order $n^{(d-1)/d}$.

20.1 Concentration in the uniform case

Theorem 20.2 (Concentration around the mean)**.** *If μ is uniform on the cube $[0,1]^d$, then for every $n \geqslant 1$ and every real $t \geqslant 0$,*

$$\mathbb{P}(|L_n - \mathbb{E}(L_n)| \geqslant t) \leqslant 2\exp\begin{cases} -\dfrac{ct^2}{\log(n)} & \text{si } d = 2, \\ -\dfrac{ct^2}{n^{(d-2)/d}} & \text{si } d > 2, \end{cases}$$

where $c > 0$ is a constant depending only on the dimension d. Thus,

$$\mathbb{P}\left(\left|\frac{L_n}{n^{(d-1)/d}} - \mathbb{E}\left(\frac{L_n}{n^{(d-1)/d}}\right)\right| \geqslant t\right) \leqslant 2\exp\begin{cases} -\dfrac{cnt^2}{\log(n)} & si\ d = 2, \\ -cnt^2 & si\ d > 2. \end{cases}$$

The proof of Theorem 20.2 relies on two lemmas.

A random variable is said to be concentrated when it stays close to its mean with high probability. Such a property can be obtained by controlling the tail of its distribution, for example through moments, as in Chebyshev's inequality based on variance, and Chernoff's inequality based on the Laplace transform (exponential moments). The Azuma-Hoeffding inequality uses information about the oscillation of the variable (the diameter of its support). We have $\mathrm{osc}(f) \leqslant 2\|f\|_\infty$, where

$$\mathrm{osc}(f) := \sup f - \inf f = \text{diamètre}(\mathrm{support}(f)).$$

Lemma 20.3 (Azuma-Hoeffding concentration inequality)**.** *Consider an integrable random variable $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$. Then, for every $r \geqslant 0$,*

$$\mathbb{P}(|Y - \mathbb{E}(Y)| \geqslant r) \leqslant 2\exp\left(-\frac{2r^2}{\mathrm{osc}(d_1)^2 + \cdots + \mathrm{osc}(d_n)^2}\right)$$

for the telescoping decomposition into a sum of martingale differences [4]

$$Y - \mathbb{E}(Y) = \sum_{k=1}^{n} \mathbb{E}(Y \,|\, \mathcal{F}_k) - \mathbb{E}(Y \,|\, \mathcal{F}_{k-1}) = \sum_{k=1}^{n} d_k$$

associated with an arbitrary interpolating filtration

$$\{\varnothing, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

Proof. Let U be a real random variable such that $\mathbb{E}(U) = 0$ and $a \leqslant U \leqslant b$. Convexity of $u \mapsto e^u$ gives, for every $t \geqslant 0$ and every $x \in [a, b]$,

$$e^{tx} \leqslant \frac{x-a}{b-a}e^{tb} + \frac{b-x}{b-a}e^{ta}.$$

Setting $p = -a/(b-a)$ and $f(u) = -pu + \log(1 - p + pe^u)$ therefore yields

$$\mathbb{E}(e^{tU}) \leqslant \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = e^{ta}\Big((1-p) + pe^{t(b-a)}\Big) = e^{f(t(b-a))}.$$

Now we have

$$f'(u) = -p + \frac{pe^u}{1 - p + pe^u} \quad \text{et} \quad f''(u) = p(1-p)\frac{e^u}{(1 - p + e^u)^2} \leqslant \frac{1}{4}.$$

[4] This is sometimes called a Doob martingale.

Since $f(0) = f'(0) = 0$, we deduce that $f(u) \leqslant u^2/8$ and therefore

$$\mathbb{E}(e^{tU}) \leqslant e^{\frac{t^2}{8}(b-a)^2}.$$

Applied to $U = d_k = \mathbb{E}(Y \,|\, \mathcal{F}_k) - \mathbb{E}(Y \,|\, \mathcal{F}_{k-1})$ conditionally on $\mathcal{F}_{k-1}$, this gives

$$\mathbb{E}(e^{td_k} \,|\, \mathcal{F}_{k-1}) \leqslant e^{\frac{t^2}{8}\operatorname{osc}(d_k)^2}.$$

Next, writing the telescoping sum $Y - \mathbb{E}(Y) = d_n + \cdots + d_1$, we obtain

$$\mathbb{E}(e^{t(Y-\mathbb{E}(Y))}) = \mathbb{E}(e^{t(d_{n-1}+\cdots+d_1)}\mathbb{E}(e^{td_n} \,|\, \mathcal{F}_{n-1})) \leqslant \cdots \leqslant e^{\frac{t^2}{8}c},$$

where $c := \sum_{k=1}^n \operatorname{osc}(d_k)^2$. Now, for every $t, r > 0$, by Markov's inequality,

$$\begin{aligned}\mathbb{P}(Y - \mathbb{E}(Y) \geqslant r) &= \mathbb{P}(e^{t(Y-\mathbb{E}(Y))} \geqslant e^{tr}) \\ &\leqslant e^{-tr}\mathbb{E}(e^{t(Y-\mathbb{E}(Y))}) \\ &\leqslant e^{-tr+\frac{ct^2}{8}}.\end{aligned}$$

Thus, for every $r \geqslant 0$,

$$\mathbb{P}(Y - \mathbb{E}(Y) \geqslant r) \leqslant \exp\left(\inf_{t>0}\left\{-tr + \frac{ct^2}{8}\right\}\right) = \exp\left(-\frac{2r^2}{c}\right).$$

Applying this to the variables Y and $-Y$ gives the desired result for $\mathbb{P}(|Y - \mathbb{E}(Y)| \geqslant r)$. □

Lemma 20.4 (Geometric lemma)**.** *There exists a constant $c_d > 0$ such that if $X_1, \ldots, X_k$ are i.i.d. uniform on $[0,1]^d$, then for every $x \in [0,1]^d$,*

$$g_k(x) := \mathbb{E}\Big(\min_{1\leqslant i\leqslant k} |X_i - x|\Big) \leqslant c_d k^{-1/d}.$$

Proof. If $B(x,r)$ denotes the ball with center x and radius $r > 0$ in $\mathbb{R}^d$, the minimum volume of $B(x,r) \cap [0,1]^d$ as x ranges over $[0,1]^d$ is attained when x is a corner of the cube $[0,1]^d$. When $r \leqslant 1$, the minimum equals $2^{-d}|B(0,r)| = 2^{-d}|B(0,1)|r^d$ (draw a picture). If $1 < r \leqslant \sqrt{d}$, the minimum is difficult to compute. It nevertheless remains at least as large as in the case $r = 1$. Thus, for every $0 < r \leqslant \sqrt{d}$, this minimum volume is at least $2^{-d}|B(0,1)|(r/\sqrt{d})^d = a_d r^d$. Hence, for every $x \in [0,1]^d$ and every $0 < r \leqslant \sqrt{d}$, using the convexity inequality $1 - u \leqslant e^{-u}$ in the last step,

$$\begin{aligned}\mathbb{P}\Big(\min_{1\leqslant i\leqslant k} |X_i - x| \geqslant r\Big) &= \prod_{i=1}^k \mathbb{P}(X_i \in B(x,r)^c) \\ &= \big(1 - |B(x,r) \cap [0,1]^d|\big)^k \\ &\leqslant \big(1 - a_d r^d\big)^k\end{aligned}$$

$$\leqslant \exp\left(-a_d k r^d\right).$$

This remains valid if $r > \sqrt{d}$ since then $\mathbb{P}(\min_{1\leqslant i\leqslant k} |X_i - x| \geqslant r) = 0$. We have

$$\mathbb{E}\Big(\min_{1\leqslant i\leqslant k} |X_i - x|\Big) = \int_0^\infty \mathbb{P}\Big(\min_{1\leqslant i\leqslant k} |X_i - x| \geqslant r\Big)\, dr$$

and

$$\int_0^\infty e^{-br^d}\, dr = \frac{\Gamma(1/d)}{db^{1/d}}.$$

□

Proof of Theorem 20.2. Apply the Azuma-Hoeffding inequality (Lemma 20.3) to $Y = L_n(X_1, \ldots, X_n)$ and the filtration $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ with $\mathcal{F}_0 = \{\varnothing, \Omega\}$. This reduces the problem to finding a uniform upper bound for

$$d_k = \mathbb{E}(L_n(X_1, \ldots, X_n) \,|\, \mathcal{F}_k) - \mathbb{E}(L_n(X_1, \ldots, X_n) \,|\, \mathcal{F}_{k-1}).$$

If $(X_1', \ldots, X_n')$ is an independent copy of $(X_1, \ldots, X_n)$, observe that

$$\mathbb{E}(L_n(X_1, \ldots, X_k, \ldots, X_n) \,|\, \mathcal{F}_{k-1}) = \mathbb{E}(L_n(X_1, \ldots, X_k', \ldots, X_n) \,|\, \mathcal{F}_k)$$

and therefore

$$d_k := \mathbb{E}(L_n(X_1, \ldots, X_k, \ldots, X_n) - L_n(X_1, \ldots, X_k', \ldots, X_n) \,|\, \mathcal{F}_k).$$

We now bound $\|d_k\|_\infty$ from above. Let L be the function assigning to a finite subset of $\mathbb{R}^d$ its minimum tour length. For every finite $S \subset \mathbb{R}^d$ and every $x \in \mathbb{R}^d$,

$$L(S) \leqslant L(S \cup \{x\}) \leqslant L(S) + 2 \min_{y\in S} |x - y|.$$

Applying this inequality to $S = \{x_1, \ldots, x_n\} \setminus \{x_k\}$ and to $x = x_k$ and $x = x_k'$ gives

$$\begin{aligned} |L_n(x_1, \ldots, x_k, \ldots, x_n) &- L_n(x_1, \ldots, x_k', \ldots, x_n)| \\ &\leqslant 2 \min_{i\neq k} |x_k' - x_i| + 2 \min_{i\neq k} |x_k - x_i| \\ &\leqslant 2 \min_{i>k} |x_k' - x_i| + 2 \min_{i>k} |x_k - x_i|. \end{aligned}$$

Setting $g_k(x) := \mathbb{E}\Big(\min_{1\leqslant i\leqslant k} |X_i - x|\Big)$ gives, for $1 \leqslant k \leqslant n-1$,

$$\begin{aligned} |d_k| &\leqslant 2\mathbb{E}\Big(\min_{i>k} |X_k - X_i| + \min_{i>k} |X_k' - X_i| \,\Big|\, \mathcal{F}_k\Big) \\ &= 2g_{n-k}(X_k) + 2\mathbb{E}(g_{n-k}(X_k')). \end{aligned}$$

The geometric lemma 20.4 gives, for $d \geqslant 2$ and $1 \leqslant k \leqslant n-1$,

$$\|d_k\|_\infty \leqslant c_d (n-k)^{-1/d}.$$

Now $\|d_n\|_\infty \leqslant c_d$ for another constant, since the X_i are bounded, hence

$$\sum_{k=1}^{n} \|d_k\|_\infty^2 \leqslant \begin{cases} c_d \log(n) & \text{pour } d = 2, \\ c_d n^{(d-2)/d} & \text{pour } d > 2. \end{cases}$$

□

20.2 Estimating the mean in the uniform case

Lemma 20.5 (A good start)**.** *If $\mu \sim \mathrm{Unif}([0,1]^d)$, then for every $n \geqslant 1$,*

$$\|L_n(X_1, \ldots, X_n)\|_\infty \leqslant c_d n^{(d-1)/d}$$

and

$$c_d^- n^{(d-1)/d} \leqslant \mathbb{E}(L_n(X_1, \ldots, X_n)) \leqslant c_d^+ n^{(d-1)/d}$$

where $0 < c_d^- \leqslant c_d^+ < \infty$ are constants depending only on d.

Proof. The cube $[0,1]^d$ is the union of $(1/\varepsilon)^d$ small cubes isometric to $[0,\varepsilon]^d$. Taking $\varepsilon = n^{-1/d}$ shows that $[0,1]^d$ can be covered by $\mathcal{O}(n)$ small cubes of diameter $\mathcal{O}(n^{-1/d})$. By the pigeonhole principle[5], for all $x_1, \ldots, x_n \in [0,1]^d$ we have, for a constant c_d which may depend on d,

$$\min_{1 \leqslant i \neq j \leqslant n} |x_i - x_j| \leqslant c_d n^{-1/d}.$$

Consequently, if ℓ_n is the maximum over $x_1, \ldots, x_n$ of the minimum tour length for $x_1, \ldots, x_n$, then

$$\ell_n \leqslant \ell_{n-1} + 2c_d n^{-1/d}$$

which gives, for a new constant c_d possibly depending on d,

$$\|L_n(X_1, \ldots, X_n)\|_\infty \leqslant c_d n^{-1/d+1} = c_d n^{(d-1)/d}.$$

Moreover, the proof of Lemma 20.4 shows that $|B(x,r) \cap [0,1]^d|$ is maximal when x is the center of the cube; hence, for every $x \in [0,1]^d$ and every $0 < r \leqslant 1/2$,

$$\mathbb{P}\Big(\min_{1 \leqslant i \leqslant n-1} |X_i - x| \geqslant r \Big) \geqslant (1 - \omega_d r^d)^{n-1}$$

with $\omega_d := |B(0,1)|$. Using the elementary inequality $(1-u)^\alpha \geqslant 1 - \alpha u$ valid whenever $0 \leqslant u \leqslant 1/\alpha$, we deduce that

$$\mathbb{E}\Big(\min_{1 \leqslant i \leqslant n-1} |X_i - x| \Big) \geqslant \int_0^{1/2} (1 - \omega_d r^d)^{n-1} \, dr \geqslant c_d n^{-1/d}.$$

[5] If n objects are placed in m boxes with $n > m$, then at least one box contains two or more objects.

Thus,

$$\min_{1\leqslant j\leqslant n}\mathbb{E}\Big(\min_{1\leqslant i\neq j\leqslant n}|X_i-X_j|\Big)=\min_{1\leqslant j\leqslant n}\mathbb{E}\Big(\mathbb{E}\Big(\min_{1\leqslant i\neq j\leqslant n}|X_i-X_j|\,|\,X_j\Big)\Big)$$
$$\geqslant c_d n^{-1/d}$$

where c_d is a constant depending on the dimension d. Now the optimal cycle through $X_1,\ldots,X_n$ has n edges and visits each of the n vertices. We can therefore write $L_n(X_1,\ldots,X_n)=\sum_{i=1}^n|X_i-X_{j_i}|$ for random integers $j_1,\ldots,j_n$, giving

$$\begin{aligned}\mathbb{E}(L_n(X_1,\ldots,X_n))&\geqslant\sum_{j=1}^n\mathbb{E}\Big(\min_{1\leqslant i\neq j\leqslant n}|X_i-X_j|\Big)\\&\geqslant nc_dn^{-1/d}\\&=c_dn^{(d-1)/d}.\end{aligned}$$

□

Theorem 20.6 (The desired result)**.** *If μ is uniform on $[0,1]^d$, then*

$$\lim_{n\to\infty}\frac{\mathbb{E}(L_n(X_1,\ldots,X_n))}{n^{(d-1)/d}}=\gamma_d$$

where $0<\gamma_d<\infty$ is a real number depending on d.

We already know that $a_n:=\mathbb{E}(L_n)\approx n^{(d-1)/d}$, which makes it natural to try to show that $n^{d/(d-1)}a_n$ converges as n tends to infinity. This nonlinear behavior prevents direct use of a subadditivity argument. It is nevertheless possible to linearize the problem by Poissonization followed by de-Poissonization. The heuristic is as follows: if N is an integer-valued random variable, then $\mathbb{E}(a_N)\approx\mathbb{E}(N^{(d-1)/d})\approx\mathbb{E}(N)^{(d-1)/d}$ which is linear in t when $N\sim\mathrm{Poi}(t^{d/(d-1)})$. Moreover, if $N\sim\mathrm{Poi}(n)$ then $a_n\approx\mathbb{E}(a_N)$.

Proof. We proceed in steps.
Poissonization. Write $L(S)$ for the minimum tour length for a finite set of points $S=\{x_1,\ldots,x_n\}\subset\mathbb{R}^d$, with the convention $L(S)=0$ if $\mathrm{card}(S)\leqslant 2$. Let P be a Poisson point process on $\mathbb{R}^d$ with Lebesgue intensity measure. Let $(Z_t)_{t\geqslant 0}$ be the process defined by $Z_t=L(P\cap[0,t]^d)$, that is, the minimum tour length for the atoms of the Poisson process P lying in the cube $[0,t]^d$. For every $n\geqslant 0$,

$$\mathrm{Loi}(P\,|\,\mathrm{card}(P\cap[0,t]^d)=n)=\mathrm{Unif}([0,t]^d)^{\otimes n}.$$

On the other hand, $L(tS)=tL(S)$ and $\mathrm{card}(P\cap[0,t]^d)\sim\mathrm{Poi}(t^d)$, giving

$$\mathbb{E}(Z_t)=\sum_{n=0}^\infty\mathbb{E}(Z_t\,|\,\mathrm{card}(P\cap[0,t]^d)=n)\mathbb{P}(\mathrm{card}(P\cap[0,t]^d)=n)$$

$$= \sum_{n=0}^{\infty} \mathbb{E}(L(P \cap [0,t]^d) \,|\, \mathrm{card}(P \cap [0,t]^d) = n) e^{-t^d} \frac{t^{dn}}{n!}$$
$$= e^{-t^d} \sum_{n=0}^{\infty} t a_n \frac{t^{dn}}{n!}$$

where

$$a_n := \mathbb{E}(L_n(X_1, \ldots, X_n)).$$

The factor t in front of a_n arises because a_n concerns $[0,1]^d$ rather than $[0,t]^d$. Lemma 20.5 implies that $a_n = \mathcal{O}(n^{(d-1)/d})$, and thus $t \mapsto \mathbb{E}(Z_t)$ is continuous and even analytic. The sequence $(a_n)_{n \geqslant 1}$ can therefore be recovered from knowledge of the function $t \mapsto \mathbb{E}(Z_t)$ near $t = 0$. However, we shall instead show that $\lim_{t\to\infty} t^{-1}\mathbb{E}(Z_t)$ exists, which will suffice.

Subadditivity. Let $C_1, \ldots, C_{k^d}$ be a partition (up to boundaries) of $[0,t]^d$ into k^d cubes similar to $[0, t/k]^d$. For all $x_1, \ldots, x_n \in \mathbb{R}^d$, the k^d tours of the sets $S_i = \{x_1, \ldots, x_n\} \cap C_i$ can be concatenated to form a tour of $\{x_1, \ldots, x_n\} \cap [0,t]^d$, at an additional cost of $\mathcal{O}(k^{d-1})$: choose a point on each of the k^d tours and connect them by a tour of cost $\mathcal{O}(k^{d-1})$ using the first bound of Lemma 20.5, yielding one large necklace-shaped tour. Thus there exists a constant $c > 0$ such that, for every $k \geqslant 1$ and every $t \geqslant 0$,

$$L(\{x_1, \ldots, x_n\} \cap [0,t]^d) \leqslant \sum_{i=1}^{k^d} L(\{x_1, \ldots, x_n\} \cap C_i) + ctk^{d-1}.$$

This gives $\mathbb{E}(Z_t) \leqslant k^d \mathbb{E}(Z_{t/k}) + ctk^{d-1}$, hence a first estimate:

$$\frac{\mathbb{E}(Z_{tk})}{(tk)^d} \leqslant \frac{\mathbb{E}(Z_t)}{t^d} + ct^{1-d}.$$

Taking $t = 1$ yields

$$0 \leqslant \gamma := \varliminf_{k\to\infty} \frac{\mathbb{E}(Z_k)}{k^d} \leqslant \mathbb{E}(Z_1) + c < \infty.$$

By definition of γ, for every $\varepsilon > 0$ we may choose k_0 sufficiently large that

$$\frac{\mathbb{E}(Z_{k_0})}{k_0^d} + ck_0^{1-d} \leqslant \gamma + \varepsilon.$$

Since $t \mapsto \mathbb{E}(Z_t)$ is continuous, we can choose $\delta > 0$ such that, for every $k_0 < t < k_0 + \delta$,

$$\frac{\mathbb{E}(Z_t)}{t^d} + ct^{1-d} \leqslant \gamma + 2\varepsilon.$$

Our first estimate shows that this second estimate holds for every $kk_0 < t < k(k_0 + \delta)$. Now, for $k > k_0/\delta$, the intervals $I_k :=]kk_0, k(k_0 + \delta)[$ and

I_{k+1} overlap. We deduce that it holds for every $t \geqslant \lfloor k_0^2/\delta \rfloor$, which implies, in particular,

$$\overline{\lim_{t\to\infty}} \frac{\mathbb{E}(Z_t)}{t^d} \leqslant \underline{\lim_{t\to\infty}} \frac{\mathbb{E}(Z_t)}{t^d} + 2\varepsilon = \gamma + 2\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we therefore obtain

$$\lim_{t\to\infty} \frac{\mathbb{E}(Z_t)}{t^d} = \gamma.$$

Thus the series expansion of $\mathbb{E}(Z_t)$ gives, after the change of variable $u = t^d$,

$$\sum_{k=0}^{\infty} e^{-u} a_k \frac{u^k}{k!} \underset{n\to\infty}{\sim} \gamma u^{(d-1)/d}.$$

De-Poissonization. We first show that $n \mapsto a_n$ is sufficiently regular. Since a cycle through $X_1, \ldots, X_m$ and a cycle through $X_{m+1}, \ldots, X_{m+n}$ can be concatenated at a cost less than $2\sqrt{d}$ (the diameter of the cube), we deduce that

$$a_{n+m} \leqslant a_n + a_m + 2\sqrt{d}$$

giving $0 \leqslant a_n - a_k \leqslant a_{n-k} + 2\sqrt{d}$ for every $0 \leqslant k \leqslant n$. Now Lemma 20.5 gives $a_n = \mathcal{O}(n^{(d-1)/d})$, from which we deduce

$$|a_n - a_k| \leqslant |a_{n-k} + 2\sqrt{d}| \leqslant c|n-k|^{(d-1)/d}.$$

If N is a Poisson random variable with mean n, then

$$\mathbb{E}(a_N) = \sum_{k=0}^{\infty} a_k e^{-n} \frac{n^k}{k!}$$

and therefore

$$\begin{aligned}
|a_n - \mathbb{E}(a_N)| &\leqslant \sum_{k=0}^{\infty} |a_n - a_k| e^{-n} \frac{n^k}{k!} \\
&\leqslant c \sum_{k=0}^{\infty} |n-k|^{(d-1)/d} e^{-n} \frac{n^k}{k!} \\
&= c\mathbb{E}(|N - \mathbb{E}(N)|^{(d-1)/d}) \\
&\leqslant c\mathbb{E}(|N - \mathbb{E}(N)|^2)^{(d-1)/(2d)} \\
&= c\mathrm{Var}(N)^{(d-1)/(2d)} \\
&= \mathcal{O}(n^{(d-1)/(2d)}) = o(n^{(d-1)/d}).
\end{aligned}$$

The Poissonization result with $t = n$ gives

$$\mathbb{E}(a_N) = \sum_{k=0}^{\infty} a_k e^{-n} \frac{n^k}{k!} \underset{n\to\infty}{\sim} \gamma n^{(d-1)/d}$$

and therefore

$$\lim_{n\to\infty} \frac{\mathbb{E}(L_n(X_1,\ldots,X_n))}{n^{(d-1)/d}} = \lim_{n\to\infty} \frac{a_n}{n^{(d-1)/d}} = \gamma.$$

The fact that $\gamma > 0$ follows from the lower bound in Lemma 20.5. □

20.3 Proof in the uniform case

We are now ready to prove Theorem 20.1 in the case where μ is uniform on the cube $[0,1]^d$. By Theorem 20.2 and the Borel-Cantelli lemma, almost surely,

$$L_n - \mathbb{E}(L_n) = \begin{cases} \mathcal{O}_{n\to\infty}(\log(n)) & \text{si } d = 2, \\ \mathcal{O}_{n\to\infty}(n^{(d-2)/(2d)}\sqrt{\log(n)}) & \text{si } d > 2 \end{cases}$$

whereas Theorem 20.6 gives, for some $0 < \gamma < \infty$,

$$\mathbb{E}(L_n(X_1,\ldots,X_n)) \sim_{n\to\infty} \gamma n^{(d-1)/d}.$$

20.4 Further reading

When $d = 2$ and $X_1,\ldots,X_n$ are i.i.d. uniform on $[0,1]^2$, one can intuitively see why $L_n(X_1,\ldots,X_n)$ should be of order $\sqrt{n}$ for large n by traversing the square in regularly spaced parallel zigzags, say horizontally: there are approximately $\sqrt{n}$ lines of unit length, each containing approximately $\sqrt{n}$ points. More rigorously, this leads to an approach to the traveling salesman problem in dimension $d = 2$ using a "space-filling" curve.

This chapter is directly inspired by Michael Steele's fine book [Ste97]. On this topic, one may also consult Joseph Yukich's book [Yuk98] and the book by Marc Mézard and Andrea Montanari [MM09]. The Azuma-Hoeffding concentration inequality is studied by Colin McDiarmid in [McD89, McD98]. Theorem 20.1 was obtained by John Hammersley and his students Jillian Beardwood and John Halton around 1959, and can be proved by reduction to the uniform case on the cube $[0,1]^d$. It is also known that $\gamma_2 \approx 0,7$ and $\gamma_d \sim_{d\to\infty} \sqrt{d2\pi e}$.

In computational complexity theory, the complexity of an algorithm is its execution cost as a function of the size (n for TSP) of the problem. A polynomial algorithm is preferable to an exponential algorithm beyond a certain size. A problem is said to be in NP when the validity of a solution can be checked by a polynomial algorithm. Combinatorial explosion means that an NP problem cannot automatically be solved by a polynomial algorithm testing every solution. A problem is NP-complete when it is at least as difficult to solve as any other NP problem, that is, every NP problem reduces to it by

a polynomial algorithm. NP-complete problems are therefore key problems. At present, all known algorithms for solving NP-complete problems are exponential, making them unusable fairly quickly. The most famous open question in computer science is whether one can find a polynomial algorithm solving an NP-complete problem. It can be proved that TSP is NP-complete, and no polynomial-time solution algorithm is known. One can nevertheless seek an approximate solution, for example with a generic stochastic algorithm such as simulated annealing, as explained in Chapter 5. Unlike TSP, the other two flagship problems of combinatorial optimization (minimum spanning tree and minimum Euclidean matching) can be solved in polynomial time. For further information, see, for example, the book by Alfred Aho, John Hopcroft, and Jeffrey Ullman [AHU75], the book by Giorgio Ausiello, Pierluigi Crescenzi, Giorgio Gambosi, and Viggo Kann [ACG$^+$99], and Charles Bordenave's lecture notes [Bor14a].

21

Random matrices

Keywords. Random matrix; covariance matrix; spectrum; Catalan numbers; Wigner's theorem; semicircle law; Marchenko-Pastur theorem; quarter-circle law.

Tools. Method of moments; characteristic function; Prohorov's theorem; Courant-Fischer formula; Hoffman-Wielandt inequality; Azuma-Hoeffding inequality; Wasserstein distance; Kolmogorov-Smirnov distance; combinatorics.

Difficulty. ***

This chapter concerns a universality phenomenon in the asymptotic behavior of the spectrum of large random matrices. We begin with a statistical motivation, whose resolution is discussed only in Section 21.5. Let $X_1, \dots, X_n$ be i.i.d. centered random column vectors in $\mathbb{R}^d$ with covariance matrix Σ. The matrix Σ is symmetric $d \times d$ and its eigenvalues are nonnegative. We have $\Sigma_{ij} = \mathbb{E}(X_{ki} X_{kj})$ for all $1 \leqslant i, j \leqslant d$ and every $1 \leqslant k \leqslant n$, or equivalently,

$$\Sigma = \mathbb{E}(X_1 X_1^\top) = \cdots = \mathbb{E}(X_n X_n^\top).$$

The sample covariance matrix $\widehat{\Sigma}_n$ is the following symmetric $d \times d$ matrix:

$$\widehat{\Sigma}_n = \frac{1}{n} \sum_{k=1}^{n} X_k X_k^\top = \frac{1}{n} (X_1 \cdots X_n)(X_1 \cdots X_n)^\top.$$

We have

$$\mathbb{E}(\widehat{\Sigma}_n) = \Sigma.$$

By the law of large numbers applied to the $d \times d$ entries,

$$\widehat{\Sigma}_n \xrightarrow[n \to \infty]{\text{p.s.}} \Sigma.$$

We wish to study the behavior of the random matrix $\widehat{\Sigma}_n$ when the data dimension $d = d_n$ depends on n and tends to ∞. For simplicity, set $\Sigma = I_d$.

This leads to the following model: given a family $(Y_{ij})_{i,j\geqslant 1}$ of i.i.d. real random variables with mean 0 and variance 1, for every integer $n \geqslant 1$ consider the symmetric random matrix $d_n \times d_n$

$$\frac{1}{n} YY^\top$$

where $Y = (Y_{ij})_{1\leqslant i\leqslant d_n, 1\leqslant j\leqslant n}$. If $n \mapsto d_n$ is constant and equal to d, then $\frac{1}{n}YY^\top$ converges almost surely to the matrix $\Sigma = I$. It is natural to seek to understand the behavior of the random matrix $\frac{1}{n}YY^\top$ when d_n depends on n, for example by studying its spectrum. Analysis of the symmetric matrix $YY^\top$ is made difficult by the dependence of its entries. This suggests first analyzing symmetric random matrices whose entries in the upper triangle are independent, leading to Wigner's theorem, the main topic of this chapter. We return to sample covariance matrices in Section 21.5.

21.1 Wigner's theorem

Let $(M_{ij})_{i\geqslant 1, j\geqslant 1}$ be independent real random variables. Consider the infinite symmetric random array obtained by setting $M_{ij} := M_{ji}$ for all $i, j \geqslant 1$:

$$\begin{pmatrix} M_{1,1} & M_{1,2} & \cdots \\ M_{2,1} & M_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

Fix an integer $n \geqslant 1$ and denote by $M := (M_{ij})_{1\leqslant i,j\leqslant n}$ the real symmetric random matrix obtained by extracting the $n\times n$ square in the upper-left corner of the array. Let

$$\lambda_{n,1}, \ldots, \lambda_{n,n}$$

be the eigenvalues of the real symmetric matrix $\frac{1}{\sqrt{n}}M$, ordered so that $\lambda_{n,1} \geqslant \cdots \geqslant \lambda_{n,n}$. We are interested in their counting measure, called the empirical spectral measure, defined by

$$\mu_n := \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_{n,k}}.$$

This is a random probability measure of the same kind as the empirical measure in the Glivenko-Cantelli theorem[1], except that the atoms $\lambda_{n,1}, \ldots, \lambda_{n,n}$ here are dependent real random variables[2]. For every Borel set $B \subset \mathbb{R}$,

[1] If $(Z_n)_{n\geqslant 1}$ are i.i.d. with distribution η, then almost surely the empirical measure $\frac{1}{n}\sum_{k=1}^{n}\delta_{Z_k}$ converges weakly to η as $n \to \infty$, and convergence of the distribution functions is uniform. This follows from the strong law of large numbers and the separability of the topology of weak convergence on $\mathbb{R}$.

[2] Nonlinear functions of M (roots of the characteristic polynomial!).

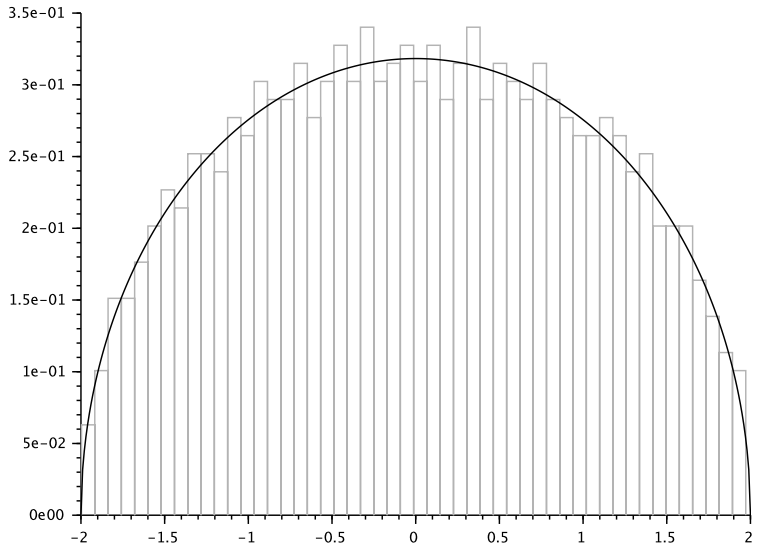


Fig. 21.1. Illustration of Wigner's Theorem 21.1: density of the semicircle law $\mu_1^{\ominus}$ and histogram of the eigenvalues of a symmetric matrix of size $n = 1000$ whose upper-triangular entries, including the diagonal, are i.i.d. real random variables uniform on the interval $[-\sqrt{3/n}, \sqrt{3/n}]$ (the variance is therefore $1/n$).

$$\mu_n(B) = \frac{\mathrm{card}\{1 \leqslant k \leqslant n : \lambda_{n,k} \in B\}}{n}.$$

For every measurable function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\int f \, d\mu_n = \frac{1}{n} \sum_{k=1}^{n} f(\lambda_{n,k}).$$

The distribution function F_n of μ_n is given, for every $x \in \mathbb{R}$, by

$$F_n(x) := \mu_n(]-\infty, x]) = \int \mathbb{1}_{]-\infty,x]} \, d\mu_n = \frac{\mathrm{card}\{1 \leqslant k \leqslant n : \lambda_{n,k} \leqslant x\}}{n}.$$

Wigner's theorem is an analogue of the law of large numbers.

Theorem 21.1 (Wigner)**.** *Suppose $(M_{ij})_{j \geqslant i \geqslant 1}$ are i.i.d. real random variables with finite, nonzero variance σ^2. Let $M = (M_{i,j})_{1 \leqslant i,j \leqslant n}$ with $M_{j,i} := M_{i,j}$. Let $\mu_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{\lambda_{n,k}}$ be the empirical measure of the eigenvalues of $\frac{1}{\sqrt{n}} M$.*

Then almost surely the sequence of probability measures $(\mu_n)_{n\geqslant 1}$ *converges weakly to the semicircle law* $\mu_\sigma^{\ominus}$ *with density*

$$x \mapsto \frac{\sqrt{4\sigma^2 - x^2}}{2\pi\sigma^2} \mathbb{1}_{[-2\sigma,2\sigma]}(x).$$

In other words, almost surely, for every continuous bounded function $f : \mathbb{R} \to \mathbb{R}$*,*

$$\int f \, d\mu_n \xrightarrow[n\to\infty]{} \int f \, d\mu_\sigma^{\ominus}.$$

Recall that weak convergence of probability measures means convergence for continuous bounded test functions. It is equivalent to pointwise convergence of characteristic functions, and also to pointwise convergence of distribution functions at every continuity point of the limit. For random variables, this corresponds to convergence in distribution.

The name "semicircle law" for $\mu_\sigma^{\ominus}$ comes from the fact that its density is the equation of a semicircle (up to a normalization constant equal to half the area of the circle). Since $\mu_\sigma^{\ominus}$ has no atoms, Wigner's theorem states that almost surely, for every interval $I \subset \mathbb{R}$,

$$\lim_{n\to\infty} \mu_n(I) = \mu_\sigma^{\ominus}(I) = \int_{I\cap[-2\sigma,2\sigma]} \frac{\sqrt{4\sigma^2 - x^2}}{2\pi\sigma^2} \, dx.$$

Thus the proportion of eigenvalues of M outside the interval $[-2\sigma\sqrt{n}, 2\sigma\sqrt{n}]$ tends to zero as $n \to \infty$. Wigner's theorem reveals a universality phenomenon: the limiting law $\mu_\sigma^{\ominus}$ depends on the distribution of the entries of the matrix M only through their variance σ^2.

The behavior of the mean of μ_n can be understood by observing that

$$\begin{aligned}\int x \, d\mu_n(x) = \frac{1}{n}\sum_{k=1}^{n} \lambda_{n,k} = \frac{1}{n}\mathrm{Tr}\left(\frac{1}{\sqrt{n}} M\right) \\ = \frac{1}{n^{3/2}} \sum_{1\leqslant i\leqslant n} M_{ii} \xrightarrow[n\to\infty]{\text{p.s.}} 0 = \int x \, d\mu_\sigma^{\ominus}(x)\end{aligned}$$

where almost-sure convergence follows from the law of large numbers. The normalization by $1/\sqrt{n}$ of M can in turn be understood by first observing that

$$\begin{aligned}\int x^2 \, d\mu_n(x) = \frac{1}{n}\sum_{k=1}^{n} \lambda_{n,k}^2 = \frac{1}{n}\mathrm{Tr}\left(\left(\frac{1}{\sqrt{n}} M\right)^2\right) \\ = \frac{1}{n^2} \sum_{1\leqslant i,j\leqslant n} M_{ij}^2 \xrightarrow[n\to\infty]{\text{p.s.}} \sigma^2 + m^2 = m^2 + \int x^2 \, d\mu_\sigma^{\ominus}(x)\end{aligned}$$

where almost-sure convergence again follows from the law of large numbers, and $m = \mathbb{E}(M_{ij})$. The second moment of μ_n converges to that of $\mu_\sigma^{\ominus}$ only when $m = 0$. Moreover, by Markov's inequality, almost surely, for every $r > 0$,

$$\mu_n([-r,r]^c) \leqslant \frac{1}{r^2}\int x^2\,d\mu_n(x) \underset{n\to\infty}{\longrightarrow} \frac{\sigma^2+m^2}{r^2}.$$

Consequently, almost surely, for every $\varepsilon > 0$ there exists a compact set $K_\varepsilon = [-r_\varepsilon, r_\varepsilon]$ such that $\mu_n(K_\varepsilon) \leqslant \varepsilon$ for every $n \geqslant 1$. Thus, almost surely, the sequence $(\mu_n)_{n\geqslant 1}$ is tight, and Prohorov's theorem[3] then states that this sequence is relatively compact for weak convergence, in the sense that if it has a unique weak limit point, then it converges weakly. Despite this reassuring observation, we shall use another approach to prove Wigner's theorem.

The idea is to reduce to the case $m = 0$, then to polynomial test functions f (Section 21.3), and then establish convergence of moments (Section 21.4), drawing inspiration from the method used above for moments of orders 1 and 2. But this first requires reducing the problem to a simpler case!

21.2 Reduction to a simpler case

The mean empirical spectral measure $\mathbb{E}\mu_n$ is the probability measure defined, for every measurable nonnegative (or bounded) function $f : \mathbb{R} \to \mathbb{R}$, by

$$\int f\,d\mathbb{E}\mu_n := \mathbb{E}\int f\,d\mu_n := \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} f(\lambda_{n,k})\right).$$

The purpose of this section is to reduce Wigner's Theorem 21.1 to a simplified version, Theorem 21.2 below, concerning the sequence of probability measures $(\mathbb{E}\mu_n)_{n\geqslant 1}$ when the matrix M has zero diagonal and bounded off-diagonal entries. Theorem 21.2 is proved later, in Section 21.4, using the method of moments from Section 21.3.

Theorem 21.2 (Simplified Wigner theorem)**.** *Let $M = (M_{ij})_{1\leqslant i,j\leqslant n}$ be a real symmetric random matrix with zero diagonal $M_{11} = \cdots = M_{nn} = 0$, such that the random variables $(M_{ij})_{1\leqslant i<j\leqslant n}$ are independent and identically distributed, centered, with unit variance and compact support $[-C, C]$, $C > 0$. Let $\mu_n = \frac{1}{n}\sum_{k=1}^{n}\delta_{\lambda_{n,k}}$ be the empirical measure of the eigenvalues of $\frac{1}{\sqrt{n}}M$. Then, for every continuous bounded function $f : \mathbb{R} \to \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{E}\int f\,d\mu_n = \int f\,d\mu_1^{\ominus}.$$

Theorem 21.2 is proved in Section 21.4.

[3] A set of probability measures is tight if and only if it is sequentially relatively compact for the topology of weak convergence. In particular, if a sequence of probability measures is tight and has a unique weak limit point, then it converges weakly.

Reduction to the centered case

In this subsection, we show that it suffices to establish Wigner's Theorem 21.1 for $M - \mathbb{E}(M)$.

Orthogonality of the eigenspaces of symmetric matrices leads to the Courant-Fischer min-max variational formulas for eigenvalues. More precisely, if $A \in \mathcal{M}_n(\mathbb{C})$ is Hermitian with eigenvalues $\lambda_1(A) \geqslant \cdots \geqslant \lambda_n(A)$, then for every $1 \leqslant k \leqslant n$, writing $\mathcal{G}_k$ for the set of linear subspaces of $\mathbb{C}^n$ with dimension k, we have

$$\lambda_k(A) = \max_{V \in \mathcal{G}_k} \min_{\substack{x \in V \\ \|x\|_2 = 1}} \langle Ax, x \rangle = \min_{V \in \mathcal{G}_{n-k+1}} \max_{\substack{x \in V \\ \|x\|_2 = 1}} \langle Ax, x \rangle.$$

The cases $k = 1$ and $k = n$ are familiar. These formulas are a standard result of linear algebra, as is the next lemma, which we state without proof. References are given in Section 21.5. Write F_A for the distribution function of the spectral measure $\mu_A := \frac{1}{n} \sum_{k=1}^n \delta_{\lambda_k(A)}$. For every $t \in \mathbb{R}$,

$$F_A(t) := \mu_A(]-\infty, t]) = \int \mathbb{1}_{]-\infty,t]} \, d\mu_A = \frac{\mathrm{card}\{1 \leqslant k \leqslant n : \lambda_k(A) \leqslant t\}}{n}.$$

Lemma 21.3 (Interlacing for additive perturbations of low rank)**.** *If $A, B \in \mathcal{M}_n(\mathbb{C})$ are Hermitian with eigenvalues $\lambda_1(A) \geqslant \cdots \geqslant \lambda_n(A)$ and $\lambda_1(B) \geqslant \cdots \geqslant \lambda_n(B)$, and if $r := \mathrm{rang}(B - A)$, then*

$$\lambda_{k+r}(A) \leqslant \lambda_k(B) \leqslant \lambda_{k-r}(A)$$

for every $1 \leqslant k \leqslant n$, with the convention $\lambda_k(A) = +\infty$ if $k < 1$ and $\lambda_k(A) = -\infty$ if $k > n$. In particular,

$$d_{\mathrm{KS}}(\mu_A, \mu_B) := \|F_A - F_B\|_\infty := \sup_{t \in \mathbb{R}} |F_A(t) - F_B(t)| \leqslant \frac{\mathrm{rang}(A - B)}{n}.$$

The notation d_{KS} denotes the Kolmogorov-Smirnov distance between the discrete probability distributions μ_A and μ_B. It is well known that convergence in Kolmogorov-Smirnov distance implies weak convergence.

Now, since the random variables $(M_{ij})_{1 \leqslant i,j \leqslant n}$ have the same mean $m := \mathbb{E}(M_{ij})$, we have $\mathrm{rang}(\mathbb{E}(M)) = \mathrm{rang}(m 1_n) \leqslant 1$ and therefore

$$d_{\mathrm{KS}}\Big(\mu_{\frac{1}{\sqrt{n}} M}, \mu_{\frac{1}{\sqrt{n}}(M - \mathbb{E}(M))}\Big) \leqslant \frac{1}{n} \underset{n \to \infty}{\longrightarrow} 0.$$

This reduces Wigner's Theorem 21.1 to the case where M has centered entries.

Reduction to the centered, standardized, bounded case

In this subsection, we show that it suffices to establish Wigner's Theorem 21.1 in the special case where the entries of M are centered, standardized,

and bounded. We proceed by truncation, recentering, and renormalization. Let M be centered, $C > 0$ real, and

$$(M_C)_{ij} := \varphi_C(M_{ij}\mathbb{1}_{\{|M_{ij}|\leqslant C\}}) \quad \text{où} \quad \varphi_C(Z) := \frac{Z - \mathbb{E}(Z)}{\sqrt{\mathbb{E}((Z - \mathbb{E}(Z))^2)}}.$$

Write $\mathcal{P}_2$ for the set of distributions on $\mathbb{R}$ with finite moment of order 2. The Wasserstein distance W_2 on $\mathcal{P}_2$ is defined, for all $\nu_1, \nu_2 \in \mathcal{P}_2$, by

$$W_2(\nu_1, \nu_2)^2 := \inf_{\substack{(X_1, X_2) \\ X_1 \sim \nu_1, X_2 \sim \nu_2}} \mathbb{E}(|X_1 - X_2|^2)$$

where the infimum runs over all pairs of random variables (X_1, X_2) with marginal distributions ν_1 and ν_2. Convergence in Wasserstein distance W_2 implies weak convergence[4]. For discrete distributions of the form

$$\nu_1 = \frac{1}{n}\sum_{k=1}^{n} \delta_{a_k} \quad \text{et} \quad \nu_2 = \frac{1}{n}\sum_{k=1}^{n} \delta_{b_k}$$

with $a_1 \leqslant \cdots \leqslant a_n$ and $b_1 \leqslant \cdots \leqslant b_n$, we find

$$W_2(\nu_1, \nu_2)^2 = \frac{1}{n}\sum_{k=1}^{n} (a_i - b_i)^2.$$

Recall that $\mathcal{M}_n(\mathbb{C})$ is a Hilbert space for the inner product $\langle A, B\rangle := \mathrm{Tr}(AB^*)$. The associated norm is called the Hilbert-Schmidt norm[5]

$$\|A\|_{\mathrm{HS}}^2 = \mathrm{Tr}(AA^*) = \sum_{i=1}^{n}\sum_{j=1}^{n} |A_{ij}|^2.$$

The next lemma, stated without proof, is a standard result of linear algebra. References are given in Section 21.5.

Lemma 21.4 (Hoffman-Wielandt inequality)**.** *If $A, B \in \mathcal{M}_n(\mathbb{C})$ are Hermitian with spectra $\lambda_1(A) \geqslant \cdots \geqslant \lambda_n(A)$ and $\lambda_1(B) \geqslant \cdots \geqslant \lambda_n(B)$, then*

$$nW_2(\mu_A, \mu_B)^2 = \sum_{k=1}^{n} (\lambda_k(A) - \lambda_k(B))^2 \leqslant \sum_{i=1}^{n}\sum_{j=1}^{n} |A_{ij} - B_{ij}|^2 = \|A - B\|_{\mathrm{HS}}^2.$$

By the Hoffman-Wielandt lemma 21.4 and the strong law of large numbers,

$$W_2(\mu_{\frac{1}{\sqrt{n}}M}, \mu_{\frac{1}{\sqrt{n}}M_C})^2$$

[4] Convergence in W_2 is equivalent to weak convergence together with convergence of moments of orders 1 and 2. In particular, if either of the first two moments fails to converge to that of the limit, there is no convergence in W_2.

[5] Also called the Schur norm or Frobenius norm.

$$\leqslant \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(M_{ij}-\varphi_C(M_{ij}))^2 \xrightarrow[n\to\infty]{\text{p.s.}} \mathbb{E}((M_{12}-\varphi_C(M_{12}))^2).$$

By the dominated convergence theorem or the monotone convergence theorem, the right-hand side can be made arbitrarily close to 0 by choosing C sufficiently large. Consequently, it suffices to establish Wigner's Theorem 21.1 for M with centered, standardized, bounded entries.

Removing the diagonal

The method used for truncation likewise allows the diagonal to be removed. Indeed, if M has centered, standardized entries bounded by C, and if $D := \operatorname{diag}(M_{ii} : 1 \leqslant i \leqslant n)$, then

$$W_2(\mu_{\frac{1}{\sqrt{n}}M}, \mu_{\frac{1}{\sqrt{n}}(M-D)})^2 \leqslant \frac{1}{n^2}\sum_{i=1}^{n}|M_{ii}|^2 \xrightarrow[n\to\infty]{\text{p.s.}} 0$$

by the law of large numbers. It therefore suffices to establish Wigner's Theorem 21.1 for M with centered, standardized, bounded entries and zero diagonal.

Reduction to the mean empirical spectral measure

In this subsection, we show that it suffices to establish Wigner's Theorem 21.1 for the mean empirical spectral measure

$$\mathbb{E}\mu_{\frac{1}{\sqrt{n}}(M_C-\mathbb{E}(M_C)-D_C)},$$

for every real $C > 0$. The next lemma, similar in nature to Chebyshev's inequality, is a reformulation of Lemma 20.3.

Lemma 21.5 (Azuma-Hoeffding concentration inequality)**.** *Consider independent random vectors $X_1, \ldots, X_n$, not necessarily of the same dimension. If $G(X_1, \ldots, X_n)$ is an integrable real random variable that is a measurable function of these vectors, then for every $t \geqslant 0$,*

$$\mathbb{P}(|G(X_1,\ldots,X_n)-\mathbb{E}(G(X_1,\ldots,X_n))| \geqslant t) \leqslant 2\exp\left(-\frac{2t^2}{c_1^2+\cdots+c_n^2}\right)$$

where $c_k := \sup_{(x,x')\in\mathcal{D}_k}|G(x)-G(x')|$ and $\mathcal{D}_k := \{(x,x') : x_i = x_i' \text{ si } i \neq k\}$.

Let us see how to use this lemma. Let $f : \mathbb{R} \to \mathbb{R}$ be of class $\mathcal{C}^1$ with compact support, and let $A, B \in \mathcal{M}_n(\mathbb{C})$ be Hermitian. Lemma 21.3 gives

$$\|F_A - F_B\|_\infty \leqslant \frac{r}{n} \quad \text{où} \quad r := \operatorname{rang}(A-B),$$

and integration by parts yields

$$\left|\int f\,d\mu_A - \int f\,d\mu_B\right| = \left|\int_{-\infty}^{+\infty} f'(t)(F_A(t)-F_B(t))\,dt\right| \leqslant \frac{r}{n}\int_{-\infty}^{+\infty}|f'(t)|\,dt := \frac{r}{n}\|f'\|_1.$$

Denote by $A(x_1,\dots,x_n) \in \mathcal{M}_n(\mathbb{C})$ the Hermitian matrix whose upper-triangular rows are $x_1,\dots,x_n$ (vectors in $\mathbb{C}^n,\mathbb{C}^{n-1},\dots,\mathbb{C}$), and set

$$G(x_1,\dots,x_n) = \int f\,d\mu_{A(x_1,\dots,x_n)}.$$

Writing $y := (x_1,\dots,x_n)$ and $y' := (x_1,\dots,x_{k-1},x'_k,x_{k+1},\dots,x_n)$ gives

$$\operatorname{rang}\{A(y)-A(y')\} \leqslant 2 \quad \text{et donc} \quad |G(y)-G(y')| \leqslant \frac{2}{n}\|f'\|_1.$$

Now $H := \frac{1}{\sqrt{n}}(M_C - \mathbb{E}(M_C) - D_C)$ is a random Hermitian matrix with values in $\mathcal{M}_n(\mathbb{C})$. Its upper-triangular rows $(H_{ij})_{1\leqslant i\leqslant j\leqslant n}$ are independent (random vectors in $\mathbb{C}^n,\mathbb{C}^{n-1},\dots,\mathbb{C}$). By Lemma 21.5, for every compactly supported function f of class $\mathcal{C}^1$ and every real $t \geqslant 0$,

$$\mathbb{P}\left(\left|\int f\,d\mu_H - \mathbb{E}\int f\,d\mu_H\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{nt^2}{8\|f'\|_1^2}\right).$$

This inequality expresses the strong concentration of μ_H around its expectation $\mathbb{E}\mu_H$. This is why fluctuations in simulations are so small once the dimension exceeds a few dozen. We have

$$\int f\,d\mu_H - \mathbb{E}\int f\,d\mu_H \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

To exploit the rapid convergence, set $t = n^{-1/2+\varepsilon}$ with fixed small $\varepsilon > 0$, so that the exponential bound on the right-hand side of the inequality is summable in n. The Borel-Cantelli lemma then implies that

$$\int f\,d\mu_H - \mathbb{E}\int f\,d\mu_H \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

This reduces the study of almost-sure weak convergence of μ_H to that of weak convergence of $\mathbb{E}\mu_H$. Thus Wigner's Theorem 21.1 follows from the simplified Wigner theorem 21.2.

21.3 Convergence of moments and weak convergence

This section is devoted to the *method of moments*. Let $\mathcal{P}$ be the set of probability distributions ν on $\mathbb{R}$ such that $\mathbb{R}[X] \subset L^1(\nu)$, equipped with the equivalence relation $\nu_1 \sim \nu_2$ if and only if $\int P\,d\nu_1 = \int P\,d\nu_2$ for every $P \in \mathbb{R}[X]$,

that is, ν_1 and ν_2 have the same moments. We say that $\nu \in \mathcal{P}$ is *determined by its moments* when its equivalence class is a singleton.

Theorem 21.6 (Convergence of moments and weak convergence)**.** *If $\nu, \nu_1, \nu_2, \dots$ are elements of $\mathcal{P}$ satisfying, for every polynomial $P \in \mathbb{R}[X]$,*

$$\lim_{n\to\infty} \int P \, d\nu_n = \int P \, d\nu,$$

and if ν is determined by its moments, then $(\nu_n)_{n\geqslant 1}$ converges weakly to ν, that is, for every continuous bounded $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int f \, d\nu_n = \int f \, d\nu.$$

Proof. By assumption, for every $P \in \mathbb{R}[X]$,

$$C_P := \sup_{n\geqslant 1} \int P \, d\nu_n < \infty.$$

Consequently, by Markov's inequality, for every real $R > 0$,

$$\nu_n([-R, R]^c) \leqslant \frac{C_{X^2}}{R^2}$$

and therefore $(\nu_n)_{n\geqslant 1}$ is tight. By Prohorov's theorem, it therefore suffices to show that if $(\nu_{n_k})_{k\geqslant 1}$ converges weakly to ν' then $\nu' = \nu$. Fix $P \in \mathbb{R}[X]$ and a real $R > 0$. Let $\varphi_R : \mathbb{R} \to [0, 1]$ be a continuous function such that $\mathbb{1}_{[-R,R]} \leqslant \varphi_R \leqslant \mathbb{1}_{[-R-1,R+1]}$. We have the decomposition

$$\int P \, d\nu_{n_k} = \int \varphi_R P \, d\nu_{n_k} + \int (1 - \varphi_R) P \, d\nu_{n_k}.$$

Now $\varphi_R P$ is continuous and bounded and $(\nu_{n_k})_{k\geqslant 1} \underset{n\to\infty}{\longrightarrow} \nu'$ weakly, hence

$$\lim_{k\to\infty} \int \varphi_R P \, d\nu_{n_k} = \int \varphi_R P \, d\nu'.$$

Moreover, by the Schwarz and Markov inequalities, we have

$$\left| \int (1 - \varphi_R) P \, d\nu_{n_k} \right|^2 \leqslant \nu_{n_k}([-R, R]^c) \int P^2 \, d\nu_{n_k} \leqslant \frac{C_{X^2} C_{P^2}}{R^2}.$$

We therefore obtain

$$\limsup_{k\to\infty} \left| \int P \, d\nu_{n_k} - \int \varphi_R P \, d\nu' \right| = \mathcal{O}\left(\frac{1}{R^2}\right) = o_{R\to\infty}(1).$$

On the other hand, we know that

$$\lim_{k\to\infty} \int P \, d\nu_{n_k} = \int P \, d\nu$$

and finally obtain

$$\lim_{R\to\infty} \int \varphi_R P \, d\nu' = \int P \, d\nu.$$

Applying this to P^2, which is nonnegative, first gives by monotone convergence that $P \in L^2(\nu') \subset L^1(\nu')$, and then by dominated convergence that

$$\int P \, d\nu' = \int P \, d\nu.$$

Since P is arbitrary and ν is determined by its moments, we have $\nu = \nu'$. □

A theorem of Weierstrass states that, for every compact set $K \subset \mathbb{R}$, the polynomials $\mathbb{R}[X]$ restricted to K are dense in $\mathcal{C}(K, \mathbb{R}, \|\cdot\|_\infty)$. Thus every compactly supported probability distribution is determined by its moments *among compactly supported distributions*. The next theorem goes further: it implies that every compactly supported distribution is determined by its moments among all distributions with finite moments of every order. This applies in particular to the semicircle law $\mu_\sigma^{\ominus}$.

Theorem 21.7 (Analyticity of the Fourier transform and moments)**.** *Let $\mu \in \mathcal{P}$. Set $\varphi(t) = \int e^{itx} \, d\mu(x)$ and $\kappa_n = \int x^n \, d\mu(x)$. The following assertions are equivalent:*

1. *φ is analytic in a neighborhood of 0,*
2. *φ is analytic on $\mathbb{R}$,*
3. *$\overline{\lim}_{n\to\infty} \left(\frac{1}{n!}|\kappa_n|\right)^{\frac{1}{n}} < \infty$.*

If these conditions hold, then μ is determined by its moments.
In particular, a compactly supported distribution is determined by its moments.

Proof. For every n, we have $\int |x|^n \, d\mu < \infty$ and therefore φ is n times differentiable on $\mathbb{R}$. Moreover, $\varphi^{(n)}$ is continuous on $\mathbb{R}$ and, for every $t \in \mathbb{R}$,

$$\varphi^{(n)}(t) = \int_{\mathbb{R}} (ix)^n e^{itx} \, d\mu(x).$$

In particular, $\varphi^{(n)}(0) = i^n \kappa_n$, and the Taylor series of φ at 0 is determined by the sequence $(\kappa_n)_{n\geqslant 1}$. The radius of convergence r of the power series $\sum_n a_n z^n$ associated with the sequence of complex numbers $(a_n)_{n\geqslant 0}$ is given by Hadamard's formula $r^{-1} = \overline{\lim}_n |a_n|^{\frac{1}{n}}$. Thus, 1⇔ 3 (take $a_n = i^n \kappa_n / n!$). Moreover, since for every $n \in \mathbb{N}$, $s, t \in \mathbb{R}$,

$$\left| e^{isx} \left(e^{itx} - 1 - \frac{itx}{1!} - \cdots - \frac{(itx)^{n-1}}{(n-1)!} \right) \right| \leqslant \frac{|tx|^n}{n!},$$

we have, for every even $n \in \mathbb{N}$ and all $s, t \in \mathbb{R}$,

$$\left|\varphi(s+t) - \varphi(s) - \frac{t}{1!}\varphi'(s) - \cdots - \frac{t^{n-1}}{(n-1)!}\varphi^{(n-1)}(s)\right| \leqslant \kappa_n \frac{|t|^n}{n!},$$

which shows that $3 \Rightarrow 2$. Since $2 \Rightarrow 1$, we indeed have $1 \Leftrightarrow 2 \Leftrightarrow 3$. If these properties hold, the preceding arguments give some $r > 0$ such that φ admits a power series expansion at every $x \in \mathbb{R}$ with radius of convergence $\geqslant r$. Step by step, we obtain that φ is determined by its derivatives at zero.

Finally, if μ has compact support, say contained in $[-C, C]$ for some constant $C > 0$, then $|\kappa_n| \leqslant C^n$ for every $n \geqslant 1$, and since Stirling's formula $n! \sim \sqrt{2\pi n}(n/e)^n$ gives $(1/n!)^{1/n} = \mathcal{O}(1/n)$, we obtain $\lim_{n\to\infty}(\frac{1}{n!}|\kappa_n|)^{1/n} = 0$, so μ is determined by its moments. □

Theorem 21.8 (Moments of the semicircle law)**.** *The semicircle law $\mu_\sigma^{\ominus}$ is determined by its moments. Moreover, for every integer $r \geqslant 0$,*

$$\int x^{2r+1}\, d\mu_\sigma^{\ominus}(r) = 0 \quad \textit{et} \quad \int x^{2r}\, d\mu_\sigma^{\ominus}(r) = \sigma^{2r}\frac{1}{r+1}\binom{2r}{r}.$$

In particular, $\mu_\sigma^{\ominus}$ has mean 0 and variance σ^2, and the even moments of the standard semicircle law $\mu_1^{\ominus}$ are the Catalan numbers.

Proof. The law $\mu_\sigma^{\ominus}$ is determined by its moments because its support $[-\sigma, \sigma]$ is compact (Lemma 21.7). The odd moments vanish because $\mu_\sigma^{\ominus}$ is symmetric. To compute the even moments, scaling and symmetry give

$$\int x^{2r}\, d\mu_\sigma^{\ominus}(x) = \frac{\sigma^{2r}}{\pi}\int_0^2 x^{2r}\sqrt{4-x^2}\, dx.$$

The change of variable $x = 2\cos(u)$ and integration by parts yield

$$\int_0^2 x^{2r}\sqrt{4-x^2}\, dx = \int_0^{\frac{\pi}{2}} 4^{r+1}\cos^{2r}(u)\sin^2(u)\, du = \frac{4^{r+1}}{2r+1}\int_0^{\frac{\pi}{2}} \cos^{2(r+1)}(u)\, du.$$

This last expression is a classical Wallis integral. □

21.4 Proof of the simplified Wigner theorem

By Theorems 21.6 and 21.8, the proof of the simplified Wigner theorem 21.2 reduces to establishing Theorem 21.9 below.

Theorem 21.9 (Convergence of moments)**.** *Under the assumptions of the simplified Wigner theorem 21.2, for every integer $r \geqslant 1$,*

$$\mathbb{E}\int x^r d\mu_n \underset{n\to\infty}{\longrightarrow} \int x^r\, d\mu_\sigma^{\ominus}.$$

Proof. The moments of $\mu_1^{\ominus}$ are given by Theorem 21.8. For every integer $r \geqslant 1$, the moment of order r of $\mathbb{E}\mu_n$ can be written (with $i_{r+1} := i_1$)

$$\begin{aligned}\mathbb{E}\int x^r \, d\mu_n(x) &= \frac{1}{n}\sum_{k=1}^{n} \lambda_{n,k}^r \\ &= \frac{1}{n^{1+r/2}}\mathrm{Tr}(M^r) \\ &= \frac{1}{n^{1+r/2}} \sum_{1\leqslant i_1,\dots,i_r\leqslant n} \mathbb{E}(M_{i_1 i_2}\cdots M_{i_r i_{r+1}}).\end{aligned}$$

Since the diagonal entries of M vanish, we have

$$\mathbb{E}\int x \, d\mu_n = \frac{1}{n^{3/2}}\sum_{i=1}^{n} \mathbb{E}M_{ii} = 0 = \int x \, d\mu_1^{\ominus}(x)$$

and since M has centered, standardized off-diagonal entries, we have

$$\mathbb{E}\int x^2 \, d\mu_n = \frac{1}{n^2}\sum_{1\leqslant i,j\leqslant n} \mathbb{E}(M_{ij}^2) = \frac{n^2-n}{n^2} \to 1 = \int x^2 \, d\mu_1^{\ominus}(x).$$

The moment of order 3 is slightly subtler. We have

$$\mathbb{E}\int x^3 \, d\mu_n = \frac{1}{n^{1+3/2}}\sum_{1\leqslant i,j,k\leqslant n} \mathbb{E}(M_{ij}M_{jk}M_{ki}).$$

If two elements among $\{\{i,j\},\{j,k\},\{k,i\}\}$ are distinct, then necessarily $\mathbb{E}(M_{ij}M_{jk}M_{ki}) = 0$ by independence and centering. Otherwise, $i = k$ or $i = j$ or $k = j$, which leads to $\mathbb{E}(M_{ij}M_{jk}M_{ki}) = 0$ because the diagonal of M vanishes. Thus the moment of order 3 of μ_n equals zero, as does the moment of order 3 of $\mu_1^{\ominus}$.

For the moment of order 4, one could proceed similarly using

$$\mathbb{E}\int x^4 \, d\mu_n = \frac{1}{n^{1+4/2}}\sum_{1\leqslant i,j,k,l\leqslant n} \mathbb{E}(M_{ij}M_{jk}M_{kl}M_{li}).$$

But a new phenomenon occurs here: the case $i = k$ and $j = l$ with $i \neq j$ gives $\mathbb{E}(M_{ij}M_{jk}M_{kl}M_{li}) = \mathbb{E}(M_{ij}^4)$, which is not a function of the first two moments of the entries of M. There is nevertheless no contradiction: these terms make no asymptotic contribution as n tends to ∞ because their number, $n(n-1)$, is negligible compared with $n^{1+4/2} = n^3$.

More generally, for every integer $r > 1$ we have

$$\mathbb{E}\int x^r \, d\mu_n(x) = \frac{1}{n^{1+r/2}} \sum_{1\leqslant i_1,\dots,i_r\leqslant n} \mathbb{E}(M_{i_1 i_2}\cdots M_{i_r i_{r+1}}).$$

We may assume that $i_k \neq i_{k+1}$ for every $1 \leqslant k \leqslant r-1$ because the diagonal of M vanishes. To each r-tuple of indices of this kind $i_1, \dots, i_r$, associate a directed multigraph; see Figures 21.2 and 21.3. Its vertices are the distinct values of these indices, and we denote their number by t. The edges of the directed multigraph are the links (i_k, i_{k+1}) with $1 \leqslant k \leqslant r-1$. They may have multiplicity and are oriented. Moreover, this directed multigraph is cyclic, of length $r+1$. Writing t for the number of distinct vertices, we call it a directed $G(r,t)$ multigraph. Two directed $G(r,t)$ multigraphs are equivalent when one can be obtained from the other by permuting the indices. Equivalent directed $G(r,t)$ multigraphs give the same value of $\mathbb{E}(M_{i_1 i_2} \cdots M_{i_r i_{r+1}})$, denoted by $\mathbb{E}(M_G)$. There are

$$n(n-1)\cdots(n-t+1)$$

directed $G(r,t)$ multigraphs in each equivalence class (the number of arrangements of t objects chosen from n). Each equivalence class contains a representative for which the t distinct values of the indices $i_1, \dots, i_r$ are successively $1, \dots, t$. To compute the contributions, we distinguish three types of directed $G(r,t)$ multigraphs, detailed below. The type is constant on each class (it is a property of the class).

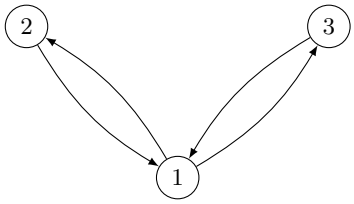


Fig. 21.2. Directed multigraph associated with $\mathbb{E}(M_{12}M_{21}M_{13}M_{31})$. We have $r = 4$, $t = 3$, $i_1 = i_3 = 1$, $i_2 = 2$, $i_4 = 3$. The successive edges are 12, 21, 13, 31. Each edge that is present, as well as its reverse, appears exactly once. The undirected skeleton is the tree $2 \leftrightarrow 1 \leftrightarrow 3$. This directed multigraph is therefore of type 1.

- type 1: each edge that is present, as well as its reverse, appears exactly once, and the undirected skeleton obtained by erasing edge orientations and multiplicities is a tree (that is, it has no cycles). For example, $\mathbb{E}(M_{12}M_{21}M_{13}M_{31}) = \mathbb{E}(M_{12}^2)\mathbb{E}(M_{13}^2) = 1$; see Figure 21.2. A longer example is given in Figure 21.3;
- type 2: at least one edge appears exactly once and its reverse does not appear, for example $\mathbb{E}(M_{12}M_{23}M_{31}) = \mathbb{E}(M_{12})\mathbb{E}(M_{23})\mathbb{E}(M_{31}) = 0$; see Figure 21.3;
- type 3: neither type 1 nor type 2. This is the case, for example, for $\mathbb{E}(M_{12}M_{21}M_{12}M_{21}) = \mathbb{E}(M_{12}^4) > 0$, since the edge 12 (as well as the edge 21) appears exactly twice. Another example is $\mathbb{E}(M_{12}M_{21}M_{13}M_{32}M_{23}M_{31}) = \mathbb{E}(M_{12}^2)\mathbb{E}(M_{13}^2)\mathbb{E}(M_{23}^2) = 1$, since its

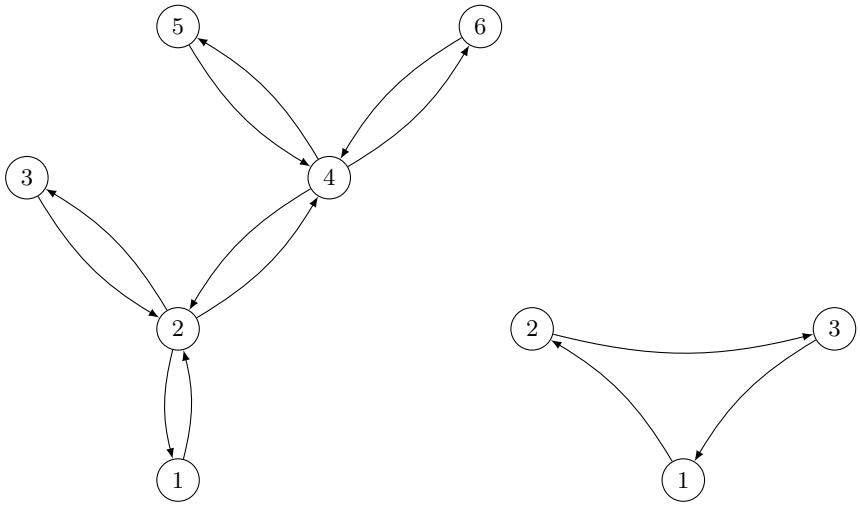
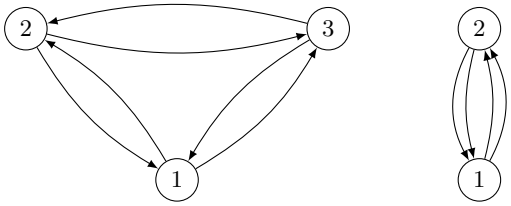


Fig. 21.3. At the upper left, the directed multigraph associated with the long example $\mathbb{E}(M_{12}M_{23}M_{32}M_{24}M_{45}M_{54}M_{46}M_{64}M_{42}M_{21})$, which is of type 1. At the upper right, the directed multigraph associated with $\mathbb{E}(M_{12}M_{23}M_{31})$, which is of type 2. At the lower left, the directed multigraph associated with $\mathbb{E}(M_{12}M_{21}M_{13}M_{32}M_{23}M_{31})$, and at the lower right, the directed multigraph associated with $\mathbb{E}(M_{12}M_{21}M_{12}M_{21})$, both of type 3.



undirected skeleton is the cycle $1 \leftrightarrow 2 \leftrightarrow 3 \leftrightarrow 1$. These two examples are illustrated in Figure 21.3.

We now reach the final stage:

- *Directed multigraphs of type 2 contribute zero.* Indeed, in this case $\mathbb{E}(M_G) = 0$ by independence and centering;
- *Directed multigraphs of type 3 have an asymptotically vanishing contribution.* Indeed, if G is of type 3, it contains at least three edges with the same endpoints or a cycle of edges with different endpoints. In either case this implies $2t \leqslant r+1$, yielding the bound $n(n-1)\cdots(n-t+1) \leqslant n^t \leqslant n^{1/2+r/2}$. Moreover, the number of equivalence classes of directed $G(r,t)$ multigraphs is at most $t^r = \mathcal{O}(1)$. Since the entries of M are bounded with values in $[-C,C]$, we also have $\mathbb{E}(M_G) \leqslant C^r = \mathcal{O}(1)$, and finally

$$\sum_{\text{cl. t. } 3} \frac{n(n-1)\cdots(n-t+1)}{n^{1+r/2}}\mathbb{E}(M_G) = \mathcal{O}(n^{-1/2}) = o_{n\to\infty}(1);$$

- *Contribution of directed multigraphs of type 1.* If G is a directed multigraph of type 1, then $\mathbb{E}(M_G) = 1$ by independence, since the off-diagonal entries of M have variance 1. This reduces the problem to determining the number N_1 of equivalence classes of directed $G(r,t)$ multigraphs of type 1. If r is odd, then $N_1 = 0$. If r is even, say $r = 2s$, then $t = 1 + r/2 = 1 + s$, because a tree always has 1 more vertices than edges. Thus classes of type 1 are in bijection with rooted plane trees with s edges[6] (and $1+s$ vertices), with parenthesizations, with excursions of the simple random walk, and so on; see Theorem 2.5. Their number is therefore $N_1 = \frac{1}{s+1}\binom{2s}{s}$ (a Catalan number), and

$$\sum_{\text{cl. t. } 1} \frac{n(n-1)\cdots(n-t+1)}{n^{1+r/2}}\mathbb{E}(M_G)$$
$$= \frac{n}{n}\cdots\frac{n-s+1}{n}\frac{1}{s+1}\binom{2s}{s}$$
$$\xrightarrow[n\to\infty]{} \frac{1}{1+s}\binom{2s}{s}.$$

□

21.5 Further reading

Return to the sample covariance matrix model from the beginning of the chapter, with $(Y_{ij})_{i,j\geqslant 1}$ i.i.d. real random variables of mean 0 and variance 1, and the rectangular matrix $Y = (Y_{ij})_{1\leqslant i\leqslant d_n, 1\leqslant j\leqslant n}$. Let $\lambda_{n,1} \leqslant \cdots \leqslant \lambda_{n,d_n}$ be the spectrum of the symmetric positive semidefinite matrix $\frac{1}{n}YY^\top$, arranged in increasing order. Suppose $\lim_{n\to\infty} d_n/n = \rho$ with $0 < \rho < \infty$. Then the Marchenko-Pastur theorem, which can be proved by the same method as Wigner's theorem, although the details are somewhat more involved, states that almost surely the empirical spectral measure $\mu_n := \frac{1}{d_n}\sum_{k=1}^{d_n}\delta_{\lambda_{n,k}}$ converges weakly as $n \to \infty$ to the Marchenko-Pastur law

$$\mu_\rho^{\oplus} := q\delta_0 + \frac{\sqrt{(b-x)(x-a)}}{2\pi\rho x}\,\mathbb{1}_{[a,b]}(x)dx.$$

with

$$q := \max(0, (1-\rho^{-1})) \quad \text{et} \quad a := (1-\sqrt{\rho})^2 \quad \text{et} \quad b := (1+\sqrt{\rho})^2.$$

[6] These should not be confused with the rooted plane binary trees of Chapter 4, which are also counted by Catalan numbers!

An illustration of the Marchenko-Pastur theorem appears in Figure 21.4. The moments of $\mu_\rho^\oplus$ are given, for every $r \geqslant 1$, by

$$\int x^r \, d\mu_\rho^\oplus(x) = \sum_{k=0}^{r-1} \frac{\rho^k}{k+1} \binom{r}{k} \binom{r-1}{k}.$$

In particular, the law $\mu_\rho^\oplus$ has mean 1 and variance ρ. The law $\mu_\rho^\oplus$ is a mixture of a Dirac mass at 0 and a distribution with a density with respect to Lebesgue measure. The atom at 0 comes from the fact that the matrix need not have full rank. It disappears when $\rho \leqslant 1$. When $\rho = 1$, the matrix Y is in a sense asymptotically square. In this case, $a = 0$ and $b = 4$. By a change of variable, almost surely the counting measure of the spectrum[7] of $(\frac{1}{n} YY^\top)^{1/2}$ converges weakly as $n \to \infty$ to the quarter-circle law with density $x \mapsto \frac{1}{\pi}\sqrt{4 - x^2}\, \mathbb{1}_{[0,2]}(x)$.

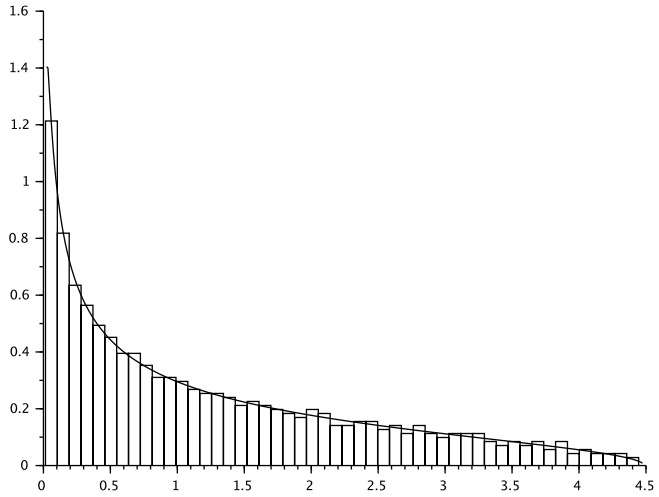


Fig. 21.4. Histogram of the nonzero eigenvalues of a sample covariance matrix with $n = 800$ and $d = 1000$, and the Marchenko-Pastur law.

The study of random matrices is a vast research area. The earliest work is mainly due to John Wishart (1920s) on sample covariance matrices of Gaussian samples, John von Neumann and Herman Goldstine (1940s) on numerical

[7] These are also the *singular values* of the rectangular matrix $\frac{1}{\sqrt{n}} Y$.

matrix analysis, and Eugene Wigner, Freeman Dyson, and Madan Lal Mehta (1950s–1960s) on energy levels of atomic nuclei in mathematical physics. The Marchenko-Pastur theorem was obtained by Vladimir Marchenko and Leonid Pastur in the 1960s. The semicircle law is an analogue of the Gaussian law in free probability theory, initiated by Dan-Virgil Voiculescu in connection with random matrices.

Wigner's Theorem 21.1 was obtained by Wigner in the 1950s under more restrictive assumptions. Its most complete version dates from the late 1970s. Weak convergence of the counting measure μ_n does not prevent an asymptotically negligible fraction of atoms from diverging as $n \to \infty$. However, the Bai-Yin theorem states that if $\mathbb{E}(M_{11}^4) < \infty$ then almost surely the support of μ_n converges to the support $[-2\sigma, 2\sigma]$ of the limit $\mu_\sigma^{\ominus}$, that is, almost surely,

$$\lim_{n\to\infty} \lambda_n = -2\sigma \quad \text{et} \quad \lim_{n\to\infty} \lambda_1 = 2\sigma.$$

The combinatorial part of the proof of Wigner's theorem is detailed, for example, in the book by Zhidong Bai and Jack Silverstein [BS10], and in the book by Greg Anderson, Alice Guionnet, and Ofer Zeitouni [AGZ10]. Wigner's theorem can also be obtained by an analytic method using the Cauchy-Stieltjes transform, in which the semicircle law appears as the solution of a fixed-point equation. Wigner's theorem is also presented in Mireille Capitaine's lecture notes [Cap12] and those of Charles Bordenave [Bor14b]. Section 21.2 is inspired by [Cha13]. Finally, beyond Theorem 21.7, there is a sufficient condition for $\mu \in \mathcal{P}$ to be determined by its moments $(\kappa_n)_{n\geqslant 1}$, known as Torsten Carleman's condition, and related to the quasi-analyticity[8] of the Fourier transform:

$$\sum_n \kappa_{2n}^{-1/(2n)} = \infty.$$

The Courant-Fischer inequalities, as well as Lemmas 21.3 and 21.4 (the Hoffman-Wielandt inequality), are proved, for example, in the book by Roger Horn and Charles Johnson [HJ13] and in the book by Rajendra Bhatia [Bha97].

[8] A function $f : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^\infty$ is quasi-analytic if it is determined by its successive derivatives at 0, among a sufficiently broad class of functions.

22

Births and assassinations

Keywords. Population dynamics.

Tools. Poisson process; Galton-Watson process; Chernoff's theorem.

Difficulty. ***

We study the extinction or survival of a population organized into clans. Each living individual gives birth to children according to a Poisson process and cannot die while its father is alive. As soon as the father disappears, all its children become exposed and die after their own random time with distribution Q. A configuration at a given time is a finite collection of genealogical trees, each with a patriarch who alone can be assassinated.

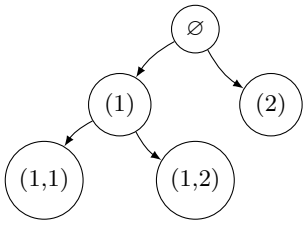


Fig. 22.1. A patriarch and some of its descendants.

The patriarch of the entire population is denoted by $(\varnothing)$. Each of its descendants is identified with a k-tuple of positive integers for $k \geqslant 1$. For example, (1) is the ancestor's first child, whereas $(1,2)$ is the second child of the ancestor's first child; see Figure 22.1. The number of coordinates gives the individual's generation. Let

$$\mathcal{N} = \cup_{n \geqslant 0} (\mathbb{N}^*)^n$$

be the set of n-tuples of positive integers, with the convention $(\mathbb{N}^*)^0 = \varnothing$. For $\mathbf{n} \in \mathcal{N}$, let $k(\mathbf{n})$ denote the number of coordinates of $\mathbf{n}$. The number of individuals that must have been born for $\mathbf{n}$ to be born is

$$\sigma(\mathbf{n}) := n_1 + \cdots + n_{k(\mathbf{n})}.$$

We adopt the conventions $k(\varnothing) = 0 = \sigma(\varnothing)$.

Let $(X^{\mathbf{n}})_{\mathbf{n}\in\mathcal{N}}$ be a family of independent Poisson processes with the same intensity λ. We denote by $(T_k^{\mathbf{n}})_{k\geqslant 1}$ the jump times of the process $X^{\mathbf{n}}$. Let $(K_{\mathbf{n}})_{\mathbf{n}\in\mathcal{N}}$ be a family of strictly positive i.i.d. real random variables with common distribution Q. We assume that the families $(X^{\mathbf{n}})_{\mathbf{n}\in\mathcal{N}}$ and $(K_{\mathbf{n}})_{\mathbf{n}\in\mathcal{N}}$ are independent.

The process starts at $t = 0$ with a single ancestor labelled $\varnothing$. This individual gives birth to children at the jump times of the Poisson process $X^{\varnothing}$; these enter the system in turn with indices $(1), (2), \ldots$ in their order of birth. Each new individual $\mathbf{n}$ entering the system begins to produce descendants at the jump times of the process $X^{\mathbf{n}}$; the descendants of $\mathbf{n}$ are indexed by $(\mathbf{n}, 1), (\mathbf{n}, 2), \ldots$, again in chronological order.

The patriarch is exposed from time 0. It gives birth to children until the random time $Y_{\varnothing} := K_{\varnothing}$ when it disappears. Each of its children becomes the patriarch of its clan and is exposed in turn. Let $\mathbf{n} = (n_1, \ldots, n_{k-1}, n_k)$ belong to $\mathcal{N}$ and let $\mathbf{n}' := (n_1, \ldots, n_{k-1})$. When $\mathbf{n}'$ is removed, at the random time $Y_{\mathbf{n}'}$, $\mathbf{n}$ becomes exposed; it continues to produce descendants until the random time $Y_{\mathbf{n}} := Y_{\mathbf{n}'} + K_{\mathbf{n}}$ and is then removed in turn.

Example 22.1. *Saying that individual* $(3, 2)$ *has been born means that*

- *individual* $\varnothing$ *had at least three children before dying:* $T_3^{\varnothing}$ *occurs before the death of* $\varnothing$ *at time* $K_{\varnothing}$*;*
- *individual* (3) *had at least two children before dying, that is,* $T_2^{(3)}$ *occurs before the death of* (3) *at time* $K_{\varnothing} + K_{(3)}$*.*

We therefore have

$$\{(3, 2) \text{ est né}\} = \left\{T_3^{\varnothing} < K_{\varnothing},\ T_3^{\varnothing} + T_2^{(3)} < K_{\varnothing} + K_{(3)}\right\}.$$

The birth-and-assassination system is said to be *stable* if, with probability 1, there is a finite time after which the system is empty. Otherwise, the system is said to be *unstable*.

Theorem 22.2 (Sufficient conditions for stability and instability)**.** *Let* $\mathcal{B}$ *be a birth-and-assassination process with intensity* λ *such that the Laplace transform* $u \mapsto \varphi(u) = \int_0^{\infty} e^{ux}\, Q(dx)$ *of* Q *is finite in a neighborhood of* 0*, and let*

$$\alpha := \min_{u>0} \frac{\lambda\varphi(u)}{u}.$$

Then the following sufficient conditions for stability and instability hold:

- *the process $\mathcal{B}$ is stable if $\alpha < 1$;*
- *the process $\mathcal{B}$ is unstable if $\alpha > 1$.*

Sections 22.1 and 22.2 are devoted to proving the two conclusions of Theorem 22.2.

22.1 A sufficient condition for stability

Introduce the mean number of descendants of $\varnothing$,

$$m := \sum_{\mathbf{n}} \mathbb{P}(\mathbf{n} \text{ est né}) \in [0, +\infty].$$

If $\mathcal{B}$ is unstable, then by definition, with strictly positive probability, there is no finite (random) time after which the system is empty. Hence the number of descendants of $\varnothing$ is infinite with positive probability, and therefore $m = \infty$. Consequently, if $m < \infty$ then $\mathcal{B}$ is stable. Thus it suffices to establish that $m < \infty$ whenever $\alpha < 1$.

To individual $\mathbf{n} = (n_1, \ldots, n_k)$ we associate its ancestral lineage, consisting of the individuals $\mathbf{n}(i) = (n_1, \ldots, n_i)$ for $0 \leqslant i \leqslant k$ (with $\mathbf{n}(0) = \varnothing$). Individual $\mathbf{n}$ is born if, for every $0 \leqslant i \leqslant k$, its ancestor $\mathbf{n}(i-1)$ in generation i has had at least n_i descendants:

$$\{\mathbf{n} \text{ est né}\} = \left\{ \sum_{i=1}^{j} T_{n_i}^{\mathbf{n}(i-1)} < \sum_{i=1}^{j} K_{\mathbf{n}(i-1)},\ j = 1, 2, \ldots, k(\mathbf{n}) \right\}.$$

In particular,

$$\mathbb{P}(\mathbf{n} \text{ est né}) \leqslant \mathbb{P}\left(\sum_{i=1}^{k(\mathbf{n})} T_{n_i}^{\mathbf{n}(i-1)} < \sum_{i=1}^{k(\mathbf{n})} K_{\mathbf{n}(i-1)} \right).$$

Markov's inequality ensures that, for every $u > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{k(\mathbf{n})} T_{n_i}^{\mathbf{n}(i-1)} < \sum_{i=1}^{k(\mathbf{n})} K_{\mathbf{n}(i-1)} \right) \leqslant \mathbb{E}\left(\exp\left(u \sum_{i=1}^{k(\mathbf{n})} K_{\mathbf{n}(i-1)} - u \sum_{i=1}^{k(\mathbf{n})} T_{n_i}^{\mathbf{n}(i-1)} \right) \right).$$

Independence of the birth and assassination processes gives

$$\mathbb{E}\left(\exp\left(u \sum_{i=1}^{k(\mathbf{n})} K_{\mathbf{n}(i-1)} - u \sum_{i=1}^{k(\mathbf{n})} T_{n_i}^{\mathbf{n}(i-1)} \right) \right) = \varphi(u)^{k(\mathbf{n})} \prod_{i=1}^{k(\mathbf{n})} \left(\frac{\lambda}{\lambda + u} \right)^{n_i}.$$

Summing over $\mathbf{n}$ yields

$$\sum_{\mathbf{n}} \mathbb{P}(\mathbf{n} \text{ est né}) \leqslant \sum_{k=1}^{\infty} \varphi(u)^k \left[\sum_{\mathbf{n},\, k(\mathbf{n})=k} \left(\frac{\lambda}{\lambda+u} \right)^{\sigma(\mathbf{n})} \right].$$

On the other hand,

$$\begin{aligned} &\sum_{\mathbf{n},\, k(\mathbf{n})=k} \left(\frac{\lambda}{\lambda+u} \right)^{\sigma(\mathbf{n})} \\ &\quad = \sum_{m \geqslant k} \operatorname{card}\{\mathbf{n} \,:\, k(\mathbf{n}) = k,\ \sigma(\mathbf{n}) = m\} \left(\frac{\lambda}{\lambda+u} \right)^m \\ &\quad = \left(\frac{\lambda}{u} \right)^k \sum_{m \geqslant k} \operatorname{card}\{\mathbf{n} \,:\, k(\mathbf{n}) = k,\ \sigma(\mathbf{n}) = m\} \left(\frac{u}{u+\lambda} \right)^k \left(\frac{\lambda}{u+\lambda} \right)^{m-k}. \end{aligned}$$

The integer $\operatorname{card}\{\mathbf{n} \,:\, k(\mathbf{n}) = k,\ \sigma(\mathbf{n}) = m\}$ represents the number of m-tuples with coordinates 0 or 1 containing exactly k 1's. The general term of the last sum above is therefore the probability that, in a Bernoulli scheme with success probability $u/(u+\lambda)$, the k^{e} success occurs at time m (this is the negative binomial distribution). Along the way, we recover

$$\operatorname{card}\{\mathbf{n} \,:\, k(\mathbf{n}) = k,\ \sigma(\mathbf{n}) = m\} = \binom{m-1}{k-1}.$$

We therefore have

$$\sum_{\mathbf{n},\, k(\mathbf{n})=k} \left(\frac{\lambda}{\lambda+u} \right)^{\sigma(\mathbf{n})} = \left(\frac{\lambda}{u} \right)^k.$$

In other words, we have shown that

$$m \leqslant \sum_{k \geqslant 1} \left(\frac{\lambda \varphi(u)}{u} \right)^k.$$

If $\alpha < 1$, then the mean number of individuals created is finite, and $\mathcal{B}$ is stable.

22.2 A sufficient condition for instability

We now prove the second part of Theorem 22.2. The main idea is to show that, if $\alpha > 1$, there exists $k \geqslant 1$ such that

$$\sum_{\mathbf{n},\, k(\mathbf{n})=k} \mathbb{P}(\mathbf{n} \text{ est né}) > 1$$

and then conclude by a branching argument.

Let $(N_i)_{i\geqslant 1}$ be a sequence of independent random variables with geometric distribution of parameter p, independent of the process $\mathcal{B}$. In generation k, consider individual $\mathbf{N} = (N_1, \dots, N_k)$ together with its ancestral lineage consisting of the individuals $\mathbf{N}(i) = (N_1, \dots, N_i)$ for $0 \leqslant i \leqslant k$. We express in two different ways the probability that $\mathbf{N}$ is born. On the one hand,

$$\begin{aligned}\mathbb{P}(\mathbf{N} \text{ est né}) &= \sum_{\mathbf{n},\, k(\mathbf{n})=k} \mathbb{P}\Big(\mathbf{N} \text{ est né} \,\Big|\, \mathbf{N}=\mathbf{n}\Big)\mathbb{P}(\mathbf{N}=\mathbf{n}) \\ &= \sum_{\mathbf{n},\, k(\mathbf{n})=k} \mathbb{P}(\mathbf{n} \text{ est né}) p^k (1-p)^{n_1+\cdots+n_k-k} \\ &= \left(\frac{p}{1-p}\right)^k \sum_{\mathbf{n},\, k(\mathbf{n})=k} \mathbb{P}(\mathbf{n} \text{ est né})(1-p)^{\sigma(\mathbf{n})}.\end{aligned}$$

On the other hand, set

$$Z_i = K_{\mathbf{N}(i-1)} - T_{N_i}^{\mathbf{N}(i-1)},$$

for every $i \geqslant 1$. As in the first part of the proof,

$$\begin{aligned}\mathbb{P}(\mathbf{N} \text{ est né}) &= \mathbb{P}\left(\sum_{i=1}^{j} T_{N_i}^{\mathbf{N}(i-1)} < \sum_{i=1}^{j} K_{\mathbf{N}(i-1)},\ j=1,\dots,k\right) \\ &= \mathbb{P}\left(\sum_{i=1}^{j} Z_i > 0,\ j=1,\dots,k\right).\end{aligned}$$

We estimate this probability using the following lemma, proved at the end of the section.

Lemma 22.3 (Asymptotic deviation)**.** *If $(X_i)_{i\geqslant 1}$ is a sequence of i.i.d. real random variables with $\mathbb{E}(X_1) < 0$, $\mathbb{P}(X_1 > 0) > 0$, and if the Laplace transform ψ of X_1 is finite in a neighborhood of 0, then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\min_{1\leqslant k\leqslant n} \sum_{i=1}^{k} X_i > 0\right) \geqslant \log(\rho), \quad \textit{où} \quad \rho := \min_{u>0} \psi(u).$$

The random variable $T_{N_i}^{\mathbf{N}(i-1)}$ has the exponential distribution[1] $\mathrm{Exp}(\lambda p)$. The random variables $(Z_i)_{i\geqslant 1}$ are i.i.d. with Laplace transform and mean

$$\mathbb{E}\big(e^{uZ_1}\big) = \varphi(u)\frac{\lambda p}{\lambda p + u} \quad \text{et} \quad \mathbb{E}(Z_1) = \varphi'(0) - \frac{1}{p\lambda}.$$

[1] Lemma 11.4 states that if $G, E_1, E_2, \dots$ are independent random variables with G geometrically distributed with parameter p and $E_1, E_2, \dots$ exponentially distributed with parameter λ, then the random sum $E_1 + \cdots + E_G$ has the exponential distribution with parameter $p\lambda$.

Thus the expectation of Z_1 is strictly negative for sufficiently small p. Moreover, writing $\alpha = 1 + \varepsilon$, we can choose p sufficiently small that

$$\min_{u>0} \mathbb{E}\big(e^{uZ_1}\big) = p \min_{u>0} \frac{\lambda\varphi(u)}{\lambda p + u} > p(1 + \varepsilon/2).$$

Lemma 22.3 applied to the sequence $(Z_n)_{n\geqslant 1}$ ensures that, for sufficiently large k,

$$\mathbb{P}\left(\min_{1\leqslant j\leqslant k} \sum_{i=1}^{j} Z_i > 0\right) > p^k(1 + \varepsilon/2)^k.$$

We therefore obtain

$$\left(\frac{p}{1-p}\right)^k \sum_{\mathbf{n},\, k(\mathbf{n})=k} \mathbb{P}(\mathbf{n} \text{ est né})(1-p)^{\sigma(\mathbf{n})} > p^k(1 + \varepsilon/2)^k.$$

Divide by p^k and let p tend to 0. By monotone convergence,

$$\sum_{\mathbf{n},\, k(\mathbf{n})=k} \mathbb{P}(\mathbf{n} \text{ est né}) \geqslant (1 + \varepsilon/2)^k > 1,$$

which gives the relation

$$\sum_{\mathbf{n},\, k(\mathbf{n})=k} \mathbb{P}(\mathbf{n} \text{ est né}) > 1.$$

Let us complete the proof of Theorem 22.2. An individual is called *special* if it is the ancestor or if it belongs to generation nk and descends from a special individual in generation $(n-1)k$ whose father has died. The relation obtained above tells us that the mean number of special individuals in generation k is strictly greater than $(1 + \varepsilon/2)^k$. Similarly, the mean number of special individuals in generation nk is greater than $(1 + \varepsilon/2)^{nk}$. Thus the process counting special individuals is a Galton-Watson process that is supercritical. With strictly positive probability, this process does not become extinct (see Theorem 3.7 on page 3.7). The same therefore holds for $\mathcal{B}$.

We finish this section with the proof of Lemma 22.3.

Proof of Lemma 22.3. Under the assumptions of Lemma 22.3, the conclusion of Lemma 22.3 can be proved at the terminal time using Chernoff's theorem:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\sum_{i=1}^{n} X_i > 0\right) = \log(\rho).$$

We now show that

$$\mathbb{P}\left(\min_{1\leqslant k\leqslant n} S_k > 0\right) \geqslant \frac{1}{n}\mathbb{P}(S_n > 0) \quad \text{où} \quad S_k = \sum_{i=1}^{k} X_i.$$

We proceed as for the rotation principle of Lemma 3.18. For every k, let $\sigma(k)$ denote the integer between 1 and n equal to k modulo n. For every $l = 1, \dots, n$, define S^l by

$$S_0^{(l)} = 0 \quad \text{et} \quad S_{k+1}^{(l)} = S_k^{(l)} + X_{\sigma(k+l)}.$$

We have $S = S^{(l)}$. The total sum $S_n^{(l)}$ does not depend on l and equals S_n. Since $(X_1, \dots, X_n)$ and $(X_l, \dots, X_n, X_1, \dots, X_{l-1})$ have the same distribution, we obtain

$$\mathbb{P}\left(\min_{1 \leqslant k \leqslant n} S_k > 0 \,\middle|\, S_n > 0 \right) = \frac{1}{n} \mathbb{E}\left(\sum_{l=1}^{n} \mathbb{1}_{\left\{ \min_{1 \leqslant k \leqslant n} S_k^{(l)} > 0 \right\}} \mid S_n > 0 \right).$$

Now the sum on the right-hand side is bounded below by 1 since, on $\{S_n > 0\}$, there is a (random) value of l for which the minimum of $S^{(l)}$ is strictly positive: the largest index at which $l \mapsto S_l$ attains its minimum. It is useful to draw a picture and compare it with Figure 3.3 on page 3.3. This yields the claimed lower bound. □

22.3 Further reading

The model presented here was introduced in an article [AK90] by David Aldous and William Krebs. The connection with a queueing model is discussed in the article [TPH86] by John Tsitsiklis, Christos Papadimitriou, and Pierre Humblet. A connection with branching random walks can be found in the article [Kor15] by Igor Kortchemski. (Herman) Chernoff's theorem, which is said to be due to Herman Rubin, is proved, for example, in the book [DZ10] by Amir Dembo and Ofer Zeitouni.

Theorem 22.2 leaves open the question of whether $\mathcal{B}$ is stable at the critical point. A precise answer can be given in an important special case. Suppose the assassination distribution is exponential. By rescaling time, we may assume without loss of generality that its parameter equals 1, giving $\varphi(u) = (1-u)^{-1}\mathbb{1}_{[0,1[}(u) + \infty \mathbb{1}_{[1,\infty[}(u)$, hence $\min_{u>0} \lambda u^{-1} \varphi(u) = 4\lambda$ (the minimum is attained at $u = 1/2$). Theorem 22.2 ensures that $\mathcal{B}$ is stable if $0 < \lambda < 1/4$ and unstable if $\lambda > 1/4$. Write $\mathbb{E}_\lambda$ for expectation with respect to the distribution of $\mathcal{B}_\lambda$ and N for the total number of individuals born, including the ancestor. The article [Bor08] by Charles Bordenave shows, in particular, that the moments of N satisfy the following integrability properties:

- if $0 < \lambda < 1/4$ then
$$\mathbb{E}_\lambda(N) = \frac{2}{1 + \sqrt{1 - 4\lambda}};$$
- if $0 < \lambda < 2/9$ then
$$\mathbb{E}_\lambda(N^2) = \frac{2}{3\sqrt{1 - 4\lambda} - 1};$$

- for every $p \geqslant 2$, $\mathbb{E}_\lambda(N^p) < \infty$ if and only if

$$0 < \lambda < \frac{p}{(p+1)^2}.$$

In particular, if $\lambda = 1/4$, the process $\mathcal{B}$ is stable since $\mathbb{E}_{1/4} N = 2$. It is also shown in [Bor08] that the distribution of N has a polynomial tail for every $0 < \lambda < 1/4$. Indeed, for every $\varepsilon > 0$, there exists C such that, for every $t > 0$,

$$\mathbb{P}_\lambda(N > t) \leqslant C t^{-\gamma(\lambda)+\varepsilon} \quad \text{avec} \quad \gamma(\lambda) = \frac{1+\sqrt{1-4\lambda}}{1-\sqrt{1-4\lambda}}.$$

The subcritical Galton-Watson processes studied in Chapter 3 do not exhibit this heavy-tail behavior: there is a constant $c > 0$ such that $\mathbb{E}_\lambda \exp(cN) < \infty$ for $\lambda < 1$. Moreover, the mean number of individuals is given by $\mathbb{E}_\lambda(N) = (1-\lambda)^{-1}$, and therefore diverges as λ tends to the critical value 1.

23

The telegraph model

Keywords. Piecewise deterministic Markov process; partial differential equation; infinitesimal generator.

Tools. Poisson process; continuous-time Markov chain; Laplace transform.

Difficulty. ***

We study the motion of a bacterium with four to six flagella distributed over its surface, which moves by rotating them like a corkscrew. When they rotate counterclockwise, the flagella form a propulsive bundle producing motion in an almost straight line. When they rotate clockwise, the flagella become uncoordinated and cause the cell to rotate without significant displacement. The cell thus alternates straight-line movements, or "*runs*", and rotations in place, or "*tumbles*". A run lasts on average about one second, while the average tumble duration is of the order of a tenth of a second. Below, we neglect the tumble time and assume that motion occurs in dimension 1. In an isotropic medium, the rate at which a tumble occurs during a run is essentially independent of the time since the last tumble: the duration of a run can therefore be modeled by an exponential distribution with parameter α.

Denote by X_t and V_t the bacterium's position and velocity at time t. After suitable rescaling, its velocity takes values in the set $\{-1,+1\}$. Let $(N_t)_{t\geqslant 0}$ be a Poisson process of intensity α and (V_0,X_0) a random variable taking values in $E=\{-1,+1\}\times\mathbb{R}$, independent of $(N_t)_{t\geqslant 0}$. Define the process $((V_t,X_t))_{t\geqslant 0}$ with values in E by setting, for every $t\geqslant 0$,

$$V_t=(-1)^{N_t}V_0 \quad \text{et} \quad X_t=X_0+\int_0^t V_s\,ds.$$

We wish to understand how the distribution of bacteria evolves over time, given its initial distribution.

First, the velocity process $(V_t)_{t\geqslant 0}$ is a Markov process on $\{-1,+1\}$ whose distribution can be determined. For every $t\geqslant 0$, write

$$p_+(t) = \mathbb{P}(V_t = +1) \quad \text{et} \quad p_-(t) = \mathbb{P}(V_t = -1).$$

Theorem 23.1 (Distribution of the velocity)**.** *The process $(V_t)_{t\geqslant 0}$ has a unique invariant distribution, namely the symmetric Rademacher distribution $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$. Moreover, for every $t \geqslant 0$, the distribution of V_t is given by*

$$p_+(t) = \frac{1}{2} + \frac{p_+(0) - p_-(0)}{2} e^{-2\alpha t} \quad et \quad p_-(t) = \frac{1}{2} - \frac{p_+(0) - p_-(0)}{2} e^{-2\alpha t},$$

which solve the differential equations

$$p'_+(t) = \alpha(p_-(t) - p_+(t)) \quad et \quad p'_-(t) = \alpha(p_+(t) - p_-(t)).$$

Proof. The formula

$$p_+(t) = \sum_{n=0}^{\infty} \mathbb{P}((-1)^n = V_0)\mathbb{P}(N_t = n) = \sum_{n=0}^{\infty} \mathbb{P}((-1)^n = V_0) e^{-\alpha t} \frac{(\alpha t)^n}{n!}$$

immediately yields the differential equation

$$p'_+(t) = -\alpha p_+(t) + \alpha p_-(t).$$

Now $p_+ + p_- = 1$, giving the differential equation satisfied by p_-. Moreover,

$$(p_+ - p_-)'(t) = -2\alpha(p_+ - p_-)(t) \quad \text{et} \quad (p_+ + p_-)'(t) = 0.$$

Solving these linear differential equations yields the result. □

In Section 23.1, we study the Markov properties of the process V and the full process (V, X), inherited from the Poisson process. We then connect this process to a partial differential equation called the telegraph equation in Section 23.2.

23.1 Markov aspects of the full process

In this section, we study the Markov semigroup and infinitesimal generator associated with the process (V, X). The approach is similar to that adopted for the growth-fragmentation process in Chapter 17 and the M/M/∞ queue in Chapter 11. For every $t \geqslant 0$ and every bounded measurable function f from E to $\mathbb{R}$, write

$$\begin{aligned} P_t f(v, x) &= \mathbb{E}(f(V_t, X_t) | V_0 = v, X_0 = x) \\ &= \mathbb{E}\left[f\left((-1)^{N_t} v, x + v \int_0^t (-1)^{N_s} \, ds \right) \right]. \end{aligned}$$

For $k \in \mathbb{N}$, denote by $\mathcal{C}_E^k$ the set of functions $f\colon E \to \mathbb{R}$ which are measurable, bounded, of class $\mathcal{C}^k$ in the second variable, with bounded successive derivatives.

Theorem 23.2 (Markov semigroup)**.** *The family $(P_t)_{t\geqslant 0}$ satisfies:*

- *for every $t \geqslant 0$, if $\mathbb{1}$ is the constant function equal to 1, then $P_t\mathbb{1} = \mathbb{1}$;*
- *if f is nonnegative, so is $P_t f$;*
- *for every $f \in \mathcal{C}^k_E$ and every $t \geqslant 0$, $P_t f \in \mathcal{C}^k_E$;*
- *for every $f \in \mathcal{C}^1_E$, $\lim_{t\to 0} P_t f = f$ uniformly;*
- *for all $t, s \geqslant 0$ and every bounded measurable function $f : E \to \mathbb{R}$,*

$$P_{t+s}f(v,x) = P_t \circ P_s f(v,x).$$

We say that $(P_t)_{t\geqslant 0}$ is a Markov semigroup.

Proof. The first two assertions are straightforward. To establish the third, write

$$\begin{aligned}|P_t f(v,x) - f(v,x)| \leqslant{}& \mathbb{E}(|f(V_t,X_t) - f(v,x)| \,|\, V_0 = v, X_0 = x)\\ \leqslant{}& \mathbb{E}\big(|f(V_t,X_t) - f(v,x)|\mathbb{1}_{\{N_t=0\}} \,|\, V_0 = v, X_0 = x\big)\\ &+ \mathbb{E}\big(|f(V_t,X_t) - f(v,x)|\mathbb{1}_{\{N_t\geqslant 1\}} \,|\, V_0 = v, X_0 = x\big)\end{aligned}$$

If $N_t = 0$, then for every $s \in [0,t]$, $V_s = v$ and $X_t = x + tv$. Hence

$$\begin{aligned}\mathbb{E}\big(|f(V_t,X_t) - f(v,x)|\mathbb{1}_{\{N_t=0\}} \,|\, V_0 = v, X_0 = x\big) &= |f(v,x+vt) - f(v,x)|\\ &\leqslant \|\partial_x f\|_\infty t,\end{aligned}$$

where we used $|v| = 1$. Moreover,

$$\begin{aligned}\mathbb{E}\big(|f(V_t,X_t) - f(v,x)|\mathbb{1}_{\{N_t\geqslant 1\}} \,|\, V_0 = v, X_0 = x\big) &\leqslant 2\|f\|_\infty \mathbb{P}(N_t \geqslant 1)\\ &= 2\|f\|_\infty\big(1 - e^{-\alpha t}\big).\end{aligned}$$

Let us prove the last property. For all $s, t \geqslant 0$, we have

$$P_{t+s}f(v,x) = \mathbb{E}\{\mathbb{E}\{f(V_{t+s},X_{t+s}) \,|\, X_0, V_0, X_t, V_t\} \,|\, X_0 = x, V_0 = v\}.$$

By definition of (V,X), write

$$V_{t+s} = (-1)^{N_{t+s}-N_t}V_t \quad \text{et} \quad X_{t+s} = X_t + V_t\int_0^s (-1)^{N_{t+u}-N_t}\,du.$$

We then use the fact that the Poisson process has independent stationary increments and, in particular, that $(N_{t+s} - N_t)_{s\geqslant 0}$ is a Poisson process of parameter α, independent of $(N_u)_{0\leqslant u\leqslant t}$, hence

$$\begin{aligned}\mathbb{E}\left(f\left((-1)^{N_{t+s}-N_t}V_t, X_t + V_t\int_0^s (-1)^{N_{t+u}-N_u}\,du\right)\,\middle|\, X_0, V_0, X_t, V_t\right)&\\ = P_s f(V_t, X_t).&\end{aligned}$$

We thus recover the semigroup identity $P_{t+s}f(v,x) = P_t \circ P_s f(v,x)$. □

Theorem 23.3 (Infinitesimal generator)**.** *For every $t \geqslant 0$ and every $f \in \mathcal{C}_E^1$,*

$$\left\| \frac{P_{t+h}f - P_t f}{h} - AP_t f \right\|_\infty \underset{h\to 0}{\longrightarrow} 0,$$

where the operator A is defined by

$$Af(v,x) := \alpha(f(-v,x) - f(v,x)) + v\partial_x f(v,x),$$

for every $f \in \mathcal{C}_E^1$. In particular, for every $t \geqslant 0$ and every $f \in \mathcal{C}_E^1$,

$$\partial_t P_t f(v,x) = AP_t f(v,x) = P_t(Af)(v,x).$$

Proof. First prove the result for $t = 0$. Let $f \in \mathcal{C}_E^1$. We have

$$\begin{aligned} P_h f(v,x) =& \mathbb{E}\big(f(V_h, X_h)\mathbb{1}_{\{N_h=0\}} \,|\, V_0 = v, X_0 = x\big) \\ &+ \mathbb{E}\big(f(V_h, X_h)\mathbb{1}_{\{N_h=1\}} \,|\, V_0 = v, X_0 = x\big) \\ &+ \mathbb{E}\big(f(V_h, X_h)\mathbb{1}_{\{N_h\geqslant 2\}} \,|\, V_0 = v, X_0 = x\big). \end{aligned}$$

We study these three expectations separately. The absolute value of the third and last expectation is bounded by

$$\|f\|_\infty \mathbb{P}(N_h \geqslant 2) = \|f\|_\infty (1 - e^{-\alpha h}(1+\alpha h)) = \mathcal{O}(h^2).$$

The first expectation equals

$$\begin{aligned} \mathbb{E}\big(f(v, x+vh)\mathbb{1}_{\{N_h=0\}}\big) &= f(v, x+vh)e^{-\alpha h} \\ &= f(v,x) + h(v\partial_x f(v,x) - \alpha f(v,x)) + o(h). \end{aligned}$$

Finally, the second expectation can be written

$$\begin{aligned} \mathbb{E}\left[f\left(-v, x + v\int_0^h (-1)^{N_u}\, du\right)\mathbb{1}_{\{N_h=1\}}\right] &= \alpha h e^{-\alpha h}(f(-v,x) + \mathcal{O}(h)) \\ &= \alpha h f(-v,x) + o(h) \end{aligned}$$

since the integral term is bounded in absolute value by h. We have thus obtained

$$P_h f(v,x) = f(v,x) + hAf(v,x) + o(h),$$

completing the case $t = 0$. To extend the result to every time $t \geqslant 0$, use the semigroup property and the first part of the proof:

$$\frac{P_{t+h}f - P_t f}{h} - AP_t f = \left(\frac{P_h - P_0}{h} - A\right) P_t f \underset{h\to 0}{\overset{\|\cdot\|_\infty}{\longrightarrow}} 0,$$

since $P_t f \in \mathcal{C}_E^1$. Similarly, the alternative expression

$$\frac{P_{t+h}f - P_t f}{h} = P_t\left(\frac{P_h - P_0}{h}\right) f;$$

yields the relation $\partial_t P_t f = P_t A f$. □

23.2 Connecting the equation and the process

Let $\phi : \mathbb{R} \to \mathbb{R}$ be of class $\mathcal{C}^2$. Define u^+ and u^- on $\mathbb{R}_+ \times \mathbb{R}$ by

$$\begin{cases} u^+(t,x) := P_t\phi(1,x) = \mathbb{E}(\phi(X_t) \,|\, V_0 = 1, X_0 = x), \\ u^-(t,x) := P_t\phi(-1,x) = \mathbb{E}(\phi(X_t) \,|\, V_0 = -1, X_0 = x), \end{cases}$$

with the following abuse of notation: we also denote by ϕ the map $(v,x) \mapsto \phi(x)$. By definition of u^+ and u^-, the last formula in Theorem 23.3 ensures that

$$\begin{aligned} \partial_t u^+(t,x) &= \partial_t P_t\phi(1,x) = AP_t\phi(1,x) \\ &= \alpha(P_t\phi(-1,x) - P_t\phi(1,x)) + \partial_x P_t\phi(1,x) \\ &= \alpha(u^-(t,x) - u^+(t,x)) + \partial_x u^+(t,x). \end{aligned}$$

We can do the same for u^- and deduce that (u^+, u^-) solves

$$\begin{cases} \partial_t u^+(t,x) = \alpha(u^-(t,x) - u^+(t,x)) + \partial_x u^+(t,x) \\ \partial_t u^-(t,x) = \alpha(u^+(t,x) - u^-(t,x)) - \partial_x u^-(t,x). \end{cases}$$

Setting

$$u(t,x) := \frac{u^+(t,x) + u^-(t,x)}{2} \quad \text{et} \quad \tilde{u}(t,x) := \frac{u^+(t,x) - u^-(t,x)}{2},$$

the pair $(u, \tilde{u})$ solves

$$\begin{cases} \partial_t u(t,x) = \partial_x \tilde{u}(t,x), \\ \partial_t \tilde{u}(t,x) = -2\alpha\tilde{u}(t,x) + \partial_x u(t,x). \end{cases}$$

Differentiating the first equation with respect to t and the second with respect to x eliminates the term $\partial_t\partial_x\tilde{u}(t,x)$ and gives

$$\partial^2_{tt} u(t,x) = -2\alpha\partial_x\tilde{u}(t,x) + \partial^2_{xx} u(t,x).$$

It then remains to replace $\partial_x\tilde{u}(t,x)$ by $\partial_t u(t,x)$ to obtain

$$\partial^2_{tt} u(t,x) - \partial^2_{xx} u(t,x) + 2\alpha\partial_t u(t,x) = 0.$$

Now, when $X_0 = x$ for some $x \in \mathbb{R}$ and $V_0 \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, we find $u(t,x) = \mathbb{E}(\phi(x + X_t))$, proving the following result.

Theorem 23.4 (Probabilistic representation of the solution of a PDE)**.** *The function u defined by $u(t,x) = \mathbb{E}(\phi(x+X_t))$ for every $t \geqslant 0$ and $x \in \mathbb{R}$, where $X_0 = 0$ and V_0 has distribution $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, solves the telegraph equation:*

$$\begin{cases} \partial^2_{tt} u(t,x) - \partial^2_{xx} u(t,x) + 2\alpha\partial_t u(t,x) = 0 & \text{pour tous } t \geqslant 0,\ x \in \mathbb{R}, \\ u(0,x) = \phi(x) & \text{pour tout } x \in \mathbb{R}, \\ \partial_t u(0,x) = 0 & \text{pour tout } x \in \mathbb{R}. \end{cases}$$

The partial differential equation in Theorem 23.4 is a superposition of two classical equations: the wave equation (or vibrating string equation) and the heat equation. The first describes propagation of a signal without damping, while the second describes a diffusion phenomenon.

23.3 An ergodic modification

In this section, we slightly modify the dynamics of the process (V, X). The velocity still takes values in $\{-1, +1\}$ and X still corresponds to the integral of V. However, we make the jump mechanism of V slightly more complicated. For simplicity, we recursively construct the positions at the jump times of V and then interpolate between them.

Let $a > 0$ and $b > 0$ be fixed real numbers. Let $(E_n)_{n\geqslant 1}$ be a sequence of i.i.d. real random variables exponential with parameter 1, independent of $(\tilde{V}_0, \tilde{X}_0) \in \{-1, +1\} \times \mathbb{R}_+$. Write $T_0 = 0$. Recursively define the following quantities:

$$T_n = T_{n-1} + \begin{cases} \dfrac{E_n}{b} & \text{si } \tilde{V}_{n-1} = 1, \\ \min\left(\dfrac{E_n}{a}, \tilde{X}_{n-1}\right) & \text{si } \tilde{V}_{n-1} = -1, \end{cases}$$

and then

$$\tilde{V}_n = -\tilde{V}_{n-1} \quad \text{et} \quad \tilde{X}_n = \tilde{X}_{n-1} + \tilde{V}_{n-1}(T_n - T_{n-1}).$$

Construct the process (X, V) by setting $V_t := \tilde{V}_{n-1}$ if $T_{n-1} \leqslant t < T_n$ and defining X by linear interpolation of $\tilde{X}$. Thus the jump rate of V is given by

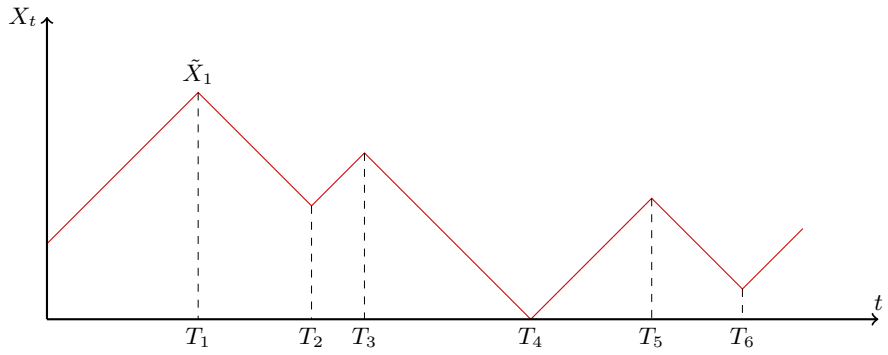


Fig. 23.1. A sample path.

$$b\mathbb{1}_{\{v=1\}} + \big(a + \infty\mathbb{1}_{\{x=0\}}\big)\mathbb{1}_{\{v=-1\}}.$$

An excursion of (V, X) is a path starting from $(+1, 0)$ and stopped at time

$$S := \inf\{t > 0 \,:\, X_t = 0\}.$$

Denote by ψ the Laplace transform of S:

$$\psi : \lambda \in \mathbb{R} \mapsto \psi(\lambda) = \mathbb{E}_{(1,0)}\big(e^{\lambda S}\big).$$

Lemma 23.5 (Length of an excursion)**.** *The domain of ψ is $]-\infty, \lambda_c]$, where λ_c is given by*

$$\lambda_c = \frac{a+b}{2} - \sqrt{ab} = \frac{(\sqrt{b} - \sqrt{a})^2}{2}.$$

Moreover, if $\lambda \leqslant \lambda_c$, then

$$\psi(\lambda) = \frac{a + b - 2\lambda - \sqrt{(a+b-2\lambda)^2 - 4ab}}{2a}.$$

In particular, $\psi(\lambda_c) = \sqrt{b/a}$ and $\mathbb{E}_{(1,0)}(S) = 2/(b-a)$.

Proof. The first jump time $T_1 = E$ of V when $(V_0, X_0) = (1, 0)$ has the exponential distribution with parameter b. If V does not jump between E and $2E$, then $S = 2E$. Otherwise, (V, X) makes an excursion above X_{T_2} with the same distribution as X. After this excursion, (V, X) is at $(-1, X_{T_2})$. Again, either X reaches 0 before V jumps, or it does not (in which case another excursion is grafted on). Figure 23.2 illustrates this decomposition. The number N of excursions included in the descent from E has distribution $\mathrm{Poi}(aE)$. We can therefore write

$$S \overset{\text{loi}}{=} 2E + \sum_{k=1}^{N} S_k,$$

where E, N, and $(S_k)_{k\geqslant 1}$ are independent. Thus, for every λ in the domain of ψ (which contains $]-\infty, 0]$),

$$\begin{aligned}\psi(\lambda) &= \mathbb{E}\left(\mathbb{E}\left(e^{2\lambda E + \lambda\sum_{k=1}^{N} S_k} | E, N\right)\right) = \big(e^{2\lambda E}\mathbb{E}\big(\psi(\lambda)^N | E\big)\big) \\ &= \mathbb{E}\Big(e^{2\lambda E} e^{aE(\psi(\lambda)-1)}\Big) = \frac{b}{b + a - 2\lambda - a\psi(\lambda)},\end{aligned}$$

yielding the formula. □

More can be said about the structure of an excursion starting from $(+1, 0)$. This will allow the following quantity to be expressed in terms of standard random variables:

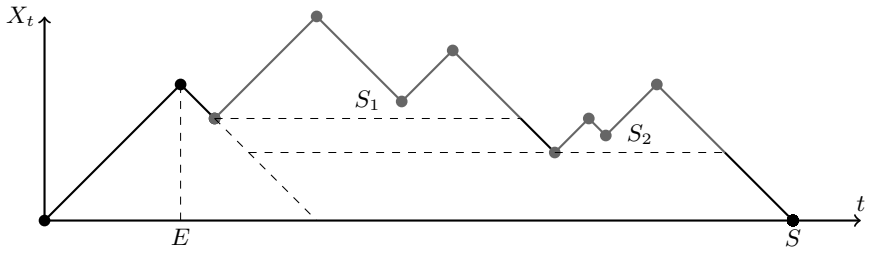


Fig. 23.2. Decomposition of an excursion. The dots mark jumps in velocity.

$$I = \int_0^S e^{\lambda X_s} g(V_s)\, ds$$

where g is a function defined on $\{-1, +1\}$. The first jump time E is exponential with parameter b. Conditionally on $\{E = x\}$, the number N of excursions "grafted" onto the return to 0 has the Poisson distribution with parameter ax. Conditionally on $\{E = x, N = k\}$, the k heights at which they occur are distributed as $(xU_{(k,k)}, xU_{(k,k-1)}, \ldots, xU_{(k,1)})$, where the real random variables $(U_i)_{i \geqslant 1}$ are i.i.d. uniform on $[0, 1]$ and $(U_{(1,k)}, \ldots, U_{(k,k)})$ denotes the sample rearranged in increasing order. Let

$$\Big((V_t^{(i)}, X_t^{(i)})_{0 \leqslant t \leqslant S^{(i)}}\Big)_{i \geqslant 1}$$

be independent paths over an excursion starting from $(+1, 0)$. Then

$$I = (g(1) + g(-1)) \int_0^E e^{\lambda y}\, dy + \sum_{i=1}^N \int_0^{S_i} e^{\lambda (EU_{(i,N)} + X_s^{(i)})} g(V_s^{(i)})\, ds.$$

We then have

$$\mathbb{E}\left(\int_0^{S_i} e^{\lambda (EU_{(i,N)} + X_s^{(i)})} g(V_s^{(i)})\, ds \,|\, E, N, (U_i)_{i \geqslant 1} \right) = \mathbb{E}(I) e^{\lambda E U_{(i,N)}}.$$

We deduce the following relation:

$$\begin{aligned} \mathbb{E}(I) &= \mathbb{E}\left(\frac{g(1) + g(-1)}{\lambda} (e^{\lambda E} - 1) + \mathbb{E}(I) \sum_{i=1}^N e^{\lambda E U_{(i,N)}} \right) \\ &= \frac{g(1) + g(-1)}{b - \lambda} + \mathbb{E}(I) \mathbb{E}\left(\sum_{i=1}^N e^{\lambda E U_i} \right). \end{aligned}$$

Finally, by independence of the random variables involved,

$$\mathbb{E}\left(\sum_{i=1}^N e^{\lambda E U_i} | E \right) = \mathbb{E}\big(N e^{\lambda E U_1} | E \big)$$

$$= \mathbb{E}\left(\frac{N}{\lambda E}(e^{\lambda E} - 1)|E\right)$$
$$= \frac{a}{\lambda}(e^{\lambda E} - 1).$$

Thus, for every $\lambda < b - a$, we obtain

$$\mathbb{E}(I) = \frac{g(1) + g(-1)}{b - a - \lambda}.$$

The ergodic theorem shows that if ν is the invariant distribution of (V, X), then

$$\int_{\{-1,+1\}\times\mathbb{R}_+} e^{\lambda x} g(v)\nu(dv, dx) = \frac{\mathbb{E}(I)}{\mathbb{E}(S)} = \frac{g(1) + g(-1)}{2}\frac{b - a}{b - a - \lambda}.$$

Thus the invariant distribution ν is the distribution $(1/2)(\delta_{-1}+\delta_{+1})\otimes\mathrm{Exp}(b-a)$.

23.4 Further reading

Many mathematical models of bacterial motion can be found, for example, in the articles by Hans Othmer, Steven Dunbar, and Wolfgang Alt [ODA88], and by Radek Erban and Hans Othmer [EO05]. Theorem 23.4 is due to Marc Kac [Kac74]. In Kac's article, the partial differential equation in Theorem 23.4 appears as a scaling limit of persistent random walks, which model the evolution of an electric current in a coaxial cable, explaining the name telegraph equation. This connection with persistent random walks is also central to an article by Samuel Herman and Pierre Vallois [HV10].

24

The Dirichlet problem

Keywords. Diffusion; Brownian motion; exit time; heat equation; harmonic function.

Tools. Markov chain; Markov property; martingale; Monte Carlo method.

Difficulty. ***

This chapter is devoted to the continuous Dirichlet problem and its connection with Brownian motion. The discrete version is discussed in Chapter 2 on random walks. The motivation is to study the temperature distribution in a solid. Consider a homogeneous solid identified with the closure $\overline{D}$ of a bounded connected open subset D of $\mathbb{R}^d$, whose possibly nonconstant surface temperature is fixed at a function b. The temperature inside the solid converges to a thermal equilibrium that we wish to express in terms of D and b. Intuitively, the temperature at an interior point of the solid is the average of the temperatures over any ball centered at that point and contained in the solid: the temperature φ satisfies the mean value formula. For every $x \in \mathbb{R}^d$ and every real $r \geqslant 0$, the ball and sphere with center x and radius r are denoted by

$$B(x,r) := \{y \in \mathbb{R}^d : |x-y| < r\} \quad \text{et} \quad S(x,r) := \{y \in \mathbb{R}^d : |x-y| = r\},$$

whereas the closed ball with center x and radius r is denoted by

$$\overline{B(x,r)} := B(x,r) \cup S(x,r) = \{y \in \mathbb{R}^d : |x-y| \leqslant r\}.$$

We sometimes write $\partial B(x,r) := S(x,r)$. We say that φ is harmonic on D if it is locally integrable and satisfies the mean value formula:

$$\forall x \in D,\ \forall r < \operatorname{dist}(x, \partial D), \quad \varphi(x) = \frac{1}{|B(0,r)|} \int_{B(0,r)} \varphi(x+y)\, dy.$$

The temperature should be continuous throughout the solid if b is continuous on the boundary ∂D of D. We can therefore formulate the Dirichlet problem as follows: a function φ defined on $\overline{D}$ with values in $\mathbb{R}$ solves the Dirichlet problem if

1. for every point x of ∂D, $\varphi(x) = b(x)$,
2. φ is continuous on $\overline{D}$,
3. φ is harmonic on D.

Many theorems address existence and/or uniqueness of the solution of the Dirichlet problem, but of course the solution is generally not explicit. In certain very special cases, however, one can give a relatively explicit integral representation of the solution. In other cases, deterministic or probabilistic numerical methods provide approximate solutions.

Theorem 24.1 (Harmonic functions)**.** *For every function $\varphi : D \to \mathbb{R}$, the following three properties are equivalent:*

1. *Mean value property on balls: φ is harmonic;*
2. *Mean value property on spheres: φ is locally integrable on D and satisfies*

$$\forall x \in D,\ \forall r < \operatorname{dist}(x, \partial D), \quad \varphi(x) = \int_{S(0,r)} \varphi(x+y)\, \sigma_r(dy),$$

   where σ_r is the uniform distribution on the sphere centered at the origin with radius r;
3. *Vanishing Laplacian: φ is of class $\mathcal{C}^\infty$ on D and $\Delta\varphi = 0$ on D.*

Moreover, if φ is harmonic on D, then all its partial derivatives are harmonic on D. Finally, if $|\varphi|$ is bounded on $\overline{D}$, the same holds for all partial derivatives of φ.

Theorem 24.1 allows the Dirichlet problem to be reformulated in terms of partial differential equations (PDEs).

Corollary 24.2 (The Dirichlet problem and PDEs)**.** *A function φ solves the Dirichlet problem on D with boundary conditions b if and only if it is of class $\mathcal{C}^\infty$ on D, continuous on $\overline{D}$, and satisfies*

$$\begin{cases} \Delta\varphi(x) = 0 & \text{pour tout } x \in D, \\ \varphi(x) = b(x) & \text{pour tout } x \in \partial D. \end{cases}$$

Proof of Theorem 24.1. Let us show that the second property implies the first. Assume without loss of generality that $x = 0$. Changing to polar coordinates gives

$$\begin{aligned} \frac{1}{|B(0,r)|} \int_{B(0,r)} \varphi(y)\, dy &= \frac{1}{r^d |B(0,1)|} \int_0^r \int_{\partial B(0,s)} \varphi(y)\, \sigma_s(dy) |\partial B(0,s)|\, ds \\ &= \frac{d}{r^d} \int_0^r \int_{\partial B(0,s)} \varphi(y)\, \sigma_s(dy) s^{d-1}\, ds. \end{aligned}$$

Thus, if φ satisfies the mean value formula on spheres, then φ satisfies the mean value formula on balls. Conversely, if φ is harmonic, the expression above gives

$$r^d \varphi(0) = d \int_0^r \int_{\partial B(0,s)} \varphi(y)\, \sigma_s(dy) s^{d-1}\, ds.$$

Differentiating with respect to r yields the mean value formula on spheres.

We now show that the last two properties are equivalent. The first step is to smooth the function φ by convolution. Let $x \in D$, let $\mathcal{V}$ be a neighborhood of x contained in D, and let $\varepsilon < \operatorname{dist}(\mathcal{V}, \partial D)$. The function

$$\Psi_\varepsilon : x \mapsto c(\varepsilon) \exp((|x|^2 - \varepsilon^2)^{-1}) \mathbb{1}_{\{|x|<\varepsilon\}}$$

is of class $\mathcal{C}^\infty$, nonnegative, has integral 1 for a suitable choice of $c(\varepsilon)$, is supported in the ball $\overline{B(0,\varepsilon)}$, and is radial. The function $\varphi * \Psi_\varepsilon$ is therefore of class $\mathcal{C}^\infty$ on $\mathcal{V}$. Moreover, for every $y \in \mathcal{V}$, the mean value formula on spheres ensures that

$$\begin{aligned}
\varphi * \Psi_\varepsilon(y) &= \int_{B(0,\varepsilon)} \varphi(y+u) \Psi_\varepsilon(u) du \\
&= \int_0^\varepsilon \left(\int_{S(0,s)} \varphi(y+sz)\, \sigma_s(dz) \right) |S(0,s)| \Psi_\varepsilon(s) ds \\
&= \varphi(y) \int_{B(0,\varepsilon)} \Psi_\varepsilon(z)\, dz,
\end{aligned}$$

since Ψ_ε is radial. It also has integral 1: φ and $\varphi * \Psi_\varepsilon$ coincide on $\mathcal{V}$, so φ is of class $\mathcal{C}^\infty$ on $\mathcal{V}$. Finally, for sufficiently small $|y|$, we have

$$\varphi(x+y) = \varphi(x) + \nabla\varphi(x) \cdot y + \frac{1}{2} (\operatorname{Hess}(\varphi)(x) y) \cdot y + o(|y|^2).$$

Integrating this relation over the sphere centered at the origin with radius r against the uniform measure and letting r tend to 0 yields $\Delta\varphi(x) = 0$, while the converse follows by integrating the relation over a small ball.

It remains to establish the automatic regularity properties of the derivatives. Harmonicity propagates by induction, differentiating the mean value formula. Moreover, for every multi-index $\alpha \in \mathbb{N}^d$, there exists $K(\varepsilon, \alpha)$ such that, for every $y \in \mathcal{V}$,

$$|\partial_\alpha \varphi(y)| = |\varphi * \partial_\alpha \Psi_\varepsilon| \leqslant K(\varepsilon, \alpha) \sup_{z \in \overline{D}} |\varphi(z)|.$$

A compactness argument yields the result. □

24.1 Probabilistic interpretation

Probability theory provides the tools needed to study the Dirichlet problem in PDE form, including a probabilistic representation of its solution. Let $(B_t)_{t \geqslant 0}$ be a Brownian motion on $\mathbb{R}^d$, and let

$$\tau := \inf\{t > 0 \,:\, B_t \notin D\}$$

be the *exit time* from D for $(B_t)_{t\geqslant 0}$. Write $\mathbb{P}_x$ for the distribution of $(B_t)_{t\geqslant 0}$ given $B_0 = x$ and $\mathbb{E}_x$ for expectation with respect to this distribution. Brownian motion is a strong Markov process, and Itô's formula ensures that if f is a function of class $\mathcal{C}^2$ from $\mathbb{R}^d$ to $\mathbb{R}$ with bounded derivatives, then

$$f(B_t) = f(B_0) + \int_0^t \frac{1}{2}\Delta f(B_s)\, ds + \int_0^t \nabla f(B_s)\, dB_s.$$

The stochastic integral (the last term) above is a martingale. The key link between probability and analysis is the following observation: $(f(B_t))_{t\geqslant 0}$ is a martingale if f is harmonic.

Theorem 24.3 (Harmonicity of the probabilistic representation)**.** *The function v defined on D by $v(x) = \mathbb{E}_x[b(B_\tau)]$ is harmonic on D.*

Establishing continuity on $\overline{D}$ of the function v given by Theorem 24.3 is delicate. It involves the geometric properties of the domain D.

Proof. Let $x \in D$ and $r < \operatorname{dist}(x, \partial D)$. Write

$$\tau_r := \inf\{t \geqslant 0 \,:\, B_t \notin B(x, r)\}$$

for the exit time from the ball with center x and radius r for Brownian motion starting from x. See Figure 24.1. The strong Markov property of $(B_t)_{t\geqslant 0}$ then ensures that

$$v(x) = \mathbb{E}_x[\mathbb{E}_x(b(B_\tau)|\mathcal{F}_{\tau_r})] = \mathbb{E}_x\big[\mathbb{E}_{X_{\tau_r}}(b(B_\tau))\big] = \mathbb{E}_x(v(B_{\tau_r})).$$

Finally, since the distribution of Brownian motion is rotation invariant, the random variable B_{τ_r} is uniform on $S(x, r)$. This property ensures that v satisfies the mean value formula on spheres. □

Remark 24.4 (Poisson kernel)**.** *The probabilistic representation above shows that the solution of the Dirichlet problem at a point x of D is the average of the boundary values weighted by the distribution of the exit location of Brownian motion starting from x. The density of this measure is called the Poisson kernel of D.*

A function h harmonic on a domain D necessarily attains its maximum on a compact set K at a boundary point of K. This maximum principle yields uniqueness of the solution of the Dirichlet problem: the difference of two solutions is harmonic and vanishes on ∂D, hence on D. The next theorem shows that uniqueness of the solution of the Dirichlet problem also follows from its probabilistic interpretation.

Theorem 24.5 (Uniqueness)**.** *If φ solves the Dirichlet problem on D with boundary condition b, then this solution is unique and is given by*

$$\forall x \in D, \quad \varphi(x) = \mathbb{E}_x[b(B_\tau)].$$

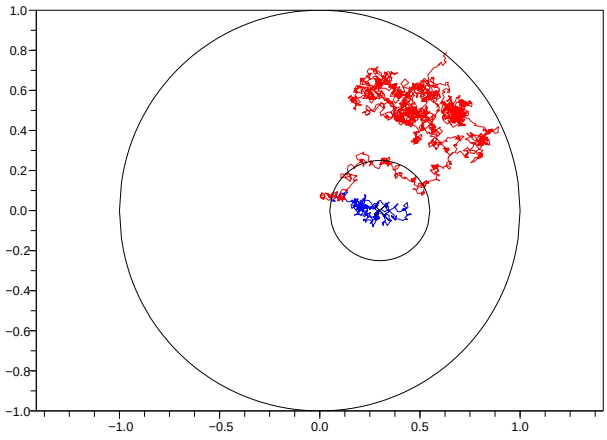


Fig. 24.1. Decomposition of a path leaving D after leaving $B(x,r)$.

Proof. Itô's formula gives the following decomposition:

$$\varphi(B_{t\wedge\tau}) = \varphi(x) + \int_0^{t\wedge\tau} \frac{1}{2}\Delta\varphi(B_s)\,ds + M_{t\wedge\tau}$$

where $(M_{t\wedge\tau})_{t\geqslant 0}$ is a bounded martingale by the automatic regularity properties of harmonic functions (Theorem 24.1). Indeed, φ is harmonic on D and bounded on $\overline{D}$. The same then holds for its derivatives. Taking expectations and letting $t \to \infty$ yields the claimed result. □

24.2 The unit ball

There are very few domains D for which an explicit expression for the solution of the Dirichlet problem is available. This section addresses Euclidean balls.

Theorem 24.6 (Poisson kernel of balls)**.** *The Dirichlet problem on the ball centered at the origin with radius r in $\mathbb{R}^d$ and boundary condition b has solution*

$$\varphi(x) = \int_{S(0,r)} b(y) P_r(x,y)\,\sigma_r(dy), \quad \text{où} \quad P_r(x,y) = \frac{r^{d-2}(r^2 - |x|^2)}{|y-x|^d}.$$

Here again, $\varphi(x)$ is the average of b over ∂D against the density $P_r(x,y)$ with respect to the uniform measure on $S(0,r)$. This measure assigns more weight to points on the sphere close to x, in agreement with the temperature interpretation.

The Dirichlet problem on a ball in $\mathbb{R}^d$ thus reduces to computing an integral over the sphere of dimension $d-1$. When $d \gg 1$, deterministic approximation methods become very difficult to implement because they require discretizing the sphere. The Monte Carlo method remains very simple to use, thanks to the following result.

Theorem 24.7 (Simulating the uniform distribution on the sphere)**.** *If X is a random vector in $\mathbb{R}^d$ with distribution $\mathcal{N}_d(0, I_d)$, the random vector $Y = X/|X|$ is uniform on $S(0,1)$.*

Proof. The random variable Y is defined almost surely because the set $\{X = 0\}$ has measure zero. Moreover, the distribution of X is invariant under every rotation: if O is an orthogonal matrix, then OX has the same distribution as X. The same therefore holds for Y. The only rotation-invariant probability distribution on the unit sphere equipped with its Borel sigma-algebra is the uniform distribution. □

Remark 24.8 (Box-Muller algorithm)**.** *If U and V are independent real random variables with U uniform on $[0,1]$ and V exponential with parameter $1/2$, then the real random variables*

$$X = \sqrt{V}\cos(2\pi U) \quad et \quad Y = \sqrt{V}\sin(2\pi U)$$

are independent with the same standard Gaussian distribution $\mathcal{N}(0,1)$. This property was long used in scientific computing software before being replaced by more efficient procedures such as Marsaglia's Ziggurat algorithm, based on a polygonal approximation method.

24.3 Numerical approaches

For a general domain D, no explicit expression for the Poisson kernel is available. Approximate solution methods must therefore be considered. To this end, we discretize D. Let $h > 0$ be the mesh size. Define the grid D_h as the intersection of D and the lattice $h\mathbb{Z}^d$:

$$D_h := D \cap h\mathbb{Z}^d = \big\{x \in D,\ \exists (i_1, \dots, i_d) \in \mathbb{Z}^d,\ x = (hi_1, \dots, hi_d)\big\}.$$

We say that $x, y \in h\mathbb{Z}^d$ are neighbors, and write $x \sim y$, if $|x - y| = h$. The boundary of D_h is

$$\partial_h D_h = \{y \in D_h^c\ :\ \exists x \in D_h,\ x \sim y\}.$$

The Laplace operator is replaced by the difference operator:

$$\Delta_h f(x) = \frac{1}{2d} \sum_{y \sim x} (f(y) - f(x)).$$

We say that φ_h is h-harmonic on D_h if $\Delta_h \varphi_h = 0$ on D_h. We also need to define a function b_h on $\partial_h D_h$ that reasonably approximates b defined on ∂D. The discrete Dirichlet problem is then: find φ_h on $D_h \cup \partial_h D_h$ such that

$$\begin{cases} \Delta_h \varphi_h(x) = 0 & \text{pour tout } x \in D_h, \\ \varphi_h(x) = b_h(x) & \text{pour tout } x \in \partial_h D_h. \end{cases}$$

This is exactly the discrete Dirichlet problem studied in Chapter 2, which has a probabilistic solution in the same spirit as the continuous Dirichlet problem, with the simple random walk replacing Brownian motion. This substitution is justified by the following observation. If f is a function of class $\mathcal{C}^2$ from $\mathbb{R}^d$ to $\mathbb{R}$, then

$$\Delta_h f(x) = \frac{h^2}{2d} \Delta f(x) + o(h^2).$$

Thus, for small h, Δ_h approximates the Laplacian. Moreover, saying that $\Delta_h \varphi_h(x) = 0$ means that $\varphi_h(x)$ is the average (under the uniform measure) of the values of φ_h at the $2d$ neighbors of x, an analogue of the mean value property.

Remark 24.9 (Deterministic and probabilistic approaches)**.** *Since D_h is finite, the discrete Dirichlet problem is a linear system of* $\operatorname{card}(D_h)$ *equations in* $\operatorname{card}(D_h)$ *unknowns, with a unique solution. Solving it is simplified by the sparsity of the associated matrix: each equation involves at most $2d+1$ unknowns. Once the calculations are complete, this direct method simultaneously determines the value of the solution φ_h at every point of the grid D_h. This method pays off when the temperature is needed at every point of the solid.*

24.4 The upper half-plane: from discrete to continuous

This section illustrates, in a special case, convergence of the solution of the discrete Dirichlet problem to that of the continuous Dirichlet problem formulated in Corollary 24.2. Although unbounded, the upper half-plane in $\mathbb{R}^2$ has an explicit Poisson kernel. It is also possible to determine the distribution of the exit location of the random walk on the lattice in $h\mathbb{Z}^2$. Let $N \in \mathbb{N}^*$. Choose the mesh size in the form $h = 1/N$. Denote by $(X_n^N, Y_n^N)_n$ the random walk on $h\mathbb{Z}^2$ starting from $(X_0^N, Y_0^N) = (0, 1)$ such that the sequences of random variables $(X_{n+1}^N - X_n^N)$ and $(Y_{n+1}^N - Y_n^N)$ are independent of one another, and each consists of independent variables with common distribution:

$$\mathbb{P}((X_{n+1}^N - X_n^N) = -1/N) = \mathbb{P}((X_{n+1}^N - X_n^N) = 1/N) = 1/2,$$
$$\mathbb{P}((Y_{n+1}^N - Y_n^N) = -1/N) = \mathbb{P}((Y_{n+1}^N - X_n^Y) = 1/N) = 1/2.$$

At each step, the walk chooses randomly among the 4 points at distance $\sqrt{2}/N$ (northwest, northeast, southwest, and southeast). This slight modification of the random walk dynamics ensures that $(X_n^N)_n$ and $(Y_n^N)_n$ are

independent simple random walks, which will simplify the subsequent argument.

We study the exit time and horizontal exit coordinate from the upper half-plane for the random walk thus defined. The exit time is the real random variable

$$T^N := \inf\big\{k \geqslant 0 \,:\, Y_k^N = 0\big\}.$$

The horizontal exit coordinate U^N is the horizontal coordinate of the random walk at time T^N,

$$U^N := X_{T^N}^N.$$

Theorem 24.10 (Convergence in distribution of the exit location)**.** *The sequence $(U^N)_{N\geqslant 1}$ converges in distribution, as N tends to infinity, to the Cauchy distribution with density:*

$$x \in \mathbb{R} \mapsto \frac{1}{\pi}\frac{1}{1+x^2}.$$

Proof. Denote by S_k the hitting time of 0 for the simple random walk $(Z_n)_{n\geqslant 0}$ on $\mathbb{Z}$ starting from k, and by G_k its generating function. The strong Markov property ensures that

$$S_k \overset{\text{loi}}{=} S_1^1 + \cdots + S_1^k$$

where $(S_1^i)_{1\leqslant i\leqslant k}$ are i.i.d. with the same distribution as S_1, hence $G_k(s) = G_1(s)^k$. Similarly,

$$S_1 \overset{\text{loi}}{=} 1 + ZS_2$$

where Z is a random variable with Bernoulli distribution $\mathrm{Ber}(1/2)$, independent of S_2, hence

$$G_1(s) = \frac{s}{2} + \frac{s}{2}G_2(s), \qquad s \in [0,1[.$$

For every $s \in [0,1[$, $G_1(s)$ solves $sx^2 - 2x + s = 0$ and therefore equals

$$G_1(s) = \frac{1-\sqrt{1-s^2}}{s}$$

since $G_1(s) < 1$. In conclusion, T^N has the same distribution as S_N, and its generating function G_N is given by $G_N(s) = G_1(s)^N$.

We now determine the characteristic function of U^N. For every $t \in \mathbb{R}$,

$$\varphi_{U^N}(t) = \mathbb{E}\Big(e^{itU^N}\Big) = \left(\frac{1-\sqrt{1-\cos(t/N)^2}}{\cos(t/N)}\right)^N,$$

since the characteristic function of X_n^N equals $\cos(t/N)^n$. Finally,

$$\varphi_{U^N}(t) = \left(\frac{1-|\sin(t/N)|}{\cos(t/N)}\right)^N \underset{N\to\infty}{\sim} \left(1-\frac{|t|}{N}\right)^N \xrightarrow[N\to\infty]{} e^{-|t|},$$

and the limit is the characteristic function of the standard Cauchy distribution. □

Remark 24.11 (Poisson kernel of the half-plane)**.** *The Dirichlet problem on* $\mathbb{R} \times \mathbb{R}_+^*$ *has the following representation: for every* $(x, y) \in \mathbb{R} \times \mathbb{R}_+^*$,

$$u(x,y) = \int_{\mathbb{R}} b(z) P_H(x,y,z)\, dz, \quad \textit{avec} \quad P((x,y),z) = \frac{1}{\pi} \frac{y}{y^2 + (x-z)^2}.$$

24.5 Recurrence, transience, and harmonic functions

Knowing the radial harmonic functions on $\mathbb{R}^d$ allows us to obtain recurrence and transience properties of Brownian motion. For $r < |x| < R$, write

$$T_{B(0,r)} := \inf \{t \geqslant 0 \,:\, B_t \in B(0,r)\}$$

and

$$\tau_{B(0,R)} := \inf \{t \geqslant 0 \,:\, B_t \notin B(0,R)\}.$$

The time $S := T_{B(0,r)} \wedge \tau_{B(0,R)}$ is the exit time from the annulus $C(0,r,R)$.

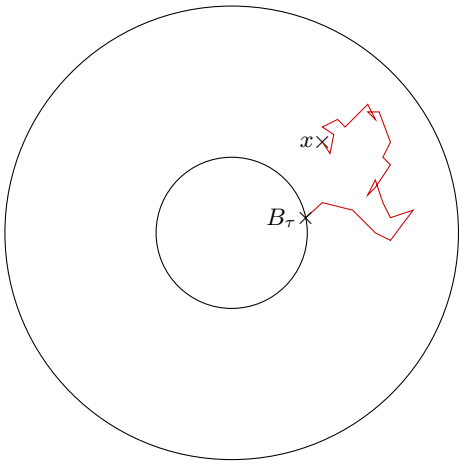


Fig. 24.2. A Brownian path starting from x, stopped upon leaving an annulus.

Theorem 24.12 (Recurrence or transience of Brownian motion)**.** *Let* $(B_t)_{t \geqslant 0}$ *be a Brownian motion on* $\mathbb{R}^d$.

1. *If* $d = 2$ *and* $r < |x| < R$, *then*
   a) $\mathbb{P}_x(T_{B(0,r)} < \tau_{B(0,R)}) = \dfrac{\log(R) - \log(|x|)}{\log(R) - \log(r)}$,
   b) $\mathbb{P}_x(T_{\{0\}} < \infty) = 0$,
   c) $\mathbb{P}_x(T_{B(0,r)} < \infty) = 1$,

d) $\mathbb{P}_x(T_{B(0,r)} < \infty \ i.s.) = 1$.

2. *If* $d \geqslant 3$ *and* $r < |x| < R$, *then*

a) $\mathbb{P}_x(T_{B(0,r)} < \tau_{B(0,R)}) = \dfrac{|x|^{2-d} - R^{2-d}}{r^{2-d} - R^{2-d}}$,

b) $\mathbb{P}_x(T_{\{0\}} < \infty) = 0$,

c) $\mathbb{P}_x(T_{B(0,r)} < \infty) = \dfrac{|x|^{2-d}}{r^{2-d}}$.

Theorem 24.12 establishes recurrence of Brownian motion in $\mathbb{R}^2$: although it does not hit any given point, it almost surely visits every neighborhood of that point infinitely often! This no longer holds in higher dimensions. The behavior is similar to that of the simple symmetric random walk on $\mathbb{Z}^d$, studied in Chapter 2.

Proof. Let φ be the function $x \mapsto \log(|x|)$, respectively $x \mapsto |x|^{2-d}$, for $d = 2$, respectively $d \geqslant 3$. In both cases, φ is a radial harmonic function on $\mathbb{R}^d$ with the origin removed. The process $(\varphi(B_{t\wedge S}))_{t\geqslant 0}$ is a bounded martingale. Moreover, S is finite almost surely because the first coordinate of (B_t) takes arbitrarily large values. The optional stopping theorem therefore ensures that

$$\varphi(x) = \mathbb{E}_x(\varphi(B_S)) = \varphi(r)\mathbb{P}_x\big(T_{B(0,r)} < \tau_{B(0,R)}\big) + \varphi(R)\mathbb{P}_x\big(T_{B(0,r)} > \tau_{B(0,R)}\big)$$

which can also be written

$$\mathbb{P}_x\big(T_{B(0,r)} < \tau_{B(0,R)}\big) = \frac{\varphi(x) - \varphi(R)}{\varphi(r) - \varphi(R)}.$$

To obtain 1.(b) and 2.(b), take the successive limits $r \to 0$ and then $R \to \infty$. Assertion 1.(d) follows from 1.(c) and the strong Markov property. □

24.6 Further reading

The continuous Dirichlet problem was studied in the nineteenth century and is named after the mathematician Peter Gustav Lejeune Dirichlet. Its probabilistic solution, due to Shizuo Kakutani, dates from the mid-twentieth century. The continuous Dirichlet problem lies at the intersection of probability (Brownian motion), potential theory, harmonic analysis, and partial differential equations, an intersection extensively developed by Paul Lévy and Joseph Leo Doob during the twentieth century. Various aspects are treated, in particular, in Richard Bass's book [Bas95] and in the book by Peter Mörters and Yuval Peres [MP10]. The discretized version provides an introduction to numerical matrix analysis (sparse matrices) and numerical analysis of partial differential equations, as well as stochastic numerical analysis, such as stochastic Euler schemes (see Section 27.6). The explicit solution of the Dirichlet problem for a ball can be found, for example, in Walter Rudin's book [Rud87]. Existence and continuity on $\overline{D}$ are discussed in Richard Bass's book [Bas95]. Our Section 24.4 is inspired by Bernard Ycart's book [Yca02]. Most results extend to differential operators far beyond the usual Laplacian.

25

The Ornstein-Uhlenbeck process

Keywords. Diffusion; Brownian motion; Gaussian process; Ornstein-Uhlenbeck process; stochastic differential equation; partial differential equation.

Tools. Gaussian vectors; Markov semigroup; Itô's formula; Hilbert space analysis; Laplace transform; Hermite polynomials; coupling; relative entropy; Poincaré inequality; logarithmic Sobolev inequality; Wasserstein distance.

Difficulty. ***

The Ornstein-Uhlenbeck process $(X_t)_{t\geqslant 0}$ on $\mathbb{R}$ is the solution of the following linear stochastic differential equation (SDE):

$$X_t = X_0 - \int_0^t X_s \, ds + \sqrt{2} B_t,$$

where $(B_t)_{t\geqslant 0}$ is a Brownian motion that is standard on $\mathbb{R}$ and independent of X_0. The factor $\sqrt{2}$ in the SDE makes some formulas neater. We also use the following abbreviated version of the SDE:

$$dX_t = -X_t \, dt + \sqrt{2} \, dB_t.$$

The Ornstein-Uhlenbeck process is the simplest ergodic Gaussian diffusion process on $\mathbb{R}$. Its remarkable properties make it indispensable. In particular, it appears as a scaling limit of several Markov chains with discrete state spaces, such as the Ehrenfest chain (Chapters 9 and 27) or the M/M/∞ queue (Chapter 11). It also enters into the definition of the Langevin process in Chapter 26.

25.1 Basic properties

Theorem 25.1 (Rescaling)**.** *Let $\lambda > 0$ and $\sigma > 0$. If X is an Ornstein-Uhlenbeck process then the process Y defined by*

$$Y_t = \frac{\sigma}{\sqrt{2\lambda}} X_{\lambda t}$$

solves the stochastic differential equation

$$Y_t = Y_0 - \lambda \int_0^t Y_s \, ds + \sigma W_t$$

where $(W_t)_{t \geqslant 0}$ *is a standard Brownian motion.*

Proof. For every $t \geqslant 0$, writing $W_t = B_{\lambda t}/\sqrt{\lambda}$,

$$\begin{aligned} \frac{\sigma}{\sqrt{2\lambda}} X_{\lambda t} &= \frac{\sigma}{\sqrt{2\lambda}} X_0 - \int_0^{\lambda t} \frac{\sigma}{\sqrt{2\lambda}} X_s \, ds + \frac{\sigma}{\sqrt{\lambda}} B_{\lambda t} \\ &= Y_0 - \lambda \int_0^t Y_s \, ds + \sigma W_t, \end{aligned}$$

and $(W_t)_{t \geqslant 0}$ is a standard Brownian motion (scaling invariance). □

Theorem 25.2 (Gaussian process)**.** *If* $X_0 = x$*, then*

$$X_t \sim \mathcal{N}(xe^{-t}, 1 - e^{-2t}).$$

More generally, the process $(X_t)_{t \geqslant 0}$ *is a Gaussian process whose covariance is given, for all* $0 \leqslant s \leqslant t$*, by*

$$\mathrm{Cov}(X_s, X_t) = (1 - e^{-2s}) e^{-(t-s)}.$$

This property actually follows from linearity of the stochastic differential equation satisfied by $(X_t)_{t \geqslant 0}$*.*

Proof. Itô's formula applied to the process $(e^t X_t)_{t \geqslant 0}$ with $X_0 = x$ gives

$$X_t = xe^{-t} + \sqrt{2} \int_0^t e^{s-t} dB_s.$$

We deduce that $X_t \sim \mathcal{N}(xe^{-t}, 1 - e^{-2t})$. To compute the covariance of X_s and X_t, similarly write

$$X_t = X_s e^{-(t-s)} + Y_t$$

where Y_t is a centered random variable independent of X_s. □

Let $\mathcal{A}$ be the set of functions of class $\mathcal{C}^\infty$ on $\mathbb{R}$ all of whose derivatives have at most polynomial growth[1]. The Ornstein-Uhlenbeck semigroup on $\mathbb{R}$ is the family $(N_t)_{t \in \mathbb{R}_+}$ of operators from $\mathcal{A}$ to $\mathcal{A}$ defined, for all $t \geqslant 0, f \in \mathcal{A}$, by

[1] A function $f : \mathbb{R} \to \mathbb{R}$ has at most polynomial growth if $|f| \leqslant P$ for some polynomial P.

$$N_t(f)(x) = \mathbb{E}(f(X_t) \,|\, X_0 = x).$$

The preceding results yield Mehler's formula: for every $t \geqslant 0$ and every $f \in \mathcal{A}$,

$$N_t(f)(x) = \int_{\mathbb{R}} f\Big(e^{-t}x + \sqrt{1-e^{-2t}}y\Big)\,\gamma(dy),$$

where γ denotes standard Gaussian measure on $\mathbb{R}$. This formula shows that $N_t(f)(x)$ is well defined for every $x \in \mathbb{R}$ and every $f \in L^1(\mathcal{N}(xe^{-t}, 1-e^{-2t}))$.

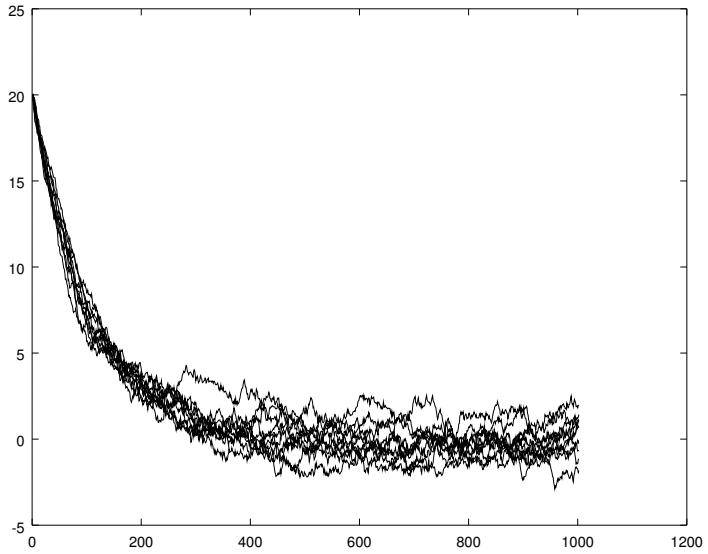


Fig. 25.1. Sample paths of the Ornstein-Uhlenbeck process starting from the same initial point far from the origin. Initially, drift dominates diffusion, quickly bringing the path back toward the origin and then keeping it nearby, where diffusion causes fluctuations. At large times the invariant distribution prevails.

Theorem 25.3 (Semigroup property)**.** *The family* $(N_t)_{t\geqslant 0}$ *satisfies the following properties:*

1. *for every function* $f \in \mathcal{A}$, $N_0 f = f$,
2. *for all* $s, t \geqslant 0$ *and* $f \in \mathcal{A}$, $(N_s \circ N_t)(f) = N_{t+s} f$.

Proof. The first assertion is obvious. For the second, write $\sigma_t = \sqrt{1-e^{-2t}}$. Let X and Y be independent real random variables with distribution γ. By definition of N_t,

$$(N_s \circ N_t)(f)(x) = \mathbb{E}\big[N_t(f)(e^{-s}x + \sigma_s X)\big]$$

$$= \mathbb{E}\big[f(e^{-t}(e^{-s}x + \sigma_s X) + \sigma_t Y)\big]$$
$$= \mathbb{E}\big[f(e^{-s-t}x + e^{-t}\sigma_s X + \sigma_t Y)\big].$$

It remains only to observe that

$$e^{-t}\sigma_s X + \sigma_t Y \sim \mathcal{N}(0, \sigma_{s+t}^2)$$

and use Mehler's formula. □

We now examine the infinitesimal evolution over a very short time.

Theorem 25.4 (Infinitesimal generator)**.** *For every $f \in \mathcal{A}$ and every $x \in \mathbb{R}$,*

$$\lim_{t\to 0^+} \frac{N_t f(x) - N_0 f(x)}{t} = (Af)(x) \quad o\grave{u} \quad (Af)(x) = f''(x) - xf'(x).$$

More generally, for every $f \in \mathcal{A}$ and every $t \geqslant 0$, the following partial differential equations hold:

$$\partial_t N_t f = A(N_t f) = N_t(Af).$$

This partial differential equation (PDE) is sometimes called the heat equation with drift. The operator A is the infinitesimal generator of the Ornstein-Uhlenbeck semigroup. It is useful to keep in mind the formal identity $N_t = e^{tA}$, which shows that $N_0 = I$ and $\partial_t N_t = AN_t$.

Proof. If $f \in \mathcal{A}$, Taylor's formula of order 2 with bounded remainder near x gives

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \mathcal{O}(h^3).$$

First observe that, for $h = (e^{-t} - 1)x + \sigma_t y$,

$$\int h\,\gamma(dy) = (e^{-t} - 1)x,$$
$$\int h^2\,\gamma(dy) = (e^{-t} - 1)^2 x^2 + \sigma_t^2,$$
$$\int |h|^3\,\gamma(dy) = \mathcal{O}(t^{3/2}).$$

The expansion above ensures that

$$N_t f(x) = f(x) + (e^{-t} - 1)xf'(x) + \frac{1}{2}\big((e^{-t} - 1)^2 x^2 + \sigma_t^2\big)f''(x) + \mathcal{O}(t^{3/2}).$$

This indeed gives the claimed limit:

$$\lim_{t\to 0} \frac{1}{t}(N_t f(x) - f(x)) = f''(x) - xf'(x).$$

This result also follows from Itô's formula: if $X_0 = x$ and $t \geqslant 0$, then

$$f(X_t) = f(x) + \int_0^t Af(X_s)\,ds + \sqrt{2}\int_0^t f'(X_s)\,dB_s,$$

and since the stochastic integral is a martingale, taking expectations gives

$$N_t f(x) = f(x) + \int_0^t N_s(Af)(x)\,ds.$$

It remains to use $N_s(Af)(x) \to Af(x)$ as $s \to 0$ to obtain the first assertion of the theorem. The second follows from the first by the semigroup property. □

By Mehler's formula, for every $x \in \mathbb{R}$ and $f \in \mathcal{A}$,

$$N_t f(x) \underset{t\to\infty}{\longrightarrow} \int f d\gamma.$$

Moreover, again by Mehler's formula, for all $t \geqslant 0$ and $f \in \mathcal{A}$,

$$\int N_t f\,d\gamma = \iint f(xe^{-t} + \sqrt{1-e^{-2t}}y)\,\gamma(dy)\,\gamma(dx) = \int f\,d\gamma.$$

In other words, the random variable X_t has the same distribution as

$$X_0 e^{-t} + \sqrt{1-e^{-2t}}Z$$

where $Z \sim \mathcal{N}(0,1)$ is a random variable independent of X_0. Thus if $X_0 \sim \gamma$, then X_t and X_0 have the same distribution $\mathcal{N}(0,1)$, and for all $t \geqslant 0$ and $f \in \mathcal{A}$,

$$\int N_t f\,d\gamma = \mathbb{E}[(N_t f)(X_0)] = \mathbb{E}(f(X_t)) = \mathbb{E}(f(X_0)) = \int f\,d\gamma.$$

The measure γ is therefore the invariant distribution of the Ornstein-Uhlenbeck process. The next result goes beyond invariance.

Theorem 25.5 (Symmetry and reversibility)**.** *The operators $(N_t)_{t\geqslant 0}$ and A are self-adjoint in $L^2(\gamma)$. In particular, for all functions $f, g \in \mathcal{A}$,*

$$\int Afg\,d\gamma = \int fAg\,d\gamma = -\int f'g'\,d\gamma.$$

The invariant distribution γ is reversible: if X_0 has distribution γ, then for every $T \geqslant 0$, the reversed process $(X_{T-t})_{t\in[0,T]}$ has the same distribution as $(X_t)_{t\in[0,T]}$.

Proof. By definition of the Ornstein-Uhlenbeck semigroup,

$$\int fN_t g\,d\gamma = \mathbb{E}\big[f(X)g\big(e^{-t}X + \sigma_t Y\big)\big],$$

where X and Y are independent random variables with distribution γ. The pair $(X, e^{-t}X + \sigma_t Y)$ is a centered Gaussian vector with covariance matrix

$$\begin{pmatrix} 1 & e^{-t} \\ e^{-t} & 1 \end{pmatrix}.$$

We may therefore interchange f and g in the preceding identity, showing that N_t is self-adjoint in $L^2(\gamma)$. Finally, integration by parts yields the formula in the theorem. □

Remark 25.6 (Simulation)**.** *The Ornstein-Uhlenbeck process has independent Gaussian increments (which are not stationary). Simulating its paths is particularly simple. Let $(t_n)_{n\geqslant 0}$ be a strictly increasing sequence of nonnegative real numbers with $t_0 := 0$. The discrete-time process $(X_{t_n})_{n\geqslant 1}$ has the same distribution as the autoregressive process $(Z_n)_{n\geqslant 0}$ defined by $Z_0 = X_0$ and*

$$Z_{t_{n+1}} = e^{-(t_{n+1}-t_n)} Z_{t_n} + \sqrt{1 - e^{-2(t_{n+1}-t_n)}}\varepsilon_n$$

for every $n \geqslant 1$, where $(\varepsilon_n)_{n\geqslant 0}$ are i.i.d. real random variables with distribution $\mathcal{N}(0,1)$, independent of Z_0. This process is a discrete-time Markov chain with state space $\mathbb{R}$ and invariant distribution $\mathcal{N}(0,1)$. This observation gives a way to simulate the paths of the Ornstein-Uhlenbeck process, illustrated in Figure 25.1. Conditionally on Z_0, the random vector $(Z_1, \ldots, Z_n)$ is Gaussian, and the Cholesky decomposition of its covariance matrix provides another efficient way to simulate the process's path.

The properties above show that the distribution of X_t is absolutely continuous with a smooth density whenever $t > 0$, regardless of X_0.

Theorem 25.7 (Evolution equation for the probability density)**.** *Let ν_t be the distribution of X_t. For every ν_0 and every $t > 0$, the law ν_t has a density of class $\mathcal{C}^\infty$ with respect to Lebesgue measure. Moreover, if ν_0 is absolutely continuous with density v_0 with respect to the measure γ, then so is ν_t, and its density v_t with respect to γ solves the partial differential equation*

$$\partial_t v_t(x) = A v_t(x).$$

Proof. If g is smooth, the symmetry property of N_t gives

$$\begin{aligned} \mathbb{E}g(X_t) &= \int \mathbb{E}(g(X_t)|X_0 = x) v_0(x) \gamma(dx) \\ &= \int N_t g(x) v_0(x) \gamma(dx) \\ &= \int g(x) N_t(v_0)(x) \gamma(dx). \end{aligned}$$

By definition, $v_t = N_t(v_0)$. By definition of N_t, $t \mapsto N_t(v_0)(x)$ is of class $\mathcal{C}^\infty$ (even if v_0 is not). Theorem 25.4 concludes the proof. □

Remark 25.8 (Fokker-Planck PDE)**.** *A change of variables shows that the density u_t of the distribution of X_t with respect to Lebesgue measure solves the following partial differential equation:*

$$\partial_t u_t(x) = \partial^2_{xx} u_t(x) + \partial_x(x u_t(x)).$$

This partial differential equation is sometimes called the Fokker-Planck equation. It expresses the time evolution of a spatial density, as do Euler's equation in fluid mechanics and the Boltzmann equation in the kinetic theory of gases.

25.2 Spectral decomposition and the Poincaré inequality

In this section, we study the action of the semigroup $(N_t)_{t\geqslant 0}$ on polynomials. Polynomials form a dense subspace of $L^2(\gamma)$. Indeed, let $f \in L^2(\gamma)$. The Laplace transform of the signed measure $\nu(dx) = f(x)\gamma(dx)$ is finite on $\mathbb{R}$ since, for $\lambda \in \mathbb{R}$, we have

$$\left| \int e^{\lambda x}\, \nu(dx) \right|^2 \leqslant \int f^2\, d\gamma \int e^{2\lambda x}\, \gamma(dx) < +\infty.$$

The Laplace transform of ν is therefore analytic in a neighborhood of 0. If f is orthogonal to all polynomials, every derivative of this transform vanishes at 0, and hence the transform itself vanishes. This implies that f vanishes γ-almost surely. We can therefore find an orthonormal basis of $L^2(\gamma)$ consisting of polynomials. Theorem 21.7 is useful here.

The Hermite polynomials $(H_n)_{n\geqslant 0}$ are defined by their generating series:

$$G(s,x) = e^{sx - \frac{1}{2}s^2} = \sum_{n=0}^{\infty} \frac{s^n}{n!} H_n(x),$$

that is, $H_n(x) = \partial_1^n G(0,x)$. We have

$$H_0(x) = 1, \quad H_1(x) = x, \quad H_2(x) = x^2 - 1, \quad \text{etc.}$$

They satisfy the three-term recurrence relation

$$H_{n+1}(x) = xH_n(x) - nH_{n-1}(x),$$

the differential recurrence relation

$$H_n'(x) = nH_{n-1}(x),$$

and the differential equation

$$H_n''(x) - xH_n'(x) + nH_n(x) = 0.$$

The recurrence relation and the facts that $H_0 = 1$ and $H_1(x) = x$ imply that the leading coefficient of H_n equals 1 for every $n \geqslant 0$.

Hermite polynomials are orthogonal for the standard Gaussian distribution γ and diagonalize the operators N_t and A of the Ornstein-Uhlenbeck process.

Theorem 25.9 (Hermite polynomials and the Ornstein-Uhlenbeck process)**.** *The sequence $(H_n/\sqrt{n!})_{n\geqslant 0}$ is an orthonormal basis of $L^2(\gamma)$. Moreover, for every integer $n \geqslant 0$ and every real $t \geqslant 0$, the polynomial H_n is an eigenvector of N_t (respectively A) associated with the eigenvalue e^{-nt} (respectively $-n$).*

Proof. Fix $s, t \in \mathbb{R}$. Apply N_t to the function $x \mapsto G(s,x)$:

$$N_t(G(s,\cdot))(x) = \exp(se^{-t}x - s^2/2)\mathbb{E}[\exp(s\sigma_t Y)]$$

where Y has distribution γ. The Laplace transform of Y equals

$$\mathbb{E}(\exp(\lambda Y)) = \exp\Big(\frac{1}{2}\lambda^2\Big)$$

hence

$$N_t(G(s,\cdot))(x) = G(se^{-t}, x).$$

By definition of the Hermite polynomials, we obtain

$$\begin{aligned} N_t(H_n)(x) &= N_t(\partial_1^n G(0,\cdot))(x) = \partial_s^n N_t(G(s,\cdot))(x)_{|s=0} \\ &= \partial_s^n G(se^{-t}, x)_{|s=0} = e^{-nt}\partial_1^n G(se^{-t}, x)_{|s=0} = e^{-nt}H_n(x). \end{aligned}$$

The Hermite polynomial H_n is therefore an eigenvector of N_t with eigenvalue e^{-nt}. Applying Theorem 25.4 to the function H_n shows that H_n is an eigenvector of A associated with eigenvalue $-n$. Symmetry of N_t with respect to γ ensures that, for all integers m and n and every $t > 0$,

$$e^{-mt}\int H_m H_n \, d\gamma = \int N_t(H_m) H_n \, d\gamma = \int H_m N_t(H_n) \, d\gamma = e^{-nt}\int H_m H_n \, d\gamma.$$

In particular, H_n and H_m are orthogonal in $L^2(\gamma)$ whenever $n \neq m$. Finally, to see that $(H_n/\sqrt{n!})_{n\geqslant 0}$ is indeed an orthonormal basis of $L^2(\gamma)$, we may use Plancherel's formula

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{s^{2n}}{n!^2}\|H_n\|^2_{L^2(\gamma)} &= \int G(s,x)^2 \, \gamma(dx) = \exp(-s^2)\int e^{2sx}\,\gamma(dx) \\ &= \exp(s^2) = \sum_{n=0}^{\infty}\frac{s^{2n}}{n!}, \end{aligned}$$

which yields the value of $\|H_n\|^2_{L^2(\gamma)} = n!$ by identifying the two series. □

Theorem 25.9 gives the decomposition of $N_t f$ in the orthonormal basis of Hermite polynomials from that of f, and in particular provides an estimate for convergence of $N_t(f)$ to a constant function in $L^2(\gamma)$.

Theorem 25.10 (Spectral decomposition and convergence in $L^2(\gamma)$)**.** *For every $f \in L^2(\gamma)$, if*

$$f = \sum_{n \geqslant 0} a_n H_n \quad \text{où} \quad a_n = \int f H_n \, d\gamma \quad \text{avec} \quad \sum_n n! a_n^2 < +\infty,$$

then, for every $t \geqslant 0$,

$$N_t f = \sum_{n \geqslant 0} e^{-nt} a_n H_n.$$

Moreover, for every $f \in L^2(\gamma)$*, the following exponential convergence holds:*

$$\|N_t f - \gamma(f)\|_{L^2(\gamma)} \leqslant e^{-t} \|f - \gamma(f)\|_{L^2(\gamma)} \quad \text{où} \quad \gamma(f) = \int f \, d\gamma.$$

Proof. The first part follows from Theorem 25.9. For the second part, first observe that $H_0(x) = 1$. Next, for every $t \geqslant 0$, since γ is the invariant measure of $(N_t)_{t \geqslant 0}$,

$$a_0 = \int f \, d\gamma = \int N_t f \, d\gamma.$$

Thus, for every $t \geqslant 0$,

$$\begin{aligned} \|N_t f - \gamma(f)\|_{L^2(\gamma)}^2 &= \int (N_t f - a_0)^2 \, d\gamma = \sum_{n \geqslant 1} a_n^2 e^{-2nt} n! \\ &\leqslant e^{-2t} \sum_{n \geqslant 1} a_n^2 n! = e^{-2t} \|f - \gamma(f)\|_{L^2(\gamma)}^2. \end{aligned}$$

□

Remark 25.11 (*Mind the gap!*)**.** *The exponential convergence rate (equal to* 1*) in Theorem 25.10 corresponds to the gap between the first two eigenvalues of* A*, namely* 0 *and* -1*. This property extends to every reversible Markov process whose infinitesimal generator has the* spectral gap *property:* 0 *is a simple eigenvalue isolated from the rest of the spectrum.*

Remark 25.12 (Conservation laws and convergence rate)**.** *The semigroup preserves the spectral decomposition. If* $f \perp \mathrm{Vect}\{H_1, \ldots, H_{k-1}\}$ *in* $L^2(\gamma)$*, then* $N_t(f) \perp \mathrm{Vect}\{H_1, \ldots, H_{k-1}\}$ *for every* $t \geqslant 0$ *and*

$$\|N_t f - \gamma(f)\|_{L^2(\gamma)} \leqslant e^{-kt} \|f - \gamma(f)\|_{L^2(\gamma)} \quad \text{où} \quad \gamma(f) = \int f \, d\gamma.$$

A probability measure μ on $\mathbb{R}$ is said to satisfy a Poincaré inequality with constant $c > 0$ if, for every compactly supported function $f : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^1$,

$$\mathrm{Var}_\mu(f) := \int f^2 \, d\mu - \left(\int f \, d\mu \right)^2 \leqslant c \int f'^2 \, d\mu.$$

The optimal constant c is the smallest possible constant. Note that

$$\mathrm{Var}_\mu(f) = \|f - \mathbb{E}_\mu(f)\|_{L^2(\mu)}^2.$$

Corollary 25.13 (Poincaré inequality)**.** *The Gaussian measure γ satisfies a Poincaré inequality with optimal constant* 1*: for every function f from $\mathbb{R}$ to $\mathbb{R}$ of class $\mathcal{C}^1$ with compact support,*

$$\mathrm{Var}_\gamma(f) \leqslant \int f'^2 \, d\gamma.$$

Proof. For every $t \geqslant 0$, we have

$$\alpha(t) := \mathrm{Var}_\gamma(N_t f) = \|N_t f - \gamma(f)\|^2_{L^2(\gamma)} = \int (N_t f)^2 \, d\gamma - \left(\int f \, d\gamma \right)^2,$$

where the second equality follows by invariance. Theorem 25.10 gives the inequality $\alpha(t) \leqslant e^{-2t}\alpha(0)$. Theorem 25.5 (symmetry and reversibility) gives

$$\alpha'(t) = \int 2 N_t f A_t N_t f \, d\gamma = -2 \int (N_t f)'^2 \, d\gamma.$$

The function $t \in \mathbb{R}_+ \mapsto e^{-2t}\alpha(0) - \alpha(t)$ is nonnegative and vanishes at zero. Its right derivative at zero is therefore nonnegative:

$$\begin{aligned} 0 &\leqslant \frac{d}{dt}\Big|_{t=0} (e^{-2t}\alpha(0) - \alpha(t)) \\ &= -2\alpha(0) - 2 \int f'^2 \, d\gamma \\ &= -2\mathrm{Var}_\gamma(f) - 2 \int f'^2 \, d\gamma. \end{aligned}$$

The Poincaré inequality with constant 1 is thus established. Optimality of the constant can be checked on the eigenfunction $f(x) = H_1(x) = x$. Note that, conversely, the exponential convergence $\alpha(t) \leqslant e^{-2t}\alpha(0)$ of Theorem 25.10 follows from this Poincaré inequality by writing

$$\alpha'(t) = -2 \int (N_t f)'^2 \, d\gamma \leqslant -2\mathrm{Var}(N_t) = -2\alpha(t).$$

Finally, we give an alternative proof of the Poincaré inequality. Let $t \geqslant 0$ and let f be smooth. Define the function β on $[0, t]$ by setting

$$\beta(s) = N_s(\Psi(N_{t-s} f)) \quad \text{où} \quad \Psi(x) = x^2.$$

Below, we write $g = N_{t-s} f$. Observe that

$$\beta(t) - \beta(0) = N_t(\Psi(f)) - \Psi(N_t f) = \mathrm{Var}_{N_t(\cdot)}(f).$$

Moreover, by Theorem 25.4, a calculation gives

$$\beta'(s) = N_s(A(\Psi(g)) - Ag\Psi'(g)).$$

A direct calculation shows that

$$A(\Psi(g)) - Ag\Psi'(g) = \Psi''(g)g'^2 = 2g'^2.$$

Hence

$$\beta'(s) = 2N_s\big((N_{t-s}f)'^2\big).$$

By Mehler's formula, the semigroup satisfies the commutation relation

$$(N_t f)' = e^{-t}N_t(f'),$$

yielding the following expression for $\beta'(s)$:

$$\beta'(s) = 2e^{-2(t-s)}N_s\big((N_{t-s}(f'^2)\big).$$

Jensen's (or Cauchy-Schwarz's) inequality ensures that

$$(N_t|f'|)^2 \leqslant N_t\big(f'^2\big).$$

Thus, using the semigroup identity, we obtain the following upper bound:

$$\beta'(s) \leqslant 2e^{-2(t-s)}N_t\big(f'^2\big).$$

Integrating between 0 and t ensures that, for every $t \geqslant 0$ and every $x \in \mathbb{R}$, the probability measure $N_t(\cdot)(x) = \mathcal{N}(xe^{-t}, 1 - e^{-2t})$ satisfies a Poincaré inequality with constant $(1 - e^{-2t})/2$. To conclude, observe that if ν satisfies a Poincaré inequality with constant c, then the image measure of ν under $x \mapsto ax + b$ satisfies a Poincaré inequality with constant a^2c. □

25.3 Entropy, coupling, and long-time behavior

We now study convergence of the distribution of X_t to its invariant distribution as $t \to +\infty$ in terms of relative entropy and Wasserstein distance. Let μ_1 and μ_2 be probability measures on $\mathbb{R}$. The relative entropy of μ_1 with respect to μ_2 is defined by[2]

$$\mathrm{Ent}(\mu_1 \,|\, \mu_2) = \begin{cases} \displaystyle\int \frac{d\mu_1}{d\mu_2} \log\left(\frac{d\mu_1}{d\mu_2}\right) d\mu_2 & \text{si } \mu_1 \ll \mu_2, \\ +\infty & \text{sinon.} \end{cases}$$

Although it is not a distance, it is nonnegative and vanishes if and only if $\mu = \nu$ (this follows from Jensen's inequality for the strictly convex function $x \in \mathbb{R}_+ \mapsto u\log(u)$ with $0\log(0) = 0$). The Wasserstein distance of order 2 between μ_1 and μ_2 is given by

[2] In physics, $\mathrm{Ent}(\mu \,|\, \gamma)$ is a free energy if μ and γ are Gibbs measures.

$$W_2(\mu_1,\mu_2) = \inf_{\substack{X_1\sim\mu_1\\ X_2\sim\mu_2}} \sqrt{\mathbb{E}\Big(|X_1-X_2|^2\Big)},$$

where the infimum runs over pairs (X_1, X_2) of random variables whose marginal distributions are μ_1 and μ_2. Wasserstein distances of all orders are defined in Chapter 17. These two quantities, $\mathrm{Ent}(\mu_1 \,|\, \mu_2)$ and $W_2(\mu_1 \,|\, \mu_2)$, have explicit expressions when μ_1 and μ_2 are Gaussian measures.

Theorem 25.14 (Distances between Gaussian distributions)**.** *If* $\mu_1 = \mathcal{N}(m_1,\sigma_1)$ *and* $\mu_2 = \mathcal{N}(m_2,\sigma_2)$ *on* $\mathbb{R}$*, then*

$$W_2(\mu_1,\mu_2)^2 = (m_1-m_2)^2 + (\sigma_1-\sigma_2)^2.$$

Moreover, if $\sigma_1 > 0$ *and* $\sigma_2 > 0$*, then*

$$\mathrm{Ent}(\mu_1 \,|\, \mu_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2-\sigma_2^2+(m_1-m_2)^2}{2\sigma_2^2}.$$

Proof. Begin with the Wasserstein distance formula. Let (X_1, X_2) be a pair of real random variables such that $X_i \sim \mu_i = \mathcal{N}(m_i, \sigma_i^2)$ for $i = 1, 2$. We have

$$\begin{aligned}\mathbb{E}(|X_1-X_2|^2) &= \mathbb{E}(X_1^2) + \mathbb{E}(X_2)^2 - 2\mathbb{E}(X_1X_2)\\ &= (m_1-m_2)^2 + \sigma_1^2 + \sigma_2^2 - 2\mathrm{Cov}(X_1,X_2).\end{aligned}$$

This formula depends only on the mean and covariance matrix of the random vector (X_1, X_2). We have

$$W_2(\mu_1,\mu_2)^2 = (m_1-m_2)^2 + \sigma_1^2 + \sigma_2^2 - 2\sup_{C\in\mathcal{C}} C_{12}$$

where $\mathcal{C}$ denotes the set of covariance matrices 2×2 whose diagonal is prescribed by $C_{11} = \sigma_1^2$ and $C_{22} = \sigma_2^2$. Since the set of covariance matrices coincides with the set of symmetric matrices having nonnegative spectrum, membership in C amounts to the additional constraint $\det(C) = \sigma_1^2\sigma_2^2 - C_{12}^2 \geqslant 0$, hence $C_{12} = \sigma_1\sigma_2$, yielding the result.

For the entropy formula, write

$$\begin{aligned}\mathrm{Ent}(\mu_1 \,|\, \mu_2) &= \int \log\left(\frac{d\mu_1}{d\mu_2}\right) d\mu_1\\ &= \int \left(\log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{(x-m_1)^2}{2\sigma_1^2} + \frac{(x-m_2)^2}{2\sigma_2^2}\right)\mu_1(dx)\\ &= \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2} + \frac{\sigma_1^2+m_1^2-2m_1m_2+m_2^2}{2\sigma_2^2}\\ &= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2-\sigma_2^2+(m_1-m_2)^2}{2\sigma_2^2}.\end{aligned}$$

□

Since the Ornstein-Uhlenbeck process is Gaussian, the preceding theorem provides an explicit expression for the distance between the two distributions at time t starting from two different Gaussian distributions. More precisely, for all x_1, x_2 and $t \geqslant 0$, using Theorem 25.14 with $\mu_1 = N_t(\cdot)(x_1) = \mathcal{N}(x_1 e^{-t}, 1 - e^{-2t})$ and $\mu_2 = N_t(\cdot)(x_2) = \mathcal{N}(x_2 e^{-t}, 1 - e^{-2t})$ yields

$$W_2(N_t(\cdot)(x_1), N_t(\cdot)(x_2))^2 = e^{-2t}(x_1 - x_2)^2 \underset{t\to\infty}{\searrow} 0$$

and

$$\mathrm{Ent}(N_t(\cdot)(x_1), N_t(\cdot)(x_2)) = \frac{e^{-2t}(x_1 - x_2)^2}{2(1 - e^{-2t})} \underset{t\to\infty}{\searrow} 0.$$

These results extend to more general initial distributions. The Wasserstein distance case is simple; the entropy case is treated in the next section.

Theorem 25.15 (Convergence in Wasserstein distance)**.** *If μ_0 and μ_0' are probability measures with finite moment of order 2, and if μ_t (respectively μ_t') is the distribution of X_t (respectively X_t') when $X_0 \sim \mu_0$ (respectively $X_0' \sim \mu_0'$), then for every $t \geqslant 0$,*

$$W_2(\mu_t, \mu_t') \leqslant e^{-t} W_2(\mu_0, \mu_0').$$

In particular, if $\mu_0' = \gamma$, then for every $t \geqslant 0$,

$$W_2(\mu_t, \gamma) \leqslant e^{-t} W_2(\mu_0, \gamma).$$

Proof. Let (X_0, X_0') be a coupling of the distributions μ_0 and μ_0', independent of the Brownian motion $(B_t)_{t\geqslant 0}$. Construct two processes $(X_t)_{t\geqslant 0}$ and $(X_t')_{t\geqslant 0}$ starting respectively from X_0 and X_0' and driven by the same Brownian motion B. We say that the processes are coupled. Then

$$X_t - X_t' = X_0 - X_0' - \int_0^t (X_s - X_s')\, ds.$$

We deduce that

$$|X_t - X_t'|^2 = e^{-2t} |X_0 - X_0'|^2.$$

By definition of $W_2(\mu_t, \mu_t')$, we obtain

$$W_2(\mu_t, \mu_t')^2 \leqslant e^{-2t} \mathbb{E}\Big(|X_0 - X_0'|^2\Big).$$

It remains to take the infimum over all couplings of μ_0 and μ_0'. □

25.4 The logarithmic Sobolev inequality

A probability measure μ on $\mathbb{R}$ is said to satisfy a logarithmic Sobolev inequality with constant $c > 0$ if, for every nonnegative compactly supported function $f : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^1$,

$$\mathrm{Ent}_\mu(f) := \int f\log(f)\,d\mu - \int f\,d\mu \log\left(\int f\,d\mu\right) \leqslant c\int \frac{f'^2}{f}\,d\mu.$$

The optimal constant c is the smallest possible constant. Note that

$$\mathbb{E}_\mu(f)\mathrm{Ent}_\mu(f) = \mathrm{Ent}(\nu\,|\,\mu) \quad \text{où} \quad d\nu = \frac{f}{\mathbb{E}_\mu(f)}\,d\mu.$$

Theorem 25.16 (Gaussian logarithmic Sobolev inequality)**.** *Standard Gaussian measure γ satisfies a logarithmic Sobolev inequality with optimal constant* $1/2$.

Proof. Proceed as in the second proof of the Poincaré inequality in Corollary 25.13. Let $t \geqslant 0$ and let f be smooth. Define $\beta : [0,t] \to \mathbb{R}$ by

$$\beta(s) = N_s(\Phi(N_{t-s}f)) \quad \text{où} \quad \Phi(x) = x\log(x).$$

Below, we write $g = N_{t-s}f$. Observe that

$$\beta(t) - \beta(0) = N_t(\Phi(f)) - \Phi(N_t f) = \mathrm{Ent}_{N_t(\cdot)}(f).$$

Moreover, by Theorem 25.4, a calculation gives

$$\beta'(s) = N_s(A(\Phi(g)) - Ag\Phi'(g)).$$

A direct calculation shows that

$$A(\Phi(g)) - Ag\Phi'(g) = \Phi''(g)g'^2 = \frac{g'^2}{g}.$$

Hence

$$\beta'(s) = N_s\left(\frac{(N_{t-s}f)'^2}{N_{t-s}f}\right).$$

By Mehler's formula, the semigroup satisfies the commutation relation

$$(N_t f)' = e^{-t}N_t(f'),$$

yielding the following expression for $\beta'(s)$:

$$\beta'(s) = e^{-2(t-s)}N_s\left(\frac{(N_{t-s}(|f'|))^2}{N_{t-s}f}\right).$$

Schwarz's inequality ensures that

$$(N_t|f'|)^2 = \left[N_t\left(\frac{|f'|}{\sqrt{f}}\sqrt{f}\right)\right]^2 \leqslant N_t\left(\frac{f'^2}{f}\right)N_t(f).$$

Thus, using the semigroup identity, we obtain the following upper bound:

$$\beta'(s) \leqslant e^{-2(t-s)} N_t\left(\frac{f'^2}{f}\right).$$

Integrating between 0 and t ensures that, for every $t \geqslant 0$ and every $x \in \mathbb{R}$, the probability measure $N_t(\cdot)(x) = \mathcal{N}(xe^{-t}, 1 - e^{-2t})$ satisfies a logarithmic Sobolev inequality with constant $(1 - e^{-2t})/2$. To conclude, observe that if ν satisfies a logarithmic Sobolev inequality with constant c, then the image measure of ν under $x \mapsto ax + b$ satisfies a logarithmic Sobolev inequality with constant $a^2 c$. Finally, the functions $x \mapsto e^{\lambda x}$ achieve equality in the logarithmic Sobolev inequality: the constant $1/2$ is optimal. □

Remark 25.17 (Alternative proof)**.** *Here is a unified alternative proof of the Poincaré inequality in Corollary 25.13 and the logarithmic Sobolev inequality in Theorem 25.16. It uses the remarkable convexity of the function* $(u, v) \mapsto \Psi(u, v) = \Phi''(u)v^2$*, with* $\Phi(u) = u^2$ *for the Poincaré inequality and* $\Phi(u) = u \log(u)$ *for the logarithmic Sobolev inequality.*

$$\begin{aligned}
N_t(\Phi(f)) - \Phi(N_t f) &= \int_0^t \frac{d}{ds} N_s(\Phi(N_{t-s} f))\, ds \\
&= \int_0^t N_s(A(\Phi(N_{t-s} f)) - A N_{t-s}(\Phi'(N_{t-s} f)))\, ds \\
&= \int_0^t N_s(\Phi''(N_{t-s} f)(N_{t-s} f)'^2)\, ds \\
&= \int_0^t e^{-2s} N_s(\Phi''(N_{t-s} f)(N_{t-s}(f'))^2)\, ds \\
&= \int_0^t e^{-2s} N_s(\Psi(N_{t-s} f, N_{t-s}(f'))\, ds \\
&\leqslant \int_0^t e^{-2s} N_s(N_{t-s}(\Psi(f, f')))\, ds \\
&= N_t(\Psi(f, f')) \int_0^t e^{-2s}\, ds \\
&= \frac{1}{2} N_t(\Psi(f, f')).
\end{aligned}$$

Finally, here is yet another proof, whose interchanges require justification:

$$\begin{aligned}
\int \Phi(f)\, d\gamma - \Phi\left(\int f\, d\gamma\right) &= -\int \left(\lim_{t\to\infty} (\Phi(N_t) - \Phi(N_0 f))\right) d\gamma \\
&= -\int \left(\lim_{t\to\infty} \int_0^t \frac{d}{ds} \Phi(N_s f)\, ds\right) d\gamma \\
&= -\int \left(\lim_{t\to\infty} \int_0^t \Phi'(N_s f) N_s A f\, ds\right) d\gamma \\
&= -\lim_{t\to\infty} \int_0^t \left(\int \Phi'(N_s f) N_s A f\, d\gamma\right) ds
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{t\to\infty} \int_0^t \left(\int \Phi''(N_s f)(N_s f)'^2 \, d\gamma \right) ds \\
&= \lim_{t\to\infty} \int_0^t e^{-2s} \left(\int \Phi''(N_s f)(N_s(f'))^2 \, d\gamma \right) ds \\
&= \lim_{t\to\infty} \int_0^t e^{-2s} \left(\int \Psi(N_s f, N_s(f')) \, d\gamma \right) ds \\
&\leqslant \lim_{t\to\infty} \int_0^t e^{-2s} \left(\int N_s(\Psi(f, f')) \, d\gamma \right) ds \\
&= \left(\lim_{t\to\infty} \int_0^t e^{-2s} \, ds \right) \int \Psi(f, f') \, d\gamma \\
&= \frac{1}{2} \int \Psi(f, f') \, d\gamma.
\end{aligned}
$$

In both cases, we used Jensen's inequality for the probability measure N_s and the convex function Ψ.

Remark 25.18 (The L^2 form of the logarithmic Sobolev inequality)**.** *Sometimes the logarithmic Sobolev inequality is preferably defined in L^2 form: for every smooth function f,*

$$
\int f^2 \log(f^2) \, d\mu - \int f^2 \, d\mu \log \left(\int f^2 \, d\mu \right) \leqslant 4c \int f'^2 \, d\mu.
$$

The logarithmic Sobolev inequality in Theorem 25.16 yields an exponential rate of convergence for the relative entropy of the distribution of X_t with respect to the invariant distribution γ.

Theorem 25.19 (Exponential convergence of entropy)**.** *Suppose the distribution μ_0 of X_0 has finite relative entropy with respect to γ. Then, writing μ_t for the distribution of X_t, the function $t \in \mathbb{R}_+ \mapsto \mathrm{Ent}(\mu_t \,|\, \gamma)$ is decreasing and, for every $t \geqslant 0$,*

$$
\mathrm{Ent}(\mu_t \,|\, \gamma) \leqslant e^{-2t} \mathrm{Ent}(\mu_0 \,|\, \gamma).
$$

Conversely, as for the Poincaré inequality (proof of Corollary 25.13), the logarithmic Sobolev inequality follows from exponential convergence of relative entropy by considering the derivative at $t = 0$ of the difference between the two sides.

Proof. Write v_t for the density of μ_t with respect to γ. Then

$$
\mathrm{Ent}(\mu_t \,|\, \gamma) = \int v_t \log(v_t) d\gamma.
$$

By Theorems 25.7 and 25.5,

$$
\frac{d}{dt} \int v_t \log(v_t) d\gamma = \int \partial_t v_t (1 + \log(v_t)) d\gamma
$$

$$= \int A v_t (1 + \log(v_t)) d\gamma$$
$$= \int v_t A(1 + \log(v_t)) d\gamma.$$

For every smooth function g, a direct calculation shows that

$$A(1 + \log(g)) = \frac{Ag}{g} - \frac{g'^2}{g^2}.$$

By the invariance property of γ, we have $\int Ag \, d\gamma = 0$. Hence

$$\frac{d}{dt}\mathrm{Ent}(\mu_t \,|\, \gamma) = -\int \frac{(\partial_x v_t)^2}{v_t} d\gamma,$$

proving that $t \mapsto \mathrm{Ent}(\mu_t \,|\, \gamma)$ is decreasing. Moreover, the logarithmic Sobolev inequality applied to the function v_t ensures that

$$\frac{d}{dt}\mathrm{Ent}(\mu_t \,|\, \gamma) \leqslant -2\mathrm{Ent}(\mu_t \,|\, \gamma).$$

Grönwall's lemma[3] finally yields the desired relation. □

Remark 25.20 (Smoothing effect)**.** *Theorem 25.7 ensures that even if the distribution of X_0 has no density, the distribution of X_t always has one whenever $t > 0$. We can therefore apply Theorem 25.19 between times $t_0 > 0$ and t to obtain exponential convergence of relative entropy to 0.*

25.5 Further reading

Everything concerning the Ornstein-Uhlenbeck process can be found in the first pages of Dominique Bakry's lecture notes [Bak94]. The expression for the Wasserstein distance between two Gaussian measures in Theorem 25.14 is established, in particular, in an article [GS84] by Clark Givens and Rae Shortt.

The Ornstein-Uhlenbeck process is the celebrated representative of an important class of stochastic processes: Kolmogorov diffusions. Let U be a real-valued function on $\mathbb{R}^d$ of class $\mathcal{C}^2$ and $(B_t)_{t \geqslant 0}$ a standard Brownian motion with values in $\mathbb{R}^d$. The following stochastic differential equation is called the Langevin equation:

$$dX_t = \sqrt{2} dB_t - \nabla U(X_t) dt.$$

[3] If $t \in [a, b] \mapsto \alpha(t) \in \mathbb{R}_+$ is a continuous differentiable function on $]a, b[$, if $t \in [a, b] \mapsto \beta(t) \in \mathbb{R}_+$ is continuous, and if $\alpha'(t) \leqslant \beta(t)\alpha(t)$ for every $t \in]a, b[$, then $\alpha(t) \leqslant \alpha(a) \exp\left(\int_a^t \beta(s)\, ds\right)$ for every $t \in [a, b]$.

The solution $(X_t)_{t\geqslant 0}$ of this stochastic differential equation is called the Kolmogorov diffusion process associated with the potential U. Like the Ornstein-Uhlenbeck process, it is a Markov process. The integral version of the stochastic differential equation is

$$X_t = X_0 + \sqrt{2}B_t - \int_0^t \nabla U(X_s)\, ds.$$

Its infinitesimal generator L is given, for a function f of class $\mathcal{C}^2$, by

$$Lf = \Delta f - \langle \nabla U, \nabla f \rangle.$$

If e^{-U} is integrable, then the probability measure

$$\mu(dx) = \frac{1}{Z} e^{-U(x)}\, dx \quad \text{avec} \quad Z = \int e^{-U(x)}\, dx,$$

is the invariant distribution associated with X. The operator L is even self-adjoint in $L^2(\mu)$. The spectral decomposition of the associated semigroup is not explicit beyond the Gaussian case, but analogues of Theorems 25.10 or 25.19 can still be obtained by comparison with the Gaussian case. For example, one can show that if there exists $\lambda > 0$ such that

$$\forall x \in \mathbb{R}^d, \quad \operatorname{Hess} U(x) \geqslant \lambda I,$$

which means that μ has a log-concave density with respect to the centered Gaussian probability measure with covariance $\lambda^{-1} I_d$, then for every $t \geqslant 0$,

$$\operatorname{Ent}(\mu_t \,|\, \mu) \leqslant e^{-2\lambda t} \operatorname{Ent}(\mu_0 \,|\, \mu)$$

where μ_t is the distribution of X_t. This is a special case of the Γ_2 criterion established by Dominique Bakry and Michel Émery [BÉ85]. Kolmogorov processes are studied in detail in the books by Gilles Royer [Roy99], by Dominique Bakry, Ivan Gentil, and Michel Ledoux [BGL14], and in the collective volume [ABC+00].

Theorem 25.14 extends to the multivariate case. More precisely, if $\mu_1 = \mathcal{N}_d(m_1, \Sigma_1)$ and $\mu_2 = \mathcal{N}_d(m_2, \Sigma_2)$ on $\mathbb{R}^d$, then

$$W_2(\mu_1, \mu_2)^2 = \|m_1 - m_2\|_2^2 + \operatorname{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2}).$$

When Σ_1 and Σ_2 commute, the last term in the formula above reduces to $\big\|\sqrt{\Sigma_1} - \sqrt{\Sigma_2}\big\|_{\mathrm{HS}}^2$, closely resembling the formula in dimension 1. Moreover, if the covariance matrices Σ_1 and Σ_2 are invertible, then

$$\operatorname{Ent}(\mu_1 \,|\, \mu_2) = \frac{1}{2}\Big(\log\left(\frac{\det \Sigma_2}{\det \Sigma_1}\right) + \operatorname{Tr}\left(\Sigma_2^{-1}\Sigma_1\right) - d + (m_1 - m_2)^T \Sigma_2^{-1} (m_1 - m_2)\Big).$$

26

Kinetic diffusion models

Keywords. Diffusion; Brownian motion; kinetic equation; partial differential equation; linear stochastic differential equation.

Tools. Gaussian vector; Wasserstein distance; relative entropy.

Difficulty. ***

26.1 Langevin model

Consider a dust particle in a fluid such as water. The particle is enormous compared with the fluid molecules. Its size is on a macroscopic scale: its movements can be observed, at least with a good magnifying glass, and measured. The molecules making up the fluid are far too small, and above all far too numerous, to be observed individually. They lie on a microscopic scale. At every instant, countless collisions occur between the particle and the surrounding molecules. On average, these collisions are distributed uniformly over the particle's surface. At any given instant, however, deviations from this uniform distribution may occur. We seek a model for the evolution of the particle's position/velocity pair (X_t, V_t) at time t derived from Newton's second law. The forces acting on the particle are:

- the particle's weight;
- buoyancy;
- friction opposing the particle's motion;
- the resultant force from collisions with water molecules.

The particle's weight and buoyancy balance, and we neglect their resultant by considering only horizontal motion. We assume that the magnitude of the friction force is proportional to the particle's velocity. It is therefore $-\lambda V_t$, since

it opposes motion. Modelling the effect of collisions with water molecules requires a little more care. Between times t and $t+\alpha$, with α large on the microscopic time scale, the particle undergoes $N\alpha$ collisions, where N, the number of collisions per unit time, is assumed very large. We adopt a "statistical" or "stochastic" approach and model the collisions by independent, identically distributed centred random variables $(\varepsilon_i)_{i\geqslant 1}$ with variance τ^2. Set $\sigma^2 = N\tau^2$. The force exerted during the time interval α is

$$\sum_{i=1}^{N\alpha} \varepsilon_i = \sqrt{\sigma^2\alpha}\, \underbrace{\frac{1}{\sqrt{\tau^2 N\alpha}} \sum_{i=1}^{N\alpha} \varepsilon_i}_{Y}.$$

If $N\alpha$ is large, the central limit theorem suggests that the distribution of Y is close to a standard Gaussian. Moreover, the random forces exerted over disjoint time intervals arise from different water molecules, since they move much less than the particle. We may therefore assume them independent. It thus seems natural to model the collisions by a process with independent stationary increments, whose increment between times t and $t+\alpha$ is centred Gaussian with variance $\sigma^2\alpha$. This process is Brownian motion $(\sigma B_t)_{t\geqslant 0}$. Here we recover the idea of Donsker's theorem (see Theorem 27.6). Since velocity is the derivative of position, we obtain the system:

$$\begin{cases} dX_t = V_t\, dt \\ dV_t = -\lambda V_t\, dt + \sigma dB_t. \end{cases}$$

This is called a kinetic equation because it involves the particle's position and velocity simultaneously. Throughout this chapter, $(Z_t)_{t\geqslant 0}$ denotes the full process defined by $Z_t = (X_t, V_t)$ for $t \geqslant 0$.

Theorem 26.1 (Langevin process)**.** *Conditionally on the event* $\{(X_0, V_0) = (x_0, v_0)\}$*, the process* $(Z_t)_{t\geqslant 0}$ *is a Gaussian process such that*

$$\mathbb{E}(X_t) = x_0 + \frac{1-e^{-\lambda t}}{\lambda} v_0$$
$$\mathrm{Cov}(X_t, X_s) = \frac{\sigma^2 s}{\lambda^2} + \frac{\sigma^2}{2\lambda^3}\Big(-2 + 2e^{-\lambda t} + 2e^{-\lambda s} - e^{-\lambda|t-s|} - e^{-\lambda(t+s)}\Big).$$

In particular, the variance of X_t *is given by*

$$\mathrm{Var}(X_t) = \frac{\sigma^2 t}{\lambda^2} + \frac{\sigma^2}{2\lambda^3}\big(-3 + 4e^{-\lambda t} - e^{-2\lambda t}\big).$$

The evolution of velocity does not depend on position.

Proof. The process $(V_t)_{t\geqslant 0}$ is an Ornstein-Uhlenbeck process, studied in Chapter 25. The process $(V_t)_{t\geqslant 0}$ is therefore Gaussian, with mean $\mathbb{E}(V_t) = e^{-\lambda t} v_0$ and covariance

$$\mathrm{Cov}(V_t, V_s) = \frac{\sigma^2}{2\lambda}\Big(e^{-\lambda|t-s|} - e^{-\lambda(t+s)}\Big).$$

The process X is Gaussian because it is the image of V under a linear map: $X_t = x_0 + \int_0^t V_s\, ds$. □

Remark 26.2 (Orders of magnitude)**.** *The constant $1/\lambda$, which has the dimension of time, is called the relaxation time. In practice it is of the order of 10^{-8} seconds. This suggests that the process $(X_t)_{t\geqslant 0}$ cannot in practice be distinguished from a Brownian motion with variance σ^2/λ^2.*

26.2 Confined Langevin process

We now slightly modify the model of the preceding section. We add a force $-\mu x$ arising from the potential energy $x \mapsto \frac{1}{2}\mu x^2$. The evolution of the velocity/position pair is then given by the following system:

$$\begin{cases} dX_t = V_t\, dt \\ dV_t = -\lambda V_t\, dt - \mu X_t\, dt + \sigma dB_t. \end{cases}$$

We call $Z = (X, V)$ a confined Langevin process[1]. The system can be rewritten in the following matrix form:

$$Z_t = Z_0 + \int_0^t MZ_s\, ds + \sigma W_t \quad \text{où} \quad M = \begin{pmatrix} 0 & 1 \\ -\mu & -\lambda \end{pmatrix} \quad \text{et} \quad W_t = \begin{pmatrix} 0 \\ B_t \end{pmatrix}.$$

The solution of this linear equation has the form

$$Z_t = e^{tM} Z_0 + \sigma \int_0^t e^{(t-s)M}\, dW_s.$$

Lemma 26.3 (Distribution at time t)**.** *If (X_0, V_0) is a Gaussian vector, so is (X_t, V_t) for every $t \geqslant 0$. Moreover, if we write, for $t \geqslant 0$,*

$$x_t = \mathbb{E}(X_t), \quad v_t = \mathbb{E}(V_t), \quad a_t = \mathbb{E}(X_t^2), \quad b_t = \mathbb{E}(X_t V_t) \quad \textit{et} \quad c_t = \mathbb{E}(V_t^2).$$

then

$$(x_t)_{t\geqslant 0}, \quad (v_t)_{t\geqslant 0}, \quad (a_t)_{t\geqslant 0}, \quad (b_t)_{t\geqslant 0}, \quad \textit{et} \quad (c_t)_{t\geqslant 0}$$

solve the linear differential equations:

$$\frac{d}{dt}\begin{pmatrix} x_t \\ v_t \end{pmatrix} = M\begin{pmatrix} x_t \\ v_t \end{pmatrix} \quad \textit{et} \quad \frac{d}{dt}\begin{pmatrix} a_t \\ b_t \\ c_t \end{pmatrix} = N\begin{pmatrix} a_t \\ b_t \\ c_t \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \sigma^2 \end{pmatrix}$$

where

$$M = \begin{pmatrix} 0 & 1 \\ -\mu & -\lambda \end{pmatrix} \quad \textit{et} \quad N = \begin{pmatrix} 0 & 2 & 0 \\ -\mu & -\lambda & 1 \\ 0 & -2\mu & -2\lambda \end{pmatrix}.$$

[1] Also called a *damped Langevin process*.

Proof. This follows immediately from Itô's formula. □

Remark 26.4 (Explicit formulas)**.** *The matrix e^{tM} can be written explicitly:*

- *if $\lambda^2 > 4\mu$, then, writing $\omega = \sqrt{\lambda^2 - 4\mu}$, $\alpha_1 = (-\lambda + \omega)/2$ and $\alpha_2 = (-\lambda - \omega)/2$,*

$$e^{tM} = \frac{e^{\alpha_2 t}}{\omega} \begin{pmatrix} -\alpha_1 & 1 \\ -\mu & \alpha_2 \end{pmatrix} + \frac{e^{\alpha_1 t}}{\omega} \begin{pmatrix} \alpha_2 & -1 \\ \mu & -\alpha_1 \end{pmatrix};$$

- *if $\lambda^2 = 4\mu$, then*

$$e^{tM} = e^{-\lambda t/2} \begin{pmatrix} 1 - \lambda t/2 & t \\ -\lambda^2 t/4 & 1 + \lambda t/2 \end{pmatrix};$$

- *if $\lambda^2 - 4\mu < 0$, then, writing $\omega = \sqrt{\mu - \lambda^2/4}$,*

$$e^{tM} = e^{-\lambda t/2} \begin{pmatrix} \cos(\omega t) + \frac{\lambda \sin(\omega t)}{2\omega} & \frac{\sin(\omega t)}{\omega} \\ -\frac{\mu \sin(\omega t)}{\omega} & \cos(\omega t) - \frac{\lambda \sin(\omega t)}{2\omega} \end{pmatrix}.$$

An explicit expression for e^{tN} is much more cumbersome to obtain. However, all the eigenvalues of N have strictly negative real parts.

The linear differential equations in Lemma 26.3 are easy to solve numerically (see Figure 26.1).

Lemma 26.5 (Density of the distribution at every time)**.** *For every initial distribution and every time $t > 0$, the distribution of the confined Langevin process (X_t, V_t) has a strictly positive density on $\mathbb{R}^2$.*

Proof. It suffices to prove the result when the initial distribution is a Dirac mass at a point (x, v); the general case follows by conditioning on the initial value. If the initial distribution is a Dirac mass, the pair (X_t, V_t) is a Gaussian vector. Its distribution therefore has a strictly positive density on $\mathbb{R}^2$ if and only if the determinant of its covariance matrix is strictly positive. For simplicity, assume $(X_0, V_0) = (0, 0)$. The determinant of the covariance matrix is then given by $d_t = a_t c_t - b_t^2$ in the notation of Lemma 26.3. Thus

$$d_t' = -2\lambda(a_t c_t - b_t^2) + \sigma^2 a_t = -2\lambda d_t + \sigma^2 a_t.$$

Since $d_0 = a_0 = 0$, the derivative of $t \mapsto d_t$ is zero at 0. Iterating gives

$$\begin{aligned} d_t'' &= -2\lambda d_t' + 2\sigma^2 b_t \\ d_t^{(3)} &= -2\lambda d_t'' + 2\sigma^2(-\mu a_t - \lambda b_t + c_t) \\ d_t^{(4)} &= -2\lambda d_t^{(3)} + 2\sigma^2(\lambda\mu a_t + (\lambda^2 - 4\mu) b_t - 3\lambda c_t) + 2\sigma^4. \end{aligned}$$

In particular, the first three derivatives of $t \mapsto d_t$ vanish at 0 and $d_0^{(4)} = 2\sigma^4$. Thus the determinant becomes strictly positive for $t > 0$ if it vanishes initially. □

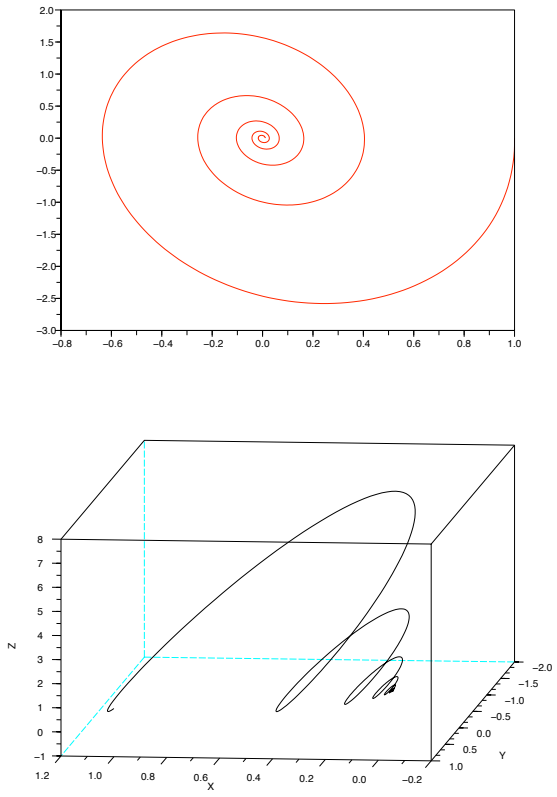


Fig. 26.1. Evolution of the mean vector (x_t, v_t) (top) and covariance (a_t, b_t, c_t) (bottom) for $\sigma = 1$, $\lambda = 1$ and $\mu = 10$.

Remark 26.6 (Hypoellipticity)**.** *In the proof of Lemma 26.5, we have*

$$d_t \underset{t\sim 0}{\sim} \frac{2\sigma^4}{4!} t^4.$$

For a diffusion on $\mathbb{R}^2$ with nondegenerate Brownian noise, the determinant would be of order t^2. This difference arises because only the velocity component is perturbed by Brownian motion. However, the subtle structure of the drift coefficient in the equation of the confined Langevin process ensures that this noise propagates to both components and that the distribution of the pair (X_t, V_t) has a density as soon as $t > 0$. This is a special case of hypoellipticity, studied in a general setting notably by Hörmander and later Malliavin.

Figure 26.2 illustrates the emergence of a density when the initial distribution is a Dirac measure on $\mathbb{R}^2$, through the time evolution of the determinant

of the covariance matrix. In particular, it converges to that of the invariant distribution, which is $\sigma^4/(4\lambda^2\mu)$.

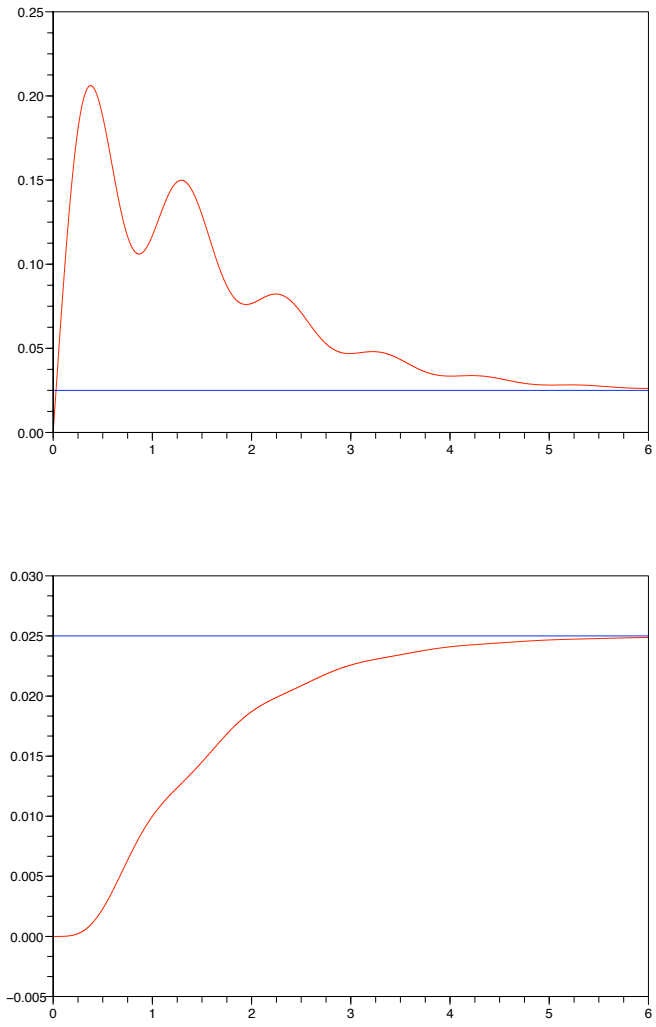


Fig. 26.2. Time evolution of the determinant of the covariance matrix starting from $(x_0, v_0, a_0, b_0, c_0) = (0, 0, 1, 1, 1)$ (top) and $(0, 0, 0, 0, 0)$ (bottom) when $\sigma = 1$, $\lambda = 1$ and $\mu = 10$.

26.3 Long-time behaviour of the confined process

As for the Ornstein-Uhlenbeck process, we associate with the confined Langevin process $Z = (X, V)$ a semigroup $(Q_t)_{t \geqslant 0}$, given for $t \geqslant 0$, $z \in \mathbb{R}^2$ and smooth f by

$$Q_t f(z) = \mathbb{E}(f(Z_t) | Z_0 = z).$$

Lemma 26.7 (Infinitesimal generator)**.** *If f is a compactly supported function of class $\mathcal{C}^\infty$ on $\mathbb{R}^2$, then for every $t \geqslant 0$ and $z \in \mathbb{R}^2$,*

$$\partial_t Q_t f(z) = Q_t(Af)(z)$$

where, writing $z = (x, v)$,

$$Af(x, v) = v \cdot \nabla_x f(x, v) - (\mu x + \lambda v) \cdot \nabla_v f(x, v) + \frac{\sigma^2}{2} \Delta_v f(x, v).$$

Proof. Apply Itô's formula to f with $Z_0 = z$:

$$f(Z_t) = f(z) + \int_0^t Af(Z_s)\, ds + M_t,$$

where $(M_t)_{t \geqslant 0}$ is a martingale. Thus,

$$\lim_{t \to 0} \frac{Q_t f(z) - f(z)}{t} = Af(z).$$

The Markov property gives the relation at every time $t \geqslant 0$. □

Theorem 26.8 (Invariant distribution)**.** *The product Gaussian probability measure*

$$\pi = \mathcal{N}\left(0, \frac{\sigma^2}{2\lambda\mu}\right) \otimes \mathcal{N}\left(0, \frac{\sigma^2}{2\lambda}\right)$$

is the invariant distribution of the process (X, V). For every compactly supported function f of class $\mathcal{C}^2$, it satisfies

$$\int Af\, d\pi = 0.$$

Proof. Since (X, V) is a Gaussian process, every invariant distribution is Gaussian. Moreover, the point $\sigma^2/(2\lambda\mu)(0, 0, 1, 0, \mu)$ is the unique fixed point of the differential equations in Lemma 26.3. Finally, note that

$$(x_t, v_t, a_t, b_t, c_t) \xrightarrow[t \to \infty]{} \frac{\sigma^2}{2\lambda\mu}(0, 0, 1, 0, \mu)$$

for all finite initial conditions, since the real parts of the eigenvalues of M and N are strictly positive. □

Remark 26.9 (Irreversibility)**.** *Unlike the Ornstein-Uhlenbeck case, one can check that the generator A of the process is not self-adjoint in $L^2(\pi)$. In other words, the semigroup $(Q_t)_{t\geqslant 0}$ is not reversible with respect to its invariant distribution π.*

Below we seek to quantify convergence of the distribution of X_t to the invariant distribution π as t tends to $+\infty$, in terms of Wasserstein distance or relative entropy as was done for the Ornstein-Uhlenbeck process in Chapter 25.

The explicit formula for the distance W_2 between two Gaussian distributions given in Section 25.5 (a multivariate version of that in Theorem 25.14) allows us to estimate the Wasserstein distance between the distribution of (X_t, V_t) and the invariant distribution when the initial distribution is Gaussian. Figure 26.3 shows the evolution of the relative entropy and squared Wasserstein distance of the distribution at time t with respect to the invariant distribution π. Figure 26.3 suggests that the relative entropy of the distribution at time t with respect to the invariant distribution decreases in stages. Unlike relative entropy, the distance W_2 is not at all a decreasing function of time!

Theorem 26.10 (Entropy decay)**.** *If π_t is the distribution of (X_t, V_t) and h_t is its density with respect to the measure π, then*

$$\frac{d}{dt}\mathrm{Ent}(\pi_t|\pi) = -\int \frac{|\nabla_v h_t|^2}{h_t}\, d\pi.$$

Proof. To simplify the calculations, assume without loss of generality that σ equals $\sqrt{2}$. By Lemma 26.5, the distribution π_t of (X_t, V_t) has a density u_t with respect to Lebesgue measure. It solves the following partial differential equation (the kinetic Fokker-Planck equation)

$$\partial_t u_t + v \cdot \nabla_x u_t - \mu x \cdot \nabla_v u_t = \Delta_v u_t + \nabla_v \cdot (\lambda v u_t),$$

which can be obtained from the infinitesimal generator (Lemma 26.7) by integration by parts. The structure of this equation is interesting: the left-hand side is conservative and describes the paths of a dynamical system in $\mathbb{R}^2$ associated with the Hamiltonian

$$(x, v) \mapsto \mu\frac{|x|^2}{2} + \lambda\frac{|v|^2}{2},$$

while the right-hand side is dissipative but acts only on the velocity variable. Write $h_t(x, v) = u_t(x, v)/\pi(x, v)$, where, with a slight abuse of notation, $\pi(x, v)$ denotes the density of the invariant distribution π at (x, v) given by Theorem 26.8. One checks that the function h then solves the following partial differential equation (PDE):

$$\partial_t h_t + v \cdot \nabla_x h_t - \mu x \cdot \nabla_v h_t = \Delta_v h_t - \lambda v \cdot \nabla_v h_t.$$

The variation of relative entropy is given by

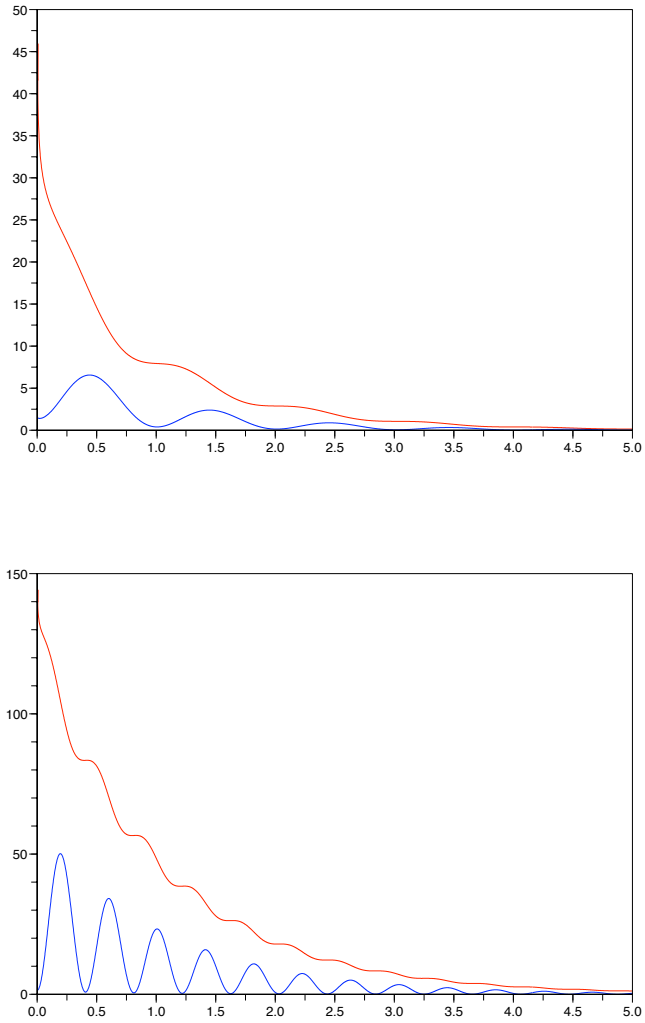


Fig. 26.3. Comparison of the evolution of relative entropy and Wasserstein distance of order 2 between the distribution at time t and the invariant distribution, $\sigma = 1$, $\lambda = 1$, $\mu = 10$, $(x_0, v_0, a_0, b_0, c_0) = (1, 0, 0, 0, 0)$ (top) and $\mu = 60$ (bottom).

$$\frac{d}{dt}\mathrm{Ent}(\pi_t|\pi) = \frac{d}{dt}\int h_t \log(h_t) d\pi = \int \partial_t h_t (1+\log(h_t))\, d\pi.$$

The expression for π in Theorem 26.8 and integration by parts

$$\begin{aligned}
-\int (v\cdot\nabla_x h_t)(1+\log(h_t))\, d\pi &= -\int v\cdot\nabla_x(h_t\log(h_t))\, d\pi \\
&\overset{\text{IPP}}{=} -\lambda\mu\int (x\cdot v) h_t \log(h_t)\, d\pi \\
\int (\mu x\cdot\nabla_v h_t)(1+\log(h_t))\, d\pi &= \int \mu x\cdot\nabla_v(h_t\log(h_t))\, d\pi \\
&\overset{\text{IPP}}{=} \lambda\mu\int (x\cdot v) h_t\log(h_t) d\pi \\
\int (\Delta_v h_t)(1+\log(h_t))\, d\pi &\overset{\text{IPP}}{=} -\int \frac{\nabla_v h_t\cdot\nabla_v h_t}{h_t}\, d\pi \\
&\quad + \lambda\int (1+\log(h_t))\nabla_v h_t\cdot v\, d\pi
\end{aligned}$$

give the claimed result. □

Obtaining explicit bounds for exponential convergence of relative entropy requires much more work than for the Ornstein-Uhlenbeck process studied in Chapter 25. A convergence result for Wasserstein distance of order 2 is easier to obtain, even though this quantity is not a decreasing function of time.

Theorem 26.11 (Convergence in Wasserstein distance)**.** *Let ν_0 and ν_0' be two probability measures on $\mathbb{R}^2$ with finite moments of order 2. Write ν_t (respectively ν_t') for the distribution of Z_t when Z_0 has distribution ν_0 (respectively ν_0'). Then there exist constants c and d such that*

$$W_2(\nu_t,\nu_t') \leqslant (c+dt)e^{-\rho t} W_2(\nu_0,\nu_0'),$$

with

$$\begin{cases} \rho = \dfrac{\lambda-\sqrt{\lambda^2-4\mu}}{2} \quad \textit{et} \quad d=0 & \textit{si } \lambda > 2\sqrt{\mu}, \\ \rho = \dfrac{\lambda}{2} & \textit{si } \lambda \leqslant 2\sqrt{\mu}. \end{cases}$$

Proof. We use coupling. Let (Z_0, Z_0') be a coupling of the distributions ν_0 and ν_0', independent of the Brownian motion $(B_t)_{t\geqslant 0}$. Write

$$Z := (X_t, V_t)_{t\geqslant 0} \quad \text{et} \quad Z' := (X_t', V_t')_{t\geqslant 0}$$

for the confined Langevin processes starting from Z_0 and Z_0', respectively, and driven by the same Brownian motion $(B_t)_{t\geqslant 0}$. Write

$$z_t := Z_t - Z_t'.$$

By linearity of the stochastic differential equation, the function $t \mapsto z_t$ depends on (Z_0, Z_0') but not on $(B_t)_{t\geqslant 0}$. Indeed, it solves the differential equation

$$\begin{cases} \dfrac{d}{dt} z_t = M z_t \\ z_0 = Z_0 - Z_0' \end{cases} \qquad \text{avec} \quad M = \begin{pmatrix} 0 & 1 \\ -\mu & -\lambda \end{pmatrix}.$$

Thus $z_t = e^{tM} z_0$. The eigenvalues of the matrix M are given by:

$$-\frac{\lambda \pm \sqrt{\lambda^2 - 4\mu}}{2} \quad \text{si} \quad \lambda \geqslant 2\sqrt{\mu} \quad \text{et} \quad -\frac{\lambda}{2} \pm i\frac{\sqrt{4\mu - \lambda^2}}{2} \quad \text{si} \quad \lambda < 2\sqrt{\mu}.$$

A Dunford decomposition of M ensures that there exist two constants c and d such that for every $t \geqslant 0$

$$|z_t| \leqslant (c + dt) e^{-\rho t} |z_0|$$

where $-\rho$ is the largest real part among the eigenvalues of M. In particular,

$$W_2(\nu_t, \nu_t')^2 \leqslant \mathbb{E}\Big(|Z_t - Z_t'|^2\Big) \leqslant (c + dt)^2 e^{-2\rho t} \mathbb{E}\Big(|Z_0 - Z_0'|^2\Big).$$

Taking the infimum over all couplings of ν_0 and ν_0' concludes the proof:

$$\inf_{\substack{(Z_0, Z_0') \\ Z_0 \sim \nu_0 \\ Z_0' \sim \nu_0'}} \mathbb{E}(|Z_0 - Z_0'|^2) =: W_2(\nu_0, \nu_0').$$

□

26.4 Further reading

In 1828, a botanist named Robert Brown[2] reported his observations: a pollen particle, visible under a microscope and immersed in water, contains grains whose movements are both rapid and irregular. The explanation that gradually became accepted was that the motion arose from collisions between the pollen particle and water molecules. Albert Einstein's work in 1905 is often presented as the first mathematical steps towards what would become the theory of Brownian motion, developed notably by Norbert Wiener[3] and Paul Lévy. However, shortly before Einstein's work, and in a very different context, Louis Bachelier had already obtained the distribution of Brownian motion in 1900 in his thesis "*Théorie de la spéculation*", which had little impact at the time. We also mention Marian Smoluchowski's work, contemporary with Einstein's. The model studied in this chapter was introduced by Paul Langevin

[2] Remembered for this discovery, which was incidental to his work.
[3] It is also called the Wiener process.

and refined by Leonard Ornstein and George Uhlenbeck in the 1930s. Edward Nelson's book [Nel67] presents the modelling approach at the start of this chapter, historical perspectives and mathematical extensions.

The Langevin process can be generalised beyond the comfortable world of Gaussian processes by considering the solution of the following stochastic differential equation:

$$\begin{cases} dX_t = V_t dt \\ dV_t = -\lambda V_t dt - \nabla U(X_t) dt + \sigma dB_t, \end{cases}$$

where the function U acts as a confining potential (think of a convex function). In this case, the invariant distribution is a product of two measures on the position and velocity spaces. When Brownian motion acts on all coordinates, the diffusion process is called nondegenerate; the Kolmogorov partial differential equation describing the evolution of its distribution at time t is called elliptic and has a solution of class $\mathcal{C}^\infty$. However, many evolution equations, such as kinetic equations describing a particle's velocity and position, fall outside this framework. In [Hör67], Lars Hörmander classifies degenerate diffusions whose distribution at time t is absolutely continuous with a smooth density.

Precise results on convergence to equilibrium for these processes are fairly recent. Denis Talay's article [Tal02] gives estimates for convergence of the time-dependent density to the density of the invariant distribution in weighted L^p spaces. Cédric Villani's work [Vil06, Vil09] offers another approach through functional inequalities (Poincaré and logarithmic Sobolev inequalities) and Lars Hörmander's theory of hypoellipticity.

The decay over time of relative entropy with respect to the invariant distribution (Theorem 26.10) holds in general for Markov models and may become very difficult to establish beyond this framework. For Ludwig Boltzmann's celebrated equation governing the evolution of the density of a gas of interacting particles, this decay is known as the H-theorem and remains a subject of study in mathematical physics. See Villani's expository article [Vil08].

27

From Markov chains to diffusion processes

Keywords. Convergence of processes; diffusion approximation; Euler scheme.

Tools. Martingale; stopping time; Ehrenfest chain; Wright-Fisher chain; M/M/∞ queue; Yule process; Ornstein-Uhlenbeck process; Brownian motion.

Difficulty. ***

Process convergence results are very useful and fairly intuitive, but require substantial mathematical machinery beyond the scope of this book. This chapter presents a short overview, if not a catalog, of the connections summarized in Figure 27.1 between processes with discrete or continuous state spaces and discrete or continuous time. Some results are only partially proved or merely stated, but the last section points to books for further study.

27.1 Laziness and sampling

Let $\mathbf{P}$ be a transition matrix on a finite state space E and $\mathbf{I}$ the identity matrix on E. For every $N \geqslant 1$,

$$\mathbf{P}_N = \left(1 - \frac{1}{N}\right)\mathbf{I} + \frac{1}{N}\mathbf{P} = \mathbf{I} + \frac{\mathbf{P} - \mathbf{I}}{N}$$

is a transition matrix. The chain $(X_n^N)_{n\in\mathbb{N}}$ associated with $\mathbf{P}_N$ remains in a state $i \in E$ for a geometrically distributed random time with parameter $(1-\mathbf{P}(i,i))/N$, then, if this probability is nonzero, jumps onto $E\backslash\{i\}$ according to the probability measure proportional to $(\mathbf{P}_N(i,j))_{j\in E\backslash\{i\}}$. We say that X^N for $N \geqslant 2$ is a *lazy chain* associated with $\mathbf{P}$. A chain of this kind is studied in Chapter 9 on the Ehrenfest model.

Set $Y_t^N = X_{\lfloor Nt\rfloor}$ for $t \in \mathbb{R}_+$. The sequence of processes $(Y^N)_{N\geqslant 1}$ converges to the continuous-time Markov process $(Y_t)_{t\geqslant 0}$ on E with generator $A = \mathbf{P} - \mathbf{I}$ and *semigroup* $(P_t)_{t\geqslant 0}$. Convergence at a given time t simply reads:

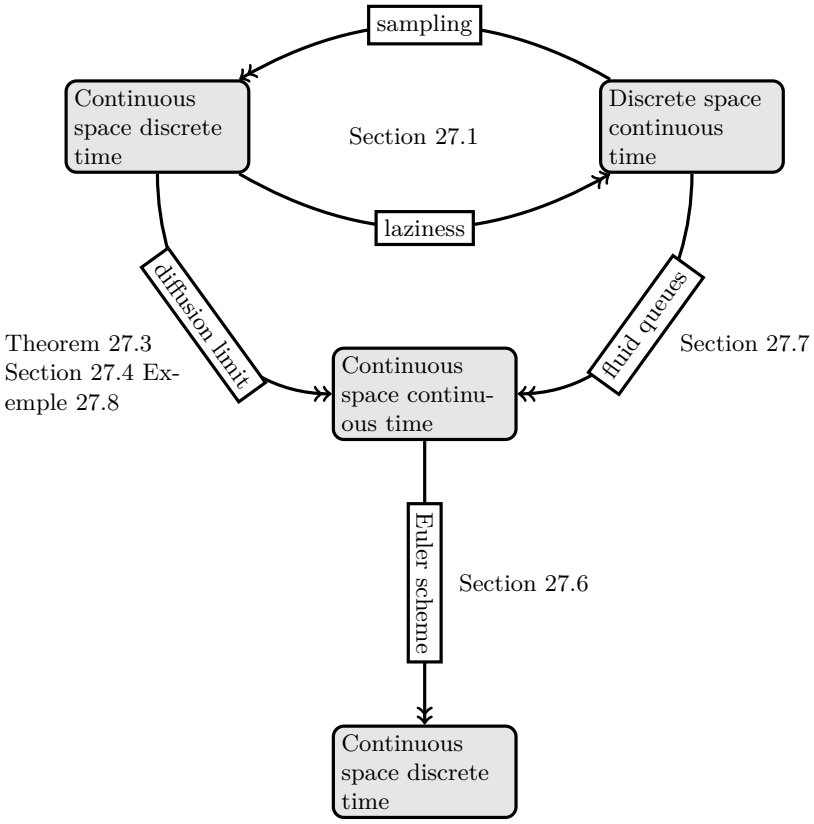


Fig. 27.1. Various connections between processes with discrete or continuous space and time presented in this chapter.

$$P_t = e^{tA} = \lim_{N\to\infty} \mathbf{P}_N^{\lfloor Nt \rfloor} = \lim_{N\to\infty} \left(\mathbf{I} + \frac{A}{N}\right)^{\lfloor Nt \rfloor}.$$

The Markov property extends this to all finite-dimensional distributions.

Conversely, if $(Y_t)_{t\in\mathbb{R}_+}$ is a continuous-time Markov chain on E with generator A, then

$$(Y_n^\delta)_{n\in\mathbb{N}} = (Y_{\delta n})_{n\in\mathbb{N}}$$

is a discrete-time Markov chain on E with transition matrix $e^{\delta A}$, called the *sampled chain*.

The Markov processes $(X_n^1)_{n\in\mathbb{N}}$, $(Y_t)_{t\in\mathbb{R}_+}$, and $(Y_n^\delta)_{n\in\mathbb{N}}$ are closely related: their recurrence and transience properties coincide, and they have the same invariant distributions. However, even if the chain $(X_n^1)_{n\in\mathbb{N}}$ is periodic, this is not the case for $(Y_t)_{t\in\mathbb{R}_+}$ and $(Y_t^\delta)_{t\in\mathbb{R}_+}$. Indeed, it is easy to show that if

there exists $n \in \mathbb{N}$ such that $\mathbb{P}_{x_0}(X_n^1 = x) > 0$, then $\mathbb{P}_{x_0}(Y_t = x) > 0$ for every $t \in \mathbb{R}_+$ and $\mathbb{P}_{x_0}(Y_m^\delta = x) > 0$ for every $m \in \mathbb{N}$.

Remark 27.1 (From Bernoulli to Poisson)**.** *An increasingly lazy Bernoulli process, suitably rescaled in time, converges to a Poisson process. This recovers, in particular, the law of small numbers encountered in Chapter 1.*

Remark 27.2 (Genealogy and laziness)**.** *In Chapter 13, the discrete-time chain describing the number of generations one must go back to find the most recent common ancestor in a population of size N converges, as $N \to \infty$, to a continuous-time pure death process in a manner similar to convergence of lazy chains.*

27.2 Convergence of a chain to a diffusion

The purpose of this section is to state a theorem ensuring convergence in distribution of a sequence of Markov chains, suitably rescaled in time and space, to a diffusion process with explicit diffusion and drift coefficients.

Theorem 27.3 (Diffusion approximation)**.** *Let $(\mathbf{P}_N)_{N\geqslant 1}$ be a sequence of transition kernels on $\mathbb{R}$, and set*

$$b_N(x) = N \int_{|y-x|\leqslant 1} (y-x)\mathbf{P}_N(x, dy)$$

and

$$a_N(x) = N \int_{|y-x|\leqslant 1} (y-x)^2\mathbf{P}_N(x, dy).$$

Suppose there exist two functions a and b with values in $]0, +\infty[$ and $\mathbb{R}$ respectively, bounded and of class $\mathcal{C}^1$, such that for all $r > 0$ and $\varepsilon > 0$,

$$\lim_{N\to\infty} \sup_{|x|\leqslant r} |b_N(x) - b(x)| = 0, \quad \textit{et} \quad \lim_{N\to\infty} \sup_{|x|\leqslant r} |a_N(x) - a(x)| = 0$$

and

$$\lim_{N\to\infty} \sup_{|x|\leqslant r} N\mathbf{P}_N(x, [x-\varepsilon, x+\varepsilon]^c) = 0.$$

Let $(Y_n^N)_{n\in\mathbb{N}}$ be the Markov chain with transition kernel $\mathbf{P}_N$ and $(Z_t^N)_{t\in\mathbb{R}_+}$ the process defined by $Z_t^N = Y_{\lfloor Nt\rfloor}^N$. If $(Y_0^N)_{N\geqslant 1}$ converges in distribution to ν, then, for every $T > 0$,

$$(Z_t^N)_{t\in[0,T]} \xrightarrow[N\to\infty]{\text{loi}} (X_t)_{t\in[0,T]}$$

where $(X_t)_{t\in[0,T]}$ is a diffusion process with initial distribution ν and generator A given by

$$Af(x) = \frac{1}{2}a(x)f''(x) + b(x)f'(x),$$

for every compactly supported function f of class $\mathcal{C}^2$.

Proof idea. The semigroup $(P_t)_{t\in\mathbb{R}_+}$ associated with the infinitesimal generator A satisfies, for every function f from $\mathbb{R}$ to $\mathbb{R}$ of class $\mathcal{C}^2$ with compact support and every $t \in \mathbb{R}_+$,

$$P_t f = f + \int_0^t P_s A f \, ds.$$

For every $N \geqslant 1$, write $\mathbf{A}_N = \mathbf{P}_N - \mathbf{I}$. Let $t > 0$. Then

$$\begin{aligned}\mathbf{P}_N^{\lfloor Nt \rfloor} f &= f + \sum_{i=0}^{\lfloor (N-1)t \rfloor} \mathbf{P}_N^i \mathbf{A}_N f = f + \frac{1}{N} \sum_{i=0}^{\lfloor (N-1)t \rfloor} \mathbf{P}_N^i (N\mathbf{A}_N) f \\ &= f + \sum_{i=0}^{\lfloor (N-1)t \rfloor} \int_{i/N}^{(i+1)/N} \mathbf{P}_N^i (N\mathbf{A}_N) f \, ds.\end{aligned}$$

We thus obtain, for every $t \in \mathbb{R}_+$,

$$\mathbf{P}_N^{\lfloor Nt \rfloor} f = f + \int_0^t \mathbf{P}_N^{\lfloor Ns \rfloor} (N\mathbf{A}_N) f \, ds.$$

Thus the family of probability measures $(\mathbf{P}_N^{\lfloor Nt \rfloor})_{t\in\mathbb{R}_+}$ is related to $N\mathbf{A}_N$ as $(P_t)_{t\in\mathbb{R}_+}$ is to A. Moreover, if f is of class $\mathcal{C}^2$ with compact support,

$$\begin{aligned}N\mathbf{A}_N f(x) &= N \int (f(y) - f(x)) \mathbf{P}_N(x, dy) \\ &= b_N(x) f'(x) + \frac{1}{2} a_N(x) f''(x) + \mathcal{O}(N\mathbf{P}_N(x, [x-1, x+1])).\end{aligned}$$

The assumptions of the theorem ensure that $N\mathbf{A}_N f$ converges to Af. It remains to deduce convergence in distribution of the associated processes. This step is beyond the scope of this book. Section 27.7 provides references for a complete proof. □

Remark 27.4 (Nothing is simple)**.** *In practice, the Markov chain Y^N usually takes values in a discrete and often finite subset I_N of $\mathbb{R}$. Its transition kernel $\mathbf{P}_N$ is therefore not defined on all of $\mathbb{R}$. One must then extend its definition outside I_N in a somewhat artificial way.*

Example 27.5 (Random walk and Brownian motion)**.** *Let S^N be the random walk on $\mathbb{Z}$ defined by*

$$S_n^N = \sum_{i=1}^n \varepsilon_i$$

where $(\varepsilon_i^N)_{i\geqslant 1}$ is a sequence of i.i.d. random variables with distribution $p_N\delta_1 + (1-p_N)\delta_{-1}$. In this case,

- *if* $p_N = 1/2$*, then* $(S^N_{\lfloor N\cdot\rfloor}/\sqrt{N})_{N\geqslant 1}$ *converges in distribution to Brownian motion starting from* 0*;*
- *if* $p_N = p \neq 1/2$*, then* $(S_{\lfloor N\cdot\rfloor}/N)_{N\geqslant 1}$ *converges in distribution to the deterministic process* $t \mapsto (2p-1)t$*;*
- *if* $p_N = 1/2 + \alpha/(2\sqrt{N})$*, then* $(S_{\lfloor N\cdot\rfloor}/N)_{N\geqslant 1}$ *converges in distribution to Brownian motion with drift* α *starting from* 0*.*

The first case of the example below is an instance of *Donsker's theorem*, an analogue of the central limit theorem for distributions of processes.

Theorem 27.6 (Donsker's invariance principle)**.** *Let* $(Y_n)_{n\geqslant 1}$ *be a sequence of i.i.d. centered random variables with variance 1. Write*

$$X^N_t = \frac{1}{\sqrt{N}}\left(\sum_{k=1}^{\lfloor Nt\rfloor} Y_k - (Nt - \lfloor Nt\rfloor)Y_{\lfloor Nt\rfloor+1}\right).$$

The sequence of processes $(X^N)_{N\geqslant 1}$ *converges in distribution to Brownian motion.*

Remark 27.7 (Linear interpolation)**.** *In Donsker's theorem, the paths of* X^N *are constructed by linear interpolation between points whose time coordinates are multiples of* $1/N$*. This precaution allows us to view convergence of* $(X^N_t)_{0\leqslant t\leqslant T}$ *as convergence of a random variable with values in* $\mathcal{C}([0,T])$ *equipped with the supremum norm and its Borel sigma-algebra.*

Example 27.8 (Ehrenfest chain and Ornstein-Uhlenbeck process)**.** *For* $N \geqslant 1$*, let* Z^N *be the Ehrenfest chain (see Chapter 9) with values in* $\{0,\dots,N\}$ *and transition matrix given by*

$$\mathbb{P}(Z^N_1 = j | Z^N_0 = i) = \begin{cases} i/N & \text{si } j = i-1, \\ 1 - i/N & \text{si } j = i+1, \\ 0 & \text{sinon.} \end{cases}$$

Define the chain $(Y^N_n)_{n\geqslant 0}$ *with values in*

$$E_N = \left\{-(1/2)\sqrt{N} + k/\sqrt{N} \,:\, 0 \leqslant k \leqslant N\right\}$$

by setting

$$Y^N_n = \sqrt{N}\left(\frac{Z^N_n}{N} - \frac{1}{2}\right).$$

Its transition kernel is given by

$$\mathbf{P}_N(x,\cdot) = \left(\frac{1}{2} + \frac{x}{\sqrt{N}}\right)\delta_{x-1/\sqrt{N}}(\cdot) + \left(\frac{1}{2} - \frac{x}{\sqrt{N}}\right)\delta_{x+1/\sqrt{N}}(\cdot),$$

for $x \in E_N$*. After extending* $\mathbf{P}_N$ *suitably to* $\mathbb{R}$*, we obtain*

$$b_N(x) = N\left(\left[\frac{1}{2} + \frac{x}{\sqrt{N}}\right]\left[x - \frac{1}{\sqrt{N}}\right] + \left[\frac{1}{2} - \frac{x}{\sqrt{N}}\right]\left[x + \frac{1}{\sqrt{N}}\right] - x\right) = -2x.$$

Similarly, $a_N(x) = 1$. *The concentration assumption on the kernel* $\mathbf{P}_N(x,\cdot)$ *is clearly satisfied, since its support converges to* $\{x\}$. *Thus, if* $(Y_0^N)_{N\geqslant 1}$ *converges in distribution to* ν, *then the sequence of processes* $(Y^N_{\lfloor N\cdot\rfloor})_{N\geqslant 1}$ *converges in distribution to the Ornstein-Uhlenbeck process with initial distribution* ν *solving the stochastic differential equation*

$$dX_t = dB_t - 2X_t\,dt.$$

27.3 Diffusion on an interval

Before giving further important examples, we collect in this section some classical results on diffusion processes with values in $\mathbb{R}$. Consider the process $(X_t)_{t\geqslant 0}$ solving the stochastic differential equation

$$dX_t = \sigma(X_t)\,dB_t + b(X_t)\,dt,$$

where σ and b are smooth functions on the interval $I =]l, r[$ with values in $]0,\infty[$ and $\mathbb{R}$ respectively. The infinitesimal generator of X is the operator acting on functions f of class $\mathcal{C}^2$ on I given by

$$Lf(x) = \frac{1}{2}\sigma^2(x)f''(x) + b(x)f'(x).$$

Let $c \in I$. Introduce

- the *scale function* p:

$$p(x) = \int_c^x \exp\left(-2\int_c^y \frac{b(z)\,dz}{\sigma^2(z)}\right)dy,$$

- the *speed measure* m, defined by

$$m(dx) = \mathbb{1}_I(x)\frac{2}{p'(x)\sigma^2(x)}\,dx = \mathbb{1}_I(x)\frac{2}{\sigma^2(x)}\exp\left(2\int_c^x \frac{b(z)\,dz}{\sigma^2(z)}\right)dx,$$

- and, for $l < a < b < r$, the function on $]a,b[$,

$$M_{a,b}(x) = -\int_a^x (p(x) - p(y))m(dy) + \frac{p(x)-p(a)}{p(b)-p(a)}\int_a^b (p(b) - p(y))m(dy).$$

The functions p and $M_{a,b}$ satisfy the following crucial relations:

$$\begin{cases} Lp(x) = 0 \quad \text{pour } x \in I, \\ p(c) = 0, \\ p'(c) = 1, \end{cases} \qquad \text{et} \qquad \begin{cases} LM_{a,b}(x) = -1 \quad \text{pour } x \in]a,b[, \\ M_{a,b}(a) = 0, \\ M_{a,b}(b) = 0. \end{cases}$$

Theorem 27.9 (Exit from an interval)**.** *Let $l < a < b < r$ and let*

$$T_{a,b} = \inf\{t \geqslant 0 \,:\, X_t \notin [a,b]\}$$

be the exit time from the interval $[a,b]$ for X. Then

$$\mathbb{P}_x\big(X_{T_{a,b}} = a\big) = 1 - \mathbb{P}_x\big(X_{T_{a,b}} = b\big) = \frac{p(b) - p(x)}{p(b) - p(a)},$$

and

$$\mathbb{E}_x(T_{a,b}) = M_{a,b}(x).$$

Remark 27.10 (Brownian motion)**.** *This generalizes the familiar Brownian motion case, in which the probability of leaving the interval $[a,b]$ at a when starting from x equals $(b-x)/(b-a)$.*

Proof. The function p is increasing by definition. Thus, for $a < y < b$, the functions

$$x \in [y,b] \mapsto \frac{p(x) - p(y)}{p(x) - p(a)} \quad \text{et} \quad x \in [a,b] \mapsto \int_a^x \frac{p(x) - p(y)}{p(x) - p(a)} m(dy)$$

are increasing. Consequently, $M_{a,b}$ is nonnegative on $[a,b]$. Set, for $n \in \mathbb{N}^*$,

$$\tau_n = \inf\left\{t \geqslant 0\,;\, \int_0^t \sigma^2(X_s)\,ds \geqslant n\right\}.$$

Since $LM_{a,b}(x) = -1$ for $x \in \,]a,b[$, Itô's formula for $M_{a,b}(X_t)$ gives

$$M_{a,b}(X_{t\wedge\tau_n\wedge T_{a,b}}) = M_{a,b}(X_0) - (t \wedge \tau_n \wedge T_{a,b}) + \int_0^{t\wedge\tau_n\wedge T_{a,b}} M'_{a,b}(X_s)\sigma(X_s)dB_s.$$

Stopping at time $\tau_n \wedge T_{a,b}$ ensures that the stochastic integral above is a martingale. Taking expectations with $X_0 = x$ and using nonnegativity of $M_{a,b}$ gives

$$\mathbb{E}_x(t \wedge \tau_n \wedge T_{a,b}) = M_{a,b}(x) - \mathbb{E}_x\big[M_{a,b}(X_{t\wedge\tau_n\wedge T_{a,b}})\big] \leqslant M_{a,b}(x) < \infty.$$

It remains to let n and then t tend to $+\infty$ to obtain, by monotone convergence, that $\mathbb{E}_x(T_{a,b})$ is finite. The time $T_{a,b}$ is therefore finite almost surely, and dominated convergence yields

$$\lim_{t\to\infty} \mathbb{E}\big[M_{a,b}(X_{t\wedge T_{a,b}})\big] = \mathbb{E}\big[M_{a,b}(X_{T_{a,b}})\big] = 0,$$

since $M_{a,b}$ is continuous and bounded on $[a,b]$ and vanishes at a and b. This gives the expression for $\mathbb{E}_x(T_{a,b})$.

Similarly, Itô's formula applied to $p(X_t)$ gives

$$p(x) = \mathbb{E}_x\big[p(X_{T_{a,b}})\big] = p(a)\mathbb{P}_x\big(X_{T_{a,b}} = a\big) + p(b)\mathbb{P}_x\big(X_{T_{a,b}} = b\big).$$

We conclude using the relation $\mathbb{P}_x\big(X_{T_{a,b}} = a\big) + \mathbb{P}_x\big(X_{T_{a,b}} = b\big) = 1$. □

27.4 Application to Wright-Fisher processes

Consider the Wright-Fisher model with mutation and selection introduced in Chapter 12. It is the Markov chain $(X_n^N)_{n\in\mathbb{N}}$ on $\{0,\dots,N\}$ with transition

$$\mathrm{Loi}(X_{n+1}^N | X_n^N = x) = \mathrm{Bin}(N, \eta_x)$$

where

$$\eta_x = \psi_x(1-u_A) + (1-\psi_x)u_B \quad \text{et} \quad \psi_x = \frac{(1+s)x}{(1+s)x + N - x}.$$

The real number s represents the selective advantage of A over B, and u_A is the probability of mutation from allele A to allele B. The proportion of A alleles in a population of size N at generation n is

$$Y_n^N := \frac{X_n^N}{N}.$$

For every $t \in \mathbb{R}_+$, set

$$Z_t^N := Y_{\lfloor Nt \rfloor}^N.$$

Theorem 27.11 (Wright-Fisher diffusion limit)**.** *Suppose the selection rate s and mutation rates u_A and u_B (which depend on N) satisfy*

$$(Ns, Nu_A, Nu_B) \underset{N\to\infty}{\longrightarrow} (\alpha, \beta_A, \beta_B).$$

Then the sequence of processes $(Z_\cdot^N)_{N\geqslant 1}$ converges in distribution to the diffusion process $(Z_t)_{t\in\mathbb{R}_+}$ solving the following stochastic differential equation with values in $[0,1]$:

$$dZ_t = \sigma(Z_t)\, dB_t + b(Z_t)\, dt,$$

where

$$\sigma^2(z) = z(1-z) \quad \textit{et} \quad b(z) = \alpha z(1-z) - \beta_A z + \beta_B(1-z).$$

Proof. Let $z \in E_N = \{0, 1/N, \dots, 1\}$. Then

$$\begin{aligned}\eta_{Nz} &= \psi_{Nz}(1-u_A) + (1-\psi_{Nz})u_B \\ &= z + \frac{-\beta_A z + \beta_B(1-z) + \alpha z(1-z)}{N} + \mathcal{O}(1/N).\end{aligned}$$

Since the distribution of NZ_1^N given $Z_0^N = z$ is $\mathrm{Bin}(N, \eta_{Nz})$, we have

$$\begin{aligned}N\mathbb{E}\big[Z_1^N - z | Z_0^N = z\big] &= N(\eta_{Nz} - z) \\ &= -\beta_A z + \beta_B(1-z) + \alpha z(1-z) + \mathcal{O}(1/N).\end{aligned}$$

Similarly,

$$N\mathbb{E}\Big[\big(Z_1^N - z\big)^2 | Z_0^N = z\Big] = \eta_{Nz}(1-\eta_{Nz}) \\ = z(1-z) + \mathcal{O}(1/N).$$

Finally, Markov's inequality yields

$$N\mathbb{P}(Z_1^N \in [z-\varepsilon, z+\varepsilon] | Z_0^N = z) \leqslant \frac{N\mathbb{E}\Big[\big(Z_1^N - z\big)^4 | Z_0^N = z\Big]}{\varepsilon^4}.$$

Now

$$N\mathbb{E}\Big[\big(Z_1^N - z\big)^4 | Z_0^N = z\Big] = \mathcal{O}\big(N^{-1}\big),$$

because the fourth centered moment of a random variable X with distribution $\mathrm{Bin}(N,p)$ is given by

$$\mathbb{E}\Big[(X-Np)^4\Big] = (3Np(1-p) + 1 - 6p(1-p))Np(1-p) = \mathcal{O}(N^2).$$

These estimates are uniform in $z \in E_N$. One final precaution is needed. The diffusion coefficient σ is not Lipschitz at 0 and 1. We must therefore use Theorem 27.3 for processes stopped when they fall below ε or rise above $1-\varepsilon$, for some small parameter $\varepsilon > 0$. □

Remark 27.12 (Time scales)**.** *To transfer absorption-time estimates from the diffusion to the Markov chain, one must multiply them by N. Hitting probabilities and invariant distribution profiles, on the other hand, are unaffected by this time rescaling.*

Let us express the scale function: for $c, x \in]0,1[$,

$$S(x) = \int_c^x y^{-2\beta_B}(1-y)^{-2\beta_A} e^{2\alpha y}\, dy,$$

and the speed measure:

$$m(dx) = 2x^{2\beta_B - 1}(1-x)^{2\beta_A - 1} e^{-2\alpha x}\, dx.$$

We can now consider various examples depending on the model parameters.

Example 27.13 (Neither mutation nor selection)**.** *If $\alpha = \beta_A = \beta_B = 0$, the diffusion has zero drift: it is a martingale, we have $\mathbb{P}_x(Z_{T_{0,1}} = 1) = x$, and the expected fixation time is given by*

$$\mathbb{E}_x(T_{0,1}) = -2[x\log(x) + (1-x)\log(1-x)].$$

This is illustrated in Figure 27.2.

Example 27.14 (Selection without mutation)**.** *Choose $\alpha > 0$ and $\beta_A = \beta_B = 0$. Then $T_{0,1}$ is finite almost surely, and the probability that allele B disappears is*

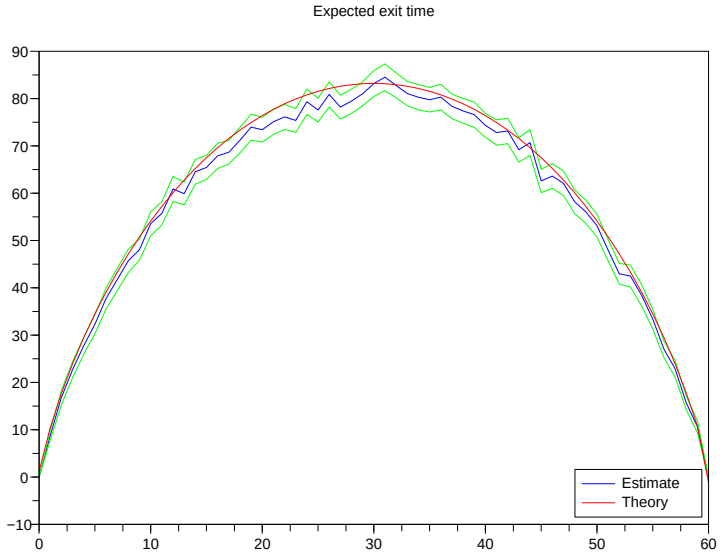


Fig. 27.2. Estimated fixation time for $N = 40$ in the absence of selection and mutation.

$$\mathbb{P}_x(Z_{T_{0,1}} = 1) = \frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha}}.$$

Figure 27.3 illustrates the estimated fixation probability of allele A. If $N = 10^5$, $s = 10^{-4}$, *and* $x = 0,5$, *then* $\alpha = 20$ *and the fixation probability of allele A is approximately* $0,999955$. *This small selective advantage s, undetectable in the laboratory or through statistical measurements, is nevertheless sufficient to have a decisive effect on allele fixation. Computing the expected fixation time is more delicate. The following differential equation:*

$$z''(x) + 2\alpha z'(x) = -\frac{1}{x(1-x)}, \quad \textit{avec } z(0) = 0 \textit{ et } z(1) = 0,$$

has no explicit solution. Figure 27.4 provides an estimate of the absorption time for $N = 30$ *and* $\alpha = 10$.

Example 27.15 (Mutation without selection)**.** *Here assume* $\alpha = 0$. *The diffusion solves*

$$dX_t = \sqrt{X_t(1 - X_t)}\, dB_t + (-\beta_A X_t + \beta_B(1 - X_t))\, dt.$$

This diffusion remains in the interval $]0,1[$, *that is,* $T_{0,1}$ *is infinite almost surely. Moreover, its invariant distribution is* ν, *namely the distribution* $\mathrm{Beta}(2\beta_B, 2\beta_A)$, *that is,*

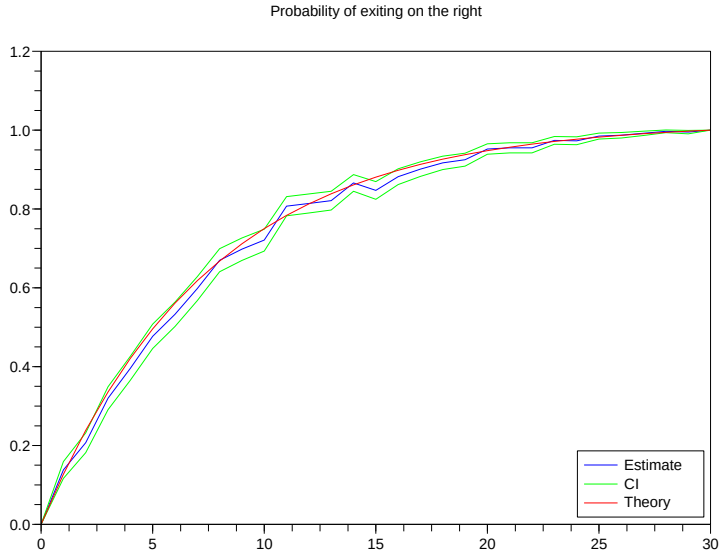


Fig. 27.3. Fixation probability of A for $N = 30$ and $\alpha = 2$.

$$\nu(dy) = \frac{\Gamma(2\beta_A + 2\beta_B)}{\Gamma(2\beta_A)\Gamma(2\beta_B)} y^{2\beta_B - 1}(1-y)^{2\beta_A - 1}\, dy.$$

In particular, if Y has distribution ν,

$$\mathbb{E}(Y) = \frac{\beta_B}{\beta_A + \beta_B} \quad et \quad \mathrm{Var}(Y) = \frac{\beta_A \beta_B}{(\beta_A + \beta_B)^2(2(\beta_A + \beta_B) + 1)}.$$

Remark 27.16 (Multiple-allele model)**.** *In the model with k distinct alleles, total mutation rate $\theta/2$, and respective mutation probabilities $(\pi_i)_{1 \leqslant i \leqslant k}$, the process of allele proportions takes values in the simplex Δ of $\mathbb{R}^k$, of dimension $k-1$, and the density of its invariant distribution with respect to Lebesgue measure on Δ is*

$$\nu(y_1, \ldots, y_K) = \frac{\Gamma(\theta)}{\Gamma(\theta\pi_1) \cdots \Gamma(\theta\pi_K)} y_1^{\Gamma(\theta\pi_1)-1} \cdots y_K^{\Gamma(\theta\pi_K)-1} \mathbb{1}_\Delta(y_1, \ldots, y_k).$$

27.5 Empirical distribution function

Let $(U_n)_{n \geqslant 1}$ be a sequence of i.i.d. random variables uniform on $[0,1]$. The empirical distribution function F^N of $U_1, \ldots, U_N$ is defined by

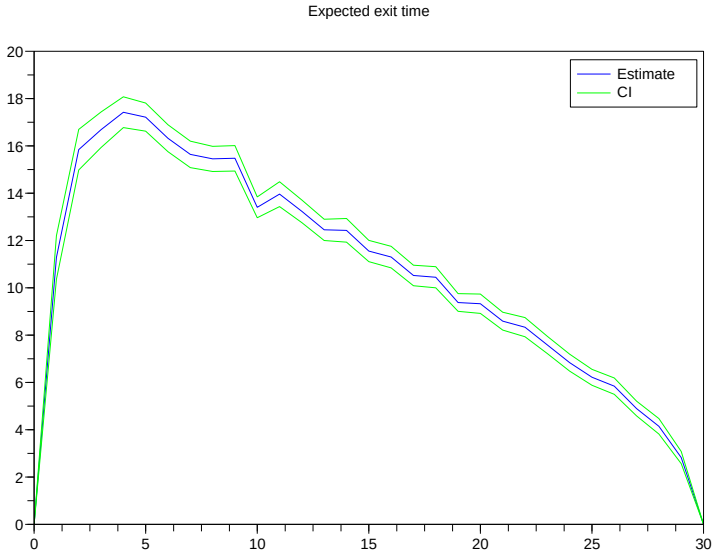


Fig. 27.4. Expected fixation time for $N = 30$ and $\alpha = 10$.

$$F^N(x) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{\{U_k \leqslant x\}} \quad \text{pour } x \in \mathbb{R}.$$

Introduce the process $(H_x^N)_{x\in[0,1]}$ defined by $H_x^N = \sqrt{N}(F^N(x) - x))$ for $x \in [0,1]$. The strong law of large numbers and the central limit theorem ensure that, for every $x \in [0,1]$,

$$F^N(x) \xrightarrow[N\to\infty]{\text{p.s.}} x \quad \text{et} \quad H_x^N \xrightarrow[N\to\infty]{\text{loi}} \mathcal{N}(0, x(1-x)).$$

Note that $H_0^N = H_1^N = 0$. Moreover, for all $0 \leqslant y \leqslant x \leqslant 1$,

$$\mathrm{Cov}(H_y^N, H_x^N) = y(1-x)$$

since $\mathrm{Cov}(\mathbb{1}_{\{U_k\leqslant y\}}, \mathbb{1}_{\{U_k\leqslant x\}}) = y(1-x)$. This calculation readily extends to convergence of any finite-dimensional marginal $(H_{x_1}^N, \ldots, H_{x_k}^N)$. We then obtain the following result.

Theorem 27.17 (Brownian bridge)**.** *The sequence of processes* $(H^N)_{N\geqslant 1}$ *converges to the Gaussian process* $(M_x)_{x\in[0,1]}$*, which is centered and has covariance*

$$\mathrm{Cov}(M_y, M_x) = y(1-x)$$

for $0 \leqslant y \leqslant x \leqslant 1$*. This process is called the Brownian bridge.*

This convergence in distribution of the empirical process, combined with continuity properties of the supremum and L^2 norms on a space of processes, recovers the Kolmogorov-Smirnov and Cramér-Von Mises theorems.

Corollary 27.18 (Kolmogorov-Smirnov and Cramér-Von Mises)**.**

$$\sup_{x\in[0,1]} \left|H_x^N\right| \xrightarrow[N\to\infty]{\text{loi}} \sup_{x\in[0,1]} |M_x| \quad et \quad \int_0^1 (H_x^N)^2\,dx \xrightarrow[N\to\infty]{\text{loi}} \int_0^1 M_x^2\,dx.$$

The remarkable properties of Brownian motion allow the distributions of the random variables in the corollary above to be determined.

Remark 27.19 (Alternative definitions)**.** *The Brownian bridge can be defined in several other ways:*

- *standard Brownian motion on the time interval* $[0,1]$*, conditioned to vanish at times 0 and 1;*
- *the process* $(M_t)_{t\in[0,1]}$ *given by* $M_t = B_t - tB_1$*, where* $(B_t)_{t\geqslant 0}$ *is a standard Brownian motion;*
- *the solution of the time-inhomogeneous stochastic differential equation*

$$dM_t = dB_t - \frac{M_t}{1-t}\,dt.$$

Unlike the second, the last formulation shows that the Brownian bridge is a Markov process.

27.6 Euler scheme

Consider the diffusion process $(X_t)_{t\geq 0}$ solving the following stochastic differential equation:

$$dX_t = \sigma(X_t)\,dB_t + b(X_t)dt.$$

Except in special cases, this equation has no explicit solution. We therefore seek to approximate it by a Markov chain that is easy to simulate. The simplest idea, and the most robust in practice, is to proceed as for ordinary differential equations, introducing an Euler scheme associated with a time discretization. Let $\alpha > 0$. The *Euler scheme* with step size α is the Markov chain $(Y_n^\alpha)_{n\geqslant 0}$ taking values in $\mathbb{R}$ and associated with the kernel

$$\mathbf{P}_\alpha(x,\cdot) = \mathcal{N}\left(x + \alpha b(x), \alpha\frac{\sigma(x)^2}{2}\right).$$

It can be extended to a diffusion process $(X_t^\alpha)_{t\in\mathbb{R}_+}$ whose coefficients are constant between times that are multiples of α: for $t \in [\alpha n, \alpha(n+1)[$,

$$X_t^\alpha = X_{n\alpha}^\alpha + \int_{n\alpha}^t b(X_{n\alpha}^\alpha)\,ds + \int_{t_\alpha}^t \sigma(X_{n\alpha}^\alpha)dB_s$$

$$= X^\alpha_{n\alpha} + (t - n\alpha)b(X^\alpha_{n\alpha}) + \sigma(X^\alpha_{n\alpha})(B_t - B_{n\alpha}).$$

Then $Y^\alpha_n = X^\alpha_{n\alpha}$ for every $n \in \mathbb{N}$.

Theorem 27.20 (Weak error)**.** *If the coefficients b and σ are functions of class $\mathcal{C}^4$ with bounded derivatives, then for every bounded measurable function f and every $t > 0$, there exists $C > 0$ such that*

$$|\mathbb{E}_x f(X_t) - \mathbb{E}_x f(X^\alpha_t)| \leqslant C\alpha.$$

Proof. We sketch the proof. Denote by $(P_t)_{t\geqslant 0}$ the semigroup associated with $(X_t)_{t\geqslant 0}$ and by L its infinitesimal generator. Write

$$\mathbf{P}^n_\alpha f - P_{\alpha n} f = \sum_{k=1}^{n} \mathbf{P}^{k-1}_\alpha (\mathbf{P}_\alpha - P_\alpha) P_{\alpha(n-k)} f.$$

The expansions of the two probability kernels P_α and $\mathbf{P}_\alpha$ agree through order 2. Indeed, since $\partial_t P_t = LP_t$,

$$P_\alpha g(x) = g(x) + \alpha Lg(x) + \mathcal{O}(\alpha^2)$$

Let Y have distribution $\mathcal{N}(0,1)$. Setting $\delta(x) = \sqrt{\alpha}\sigma(x)Y + \alpha b(x)$, we have

$$\begin{aligned}\mathbf{P}_\alpha g(x) &= \mathbb{E} g(x + \alpha b(x) + \sqrt{\alpha}\sigma(x)Y) \\ &= \mathbb{E}\left(g(x) + \delta(x)g'(x) + \frac{\delta(x)^2}{2} g''(x) + \frac{\delta(x)^3}{6} g'''(x) + \mathcal{O}(\delta(x)^3) \right) \\ &= g(x) + \alpha Lg(x) + \mathcal{O}(\alpha^2)\end{aligned}$$

since the moments of orders 1 and 3 of Y vanish. The error between the diffusion and the Euler scheme over one time step α is therefore of order α^2. After n iterations, the error is thus of order α (since αn is of order t). A major difficulty remains: controlling the derivatives of $P_{\alpha(n-k)} f$ to justify the expansion above. This follows from regularity properties of the solution of the parabolic partial differential equation $\partial_t u = Lu$. □

One can also control the pathwise discrepancy between the processes X and X^α, assuming they are constructed using the same Brownian motion.

Theorem 27.21 (Strong error)**.** *If there exists $C > 0$ such that, for all $x, y \in \mathbb{R}$,*

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leqslant C|x - y|,$$

then, for every $p \geqslant 1$ and every $T > 0$, there exists $K_p(T)$ such that

$$\mathbb{E}\left(\sup_{0\leqslant t\leqslant T} |X_t - X^\alpha_t|^p \right) \leqslant K_p(T)\alpha^{p/2}.$$

The proof is technical but fairly direct. It relies essentially on Itô's formula, Grönwall's lemma, and the Burkholder-Davis-Gundy inequality.

27.7 Further reading

Many books are devoted to stochastic calculus and diffusion processes. In particular, the book by Iosif Gikhman and Anatoliy Skorokhod [GS72] studies diffusions on an interval of $\mathbb{R}$ and their behavior at the boundary of the domain. The various versions and subtleties of process convergence results are collected in the book [EK86] by Stewart Ethier and Thomas Kurtz. Section 7.4 of this very dense work addresses, in particular, diffusion approximation and provides a proof of Theorem 27.3.

Let us mention two further classical examples of convergence of continuous-time Markov processes with discrete state spaces to a diffusion.

For $N \in \mathbb{N}^*$, let $Z^{N\lambda,\mu}$ be the M/M/∞ queue, studied in Chapter 11, with arrival rate $N\lambda$ and service rate μ. For all $N \geqslant 1$ and $t \geqslant 0$, set

$$X_t^N = \sqrt{N}\left(\frac{Z_t^{N\lambda,\mu}}{N} - \frac{\lambda}{\mu}\right) \quad \text{avec} \quad Z_0^{N\lambda,\mu} = \lfloor N\lambda/\mu + \sqrt{N}x \rfloor.$$

The sequence $(X^N_{\lfloor N\cdot \rfloor})_{N\geqslant 1}$ converges in distribution to the Ornstein-Uhlenbeck process solving

$$dX_t = \sqrt{2\lambda}\, dB_t - \mu X_t\, dt \quad \text{avec} \quad X_0 = x,$$

This diffusion approximation of a queue is widely used in practice. This model is called a fluid queue. See Philippe Robert's book [Rob00].

Similarly, let $(Y^N)_{N\geqslant 1}$ be a sequence of birth-and-death processes with respective rates $(\lambda_N n)_{n\in\mathbb{N}}$ and $(\mu_N n)_{n\in\mathbb{N}}$. Assume

$$N(\lambda_N - \mu_N) = b + \mathcal{O}(1/N) \quad \text{et} \quad \lambda_N + \mu_N = 2a + \mathcal{O}(1/N).$$

Then the sequence of processes $(Y^N_{\lfloor N\cdot \rfloor}/N)_{N\geqslant 1}$ converges to the diffusion X solving the following differential equation:

$$dX_t = \sqrt{2aX_t}\, dB_t + bX_t\, dt.$$

Euler schemes associated with diffusion processes have been extensively studied. We mention the survey books by Peter Kloeden and Eckhard Platen [KP95], and by Nicolas Bouleau and Dominique Lépingle [BL94] for the proof of Theorem 27.21, as well as the book by Carl Graham and Denis Talay [GT13]. Theorem 27.20 can be improved by obtaining an expansion of the weak error in powers of α. This allows convergence to be accelerated: combining schemes with step sizes 2α and α yields an error of order α^2! This method, called Romberg extrapolation, is also used in numerical analysis of ordinary differential equations. The two foundational works in this direction are due to Denis Talay and Luciano Tubaro [TT90], who assume that the test function is very smooth, and then Vlad Bally and Denis Talay [BT95], who use Malliavin calculus to obtain the result for bounded measurable functions when the coefficients σ and b are smooth. Denis Talay's contribution [Tal96] also provides a survey of weak-error estimates.

Suggested reading

This short chapter offers a necessarily incomplete and subjective selection of suggested reading on probability at master's level. Among the most accessible works are the indispensable classic by William Feller [Fel68, Fel71], the more recent book by Alan Gut [Gut13], and the books by Philippe Barbe and Michel Ledoux [BL98], Jean Jacod and Philip Protter [JP02, JP03], Daniel Revuz [Rev97], Vivek Borkar [Bor95], Paul Toulouse [Tou99], David Williams [Wil91], Rick Durrett [Dur10a], Pierre Brémaud [Bré99], and Albert Shiryaev [Shi96]. We also mention the monumental treatise by Claude Dellacherie and Paul-André Meyer [DM80].

Discrete-time martingales are the main subject of the books by Jacques Neveu [Nev72, Nev75], Peter Hall and Christopher Heyde [HH80], and David Williams [Wil91].

Markov chains with discrete time and state space are the main subject of the books by Paoli Baldi, Laurent Mazliak and Pierre Priouret [BMP08], Olle Häggström [Häg02], and David Levin, Yuval Peres and Elizabeth Wilmer [LPW09]. The book by James Norris [Nor98b] also covers continuous-time Markov processes with countable state spaces and martingales. We also mention the books by Michel Benaïm and Nicole El Karoui [BEK05], Jean-François Delmas and Benjamin Jourdain [DJ06], Étienne Pardoux [Par07], and Thierry Bodineau [Bod14], which also present important or classical applications (physics, biology, financial mathematics, etc.). For a similar approach that also connects with the modelling component of the mathematics agrégation examination, see [BC07].

The book by Christiane Cocozza-Thivent [CT97] presents applications of probability to reliability. Extreme value theory is covered, for example, in the book by Paul Embrechts, Claudia Klüppelberg and Thomas Mikosch [EKM97], while the theory of large deviation principles is covered, for example, in the book by Amir Dembo and Ofer Zeitouni [DZ10].

Stochastic calculus is covered, for example, in the books by Jean-François Le Gall [LG13], Bert Øksendal [Øks03], Daniel Revuz and Marc Yor [RY99], Kai Lai Chung and David Williams [CW14], David Williams and Chris Rogers

[RW00], and Ioannis Karatzas and Steven Shreve [KS91]. Finally, the book by Carl Graham and Denis Talay [GT13] also presents results on the approximation of diffusions by Euler schemes.

References

AB99. R. Albert and A.-L. Barabási – "Emergence of scaling in random networks", *Science* (1999), no. 286, p. 509–512. 208

AB02. — , "Statistical mechanics of complex networks", *Rev. Modern Phys.* **74** (2002), no. 1, p. 47–97. 208

ABC+00. C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto and G. Scheffer – *Sur les inégalités de Sobolev logarithmiques*, Panoramas et Synthèses [Panoramas and Syntheses], vol. 10, Société Mathématique de France, Paris, 2000, With a preface by Dominique Bakry and Michel Ledoux. 336

ABR08. L. Addario-Berry and B. A. Reed – "Ballot theorems, old and new", Horizons of combinatorics, Bolyai Soc. Math. Stud., vol. 17, Springer, Berlin, 2008, p. 9–35. 52

ABT03. R. Arratia, A. D. Barbour and S. Tavaré – *Logarithmic combinatorial structures: a probabilistic approach*, EMS Monographs in Mathematics, European Mathematical Society (EMS), Zürich, 2003. 193

ACG+99. G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela and M. Protasi – *Complexity and approximation*, Springer-Verlag, Berlin, 1999, Combinatorial optimization problems and their approximability properties, With 1 CD-ROM (Windows and UNIX). 272

AF01. D. J. Aldous and J. A. Fill – "Reversible Markov Chains and Random Walks on Graphs", unpublished lecture notes available online, 1994-2001. 16, 35

AG13a. A. Asselah and A. Gaudillière – "From logarithmic to subdiffusive polynomial fluctuations for internal dla and related growth models", *The Annals of Probability* **41** (2013), no. 3A, p. 1115–1159. 87

AG13b. — , "Sublogarithmic fluctuations for internal dla", *The Annals of Probability* **41** (2013), no. 3A, p. 1160–1179. 87

AGZ10. G. W. Anderson, A. Guionnet and O. Zeitouni – *An introduction to random matrices*, Cambridge Studies in Advanced Mathematics, vol. 118, Cambridge University Press, Cambridge, 2010. 290

AHU75. A. V. Aho, J. E. Hopcroft and J. D. Ullman – *The design and analysis of computer algorithms*, Addison-Wesley Publishing Co.,

Reading, Mass.-London-Amsterdam, 1975, Second printing, Addison-Wesley Series in Computer Science and Information Processing. 272

AK90. D. Aldous and W. B. Krebs – "The "birth-and-assassination" process", *Statist. Probab. Lett.* **10** (1990), no. 5, p. 427–430. 297

Ald85. D. J. Aldous – "Exchangeability and related topics", École d'été de probabilités de Saint-Flour, XIII—1983, Lecture Notes in Math., vol. 1117, Springer, Berlin, 1985, p. 1–198. 193

Ald91. — , "Threshold limits for cover times", *J. Theoret. Probab.* **4** (1991), no. 1, p. 197–211. 16

AN04. K. B. Athreya and P. E. Ney – *Branching processes*, Dover Publications Inc., Mineola, NY, 2004, Reprint of the 1972 original [Springer, New York; MR0373040]. 52, 237

And87. D. André – "Solution directe du problème résolu par M. Bertrand", *Comptes Rendus Acad. Sci. Paris* **105** (1887), p. 436–437. 35

AS80. D. B. Arnold and M. R. Sleep – "Uniform random generation of balanced parenthesis strings", *ACM Trans. Program. Lang. Syst.* **2** (1980), p. 122–128. 63

Asm00. S. Asmussen – *Ruin probabilities*, Advanced Series on Statistical Science & Applied Probability, vol. 2, World Scientific Publishing Co. Inc., River Edge, NJ, 2000. 246

Bak94. D. Bakry – "L'hypercontractivité et son utilisation en théorie des semigroupes", Lectures on probability theory (Saint-Flour, 1992), Lecture Notes in Math., vol. 1581, Springer, Berlin, 1994, p. 1–114. 335

Bas95. R. F. Bass – *Probabilistic techniques in analysis*, Probability and its Applications (New York), Springer-Verlag, New York, 1995. 318

Bax82. R. J. Baxter – *Exactly solved models in statistical mechanics*, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London, 1982. 76

BB03. F. Baccelli and P. Brémaud – *Elements of queueing theory*, second ed., Applications of Mathematics (New York), vol. 26, Springer-Verlag, Berlin, 2003, Palm martingale calculus and stochastic recurrences, Stochastic Modelling and Applied Probability. 148

BC03. C. Biernacki and S. Chrétien – "Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM", *Statist. Probab. Lett.* **61** (2003), no. 4, p. 373–382. 106

BC07. B. Bercu and D. Chafaï – *Modélisation stochastique et simulation, cours et applications*, Collection Sciences Sup, Dunod, 2007, Mathématiques appliquées/SMAI. 76, 87, 98, 365

BD02. P. J. Brockwell and R. A. Davis – *Introduction to time series and forecasting*, second ed., Springer Texts in Statistics, Springer-Verlag, New York, 2002, With 1 CD-ROM (Windows). 98

BDMT11. V. Bansaye, J.-F. Delmas, L. Marsalle and V. C. Tran – "Limit theorems for Markov processes indexed by continuous time Galton-Watson trees", *Ann. Appl. Probab.* **21** (2011), no. 6, p. 2263–2314. 237

BÉ85. D. Bakry and M. Émery – "Diffusions hypercontractives", Séminaire de probabilités, XIX, 1983/84, Lecture Notes in Math., vol. 1123, Springer, Berlin, 1985, p. 177–206. 336

BEK05. M. Benaïm and N. El Karoui – *Promenade aléatoire: chaînes de markov et simulations ; martingales et stratégies*, Mathématiques appliquées, Les éd. de l'École polytechnique, 2005. 209, 365

Ber76. C. Berge – *Graphs and hypergraphs*, revised ed., North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York., 1976, Translated from the French by Edward Minieka, North-Holland Mathematical Library, Vol. 6. 63

Ber06. J. Bertoin – *Random fragmentation and coagulation processes*, Cambridge Studies in Advanced Mathematics, vol. 102, Cambridge University Press, Cambridge, 2006. 179, 193, 237

Ber09. N. Berestycki – *Recent progress in coalescent theory*, Ensaios Matemáticos [Mathematical Surveys], vol. 16, Sociedade Brasileira de Matemática, Rio de Janeiro, 2009. 179, 193

BGL14. D. Bakry, I. Gentil and M. Ledoux – *Analysis and geometry of Markov diffusion operators*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 348, Springer, Cham, 2014. 336

Bha97. R. Bhatia – *Matrix analysis*, Graduate Texts in Mathematics, vol. 169, Springer-Verlag, New York, 1997. 290

BHJ92. A. D. Barbour, L. Holst and S. Janson – *Poisson approximation*, Oxford Studies in Probability, vol. 2, The Clarendon Press Oxford University Press, New York, 1992, Oxford Science Publications. 15

BL94. N. Bouleau and D. Lépingle – *Numerical methods for stochastic processes*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York, 1994, A Wiley-Interscience Publication. 363

BL98. P. Barbe and M. Ledoux – *Probabilités*, De la licence à l'agrégation, Belin, 1998. 365

BMP08. P. Baldi, L. Mazliak and P. Priouret – *Martingales et chaînes de Markov. Théorie élémentaire et exercices corrigés*, Hermann, 2008. 365

BMR14. S. Bubeck, E. Mossel and M. Z. Rácz – "On the influence of the seed graph in the preferential attachment model", arXiv:1401.4849, 2014. 208

Bod14. T. Bodineau – "Promenade aléatoire: chaînes de Markov et martingales", Lecture notes, École Polytechnique, Palaiseau, France, 2014. 76, 208, 365

Bol80. B. Bollobás – "A probabilistic proof of an asymptotic formula for the number of labelled regular graphs", *European J. Combin.* **1** (1980), no. 4, p. 311–316. 63

Bol89. E. Bolthausen – "A note on the diffusion of directed polymers in a random environment", *Comm. Math. Phys.* **123** (1989), no. 4, p. 529–534. 259

Bor95. V. S. Borkar – *Probability theory*, Universitext, Springer-Verlag, New York, 1995, An advanced course. 365

Bor08. C. Bordenave – "On the birth-and-assassination process, with an application to scotching a rumor in a network", *Electron. J. Probab.* **13** (2008), p. no. 66, 2014–2030. 297, 298

Bor14a. C. Bordenave – "Lecture notes on random graphs and probabilistic combinatorial optimization", Available on the author's website, 2014. 63, 222, 223, 272

Bor14b. — , "A short course on random matrices", Available on the author's website, 2014. 290

BPPC99. C. Bouzitat, G. Pagès, F. Petit and F. Carrance – *En passant par hasard... les probabilités de tous les jours*, Vuibert, 1999. 15

BR11. M. Benaïm and O. Raimond – "Self-interacting diffusions IV: Rate of convergence", *Electron. J. Probab.* **16** (2011), p. no. 66, 1815–1843. 209

Bré99. P. Brémaud – *Markov chains*, Texts in Applied Mathematics, vol. 31, Springer-Verlag, New York, 1999, Gibbs fields, Monte Carlo simulation, and queues. 365

BRST01. B. Bollobás, O. Riordan, J. Spencer and G. Tusnády – "The degree sequence of a scale-free random graph process", *Random Structures Algorithms* **18** (2001), no. 3, p. 279–290. 208

BS10. Z. Bai and J. W. Silverstein – *Spectral analysis of large dimensional random matrices*, second ed., Springer Series in Statistics, Springer, New York, 2010. 290

BSZ11. N. Berestycki, O. Schramm and O. Zeitouni – "Mixing times for random k-cycles and coalescence-fragmentation chains", *Ann. Probab.* **39** (2011), no. 5, p. 1815–1843. 63

BT95. V. Bally and D. Talay – "The Euler scheme for stochastic differential equations: error analysis with Malliavin calculus", *Math. Comput. Simulation* **38** (1995), no. 1-3, p. 35–41, Probabilités numériques (Paris, 1992). 363

Cap12. M. Capitaine – "Introduction aux grandes matrices aléatoires", Available on the author's website, 2012. 290

Cat99. O. Catoni – "Simulated annealing algorithms and Markov chains with rare transitions", Séminaire de Probabilités, XXXIII, Lecture Notes in Math., vol. 1709, Springer, Berlin, 1999, p. 69–119. 77

CD04. A. Charpentier and M. Denuit – *Mathématiques de l'assurance non-vie - concepts fondamentaux de théorie du risque, tome 1*, Economica, 2004. 246

CDKM14. N. Curien, T. Duquesne, I. Kortchemski and I. Manolescu – "Scaling limits and influence of the seed graph in preferential attachment trees", preprint arXiv:1406.1758, 2014. 208

CDM13. D. Chafaï, Y. Doumerc and F. Malrieu – "Processus des restaurants chinois et loi d'ewens", *RMS* **123** (2013), no. 3, p. 56–74. 192

CDPP09. P. Caputo, P. Dai Pra and G. Posta – "Convex entropy decay via the Bochner-Bakry-Emery approach", *Ann. Inst. Henri Poincaré Probab. Stat.* **45** (2009), no. 3, p. 734–753. 149

CF51. K. L. Chung and W. H. J. Fuchs – "On the distribution of values of sums of random variables", *Mem. Amer. Math. Soc.* **1951** (1951), no. 6, p. 12. 34

CH02. P. Carmona and Y. Hu – "On the partition function of a directed polymer in a Gaussian random environment", *Probab. Theory Related Fields* **124** (2002), no. 3, p. 431–457. 259

Cha06. D. Chafaï – "Binomial-Poisson entropic inequalities and the $M/M/\infty$ queue", *ESAIM Probab. Stat.* **10** (2006), p. 317–339 (electronic). 148

Cha13. — , "Introduction aux matrices aléatoires", Aléatoire, Ed. Éc. Polytech., Palaiseau, 2013, p. 87–122. 290

Clo13. B. Cloez – "Comportement asymptotique de processus avec sauts et applications pour des modèles avec branchement", PhD thesis, Université Paris-Est, 2013. 237

CMP10. D. Chafaï, F. Malrieu and K. Paroux – "On the long time behavior of the TCP window size process", *Stochastic Processes and their Applications* **8** (2010), no. 120, p. 1518–1534. 237

CSY03. F. Comets, T. Shiga and N. Yoshida – "Directed polymers in a random environment: path localization and strong disorder", *Bernoulli* **9** (2003), no. 4, p. 705–723. 259

CT97. C. Cocozza-Thivent – *Processus stochastiques et fiabilité des systèmes*, Mathématiques et Applications, Springer, 1997. 246, 365

CV06. F. Comets and V. Vargas – "Majorizing multiplicative cascades for directed polymers in random media", *ALEA Lat. Am. J. Probab. Math. Stat.* **2** (2006), p. 267–277. 259

CW14. K. L. Chung and R. J. Williams – *Introduction to stochastic integration*, second ed., Modern Birkhäuser Classics, Birkhäuser/Springer, New York, 2014. 365

CY06. F. Comets and N. Yoshida – "Directed polymers in random environment are diffusive at weak disorder", *Ann. Probab.* **34** (2006), no. 5, p. 1746–1770. 259

Dav90. B. Davis – "Reinforced random walk", *Probab. Theory Related Fields* **84** (1990), no. 2, p. 203–229. 209

DCD83. D. Dacunha-Castelle and M. Duflo – *Probabilités et statistiques. Tome 2*, Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree], Masson, Paris, 1983, Problèmes à temps mobile. [Movable-time problems]. 98

DCS12. H. Duminil-Copin and S. Smirnov – "The connective constant of the honeycomb lattice equals $\sqrt{2+\sqrt{2}}$", *Ann. of Math. (2)* **175** (2012), no. 3, p. 1653–1665. 223

Dem05. A. Dembo – "Favorite points, cover times and fractals", Lectures on probability theory and statistics, Lecture Notes in Math., vol. 1869, Springer, Berlin, 2005, p. 1–101. 16

Dev12. L. Devroye – "Simulating size-constrained Galton-Watson trees", *SIAM J. Comput.* **41** (2012), no. 1, p. 1–11. 64

DF91. P. Diaconis and W. Fulton – "A growth model, a game, an algebra, Lagrange inversion, and characteristic classes", *Rend. Sem. Mat. Univ. Politec. Torino* **49** (1991), no. 1, p. 95–119 (1993), Commutative algebra and algebraic geometry, II (Italian) (Turin, 1990). 87

DGR02. V. Dumas, F. Guillemin and P. Robert – "A Markovian analysis of additive-increase multiplicative-decrease algorithms", *Adv. in Appl. Probab.* **34** (2002), no. 1, p. 85–111. 236

DJ06. J.-F. Delmas and B. Jourdain – *Modèles aléatoires*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 57, Springer-Verlag, Berlin, 2006, Applications aux sciences de l'ingénieur et du vivant. [Applications to engineering and the life sciences]. 52, 98, 165, 179, 365

DLM99. B. Delyon, M. Lavielle and E. Moulines – "Convergence of a stochastic approximation version of the EM algorithm", *Ann. Statist.* **27** (1999), no. 1, p. 94–128. 106

DLR77. A. P. Dempster, N. M. Laird and D. B. Rubin – "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Statist. Soc. Ser. B* **39** (1977), no. 1, p. 1–38, With discussion. 106

DM47. A. Dvoretzky and T. Motzkin – "A problem of arrangements", *Duke Math. J.* **14** (1947), p. 305–313. 52

DM80. C. Dellacherie and P.-A. Meyer – *Probabilités et potentiel. Chapitres V à VIII*, revised ed., Actualités Scientifiques et Industrielles [Current Scientific and Industrial Topics], vol. 1385, Hermann, Paris, 1980, Théorie des martingales. [Martingale theory]. 365

DMWZZ04. P. Diaconis, E. Mayer-Wolf, O. Zeitouni and M. P. W. Zerner – "The Poisson-Dirichlet law is the unique invariant distribution for uniform split-merge transformations", *Ann. Probab.* **32** (2004), no. 1B, p. 915–938. 62

DP13. A. V. Doumas and V. G. Papanicolaou – "Asymptotics of the rising moments for the coupon collector's problem", *Electron. J. Probab.* **18** (2013), p. no. 41, 15. 15

Drm09. M. Drmota – *Random trees*, SpringerWienNewYork, Vienna, 2009, An interplay between combinatorics and probability. 63

DS81. P. Diaconis and M. Shahshahani – "Generating a random permutation with random transpositions", *Z. Wahrsch. Verw. Gebiete* **57** (1981), no. 2, p. 159–179. 62

DS84. P. G. Doyle and J. L. Snell – *Random walks and electric networks*, Carus Mathematical Monographs, vol. 22, Mathematical Association of America, Washington, DC, 1984. 34

DS87. P. Diaconis and M. Shahshahani – "Time to reach stationarity in the Bernoulli-Laplace diffusion model", *SIAM J. Math. Anal.* **18** (1987), no. 1, p. 208–218. 121

DST14. M. Disertori, C. Sabot and P. Tarrès – "Transience of Edge-Reinforced Random Walk", arXiv:1403.6079, 2014. 209

Duf97. M. Duflo – *Random iterative models*, Applications of Mathematics (New York), vol. 34, Springer-Verlag, Berlin, 1997, Translated from the 1990 French original by Stephen S. Wilson and revised by the author. 106

Dur64. R. Durstenfeld – "Algorithm 235: Random permutation", *Communications of the ACM* **7** (1964), no. 7, p. 420. 62

Dur08. R. Durrett – *Probability models for DNA sequence evolution*, second ed., Probability and its Applications (New York), Springer, New York, 2008. 165, 179, 192

Dur10a. R. Durrett – *Probability: theory and examples*, fourth ed., Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2010. 365

Dur10b. — , *Random graph dynamics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2010. 208

Dwa69. M. Dwass – "The total progeny in a branching process and a related random walk.", *J. Appl. Probability* **6** (1969), p. 682–686. 52

DZ10. A. Dembo and O. Zeitouni – *Large deviations techniques and applications*, Stochastic Modelling and Applied Probability, vol. 38, Springer-Verlag, Berlin, 2010, Corrected reprint of the second (1998) edition. 297, 365

EG60. P. Erdős and T. Gallai – "Gráfok előírt fokszámú pontokkal", *Matematikai Lapok* (1960), no. 11, p. 264–274. 63

EK86. S. N. Ethier and T. G. Kurtz – *Markov processes*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986, Characterization and convergence. 363

EKM97. P. Embrechts, C. Klüppelberg and T. Mikosch – *Modelling extremal events*, Applications of Mathematics (New York), vol. 33, Springer-Verlag, Berlin, 1997, For insurance and finance. 365

EO05. R. Erban and H. G. Othmer – "From individual to collective behavior in bacterial chemotaxis", *SIAM J. Appl. Math.* **65** (2004/05), no. 2, p. 361–391 (electronic). 307

ER59. P. Erdős and A. Rényi – "On random graphs. I", *Publ. Math. Debrecen* **6** (1959), p. 290–297. 222

ER61a. — , "On a classical problem of probability theory", *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), p. 215–220. 15

ER61b. — , "On the evolution of random graphs", *Bull. Inst. Internat. Statist.* **38** (1961), p. 343–347. 222

Ewe04. W. J. Ewens – *Mathematical population genetics. I*, second ed., Interdisciplinary Applied Mathematics, vol. 27, Springer-Verlag, New York, 2004, Theoretical introduction. 165, 179, 192

Fel68. W. Feller – *An introduction to probability theory and its applications. Vol. I*, Third edition, John Wiley & Sons Inc., New York, 1968. 14, 15, 34, 131, 365

Fel71. — , *An introduction to probability theory and its applications. Vol. II.*, Second edition, John Wiley & Sons Inc., New York, 1971. 14, 15, 365

Fer89. T. S. Ferguson – "Who solved the secretary problem?", *Statist. Sci.* **4** (1989), no. 3, p. 282–296, With comments and a rejoinder by the author. 132

FF04. D. Foata and A. Fuchs – *Processus stochastiques : processus de poisson, chaînes de markov et martingales*, Collection Sciences Sup, Dunod, 2004. 246

Fil98. J. A. Fill – "An interruptible algorithm for perfect sampling via Markov chains", *Ann. Appl. Probab.* **8** (1998), no. 1, p. 131–162. 77

FKG71. C. M. Fortuin, P. W. Kasteleyn and J. Ginibre – "Correlation inequalities on some partially ordered sets", *Comm. Math. Phys.* **22** (1971), p. 89–103. 259

Fre65. D. A. Freedman – "Bernard Friedman's urn", *Ann. Math. Statist* **36** (1965), p. 956–970. 87

FY48. R. A. Fisher and F. Yates – *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, London, 1948, 3d ed. 62

Gil59. E. N. Gilbert – "Random graphs", *Ann. Math. Statist.* **30** (1959), p. 1141–1144. 222

GJ87. J. Glimm and A. Jaffe – *Quantum physics*, second ed., Springer-Verlag, New York, 1987, A functional integral point of view. 35

GR10. C. Graham and P. Robert – "A multi-class mean-field model with graph structure for TCP flows", Progress in industrial mathematics at ECMI 2008, Math. Ind., vol. 15, Springer, Heidelberg, 2010, p. 125–131. 237

Gri99. G. Grimmett – *Percolation*, second ed., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 321, Springer-Verlag, Berlin, 1999. 222

Gri10. — , *Probability on graphs*, Institute of Mathematical Statistics Textbooks, vol. 1, Cambridge University Press, Cambridge, 2010, Random processes on graphs and lattices. 222

GRZ04. F. Guillemin, P. Robert and B. Zwart – "AIMD algorithms and exponential functionals", *Ann. Appl. Probab.* **14** (2004), no. 1, p. 90–117. 237

GS72. Ĭ. Ī. Gīhman and A. V. Skorohod – *Stochastic differential equations*, Springer-Verlag, New York, 1972, Translated from the Russian by Kenneth Wickwire, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 72. 363

GS84. C. R. Givens and R. M. Shortt – "A class of Wasserstein metrics for probability distributions", *Michigan Math. J.* **31** (1984), no. 2, p. 231–240. 335

GS01. G. R. Grimmett and D. R. Stirzaker – *Probability and random processes*, third ed., Oxford University Press, New York, 2001. 148

GT13. C. Graham and D. Talay – *Stochastic simulation and Monte Carlo methods*, Stochastic Modelling and Applied Probability, vol. 68, Springer, Heidelberg, 2013, Mathematical foundations of stochastic simulation. 363, 366

Gut13. A. Gut – *Probability: a graduate course*, second ed., Springer Texts in Statistics, Springer, New York, 2013. 365

Häg02. O. Häggström – *Finite markov chains and algorithmic applications*, London Mathematical Society student texts, Cambridge University Press, 2002. 77, 365

Har60. T. E. Harris – "A lower bound for the critical probability in a certain percolation process", *Proc. Cambridge Philos. Soc.* **56** (1960), p. 13–20. 223

Har02. — , *The theory of branching processes*, Dover Phoenix Editions, Dover Publications Inc., Mineola, NY, 2002, Corrected reprint of the 1963 original [Springer, Berlin; MR0163361 (29 #664)]. 52, 237

Has70. W. K. Hastings – "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika* (1970), no. 57, p. 97–109. 76

HH80. P. Hall and C. C. Heyde – *Martingale limit theory and its application*, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980, Probability and Mathematical Statistics. 87, 365

HJ06. O. Häggström and J. Jonasson – "Uniqueness and non-uniqueness in percolation theory", *Probab. Surv.* **3** (2006), p. 289–344. 223

HJ13. R. A. Horn and C. R. Johnson – *Matrix analysis*, second ed., Cambridge University Press, Cambridge, 2013. 290

HJV07. P. Haccou, P. Jagers and V. A. Vatutin – *Branching processes: variation, growth, and extinction of populations*, Cambridge Studies in Adaptive Dynamics, Cambridge University Press, Cambridge, 2007. 52

Hol01. L. Holst – "Extreme value distributions for random coupon collector and birthday problems", *Extremes* **4** (2001), no. 2, p. 129–145 (2002). 15

Hop84. F. M. Hoppe – "Pólya-like urns and the Ewens' sampling formula", *J. Math. Biol.* **20** (1984), no. 1, p. 91–94. 193

Hör67. L. Hörmander – "Hypoelliptic second order differential equations", *Acta Math.* **119** (1967), p. 147–171. 348

HV10. S. Herrmann and P. Vallois – "From persistent random walk to the telegraph noise", *Stoch. Dyn.* **10** (2010), no. 2, p. 161–196. 307

IS88. J. Z. Imbrie and T. Spencer – "Diffusion of directed polymers in a random environment", *J. Statist. Phys.* **52** (1988), no. 3-4, p. 609–626. 259

JK77. N. L. Johnson and S. Kotz – *Urn models and their application*, John Wiley & Sons, New York-London-Sydney, 1977, An approach to modern discrete probability theory, Wiley Series in Probability and Mathematical Statistics. 208

JKB97. N. L. Johnson, S. Kotz and N. Balakrishnan – *Discrete multivariate distributions*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1997, A Wiley-Interscience Publication. 193

JLS12. D. Jerison, L. Levine and S. Sheffield – "Logarithmic fluctuations for internal DLA", *J. Amer. Math. Soc.* **25** (2012), no. 1, p. 271–301. 87

JLS13. D. Jerison, L. Levine and S. Sheffield – "Internal dla in higher dimensions", *Electron. J. Probab.* **18** (2013), p. no. 98, 1–14. 87

JP02. J. Jacod and P. Protter – *L'essentiel en théorie des probabilités*, Enseignement des mathématiques, Cassini, 2002. 365

JP03. — , *Probability essentials*, second ed., Universitext, Springer-Verlag, Berlin, 2003. 365

Kac47. M. Kac – "Random walk and the theory of Brownian motion", *Amer. Math. Monthly* **54** (1947), p. 369–391. 121

Kac74. — , "A stochastic model related to the telegrapher's equation", *Rocky Mountain J. Math.* **4** (1974), p. 497–509, Reprinting of an article published in 1956, Papers arising from a Conference on Stochastic Differential Equations (Univ. Alberta, Edmonton, Alta., 1972). 307

KB97. S. Kotz and N. Balakrishnan – "Advances in urn models during the past two decades", Advances in combinatorial methods and applications to probability and statistics, Stat. Ind. Technol., Birkhäuser Boston, Boston, MA, 1997, p. 203–257. 208

KDS+15. N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski and N. Chopin – "On particle methods for parameter estimation in state-space models", *Stat. Sci.* **30** (2015), no. 3, p. 328–351 (English). 98

Kes80. H. Kesten – "The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$", *Comm. Math. Phys.* **74** (1980), no. 1, p. 41–59. 223

Kin80. J. F. C. Kingman – *Mathematics of genetic diversity*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 34, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1980. 192

Kin93. — , *Poisson processes*, Oxford Studies in Probability, vol. 3, The Clarendon Press Oxford University Press, New York, 1993, Oxford Science Publications. 179, 193

KL04. E. Kuhn and M. Lavielle – "Coupling a stochastic approximation version of EM with an MCMC procedure", *ESAIM Probab. Stat.* **8** (2004), p. 115–131 (electronic). 106

Knu05. D. E. Knuth – *The art of computer programming. Vol. 1–4*, Addison-Wesley, 2005. 62, 63

Kor15. I. Kortchemski – "Predator-Prey Dynamics on Infinite Trees: A Branching Random Walk Approach", Preprint arXiv:1312.4933 forthcoming in Journal of Theoretical Probability, DOI 10.1007/s10959-015-0603-2, 2015. 297

KP95. P. E. Kloeden and E. Platen – "Numerical methods for stochastic differential equations", Nonlinear dynamics and stochastic mechanics, CRC Math. Model. Ser., CRC, Boca Raton, FL, 1995, p. 437–461. 363

KPSC08. A. M. Krieger, M. Pollak and E. Samuel-Cahn – "Beat the mean: sequential selection by better than average rules", *J. Appl. Probab.* **45** (2008), no. 1, p. 244–259. 131

KS66. H. Kesten and B. P. Stigum – "A limit theorem for multidimensional Galton-Watson processes", *Ann. Math. Statist.* **37** (1966), p. 1211–1223. 52

KS76. J. G. Kemeny and J. L. Snell – *Finite Markov chains*, Springer-Verlag, New York, 1976, Reprinting of the 1960 original, Undergraduate Texts in Mathematics. 121

KS91. I. Karatzas and S. E. Shreve – *Brownian motion and stochastic calculus*, second ed., Graduate Texts in Mathematics, vol. 113, Springer-Verlag, New York, 1991. 366

Lac10. H. Lacoin – "New bounds for the free energy of directed polymers in dimension $1 + 1$ and $1 + 2$", *Comm. Math. Phys.* **294** (2010), no. 2, p. 471–503. 259

Law13. G. F. Lawler – *Intersections of random walks*, Modern Birkhäuser Classics, Birkhäuser/Springer, New York, 2013, Reprint of the 1996 edition. 35

LBG92. G. F. Lawler, M. Bramson and D. Griffeath – "Internal diffusion limited aggregation", *Ann. Probab.* **20** (1992), no. 4, p. 2117–2140. 87

LG13. J.-F. Le Gall – *Mouvement brownien, martingales et calcul stochastique*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 71, Springer, Heidelberg, 2013. 365

Lig85. T. M. Liggett – *Interacting particle systems*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 276, Springer-Verlag, New York, 1985. 259

LL10. G. F. Lawler and V. Limic – *Random walk: a modern introduction*, Cambridge Studies in Advanced Mathematics, vol. 123, Cambridge University Press, Cambridge, 2010. 35

LP15. R. Lyons and Y. Peres – *Probability on trees and networks*, Cambridge University Press, 2015. 64

LPW09. D. A. Levin, Y. Peres and E. L. Wilmer – *Markov chains and mixing times*, American Mathematical Society, Providence, RI, 2009, With a chapter by James G. Propp and David B. Wilson. 16, 35, 77, 121, 365

LW09. Q. Liu and F. Watbled – "Exponential inequalities for martingales and asymptotic properties of the free energy of directed polymers in

a random environment", *Stochastic Process. Appl.* **119** (2009), no. 10, p. 3101–3132. 259

Lé08. T. Lévy – "Probabilistic methods in statistical physics", Lecture notes - Tsinghua University, October 2008. 222

Mah09. H. M. Mahmoud – *Pólya urn models*, Texts in Statistical Science Series, CRC Press, Boca Raton, FL, 2009. 208

Mar04. J.-F. Marckert – "Structures arborescentes, algorithmes", Habilitation thesis, 2004. 52

Mau03. A. Maurer – "A bound on the deviation probability for sums of non-negative random variables", *JIPAM. J. Inequal. Pure Appl. Math.* **4** (2003), no. 1, p. Article 15, 6 pp. (electronic). 132

McD89. C. McDiarmid – "On the method of bounded differences", Surveys in combinatorics, 1989 (Norwich, 1989), London Math. Soc. Lecture Note Ser., vol. 141, Cambridge Univ. Press, Cambridge, 1989, p. 148–188. 271

McD98. — , "Concentration", Probabilistic methods for algorithmic discrete mathematics, Algorithms Combin., vol. 16, Springer, Berlin, 1998, p. 195–248. 271

Mik09. T. Mikosch – *Non-life insurance mathematics*, second ed., Universitext, Springer-Verlag, Berlin, 2009, An introduction with the Poisson process. 246

MM09. M. Mézard and A. Montanari – *Information, physics, and computation*, Oxford Graduate Texts, Oxford University Press, Oxford, 2009. 271

MP10. P. Mörters and Y. Peres – *Brownian motion*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2010, With an appendix by Oded Schramm and Wendelin Werner. 318

MR95. R. Motwani and P. Raghavan – *Randomized algorithms*, Cambridge University Press, Cambridge, 1995. 15, 62

MRR+53. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller – "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics* (1953), no. 21, p. 1087–1092. 76

MV12. S. Méléard and D. Villemonais – "Quasi-stationary distributions and population processes", *Probab. Surv.* **9** (2012), p. 340–410. 52

MW90. B. D. McKay and N. C. Wormald – "Uniform generation of random regular graphs of moderate degree", *J. Algorithms* **11** (1990), no. 1, p. 52–67. 63

Mé13. S. Méléard – "Modèles aléatoires en ecologie et evolution", Lecture notes, École Polytechnique, 2013. 165, 179, 193

Nel67. E. Nelson – *Dynamical theories of Brownian motion*, Princeton University Press, Princeton, N.J., 1967. 348

Nev72. J. Neveu – *Martingales à temps discret*, Masson et Cie, éditeurs, Paris, 1972. 365

Nev75. J. Neveu – *Discrete-parameter martingales*, revised ed., North-Holland Publishing Co., Amsterdam-Oxford; American Elsevier Publishing Co., Inc., New York, 1975, Translated from the French by T. P. Speed, North-Holland Mathematical Library, Vol. 10. 365

Nor98a. J. R. Norris – *Markov chains*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 2, Cambridge University Press, Cambridge, 1998, Reprint of 1997 original. 34, 121, 148, 246

Nor98b. J. R. Norris – *Markov chains*, Cambridge University Press, 1998. 365

NS60. D. J. Newman and L. Shepp – "The double dixie cup problem", *Amer. Math. Monthly* **67** (1960), p. 58–61. 15

ODA88. H. G. Othmer, S. R. Dunbar and W. Alt – "Models of dispersal in biological systems", *J. Math. Biol.* **26** (1988), no. 3, p. 263–298. 307

OK08. T. J. Ott and J. H. B. Kemperman – "Transient behavior of processes in the TCP paradigm", *Probab. Engrg. Inform. Sci.* **22** (2008), no. 3, p. 431–471. 236

Øks03. B. Øksendal – *Stochastic differential equations*, sixth ed., Universitext, Springer-Verlag, Berlin, 2003, An introduction with applications. 365

Par07. E. Pardoux – *Processus de Markov et applications*, Collection Sciences Sup, Dunod, 2007, Mathématiques appliquées/SMAI. 77, 98, 365

Pem07. R. Pemantle – "A survey of random processes with reinforcement", *Probab. Surv.* **4** (2007), p. 1–79. 208, 209

Per07. B. Perthame – *Transport equations in biology*, Frontiers in Mathematics, Birkhäuser Verlag, Basel, 2007. 237

Pit06. J. Pitman – *Combinatorial stochastic processes*, Lecture Notes in Mathematics, vol. 1875, Springer-Verlag, Berlin, 2006, Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. 179, 193

Pre00. J. Preater – "Sequential selection with a better-than-average rule", *Statist. Probab. Lett.* **50** (2000), no. 2, p. 187–191. 131

PW96. J. D. Propp and D. B. Wilson – "Exact sampling with coupled Markov chains and applications to statistical mechanics", *Random Structures Algorithms* **9** (1996), no. 1-2, p. 223–252. 77

PW98. — , "Coupling from the past: a user's guide", Microsurveys in discrete probability (Princeton, NJ, 1997), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 41, Amer. Math. Soc., Providence, RI, 1998, p. 181–192. 77

RC04. C. P. Robert and G. Casella – *Monte Carlo statistical methods*, second ed., Springer Texts in Statistics, Springer-Verlag, New York, 2004. 76

Rev97. D. Revuz – *Probabilités*, Hermann, 1997. 365

Rob00. P. Robert – *Réseaux et files d'attente: méthodes probabilistes*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 35, Springer-Verlag, Berlin, 2000. 148, 363

Rob05. — , "Réseaux de communication, algorithmes et probabilités", Lecture notes, École Polytechnique, Palaiseau, France, 2005. 237

Rob10. — , "Mathématiques et réseau de communication", Dossiers pour la science (2010) hal:inria-00472217, 2010. 237

Roy99. G. Royer – *Une initiation aux inégalités de Sobolev logarithmiques*, Cours Spécialisés [Specialized Courses], vol. 5, Société Mathématique de France, Paris, 1999. 336

RRS05. S. Robin, F. Rodolphe and S. Schbath – *DNA, words and models*, Cambridge University Press, Cambridge, 2005, Translated from the 2003 French original. 98

Rud87. W. Rudin – *Real and complex analysis*, third ed., McGraw-Hill Book Co., New York, 1987. 318

Rug01. C. Ruget (ed.) – *Mathématiques en situation*, SCOPOS, vol. 11, Springer-Verlag, Berlin, 2001, Issues de l'épreuve de modélisation de l'agrégation. [From the Examination in Modelling for the Agrégation]. 52

RW00. L. C. G. Rogers and D. Williams – *Diffusions, Markov processes, and martingales. Vol. 1*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 2000, Foundations, Reprint of the second (1994) edition. 366

RY99. D. Revuz and M. Yor – *Continuous martingales and Brownian motion*, third ed., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293, Springer-Verlag, Berlin, 1999. 365

Ré85. J.-L. Rémy – "Un procédé itératif de dénombrement d'arbres binaires et son application à leur génération aléatoire", *R.A.I.R.O. Informatique Théorique* **19** (1985), no. 2, p. 179–195. 63

She07. S. Sheffield – "Gaussian free fields for mathematicians", *Probab. Theory Related Fields* **139** (2007), no. 3-4, p. 521–541. 35

Shi96. A. N. Shiryaev – *Probability*, second ed., Graduate Texts in Mathematics, vol. 95, Springer-Verlag, New York, 1996, Translated from the first (1980) Russian edition by R. P. Boas. 365

SM02. J. Siltaneva and E. Mäkinen – "A comparison of random binary tree generators", *The Computer Journal* **45** (2002), no. 6, p. 653–660. 63

Spi70. F. Spitzer – *Principes des cheminements aléatoires*, Traduit de l'anglais par E. Baverez et J.-L. Guignard. Centre Interarmés de Recherche Opérationnelle, vol. 2, Dunod, Paris, 1970. 34

Sta83. A. J. Stam – "Generation of a random partition of a finite set by an urn model", *J. Combin. Theory Ser. A* **35** (1983), no. 2, p. 231–240. 63

Ste94. J. M. Steele – "Le Cam's inequality and Poisson approximations", *Amer. Math. Monthly* **101** (1994), no. 1, p. 48–54. 15

Ste97. — , *Probability theory and combinatorial optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 69, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. 15, 271

Tal96. D. Talay – "Probabilistic numerical methods for partial differential equations: elements of analysis", Probabilistic models for nonlinear partial differential equations (Montecatini Terme, 1995), Lecture Notes in Math., vol. 1627, Springer, Berlin, 1996, p. 148–196. 363

Tal02. — , "Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme", *Markov Process. Related Fields* **8** (2002), no. 2, p. 163–198, Inhomogeneous random systems (Cergy-Pontoise, 2001). 348

Tar11. P. Tarrès – "Localization of reinforced random walks", arXiv:1103.5536, 2011. 209

Tav04. S. Tavaré – "Ancestral inference in population genetics", Lectures on probability theory and statistics, Lecture Notes in Math., vol. 1837, Springer, Berlin, 2004, p. 1–188. 165, 179, 193

Tou99. P. Toulouse – *Thèmes de probabilités et statistique*, Dunod, 1999. 365

TPH86. J. N. Tsitsiklis, C. H. Papadimitriou and P. Humblet – "The performance of a precedence-based queueing discipline", *J. Assoc. Comput. Mach.* **33** (1986), no. 3, p. 593–602. 297

TT90. D. Talay and L. Tubaro – "Expansion of the global error for numerical schemes solving stochastic differential equations", *Stochastic Anal. Appl.* **8** (1990), no. 4, p. 483–509 (1991). 363

TVW10. A. Tripathi, S. Venugopalan and D. B. West – "A short constructive proof of the Erdős-Gallai characterization of graphic lists", *Discrete Math.* **310** (2010), no. 4, p. 843–844. 63

vdH14. R. van der Hofstad – "Random Graphs and Complex Networks", Lecture notes available on Internet, 2014. 63, 208, 223

Vil03. C. Villani – *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003. 237

Vil06. — , "Hypocoercive diffusion operators", International Congress of Mathematicians. Vol. III, Eur. Math. Soc., Zürich, 2006, p. 473–498. 348

Vil08. — , "H-theorem and beyond: Boltzmann's entropy in today's mathematics", Boltzmann's legacy, ESI Lect. Math. Phys., Eur. Math. Soc., Zürich, 2008, p. 129–143. 348

Vil09. — , "Hypocoercivity", *Mem. Amer. Math. Soc.* **202** (2009), no. 950, p. iv+141. 348

Wil91. D. Williams – *Probability with martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, 1991. 98, 365

Wil96. D. B. Wilson – "Generating random spanning trees more quickly than the cover time", In Proceedings of the 28'th annual ACM symposium on Theory of computing pages 296–303, 1996. 63

WL01. F. Wang and D. P. Landau – "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States", *Physical Review Letters* **86** (2001), p. 2050. 77

Yca02. B. Ycart – *Modèles et algorithmes markoviens*, Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 39, Springer-Verlag, Berlin, 2002. 318

Yuk98. J. E. Yukich – *Probability theory of classical Euclidean optimization problems*, Lecture Notes in Mathematics, vol. 1675, Springer-Verlag, Berlin, 1998. 271

Index

actuarial science, 246
algorithm
  Arnold–Sleep, 61
  BFS, 221
  breadth-first search, 221
  configuration, 60
  EM, 103
  Fisher–Yates–Knuth, 55, 184
  forward-backward, 91
  Kiefer–Wolfowitz, 106
  Metropolis–Hastings, 68, 106
  Propp–Wilson, 72
  quicksort, 54
  randomized quicksort, 62
  rejection, 55
  Robbins–Monro, 106
  Rémy's, 63
  segmentation, 98
  simulated annealing, 70, 262, 263
  Stam's, 58
aperiodic, *see* Markov chain
Arnold–Sleep, *see* algorithm
Avogadro, *see* Avogadro's number
Avogadro's number, 115
Azuma–Hoeffding, *see* inequality
Azuma-Hoeffding, *see* inequality

ballot, *see* theorem
Barabási–Albert, *see* random graph
Barbour–Eagleson, *see* inequality
Bearwood-Halton-Hammersley, *see* theorem
Bell, *see* numbers, *see* numbers
Berry–Esseen, *see* central limit theorem, *see* central limit theorem, *see* theorem
Birkhoff, *see* theorem
birth and death, *see* process
bistochastic, *see* doubly stochastic
bits, 3
Bollobás, *see* configuration algorithm, *see* configuration model
Boltzmann, *see* entropy, *see* kinetic theory of gases, *see* entropy
Borel–Cantelli, *see* lemma
branching, *see* property
Brownian bridge, 360
Brownian motion, 311, 319, 347, 353, 355
Burkholder-Davis-Gundy, *see* inequality

Cannings, *see* model
card shuffling, 31
Carleman, *see* condition
Catalan, *see* numbers, *see* numbers, *see* numbers
Cauchy, *see* law
Cayley, *see* graph
Cesàro, *see* lemma
Chapman–Kolmogorov, *see* equation, *see* equation
Chebyshev, *see* inequality, *see* inequality
Chen–Stein, *see* method
Chernoff, *see* inequality, *see* theorem
Chinese restaurant, *see* process
Clopper–Pearson, *see* confidence interval

coalescence of trajectories, 73
coin tossing, 1, 38, 157, 187, 191, 199, 212
comparaison série-intégrale, 187
competing exponential clocks, 137
condition
  Carleman, 290
  Doeblin's, 68
  Lindeberg, 125
  Lyapunov, 83, 84
configuration, *see* model
constant
  connective, 217
  Euler's, 10
coupling, *see* method, *see* method, *see* method
coupon collector, 1, 9, 25, 33, 109
Courant-Fischer, *see* variational formula
covariance, 273, 320
  sample, 273, 289
Cramér–Lundberg, *see* model
Cramér-Von Mises, *see* theorem
cutoff, 34, 115

de Moivre–Laplace, *see* central limit theorem
de-Poissonization, 268
decomposition
  Dunford, 347
derangement, 55
dice rolling, 9
Dirichlet, *see* problem, *see* problem
discrete Laplacian, 28
disorder, 250
distance
  coupling, 237
  Fréchet, 237
  Kolmogorov-Smirnov, 278
  Mallows, 237
  total variation, 4, 146
  Wasserstein, 15, 145, 230, 279, 329, 331, 344, 346
distribution
  Bernoulli, 1
  beta, 161, 193, 197
  beta-binomial, 3, 197
  binomial, 1
  Ewens, 66, 176, 183
  geometric, 2
  Gumbel, 12, 114, 127, 129, 173
  invariant, 32, 33, 68, 69, 75, 77, 93, 94, 108, 109, 111, 113, 116, 119, 121, 145, 146, 148, 149, 160–162, 228, 229, 342–345, 348
  multinomial, 25, 40, 167
  negative binomial, 2
  offspring, 37
  Pareto, 130, 208
  Pascal, 2
  Poisson, 8, 54
  Rademacher, 17
  reversible, 68, 108, 145, 344
  semicircle, 22
  uniform
    finite graphs, 58
    matchings, 57
    partitions of a finite set, 58
    rooted plane binary trees, 60
    spanning tree, 63
    symmetric group, 55, 184
  Zipf, 208
DNA, 89, 174
Dobinski, *see* formula
Doeblin, *see* condition
Donsker, *see* theorem
Doob, *see* optional stopping theorem
doubly stochastic, 33, 108
Dunford, *see* decomposition
Désiré André, *see* formula

Ehrenfest, *see* model, *see* urn
EM, *see* algorithm
energy, 65
  free, 249
entropy, 107, 258
  Boltzmann, 66, 155
  relative, 104, 334, 344
  Shannon, 155
equality in distribution, 48
equation
  Chapman–Kolmogorov, 142, 228, 235
  Fokker-Planck, 325
  heat, 304, 318
  heat with drift, 322
  kinetic Fokker-Planck, 344
  master, 201
  partial differential, 300, 322

stochastic differential, 159, 161, 319, 320, 335, 336, 347, 354, 356, 358, 361
telegraph, 300, 303, 307
wave, 304
Erdős–Gallai, *see* theorem
Erdős–Rényi, *see* random graph, *see* random graph
error
strong, 362
weak, 362
Euler, *see* constant, *see* scheme
Euler scheme, 361
Ewens, *see* sampling formula, *see* distribution
exact sampling, 77
exchangeability, 165, 197
excursion, 305
exponential family, 67

Fekete, *see* lemma, *see* lemma
Fisher–Yates, *see* algorithm, *see* algorithm
fixation, 153, 156
FKG, *see* inequality
Fokker-Planck, *see* equation, *see* equation
formula
Bayes', 96
Courant-Fischer variational, 278
discrete Itô, 22
Dobinski's, 57
Désiré André's, 23
Ewens sampling, 176
Hadamard, 283
Isserlis', 56
Itô, 312, 320, 323, 355, 362
Kantorovich–Rubinstein, 147
mean value, 309
Mehler, 321
Plancherel, 326
Fourier, *see* transform
Friedman, *see* urn
Frobenius, *see* norm
Fréchet, *see* distance
function
acceptance-rejection, 69, 71
beta, 197
characteristic, 13, 171, 241, 316
contracting, 73
discrete Green, 30
gamma, 197
generating, 21, 39, 143, 215, 216, 235, 316
harmonic, 22, 28, 309
partition, 65, 248
reinforcement, 202
scale, 354

Galton–Watson, *see* process, *see* tree
gambler's ruin, *see* problem
Gaussian free field, 30, 66
Gaussian vector, 95, 320
generating series, 325
giant connected component, 221
Gibbs, *see* measure, *see* Gibbs sampler
Gibbs sampler, 70
graph
d-regular, 59
Barabási–Albert random, 199
Bethe, 213
Cayley, 184
complete, 220
dual, 218
Erdős–Rényi random, 58, 220
finite, 58
preferential attachment, 199
random, 66, 212
regular, 69
Green, *see* function
Grönwall, *see* lemma
Gumbel, *see* distribution

Hadamard, *see* formula
Hamiltonian, 65, *see* Hamiltonian cycle
Hamiltonian cycle, 261
Hardy–Weinberg, *see* law
harmonic, *see* function, *see* measure
Hermite, *see* polynomial
Hilbert-Schmidt, *see* norm
Hoffman-Wielandt, *see* inequality
Hoppe, *see* urn
hypoellipticity, 341
Hörmander, 341

immigration, *see* Galton–Watson process
inequality

Azuma–Hoeffding, 85, 208
Azuma-Hoeffding, 131, 264, 266, 280
Barbour–Eagleson, 15
Burkholder-Davis-Gundy, 362
Chebyshev, 264, 280
Chebyshev's, 117
Chernoff, 264
Chernoff's, 118
FKG, 251, 253
functional, 348
Hoffman-Wielandt, 279
Jensen, 251, 256, 329, 334
Jensen's, 67, 104
Le Cam, 189
Le Cam's, 8, 15
logarithmic Sobolev, 331–335, 348
Markov, 140, 141, 187, 221, 265, 282
Markov's, 12
Maurer, 132
Poincaré, 327–329, 348
Schwarz, 282, 329, 332
Schwarz's, 112
infinitesimal generator, 144, 147, 226, 234, 246, 322
internal diffusion limited aggregation, 80, 87
interval
Clopper–Pearson confidence, 3
prediction, 11, 12
Wald confidence, 2
inverse temperature, 65, 248
irreversibility, 107, 344
Ising, *see* model
Isserlis, *see* formula
Itô, *see* formula

Jensen, *see* inequality, *see* inequality, *see* inequality
Jordan, *see* curve
Jordan curve, 218

Kalman, *see* Kalman filter
Kalman filter, 94
Kantorovich–Rubinstein, *see* formula
kernel
discrete Poisson, 28
Poisson, 312, 313, 317
proposal or exploration, 69
Kingman, *see* coalescent
Kingman's coalescent, 177
Knuth, *see* algorithm, *see* algorithm
Kolmogorov, *see* zero-one law, *see* diffusion process
Kolmogorov-Smirnov, *see* distance, *see* theorem
Kullback-Leibler, *see* divergence
Kullback-Leibler divergence, 334

Langevin, *see* process
Laplace, *see* transform, *see* principle, *see* transform
law
Beta, 358
Cauchy, 316
Hardy–Weinberg, 151, 152
invariant, 300, 307, 323, 329, 334–336, 350, 357–359
large numbers, 2, 21, 273, 275–277, 360
Markov chains, 94
martingales, 97, 98
Mendel, 151
of large numbers, 241, 242
Markov chains, 160
martingales, 125
of rare events, 148
of small numbers, 148, 220, 225
quarter-circle, 289
Rademacher, 300
reversible, 323
semicircle, 276
small numbers, 9, 351
Strassen's iterated logarithm, 204, 242
zero-one
for exponential distributions, 139, 208
for polymers, 252
Kolmogorov, 219, 252
Le Cam, *see* inequality, *see* inequality
lemma
Borel–Cantelli, 3, 12, 73, 141, 187, 191
Borel-Cantelli, 271, 281
Cesàro's, 84
Fekete, 217, 256
Grönwall, 335, 362
Rényi, 137

Scheffé's, 46
Slutsky, 189
Wald, 241
life and death, *see* birth-and-death process
likelihood, 97, 101
loi
Beta, 3

Malliavin, 341
Mallows, *see* distance
Marchenko-Pastur, *see* theorem
Markov, *see* inequality, *see* property, *see* property, *see* inequality, *see* semigroup
Markov chain
aperiodic, 32, 33, 68, 90, 109, 160
embedded, 231
hidden, 89
lazy, 109, 349
sampled, 350
martingale, 2, 22, 28, 42, 84, 97, 125, 155, 157, 163, 196, 264, 312, 323, 343, 355, 357
bounded, 43, 155, 313, 318
bounded in L^2, 46, 128, 174, 188, 236, 255
nonnegative, 42, 43, 131, 155, 221, 243, 252
matching, 56, 60
minimum Euclidean, 261
matrix
adjacency, 58
covariance, 273
random, 273
replacement, 199
sample covariance, 273, 289
Vandermonde, 164
Maurer, *see* inequality
maximum likelihood, 97, 101
maximum likelihood estimator, 97, 101
measure
empirical spectral, 274
Gibbs, 65, 248, 329
harmonic, 28
speed, 354
memorylessness, 137, 176, 207, 233
Mendel, *see* law
method
Chen–Stein, 15
coupling, 7, 15, 19, 81, 109, 146, 157, 213, 218, 221, 230, 231, 279, 331, 346
coalescent, 74
coupling from the past, 74
inverse transform sampling, 53
moments, 83, 281, 283
Monte Carlo, 106, 263
of moments, 229
rejection sampling, 55
second moment, 12, 187
Metropolis–Hastings, *see* algorithm
mixture
of distributions, 100, 197
of Gaussian distributions, 102
model
Cannings, 162, 163
configuration, 60
Cramér–Lundberg, 246
Ehrenfest, 107
Erdős–Rényi, *see* random graph
Ising, 66, 76
Moran, 152
Wright–Fisher, 19, 40, 151, 155, 168
Wright-Fisher, 124, 356
moment, *see* method of moments
moments, *see* method of moments
determination, 283
exponential, 264
monotonicity, 74
Monte Carlo, *see* method
Moran, *see* model
most recent common ancestor, 168
multigraph, 59

network, 222
norm
Frobenius, 279
Hilbert-Schmidt, 279
Schur, 279
numbers
Bell, 57, 191
Catalan, 22, 24, 35, 62, 284
Stirling, first kind, 192
Stirling, second kind, 9, 170, 191, 192

order statistic, 123, 137
Ornstein–Uhlenbeck, *see* process

Ornstein-Uhlenbeck, *see* semigroup
oscillation, 264

PDMP, *see* piecewise deterministic Markov process, *see* piecewise deterministic Markov process
percolation, 212
phase transition, 250
phenomenon
  threshold, 14, 39, 115, 214, 249, 250
phylogeny, 179
Plancherel, *see* formula
Pochhammer, *see* polynomial
Pochhammer symbol, 192
Poincaré, *see* Poincaré sieve, *see* Poincaré sieve, *see* inequality
Poincaré sieve, 11, 56
Poisson, *see* process, *see* kernel
Poissonization, 189, 268
polymer, 66, 247, 249
polynomial
  Hermite, 325
  Pochhammer, 192
preferential attachment, 199
principle
  Donsker invariance, 338, 353
  inclusion-exclusion, 11, 56, 192
  Laplace, 70
  maximum, 312
  pigeonhole, 267
  rotation, 49, 297
probability
  critical, 212
  extinction, 41
  percolation, 212
problem
  Dirichlet, 310
  discrete Dirichlet, 28
  gambler's ruin, 19, 28, 81
  marriage, 132
  optimal stopping, 132
  secretary, 132
  stochastic control, 132
  traveling salesman, 71, 261
process
  autoregressive, 17, 95, 231, 324
  Bernoulli, 2
  birth and death, 147, 170
  Chinese restaurant, 124, 175, 176, 181
  compound Poisson, 241
  confined Langevin, 339
  continuous-time Markov, 147
  Galton–Watson, 168, 221, 233
  Galton–Watson with immigration, 50
  Galton-Watson, 296, 298
  Gaussian, 320
  Kolmogorov diffusion, 336
  Ornstein–Uhlenbeck, 121, 148
  Ornstein-Uhlenbeck, 319, 338, 354
  piecewise deterministic Markov, 226, 240
  Poisson, 2, 112, 136, 175, 225, 239, 291, 300
  Poisson point, 268
  pure death, *see* Yule process
  TCP window-size, 226
  thinned Poisson, 139
  Wiener, 347
  Yule, 138, 139, 147, 233
Prohorov, *see* theorem, *see* theorem
property
  branching, 38, 46
  Markov, 46, 252, 255, 256, 300, 312
Propp–Wilson, *see* algorithm
Pólya, *see* urn

queue, 135, 363
  fluid, 363

Rademacher, *see* distribution
random environment, 248, 250
random walk, 18, 24, 48, 221, 242, 248, 315
  on the discrete cube, 108
  on the symmetric group, 31
  persistent, 307
  vertex-reinforced, 203
randomized combinatorial optimization, 261
reinforcement, 195
  geometric, 204
  linear, 204
reversibility, 107
reversible, *see* distribution
Rémy, *see* algorithm
Rényi, *see* lemma

Scheffé, *see* lemma

scheme
  Euler, 361
Schur, *see* norm
Schwarz, *see* inequality, *see* inequality, *see* inequality, *see* inequality
second law of thermodynamics, 107
self-avoiding loop, 218
self-avoiding path, 212
semicircle, *see* distribution
semigroup
  Markov, 227, 349
  Ornstein-Uhlenbeck, 320
Shakespeare, 3
Shannon, *see* entropy
simulated annealing, *see* algorithm
singular values, 289
Slutsky, *see* lemma
Sobolev, *see* inequality
spectral gap, 327
spectrum
  gap, 327
  of transition matrix, 164, 165, 229
  sample covariance matrix, 289
  transition matrix, 111, 325
spin, 30, 66
Stam, *see* algorithm
Stirling, *see* numbers
Strassen, *see* law of the iterated logarithm, *see* law of the iterated logarithm
subadditivity, 217, 256, 268, 269
submartingale
  bounded in L^1, 51
sum-integral comparison, 85, 127
supermartingale, 244
symmetric group, 32, 54, 55, 66, 181, 261

theorem
  ballot, 23
  Bearwood-Halton-Hammersley, 262
  Berry–Esseen, 15, 189
  Birkhoff, 307
  centered series, 188
  central limit, 2, 4, 21, 35, 118, 189, 338, 353, 360
    Berry–Esseen, 4, 14
    de Moivre–Laplace, 4
    Lindeberg–Lévy, 189
    Lindeberg-Lévy, 125
    Markov chains, 94
    martingales, 83, 87, 97, 98, 125, 189
  centred series, 124, 127
  Chernoff, 296, 297
  Cramér-Von Mises, 361
  Donsker, 338, 353
  Erdős–Gallai, 59
  ergodic
    Markov chains, 94, 160
    Markov chains (fluctuations), 94
    Markov processes, 307
  Kolmogorov-Smirnov, 361
  Marchenko-Pastur, 288
  optional stopping, 22, 28, 155, 221
  Prohorov, 277, 282
  simplified Wigner, 277
  Weierstrass, 283
  Wigner, 275
  Yaglom's, 43
theory
  Boltzmann's kinetic theory of gases, 107
  free probability, 290
  Mendelian, 151
  optimal transport, 237
tightness, 277
time
  bounded stopping, 244
  coalescence, 74, 110, 113, 114
  contraction, 73
  coupling, 74
  cover, 16
  exit, 19, 312
  explosion, 233
  fixation, 153, 156
  hitting, 28, 48, 81, 153
  relaxation, 339
  return, 119
total variation, *see* distance
transform
  Fourier, 290
  Laplace, 46, 129, 221, 222, 235, 244, 264, 305, 325
traveling salesman, *see* problem, *see* problem
tree, 212
  binary, 60
  Galton–Watson, 38, 214, 221

conditioned on fixed size, 64
minimum spanning, 261
plane, 60
regular, 213
rooted, 60
uniform spanning, 63

urn
Ehrenfest, 107
Friedman, 87
generalized Pólya, 199, 204, 206
Hoppe, 193
Pólya, 87, 195

Vandermonde, *see* identity, *see* matrix
Vandermonde's identity, 26
von Neumann, 3

Wald, *see* confidence interval, *see* lemma
Wasserstein, *see* distance, *see* distance, *see* distance, *see* distance
Weierstrass, *see* theorem
Wiener, *see* process
Wigner, *see* theorem
simplified theorem, 277
theorem, 275
Wright–Fisher, *see* model

Yaglom, *see* theorem
Yule, *see* process, *see* process

A

Main notation and abbreviations

$d_{\mathrm{VT}}(\mu,\nu)$	total variation distance
$W_p(\mu,\nu)$	Wasserstein distance of order p
$L^p(\mu)$	space of functions f s.t. $\lvert f\rvert^p$ is μ-integrable
$L^p(E)$	space of functions f s.t. $\lvert f\rvert^p$ is Lebesgue integrable on E
L^p_{loc}	space of functions f s.t. $\lvert f\rvert^p$ is locally integrable
$\lVert f\rVert_p$	$L^p(\mu)$ norm of the function f
$\lVert M\rVert_{\mathrm{HS}}$	Hilbert–Schmidt norm $\sqrt{\mathrm{Tr}(MM^*)}$ of the matrix M
$\mathrm{Tr}(M)$	trace of the matrix M
$\lvert x\rvert$	norm, modulus, or absolute value of x
$\mathcal{S}_n$	symmetric group
$\lfloor x\rfloor$	greatest integer less than or equal to x (integer part)
$\lceil x\rceil$	least integer greater than or equal to x
log	natural logarithm
$X \overset{\mathrm{loi}}{=} Y$	equality in distribution
$X \sim \mu$	the random variable X has distribution μ
$x := a$	x is defined by a
$a =: x$	x is defined by a
$a \wedge b$	$\min(a,b)$
$a \vee b$	$\max(a,b)$
$\nu \ll \mu$	ν is absolutely continuous with respect to μ
$\mathrm{Ber}(p)$	Bernoulli distribution $n \mapsto p\mathbf{1}_{n=1} + (1-p)\mathbf{1}_{n=0}$
$\mathrm{Bin}(n,p)$	binomial distribution $k \mapsto \binom{n}{k} p^k (1-p)^{n-k} \mathbf{1}_{k\in\mathbb{N}\cap[0,n]}$
$\mathrm{Geo}(p)$	geometric distribution $n \mapsto (1-p)^{n-1} p \mathbf{1}_{n\in\mathbb{N}^*}$
$\mathrm{Geo}_{\mathbb{N}}(p)$	geometric distribution $n \mapsto (1-p)^{n} p \mathbf{1}_{n\in\mathbb{N}}$
$\mathrm{Exp}(\lambda)$	exponential distribution $x \mapsto \lambda e^{-\lambda x} \mathbf{1}_{x\geqslant 0}$
$\mathrm{Gamma}(a,\lambda)$	Gamma distribution $x \mapsto \lambda^a \Gamma(a)^{-1} x^{a-1} e^{-\lambda x} \mathbf{1}_{x\geqslant 0}$
Cauchy	Cauchy distribution $x \in \mathbb{R} \mapsto \frac{1}{\pi(1+x^2)}$
$\mathrm{Mul}(n,p)$	multinomial distribution $k \mapsto \binom{n}{k_1,\dots,k_d} p_1^{k_1} \cdots p_d^{k_d} \mathbf{1}_{k\in\mathbb{N}^n : k_1+\cdots+k_d=n}$
$\mathrm{Poi}(\lambda)$	Poisson distribution $n \mapsto e^{-\lambda} \frac{\lambda^n}{n!} \mathbf{1}_{n\in\mathbb{N}}$
$\mathrm{Unif}(I)$	uniform distribution on the set I
$\mathcal{N}(m,\sigma^2)$	Gaussian distribution $x \mapsto (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(x-m)^2}$
$\mathcal{N}(m,\Sigma)$	Gaussian distribution $x \mapsto (2\pi)^{-n/2} \det(\Sigma)^{-1/2} e^{-\frac{1}{2}\left\langle \Sigma^{-1}(x-m),(x-m)\right\rangle}$
$\chi^2(n)$	chi-squared distribution with n degrees of freedom $= \mathrm{Gamma}(n/2, 1/2)$
$t(n)$	Student's distribution with n degrees of freedom $= \mathcal{N}(0,1)/\sqrt{\chi^2(n)/n}$
$x_n = o_{n\to\infty}(y_n)$	$x_n/y_n \to 0$ as $n \to \infty$
$x = \mathcal{O}(y)$	$\lvert x\rvert \leqslant C\lvert y\rvert$ for some constant $C > 0$
$a_n \sim b_n$	$a_n/b_n \to 1$ as $n \to \infty$
i.i.d.	independent and identically distributed
a.s.	almost surely
iff	if and only if
r.v.	random variable(s)
real r.v.	real random variable(s)
s.t.	such that
LLN	Law of Large Numbers
CLT	Central Limit Theorem