Universitext

Nicolas Lanchier

Stochastic Modeling

Springer

Universitext

Universitext

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

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

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

Nicolas Lanchier

Stochastic Modeling

Springer

Nicolas Lanchier
School of Mathematical
and Statistical Sciences
Arizona State University
Tempe, AZ, USA

ISSN 0172-5939 ISSN 2191-6675 (electronic)
Universitext
ISBN 978-3-319-50037-9 ISBN 978-3-319-50038-6 (eBook)
DOI 10.1007/978-3-319-50038-6

Library of Congress Control Number: 2016962183

Mathematics Subject Classification (2010): 60-01, 60-02, 60F, 60G, 60J, 60K

Printed on acid-free paper

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

Têtes de plomb plaquées au sol
Sous les giffles de la lourde cascade des féeries

L'infini roulé sur les velours de sable
Cadence molle des tambours cosmiques
Pulsations pourpres

La peau se dilate
Les pores s'écarquillent
Pour percer les cieux de glace
Et mordre la chair-soleil

Décupler – goût toucher vision

Rouen, July 1995.

To my parents and my two daughters . . .
on both sides of the Atlantic Ocean.

Preface

This textbook originated from independent collections of lecture notes that slowly evolved and eventually merged together. The first embryonic version was a set of lecture notes on special stochastic processes with an emphasis on percolation models and interacting particle systems. Then came two more manuscripts both focusing on the general theory of stochastic processes developed as lecture notes for two courses on this topic, one for advanced undergraduate students and later its counterpart for graduate students based on measure theory. Since most of the proofs about percolation models, interacting particle systems, and other special stochastic models involve results from martingale theory, Poisson processes, and discrete-time and continuous-time Markov chains, it became natural to merge these lecture notes into the same document.

Material and structure

There are seventeen chapters that, together, form three coherent parts.

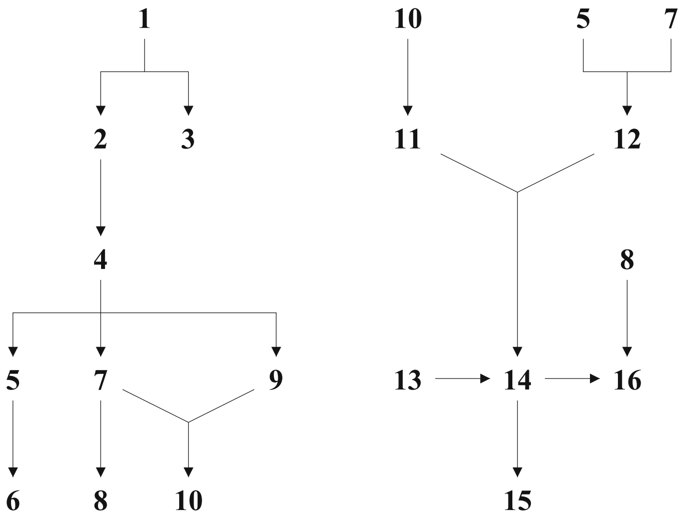
Part I. Probability theory — Chapters 1–3 cover the classical material on probability theory seen at the undergraduate level, from random variables to conditional probability and standard limit theorems. The main objective of this part is to redefine the main probability concepts with an emphasis on their connection to concepts from measure theory. Because the reader is expected to be familiar with basic probability and because the primary focus is not measure theory, the style is intentionally brief, focusing more on intuition rather than the technical details.

Part II. Stochastic processes — Chapters 4–10 cover the classical material on stochastic processes and, in contrast to the first three chapters, are mostly self-contained. This second part presents the main results about martingales, discrete-time Markov chains, Poisson processes, and continuous-time Markov chains. The theory is interspersed with examples of applications, including the popular gambler's ruin chain, branching processes, symmetric random walks on graphs, and birth and death processes, all of which illustrate the main theorems.

Part III. Special models — Chapters 11–17, which are less traditional and more research oriented, focus on models that arise in physics, biology, and sociology with an emphasis on minimal mathematical models that have been used historically to develop new techniques in the field of stochastic processes: the logistic growth process, the Wright–Fisher model, Kingman's coalescent, and some of the simplest percolation models and interacting particle systems. The proofs mostly rely on the subtle combination of results established in the second part and tailored-made arguments. Even though the proofs are model specific and involve a wide variety of techniques, the models covered in this last part form a coherent unit. While going through these seven chapters in the last part, the reader will discover how the models are connected to one another. The end of this part focuses on numerical simulations and summarizes these connections by exploring various algorithms that can be used to construct the models.

For an overview of the main topics mentioned in this textbook organized chronologically from the origin of probability theory in the seventeenth century until the end of the twentieth century along with the names of some of the main contributors, we refer the reader to the timeline at the very end of this textbook.

The logical dependence among the first sixteen chapters is shown in the following directed graph that we have broken down into two connected components to avoid having arrows crossing each other.



Note that Chapter 17 does not appear in the picture even though it gives an overview of the main models presented in this textbook and is therefore related to a number of other chapters. In fact, the reader can go directly to this chapter after reading the definition of the main models because there is no specific need to know about their analysis or the theory that supports them. We also point out that the following seven sections from the part on stochastic processes deal with more specific topics that are not required to understand the rest of the textbook:

Sections 5.3–5.5 from the chapter on martingale theory;

Section 6.4 on the number of individuals in the branching process;

Sections 8.2–8.3 about symmetric random walks;

Section 9.4 on the conditioning property for Poisson processes.

Each chapter starts with a brief overview illuminating the key steps and some additional readings on the topic. Diagrams summarizing the main concepts and results along with their connection are given at the end of the chapters covering the standard material. Exercises, with an emphasis on the classics of probability theory and real-world problems, are given at the end of most of the chapters. Exercises from the first two parts are purely analytical and rely on the techniques covered in the textbook, whereas the exercises in the third part on special models are more research oriented and consist of mixtures of analytical and simulation-based problems.

Prerequisites and course planning

There is a constant balance between mathematical rigor and brevity in order to cover the maximum number of topics in a minimum number of pages, but the reader interested in a specific topic is directed to references providing additional details or results. Even though this textbook is not that long, it seems to be ambitious to teach all the material in only one semester. The first two parts can typically be used to teach a general course on **random processes** for advanced undergraduate and graduate students in applied mathematics. The last part focusing on special stochastic processes can be used to teach a more research oriented **stochastic modeling** course for advanced graduate students in mathematics and applied sciences, while also using the beginning of the book as a refresher. Here are more details on the prerequisites and possible course planning for each of these two parts.

Random processes — The prerequisite for this material is a good knowledge of undergraduate probability: the basics of combinatorics and finite probability, conditional probability, discrete and continuous random variables, properties of expectation, and limit theorems. This is basically an undergraduate version, not relying on measure theory, of the material covered in the first part of this textbook, which is treated for example in *A First Course in Probability* by Sheldon Ross. All the theory in the first ten chapters can be covered in one semester, which should leave some time to also treat some of the examples and exercises. My preference, however, is to skip some of the seven more specific sections listed above in order to more frequently alternate between theory and illustrative practice problems, devoting about half of the lectures to each of these two aspects. The point here is that there are enough examples and exercises so that instructors can adjust the balance between theory and practice to better fit the background and taste of their students.

Stochastic modeling — As previously mentioned, the third part on special processes forms a coherent unit. In particular, skipping any aspect of this part might appear as a missing piece of the puzzle. For a course based on this material, it seems to be more appropriate that the students submit a short research project on the topic of their choice using the material covered in class rather than prepare for a final exam. The prerequisite for the material on special processes is precisely the material covered in the first two parts, with again the exception of the seven more specific sections mentioned above. My suggestion is to cover the prerequisite without treating any of the examples or exercises during the first half of the semester and the part on special processes during the second half of the semester. In my experience, it is also good to give some homework chosen from the exercise sections while covering the prerequisites, but let the students focus on their final research project while lecturing in special processes. In the case when students are already comfortable with the theory of random processes, the instructor can instead go directly to the third part and cover some of the most difficult and more research-oriented exercises.

Finally, while the first two parts have been specifically written for teaching purposes, the reader can also use the last part on special processes as a starting point outside the classroom to learn about a research topic.

Acknowledgements

The material of this textbook has co-evolved in parallel with teaching and the style has largely been shaped by fruitful interactions with students. The final version was submitted while Eric Foxall was working with me on various research projects, and our discussions about probability theory clearly had some beneficial influence on some of the aspects of this textbook. Going backward in time, the first set of lecture notes on special stochastic processes was initiated about eight years ago and is largely the result of my discussions with Claudia Neuhauser and Rick Durrett. Going back a few more years earlier, I am obviously grateful to the very first people who taught me probability theory while I was a graduate student in France, in particular my former advisor Claudio Landim who introduced me to the beautiful field of interacting particle systems. Finally, special thanks to four anonymous reviewers whose comments considerably helped improve the structure and clarity of this textbook, as well as to Donna Chernyk for coordinating the review process.

Phoenix, AZ, USA
June 2016

Nicolas Lanchier

Contents

Part I Probability theory

Part II Stochastic processes

Part III Special models

Part I
Probability theory

Chapter 1
Basics of measure and probability theory

The first use of mathematics to solve probability problems goes back to 1654 with the works of Fermat and Pascal. Their joint effort was motivated by questions raised by Antoine Gombaud, Chevalier de Méré, who was interested in betting strategies in the context of dice games. One of his main questions was: Is the probability of getting at least one double six when rolling two fair dice 24 times larger or smaller than one-half? A few years later, Huygens wrote the first book in probability theory [46] while, since the beginning of the 18th century, a number of mathematicians, including Bernoulli, de Moivre, Laplace, Poisson and Chebyshev, have made major discoveries in this field. But the most important turn in the history of probability since Fermat and Pascal is certainly the publication of Kolmogorov's monograph [56] in 1933 which defines the axiomatic foundations of probability theory and marks the beginning of modern probability. For an English translation, see [58]. Though this point of view might be simplistic, the main idea of his work was to redefine probability concepts from their analog in measure theory.

Following Kolmogorov's work, this first chapter gives a brief overview of the most basic concepts, definitions, and results of measure theory along with their connections with probability. Since the main topics of this textbook are probability theory and stochastic processes, most of the proofs are omitted and we refer the reader to Rudin [88] for the details. Instead, the emphasis is on explaining why measure theory is a key to developing probability theory. Measure theory and probability theory mostly differ in their terminology and interpretation so we will explain both at the same time. For a brief overview of the main concepts of measure theory that also underlines the connections with probability, see, for example, [8, 28, 79].

We start with a simple probability question showing the limitations of the Riemann integral [84] and the reason why measure theory is essential.

Question — What is the probability that a real number chosen uniformly at random in the unit interval is a rational number?

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_1

Applying concepts seen at the undergraduate level in order to compute this probability leads to the integral of a function that is not Riemann integrable, and therefore not even defined. In particular, to be able to answer this simple question, one needs to define a more general integral.

Motivated by the limitations of the Riemann integral, the second section introduces the first main concepts of measure theory: σ-algebras, measurable spaces, measurable functions, positive measures and finally the abstract integral with a construction due to Lebesgue [66]. The reader also learns that it is reasonable to interpret these concepts as events, random variables, probability measures, and expected values, which are the first concepts of probability theory covered at the undergraduate level. Along the way, some basic results are given in order to answer the question above and develop the theory later. For an exposition of the main probability concepts covered in this chapter and the next two chapters at the undergraduate level, we refer to the classical textbook of Ross [86]. See also [25] for detailed solutions to a number of interesting exercises and many historical notes.

In the last section, we turn our attention to some of the earliest and most important results in measure theory. Having a sequence of random variables with a pointwise limit, is the expected value of the limit equal to the limit of the expected values? In other words, do the limit and the expected value commute? We will see that, at least in the following two contexts, the answer is yes.

Monotone convergence — Limit and expected value commute when the sequence of random variables is monotone.

Dominated convergence — Limit and expected value commute when the sequence is dominated by an integrable random variable.

We also state another useful result, Fubini's theorem [37], which gives conditions under which two expected values or integrals commute. Counterexamples are also given showing that the assumptions of the previous theorems cannot be omitted.

1.1 Limitations of the Riemann integral

The Riemann integral was introduced by Bernhard Riemann in his qualification to become an instructor [84]. This short section shows the limitations of this integral through a simple example and the reason why measure theory is fundamental in probability theory. Consider the following problem: Find the probability that a number chosen uniformly at random in the unit interval is rational. To answer this question using tools from undergraduate probability, we let

$$X \sim \text{Uniform}\,(0,1) \quad \text{and} \quad f_X = \mathbf{1}_{(0,1)} = \text{density function of } X$$

where $\sim$ means that the law of the random variable on the left-hand side is given by the right-hand side. Then, the probability to be found is

$$P(X \in \mathbb{Q}) = \int_{\mathbb{Q}} f_X(x)\,dx = \int_0^1 \mathbf{1}_{\mathbb{Q}}(x)\,dx. \tag{1.1}$$

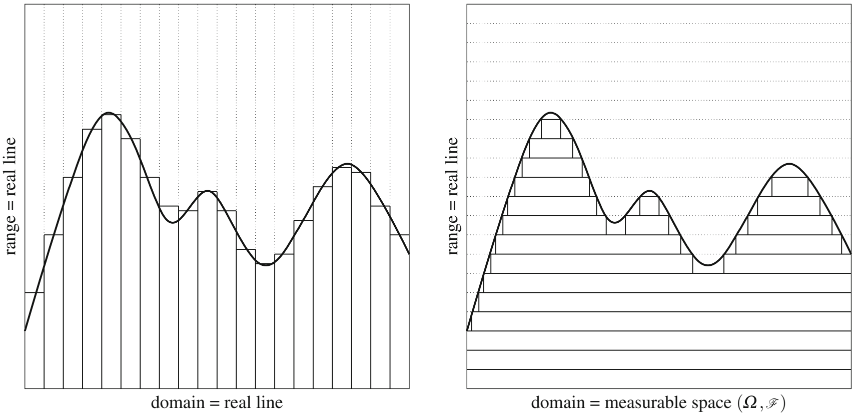


Fig. 1.1 Partitioning the domain (construction of the Riemann integral) versus partitioning the range (construction of the Lebesgue/abstract integral).

The function $\mathbf{1}_{\mathbb{Q}}$ is known as the Dirichlet function. To find its integral, recall that the Riemann sum of a function is defined from a tagged partition of the domain of the function as illustrated on the left-hand side of Figure 1.1. The function is said to be Riemann-integrable if and only if the sequence of the Riemann sums converges to a limit that does not depend on the choice of the sequence of tagged partitions as the partitions get finer and finer. In particular, since each open interval contains infinitely many rational numbers and infinitely many irrational numbers, the Dirichlet function is not Riemann-integrable so the integral in (1.1) does not exist. Using the framework of measure theory, however, we can properly define and compute this integral. This is done at the end of the next section.

1.2 Construction of the abstract integral

This section is devoted to the construction of the abstract integral, which was introduced by Henri Lebesgue in his Ph.D. dissertation [66] in an attempt to generalize the Riemann integral and make it more flexible. Like the Riemann integral, the abstract integral is a linear operator defined on a set of functions, but it is also more powerful for the following two reasons.

1. The abstract integral is much more general than the Riemann integral because there is in fact one integral associated to each so-called positive measure.
2. The special case of the Lebesgue measure gives rise to an integral that extends the Riemann integral to a much larger set of functions. This set of functions includes in particular the Dirichlet function.

To construct the abstract integral, we first need a few definitions.

Definition 1.1. Let Ω be a set. From the point of view of probability theory, we think of this set as a **sample space**: set of the **outcomes** of an experiment.

1. A collection $\mathscr{F}$ of subsets of Ω is said to be a **σ-algebra** whenever
   - $\Omega \in \mathscr{F}$
   - for all $A \in \mathscr{F}$, we have $A^c \in \mathscr{F}$
   - for each sequence $(A_n) \subset \mathscr{F}$, we have $\bigcup_n A_n \in \mathscr{F}$.
2. The pair $(\Omega, \mathscr{F})$ is then called a **measurable space**.
3. Members of the σ-algebra are called **measurable sets** in measure theory and are interpreted as **events** in probability theory.

It follows from the definition that the σ-algebra also contains the empty set and is stable under countable intersections. More generally, any set obtained from elementary set operations involving countably many measurable sets is again measurable. In the context of probability theory, the σ-algebra or set of events represents the information available: the largest σ-algebra, which consists of all subsets of Ω, means perfect information, whereas the smallest one, which reduces to the sample space and the empty set, means no information. For any collection $\mathscr{H}$ of subsets of Ω, there is a smallest σ-algebra that contains $\mathscr{H}$. It is called the **σ-algebra generated by the collection** $\mathscr{H}$ and it is traditionally written $\sigma(\mathscr{H})$. The usual σ-algebra on the real line is the one generated by the open intervals:

$$\mathscr{B} = \sigma\{(a,b) : a,b \in \mathbb{R} \text{ and } a < b\} \tag{1.2}$$

and is called the **Borel σ-algebra**. More generally, the Borel σ-algebra on a topological space is the one generated by the open sets. In view of the large collection of sets that can be produced using elementary set operations involving countably many open intervals, it seems that any subset of the real line is a Borel set. In fact, one can prove using transfinite induction that the cardinality of the set of Borel measurable sets is $\mathfrak{c}$, the cardinality of the set of real numbers. Since the cardinality of the set of subsets of the real line is $2^{\mathfrak{c}} > \mathfrak{c}$, it follows that the number of subsets of the real line that are not Borel sets is uncountable.

The next step to construct the Lebesgue integral is to define the concept of **measurability** of a function, which is the analog of continuity in topology.

Definition 1.2. A function $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ is called a **measurable function** in measure theory and a **random variable** in probability theory if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{F} \quad \text{for all} \quad B \in \mathscr{B}. \tag{1.3}$$

The set of all measurable functions is denoted by $\mathscr{M}(\Omega, \mathscr{F})$.

Using that the Borel σ-algebra (1.2) is generated by the open intervals, it can be proved that a function is a measurable function/random variable if and only if the inverse image of any open interval is measurable. In fact, to prove that X is measurable, it is even enough to prove that

$$X^{-1}((-\infty,a]) = \{\omega \in \Omega : X(\omega) \le a\} \in \mathscr{F} \quad \text{for all } a \in \mathbb{R}. \tag{1.4}$$

Using this characterization, one can prove that the positive part, negative part, and absolute value of a measurable function are again measurable. Likewise, the supremum, infimum, limit superior, and limit inferior of a sequence of measurable functions are measurable. It is easy to check that the collection of the inverse images of the Borel sets by a function X, i.e.,

$$\sigma(X) = \{A \subset \Omega : A = X^{-1}(B) \text{ for some } B \in \mathscr{B}\}$$

is the smallest σ-algebra that makes X measurable. This σ-algebra is called the **σ-algebra generated by the function** X. In the context of probability theory, the measurability of a random variable is a natural assumption that expresses the respect for the information. Indeed, the information that can be observed from the random variable X is represented by $\sigma(X)$ and we will see later that the probability of $A \in \sigma(X)$ is only well-defined when the set A is an event. In particular, to have a well-defined theory, all sets in $\sigma(X)$ must be measurable, which is exactly the definition of the measurability (1.3) of a function. Here are two other properties regarding random variables that will be useful later to study stochastic processes.

- Letting (X_n) be a sequence of random variables and T be a positive integer-valued random variable, the function X_T is again a random variable.
- Let X and Y be real random variables. Then, the function Y is $\sigma(X)$-measurable if and only if $Y = \phi(X)$ for some measurable function ϕ.

The first property will ensure suitable measurability properties when dealing with stochastic processes and so-called stopping times, while the second property is the key ingredient to check in practice if a process is Markovian. These properties appear as exercises in this chapter. Returning to the general context of measure theory, key measurable functions to define the integral are called simple functions.

Definition 1.3. A function $s : \Omega \to \mathbb{R}$ is a **simple function** whenever its range is finite. In particular, letting $a_1, a_2, \ldots, a_n \neq 0$ be the distinct values of s,

$$s = a_1 \mathbf{1}_{A_1} + \cdots + a_n \mathbf{1}_{A_n} \quad \text{where} \quad A_i = \{\omega \in \Omega : s(\omega) = a_i\}. \tag{1.5}$$

Note that, the set Ω being equipped with a σ-algebra $\mathscr{F}$, the simple function s is measurable if and only if all the sets A_i are measurable. Another result that will be useful later is the fact that any positive measurable function X can be written as the pointwise limit of a nondecreasing sequence (s_n) of simple measurable functions, and an example of such a sequence is

$$s_n(\omega) = \min(n, 2^{-n}\lfloor 2^n X(\omega)\rfloor) \quad \text{where} \quad \lfloor \cdot \rfloor \text{ refers to the integer part.} \tag{1.6}$$

From now on, we write $\mathscr{S}(\Omega,\mathscr{F})$ for the set of all simple measurable functions.

In order to construct the abstract integral introduced in Lebesgue's Ph.D. dissertation, the last step is to define positive measures.

Definition 1.4. Let $(\Omega, \mathscr{F})$ be a measurable space.

1. A **positive measure** on $(\Omega, \mathscr{F})$ is a function $\mu : \mathscr{F} \to [0, \infty]$ which is **σ-additive**, i.e., for each sequence $(A_n) \subset \mathscr{F}$,
$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n) \quad \text{whenever} \quad A_i \cap A_j = \varnothing \text{ for } i \neq j.$$
2. If in addition $\mu(\Omega) = 1$ then μ is called a **probability measure**.
3. The triplet $(\Omega, \mathscr{F}, \mu)$ is then called a **measure space** in measure theory and a **probability space** in probability theory.

When dealing with a probability measure, we use the notation P for probability as opposed to μ for a general positive measure. Throughout this textbook and more generally across the literature on probability theory and stochastic processes, the expression **almost surely** is copiously used. Having defined positive and probability measures, we can now rigorously define this concept.

Definition 1.5 (Almost surely). Letting $(\Omega, \mathscr{F}, \mu)$ be a measure space, we say that a property holds almost everywhere for general measure spaces and almost surely for probability spaces if the set of elements ω for which the property does not hold has measure zero. For instance, we say that two real random variables X and Y are equal almost surely whenever the following holds:
$$P(\{\omega \in \Omega : X(\omega) \neq Y(\omega)\}) = 0.$$

From now on, any property involving random variables written with no further precision will be implicitly assumed to hold almost surely. For example, having two random variables X and Y, we will often write
$$\begin{array}{lll} X = Y & \text{instead of} & X = Y \text{ almost surely} \\ X \leq Y & \text{instead of} & X \leq Y \text{ almost surely.} \end{array}$$

Probability measures have a number of very useful properties that can typically be guessed using a so-called Venn diagram. We only state without proof some of the main properties. Letting P be a probability measure,

- **Empty set**: the probability of the empty event is zero, i.e., $P(\varnothing) = 0$.
- **Additivity**: for all events $A_1, A_2, \ldots, A_n \in \mathscr{F}$,
$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \quad \text{whenever} \quad A_i \cap A_j = \varnothing \text{ for } i \neq j.$$
- **Complement**: $P(A^c) = 1 - P(A)$ for every event $A \in \mathscr{F}$.
- **Monotonicity**: $P(A) \leq P(B)$ whenever $A \subset B$.

- **Monotone convergence**: Let $(A_n) \subset \mathscr{F}$. Then,

$$P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n) \quad \text{when } (A_n) \text{ is monotone.} \tag{1.7}$$

- **Inclusion-exclusion identity**: Let $A_1, A_2, \dots, A_n \in \mathscr{F}$. Then,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \cdots < i_k} P(A_{i_1} \cap \cdots \cap A_{i_k}). \tag{1.8}$$

We point out that the monotone convergence property can be viewed as a weak version of the monotone convergence theorem stated below. Note also that the inclusion–exclusion identity states that the probability of a finite union of events is equal to the sum of the probabilities of these events, minus the sum of the probabilities of all the double intersections, plus the sum of the probabilities of all the triple intersections, and so on. Figure 1.2 gives the intuition behind this equation in the case of three events. Before constructing the integral, we also point out that, for a number of random experiments such as flipping a coin, rolling a die, or playing poker, the set of outcomes is finite and it is natural to assume that all the outcomes have the same probability. In this case, by additivity,

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} \frac{1}{\operatorname{card}(\Omega)} = \frac{\operatorname{card}(A)}{\operatorname{card}(\Omega)}.$$

In particular, computing a probability reduces in this context to a counting or combinatorial problem so we have reported the most important counting rules, namely the number of arrangements, permutations, and combinations with and without repetition in the table of Figure 1.3.

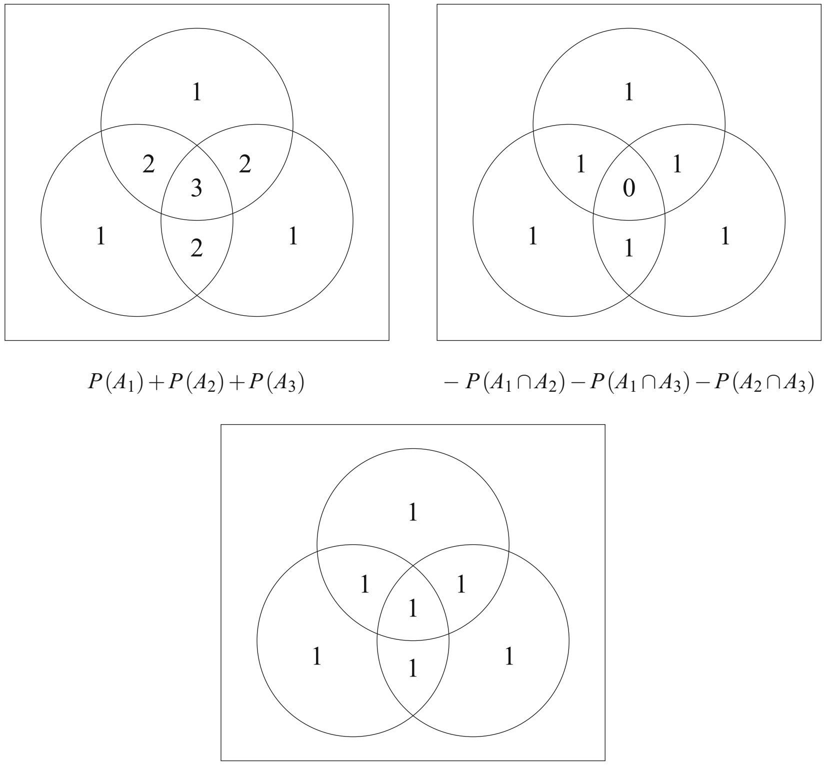
Having defined positive measures, we can now construct the integral. Provided it exists, the integral of a function X with respect to a measure μ is written

$$\int X \, d\mu = \int X(\omega) \, \mu(d\omega).$$

The construction of the abstract integral is divided into four steps starting with indicator functions, then simple functions, then positive functions, and finally integrable functions. This construction is rarely used to compute integrals in practice, but it is essential to develop the theory. In particular, a number of results in measure theory can be proved conveniently by first considering indicator functions and then following the construction to extend the result to more general functions. The common assumption across all four steps of the construction is that we are dealing with measurable functions with value in the real line equipped with its Borel σ-algebra.

Step 1 — Indicator functions. As a starting point, the integral of $X = \mathbf{1}_A$ with respect to a positive measure μ is assumed to be the measure $\mu(A)$, i.e.,

$$\int X \, d\mu = \int \mathbf{1}_A \, d\mu = \mu(A).$$



$P(A_1)+P(A_2)+P(A_3)$

$-P(A_1 \cap A_2)-P(A_1 \cap A_3)-P(A_2 \cap A_3)$

$+P(A_1 \cap A_2 \cap A_3)$

Fig. 1.2 Intuition behind the inclusion–exclusion identity with $n=3$. The picture on the left-hand side overestimates the probability of the intersections of two and three events while the picture in the middle underestimates the probability of the intersection of all three events.

Step 2 — Simple functions. Having a finite range, each simple function X can be written as the linear combination of indicator functions. In order for the integral to be linear, we assume that the integral of X is the linear combination of the integrals of the indicator functions:

$$\int X\,d\mu = \int \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}\,d\mu = \sum_{i=1}^{n} a_i\,\mu(A_i).$$

This definition is coherent because it can be proved that

$$a_1 \mathbf{1}_{A_1} + \cdots + a_n \mathbf{1}_{A_n} = b_1 \mathbf{1}_{B_1} + \cdots + b_m \mathbf{1}_{B_m} \quad \text{implies that}$$
$$a_1\,\mu(A_1) + \cdots + a_n\,\mu(A_n) = b_1\,\mu(B_1) + \cdots + b_m\,\mu(B_m).$$

Step 3 — Positive functions. Motivated by (1.6), i.e., the fact that any positivemeasurable function X can be written as the pointwise limit of a nondecreasing sequence of simple measurable functions, we define the integral of X as the supremum

$$\int X\,d\mu = \sup_{s \in \mathscr{S}(\Omega,\mathscr{F})} \int s\,d\mu.$$

Permutation without repetition # rankings of n distinct blocks $n(n-1)\cdots 1 = n!$	**Arrangement without repetition** # rows of k blocks chosen from a set of n distinct blocks $n\cdots(n-k+1) = \frac{n!}{(n-k)!}$	**Combination without repetition** # subsets of k blocks chosen from a set of n distinct blocks $\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$
Permutation with repetition # rankings of $n = n_1 + \cdots + n_k$ blocks with n_i blocks of color i $\binom{n}{n_1,\ldots,n_k} = \frac{n!}{n_1!\cdots n_k!}$	**Arrangement with repetition** # rows of k blocks, each of which can be of n different colors $n\ldots n = n^k$	**Combination with repetition** # subsets of k blocks, each of which can be of n different colors $\binom{n+k-1}{n-1}$

Fig. 1.3 Brief summary of the main counting rules.

Step 4 — Integrable functions. In order again for the integral to be linear, we define the integral of a general measurable function X as the difference between the integral of two positive functions: its positive part and its negative part, i.e.,

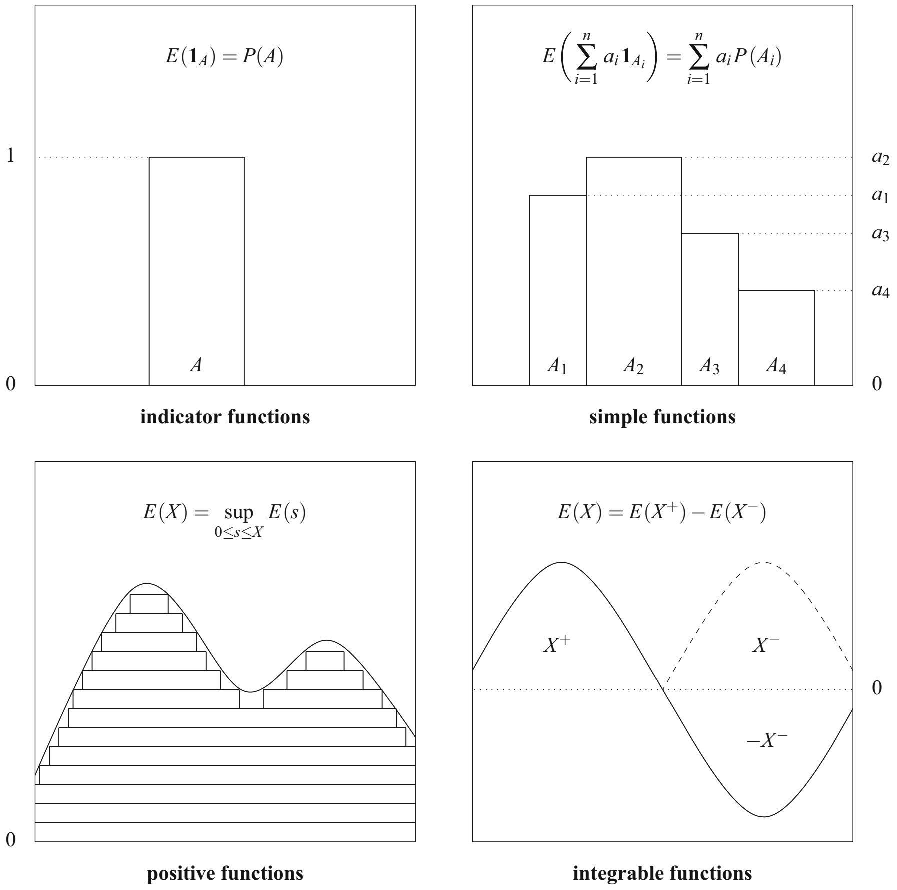
$$\int X\,d\mu = \int X^+\,d\mu - \int X^-\,d\mu.$$

Note that the last two integrals might be infinite, thus leading to an indetermination, so we need to restrict the definition to functions with an absolute value that has a finite integral. Such functions are called **integrable functions**.

By convention, we assume that $0\cdot\infty = \infty\cdot 0 = 0$. In particular, the integral of the function identically equal to zero over a set with infinite measure is zero. In probability theory, the integral of a random variable X is nothing else than the **expected value** of this random variable and we write

$$E(X) = E_P(X) = \int X\,dP$$

where the measure P is the underlying probability measure. Since this textbook ismainly about probability theory, we adopt from now on the expected value notation, but will switch sometimes to the integral notation when dealing with specific measures which are not probability measures or results that will be needed in the general context of positive measures. Figure 1.4 gives an illustration of the four steps of the construction using the expected value notation. The expected value has


$E(\mathbf{1}_A) = P(A)$
1
0
A
$E\left(\sum_{i=1}^{n} a_i \mathbf{1}_{A_i}\right) = \sum_{i=1}^{n} a_i P(A_i)$
a_2
a_1
a_3
a_4
0
A_1
A_2
A_3
A_4
$E(X) = \sup_{0 \leq s \leq X} E(s)$
0
$E(X) = E(X^+) - E(X^-)$
X^+
X^-
0
$-X^-$

indicator functions

simple functions

positive functions

integrable functions

Fig. 1.4 Construction of the abstract integral.

a number of nice properties that directly follow from the construction of the integral. The main three properties we will use across this textbook are

- **Linearity**: for all X, Y integrable and $a, b \in \mathbb{R}$,
$$E(aX + bY) = aE(X) + bE(Y).$$
- **Monotonicity**: $X \leq Y$ implies that $E(X) \leq E(Y)$.
- **Jensen's inequality**: $\phi(E(X)) \leq E(\phi(X))$ for all convex functions ϕ.

We point out that linearity and monotonicity are true for general positive measures whereas Jensen's inequality only holds for probability measures. Later on, the set of integrable random variables will be written

$$L^1(\Omega, \mathscr{F}, P) = \{X \in \mathscr{M}(\Omega, \mathscr{F}) : E|X| < \infty\}.$$

On this set, it is natural to define the operator

$$X \mapsto \|X\|_1 = E\,|X| \quad \text{for all} \quad X \in L^1(\Omega, \mathscr{F}, P).$$

It follows from the linearity of the integral that $\|\cdot\|_1$ is absolutely homogeneous and satisfies the triangle inequality so this is a seminorm. Note, however, that

$$\|X\|_1 = E\,|X| = 0 \quad \text{if and only if} \quad X = 0 \text{ almost surely.}$$

This shows that $\|\cdot\|_1$ is not a norm on the set of integrable random variables but that it indeed defines a norm on a certain set of equivalence classes of integrable random variables where two random variables X and Y belong to the same class if and only if they are equal almost surely. More generally, one can define

$$L^p(\Omega, \mathscr{F}, P) = \{X \in \mathscr{M}(\Omega, \mathscr{F}) : E\,|X|^p < \infty\} \quad \text{for all} \quad 1 < p < \infty.$$

In this case, the set of equivalence classes of random variables, where again two random variables are equivalent if and only if they are equal almost surely, is a normed space for the norm

$$\|X\|_p = (E\,|X|^p)^{1/p} \quad \text{for all} \quad X \in L^p(\Omega, \mathscr{F}, P).$$

These function spaces of equivalence classes are called **Lebesgue spaces** and will appear later when studying the convergence of random variables. Having a pair of conjugate exponents, i.e., two real numbers

$$p, q > 1 \quad \text{such that} \quad 1/p + 1/q = 1$$

we have the following useful result called **Hölder's inequality**:

$$\|XY\|_1 \le \|X\|_p \, \|Y\|_q \quad \text{for all} \quad X, Y \in \mathscr{M}(\Omega, \mathscr{F}). \tag{1.9}$$

Figure 1.5 gives a summary of the main concepts introduced in this section including the terminology from both measure theory and probability theory.

Returning for a moment to the general context of measure theory, a particular case of interest in analysis is the integral with respect to the **Lebesgue measure**, usually written λ, which can be proved to be the unique measure on the Borel sets that basically measures the length of the intervals, i.e.,

$$\lambda\,((a,b]) = b - a \quad \text{for all} \quad a \le b.$$

It turns out that the integral with respect to the Lebesgue measure coincides with the Riemann integral on the set of Riemann-integrable functions, but the set of functions that are Lebesgue-integrable is much larger. To illustrate this aspect, we nowanswer the question raised at the beginning of this chapter and compute the integral (1.1).

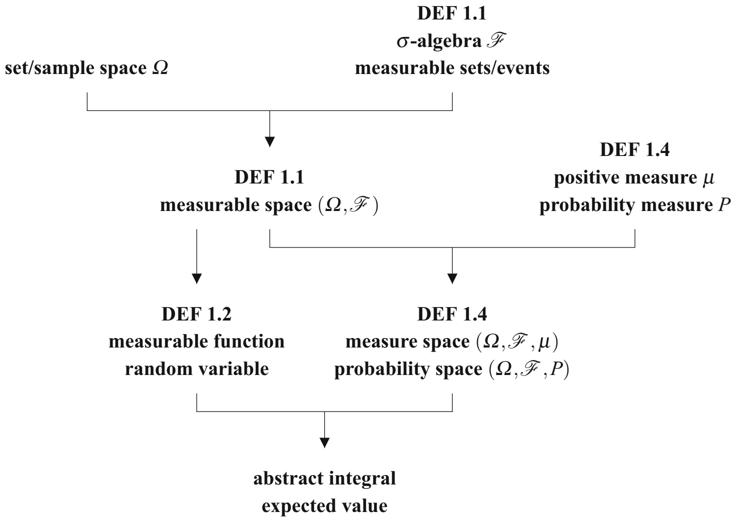


Fig. 1.5 Main concepts introduced in Section 1.2.

Since the set of rational numbers is countable, there is a sequence of rational numbers (x_n) such that

$$\mathbb{Q}\cap(0,1)=\{x_1,x_2,\ldots,x_n,\ldots\} \quad \text{and} \quad x_i\neq x_j \quad \text{for all} \quad i\neq j.$$

Then, by σ-additivity, we have

$$P(X\in\mathbb{Q})=\lambda(\mathbb{Q}\cap(0,1))=\textstyle\sum_i \lambda(\{x_i\})=\sum_i (x_i-x_i)=0.$$

More generally, our reasoning shows that the Lebesgue measure of any countable set is equal to zero but we point out that the converse is not true. For example, the Cantor set is an example of uncountable set with Lebesgue measure zero. The example of the Dirichlet function reveals another important shortcoming of the Riemann integral, namely, the functions

$$X_n=\mathbf{1}\{x_1,x_2,\ldots,x_n\} \quad \text{for} \quad n\in\mathbb{N}^*$$

define an increasing sequence of Riemann-integrable functions with integral equal to zero that converges pointwise to a function that is not Riemann-integrable. In contrast, the abstract integral is consistent in such a context in the sense that the sequence of the integrals converges to the integral of the limit. This aspect is discussed in detail in the following section.

1.3 Main properties of the integral

In this section, we state additional properties of the abstract integral that are particularly useful in both analysis and probability theory. The first properties give sufficient conditions under which, having a sequence of measurable functions converging to a pointwise limit, the limit and integral commute: the monotone convergence theorem and the dominated convergence theorem. Both theorems are stated for a probability measure because it is in this context that they will be used later, but they hold more generally for positive measures.

Theorem 1.1 (Monotone convergence). *Let (X_n) be a nondecreasing sequence of positive random variables and let X be its limit. Then,*

$$X \in \mathscr{M}(\Omega, \mathscr{F}) \quad \textit{and} \quad \lim_{n\to\infty} E(X_n) = E(X).$$

Applying the monotone convergence theorem to

$$Z_n(\omega) = \inf_{i \geq n} X_i(\omega) \quad \text{for all} \quad \omega \in \Omega$$

where (X_n) is a sequence of positive random variables not necessarily monotone, we obtain the following result which is not only useful in practice but also the key to proving the dominated convergence theorem.

Lemma 1.1 (Fatou's lemma). *For any sequence (X_n) of positive random variables, the limit inferior of the sequence is a random variable and*

$$E(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} E(X_n).$$

Using Fatou's lemma, we obtain the following.

Theorem 1.2 (Dominated convergence). *Let (X_n) be a sequence of random variables with pointwise limit X and dominated by an integrable random variable:*

$$|X_n| \leq Y \quad \textit{for all} \;\; n \geq 1 \;\; \textit{and some} \;\; Y \in L^1(\Omega, \mathscr{F}, P).$$

Then, the limit X is integrable,

$$\lim_{n\to\infty} E\,|X_n - X| = 0 \quad \textit{and} \quad \lim_{n\to\infty} E(X_n) = E(X).$$

For detailed proofs of the previous three results, see [88, Chapter 1].

Before moving to the next property, we give a simple counterexample showing that limit and integral do not always commute when the assumptions of the monotone convergence theorem and dominated convergence theorem are not satisfied. To see this, consider the integral with respect to the Lebesgue measure and the sequence of functions such that the graph of X_n connects the points

$$(2^{-(n+1)}, 0) \to ((1/2)(2^{-(n+1)} + 2^{-n}), 2^{n+2}) \to (2^{-n}, 0).$$

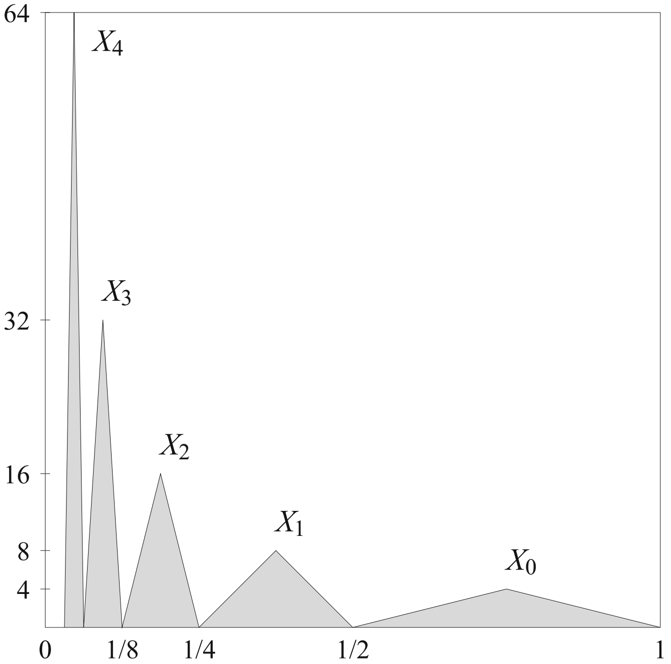


Fig. 1.6 Example of a sequence for which limit and integral do not commute.

See Figure 1.6. This sequence is neither monotone nor dominated by an integrable function and, since it converges to the function identically equal to zero,

$$\lim_{n\to\infty}\int_0^1 X_n\,d\lambda = 1 \neq 0 = \int_0^1 \lim_{n\to\infty} X_n\,d\lambda$$

showing that limit and integral do not commute.

We now state Fubini's theorem, which was proved in [37]. While the monotone and dominated convergence theorems give sufficient conditions under which limit and integral commute, this result gives sufficient conditions under which two integrals commute when looking at functions of two variables. Again, we only state the theorem and give a counterexample and refer the reader to [88, chapter 8] for a detailed proof. To begin with, we consider two measure spaces

$$(S,\mathscr{S},\mu_S) \quad \text{and} \quad (T,\mathscr{T},\mu_T).$$

To extend the concepts seen so far to functions of two variables defined on the Cartesian product $\Omega = S\times T$, we first introduce the σ-algebra

$$\mathscr{F} = \mathscr{S}\times\mathscr{T} = \sigma\{A\times B : A\in\mathscr{S} \text{ and } B\in\mathscr{T}\}$$

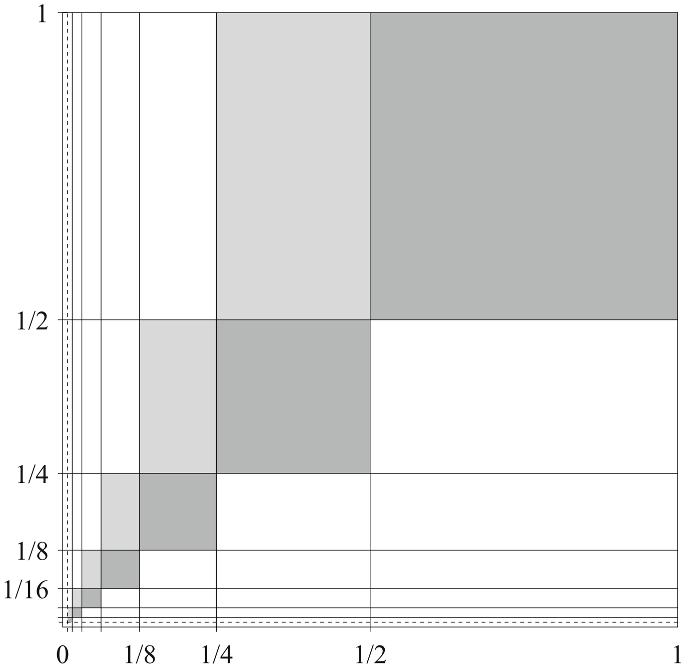


Fig. 1.7 Domain of the function of two variables for which the order of integration matters.

generated by the so-called rectangles. Then, it can be proved that there exists a unique measure μ on the σ-algebra $\mathscr{F}$ such that

$$\mu(A \times B) = \mu_S(A)\,\mu_T(B) \quad \text{for all} \quad A \in \mathscr{S} \text{ and } B \in \mathscr{T}.$$

This measure is called **product measure** and is denoted by $\mu = \mu_S \times \mu_T$. In contrast with the monotone and dominated convergence theorems, Fubini's theorem only holds for so-called σ-finite spaces, which are defined as follows.

Definition 1.6. A measure space $(\Omega, \mathscr{F}, \mu)$ is said to be **σ-finite** whenever Ω can be written as a countable union of measurable sets with finite measure.

Theorem 1.3 (Fubini). *Let $X : (\Omega, \mathscr{F}, \mu) \to (\mathbb{R}, \mathscr{B})$ be a positive measurable function or an integrable function on a σ-finite measure space. Then,*

- *For all $s \in S$ and all $t \in T$,*

$$X(s, \cdot\,) \in \mathscr{M}(T, \mathscr{T}) \quad \textit{and} \quad X(\,\cdot\,, t) \in \mathscr{M}(S, \mathscr{S}).$$

- *We have $\phi \in \mathscr{M}(S, \mathscr{S})$ and $\psi \in \mathscr{M}(T, \mathscr{T})$ where*

$$\phi(s) = \int_T X(s,t)\,\mu_T(dt) \quad \textit{and} \quad \psi(t) = \int_S X(s,t)\,\mu_S(ds).$$

- *The two integrals commute in the sense that*

$$\int_S \int_T X \, d\mu_T \, d\mu_S = \int_{S \times T} X \, d\mu = \int_T \int_S X \, d\mu_S \, d\mu_T .$$

In practice, we often use that the double integral on the left-hand side equals the double integral on the right-hand side.

As previously mentioned, we conclude this section with a counterexample showing that, when the function X is neither positive nor integrable, the two integrals may not commute. To construct a counterexample, recall the sequence (X_n) introduced above as a counterexample of the monotone convergence and dominated convergence theorems. Then, consider the function X defined on the unit square as

$$X(s,t) = \begin{cases} +X_n(s)X_n(t) & \text{for } 2^{-(n+1)} < s,t < 2^{-n} \\ -X_{n+1}(s)X_n(t) & \text{for } 2^{-(n+2)} < s < 2^{-(n+1)} < t < 2^{-n} \end{cases}$$

and equal to zero for all other values of s and t. Figure 1.7 shows a partition of the unit square where the dark grey squares represent the region where X is positive, the pale grey rectangles the region where it is negative, and the white rectangles the region where it is equal to zero. Using that the integral of each X_n is equal to one, some basic algebra shows that, one variable being fixed, the integral of X with respect to the other variable depends on whether or not the corresponding cross-section intersects a dark or a pale grey rectangle and we find

$$\int_0^1 X(s,t) \, ds = 0 \quad \text{and} \quad \int_0^1 X(s,t) \, dt = \begin{cases} 0 & \text{for } s < 1/2 \\ X_1(s) & \text{for } s > 1/2. \end{cases}$$

It directly follows that

$$\int_0^1 \int_0^1 X(s,t) \, dt \, ds = \int_0^1 X_1(s) \, ds = 1 \neq 0 = \int_0^1 \int_0^1 X(s,t) \, ds \, dt$$

showing that the two integrals do not commute.

1.4 Exercises

Measure theory

Exercise 1.1. Let (X_n) be a sequence of real random variables and T be a positive integer-valued random variable. Prove that

$$Z : (\Omega, \mathscr{F}, P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R})) \quad \text{with} \quad Z(\omega) = X_{T(\omega)}(\omega)$$

is a random variable.

Exercise 1.2. Let (X_n) be a sequence of real measurable functions. Prove that its set of convergence is measurable, that is,

$$B = \{\omega \in \Omega : \limsup_{n\to\infty} X_n = \liminf_{n\to\infty} X_n\} \in \mathscr{F}.$$

Exercise 1.3. Let X and Y be two real random variables. Show that the following two properties are equivalent:

1. the function Y is $\sigma(X)$-measurable,
2. there is $\phi : (\mathbb{R}, \mathscr{B}(\mathbb{R})) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ measurable such that $Y = \phi(X)$.

Hint: Prove first the result when Y is a simple function, then deduce the general result using that Y is the pointwise limit of a sequence of simple functions.

Exercise 1.4. Let $(A_n) \subset \mathscr{F}$ be a nondecreasing sequence of events. Prove the weak version of the monotone convergence theorem, namely,

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n).$$

Hint: Define $B_n = A_n \setminus A_{n-1}$ and use σ-additivity.

Symmetric probability spaces

Exercise 1.5 (Chevalier de Méré). Prove that the probability of getting at least one double six when rolling two fair dice 24 times is smaller than $1/2$.

Exercise 1.6 (Poker probabilities). Recall that a poker hand is a set of five cards chosen from an ordinary deck of 52 cards. Assuming that all the poker hands are equally likely, find the probability of being dealt

1. one pair, i.e., two cards from the same denomination and three cards from three other denominations,
2. two pairs, i.e., two cards from the same denomination, two other cards from another denomination, and one more card from a third denomination,
3. three of a kind, i.e., three cards from the same denomination and two cards from two other denominations,
4. a full house, i.e., three cards from the same denomination and two other cards from another denomination,
5. four of a kind, i.e., four cards from the same denomination and one card necessarily from another denomination,
6. a flush, i.e., five cards from the same suit.

Exercise 1.7. How many poker hands are needed so that the probability that at least one of them contains exactly one pair exceeds 0.9 under the natural assumption that all the poker hands are independent and equally likely?

Exercise 1.8 (Poker dice). Rolling five fair dice simultaneously and assuming that all the outcomes are equally likely, find the probability of

1. one pair, i.e., exactly two dice land one the same value while the other three dice land on three other values,
2. two pairs, i.e., two dice land one the same value, two other dice land on another value and the last die land on a third value,
3. three alike, i.e., exactly three dice land on the same value and the other two dice on two other values,
4. a full house, i.e., exactly three dice land on the same value and the other two dice land on another value,
5. four alike, i.e., exactly four dice land on the same value and the last die lands on another value,
6. five alike, i.e., all five dice land on the same value.

Hint: It is easier to compute these probabilities thinking of the dice as being distinguishable, which gives 6^5 outcomes that are equally likely.

Exercise 1.9 (Monty Hall problem). Suppose that you are given the choice among three doors: Behind one door there is a car while behind the other two doors are goats. You pick a door at random. Then, someone opens another door which reveals a goat and you are given the opportunity to keep the same door or choose the other door. Is there a better strategy to find the car?

Exercise 1.10 (Birthday problem). Assume that a host invites n guests and that the birthdays of all $n+1$ people are independent and uniformly distributed across all the 365 days of the year.

1. Prove the following: The probability that two of the guests share the same birthday is larger than one-half if and only if $n \geq 23$.
2. How large should n be so that the probability that at least one of the guests has the same birthday as the host is larger than one-half?

Exercise 1.11. An old man just turned 100. Throughout his life, he collected all the candles from his birthday cakes, which he keeps in a box. If he takes one candle out of the box at random, what is the probability that it is from his 100th birthday?

Exercise 1.12 (Anagrams). Let W be a word of length n using k different letters and we denote by n_i the number of times the letter i is used. Let also W' be an anagram of W obtained by choosing a permutation of the n letters uniformly at random. Find the probability that the two words W and W' are identical.

Exercise 1.13. Assume that n people, including Jules and Jim, sit in a row, with all the possible permutations being equally likely.

1. Find the probability that Jules and Jim are next to each other.
2. What does this probability become if the n people sit at a round table?

Exercise 1.14. Player i flips a fair coin n_i times where $i = 1,2$. Prove that the probability that both players obtain the same number of heads is equal to the probability that the total number of heads is n_1.

Hint: This can be done without computing the probability explicitly.

Exercise 1.15. Assume that the numbers $1,2,\ldots,n$ are randomly given to players labeled $1,2,\ldots,n$. Initially, player 1 and player 2 compare their numbers. The one with the largest number wins and compares her number with player 3, and so on. Find the probability that player 1 wins m times.

Hint: Use that, for every subset of numbers chosen uniformly at random, all the possible permutations of these numbers are equally likely.

Exercise 1.16. Let $A = \{1,2,\ldots,m\}$ and $B = \{1,2,\ldots,n\}$ and define

$$\phi(x) \sim \text{Uniform}(B) \text{ be independent for all } x \in A.$$

1. Find the probability that the function ϕ is increasing when $m \leq n$.
2. Deduce the probability that ϕ is nondecreasing for all values of m and n.

Hint: For the second question, use the new function $\psi(x) = \phi(x) + x$.

Exercise 1.17. Recall that a chessboard is an 8 × 8 grid of squares. Assuming that eight rooks are randomly placed on a chessboard, compute the probability that none of the rooks can capture any of the others (rooks can move any number of squares horizontally or vertically).

Exercise 1.18. Take $s_n \leq n$ shoes at random from a drawer containing n pairs.

1. Use Stirling formula (8.2) to study the asymptotic behavior of the probability of the event A_n that no complete pair is selected when $s_n = n$.
2. Let X_n be the number of complete pairs. Compute

$$E(X_n) \quad \text{when} \quad s_n = \lfloor \mu n \rfloor \ \text{ with } \ \mu \in (0,1).$$

   Study again the asymptotic behavior as $n \to \infty$.

Exercise 1.19 (Coupon collector's problem). There are n different types of coupons and each time one buys a coupon, it is equally likely to be of each type.

1. Compute the expected number of types collected X_n after one buys $\lfloor \mu n \rfloor$ coupons and study the asymptotic behavior as $n \to \infty$.
2. Compute the expected number of coupons T_n that one has to buy to have the full collection and study the asymptotic behavior as $n \to \infty$.

Exercise 1.20. Consider a tournament in which each of the n contestants plays exactly once with each of the other contestants. Use a probabilistic argument to prove the following result: it is possible that, for each set of k players, there is a player who beats each member of that set whenever

$$\binom{n}{k}\left(1-\left(\frac{1}{2}\right)^k\right)^{n-k} < 1.$$

Hint: Compute the probability that, for each set of k players, there is a player who beats each other member of that set under the assumption that each game is equally likely to be won by each of the two opponents.

Exercise 1.21 (Ramsey numbers). Let $K_n = (V,E)$ be the complete graph with n vertices, i.e., any two vertices are connected by an edge, and color each edge of the graph using one of two possible colors. Ramsey theorem states that for every integer k, there exists $R(k)$ finite such that the following holds: for every coloring, there exists a subset $S \subset V$ of size k that is monochromatic, i.e., all the edges connecting two vertices in S have the same color, whenever $n \geq R(k)$. The objective is to give a lower bound for the Ramsey number $R(k)$.

1. Assume that the edges of the graph are independently equally likely to be of each of the two colors. Letting $S \subset V$ with k vertices, compute the probability of the event B_S that S is monochromatic.
2. Deduce that there is a coloring for which no subset of size k is monochromatic whenever the following condition holds:

$$\binom{n}{k} 2^{1-\binom{k}{2}} < 1.$$

3. Letting $\lfloor \cdot \rfloor$ refer to the integer part, conclude that

$$R(k) > \lfloor 2^{k/2} \rfloor \quad \text{for all} \quad k \geq 3.$$

Inclusion–exclusion identity

Exercise 1.22. Let $A_1, A_2, \ldots, A_n \subset \Omega$ be events.

1. For all $\omega \in \Omega$, compute

$$\phi(\omega) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \cdots < i_k} \mathbf{1}\{\omega \in A_{i_1} \cap \cdots \cap A_{i_k}\}.$$

2. Deduce the inclusion–exclusion identity (1.8).

Exercise 1.23. Compute the probability that a bridge hand (13 cards chosen at random from an ordinary deck of 52 cards) is void in at least one suit.

Exercise 1.24. There are n married couples arranged at random in a row.

1. Find the probability that no husband sits next to his wife.
2. Compute this probability explicitly when $n = 3$.

Exercise 1.25 (Matching problem). Assume that n married couples are at a party and that the husbands and wives are randomly paired for a dance. We say that a match occurs each time a husband is paired with his wife and we let X denote the total number of matches.

1. Find the probability of the event A that there is no match. What is the limit of this probability when the number of couples goes to infinity?
2. Deduce the probability mass function of the number X of matches.

Exercise 1.26. There are n players playing $m \geq n$ games where each game is independently won by each of the players with the same probability $1/n$.

1. Assuming that each player scores one point for each game she wins, find the total number of possible scorings at the end of the m games.
2. Use the inclusion–exclusion identity to compute the probability that each player gets at least one point at the end of the m games.
3. Compute the expected value of the number of players with no point.

Exercise 1.27. Let $A \subset B$ be two finite sets and define

$$\phi(x) \sim \text{Uniform}\,(B) \text{ be independent for all } x \in A.$$

Let $\text{card}(A) = m$ and $\text{card}(B) = n$.

1. Find the probability that the function ϕ is injective when $m \leq n$.
2. Letting $B = \{y_1, y_2, \ldots, y_n\}$ and

$$A_i = \{\phi(x) \neq y_i \text{ for all } x \in A\} \quad \text{for} \quad i = 1, 2, \ldots, n,$$

use the inclusion–exclusion identity to compute the probability that the function ϕ is surjective when $m \geq n$.

Additional probability problems

Exercise 1.28. Two contestants A and B play a series of games until one has won a total of n games. Assuming that contestant A wins each game independently with probability p, find the probability that a total of x games are played.

Exercise 1.29. Take n points $X_1, X_2, \ldots, X_n$ uniformly at random on a circle with radius one. Find the probability of the following two events:

1. the chord connecting X_1 and X_2 has length at least one,
2. all n points are in the same semicircle.

Hint: For the second question, fix an orientation of the circle and use the event A_i that all the points are in the semi-circle starting at point X_i.

Exercise 1.30. You have an appointment at noon with a disorganized friend who will come at the appointment late. At which time should you go to the appointment in order to minimize the expected amount of time the first one at the appointment will have to wait in each of the following two contexts?

1. The time at which your friend will be at the appointment is uniformly distributed between noon and 1:00pm.
2. Your friend will be late by an amount of time that is exponentially distributed with mean 30 minutes.

Exercise 1.31. Assuming that, tossing a coin $n+m$ times, we get n heads and m tails and that all permutations of the heads and tails are equally likely, compute the probability of r runs, i.e., r sequences of consecutive heads.

Exercise 1.32. Let c_n and p_n be, respectively, the number of ways and the probability of tossing a fair coin n times such that there is no successive heads.

1. Prove that $c_n = c_{n-1} + c_{n-2}$ for all $n > 2$.
2. Deduce the probability p_n.

Hint: For the first question, decompose each sequence of n coin flips based on whether they start with heads or they start with tails.

Exercise 1.33. Let $X, Y \sim \text{Uniform}\{1, 2, \ldots, n\}$ be independent and

$$Q_d = \lim_{n \to \infty} Q_d(n) \quad \text{where} \quad Q_d(n) = P(\gcd(X,Y) = d).$$

Prove that $Q_d = (1/d)^2 Q_1$ for all $d \in \mathbb{N}^*$. Deduce that

$$Q_1 = \lim_{n \to \infty} Q_d(1) = 6/\pi^2 \approx 0.608.$$

Chapter 2
Distribution and conditional expectation

Following the same spirit as in the previous chapter, we continue to explore the implications of applying the results from measure theory to the field of probability theory. This chapter introduces in particular concepts that will be key later to studying various stochastic processes.

In the first section, we state the Radon–Nikodým theorem [78] without proof, and again refer the reader to Rudin [88] for the details. This result basically says that, under certain assumptions and having two positive measures, there is a unique measurable function ϕ such that the measure of a set for the first measure can be expressed using the function ϕ and the second measure. The function ϕ is called to the Radon–Nikodým derivative of the first with respect to the second measure. A nice consequence of the theorem is that positive measures can be expressed using a measurable function and a well-known reference measure such as the Lebesgue measure. This result is important in probability theory because it is the main ingredient to define more rigorously key probability concepts seen at the undergraduate level. The theorem is used in particular in this chapter to define the distribution of a random variable and the concept of conditional expectation.

Distribution — In the second section, we define the distribution of a random variable and prove the change of variable formula which shows how quantities related to the random variable can be computed using the distribution rather than the more mysterious underlying probability measure. The random variable is said to be continuous or discrete depending on whether its distribution satisfies a property called absolute continuity with respect to the Lebesgue or the counting measure. In particular, the Radon–Nikodým theorem implies that the density function of a continuous random variable and the probability mass function of a discrete random variable can both be seen as the Radon–Nikodým derivative of their distribution with respect to a reference measure. This not only unifies discrete and continuous random variables under the same umbrella, but also, together with the change of variable formula, this gives a powerful computational tool to study random variables.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_2

Conditional expectation — In the third section, we introduce the important concept of conditional expectation which will be important later for defining and studying martingales and Markov chains. Roughly speaking, the conditional expectation of a random variable with respect to a σ-algebra is the best guess we can make about the random variable given the information contained in the σ-algebra. The existence and uniqueness of the conditional expectation follows from the Radon–Nikodým theorem. We show how important formulas seen at the undergraduate level can be derived from this more abstract concept of conditional expectation, and also how to compute the conditional expectation in practice by breaking down the random variable under consideration into pieces that are either measurable with respect to or independent of the conditioning σ-algebra.

2.1 Radon–Nikodým theorem

This theorem is a general result in measure theory that has interesting applications in probability theory discussed in the next sections. To motivate the theorem, note that, given a positive measurable function ϕ on $(\Omega, \mathscr{F}, \mu)$,

$$\nu(A) = \int_A \phi \, d\mu = \int \phi \, \mathbf{1}_A \, d\mu \quad \text{for all} \quad A \in \mathscr{F}$$

defines a new measure ν on $(\Omega, \mathscr{F})$. Indeed, for each sequence (A_n) of mutually exclusive measurable sets, we have

$$\nu\left(\bigcup_{i=1}^{\infty} A_j\right) = \int \lim_{n \to \infty} \sum_{i=1}^{n} (\phi \, \mathbf{1}_{A_i}) \, d\mu = \lim_{n \to \infty} \sum_{i=1}^{n} \int \phi \, \mathbf{1}_{A_i} \, d\mu = \sum_{i=1}^{\infty} \nu(A_i)$$

according to the monotone convergence theorem. The Radon–Nikodým theorem is in some sense the converse of the previous statement as it gives the existence and uniqueness of ϕ under certain conditions on the measure ν.

Definition 2.1 (Absolute continuity). The measure ν is said to be absolutely continuous with respect to μ, which we write $\nu \ll \mu$, whenever

$$\text{for all } A \in \mathscr{F}, \ \ \mu(A) = 0 \ \text{ implies that } \ \nu(A) = 0. \tag{2.1}$$

This definition is motivated by the fact that

$$\mu(A) = 0 \quad \text{implies that} \quad \nu(A) = \int_A \phi \, d\mu = 0.$$

In particular, given two positive measures μ and ν, the absolute continuity of ν with respect to μ is a necessary condition for the existence of the function ϕ above. The Radon–Nikodým theorem states that this condition is also sufficient.

Theorem 2.1 (Radon–Nikodým). *Let μ and ν be two σ-finite measures such that ν is absolutely continuous with respect to μ. Then,*

- *There is $\phi : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ measurable positive such that*

  $$\nu(A) = \int_A \phi \, d\mu \quad \textit{for all} \quad A \in \mathscr{F}.$$

  The function ϕ is written $\phi = d\nu/d\mu$ and is called the **Radon–Nikodým derivative** *of the measure ν with respect to the measure μ.*
- *It is unique in the sense that two such derivatives are equal μ-almost everywhere: the set where both functions differ has measure zero for μ.*

This version of the theorem was proved by Nikodým in [78] and extends an earlier result due to Radon who proved the theorem in the special case where the underlying space is $\mathbb{R}^n$ rather than a general measure space. To understand the assumption of the theorem, assume for a moment that the measures ν and μ are simply nonnegative functions defined on the real line. Then, there exists a function ϕ that satisfies $\nu = \phi \, \mu$ if and only if

$$\mu(x) = 0 \quad \text{implies} \quad \nu(x) = 0 \quad \text{for all} \quad x \in \mathbb{R}.$$

This last condition can be viewed as the analog of the absolute continuity for positive measures (2.1). For a proof of the theorem, we refer to [88, Chapter 6]. In the next two sections, we use this theorem to redefine more rigorously various key concepts of probability theory.

2.2 Induced measure and distribution

Having a real random variable X on a probability space, one can define a probability measure ν_X on the Borel σ-algebra by setting

$$\nu_X(B) = P(X \in B) = \int_\Omega \mathbf{1}_{X^{-1}(B)} \, dP \quad \text{for all} \quad B \in \mathscr{B}.$$

It is called the **measure induced** by X in measure theory and **distribution** of X in probability theory. To study a random variable in practice, probabilists do not work with the probability measure P but with its distribution by using the following result called **change of variables formula**.

Theorem 2.2. *Let $X : (\Omega, \mathscr{F}, P) \to (\mathbb{R}, \mathscr{B})$ and $h : \mathbb{R} \to \mathbb{R}$ be measurable. Then, whenever h is positive or integrable,*

$$E(h(X)) = \int_{\mathbb{R}} h \, d\nu_X = \int_{\mathbb{R}} h(x) \, \nu_X(dx). \tag{2.2}$$

Proof. The steps of the proof follow the construction of the integral.

Step 1 — Assume first that $h = \mathbf{1}_B$ for some Borel set B. Then,

$$E(\mathbf{1}_B(X)) = E(\mathbf{1}\{X \in B\}) = P(X \in B) = \nu_X(B) = \int_{\mathbb{R}} \mathbf{1}_B \, d\nu_X.$$

Step 2 — In case $h = a_1 \mathbf{1}_{B_1} + \cdots + a_n \mathbf{1}_{B_n}$ is a simple measurable function, we use the previous step and the linearity of the integral to obtain

$$E(h(X)) = \sum_{i=1}^{n} a_i E(\mathbf{1}_{B_i}(X)) = \sum_{i=1}^{n} a_j \int_{\mathbb{R}} \mathbf{1}_{B_i} \, d\nu_X = \int_{\mathbb{R}} h \, d\nu_X.$$

Step 3 — Assume that h is positive. Recall from (1.6) that

$$s_n(x) = \min(n, 2^{-n} \lfloor 2^n h(x) \rfloor) \quad \text{for all} \quad x \in \mathbb{R}$$

defines a nondecreasing sequence of simple measurable functions with pointwise limit h. In particular, by the monotone convergence theorem,

$$E(h(X)) = \lim_{n \to \infty} E(s_n(X)) = \lim_{n \to \infty} \int_{\mathbb{R}} s_n \, d\nu_X = \int_{\mathbb{R}} h \, d\nu_X.$$

Step 4 — Finally, when h is integrable, we write $h = h^+ - h^-$. Since h^+ and h^- are both positive (so the previous step applies) and integrable,

$$E(h(X)) = E(h^+(X)) - E(h^-(X)) = \int_{\mathbb{R}} h^+ \, d\nu_X - \int_{\mathbb{R}} h^- \, d\nu_X = \int_{\mathbb{R}} h \, d\nu_X.$$

This completes the proof. □

Note that the probability that $X \in A$ and the expected value of X can be computed by taking $h = \mathbf{1}_A$ and $h = \text{id}$ respectively, i.e.,

$$P(X \in A) = \int_{\mathbb{R}} \mathbf{1}_A \, d\nu_X = \int_A d\nu_X \quad \text{and} \quad E(X) = \int_{\mathbb{R}} x \, d\nu_X.$$

In practice, it is convenient to express the distribution ν_X as a measurable function times a standard measure such as the Lebesgue measure. This idea is again related to the Radon–Nikodým theorem, which we now use to show the connection between the theory we have presented so far and undergraduate probability classes.

Probability density function. A random variable X is said to be **continuous** whenever its distribution ν_X is absolutely continuous with respect to the Lebesgue measure. Since the Lebesgue measure is σ-finite, the Radon–Nikodým theorem applies and gives the existence of a measurable function ϕ_X such that $d\nu_X = \phi_X \, d\lambda$. The derivative ϕ_X is then called the probability density function of the random variable and the change of variable formula (2.2) becomes

$$E(h(X)) = \int h \, \phi_X \, d\lambda = \int_{\mathbb{R}} h(x) \, \phi_X(x) \, dx. \tag{2.3}$$

Probability mass function. A random variable X is said to be **discrete** whenever its range S is either finite or countable, in which case its distribution ν_X is absolutely continuous with respect to the counting measure on the set S, that is,

$$\nu_X \ll \mu_S \quad \text{where} \quad \mu_S(A) = \operatorname{card}(A) \ \text{ for all } \ A \subset S.$$

Using once more the Radon–Nikodým theorem gives the existence of a measurable function ϕ_X such that $d\nu_X = \phi_X \, d\mu_S$. In this case, the derivative ϕ_X is called the probability mass function and the change of variable formula (2.2) becomes

$$E(h(X)) = \int h \, \phi_X \, d\mu_S = \sum_{x \in S} h(x) \, \phi_X(x). \tag{2.4}$$

In conclusion, random variables are characterized by their distributions which, in turn, are characterized by their Radon–Nikodým derivative ϕ_X with respect to some standard measures, either the Lebesgue measure or the counting measure. To this extent, measure theory and the abstract integral unify both discrete and continuous random variables by interpreting both probability density and probability mass functions as Radon–Nikodým derivatives. In practice, we deal with the integral with respect to the Lebesgue measure (2.3) or with respect to the counting measure (2.4) instead of the somewhat mysterious probability measure P. For a list of the most common distributions along with their interpretation and probability mass/density functions, we refer to Figure 2.1.

2.3 Conditional expectation

To study stochastic processes later, the next step is to define conditional expectation since this is a key concept to express certain dependency relationships among random variables and define martingales and Markov chains. This concept is introduced in the following definition. The fact that the conditional expectation exists and is unique follows from the Radon–Nikodým theorem.

Definition 2.2. Let $X \in L^1(\Omega, \mathscr{F}, P)$ and let $\mathscr{G} \subset \mathscr{F}$ be a σ-algebra.

- The **conditional expectation of** X **given** $\mathscr{G}$ is any

  $$Z \in \mathscr{M}(\Omega, \mathscr{G}) \quad \text{such that} \quad E(X \mathbf{1}_A) = E(Z \mathbf{1}_A) \quad \text{for all} \quad A \in \mathscr{G}.$$

  The variable Z is called a **version** of $E(X \,|\, \mathscr{G})$.
- Having a second random variable Y, we let $E(X \,|\, Y) = E(X \,|\, \sigma(Y))$.
- Also, we define the **conditional probability** as

  $$P(X \in B \,|\, \mathscr{G}) = E(\mathbf{1}\{X \in B\} \,|\, \mathscr{G}).$$

Name/parameters	Interpretation/origin	Probability mass/density function	Mean	Variance
Uniform (S), S finite	Equally likely outcomes.	$\phi(x) = \dfrac{1}{\mathrm{card}(S)}$ for all $x \in S$		
Uniform (S), $0 < \lambda(S) < \infty$	Equally likely outcomes.	$\phi(x) = \dfrac{1}{\lambda(S)}$ for all $x \in S$		
Bernoulli (p)	Coin flip - success/failure.	$\phi(1) = 1 - \phi(0) = p$	p	$p(1-p)$
Binomial (n,p)	Number of successes in a sequence of n independent Bernoulli trials. Discrete-time analog of Poisson.	$\phi(x) = \dbinom{n}{x} p^x (1-p)^{n-x}$ for all $x \in \{0,1,\dots,n\}$	np	$np(1-p)$
Poisson (μ)	Number of Poisson events in a time interval of length one. Continuous-time analog of binomial.	$\phi(x) = \dfrac{\mu^x}{x!} e^{-\mu}$ for all $x \in \mathbb{N}$	μ	μ
Geometric (p)	Time to the first/next success in a sequence of independent Bernoulli trials. Discrete-time analog of exponential.	$\phi(x) = (1-p)^{x-1} p$ for all $x \in \mathbb{N}^*$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Exponential (μ)	Time to the first/next Poisson event. Continuous-time analog of the geometric random variable.	$\phi(x) = \mu e^{-\mu x}$ for all $x \in \mathbb{R}_+$	$\dfrac{1}{\mu}$	$\dfrac{1}{\mu^2}$
Negative Binomial (n,p)	Time to the nth success in a sequence of independent Bernoulli trials. Discrete-time analog of gamma.	$\phi(x) = \dbinom{x-1}{n-1} p^n (1-p)^{x-n}$ for all $x \in \{n, n+1, \dots\}$	$\dfrac{n}{p}$	$\dfrac{n(1-p)}{p^2}$
Gamma (n,μ)	Time to the nth Poisson event. Continuous-time analog of the negative binomial.	$\phi(x) = \dfrac{\mu e^{-\mu x} (\mu x)^{n-1}}{n!}$ for all $x \in \mathbb{R}_+$	$\dfrac{n}{\mu}$	$\dfrac{n}{\mu^2}$
Normal (μ,σ^2)	Universal limit in the central limit theorem.	$\phi(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$ for all $x \in \mathbb{R}$	μ	σ^2

Fig. 2.1 Most common distributions.

Theorem 2.3. *The conditional expectation exists and is unique in the sense that two versions of the conditional expectation are equal P-almost surely.*

Proof. Assuming first that X is positive, since

$$\nu(A) = E(X\,\mathbf{1}_A) = \int_A X\,dP \quad \text{for all} \quad A \in \mathscr{G}$$

defines a finite measure $\nu \ll P$ on the space $(\Omega, \mathscr{G})$, there exists Z that satisfies the statement of the theorem, namely the Radon–Nikodým derivative $d\nu/dP$, and two such random variables are equal P-almost surely. In the general case where the random variable X is integrable, the first part of the proof applies to its positive part and its negative part. In particular, there exist two random variables Z_+ and Z_- measurable with respect to $\mathscr{G}$ such that

$$\begin{aligned} E(X\,\mathbf{1}_A) &= E(X^+\,\mathbf{1}_A) - E(X^-\,\mathbf{1}_A) \\ &= E(Z_+\,\mathbf{1}_A) - E(Z_-\,\mathbf{1}_A) = E((Z_+ - Z_-)\,\mathbf{1}_A) \end{aligned}$$

for all $A \in \mathscr{G}$. The uniqueness P-almost surely again follows from the uniqueness of the Radon–Nikodým derivative. □

The conditional expectation has several interesting properties. For instance, it can be proved that the conditional expectation inherits the following properties from the unconditional expectation:

- The function $X \mapsto E(X \,|\, \mathscr{G})$ is linear and nondecreasing.
- Jensen's inequality:

$$\phi(E(X \,|\, \mathscr{G})) \leq E(\phi(X) \,|\, \mathscr{G}) \quad \text{for all convex functions } \phi.$$

- Monotone convergence:

$$\lim_{n \to \infty} E(X_n \,|\, \mathscr{G}) = E(X \,|\, \mathscr{G}) \quad \text{whenever} \quad X_n \uparrow X.$$

- Dominated convergence:

$$\lim_{n \to \infty} E(X_n \,|\, \mathscr{G}) = E(X \,|\, \mathscr{G}) \quad \text{whenever} \quad X_n \to X \;\text{ and }\; |X_n| \leq Y \in L^1.$$

The conditional expectation will be used later in two contexts.

1. From the concept of conditional expectation, we can recover important formulas seen in undergraduate probability classes. These formulas show how to compute the probability of an event or the expected value of a random variable by conditioning on a partition of the sample space or on another random variable.
2. We also show how to compute the conditional expectation of a random variable in practice by breaking down this variable into pieces that are measurable with respect to the σ-algebra and pieces that are independent of the σ-algebra. These results will be our main tools to study martingales.

We now explore these two aspects.

Computing by conditioning. The most trivial but also one of the most useful properties, obtained by taking $A = \Omega$ in Definition 2.2, states that

$$E(E(X\,|\,Y)) = E(X) \quad \text{for all} \quad X \in L^1(\Omega, \mathscr{F}, P). \tag{2.5}$$

This equation is most useful looking at it backward, namely, it is used in practice to compute the expected value of a random variable, i.e., the right-hand side of (2.5), by conditioning on another random variable, i.e., the left-hand side. To deduce how to compute probability and expected value by conditioning, recall that the conditional probability of an event and the conditional expectation of a discrete random variable given an event are defined as

$$P(A\,|\,B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad E(X\,|\,B) = \frac{E(X\,\mathbf{1}_B)}{P(B)}. \tag{2.6}$$

Now, assume that we have a σ-algebra $\mathscr{G}$ generated by a partition (B_n) of events that all have a strictly positive probability. The $\mathscr{G}$-measurability of the conditional expectation implies that

$$Z = E(X\,|\,\mathscr{G}) = a_1\,\mathbf{1}_{B_1} + \cdots + a_n\,\mathbf{1}_{B_n} + \cdots \quad \text{for some} \quad a_n \in \mathbb{R},$$

but since $B_n \in \mathscr{G}$, by definition of the conditional expectation,

$$E(X\,\mathbf{1}_{B_n}) = E(Z\,\mathbf{1}_{B_n}) = a_n\,P(B_n).$$

In particular, we obtain the following expression:

$$E(X\,|\,\mathscr{G}) = \sum_{n=1}^{\infty} \frac{E(X\,\mathbf{1}_{B_n})}{P(B_n)}\,\mathbf{1}_{B_n}. \tag{2.7}$$

Recall from our convention that, even if it is not specified, this equation only holds almost surely. Setting $X = \mathbf{1}_A$ in (2.7), taking the expected value, using (2.5) and recalling the definition of conditional probability from (2.6), we deduce

$$P(A) = E(E(\mathbf{1}_A\,|\,\mathscr{G})) = \sum_{n=1}^{\infty} \frac{E(\mathbf{1}_A\,\mathbf{1}_{B_n})}{P(B_n)}\,E(\mathbf{1}_{B_n}) = \sum_{n=1}^{\infty} P(A\,|\,B_n)\,P(B_n). \tag{2.8}$$

Now, let Y be a discrete random variable. Since the σ-algebra $\sigma(Y)$ is generated by a countable partition, we deduce from (2.7) that

$$E(X\,|\,Y) = \sum_y \frac{E(X\,\mathbf{1}\{Y = y\})}{P(Y = y)}\,\mathbf{1}\{Y = y\}$$

where the sum is over the range of Y. Taking the expected value, using (2.5) and recalling the definition of conditional expectation from (2.6), we get

$$E(X) = E(E(X\,|\,Y)) = \sum_y E(X\,|\,Y=y)\,P(Y=y). \tag{2.9}$$

Equations (2.8)–(2.9) are two important formulas seen at the undergraduate level. They are useful in practice to compute unconditional probability and expected value provided one is able to find natural partitions or random variables that make the conditional objects easier to compute than their unconditional counterparts. The exercises at the end of this chapter give a wide variety of examples of application of these equations. Note that these two formulas can be proved more easily by simply using the σ-additivity of the probability. The approach we used is to show how they can also be derived from the general definition of conditional expectation.

Computing conditional expectations. Finally, we give some tools that will be useful later to prove that a stochastic process is a martingale and/or a Markov chain. To better understand the real meaning of the conditional expectation hidden behind its somewhat mysterious definition, one can think of the conditional expectation as the best possible approximation of X given the information contained in the σ-algebra $\mathscr{G}$, namely, the best possible approximation by a random variable which is $\mathscr{G}$-measurable. The larger the σ-algebra, the better the approximation. With this in mind, it is clear intuitively and easy to prove that

$$\begin{array}{rlll} \textbf{perfect information} & E(X\,|\,\mathscr{G}) = X & \text{when} & \mathscr{G} = \mathscr{F} \\ \textbf{no information} & E(X\,|\,\mathscr{G}) = E(X) & \text{when} & \mathscr{G} = \{\varnothing, \Omega\}. \end{array} \tag{2.10}$$

Following along these lines, we also prove that, when $\mathscr{H} \subset \mathscr{G}$,

$$E(E(X\,|\,\mathscr{G})\,|\,\mathscr{H}) = E(E(X\,|\,\mathscr{H})\,|\,\mathscr{G}) = E(X\,|\,\mathscr{H})$$

indicating that, after the double conditioning, the only information available comes from the smaller σ-algebra. It is convenient to call this property the **projection rule** since one can think of the conditional expectation as the projection of a random variable onto a subset of measurable functions, so the property above simply says that projecting twice has the same result as projecting onto the smaller subset.

To compute the conditional expectation of a random variable in practice, the trick is to break it down into pieces for which the information is perfect and pieces for which there is no information and then use (2.10). To make this precise, we prove a couple of lemmas showing basically that the pieces that are perfectly known under the conditioning can be moved outside the conditional expectation.

Lemma 2.1. *Let $X, Y \in L^1(\Omega, \mathscr{F}, P)$. Then,*

$$E(X+Y\,|\,\mathscr{G}) = X + E(Y\,|\,\mathscr{G}) \quad \textit{when } X \textit{ is } \mathscr{G}\textit{-measurable}.$$

Proof. This follows from (2.10) and the linearity of the conditional expectation. □

Lemma 2.2. *Let $Y, XY \in L^1(\Omega, \mathscr{F}, P)$. Then,*

$$E(XY\,|\,\mathscr{G}) = X\,E(Y\,|\,\mathscr{G}) \quad \textit{when } X \textit{ is } \mathscr{G}\textit{-measurable}.$$

Proof. Let $A, B \in \mathscr{G}$ and set $X = \mathbf{1}_A$. Since $A \cap B \in \mathscr{G}$,

$$E(\mathbf{1}_B E(Y \mid \mathscr{G}) \mathbf{1}_A) = E(E(Y \mid \mathscr{G}) \mathbf{1}_{A \cap B}) = E(Y \mathbf{1}_{A \cap B}) = E(\mathbf{1}_B Y \mathbf{1}_A)$$

which shows the result when X is an indicator function. We conclude following the same steps and using the same ingredients as in the proof of Theorem 2.2. More precisely, we extend the result to simple random variables using that the conditional expectation is linear, then to positive random variables using the monotone convergence theorem, and finally to integrable random variables by looking at negative and positive parts separately. □

To study branching processes later, we will also need the following result.

Lemma 2.3. *Let $(X_n) \subset L^1(\Omega, \mathscr{F}, P)$ be a sequence of identically distributed random variables and let T be an integer-valued random variable. Then,*

$$E(X_1 + X_2 + \cdots + X_T \mid T) = T\, E(X_1 \mid T).$$

Proof. Letting $A_n = \{T = n\}$ for all $n \in \mathbb{N}$, we have

$$\begin{aligned} E((X_1 + \cdots + X_T) \mathbf{1}_{A_n}) &= E((X_1 + \cdots + X_n) \mathbf{1}_{A_n}) \\ &= E(nX_1 \mathbf{1}_{A_n}) = E(n E(X_1 \mid T) \mathbf{1}_{A_n}) = E(T E(X_1 \mid T) \mathbf{1}_{A_n}). \end{aligned}$$

Since $\sigma(T)$ is generated by the partition (A_n), the previous set of equations remains true replacing A_n by any $A \in \sigma(T)$, which proves the result. □

The previous three lemmas show how to deal in practice with the pieces that are measurable with respect to the conditioning. We now look at the other extreme: the pieces of the random variable for which we have no information. In the second part of (2.10), we have no information about the random variable because the σ-algebra does not give any information. More generally, we have no information about the random variable whenever it is independent of the σ-algebra, so we define independence and show that the second statement in (2.10) extends to this case. The reader should be already familiar with the independence of events. In contrast, the next definition is about independence of σ-algebras and random variables.

Definition 2.3. Two σ-algebras $\mathscr{G}, \mathscr{H} \subset \mathscr{F}$ are said to be **independent** if each event in $\mathscr{G}$ and each event in $\mathscr{H}$ are pairwise independent, i.e.,

$$P(A \cap B) = P(A) P(B) \quad \text{for all} \quad A \in \mathscr{F} \text{ and } B \in \mathscr{G}.$$

Two random variables X and Y are independent when

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B) \quad \text{for all} \quad A, B \in \mathscr{B}(\mathbb{R})$$

meaning that $\sigma(X)$ and $\sigma(Y)$ are independent.

The definition will become clear later when dealing with concrete examples of random variables independent from a σ-algebra in the context of stochastic processes.

Before we state our next result about conditional expectation, recall that, having a larger collection, finite or countable, of events (A_n), these events are said to be independent whenever

$$P\Big(\bigcap_{i\in I} A_i\Big) = \prod_{i\in I} P(A_i) \quad \text{for all} \quad I \subset \mathbb{N}^* \text{ finite.}$$

Recall also that a collection of events may be pairwise independent but not independent, showing that independence is a somewhat subtle concept. To give a simple counterexample, flip a fair coin twice and let X_1 and X_2 be respectively the outcome of the first and second flip. Both flips are assumed to result in independent outcomes. Then, define the events

$$A_1 = \{X_1 = H\} \qquad A_2 = \{X_2 = H\} \qquad A_3 = \{X_1 \neq X_2\}.$$

These events are pairwise independent. Indeed, they all have probability one-half, and the intersection of any two has probability one-fourth since

$$A_1 \cap A_2 = \{HH\} \qquad A_1 \cap A_3 = \{HT\} \qquad A_2 \cap A_3 = \{TH\}.$$

However, they are not independent because

$$P(A_1 \cap A_2 \cap A_3) = P(\varnothing) = 0 \neq (1/2)^3 = P(A_1)P(A_2)P(A_3).$$

Returning to conditional expectation, we now prove that, when the random variable and the conditioning σ-algebra are independent, suggesting that the σ-algebra does not provide any information about the random variable, our best guess is again the unconditional expectation, i.e., we can remove the conditioning. This is proved in the following lemma.

Lemma 2.4. *Let X and $\mathscr{G}$ be independent. Then, $E(X\,|\,\mathscr{G}) = E(X)$.*

Proof. The joint distribution of two independent random variables is equal to the product of their distributions. Therefore, if X and Y are independent, by applying Fubini's theorem to the function $h(x,y) = xy$, we obtain

$$E(XY) = \int h\,d\nu_{X,Y} = \int x\,\nu_X(dx) \int y\,\nu_Y(dy) = E(X)E(Y).$$

In particular, since X and $Y = \mathbf{1}_A$ are independent for all $A \in \mathscr{G}$,

$$E(X\,\mathbf{1}_A) = E(XY) = E(X)E(Y) = E(X)E(\mathbf{1}_A) = E(E(X)\,\mathbf{1}_A).$$

Moreover, since $Z = E(X)$ is constant, it is $\mathscr{G}$-measurable. □

Lemmas 2.1–2.4 are the main tools to compute the conditional expectation. As previously explained, in order to be able to apply these lemmas in practice, the first step is to break down the random variable into pieces that are measurable with respect to the σ-algebra and pieces that are independent of the σ-algebra.

2.4 Exercises

Probability and conditional probability

Exercise 2.1. Let A and B be two events.

1. Prove that the occurrence of A makes the occurrence of B more likely if and only if the occurrence of B makes the occurrence of A more likely.
2. Prove that A and B are independent if and only if A and B^c are independent.

Exercise 2.2. Suppose that two balls numbered 1 and 2 are independently black or white with the same probability of one-half.

1. Given that ball 2 is black, find the probability that the other ball is black.
2. Given that at least one of the two balls is black, find the probability that the other ball is also black.

Exercise 2.3. An urn contains three white and three black balls. A fair die is rolled and that number of balls is randomly chosen from the urn. Find the probability that all the selected balls are white.

Exercise 2.4. An urn contains a white and b black balls. Take one ball at random, paint it black in case it is white, and put the ball back in the urn. Then, take again a ball at random. By what factor does the probability that the second ball is white decreases in comparison with the probability that the first ball is white?

Exercise 2.5. An urn contains nine white balls and six black balls. Take three balls at random, paint the white ones black, and put the balls back in the urn. Then, take again three balls at random. Find the probability that these last three balls are all white.

Exercise 2.6. Assume that a parallel system with n components is operational if and only if at least one of its component is working. Compute the conditional probability that component 1 is working given that the system is operational under the condition that the components work independently with probability p.

Exercise 2.7. Consider $n \geq 2$ individuals labeled $1, 2, \ldots, n$. Individual 1 creates a rumor that she tells individual 2. Then, each individual independently tells the next individual either the information she learned with probability p or the opposite of the information she learned with probability $q = 1 - p$.

1. Condition on whether or not the information told by individual 2 is the one she learned to express p_n as a function of p_{n-1} where p_n be the probability that the information received by individual n is the rumor created by individual 1.
2. Deduce an explicit expression for p_n.
3. Compute p_n when $p = 0$, and in the limit as $n \to \infty$ when $p \in (0, 1)$.

Exercise 2.8 (Craps). At this game, the player throws two fair dice.

- If the sum is 7 or 11 then she wins.
- If the sum is 2, 3 or 12 then she loses.
- If the sum is anything else, she continues throwing the dice until the sum is either that number again, in which case she wins, or 7, in which case she loses.

Find the probability of winning at the game of craps.

Hint: Compute first the probability that the sum i appears before the sum 7 by conditioning on the value of the first roll.

Exercise 2.9. Players A and B play until they are two points apart. Find the probability that A is the first player to have two more points than the other player if each point is independently won by A with probability p.

Exercise 2.10. Let A and B be two tennis players. Assume that each point is scored independently by player A with probability $p = 3/5$. Compute the probability that A wins a game, where the winner of a game is the first player with four points and two more points than the other player.

Exercise 2.11. Two players A and B take turn playing a game until one of the two players wins. Independently at each step, player A wins with probability p while player B wins with probability q. Find the values of p and q for which the game is fair, i.e., each player is equally likely to be the first one to win.

Hint: Condition on whether A wins the first game or not.

Exercise 2.12. Three evenly matched players A, B and C play a series of games. The winner of each game plays the next game with the waiting player until a player wins two games in a row and is declared the overall winner. Find the probability of each of the players being the overall winner when A and B play the first game.

Exercise 2.13. A player plays alternatively with two opponents until she wins twice in a row. Assuming that she wins independently each game with probability p against opponent 1 and with probability $q > p$ against opponent 2, should she start playing with opponent 1 or with opponent 2 if her objective is to minimize the expected number of games she has to play?

Exercise 2.14. Consider a contest with 2^n evenly matched players. The players are randomly paired off against each other, then the 2^{n-1} winners are again paired off, and so on, until a single winner remains. Find the probability that two randomly chosen contestants play each other.

Exercise 2.15 (The ballot problem). In an election, candidates A and B receive respectively a and b votes with $a > b$. Let $p(a,b)$ be the probability that candidate A is always ahead of B when all the orderings of the votes are equally likely.

1. Express $p(a,b)$ as a function of $p(a-1,b)$ and $p(a,b-1)$ by conditioning on the candidate who receives the last vote.
2. Deduce that $p(a,b) = (a-b)/(a+b)$.

Exercise 2.16 (The best prize problem). Consider the game where n distinct prizes are presented one by one to a player. All $n!$ possible permutations are assumed to be equally likely. Each time a prize is revealed, the player learns about the relative rank of that prize compared to the ones already seen and, based on this information, must decide to either leave with that prize or wait for a hopefully better prize. A natural strategy is to reject the first m prizes and then accept the first one that is better than the first m prizes provided there is one.

1. Letting B be the event that the best prize is selected and P_m be the probability under the strategy described above, prove that
$$P_m(B) = \frac{m}{n}\left(\frac{1}{m} + \frac{1}{m+1} + \cdots + \frac{1}{n-1}\right).$$
2. Deduce that, as $n \to \infty$, the maximal probability of selecting the best prize under the strategy described above converges to $1/e \approx 0.368$.

Exercise 2.17. Let $m \leq n$ and let
$$U_1, \ldots, U_m \sim \text{Uniform}\{1,2,\ldots,n\}$$
be independent. Condition on the event that the random variables are all distinct to compute the probability of the event
$$A = \{U_1 < U_2 < \cdots < U_m\}.$$

Exercise 2.18. Compute $P(A \subset B)$ where A and B are chosen uniformly at random among the 2^n possible subsets of a set with n elements.

Hint: Condition on the cardinal of B.

Exercise 2.19 (Euler ϕ function). For all $n \in \mathbb{N}^*$, let $\phi(n)$ be the number of integers less than or equal to n which are prime to n. We want to show that
$$\phi(n) = n \prod_{p \in P_n}\left(1 - \frac{1}{p}\right) \quad \text{where} \quad P_n = \{p : p \text{ is prime and divides } n\}.$$

Let $X \sim \text{Uniform}\{1,2,\ldots,n\}$ and $A_p = \{p \text{ divides } X\}$ for each $p \leq n$.

1. Compute $P(A_p)$ when p divides n.
2. Let $p_1, \ldots, p_k$ be distinct prime divisors of n. Prove that
$$A_{p_1}, A_{p_2}, \ldots, A_{p_k} \quad \text{are independent.}$$
3. Use the previous two questions to conclude.

Expected value and conditional expectation

Exercise 2.20. Players $1,2,\dots,n$ take turns flipping a coin having probability p of turning up heads, with the successive flips being independent.

1. Thinking of the geometric random variable with success probability p as the first flip resulting in heads, compute its expected value by conditioning on the outcome of the first flip.
2. Use the same idea to find the probability mass function of the random variable X referring to the first player who gets heads.

Exercise 2.21. Let (X_n) be a sequence of independent and identically distributed discrete random variables with finite range, say $\{1,2,\dots,m\}$. Find the expected value of the total number T of random variables one needs to observe until the first outcome appears again.

Hint: Condition on the outcome of the first random variable.

Exercise 2.22 (Matching rounds problem). Referring to Exercise 1.25,

1. Find $E(X)$ and $\mathrm{Var}(X)$ where

$$X = \text{number of husbands paired with their wife.}$$

Now, assume that the couples for which a match occurs depart while the others are again randomly paired. This continues until there is no couple left.

2. Prove that there are in average n rounds in this process.

Exercise 2.23. Let (X_i) be a sequence of independent and identically distributed discrete random variables and fix a pattern $(x_1,x_2,\dots,x_n)$ such that,

$$(x_{n-k+1},x_{n-k+2},\dots,x_n) \neq (x_1,x_2,\dots,x_k) \quad \text{for all} \quad k < n.$$

Let T be the number of random variables until the pattern appears.

1. Prove that $T = i+n$ if and only if

$$T > i \quad \text{and} \quad (X_{i+1},X_{i+2},\dots,X_{i+n}) = (x_1,x_2,\dots,x_n).$$

2. Deduce that $E(T) = (p_{x_1}\, p_{x_2} \cdots p_{x_n})^{-1}$ where $p_x = P(X_i = x)$.

Exercise 2.24. Assume that an *a priori* biased coin whose probability of landing on heads is given by p is continually flipped.

1. Find the expected value of the time T_{HT} until the pattern HT appears.
2. Find the expected value of the time T_{HH} until the pattern HH appears.
3. More generally, find the expected value of the time T_n until H appears n times in a row by conditioning on T_{n-1}.

Exercise 2.25 (Compound random variable). Let (X_n) be a sequence of identically distributed random variables with

$$\mu = E(X_n) < \infty \quad \text{and} \quad \sigma^2 = \mathrm{Var}(X_n) < \infty$$

and let T be an independent nonnegative integer-valued random variable also with finite mean and finite variance.

1. Prove that $E(X_1 + X_2 + \cdots + X_T) = \mu\, E(T)$.
2. Assuming in addition that the X_n are independent, show that

$$\mathrm{Var}\,(X_1 + X_2 + \cdots + X_T) = \sigma^2\, E(T) + \mu^2\, \mathrm{Var}\,(T).$$

Hint: For both parts, condition on the random variable T.

Exercise 2.26. Let $p \in (0,1)$ and $X, Y \sim \mathrm{Bernoulli}\,(p)$ be independent.

1. Compute $E(X \,|\, Z)$ where $Z = \mathbf{1}\{X + Y = 0\}$.
2. Deduce that $E(X \,|\, Z)$ and $E(Y \,|\, Z)$ are not independent.

Chapter 3
Limit theorems

In this chapter, we move one more step toward the world of stochastic processes by considering sequences of random variables, the convergence of these sequences, and limit theorems in the special case when these random variables are independent and identically distributed.

In contrast with sequences of real numbers, there are different notions or levels of convergence for sequences of random variables. In the first section, we define convergence in distribution, convergence in probability, convergence almost surely and convergence in mean, and show how these different notions of convergence are related. We also give counterexamples showing that the global hierarchical picture we get about the relationships among these notions of convergence is in some sense optimal. Along the way, we prove Markov's inequality and the two Borel–Cantelli lemmas [9, 15] which are not only key tools to study the convergence of random variables but also very useful results in probability theory in general.

Starting from the second section, we specialize in the case of sequences (X_n) of independent and identically distributed random variables. More precisely, we are interested in the convergence of the sequence of empirical means

$$S_n/n \quad \text{where} \quad S_n = X_1 + X_2 + \cdots + X_n \quad \text{for all} \quad n \in \mathbb{N}^*.$$

Using Chebyshev's inequality [17], which follows from Markov's inequality, it can be proved that, when the random variables have finite mean and finite variance, the empirical mean converges in probability to its theoretical mean, a result known as the weak law of large numbers. In fact, the assumption on the finiteness of the variance can be relaxed and the convergence is almost sure:

Strong law of large numbers — Assuming that the random variables are independent and identically distributed with finite mean μ, the empirical mean converges almost surely to the theoretical mean μ.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_3

This chapter only gives a proof when the variance is finite, but we will show the result in its full generality later using reverse martingales. The next natural question following the law of large numbers is: How much does the partial sum S_n deviate from its mean? By the strong law of large numbers, the deviation cannot be of order n or larger. The exact answer, given by the central limit theorem, is that the deviation is of the order of $\sqrt{n}$. More precisely,

Central limit theorem — Divided by $\sqrt{n}$, the difference between S_n and its mean converges in distribution to a centered normal distribution.

Note that the limit is always normally distributed regardless of the common distribution of the random variables X_n, a quite remarkable result that suggests some kind of universality property of the normal distribution. Classical applications of the law of large numbers and central limit theorem in the context of random games, but also more exotic applications such as the Weierstrass theorem, are also given across this chapter.

The law of large numbers and central limit theorem are often viewed as the two most important results in probability theory. This is certainly true from a historical perspective since it took more than two centuries for the mathematics community to obtain rigorous proofs of the versions of the theorems given in this chapter. The law of large numbers was first proved in the 18th century by Daniel Bernoulli [4] in the special case of a binary random variable, i.e., independent coin flips. His result was later improved by Simeon Poisson [83] and Pafnuty Chebyshev [16]. The early version of the central limit theorem, again for a binary random variable, is due to Abraham de Moivre [72] and was improved later by Pierre-Simon Laplace [64]. A more recent contribution to the law of large numbers and to the central limit theorem was made by Paul Lévy [67]. The versions of the two theorems given in this chapter are the standard versions traditionally found in probability textbooks, but we point out that there are also more recent refinements of these two fundamental results.

Further reading

- Billingsley [7] is one of the main references on sequences of random variables and limit theorems.
- For general textbooks on probability theory and stochastic processes that also cover this topic, we refer the reader to [28, 34, 35, 90] where not only the strong law of large numbers and the central limit theorem but also more general limit theorems are given.
- For a treatment at the undergraduate level, see Ross [86].

3.1 Levels of convergence

We distinguish four different notions of convergence of the sequence of random variables (X_n) to a random variable X.

Definition 3.1. The sequence (X_n) is said to converge to X

- in distribution: $X_n \xrightarrow{d} X$ when

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{for all } x \text{ that are continuity points of } F_X$$

  where F_X is the distribution function: $F_X(x) = P(X \leq x)$,

- in probability: $X_n \xrightarrow{p} X$ when $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$ for all $\varepsilon > 0$,
- almost surely: $X_n \xrightarrow{a.s.} X$ when $P(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1$,
- in L^p for $1 \leq p < \infty$: $X_n \xrightarrow{L^p} X$ when $\lim_{n\to\infty} E|X_n - X|^p = 0$.

The next four lemmas show how these are related.

Lemma 3.1. *Convergence almost surely implies convergence in probability.*

Proof. Assume that $X_n \xrightarrow{a.s.} X$ and fix $\varepsilon > 0$. Since

$$A_n = \bigcup_{i=n}^{\infty} \{|X_i - X| \geq \varepsilon\} \downarrow A_\infty = \bigcap_{i=1}^{\infty} A_i \quad \text{as } n \to \infty,$$

by inclusion of events and monotone convergence, we get

$$\begin{aligned} \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) &\geq \lim_{n\to\infty} P(A_n) = P(A_\infty) \\ &= P(\omega : \text{for all } n \text{ there is } i \geq n \text{ such that } |X_i(\omega) - X(\omega)| \geq \varepsilon) = 0 \end{aligned}$$

which proves that $X_n \xrightarrow{p} X$. □

Lemma 3.2. *Convergence in L^p for some $1 < p < \infty$ implies convergence in L^1.*

Proof. Assume that $X_n \xrightarrow{L^p} X$ for some $1 < p < \infty$. Applying Jensen's inequality with the convex function $\phi(x) = x^p$, we deduce that

$$\lim_{n\to\infty} E|X_n - X| \leq \lim_{n\to\infty} (E|X_n - X|^p)^{1/p} = 0$$

therefore $X_n \xrightarrow{L^1} X$. □

Lemma 3.3. *Convergence in L^1 implies convergence in probability.*

Proof. For any $\varepsilon > 0$ and $Y \geq 0$ integrable,

$$\begin{aligned} E(Y) &= E(Y\mathbf{1}\{Y < \varepsilon\}) + E(Y\mathbf{1}\{Y \geq \varepsilon\}) \\ &\geq E(Y\mathbf{1}\{Y \geq \varepsilon\}) \geq E(\varepsilon\mathbf{1}\{Y \geq \varepsilon\}) = \varepsilon P(Y \geq \varepsilon). \end{aligned} \tag{3.1}$$

Taking $Y = |X_n - X|$ and assuming that $X_n \xrightarrow{L^1} X$, we deduce that

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) \leq \lim_{n\to\infty} \varepsilon^{-1} E|X_n - X| = 0$$

therefore $X_n \xrightarrow{p} X$. □

Inequality (3.1) is known as **Markov's inequality**.

Lemma 3.4. *Convergence in probability implies convergence in distribution.*

$$\begin{tikzcd}
& X_n \xrightarrow{a.s.} X \arrow[dr] & & \\
& X_{n_k} \xrightarrow{a.s.} X \text{ for a subsequence} & X_n \xrightarrow{p} X \arrow[l] \arrow[r] & X_n \xrightarrow{d} X \\
X_n \xrightarrow{L^p} X \text{ with } p > 1 \arrow[r] & X_n \xrightarrow{L^1} X \arrow[ur] & &
\end{tikzcd}$$

Fig. 3.1 Relationships among the different notions of convergence.

Proof. Let $X_n \xrightarrow{p} X$. First, we note that

$$\{Y \leq z\} \subset \{Z \leq z + \varepsilon\} \cup \{|Y - Z| > \varepsilon\} \quad \text{for all} \quad z \in \mathbb{R}.$$

Taking the probability on both sides gives

$$\begin{array}{ll} F_{X_n}(x) \leq F_X(x+\varepsilon) + P(|X_n - X| > \varepsilon) & \text{when} \quad (Y,Z,z) = (X_n, X, x) \\ F_X(x-\varepsilon) \leq F_{X_n}(x) + P(|X_n - X| > \varepsilon) & \text{when} \quad (Y,Z,z) = (X, X_n, x-\varepsilon) \end{array}$$

from which it follows that

$$\begin{aligned} F_X(x-\varepsilon) &= F_X(x-\varepsilon) - \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) \leq \lim_{n\to\infty} F_{X_n}(x) \\ &\leq F_X(x+\varepsilon) + \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = F_X(x+\varepsilon). \end{aligned}$$

In particular, when F_X is continuous at x,

$$\begin{aligned} F_X(x) &= \lim_{\varepsilon\downarrow 0} F_X(x-\varepsilon) \\ &\leq \lim_{n\to\infty} F_{X_n}(x) \leq \lim_{\varepsilon\downarrow 0} F_X(x-\varepsilon) = F_X(x) \end{aligned}$$

which proves that $X_n \xrightarrow{d} X$. □

Lemmas 3.1–3.4 as well as Lemma 3.6 that will be proved in a moment are summarized in the diagram of Figure 3.1 and we now give counterexamples to show that the picture is optimal in the sense that no other implications hold in general.

To begin with, let X be any discrete random variable taking at least two values, say $a \neq b$, with positive probability and let $X_1, X_2, \ldots$ be independent with the same distribution as X. Then (X_n) converges in distribution to X but

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 2P(X = a)P(X = b) \neq 0$$

for all $0 < \varepsilon < |a-b|$ so there is no convergence in probability: convergence in distribution does not imply convergence in probability.

To construct additional counterexamples, and also prove the strong law of large numbers in the next section, we need the two Borel–Cantelli lemmas that were proved by Emile Borel in [9] and Francesco Cantelli in [15]. Having a sequence of events (A_n), this result is concerned with $\limsup_{n\to\infty} A_n$, which is the event of infinitely many occurrences, i.e., A_n occurs for infinitely many integers n.

Lemma 3.5 (Borel–Cantelli Lemmas). *We have*

1. $\sum_n P(A_n) < \infty$ *implies that* $P(\limsup_{n\to\infty} A_n) = 0$,
2. whenever the events A_n *are independent,*

$$\textstyle\sum_n P(A_n) = \infty \quad \textit{implies that} \quad P(\limsup_{n\to\infty} A_n) = 1.$$

Proof. Let $N = \sum_n \mathbf{1}_{A_n}$ and apply Fubini's theorem to get

$$E(N) = E\left(\sum_{n=1}^{\infty} \mathbf{1}_{A_n}\right) = \sum_{n=1}^{\infty} E(\mathbf{1}_{A_n}) = \sum_{n=1}^{\infty} P(A_n) < \infty.$$

This shows that $P(\limsup_{n\to\infty} A_n) = P(N = \infty) = 0$. Now, assume that the events A_n are independent and that the sum of their probabilities is infinite. Then,

$$\begin{aligned} \lim_{m\to\infty} P(A_n^c \cap \cdots \cap A_{n+m}^c) &= \lim_{m\to\infty} (1 - P(A_n)) \cdots (1 - P(A_{n+m})) \\ &\leq \lim_{m\to\infty} \exp(-P(A_n)) \cdots \exp(-P(A_{n+m})) \\ &= \lim_{m\to\infty} \exp(-(P(A_n) + \cdots + P(A_{n+m}))) = 0. \end{aligned}$$

This implies that

$$P(\limsup_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n \cup \cdots \cup A_{n+m} \cup \cdots) = 1$$

and completes the proof. □

Now, let (X_n) be a sequence of independent random variables with

$$P(X_n = 0) = 1 - p_n \quad \text{and} \quad P(X_n = x_n) = p_n \tag{3.2}$$

where $p_n \to 0$ and $x_n \to \infty$ as $n \to \infty$. Then, for all $\varepsilon > 0$,

$$\begin{aligned} \lim_{n\to\infty} P(|X_n - 0| \geq \varepsilon) &= \lim_{n\to\infty} P(X_n \geq \varepsilon) \\ &\leq \lim_{n\to\infty} P(X_n = x_n) = p_n \end{aligned}$$

which proves convergence in probability to zero. Also,

$$\lim_{n\to\infty} E\,|X_n - 0| = \lim_{n\to\infty} E(X_n) = \lim_{n\to\infty} x_n\, p_n$$

showing that there is convergence in mean to zero if and only if $x_n p_n \to 0$. To study the convergence almost surely, we use the two Borel–Cantelli lemmas: since

$$\textstyle\sum_n P(X_n = x_n) = \sum_n p_n$$

and the random variables X_n are independent,

$$P(\limsup_{n\to\infty} \{X_n = x_n\}) = \mathbf{1}\{\textstyle\sum_n p_n = \infty\}$$

so there is almost sure convergence if and only if the sum of the p_n is finite. Specific choices for the sequences (p_n) and (x_n) provide counterexamples.

- Take $p_n = 1/n \to 0$ and $x_n = \sqrt{n} \to \infty$ so that

  $$\lim_{n\to\infty} x_n p_n = \lim_{n\to\infty} 1/\sqrt{n} = 0 \quad \text{and} \quad \textstyle\sum_n p_n = \infty.$$

  In this case, the sequence of random variables (3.2) converges to zero in probability and in mean but not almost surely so convergence in probability and convergence in mean do not imply convergence almost surely.
- Take $p_n = 1/n^2 \to 0$ and $x_n = n^3 \to \infty$ so that

  $$\lim_{n\to\infty} x_n p_n = \infty \quad \text{and} \quad \textstyle\sum_n p_n < \infty.$$

  In this case, the sequence of random variables (3.2) converges to zero in probability and almost surely but not in mean so convergence in probability and convergence almost surely do not imply convergence in mean.

Another consequence of the first Borel–Cantelli lemma that will be useful later to study uniformly integrable random variables is the following result.

Lemma 3.6. *Convergence in probability implies convergence almost surely of a subsequence to the same limit.*

Proof. We construct a subsequence (X_{n_k}) recursively by letting $n_0 = 1$ and

$$n_k = \min\{n > n_{k-1} : P(|X_n - X| > 2^{-k}) \leq 2^{-k}\} \quad \text{for all} \quad k > 0.$$

For this subsequence, we have

$$\textstyle\sum_k P(|X_{n_k} - X| > 2^{-k}) \leq \sum_k (1/2)^k < \infty$$

therefore, by the first Borel–Cantelli lemma,

$$P(|X_{n_k} - X| > 2^{-k} \text{ infinitely often}) = 0,$$

showing that $X_{n_k} \xrightarrow{a.s.} X$. □

3.2 Strong law of large numbers

From now on, we assume that the random variables X_n are independent and identically distributed with finite mean and variance

$$\begin{aligned} \mu &= E(X_n) < \infty \\ \sigma^2 &= \operatorname{Var}(X_n) = E((X_n - E(X_n))^2) < \infty. \end{aligned} \tag{3.3}$$

We are interested in the partial sum $S_n = X_1 + \cdots + X_n$. The first step is to prove Chebyshev's inequality, which was first proved in [17].

Lemma 3.7 (Chebyshev's inequality). *For any random variable X with finite mean μ and finite variance σ^2, we have*

$$P(|X - \mu| \geq \varepsilon) \leq (\sigma/\varepsilon)^2 \quad \textit{for all} \quad \varepsilon > 0. \tag{3.4}$$

Proof. Taking $Y = (X - \mu)^2$ in Markov's inequality (3.1), we obtain

$$\begin{aligned} P(|X - \mu| \geq \varepsilon) &= P((X - \mu)^2 \geq \varepsilon^2) \\ &\leq \varepsilon^{-2} E((X - \mu)^2) = (\sigma/\varepsilon)^2. \end{aligned}$$

This completes the proof. □

Chebyshev's inequality is the key to proving a weak version of the law of large numbers, but before going into the proof we give a beautiful example of the application of this inequality in analysis.

Example 3.1 (Weierstrass theorem). This example gives a probabilistic proof of Weierstrass theorem which relies in part on Chebyshev's inequality. Recall that this theorem states that any continuous function ϕ defined on a bounded closed interval is the uniform limit of a sequence of polynomials. Without loss of generality, we may assume that $\phi : [0,1] \to \mathbb{R}$. The usual constructive proof is to show that ϕ is the uniform limit of the Bernstein polynomials

$$B_n(x) = \sum_{k=0}^{n} \binom{n}{k} \phi(k/n) x^k (1-x)^{n-k} \quad \text{for all} \quad x \in \mathbb{R}.$$

The first step is to observe that the expression of $B_n(x)$ resembles the probability mass function of a binomial random variable. Indeed, the analysis problem can be turned into a probability problem by letting (X_n) be a sequence of independent Bernoulli (x) and using that

$$Z_n = (1/n)(X_1 + \cdots + X_n) \sim (1/n) \text{ Binomial}(n,x),$$

to re-express $B_n(x)$ as follows:

$$\begin{aligned}E(\phi(Z_n)) &= \sum_{k=0}^{n} \phi(k/n)\, P(X_1+\cdots+X_n=k) \\ &= \sum_{k=0}^{n} \binom{n}{k}\, \phi(k/n)\, x^k\,(1-x)^{n-k} = B_n(x).\end{aligned}$$

To find a good uniform upper bound for

$$\sup_x |B_n(x)-\phi(x)| = \sup_x |E(\phi(Z_n))-\phi(x)|$$

the basic idea is to decompose the expected value on the right-hand side distinguishing on whether Z_n is close to x or not. This gives a first term which is small because ϕ is continuous and a second term which is small because Z_n is close to x with high probability. To turn this heuristics into a proof, we fix $\varepsilon > 0$ small. Since ϕ is continuous and therefore bounded on $[0,1]$,

- there is $\delta > 0$ such that $|\phi(x)-\phi(y)| \leq \varepsilon$ when $|x-y| \leq \delta$,
- $m = \sup_{x\in[0,1]} |\phi(x)| < \infty$.

To see that $|Z_n - x| \leq \delta$ with high probability, note that $E(Z_n) = x$ and

$$\mathrm{Var}\,(Z_n) = (1/n)^2\,\mathrm{Var}\,(\mathrm{Binomial}\,(n,x)) = x\,(1-x)/n \leq 1/4n$$

therefore Chebyshev's inequality (3.4) implies that

$$P(|Z_n-x|>\delta) = P(|Z_n-E(Z_n)|>\delta) \leq (1/\delta)^2\,\mathrm{Var}\,(Z_n) \leq 1/4n\delta^2.$$

Decomposing on whether $|Z_n - x|$ is larger or smaller than δ, we get

$$\begin{aligned}|B_n(x)-\phi(x)| &= |E(\phi(Z_n))-\phi(x)| \\ &\leq |E(((\phi(Z_n))-\phi(x))\,\mathbf{1}\{|Z_n-x|>\delta)| \\ &\qquad + |E(((\phi(Z_n))-\phi(x))\,\mathbf{1}\{|Z_n-x|\leq\delta)| \\ &\leq 2m\,P(|Z_n-x|>\delta)+\varepsilon\,P(|Z_n-x|\leq\delta) \leq m/2n\delta^2+\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ can be chosen arbitrarily small, this shows uniform convergence on the unit interval and completes the proof of the Weierstrass theorem. □

We now return to the general context (3.3). Since

$$E(S_n/n) = \mu \quad \text{and} \quad \mathrm{Var}(S_n/n) = \sigma^2/n,$$

it follows from Chebyshev's inequality (3.4) that

$$\lim_{n\to\infty} P(|S_n/n-\mu| \geq \varepsilon) \leq \lim_{n\to\infty} (1/n)(\sigma/\varepsilon)^2 = 0.$$

This shows that $S_n/n \xrightarrow{p} \mu$, a result known as the **weak law of large numbers**. To prove the strong law of large numbers, which states that we have in fact convergence almost surely, we use the first Borel–Cantelli lemma 3.5.

Theorem 3.1 (Strong law of large numbers). *Let (X_n) be independent identically distributed random variables with finite mean μ. Then,*

$$\frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu.$$

As we do for the weak law of large numbers, we only prove the result when the variance is finite but refer to the end of Chapter 5 for a proof in the general case using Kolmogorov's zero-one law and reverse martingales.

Proof. Since $\operatorname{Var}(S_{n^2}/n^2) = (\sigma/n)^2$, it follows from (3.4) that

$$\sum_{n=1}^{\infty} P\left(\left|\frac{S_{n^2}}{n^2} - \mu\right| \geq \varepsilon\right) \leq \sum_{n=1}^{\infty} \left(\frac{\sigma}{n\varepsilon}\right)^2 < \infty$$

so the first Borel–Cantelli lemma implies that

$$\begin{aligned} & P(|S_{n^2}/n^2 - \mu| \geq \varepsilon \text{ infinitely often for some } \varepsilon \in \mathbb{Q}_+^*) \\ & \quad \leq \ \textstyle\sum_{\varepsilon \in \mathbb{Q}_+^*} P(|S_{n^2}/n^2 - \mu| \geq \varepsilon \text{ infinitely often}) = 0 \end{aligned}$$

showing that $S_{n^2}/n^2 \xrightarrow{a.s.} \mu$. We easily deduce the theorem when the random variables are nonnegative using that, in this case, the sequence of the partial sums (S_n) is monotone. In the general case, invoking the linearity of the expected value, we deduce that, regardless of the sign of the X_n,

$$\begin{aligned} \lim_{n\to\infty} S_n/n & = \lim_{n\to\infty} (1/n)((X_1^+ + \cdots + X_n^+) - (X_1^- + \cdots + X_n^-)) \\ & = E(X_1^+) - E(X_1^-) = E(X_1^+ - X_1^-) = E(X_1) = \mu \end{aligned}$$

almost surely, which completes the proof. □

In words, the law of large numbers states that the empirical mean S_n/n, which is the average over n independent realizations of the same random variable, converges almost surely to the theoretical mean. Note that the result is universal in the sense that it holds regardless of the distribution of the random variables. The law of large numbers can also be used to estimate empirically the probability of an event, which gives an intuitive idea of the meaning of a probability. To see this, assume that one repeats a given random experiment a large number of times and let A be an event of interest attached to this experiment. To estimate its probability, consider the Bernoulli random variables

$$X_n = \mathbf{1}\{A \text{ occurs at step } n\}.$$

Assuming that the consecutive outcomes are independent so that the random variables X_n are independent, the law of large numbers implies that

$$(1/n)(X_1 + \cdots + X_n) \xrightarrow{a.s.} E(X_1) = P(X_1 = 1) = P(A).$$

This means that the fraction of realizations that result in the event A approaches the probability of the event A as the number of trials get larger. This can be used, for example, to estimate the number π as shown in the next example.

Example 3.2 (Buffon's needle and the number π). Assuming that someone drops a needle of length one onto a floor made of parallel strips of wood, each of which have the same width of one, we are interested in the probability that the needle lies across a line between two strips.

The first step to answer this question is to have a precise description of the position of the needle relative to the lines between strips in terms of random variables. There are two quantities that characterize the position of the needle: the acute angle θ between the needle and the lines and the distance D between the center of the needle and the closest line. It is natural to assume that these quantities are independent uniform random variables, i.e.,

$$\theta \sim \text{Uniform}\,(0,\pi/2) \quad \text{and} \quad D \sim \text{Uniform}\,(0,1/2).$$

Letting A be the event that the needle lies across a line, some basic trigonometry implies that the distance between the two extremities of the needle in the direction perpendicular to the lines is equal to $\sin(\theta)$, therefore

$$A \text{ occurs} \quad \text{if and only if} \quad 2D < \sin(\theta),$$

from which it follows that

$$P(A \,|\, \theta = x) = P(2D < \sin(x)) = \sin(x) \quad \text{for all} \quad x \in (0,\pi/2).$$

Letting f_θ be the density function of θ, we conclude that

$$\begin{aligned} P(A) &= \int P(A \,|\, \theta = x)\, f_\theta(x)\, dx \\ &= \frac{2}{\pi} \int P(A \,|\, \theta = x)\, \mathbf{1}\{x \in (0,\pi/2)\}\, dx = \frac{2}{\pi} \int_0^{\frac{\pi}{2}} \sin(x)\, dx = \frac{2}{\pi}. \end{aligned}$$

Using the law of large numbers, we now design an experiment to estimate π. Drop n needles of length one on the floor and consider the random variables

$$X_i = \mathbf{1}\{\text{needle } i \text{ lies across a line between two strips}\}$$

for all $i = 1,2,\ldots,n$. Then, under the reasonable assumption that these random variables are independent, the law of large numbers implies that

$$\begin{aligned} &2/\lim_{n\to\infty}(1/n)(X_1 + X_2 + \ldots + X_n) = 2/E(X_1) \\ &\quad = 2/P(\text{the first needle lies across a line}) = 2/P(A) = \pi \end{aligned}$$

almost surely. In particular, one way to find an approximation of π is to drop a large number of needles on the floor and be patient enough to count the number of needles that lie across a line between two strips. Calling this number k, twice the number of needles divided by k should be close to π. Of course, the larger the number of needles, the better the approximation. $\square$

3.3 Central limit theorem

The next natural question is: how much does S_n deviate from its mean when the integer n is large? The answer is given by the central limit theorem, which looks at the fluctuations around the mean and, like the law of large numbers, gives a universal estimate which holds regardless of the distribution.

The central limit theorem gives a convergence in distribution, a type of convergence that can be studied using Lévy's convergence theorem below which involves characteristic functions. Recall that the **characteristic function** of a real random variable X is the function defined on the real line by

$$\phi_X(\theta) = E(e^{i\theta X}) \quad \text{for all} \quad \theta \in \mathbb{R} \quad \text{where} \quad i = \sqrt{-1} \in \mathbb{C}.$$

Theorem 3.2 (Lévy's convergence theorem). *Let (X_n) be a sequence of random variables with characteristic functions ϕ_{X_n} and assume that*

- $\phi(\theta) = \lim_{n\to\infty} \phi_{X_n}(\theta)$ *exists for all $\theta \in \mathbb{R}$ and*
- *the function ϕ is continuous at zero.*

Then ϕ is the characteristic function of a random variable X and $X_n \xrightarrow{d} X$.

We omit the proof because it requires the introduction of a number of definitions and intermediate results that will not be used later. For a complete proof, we refer the reader to Chapters 16–18 of Williams [95]. The second ingredient we need to prove the central limit theorem is the following lemma that gives a Taylor expansion of the characteristic function of a square integrable random variable.

Lemma 3.8. *Assume that $E(X^2) < \infty$. Then,*

$$\phi_X(\theta) = 1 + i\theta E(X) - (1/2)\,\theta^2 E(X^2) + o(\theta^2) \quad \textit{for all } \theta \textit{ small.}$$

Proof. The basic idea is to control the remainder

$$R(x) = e^{ix} - \left(1 + ix - \frac{x^2}{2}\right) \quad \text{for} \quad x \in \mathbb{R}$$

and then set $x = \theta X$, take the expected value and apply the dominated convergence theorem. First, we use an integration by parts to get

$$\int_0^x (x-s)^n e^{is}\, ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1} \int_0^x (x-s)^{n+1} e^{is}\, ds.$$

Taking $n = 0$, we obtain

$$x + i\int_0^x (x-s)\, e^{is}\, ds = \int_0^x e^{is}\, ds = -i\, e^{ix} + i = \frac{e^{ix}-1}{i} \tag{3.5}$$

while, taking $n = 1$, we obtain

$$\int_0^x (x-s)\,e^{is}\,ds = \frac{x^2}{2} + \frac{i}{2}\int_0^x (x-s)^2\,e^{is}\,ds. \tag{3.6}$$

Using (3.5) and then (3.6), we deduce that

$$\begin{aligned} e^{ix} &= 1 + i\left(\frac{e^{ix}-1}{i}\right) = 1 + ix - \int_0^x (x-s)\,e^{is}\,ds \\ &= 1 + ix - \frac{x^2}{2} - \frac{i}{2}\int_0^x (x-s)^2\,e^{is}\,ds \end{aligned}$$

which, rearranging the terms, gives

$$R(x) = e^{ix} - \left(1 + ix - \frac{x^2}{2}\right) = -\frac{i}{2}\int_0^x (x-s)^2\,e^{is}\,ds. \tag{3.7}$$

Now, on the one hand, we have

$$|R(x)| \leq \frac{1}{2}\left|\int_0^x (x-s)^2\,ds\right| \leq \frac{|x|^3}{6} \tag{3.8}$$

while, on the other hand, a new integration by parts gives

$$\begin{aligned} -\frac{i}{2}\int_0^x (x-s)^2\,e^{is}\,ds &= \frac{x^2}{2} - \int_0^x (x-s)\,e^{is}\,ds \\ &= \int_0^x (x-s)\,ds - \int_0^x (x-s)\,e^{is}\,ds = \int_0^x (x-s)(1-e^{is})\,ds \end{aligned}$$

from which it follows that

$$|R(x)| = \left|\int_0^x (x-s)(1-e^{is})\,ds\right| \leq 2\left|\int_0^x (x-s)\,ds\right| \leq x^2. \tag{3.9}$$

Combining (3.7)–(3.9), setting as previously mentioned $x = \theta X$, taking the expected value, and using Jensen's inequality, we deduce that

$$\begin{aligned} &|\phi_X(\theta) - 1 - i\,\theta E(X) + (1/2)\,\theta^2 E(X^2)| \\ &\quad = |E(R(\theta X))| \leq E|R(\theta X)| \leq \theta^2 E(\min((1/6)\,|\theta X^3|, X^2)). \end{aligned}$$

Since the random variable in the last expected value converges almost surely to zero as $\theta \to 0$ and is bounded by X^2 which is integrable by assumption, it follows from the dominated convergence theorem that the expected value on the right-hand side goes to zero as $\theta \to 0$. This completes the proof. □

With Lévy's convergence theorem and Lemma 3.8 in hands, we are now ready to prove the central limit theorem.

Theorem 3.3 (Central limit theorem). *Let (X_n) be independent identically distributed random variables with finite mean μ and variance σ^2. Then,*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} X \sim \text{Normal}\,(0,1).$$

Recalling the density function of X, this means that

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt \quad \textit{for all} \quad x \in \mathbb{R}.$$

Proof. If this holds for the centered random variables then this holds in the general case so it suffices to prove the result when $\mu = 0$. Since the summands in S_n are independent and identically distributed random variables,

$$\begin{aligned} \phi_{S_n/\sigma\sqrt{n}}(\theta) &= E(e^{i\theta(X_1+\cdots+X_n)/\sigma\sqrt{n}}) \\ &= E(e^{i\theta X_1/\sigma\sqrt{n}})\cdots E(e^{i\theta X_n/\sigma\sqrt{n}}) \\ &= \phi_{X_1}(\theta/\sigma\sqrt{n})\cdots\phi_{X_n}(\theta/\sigma\sqrt{n}) = (\phi_{X_1}(\theta/\sigma\sqrt{n}))^n. \end{aligned} \tag{3.10}$$

Moreover, according to Lemma 3.8,

$$\begin{aligned} \phi_{X_1}(\theta) &= 1 + i\theta\,E(X_1) - (1/2)\,\theta^2\,E(X_1^2) + o(\theta^2) \\ &= 1 - (1/2)\,\theta^2\,\sigma^2 + o(\theta^2). \end{aligned} \tag{3.11}$$

Combining (3.10)–(3.11), we deduce that

$$\begin{aligned} \lim\nolimits_{n\to\infty} \phi_{S_n/\sigma\sqrt{n}}(\theta) &= \lim\nolimits_{n\to\infty} (\phi_{X_1}(\theta/\sigma\sqrt{n}))^n \\ &= \lim\nolimits_{n\to\infty} (1 - (1/2)\,\theta^2/n + o(1/n))^n = e^{-\theta^2/2}. \end{aligned}$$

Theorem 3.2 implies convergence in distribution to a random variable and, in order to conclude, it remains to prove that the limit above is indeed the characteristic function of the standard normal random variable X. To compute the characteristic function, we first use that the sinus function is odd to get

$$\phi_X(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\theta x}\,e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(\theta x)\,e^{-x^2/2}\,dx.$$

Taking the derivative and integrating by parts, we obtain

$$\begin{aligned} \phi_X'(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -x\,\sin(\theta x)\,e^{-x^2/2}\,dx \\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -\theta\,\cos(\theta x)\,e^{-x^2/2}\,dx = -\theta\,\phi_X(\theta). \end{aligned}$$

We easily deduce that $\phi_X(\theta)\, e^{\theta^2/2}$ is constant therefore

$$\phi_X(\theta) = \phi_X(0)\, e^{-\theta^2/2} = \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right) e^{-\theta^2/2} = e^{-\theta^2/2}$$

and the proof is complete. □

Looking again at the special case of independent Bernoulli random variables, the central limit theorem implies that, after a large number n of realizations of a random experiment, the deviation between the number of times one observes a given event A and its theoretical mean $n\,P(A)$ is of the order of $\sqrt{n}$. We conclude with two examples of application of the central limit theorem.

Example 3.3. At each step of a game, a gambler gives one dollar to pick a number uniformly at random between zero and one and increase her fortune by twice that number. After n steps, the gambler's winning is given by

$$X_n = (2U_1 - 1) + \cdots + (2U_n - 1) \quad \text{where} \quad U_i \sim \text{Uniform}\,(0,1).$$

We are interested in the probability that the gambler wins at least \$5 after playing this game 100 times. Since $2U_i - 1$ has mean zero and variance

$$\sigma^2 = E((2U_i - 1)^2) = \frac{1}{2}\int_{-1}^{1} x^2\, dx = 1/3,$$

by the central limit theorem,

$$P(X_{100} > 5) = P(X_{100}/10\,\sigma > \sqrt{3}/2) \approx \frac{1}{\sqrt{2\pi}} \int_{\sqrt{3}/2}^{\infty} e^{-x^2/2}\, dx \approx 0.194,$$

so the gambler wins more than \$5 with probability about 0.194. □

Example 3.4. Assume now that the gambler picks a number uniformly at random between zero and one and that her fortune is multiplied by twice that number. In particular, after n steps, the gambler's fortune is now given by

$$X_n = X_0\,(2U_1)\cdots(2U_n) = X_0\; 2^n\, U_1 \cdots U_n \quad \text{where} \quad U_i \sim \text{Uniform}\,(0,1)$$

and where X_0 is the gambler's initial fortune. By independence,

$$E(X_n) = X_0\, E\left(\prod_{i=1}^{n} (2U_i)\right) = X_0 \prod_{i=1}^{n} E(2U_i) = X_0$$

so the expected value of the gambler's fortune stays constant. We are interested in the probability that the gambler has more than one dollar after n steps. In order to apply the central limit theorem, the idea is to look at $\ln(X_n)$ to turn the product into a sum of random variables. Note that

$$P(-\ln(U_i) > t) = P(\ln(U_i) < -t) = P(U_i < e^{-t}) = e^{-t}$$

which shows that the random variables $-\ln(U_i)$ are independent and exponentially distributed with the same parameter of one. Since these random variables have mean and variance both equal to one, we deduce that

$$\begin{aligned} P(X_n > 1) &= P(\ln(X_n) > 0) \\ &= P(\ln(X_0) + \ln(2^n) + \sum_i \ln(U_i) > 0) \\ &= P(-\sum_i \ln(U_i) < \ln(X_0) + n\ln(2)) \\ &= P((-\sum_i \ln(U_i) - n)/\sqrt{n} < (\ln(X_0) + n\ln(2) - n)/\sqrt{n}). \end{aligned}$$

For instance, if the gambler starts with $X_0 = 10^9$ dollars, since

$$\ln(10^9) + 100\ln(2) - 100 \approx -9.962 \approx -10,$$

the probability that she has at least \$1 left at step $n = 100$ is about

$$P\left(\frac{-\sum_i \ln(U_i) - 100}{10} < -1\right) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1} e^{-x^2/2}\,dx \approx 0.159.$$

The result is quite surprising. It shows that, after playing 100 times, a gambler who starts as a billionaire is likely to be left with less than one dollar. The reason for this is that, at each step, the gambler cannot do better than doubling her fortune, whereas many times she will see her fortune being divided by quite large numbers. The fact that the expected value of the gambler's fortune remains constant over time does not contradict this result. This simply means that the ultimate fortune can be extremely large, but this only happens with a very small probability. □

3.4 Exercises

Exercise 3.1 (Gossip process). The following problem is inspired from a model studied in [36]. Consider a population of N individuals. At each time step, one of the individuals tells a rumor to another individual, with each pair of individuals being equally likely to be selected. If the teller already heard about a rumor, she spreads one of them, whereas if she did not, she creates a rumor, and we assume that all the rumors created are different from one another.

1. Compute the expected value of the number X of rumors ever created.
2. Prove that
$$\mathrm{Var}(X) = \frac{N}{4}\left(\frac{N-2}{2N-3}\right).$$
3. Conclude that, for all $\alpha > 1/2$,
$$\lim_{N\to\infty} P(|X - N/2| \geq N^{\alpha}) \leq \lim_{N\to\infty} N^{-2\alpha}\,\mathrm{Var}(X) = 0.$$

Hint: For the first two questions, write X as a sum of N random variables.

Exercise 3.2 (Roulette). At the game of roulette, there are two green numbers, eighteen red numbers and eighteen black numbers. When the gambler bets on red, she either wins her bet if a red number appears or loses her bet otherwise. This leads to a probability $18/38 < 1/2$ of winning. Now, consider the following betting strategy: Bet one dollar on red. If a red number appears, then quit. Otherwise, play once more by betting two dollars on red.

1. Prove that, when following this strategy, the probability of winning becomes larger than one half.
2. Use the law of large numbers to argue that, in spite of this result, this game is unfavorable.

Exercise 3.3. Consider a community of n families that are independently of size s with the same probability p_s and finite expected size. Let B_s be the event that an individual chosen uniformly at random is from a family with s members. Use the law of large numbers to show that

$$\lim_{n\to\infty} P(B_s) = sp_s/\mu \quad \text{where} \quad \mu = \sum_{s=1}^{n} s\,p_s < \infty.$$

Hint: Condition on $X_1, X_2, \ldots, X_n$ where X_i is the size of family i.

Exercise 3.4. Consider a particle that moves on the integers. The particle starts at zero and, at each time step, jumps independently one unit to the right with probability p and one unit to the left with probability q with $p+q=1$. Use the law of large numbers to prove that, when $p \neq q$, the number of times the particle visits zero is almost surely finite.

Exercise 3.5 (Stirling's formula). This exercise gives a probabilist proof of Stirling's formula used later to study symmetric random walks. Let X_i be independent Poisson random variables with mean $\mu = 1$.

1. Use the central limit theorem to show that

$$\lim_{n\to\infty} \sqrt{2\pi n}\, P(S_n = n) = 1 \quad \text{where} \quad S_n = X_1 + X_2 + \cdots + X_n.$$

2. Deduce Stirling's formula: $n! \sim \sqrt{2\pi n}\,(n/e)^n$ as $n \to \infty$.

Exercise 3.6. Letting X_i be independent Poisson random variables with mean $\mu = 1$, follow the same argument as in Exercise 3.5 to show that

$$1 + n + \frac{n^2}{2} + \frac{n^3}{6} + \frac{n^4}{24} + \cdots + \frac{n^n}{n!} \sim (1/2)\,e^n \quad \text{as} \quad n \to \infty.$$

Part II
Stochastic processes

Chapter 4
Stochastic processes: general definition

This short chapter introduces the main topic of this textbook, namely, stochastic processes. As was true for the law of large numbers and the central limit theorem, stochastic processes are collections of random variables defined on the same probability space. These random variables, however, are not usually assumed to be independent. Instead, they are connected by some dependency relationships that can typically be expressed using conditional expectation or probability. In this chapter, we give the general definition of a stochastic process and define martingales and Markov chains along with a couple of examples.

A stochastic process is a collection of random variables

$$\{X_i : i \in I\} \quad \text{where } I \text{ can be finite or infinite}$$

defined on the same probability space $(\Omega, \mathscr{F}, P)$ and with value in $(S, \mathscr{S})$, a topological space equipped with its Borel σ-algebra. In most of our examples, the processes will have value in a subset of the real line: $S \subset \mathbb{R}$.

- The set Ω is called the **set of realizations**.
- The set S is called the **state space** of the process.
- For all $\omega \in \Omega$, the set $\{X_i(\omega) : i \in I\}$ is called a **sample path**.

The random variables that constitute a stochastic process are often ordered, meaning that the index set is a totally ordered set to describe random phenomena that evolve in time, but they can also model more general random structures. For example, starting from a deterministic graph, i.e., a set of vertices along with a set of edges that connect some of these vertices, one can define a collection of independent Bernoulli random variables indexed by the set of edges and consider the random subgraph induced by the set of edges for which a success occurs. This random subgraph is an example of a stochastic process.

In the event the stochastic process indeed models a random phenomenon that evolves over time, the temporal structure is encoded in the index set I. To model discrete time we take $I = \mathbb{N}$, whereas in continuous time we have $I = \mathbb{R}_+$. It is natural

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_4

to assume that the future, which is random and yet still to be discovered, depends on the past and present, which are both known. The relationship between the fixed past and the random future is expressed mathematically assuming that the future random variables are measurable with respect to the past random variables, therefore using conditional expectation or conditional probability given the σ-algebra containing the past information. To make this precise, we focus for simplicity on the discrete-time setting, in which case we have the following definitions.

- **Filtration** is a nondecreasing sequence $(\mathscr{G}_n)$ of sub-σ-algebras of $\mathscr{F}$, with $\mathscr{G}_n$ representing the information available at time n. Obviously, as time evolves, more information becomes available, which explains why it is assumed that a filtration is nondecreasing.
- We say that (X_n) is **adapted** to the filtration $(\mathscr{G}_n)$ when X_n is $\mathscr{G}_n$-measurable for all times n, which models the fact that the sample path of the process up to time n must be part of the information available at that time.
- The **natural filtration** of the process (X_n) is the special filtration

  $$\mathscr{F}_n = \sigma(X_0, X_1, \ldots, X_n) \quad \text{for all} \quad n \in \mathbb{N}.$$

  By construction, this is the smallest filtration (X_n) is adapted to; therefore, the natural filtration is the information available at time n under the assumption that nothing else other than the process itself is known.

The two most common classes of discrete-time stochastic processes are martingales and Markov chains. They can be defined, respectively, using the conditional expectation and the conditional probability with respect to a given filtration.

Definition 4.1 (Martingale). The stochastic process (X_n) is a martingale with respect to a filtration $(\mathscr{G}_n)$ if it is adapted to this filtration and

$$E(X_{n+1} \,|\, \mathscr{G}_n) = X_n \quad \text{for all} \quad n \in \mathbb{N}.$$

Even though it is not specified, we point out that this equation, which involves two random variables, holds almost surely. Similarly, the stochastic process is called a supermartingale and a submartingale when the equality above is replaced by $\leq$ and $\geq$, respectively. When the filtration is not specified, this implicitly means that the process is a martingale with respect to its natural filtration $(\mathscr{F}_n)$.

It is convenient to think of the random variable X_n as the fortune of a gambler after she has played n times a fair game (martingale), an unfavorable game (supermartingale), or a favorable game (submartingale). In this context, the definition means that regardless of the past, on average the fortune of the gambler will not increase or decrease when she bets once more on a fair game.

Definition 4.2 (Markov chain). The stochastic process (X_n) is a discrete-time Markov chain whenever

$$P(X_{n+1} \in B \,|\, \mathscr{F}_n) = P(X_{n+1} \in B \,|\, X_n) \quad \text{for all} \quad B \in \mathscr{S}$$

where $(\mathscr{F}_n)$ is the natural filtration of the process. Again, even though it is not specified, this equation holds almost surely.

In words, thinking of n as the present time, the definition of a Markov chain means that the future of the process depends on the past only through the present, or equivalently, given the present, past and future are independent. Intuitively, this means that the process is memoryless, that is, it forgets the past and makes its next move based only on where it is at the current time.

It is important to identify the type of stochastic processes we are dealing with here in order to analyze them, because the techniques available vary strongly among the different classes of stochastic processes.

This chapter only gives the definition of discrete-time martingales and Markov chains, but these processes can also be defined in a continuous-time setting. In fact, one chapter will be devoted to the general theory of Markov chains in continuous time while other chapters will focus on examples of such processes. Looking at the state space, this textbook emphasizes discrete models, meaning that the state space is either finite or countable. Percolation models and interacting particle systems appear as exceptions because they describe spatial configurations on infinite graphs and have therefore an uncountable state space. These models, however, are typically viewed as examples of discrete stochastic models because the underlying spatial structure is discrete. Although it is not the focus of this textbook, the reader should be aware that there are many examples of continuous-time stochastic process that have an uncountable state space and continuous sample paths. Examples of such processes are the diffusion processes, such as the popular Brownian motion, which will be briefly discussed later.

We now give a couple of concrete examples to show what martingales and Markov chains may look like in practice.

Example 4.1. Flip a fair coin infinitely often and let

$$Y_n = \begin{cases} -1 & \text{if the } n\text{th flip lands on tails} \\ +1 & \text{if the } n\text{th flip lands on heads.} \end{cases}$$

Then, the sequence (Y_n) is a stochastic process with state space $S = \{-1, +1\}$. This example is not particularly interesting because of the lack of dependence among the random variables, but we can use it to construct the more exciting random walk on the integers. □

Example 4.2 (Random walk on the integers). There is a particle located initially at the origin. At each time step, it jumps one unit to the left or one unit to the right independently of its current position with probability one-half. Let X_n be its position at time n. Referring to the previous example,

$$X_0 = 0 \quad \text{and} \quad X_n = Y_1 + Y_2 + \cdots + Y_n \quad \text{in distribution.}$$

In particular, the properties of the conditional expectation give

$$E(X_{n+1} \,|\, \mathscr{F}_n) = E(X_n + Y_{n+1} \,|\, \mathscr{F}_n) = X_n + E(Y_{n+1}) = X_n$$

indicating that (X_n) is a martingale, which is expected because the coin used to determine the moves is a fair coin. In addition, we have

$$P(X_{n+1} = x \,|\, \mathscr{F}_n) = (1/2)\,\mathbf{1}\{|X_n - x| = 1\} = P(X_{n+1} = x \,|\, X_n)$$

indicating that (X_n) is also a Markov chain. Note that the leftmost and rightmost terms are indeed equal because the middle term is a function of X_n and is therefore $\sigma(X_n)$-measurable. We point out that, to prove in practice that a process is a Markov chain, it suffices to show that the conditional probability on the left-hand side can be written as a function of X_n only. □

It is easy to slightly perturb our particle so that its position still defines a Markov chain but not a martingale, as shown in the next example.

Example 4.3. Assume that the coin in Example 4.1 is no longer fair. To fix the ideas, assume that the coin has a probability $p > 1/2$ of landing on heads and that the particle moves according to the following rules:

- the particle jumps right if its current position is even and the coin lands on heads or if the particle's current position is odd and the coin lands on tails
- the particle jumps left if its current position is even and the coin lands on tails or if the particle's current position is odd and the coin lands on heads.

Letting $q = 1 - p$, we get

$$\begin{aligned} P(X_{n+1} = x \,|\, \mathscr{F}_n) &= p\,\mathbf{1}\{X_n = x - 1\} + q\,\mathbf{1}\{X_n = x + 1\} \quad \text{when } x \text{ is odd} \\ &= q\,\mathbf{1}\{X_n = x - 1\} + p\,\mathbf{1}\{X_n = x + 1\} \quad \text{when } x \text{ is even} \end{aligned}$$

therefore the process is a Markov chain. However,

$$\begin{aligned} E(X_{n+1} \,|\, \mathscr{F}_n) &= X_n + (p - q) > X_n \quad \text{when} \quad X_n \text{ is even} \\ &= X_n - (p - q) < X_n \quad \text{when} \quad X_n \text{ is odd} \end{aligned}$$

showing that (X_n) is neither a supermartingale nor a submartingale. □

Modifying the random walk so that it is still a martingale but no longer a Markov chain is a little more tricky. This is done in the next example.

Example 4.4 (Random walk with betting). We return to our fair coin and assume that a gambler bets before each coin flip: she wins her bet each time the coin lands on heads and loses her bet otherwise. Letting X_n be the gambler's winning at time n, we recover the random walk on the integers assuming that the gambler bets one dollar at each step of the game. Now, assume that the gambler is allowed to place any bet B_n on the outcome of the nth flip. Then,

$$X_0 = 0 \quad \text{and} \quad X_n = B_1 Y_1 + B_2 Y_2 + \cdots + B_n Y_n$$

where we assume that B_n is $\mathscr{F}_{n-1}$-measurable because the gambler's bet on the nth coin flip must be fixed before that flip. In particular, according to the basic properties of the conditional expectation,

$$\begin{aligned} E(X_{n+1} \mid \mathscr{F}_n) &= E(X_n + B_{n+1} Y_{n+1} \mid \mathscr{F}_n) = X_n + E(B_{n+1} Y_{n+1} \mid \mathscr{F}_n) \\ &= X_n + B_{n+1} E(Y_{n+1}) = X_n \end{aligned}$$

so (X_n) is a martingale. Now, let $a > b > 0$ and assume that the gambler feels lucky and bets a after each winning but only b after each loss. Then,

$$\begin{aligned} P(X_{n+1} = X_n + a \mid \mathscr{F}_n) &= P(Y_{n+1} = 1 \text{ and } B_{n+1} = a \mid \mathscr{F}_n) \\ &= P(Y_{n+1} = 1)\,\mathbf{1}\{B_{n+1} = a\} = (1/2)\,\mathbf{1}\{Y_n = 1\} \\ &= (1/2)\,\mathbf{1}\{X_n > X_{n-1}\} \end{aligned}$$

therefore (X_n) is not a Markov chain. □

Chapter 5
Martingales

Martingales are extensively used in physics, biology, sociology and economics, among other fields. In particular, we will later use martingales in a biological context as models of fair competition involving species that have the same fitness. However, to explain the theory, it is more convenient to think of a martingale as the process that keeps track of the fortune of a gambler playing a fair game. Similarly, we think of submartingales and supermartingales as the processes that, respectively, keep track of the fortune of a gambler playing a favorable and an unfavorable game.

Although the use of mathematics to study gambling strategies for games of chance goes back to the 17th century with the works of Fermat and Pascal, it was only in 1934 that the rigorous definition of martingales viewed as stochastic processes was introduced by Paul Lévy. The term martingale which describes these processes was proposed a few years later by Jean Ville [94]. In common French, the word *martingale* refers to a betting strategy that guarantees the gambler to beat a fair game. The first important results in martingale theory, including most of the results presented in this chapter, are due to Joseph Doob [23, chapter 7].

An immediate consequence of the definition of supermartingales is that the expected value of the process is nonincreasing. In particular, for unfavorable games, the expected value of the gambler's fortune at time n cannot exceed the gambler's initial fortune. In the context of fair games, the expected value stays constant. This leads to a first natural question:

Question 1 — Does this remain true replacing n by a random time T?

The optional stopping theorem states that the answer is yes, but only under a couple of assumptions on the martingale and the random time. First, time T must satisfy a certain measurability condition. Thinking here of this time as the time gambler decides to quit, this condition basically indicates that the gambler's decision cannot be based on the outcomes of future games since the gambler does not know about these outcomes yet. Such a time is called a *stopping time*. The theorem then says that, in the context of unfavorable games, the expected value at time T cannot exceed the initial gambler's fortune if time T is almost surely finite and the process is bounded or also if time T has a finite expected value and the process has uniformly bounded

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_5

increments. This implies in particular that, no matter the gambler's strategy, she cannot beat an unfavorable game. This result is applied to two popular problems: the gambler's ruin and the abracadabra problem.

As is true for any other stochastic process that evolves over time, another natural question is about the convergence of the process.

Question 2 — Does the process converge? More precisely, does it converge in distribution, almost surely, in mean, and so on?

As we will see later, the best we can generally expect for Markov chains is convergence of the process in distribution. For submartingales and supermartingales, the martingale convergence theorem gives the following remarkable result under some mild assumptions: The process converges almost surely to a random variable. Focusing on supermartingales to fix the ideas, it can be proved that each upcrossing of a given interval—defined as the event that the process goes from below a to above b where $a < b$—has a certain cost. This is then used to show that, provided the process is bounded in L^1, the number of upcrossings of any interval is almost surely finite, which implies almost sure convergence to a random variable. In particular, almost all the sample paths of the process converge to a fixed value.

Then, we introduce the concept of uniform integrability. For martingales that are uniformly integrable, the convergence theory gives stronger results. In particular, we prove that a supermartingale or submartingale converges not only almost surely but also in mean if and only if it is uniformly integrable. For martingales, convergence in mean is also equivalent to the fact that the process can be represented as the conditional expectation of a fixed integrable random variable given the members of the filtration. Such martingales are called *regular martingales*. Special cases of regular martingales are *reverse martingales*, which are used in this chapter to prove the strong law of large numbers in its full generality, i.e., when the common variance of the random variables is not necessarily finite.

To conclude the chapter, we look at martingales bounded in L^p for some $p > 1$. Not only are these processes uniformly integrable and therefore they converge almost surely and in mean, but they also converge in L^p. This will be proved using the so-called Doob's inequality.

Further reading

The literature on martingales is rather copious.

- For references specifically about martingale theory and its applications, we refer the reader to [77, 95].
- General references on stochastic processes that offer a thorough treatment of martingale theory include [12, 28].
- For a brief review with a number of exercises with solutions, see [3].
- For a treatment of martingale theory at the undergraduate level not relying on measure theory, we refer to [28, 49, 85].

5.1 Optional stopping theorem

Assume that (X_n) is a supermartingale with respect to a filtration $(\mathscr{G}_n)$, which models an unfavorable game. Then, a simple induction implies that

$$E(X_n) = E(E(X_n \,|\, \mathscr{G}_{n-1})) \leq E(X_{n-1}) \leq \cdots \leq E(X_0) \tag{5.1}$$

and a natural question arises: Is this still true looking more generally at random times instead of deterministic times? This is the motivation behind the optional stopping theorem. To state the result, we need to define stopping times.

Definition 5.1. A **stopping time** for the filtration $(\mathscr{G}_n)$ is a random variable

$$T : \Omega \to \bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\} \quad \text{such that} \quad \{T = n\} \in \mathscr{G}_n \quad \text{for all} \quad n \geq 0.$$

Thinking of a stopping time as the (random) time at which a gambler stops playing, the measurability condition in the definition means that the decision to stop at time n must be based on the information available at that time. The stopping times we will consider in practice are hitting times, i.e., the first time at which the process hits a certain Borel set $B \subset \mathbb{R}$. In this case,

$$\{T = n\} = \{X_1 \notin B, X_2 \notin B, \ldots, X_{n-1} \notin B, X_n \in B\} \in \mathscr{F}_n$$

so T is indeed a stopping time with respect to $(\mathscr{F}_n)$. Note also that the process stopped at time T, defined as $X_n^T = X_{n \wedge T}$, is again a supermartingale since

$$\begin{aligned} E(X_{n+1}^T - X_n^T \,|\, \mathscr{G}_n) &= E((X_{n+1} - X_n)\,\mathbf{1}\{T > n\} \,|\, \mathscr{G}_n) \\ &= \mathbf{1}\{T > n\}\; E(X_{n+1} - X_n \,|\, \mathscr{G}_n) = \mathbf{1}\{T > n\}\,(E(X_{n+1} \,|\, \mathscr{G}_n) - X_n) \leq 0 \end{aligned}$$

where we have used the fact that

$$\{T > n\} = \{T = 1\}^c \cap \{T = 2\}^c \cap \cdots \cap \{T = n\}^c \in \mathscr{G}_n.$$

Using this, we can now state and prove the optional stopping theorem. There are several versions of this theorem: All state that (5.1) still holds, replacing the deterministic time n by a stopping time T, but assuming different conditions. We will only give the two versions that we will use later.

Theorem 5.1 (Optional stopping theorem). *Assume that*

- *the process (X_n) is a supermartingale for the filtration $(\mathscr{G}_n)$ and*
- *time T is a stopping time for the same filtration.*

Then, we have $E(X_T) \leq E(X_0)$ in each of the following two cases:

1. *There exists $c < \infty$ such that for all $n \geq 0$*

$$|X_n^T| \leq c \quad \text{and} \quad P(T < \infty) = 1.$$

2. There exists $c < \infty$ such that for all $n \geq 0$

$$E(|X_{n+1} - X_n| \, \mathbf{1}\{T > n\} \, | \, \mathscr{G}_n) \leq c \quad \textit{almost surely} \quad \textit{and} \quad E(T) < \infty.$$

Proof. Because the process stopped at time T is again a supermartingale and also because, in both cases, the stopping time is almost surely finite, we have

$$\begin{aligned} \lim_{n\to\infty} E\,(X_n^T) &\leq \lim_{n\to\infty} E\,(X_0^T) = E(X_0^T) = E(X_0) \\ \lim_{n\to\infty} X_n^T &= \lim_{n\to\infty} (X_n \, \mathbf{1}\{T > n\} + X_T \, \mathbf{1}\{T \leq n\}) = X_T \end{aligned}$$

almost surely. In particular, it suffices to prove that

$$E\,(\lim_{n\to\infty} X_n^T) = \lim_{n\to\infty} E\,(X_n^T). \tag{5.2}$$

In the first case, since the process is bounded, this directly follows from the dominated convergence theorem. To deal with the second case, let

$$Y = |X_0| + \sum_{k=0}^{\infty} |X_{k+1} - X_k| \, \mathbf{1}\{T > k\}$$

and note that $|X_n^T| \leq Y$ for all $n \geq 0$. Using the fact that the process has uniformly bounded increments and that the stopping time has finite mean, we get

$$\begin{aligned} E(Y) &\leq E|X_0| + \textstyle\sum_k E\,(|X_{k+1} - X_k| \, \mathbf{1}\{T > k\}) \\ &= E|X_0| + \textstyle\sum_k E\,(E(|X_{k+1} - X_k| \, \mathbf{1}\{T > k\} \, | \, \mathscr{G}_k)) \\ &\leq E|X_0| + c \textstyle\sum_k P(T > k) \\ &= E|X_0| + c\,E(T) < \infty \end{aligned}$$

where the last equation comes from

$$\begin{aligned} E(T) = \sum_{i=1}^{\infty} i P(T = i) &= \sum_{i=1}^{\infty} \sum_{k=0}^{i-1} P(T = i) \\ &= \sum_{k=0}^{\infty} \sum_{i=k+1}^{\infty} P(T = i) = \sum_{k=0}^{\infty} P(T > k). \end{aligned} \tag{5.3}$$

This shows that the random variable Y is integrable so (5.2) again follows from the dominated convergence theorem. This completes the proof. □

More generally, we can prove that if S and T are two stopping times with $S \leq T$ almost surely then, under the assumptions of the theorem,

$$E(X_T) \leq E(X_S) \leq E(X_0).$$

We point out that the result also holds for submartingales with the inequality reversed, and that the inequality becomes an equality for martingales since martingales are both supermartingales and submartingales. Returning to the statement for supermartingales, the optional stopping theorem says: no matter the strategy the player follows, i.e., no matter the stopping time T, she cannot beat an unfavorable game in the sense that, at least on average, her fortune at the time she quits will be less than her fortune at the time she started playing.

Before giving examples of application of the theorem, we come back for a moment to the betting random walk introduced in Example 4.4. As previously mentioned, the French word *martingale* means: a strategy that guarantees the gambler to beat a fair game. There indeed exists such a strategy for the betting random walk: Keep doubling your bet as long as you lose and quit just after the first time you win which, in equation, can be written as $B_1 = 1$ and

$$B_n = 2^{n-1}\,\mathbf{1}\{Y_1 = Y_2 = \cdots = Y_{n-1} = -1\} \quad \text{for all} \quad n > 1.$$

Then, if the first heads appears at step n, the final winning is

$$-1-2-4-\cdots-2^{n-2}+2^{n-1} = 1 \quad \text{so} \quad E(X_T) = 1 \neq 0 = E(X_0)$$

even though the process is a martingale. The trouble here is that the martingale does not have uniformly bounded increments, in which case the conclusion of the optional stopping theorem may not hold.

Example 5.1 (Gambler's ruin chain). In this process, the gambler bets one dollar at each step, wins or loses her bet with probability p and $q = 1-p$, and quits when she is ruined or has reached some target N. In particular, the gambler's ruin chain can be seen as the random walk with betting stopped at time T with

$$B_n = 1 \quad \text{and} \quad P(Y_n = 1) = p \quad \text{and} \quad T = \inf\{n : X_n \in \{0,N\}\}.$$

Our objective is to compute the probability

$$p_x = P_x(X_T = N) = P(X_T = N \,|\, X_0 = x) \quad \text{where} \quad 0 \le x \le N$$

that the gambler quits a winner when she starts with x dollars. The winning probability can be computed using a so-called **first-step analysis**, which consists in conditioning on the possible values of the process after the first update. Using such a conditioning, we obtain

$$\begin{aligned} p_x &= \textstyle\sum_{y=x\pm 1} P(X_T = N \,|\, X_1 = y)\, P_x(X_1 = y) \\ &= \textstyle\sum_{y=x\pm 1} P(X_T = N \,|\, X_0 = y)\, P_x(X_1 = y) = p\,p_{x+1} + q\,p_{x-1} \end{aligned}$$

therefore $(p+q)\,p_x = p\,p_{x+1} + q\,p_{x-1}$ for all $0 < x < N$ and

$$\begin{aligned} p_{x+1} - p_x &= (q/p)\,(p_x - p_{x-1}) \\ &= c\,(p_x - p_{x-1}) = \cdots = c^x\,(p_1 - p_0) = c^x\,p_1 \end{aligned}$$

where $c = q/p$. Summing, we obtain

$$p_x = \sum_{z=0}^{x-1} (p_{z+1} - p_z) = \begin{cases} (1-c^x)(1-c)^{-1}\,p_1 & \text{when} \quad p \neq q \\ x\,p_1 & \text{when} \quad p = q. \end{cases}$$

Using that $p_N = 1$ to extract p_1, we conclude that

$$p_x = \begin{cases} (1-c^x)(1-c^N)^{-1} & \text{when} \quad p \neq q \\ x/N & \text{when} \quad p = q. \end{cases} \tag{5.4}$$

We now give an alternative less computational and more elegant proof using the first version of the optional stopping theorem. This approach is particularly useful when dealing with more complicated processes, in which case a first-step analysis can lead to tedious calculations whereas the use of the optional stopping theorem does not get much more complicated.

Proof of (5.4). Since it takes less than N jumps in one direction to bring the process to either zero or the target, the monotone convergence theorem implies that

$$\begin{aligned} P(T=\infty) &= \lim_{k\to\infty} P(T > kN) \\ &\leq \lim_{k\to\infty} (1-(p\vee q)^N)^k \leq \lim_{k\to\infty} (1-(1/2)^N)^k = 0 \end{aligned}$$

showing that T is almost surely finite. Now, when $p = q$, the game is fair and the process is a bounded martingale, so the optional stopping theorem gives

$$x = E_x(X_0) = E_x(X_T) = N p_x + 0\,(1-p_x) = N p_x$$

indicating that $p_x = x/N$. When $p \neq q$, letting again $c = q/p$, we have

$$\begin{aligned} E(c^{X_{n+1}} \,|\, \mathscr{F}_n) &= E(c^{X_n + Y_{n+1}} \,|\, \mathscr{F}_n) \\ &= c^{X_n}\, E(c^{Y_{n+1}}) = c^{X_n}\,(pc + qc^{-1}) = c^{X_n}. \end{aligned}$$

In particular, the process (c^{X_n}) is a martingale, therefore,

$$c^x = E_x(c^{X_0}) = E_x(c^{X_T}) = p_x\, c^N + (1-p_x)\, c^0 = p_x\,(c^N - 1) + 1$$

showing that $p_x = (1-c^x)(1-c^N)^{-1}$. □

For example, at the game of roulette introduced in Exercise 3.2, if at each step the gambler bets one dollar on either red or black, we have

$$p = 18/38 \quad \text{and} \quad q = 20/38 \quad \text{and} \quad c = q/p = 20/18 = 10/9.$$

Assuming that the gambler starts with \$10 and plays until she is either ruined or has doubled her fortune, the probability that she quits as a winner is given by

$$p_{10} = P_{10}(X_T = 20) = \frac{1-(10/9)^{10}}{1-(10/9)^{20}} = \frac{1}{1+(10/9)^{10}} \approx 0.259.$$

Another natural question is: With how much money should the gambler start so that the probability that she reaches \$20 is at least one-half? It suffices to solve

$$p_x = P_x(X_T = 20) = \frac{1-(10/9)^x}{1-(10/9)^{20}} > 1/2.$$

Some basic algebra gives

$$\begin{aligned} x &> \log(1-(1/2)(1-c^{20}))/\log(c) \\ &= \log(1-(1/2)(1-(10/9)^{20}))/\log(10/9) \approx 14.51 \end{aligned}$$

so she should start with at least \$15. Note that (5.4) also suggests that, even when the game is only slightly unfavorable, things get really bad as the number of games gets larger. For instance, if the target is \$100, the probability of winning starting with \$50 and the amount of money one should start with to make this probability larger than one-half are now given by

$$\begin{aligned} p_{50} &= P_{50}(X_T = 100) = 1/(1+(10/9)^{50}) \\ x &> \log(1-(1/2)(1-(10/9)^{100}))/\log(10/9) \end{aligned}$$

which gives $p_{50} \approx 0.0051$ and $x \geq 94$ dollars. □

For numerical simulations of the gambler's ruin chain in Matlab, we refer the reader to the simulation chapter at the end of this book.

Example 5.2. Consider an infinite sequence of letters chosen independently and uniformly at random, i.e., we let (X_n) be a sequence of independent random variables uniformly distributed on the set of the twenty-six letters of the Latin alphabet. We are interested in the expected value of the first time

$$T_A = \inf\{n : X_n = A\}$$

the letter A appears in the sequence. Noting that T_A is the geometric random variable with success probability $1/26$, its expected value is simply equal to 26. This problem can also be solved using the optional stopping theorem. This approach has the advantage that it can be extended to find more generally the expected value of the first time a given word appears.

To find the expected value, the first step is to construct a martingale by turning the problem into a fair game. Assume that a gambler bets one dollar on each of the outcomes of the sequence, loses her bet if the outcome is different from A, and wins otherwise. To make this game fair, it is natural to assume that she wins \$26 each time an A appears, suggesting that the process

$$Y_n = \sum_{k=1}^{n} (26\,\mathbf{1}\{X_k = A\} - 1)$$

is a martingale with respect to $\mathscr{F}_n = \sigma(X_1, X_2, \ldots, X_n)$. Indeed, using that the successive letters are chosen independently, we get

$$\begin{aligned} E(Y_{n+1} \,|\, \mathscr{F}_n) &= E(Y_n + 26\,\mathbf{1}\{X_{n+1} = A\} - 1 \,|\, \mathscr{F}_n) \\ &= Y_n + 26\,P(X_{n+1} = A) - 1 = Y_n + 26/26 - 1 = Y_n. \end{aligned}$$

Since the process also has bounded increments, we can apply the optional stopping theorem with the stopping time T_A. This implies that

$$0 = E(Y_0) = E(Y_{T_A}) = 26\, P(X_{T_A} = A) - E(T_A) = 26 - E(T_A)$$

which again gives $E(T_A) = 26$. □

Example 5.3. We now look at the expected value of the time T_{AB} it takes until we see the two-letter word AB, which we define formally as

$$T_{AB} = \inf\{n : (X_{n-1}, X_n) = (A, B)\}.$$

To get a fair game, we now assume that the gambler gets a payoff of \$$26^2$ each time the word appears. However, if the gambler plays n times and then stops, she also gets a partial payoff of \$26 if the nth letter is A, suggesting that

$$Y_n = \sum_{k=2}^{n} 26^2\, \mathbf{1}\{(X_{k-1}, X_k) = (A, B)\} + 26\, \mathbf{1}\{X_n = A\} - n$$

is a martingale. Indeed,

$$\begin{aligned} Y_{n+1} = Y_n &+ 26^2\, \mathbf{1}\{(X_n, X_{n+1}) = (A, B)\} \\ &+ 26\, \mathbf{1}\{X_{n+1} = A\} - 26\, \mathbf{1}\{X_n = A\} - 1. \end{aligned}$$

Taking the conditional expectation on both sides,

$$\begin{aligned} E(Y_{n+1} \mid \mathscr{F}_n) &= Y_n + 26^2\, \mathbf{1}\{X_n = A\}\, P(X_{n+1} = B) \\ &\qquad + 26\, P(X_{n+1} = A) - 26\, \mathbf{1}\{X_n = A\} - 1 \\ &= Y_n + 26\, \mathbf{1}\{X_n = A\} + 1 - 26\, \mathbf{1}\{X_n = A\} - 1 = Y_n \end{aligned}$$

so (Y_n) is a martingale with bounded increments. In addition,

$$E(T_{AB}) = \textstyle\sum_n P(T_{AB} > n) \leq 2 \sum_n P(T_{AB} > 2n) \leq 2 \sum_n (1 - 26^{-2})^n < \infty$$

so the optional stopping theorem is applicable, and we get

$$\begin{aligned} 0 = E(Y_0) = E(Y_{T_{AB}}) &= 26^2\, P((X_{T_{AB}-1}, X_{T_{AB}}) = (A, B)) - E(T_{AB}) \\ &= 26^2 - E(T_{AB}). \end{aligned}$$

In conclusion, $E(T_{AB}) = 26^2$. □

Example 5.4 (The abracadabra problem). The previous two examples can be generalized to any word with finite length, and a classical application of the optional stopping theorem is to find the expected value of the time T until the word abracadabra appears in our random sequence:

$$T = \inf\{n : (X_{n-10}, \ldots, X_n) = (A, B, R, A, C, A, D, A, B, R, A)\}.$$

In this case, things become more complicated, not because the word is longer but because there is repetition of some pieces of the word. Following the same approach as before, it is natural to consider the process

$$\begin{aligned} Y_n = \sum_{k=11}^{n} 26^{11}\, \mathbf{1}\{(X_{k-10}, X_{k-9}, \ldots, X_k) = (x_1, x_2, \ldots, x_{11})\} \\ + \sum_{k<11} 26^k\, \mathbf{1}\{(X_{n-k+1}, \ldots, X_n) = (x_1, \ldots, x_k)\} - n \end{aligned}$$

where x_i is the ith letter in the word. Then,

$$\begin{aligned} Y_{n+1} = Y_n + 26^{11}\, &\mathbf{1}\{(X_{n-9}, X_{n-8}, \ldots, X_{n+1}) = (x_1, x_2, \ldots, x_{11})\} \\ &+ \textstyle\sum_{k=1,2,\ldots,10} 26^k\, \mathbf{1}\{(X_{n-k+2}, \ldots, X_{n+1}) = (x_1, \ldots, x_k)\} \\ &\quad - \textstyle\sum_{k=1,2,\ldots,10} 26^k\, \mathbf{1}\{(X_{n-k+1}, \ldots, X_n) = (x_1, \ldots, x_k)\} - 1 \end{aligned}$$

from which we can again deduce that (Y_n) is a martingale. It is also clear that this martingale has bounded increments while the same argument as before also implies that the stopping time T has finite mean. In particular, we can apply the second version of the optional stopping theorem to get

$$\begin{aligned} 0 &= E(Y_0) = E(Y_T) \\ &= 26^{11}\, P((X_{T-10}, X_{T-9}, \ldots, X_T) = (x_1, x_2, \ldots, x_{11})) \\ &\quad + \textstyle\sum_{k=1,2,\ldots,10} 26^k\, P((X_{T-k+1}, \ldots, X_T) = (x_1, \ldots, x_k)) - E(T). \end{aligned}$$

Now, observe that the first and last letters in the word abracadabra are identical, as well as the first four letters and the last four letters:

$$\underline{\text{A}}\text{BRACADABR}\underline{\text{A}} \qquad \underline{\text{ABRA}}\text{CAD}\underline{\text{ABRA}}$$

from which it follows that

$$(X_{T-k+1}, \ldots, X_T) = (x_1, \ldots, x_k) \quad \text{if and only if} \quad k \in \{1, 4, 11\}.$$

In conclusion, we obtain the expected value

$$E(T) = \textstyle\sum_k 26^k\, P((X_{T-k+1}, \ldots, X_T) = (x_1, \ldots, x_k)) = 26 + 26^4 + 26^{11}.$$

More generally, if we fix $(x_1, x_2, \ldots, x_m) \in \{A, B, \ldots, Z\}^m$ and let T be the first time this word appears, the same argument shows that

$$E(T) = \textstyle\sum_{k\in\Theta} 26^k$$

where the sum is over $\Theta = \{k \leq m : (x_1, \ldots, x_k) = (x_{m-k+1}, \ldots, x_m)\}$. $\square$

5.2 Martingale convergence theorem

Another remarkable result about martingales is the martingale convergence theorem which states that supermartingales and submartingales bounded in L^1 converge almost surely to an integrable random variable X_∞. This does not mean that the process converges almost surely to a fixed value but that almost all the sample paths converge, with possibly different limits for different sample paths. In particular, the process does not oscillate or drift off to infinity.

To fix the idea, we prove the result for supermartingales. The first step is to show the upcrossing inequality which gives a bound for the **number of upcrossings** of an interval. Considering instead a submartingale, the same inequality holds but for the number of downcrossings of the interval. To state this result, we let (X_n) for $n \in \bar{\mathbb{N}}$ be a supermartingale defined at time infinity, fix $a < b$, and let

$$u(a,b) = \text{number of upcrossings of the interval } (a,b).$$

More precisely, letting $T_0 = 0$ and defining

$$T_{2k+1} = \inf\{n \geq T_{2k} : X_n \leq a\} \quad \text{and} \quad T_{2k+2} = \inf\{n \geq T_{2k+1} : X_n \geq b\}$$

for all $k \geq 0$, the number of upcrossings is given by

$$u(a,b) = \sup\{k : T_{2k} < \infty\}.$$

In other words, the number of upcrossings is the (random) number of times the supermartingale goes from below a to above b. Due to stochasticity, there may be some accidents, i.e., the presence of upcrossings, but since a supermartingale is expected to decrease, these upcrossings have a cost and it can be proved that the number of upcrossings is almost surely finite. The fact that the supermartingale is bounded in L^1 also prevents the sample paths from drifting off to minus infinity; therefore, almost all the sample paths must converge. We now turn this heuristic argument into a proof, starting with the upcrossing inequality.

Lemma 5.1 (Upcrossing inequality). *For all $k \in \mathbb{N}$,*

$$P(u(a,b) > k) \leq (b-a)^{-1} \, E\,[(X_\infty - a)^- \, \mathbf{1}\{u(a,b) = k\}]. \tag{5.5}$$

Proof. Since the process $(X_n - a)$ is also a supermartingale, it suffices to prove the result when $a = 0$. Let

$$\begin{aligned}(\sigma_1, \sigma_2) &= \text{time interval when the } (k+1)\text{th upcrossing occurs} \\ &= (T_{2k+1}, T_{2k+2})\end{aligned}$$

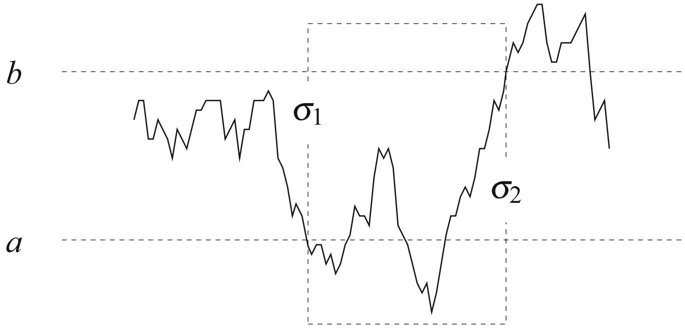


Fig. 5.1 Picture of the $(k+1)$th upcrossing of the interval (a,b).

as shown in Figure 5.1. Note that

$$\{u(0,b) > k\} = \{\sigma_2 < \infty\} \subset \{\sigma_1 < \infty\} \cap \{X_{\sigma_2} \geq b\}$$

from which it follows that

$$\begin{aligned} P(u(0,b) > k) &= E(\mathbf{1}\{u(0,b) > k\}) \leq (1/b)\,E\,[X_{\sigma_2}\,\mathbf{1}\{u(0,b) > k\}] \\ &\leq (1/b)\,E\,[X_{\sigma_2}^+\,\mathbf{1}\{u(0,b) > k\}] \\ &\leq (1/b)\,E\,[X_{\sigma_2}^+\,\mathbf{1}\{\sigma_1 < \infty\}]. \end{aligned} \tag{5.6}$$

Since (X_n) is a supermartingale and $X_{\sigma_1} \leq 0$ on the event $\sigma_1 < \infty$, it follows from the optional stopping theorem that

$$\begin{aligned} &E\,[X_{\sigma_2}^+\,\mathbf{1}\{\sigma_1 < \infty\}] - E\,[X_{\sigma_2}^-\,\mathbf{1}\{\sigma_1 < \infty\}] \\ &\quad = \; E\,[X_{\sigma_2}\mathbf{1}\{\sigma_1 < \infty\}] \leq E\,[X_{\sigma_1}\,\mathbf{1}\{\sigma_1 < \infty\}] \leq 0. \end{aligned} \tag{5.7}$$

Combining (5.6) and (5.7), we deduce that

$$\begin{aligned} P(u(0,b) > k) &\leq (1/b)\,E\,[X_{\sigma_2}^-\,\mathbf{1}\{\sigma_1 < \infty\}] \\ &\leq (1/b)\,E\,[X_{\sigma_2}^-\,\mathbf{1}\{\sigma_1 < \infty, X_{\sigma_2} \leq 0\}] \\ &\leq (1/b)\,E\,[X_{\sigma_2}^-\,\mathbf{1}\{\sigma_1 < \infty, \sigma_2 = \infty\}] \\ &\leq (1/b)\,E\,[X_\infty^-\,\mathbf{1}\{u(0,b) = k\}] \end{aligned}$$

which completes the proof. □

Using the upcrossing inequality and some measure theory, we can now prove that the number of upcrossings is almost surely finite and deduce the main result of this section: the martingale convergence theorem. We again state and prove the theorem for supermartingales but the result obviously holds for submartingales.

Theorem 5.2 (Martingale convergence theorem). *Let (X_n) be a supermartingale bounded in L^1, i.e., such that $\sup_n E\,|X_n| < \infty$. Then,*

$$X_n \xrightarrow{a.s.} X_\infty \quad \textit{where} \quad X_\infty \in L^1(\Omega, \mathscr{F}, P).$$

Proof. To begin with, we fix $n > 0$ and denote by $u_n(a,b)$ the number of upcrossings that occur by time n. Applying the upcrossing inequality (5.5) to the process stopped at time n and using the same trick as in (5.3) to re-express the expected value using Fubini's theorem, we obtain

$$\begin{aligned} E(u_n(a,b)) &= \textstyle\sum_k P(u_n(a,b) > k) \\ &\leq (b-a)^{-1} \textstyle\sum_k E\left[(X_n - a)^- \mathbf{1}\{u_n(a,b) = k\}\right] \\ &= (b-a)^{-1} E((X_n - a)^-). \end{aligned}$$

From the monotone convergence theorem, we deduce that

$$\begin{aligned} E(u(a,b)) &= \lim_{n\to\infty} E(u_n(a,b)) \\ &\leq (b-a)^{-1} \lim_{n\to\infty} E((X_n - a)^-) \\ &\leq (b-a)^{-1} \sup_n E(X_n^- + |a|) < \infty \end{aligned}$$

so $P(u(a,b) < \infty) = 1$ for all $a < b$. Since $\mathbb{Q}$ is countable, we also have

$$\begin{aligned} & P(\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n) \\ &\quad = P(\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n \text{ for some } a, b \in \mathbb{Q}) \\ &\quad = P(u(a,b) = \infty \text{ for some } a, b \in \mathbb{Q} \text{ with } a < b) = 0 \end{aligned}$$

which proves almost sure convergence:

$$X_n \xrightarrow{a.s.} X_\infty \quad \text{where} \quad E|X_\infty| \leq \sup_n E|X_n| < \infty.$$

This completes the proof. □

To comment on the assumptions of the theorem, note that the symmetric random walk on the integers does not converge almost surely. The reason is that the process is not bounded in L^1. Similarly, the random walk on a finite interval with reflecting boundaries does not converge almost surely. In this case, the reason is that the process is no longer a martingale. However, the theorem holds in the presence of absorbing boundaries, which corresponds to the gambler's ruin chain.

A beautiful application of the martingale convergence theorem is the so-called Kolmogorov's zero-one law which states that tail events, which are events that are not affected by any finite subset of an infinite collection of independent random variables, either almost surely occur or almost surely do not occur. To make this statement precise, we consider a sequence (X_n) of independent random variables and consider the σ-algebra

$$\mathscr{F}_+ = \sigma\left(\bigcap_{n=1}^{\infty} \mathscr{F}_+^n\right) \quad \text{where} \quad \mathscr{F}_+^n = \sigma(X_n, X_{n+1}, X_{n+2}, \ldots).$$

Then, we call A a **tail event** whenever $A \in \mathscr{F}_+$. For example, the event that

$$S_n = X_1 + X_2 + \cdots + X_n \xrightarrow{a.s.} \infty$$

is a tail event because, for each n, whether this event occurs or not does not depend on the first n random variables. The **Kolmogorov's zero-one law** states that such an event has probability either zero or one.

Theorem 5.3. *For all $A \in \mathscr{F}_+$, we have $P(A) \in \{0,1\}$.*

To be able to prove the strong law of large numbers later, we prove the following slightly stronger result: Each random variable X such that

$$X \text{ is } \mathscr{F}_+\text{-measurable} \quad \text{and} \quad X \in L^1(\Omega, \mathscr{F}, P),$$

is almost surely constant.

Proof. Letting $(\mathscr{F}_-^n)$ be the natural filtration of (X_n),

$$Y_n = E(X \,|\, \mathscr{F}_-^n) \quad \text{and} \quad \mathscr{F}_- = \sigma\left(\bigcup_{n=1}^{\infty} \mathscr{F}_-^n\right)$$

the basic idea is to apply the martingale convergence theorem to (Y_n). To see that this is indeed a martingale, note that, since the filtration $(\mathscr{F}_-^n)$ is nondecreasing, the projection rule for conditional expectation implies that

$$E(Y_{n+1} \,|\, \mathscr{F}_-^n) = E(E(X \,|\, \mathscr{F}_-^{n+1}) \,|\, \mathscr{F}_-^n) = E(X \,|\, \mathscr{F}_-^n) = Y_n.$$

Since the martingale (Y_n) is also bounded in L^1 because X is integrable, it follows from the martingale convergence theorem that

$$Y_n = E(X \,|\, \mathscr{F}_-^n) \xrightarrow{a.s.} E(X \,|\, \mathscr{F}_-).$$

Recalling that the random variable X is $\mathscr{F}_+$-measurable, the left-hand side and right-hand side are respectively equal to

$$\begin{aligned} E(X \,|\, \mathscr{F}_-^n) = E(X) &\quad \text{because} \quad \mathscr{F}_+^{n+1} \text{ is independent of } \mathscr{F}_-^n \\ E(X \,|\, \mathscr{F}_-) = X &\quad \text{because} \quad \mathscr{F}_+ \subset \mathscr{F}_-. \end{aligned}$$

This, together with the almost sure convergence, implies $X = E(X)$ almost surely, showing in particular that X is almost surely constant. To see that this indeed implies the zero-one law, we simply apply the result to the random variable

$$X = \mathbf{1}_A \quad \text{where} \quad A \in \mathscr{F}_+$$

which gives $P(A) = E(X) = X = \mathbf{1}_A \in \{0,1\}$. □

5.3 Uniformly integrable and regular martingales

In this section, we look at special cases of martingales for which the convergence theory is simplified. In particular, in addition to converging almost surely, they converge in mean. To begin with, we study general uniformly integrable collections of random variables which turn out to be particularly interesting when applied to the special case of martingales. To motivate the definition of uniform integrability, we first observe that a random variable X is integrable if and only if

$$\lim_{M\to\infty} E(|X|\,\mathbf{1}\{|X|>M\}) = 0. \tag{5.8}$$

Indeed, assuming that (5.8) holds, there exists $M<\infty$ such that

$$E\,|X| = E(|X|\,\mathbf{1}\{|X|\le M\}) + E(|X|\,\mathbf{1}\{|X|\le M\}) \le M+1.$$

Conversely, if the random variable X is integrable, then it is almost surely finite, so the monotone convergence theorem implies that

$$\lim_{M\to\infty} E(|X|\,\mathbf{1}\{|X|>M\}) = E(\lim_{M\to\infty} |X|\,\mathbf{1}\{|X|>M\}) = 0.$$

Having a general infinite collection of random variables, the uniform integrability of this collection is defined by generalizing (5.8) as follows.

Definition 5.2. The collection $\{X_i : i\in I\}$ is uniformly integrable if

$$\lim_{M\to\infty} \sup_{i\in I} E(|X_i|\,\mathbf{1}\{|X_i|>M\}) = 0.$$

A notion related to uniform integrability is equi-integrability.

Definition 5.3. The collection $\{X_i : i\in I\}$ is equi-integrable if, for all $\varepsilon>0$, there exists a small $\delta_\varepsilon>0$ such that

$$P(B)\le\delta_\varepsilon \quad \text{implies that} \quad E(|X_i|\,\mathbf{1}_B)\le\varepsilon \ \text{ for all } \ i\in I.$$

To ultimately study the convergence in mean, we start by proving that a collection of random variables is uniformly integrable if and only if it is bounded in L^1 and equi-integrable. More precisely, we have the following result.

Theorem 5.4. *The following two statements are equivalent.*

(a) (X_i) *is uniformly integrable.*

(b) (X_i) *is equi-integrable and bounded in* $L^1(\Omega,\mathscr{F},P)$.

Proof of (a) implies (b). There exists $M<\infty$ such that

$$\begin{aligned}\sup_{i\in I} E\,|X_i| \le\ & \sup_{i\in I} E(|X_i|\,\mathbf{1}\{|X_i|\le M\}) \\ & + \sup_{i\in I} E(|X_i|\,\mathbf{1}\{|X_i|>M\}) \le M+1<\infty\end{aligned}$$

so (X_i) is bounded in L^1. To prove equi-integrability, let $\varepsilon > 0$. Then, uniform integrability implies that there exists $M < \infty$ such that

$$E(|X_i|\,\mathbf{1}\{|X_i| > M\}) \leq \varepsilon/2 \quad \text{for all} \quad i \in I.$$

In particular, taking $\delta_\varepsilon = \varepsilon/2M > 0$, we get

$$\begin{aligned} E(|X_i|\,\mathbf{1}_B) &= E(|X_i|\,\mathbf{1}\{B, |X_i| \leq M\}) + E(|X_i|\,\mathbf{1}\{|X_i| > M\}) \\ &\leq M P(B) + \varepsilon/2 \leq \varepsilon \end{aligned}$$

whenever $P(B) \leq \delta_\varepsilon$. □

Proof of (b) implies (a). Let

$$\varepsilon > 0 \quad \text{and} \quad m = \sup_{i \in I} E\,|X_i| < \infty \quad \text{and} \quad M = \delta_\varepsilon / m.$$

Then, by Markov's inequality (3.1),

$$\sup_{i \in I} P(|X_i| > M) \leq (1/M) \sup_{i \in I} E\,|X_i| = m/M \leq \delta_\varepsilon$$

so the equi-integrability condition implies that

$$\sup_{i \in I} E(|X_i|\,\mathbf{1}\{|X_i| > M\}) \leq \varepsilon.$$

Since ε is arbitrary, this shows uniform integrability. □

Using the theorem, we can show the next result, which is the key to proving that a martingale converges in mean if and only if it is uniformly integrable.

Theorem 5.5. *The following two statements are equivalent.*

(a) $X_n \xrightarrow{p} X$ *and* (X_n) *is uniformly integrable.*

(b) $X_n \xrightarrow{L^1} X$.

Proof of (a) implies (b). Let $\varepsilon > 0$. Using Lemma 3.6 to get a subsequence (X_{n_k}) that converges almost surely to X and applying Fatou's lemma, we get

$$E\,|X| \leq \liminf_{k \to \infty} E|X_{n_k}| \leq \sup_n E|X_n| < \infty$$

where finiteness follows from the previous theorem. This shows that X is integrable so the sequence $(X_n - X)$ is uniformly integrable. In particular, this sequence is equi-integrable and since, invoking again convergence in probability,

$$P(|X_n - X| \geq \varepsilon) \leq \delta_\varepsilon \quad \text{for all } n \text{ large},$$

it follows from the previous theorem that

$$\begin{aligned} E\,|X_n - X| = E(|X_n - X|\,&\mathbf{1}\{|X_n - X| < \varepsilon\}) \\ &+ E(|X_n - X|\,\mathbf{1}\{|X_n - X| \geq \varepsilon\}) \leq \varepsilon + \varepsilon = 2\varepsilon \end{aligned}$$

for all n large. This shows that $X_n \xrightarrow{L^1} X$. □

Proof of (b) implies (a). Convergence in probability follows from Lemma 3.3. To prove uniform integrability, we use the result of Theorem 5.4. The fact that the sequence (X_n) is bounded in L^1 is clear. To show equi-integrability, we let $\varepsilon > 0$, and fix an integer $n_0 < \infty$ such that

$$E\,|X_n - X| \leq \varepsilon \quad \text{for all} \quad n \geq n_0.$$

Since n_0 is finite, there exists $\delta_\varepsilon > 0$ such that

$$P(B) \leq \delta_\varepsilon \quad \text{implies that} \quad E(|X_n|\,\mathbf{1}_B) \vee E(|X|\,\mathbf{1}_B) \leq \varepsilon \ \text{ for all } \ n \leq n_0.$$

This, together with the choice of n_0, implies that if $P(B) \leq \delta_\varepsilon$ then

$$E(|X_n|\,\mathbf{1}_B) \leq E\,|X_n - X| + E\,(|X|\,\mathbf{1}_B) \leq \varepsilon + \varepsilon = 2\varepsilon \quad \text{for all} \quad n > n_0,$$

showing that (X_n) is equi-integrable and so uniformly integrable. □

Specializing in martingales rather than arbitrary sequences of random variables and using the previous two theorems as well as the martingale convergence theorem, we deduce that uniform integrability is equivalent to convergence in mean. More precisely, we have the following result, which we state for supermartingales but obviously which also holds for martingales and submartingales.

Theorem 5.6. *For a supermartingale (X_n), the following are equivalent.*

(a) (X_n) is uniformly integrable.

(b) (X_n) converges almost surely and in mean.

(c) (X_n) converges in mean.

Proof of (a) implies (b). Assuming uniform integrability, Theorem 5.4 implies that the process is bounded in L^1; therefore the martingale convergence theorem implies that it converges almost surely. Now, according to Lemma 3.1, the process also converges in probability which, together with uniform integrability, gives convergence in mean according to Theorem 5.5. □

Proof of (b) implies (c). Obvious. □

Proof of (c) implies (a). This follows from Theorem 5.5. □

We now give two simple examples of uniformly integrable collections of random variables. The second example will motivate the definition of regular martingales.

Example 5.5. Let $p > 1$ and let (X_i) be bounded in L^p. Since $1 - p < 0$,

$$|X_i| \geq M > 0 \quad \text{implies that} \quad |X_i| \leq M^{1-p}\,|X_i|^p$$

from which it follows that

$$\begin{aligned} &\lim_{M\to\infty} \sup_{i\in I} E(|X_i| \mathbf{1}\{|X_i| > M\}) \\ &\quad\le \lim_{M\to\infty} \sup_{i\in I} M^{1-p} E(|X_i|^p \mathbf{1}\{|X_i| > M\}) \\ &\quad\le (\lim_{M\to\infty} M^{1-p})(\sup_{i\in I} E\,|X_i|^p) = 0 \end{aligned}$$

which shows uniform integrability. □

Example 5.6. Let $\mathbf{G}$ be a collection of sub-σ-algebras of $\mathscr{F}$. Then,

$$\{E(X\,|\,\mathscr{G}) : \mathscr{G} \in \mathbf{G}\} \quad \text{where} \quad X \in L^1(\Omega, \mathscr{F}, P)$$

is a collection of uniformly integrable random variables. To simplify the notation and make the proof more readable, for each random variable Y, we write

$$Y_{\mathscr{G}} = E(Y\,|\,\mathscr{G}) \quad \text{for all} \quad \mathscr{G} \in \mathbf{G}.$$

Let $\varepsilon > 0$. Since $X \in L^1$, there exists $\delta_\varepsilon > 0$ such that

$$P(B) \le \delta_\varepsilon \quad \text{implies that} \quad E(|X|\,\mathbf{1}_B) \le \varepsilon. \tag{5.9}$$

Taking $M > (1/\delta_\varepsilon) E\,|X|$, by Markov's inequality (3.1),

$$P(|X|_{\mathscr{G}} > M) \le (1/M) E(|X|_{\mathscr{G}}) = (1/M) E\,|X| \le \delta_\varepsilon. \tag{5.10}$$

Combining (5.9)–(5.10), we deduce that

$$\begin{aligned} &E(|X_{\mathscr{G}}|\,\mathbf{1}\{|X_{\mathscr{G}}| > M\}) \le E(|X|_{\mathscr{G}}\,\mathbf{1}\{|X_{\mathscr{G}}| > M\}) \\ &\qquad = E(|X|\,\mathbf{1}\{|X_{\mathscr{G}}| > M\}) \le E(|X|\,\mathbf{1}\{|X|_{\mathscr{G}} > M\}) \le \varepsilon \end{aligned}$$

which shows uniform integrability. □

Motivated by the previous example, we now define regular martingales. The martingale introduced in the proof of the Kolmogorov's zero-one law is an example of such a martingale. More generally, we have the following definition.

Definition 5.4. The process (X_n) is a **regular martingale** if

$$X_n = E(X\,|\,\mathscr{G}_n) \ \text{ for all } \ n \ge 0 \quad \text{for some} \quad X \in L^1(\Omega, \mathscr{F}, P)$$

where the sequence $(\mathscr{G}_n)$ is a filtration.

Since a filtration is nondecreasing, the process (X_n) reveals a little bit more about the random variable X at each time step. This is indeed a martingale since

$$E(X_{n+1}\,|\,\mathscr{G}_n) = E(E(X\,|\,\mathscr{G}_{n+1})\,|\,\mathscr{G}_n) = E(X\,|\,\mathscr{G}_n) = X_n.$$

According to Example 5.6, regular martingales are uniformly integrable. In fact, the class of regular martingales coincides with the class of uniformly integrable martingales, meaning in particular that if a martingale is uniformly integrable, then it has the very specific representation of a regular martingale. Combining this together with Theorem 5.6 gives the following result.

Theorem 5.7. *For a martingale (X_n), the following are equivalent.*

(a) (X_n) is uniformly integrable.
(b) (X_n) converges almost surely and in mean.
(c) (X_n) converges in mean.
(d) (X_n) is a regular martingale.

Proof. In view of Theorem 5.6 and Example 5.6, we only need to prove that one of the first three statements implies the last one. Assume, for example, convergence in mean to a random variable X. Then, from the definition of a martingale and conditional expectation, we have

$$E(X_n \mathbf{1}_A) = E(X_m \mathbf{1}_A) \quad \text{for all} \quad A \in \mathscr{G}_n \ \text{ and } \ m > n.$$

But since $X_m \xrightarrow{L^1} X$, we also have

$$\begin{aligned} &\lim_{m\to\infty} |E(X_m \mathbf{1}_A) - E(X \mathbf{1}_A)| \\ &\qquad \leq \ \lim_{m\to\infty} E\,|(X_m - X)\,\mathbf{1}_A| \leq \lim_{m\to\infty} E\,|X_m - X| = 0 \end{aligned}$$

therefore $E(X_n \mathbf{1}_A) = E(X\,\mathbf{1}_A)$ and so $X_n = E(X\,|\,\mathscr{G}_n)$ for all n. □

5.4 Reverse martingales and the law of large numbers

This section gives one more example of martingales, called reverse or backwards martingales, that turn out to be very useful in some contexts. For example, this concept allows us to relatively easily prove the strong law of large numbers in its full generality . Reverse martingales are indexed by the nonpositive integers: a process (X_{-n}) adapted to a filtration $(\mathscr{G}_{-n})$ is a **reverse martingale** if

$$E(X_{-n}\,|\,\mathscr{G}_{-n-1}) = X_{-n-1} \quad \text{for all} \quad n \geq 0.$$

From Theorem 5.7, we deduce the following useful result.

Theorem 5.8. *Almost surely and in mean,*

$$\lim_{n\to\infty} X_{-n} = X_{-\infty} = E(X_0\,|\,\mathscr{G}_{-\infty}) \quad \textit{where} \quad \mathscr{G}_{-\infty} = \bigcap_{n=0}^{\infty} \mathscr{G}_{-n}.$$

Proof. Since $\mathscr{G}_0 \supset \mathscr{G}_{-n}$ for all $n \geq 0$,

$$E(X_0\,|\,\mathscr{G}_{-n}) = X_{-n} \quad \text{for all} \quad n \geq 0$$

therefore (X_{-n}) is a regular martingale. In particular, convergence almost surely and in mean follows directly from Theorem 5.7. Calling $X_{-\infty}$ the limit,

$$\begin{aligned} E(X_0 \mathbf{1}_A) &= E(X_{-n} \mathbf{1}_A) = \lim_{n\to\infty} E(X_{-n} \mathbf{1}_A) \\ &= E(X_{-\infty} \mathbf{1}_A) \quad \text{for all} \quad A \in \mathscr{G}_{-\infty} \subset \mathscr{G}_{-m} \end{aligned}$$

which shows that $X_{-\infty} = E(X_0 \,|\, \mathscr{G}_{-\infty})$. □

To conclude, we now use the previous theorem along with Kolmogorov's zero-one law to prove the strong law of large numbers in its full generality, i.e., when the variance is not necessarily finite.

Theorem 5.9 (Strong law of large numbers). *Let (X_n) be independent identically distributed random variables with finite mean μ. Then,*

$$\frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu.$$

Proof. Let $Y_{-n} = S_n/n$ and

$$\mathscr{G}_{-n} = \sigma(S_n, S_{n+1}, S_{n+2}, \ldots) = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots).$$

Since the law of $(X_1, X_2, \ldots, X_n)$ is invariant by permutation,

$$(X_i, S_n, X_{n+1}, X_{n+2}, \ldots) = (X_j, S_n, X_{n+1}, X_{n+2}, \ldots) \quad \text{in distribution}$$

for all $i, j \leq n$, therefore $E(X_i \,|\, \mathscr{G}_{-n}) = E(X_j \,|\, \mathscr{G}_{-n})$. It follows that

$$E(X_1 \,|\, \mathscr{G}_{-n}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i \,|\, \mathscr{G}_{-n}) = E(S_n/n \,|\, \mathscr{G}_{-n}) = E(Y_{-n} \,|\, \mathscr{G}_{-n}) = Y_{-n}.$$

This implies that (Y_{-n}) is a reverse martingale so Theorem 5.8 gives

$$\lim_{n\to\infty} Y_{-n} = Y_{-\infty} = E(Y_{-1} \,|\, \mathscr{G}_{-\infty}) \quad \text{where} \quad \mathscr{G}_{-\infty} = \bigcap_{n=0}^{\infty} \mathscr{G}_{-n}$$

almost surely and in mean. In particular, the limit is $\mathscr{G}_{-\infty}$-measurable hence Kolmogorov's zero-one law implies that it is constant and so equal to its mean. Taking the expected value in the previous equations, we deduce

$$Y_{-\infty} = E(Y_{-\infty}) = E(E(Y_{-1} \,|\, \mathscr{G}_{-\infty})) = E(Y_{-1}) = E(X_1) = \mu$$

therefore $S_n/n = Y_{-n} \xrightarrow{a.s.} Y_{-\infty} = \mu$. □

5.5 Doob's inequality and convergence in L^p

To conclude, we look at martingales bounded in L^p for some $1 < p < \infty$ for which the convergence theory is even simpler. In this case, we already know that the process is uniformly integrable which, in turn, implies convergence almost surely and in mean to an integrable random variable X_∞. The main result of this section shows in fact that boundedness in L^p is enough to also have convergence in L^p. Note that this does not hold for $p = 1$. The first step to prove this result is Doob's inequality.

Theorem 5.10 (Doob's inequality). *Let (X_n) be a uniformly integrable submartingale. Then, for all $M > 0$*

$$M\,P(X_n^* \geq M) \leq E(|X_n|\,\mathbf{1}\{X_n^* \geq M\}) \quad \textit{where} \quad X_n^* = \max\nolimits_{i \leq n} |X_i|.$$

Proof. Define $T_n = \inf\{i : |X_i| \geq M \text{ or } i = n\}$ and observe that

$$|X_{T_n}| \geq M \quad \text{on the event} \quad \{X_n^* \geq M\}.$$

Since in addition $T_n \leq n$ and the process $|X_n|$ is a submartingale by Jensen's inequality, it follows from the optional stopping theorem that

$$\begin{aligned} M\,P(X_n^* \geq M) &= E(M\,\mathbf{1}\{X_n^* \geq M\}) \\ &\leq E(|X_{T_n}|\,\mathbf{1}\{X_n^* \geq M\}) \leq E(|X_n|\,\mathbf{1}\{X_n^* \geq M\}). \end{aligned}$$

This completes the proof. □

To deduce from Doob's inequality that submartingales bounded in L^p also converge in the Lebesgue space L^p, we need the following lemma.

Lemma 5.2. *Let X and Y be two positive random variables such that*

$$M\,P(X > M) \leq E(Y\,\mathbf{1}\{X > M\}) \quad \textit{for all} \quad M > 0.$$

Then, for all $1 < p < \infty$,

$$\|X\|_p \leq q\,\|Y\|_p \quad \textit{where} \quad q = p/(p-1)$$

is the exponent conjugate to p.

Proof. By Fubini's theorem,

$$E(X^p) = \int_0^\infty P(X^p > t)\,dt.$$

Setting $t = s^p$, we deduce that

$$E(X^p) = \int_0^\infty P(X > s)\,d(s^p) \leq \int_0^\infty (1/s)\,E(Y\,\mathbf{1}\{X > s\})\,d(s^p)$$

and, using again Fubini's theorem, we get

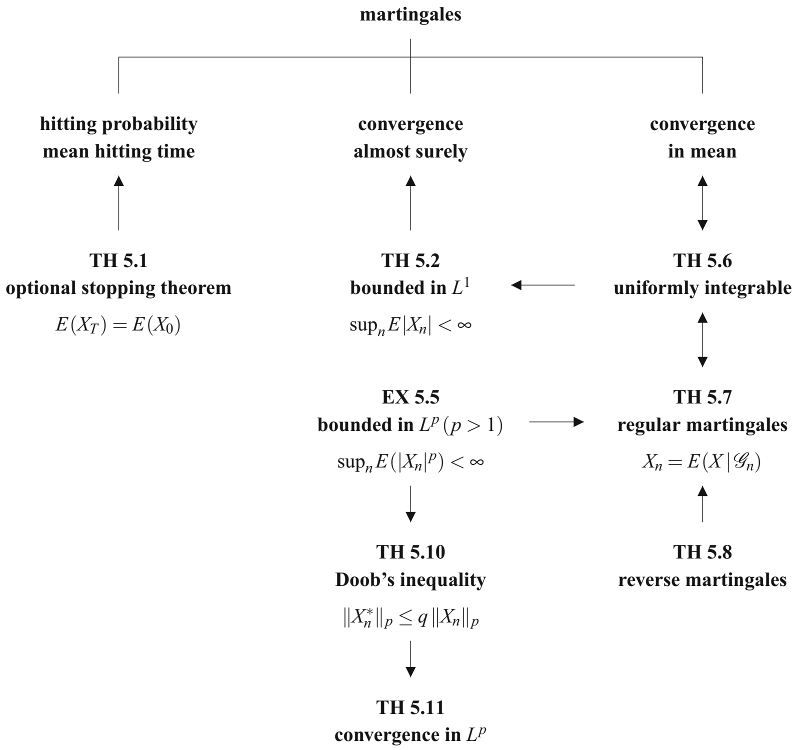


Fig. 5.2 Summary of Chapter 5.

$$\begin{aligned} E(X^p) &\leq \int_0^\infty \int_\Omega (1/s)\,Y\,\mathbf{1}\{X > s\}\,dP\,d(s^p) = \int_\Omega \int_0^X (1/s)\,Y\,d(s^p)\,dP \\ &= E\left(Y \int_0^X (1/s)\,d(s^p)\right) = \left(\frac{p}{p-1}\right) E(YX^{p-1}) = qE(YX^{p-1}). \end{aligned}$$

Then, using Hölder's inequality (1.9) and $q(p-1) = p$,

$$\begin{aligned} E(X^p) &\leq q\,\|Y\|_p\,\|X^{p-1}\|_q = q\,\|Y\|_p\,E(X^{q(p-1)})^{1/q} \\ &= q\,\|Y\|_p\,E(X^p)^{1/q} = q\,\|Y\|_p\,E(X^p)^{1-1/p} \end{aligned}$$

therefore $\|X\|_p \leq q\,\|Y\|_p$ and the proof is complete. □

By Doob's inequality, $X = X_n^*$ and $Y = |X_n|$ satisfy the assumptions of the lemma whenever (X_n) is a uniformly integrable submartingale. In particular,

$$\|X_n^*\|_p \leq q\,\|X_n\|_p \quad \text{where} \quad 1/p + 1/q = 1. \tag{5.11}$$

Using this inequality, we can now prove the main result of this section.

Theorem 5.11. *For a submartingale (X_n), the following are equivalent.*

(a) (X_n) is bounded in $L^p(\Omega, \mathscr{F}, P)$,

(b) (X_n) is uniformly integrable and its limit $X_\infty \in L^p(\Omega, \mathscr{F}, P)$,

(c) $X^* = \sup_n |X_n| \in L^p(\Omega, \mathscr{F}, P)$,

and in each case $X_n \to X_\infty$ almost surely and in $L^p(\Omega, \mathscr{F}, P)$.

Proof of (a) implies (b). The fact that the process is uniformly integrable follows from Example 5.5. This also implies that the process converges almost surely and in mean to an integrable random variable X_∞ according to Theorem 5.6. In addition, applying Fatou's lemma, we obtain

$$\begin{aligned} E(|X_\infty|^p) &= E(\liminf_{n\to\infty} |X_n|^p) \\ &\leq \liminf_{n\to\infty} E(|X_n|^p) \leq \sup_n(|X_n|^p) < \infty \end{aligned}$$

showing that $X_\infty \in L^p(\Omega, \mathscr{F}, P)$. □

Proof of (b) implies (c). Since (X_n^*) is nondecreasing and converges to X^*, the monotone convergence theorem and inequality (5.11) imply that

$$\|X^*\|_p = \lim_{n\to\infty} \|X_n^*\|_p \leq q \lim_{n\to\infty} \|X_n\|_p \leq q \sup_n \|X_n\|_p < \infty$$

showing that $X^* \in L^p(\Omega, \mathscr{F}, P)$. □

Proof of (c) implies (a). We have

$$\sup_n E(|X_n|^p) \leq E((\sup_n |X_n|)^p) < \infty$$

showing that the process is bounded in $L^p(\Omega, \mathscr{F}, P)$. □

Proof of convergence in L^p. We have already proved that

$$X_n \xrightarrow{a.s.} X_\infty \in L^p(\Omega, \mathscr{F}, P) \quad \text{and} \quad X^* = \sup_n |X_n| \in L^p(\Omega, \mathscr{F}, P).$$

In particular, we have

$$|X_n - X_\infty|^p \xrightarrow{a.s.} 0 \quad \text{and} \quad |X_n - X_\infty|^p \leq 2\,(X^*)^p \in L^1(\Omega, \mathscr{F}, P)$$

so the result follows from the dominated convergence theorem. □

Even though the previous theorem is proved for submartingales, the result also holds for supermartingales, and obviously for martingales as well. This theorem and the other main results of this chapter are summarized in the diagram of Figure 5.2. We conclude with an example of the application of the previous theorem.

Example 5.7 (Random harmonic series). It is known that

$$\lim_{n\to\infty} 1/1 + 1/2 + \cdots + 1/n \geq \int_1^\infty \frac{dx}{x} = \infty.$$

However, letting $U_i \sim \text{Uniform}\{-1,1\}$ be independent and using Theorem 5.11, one can prove that the random harmonic series

$$X_n = (1/1)\,U_1 + (1/2)\,U_2 + \cdots + (1/n)\,U_n$$

converges almost surely to a square-integrable random variable X_∞. To prove this result, we first observe that

$$\begin{aligned} E(X_n \,|\, \mathscr{F}_{n-1}) &= E(X_{n-1} + (1/n)\,U_n \,|\, \mathscr{F}_{n-1}) \\ &= E(X_{n-1} \,|\, \mathscr{F}_{n-1}) + (1/n)\,E(U_n \,|\, \mathscr{F}_{n-1}) \\ &= X_{n-1} + (1/n)\,E(U_n) = X_{n-1} \end{aligned}$$

showing that (X_n) is a martingale. This can also be proved by observing that the process is the betting random walk of Example 4.4 with $B_n = 1/n$. In addition, since the uniform random variables U_i are centered and independent,

$$\begin{aligned} E(X_n^2) &= \text{Var}(X_n) = \text{Var}\left(\sum_{i=1}^n \frac{U_i}{i}\right) \\ &= \sum_{i=1}^n \left(\frac{1}{i}\right)^2 \text{Var}(U_i) = \sum_{i=1}^n \left(\frac{1}{i}\right)^2 \leq \frac{\pi^2}{6}. \end{aligned}$$

This shows that (X_n) is bounded in L^2 which implies convergence in L^2 and almost surely to a square-integrable random variable. □

5.6 Exercises

Special martingales

Exercise 5.1. Letting (X_n) be a sequence of independent random variables with the same mean μ, prove that the following processes are martingales

$$S_n = X_1 + \cdots + X_n - n\mu \quad \text{and} \quad \pi_n = \mu^{-n} X_1 X_2 \cdots X_n \ \text{ when } \ \mu \neq 0.$$

Exercise 5.2 (Wald's first equation). Assume that the X_n in Exercise 5.1 are identically distributed and let T be an integrable stopping time.

1. Prove that the conditional increments of (S_n) are bounded in L^1.
2. Deduce Wald's equation: $E(X_1 + \cdots + X_T) = E(T)\,E(X_1)$.

Exercise 5.3 (Wald's second equation). Let (X_n) be a sequence of independent and identically distributed random variables with mean zero and variance σ^2 and let T be an integrable stopping time. The goal of this problem is to show that

$$E((X_1 + X_2 + \cdots + X_T)^2) = \sigma^2 E(T).$$

1. Prove that the increments of a martingale (S_n) are orthogonal, i.e.,

$$E[(S_l - S_k)(S_j - S_i)] = 0 \quad \text{for all} \quad i \leq j \leq k \leq l.$$

2. Deduce that

$$E[(S_n - S_0)^2] = \sum_{k=1}^{n} E[(S_k - S_{k-1})^2].$$

3. Use Exercise 5.1 to conclude that

$$E((X_1 + X_2 + \cdots + X_T)^2) = \sigma^2 E(T).$$

Exercise 5.4 (Exponential martingale). Let (X_n) be a sequence of independent and identically distributed random variables and let

$$\phi(\theta) = E(e^{\theta X_i}) < \infty \quad \text{and} \quad S_n = X_1 + X_2 + \cdots + X_n.$$

Show that the process (Z_n) where $Z_n = e^{\theta S_n}/\phi(\theta)^n$ is a martingale.

Exercise 5.5. Letting (X_n) be a sequence of independent and identically distributed random variables with mean zero and finite variance σ^2, prove that the following process is a martingale

$$Y_n = S_n^2 - \sigma^2 n \quad \text{where} \quad S_n = X_1 + X_2 + \cdots + X_n.$$

Optional stopping theorem

Exercise 5.6. Consider a particle moving on the set of the integers as follows: At each time step, the particle jumps

- one unit to the right with probability $p > 1/2$ or
- one unit to the left with probability $q = 1 - p < 1/2$.

Let X_n be the position of the particle at time n.

1. Prove that there is $a < 1$ to be determined such that (a^{X_n}) is a martingale.
2. Find the probability that the process starting at $x \in (0, N)$ hits N before 0.
3. Conclude that the probability that the particle starting at position $x > 0$ never visits site zero is given by

$$P_x(X_n > 0 \text{ for all } n \geq 0) = 1 - (q/p)^x.$$

Exercise 5.7. Assuming that two players A and B play until they are N points apart and that each point is independently won by A with probability p, find the probability that player A has more points than B at the end of the game.

Hint: Use the probability of winning in the gambler's ruin chain.

Exercise 5.8 (Duration of fair games). Let T be the number of games played in the gambler's ruin chain when the probability of winning a game is one-half and let X_n be the gambler's fortune after n games.

1. Prove that the process $(X_n^2 - n)$ stopped at time T is a martingale.
2. Deduce that $E(T) = x(N-x)$ when starting with x dollars.

Hint: Use that the probability of winning is x/N.

Exercise 5.9 (Duration of unfair games). Let T be the number of games played in the gambler's ruin chain when the probability of winning a game is $p \neq 1/2$ and let X_n be the gambler's fortune after n games.

1. Prove that $(X_n - n(p-q))$ stopped at time T where $q = 1-p$ is a martingale.
2. Deduce that, when starting with x dollars,

$$E(T) = \left(\frac{x}{q-p}\right) - \left(\frac{N}{q-p}\right)\left(\frac{1-c^x}{1-c^N}\right) \quad \text{where} \quad c = q/p.$$

Hint: Use that the probability of winning is $(1-c^x)/(1-c^N)$.

Exercise 5.10. Three players start with a, b, and c coins, respectively. At each time step, a randomly chosen player gives one of her coins to another randomly chosen player. When one of the players runs out of coins, only the two players left keep playing and the game stops when one player has all the coins.

1. Let X_n, Y_n and Z_n be the number of coins each player has at step n and T be the total number of steps until one player has all the coins. Prove that

$$M_n = X_{n\wedge T}\, Y_{n\wedge T} + X_{n\wedge T}\, Z_{n\wedge T} + Y_{n\wedge T}\, Z_{n\wedge T} + n \wedge T$$

   defines a martingale.
2. Deduce the expected number of steps until one player has all the coins.

Exercise 5.11 (Matching rounds problem). Referring to Exercise 2.22 about the matching rounds problem, let

$$\begin{aligned} R_n &= \text{number of rounds until all the couples have been paired} \\ X_i &= \text{number of matches at round } i = 1, 2, \ldots, R_n. \end{aligned}$$

1. Prove that the process (Z_k) is a martingale where

$$Z_k = (X_1 - 1) + (X_2 - 1) + \cdots + (X_{k \wedge R_n} - 1).$$

2. Deduce that $E(R_n) = n$.

Exercise 5.12. Rolling a fair six-sided die continually, find the expected value of the number of rolls until the pattern 121 appears twice in a row, i.e., letting X_i be the outcome of the ith roll, find the expected value of

$$T = \inf\{n : (X_{n-5},\dots,X_n) = (1,2,1,1,2,1)\}.$$

Hint: Use Example 5.4.

Exercise 5.13. A coin with probability p of landing on heads is continually flipped. Find the expected value of the time until the pattern HTH appears, i.e., letting X_i be the outcome of the ith flip, find the expected value of

$$S = \inf\{n : (X_{n-2},X_{n-1},X_n) = (H,T,H)\}.$$

Hint: Use the same approach as in Example 5.4 to find a martingale.

Exercise 5.14 (The ballot problem). This exercise gives an alternative proof of the ballot problem in Exercise 2.15 using martingales. Let S_k be the difference between the number of votes for A and the number of votes for B after k votes.

1. Prove that $X_k = (n-k)^{-1} S_{n-k}$ where $n = a+b$ is a martingale.
2. Use the stopping time

   $$T = \inf\{k : X_k = 0\} \wedge (n-1)$$

   to deduce that $p(a,b) = (a-b)/(a+b)$.

Exercise 5.15 (Voter model). Let $G = (V,E)$ be a finite connected graph and consider the following discrete-time version of the so-called voter model on this graph. Each vertex is occupied by an individual who has opinion 0 or opinion 1. At each time step, an edge is chosen uniformly at random, then one of the two vertices again chosen uniformly at random mimics the opinion of the other vertex. Consider the discrete-time process (Z_n) defined as

$$Z_n : V \to \{0,1\} \quad \text{where} \quad Z_n(x) = \text{opinion of vertex } x \text{ at time } n.$$

Let its natural filtration be $(\mathscr{G}_n)$ and consider the new process (X_n) that keeps track of the number of individuals with opinion 1 at time n.

1. Prove that (X_n) is a martingale with respect to the filtration $(\mathscr{G}_n)$.
2. Show that the probability that all that individuals have opinion 1 eventually equals the initial fraction of individuals with that opinion.

Convergence of the martingales

Exercise 5.16 (Pólya's urn process). An urn contains initially (time 0) one black ball and one white ball. At each time step, we choose a ball at random and then replace it in the urn along with one more ball of the color drawn. Let X_n be the proportion of black balls after the nth draw.

1. Prove that (X_n) is a martingale.
2. Deduce almost sure convergence to $X_\infty \sim \text{Uniform}\,(0,1)$.

Exercise 5.17. Assume that, at each step of a game, a gambler is equally likely to either win or lose one dollar. Prove that, if the gambler keeps playing as long as she has some money, she will lose all her money eventually.

Exercise 5.18. At each step of a game, a player picks a number uniformly between zero and one and her fortune is multiplied by twice this number.

1. Prove that the gambler's fortune X_n at step n defines a martingale.
2. Find the distribution of the random variable

$$-\ln(U) \quad \text{where} \quad U \sim \text{Uniform}\,(0,1).$$

3. Use the law of large numbers to study $\ln(X_n)$ and deduce that, though the game is fair, the fortune of the player goes to zero almost surely.

Chapter 6
Branching processes

Branching processes, also called Galton–Watson processes after the name of their inventors (Francis Galton and Henry William Walton), are some of the simplest stochastic processes describing single-species dynamics. Such processes ignore the presence of a spatial structure but include general offspring distributions. The first reference to these processes is [39].

The first section gives a precise description of the model. The main assumption is that the number of offspring produced by each individual is independent and identically distributed across individuals and generations, which implies that the number of individuals evolves according to a discrete-time Markov chain.

The second section exhibits some useful connections with martingales suggesting that the process experiences a phase transition when the expected value μ of the offspring distribution is equal to one. From these connections, we prove exponentially fast extinction when $\mu < 1$ and, more interestingly, almost sure extinction in the critical case $\mu = 1$ using the martingale convergence theorem.

When the expected number of offspring $\mu > 1$, referred to as the supercritical regime, the population survives in the long run with positive probability. The third section proves this result by giving an implicit expression of the probability of survival involving the probability-generating function of the offspring distribution.

In the last section, we prove the so-called law of total variance and study the evolution of the mean and variance of the number of individuals.

Further reading

- The first classical reference on Galton–Watson processes and more sophisticated branching processes is Harris [42].
- For additional results, see Athreya and Ney [2] as well as Jagers [48] which also has an emphasis on biological applications of branching processes.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_6

6.1 Model description

Branching processes are discrete-time processes (X_n) that were originally introduced to model the survival of a family name. In this context, rather than time, the index n refers to the generation and X_n to the number of individuals from generation n. Here, generation can be defined inductively by assuming that the offspring of an individual from generation n is by definition an individual from generation $n+1$, so individuals from the same generation may not live at the same time. One can also think of the index n as an actual time when looking at some species with a fixed life cycle. The model relies on the following two simple assumptions.

- All the individuals produce a random number of offspring according to a fixed distribution, say k offspring with probability p_k.
- The random numbers of offspring produced by the different individuals at the same or at different generations are independent.

Mathematically, the process (X_n) is a discrete-time Markov chain with the state space the set of all nonnegative integers. To define the dynamics, we let $(Y_{n,i})$ be a collection of independent identically distributed random variables with

$$P(Y_{n,i} = k) = p_k \quad \text{for all} \quad n,i \in \mathbb{N}.$$

Thinking of $Y_{n,i}$ as the number of offspring produced by individual i at generation n if there are indeed at least i individuals at that generation,

$$X_{n+1} = Y_{n,1} + Y_{n,2} + \cdots + Y_{n,X_n}. \tag{6.1}$$

Since the random variables $Y_{n,i}$ are independent and X_{n+1} can be written using these random variables and X_n, the process is indeed a Markov chain. For realizations of branching processes, we refer the reader to Figure 6.1.

6.2 Connection with martingales

To study branching processes, the first step is to exhibit some connections with martingales, which is a powerful tool to understand their behavior. To begin with, we note that, according to (6.1) and Lemma 2.3,

$$E(X_{n+1} \,|\, \mathscr{F}_n) = E(X_{n+1} \,|\, X_n) = E(Y_{n,1} + \cdots + Y_{n,X_n} \,|\, X_n) = \mu X_n$$

where $\mu = E(Y_{n,i})$. This implies that

$$E((1/\mu)^{n+1} X_{n+1} \,|\, \mathscr{F}_n) = \mu\,(1/\mu)^{n+1} X_n = (1/\mu)^n X_n$$

indicating that the process $((1/\mu)^n X_n)$ is a martingale. This suggests that the process exhibits three types of behaviors depending on whether μ is smaller than, equal to, or larger than one. These three types of behaviors are referred to, respectively, as subcritical, critical, and supercritical regimes.

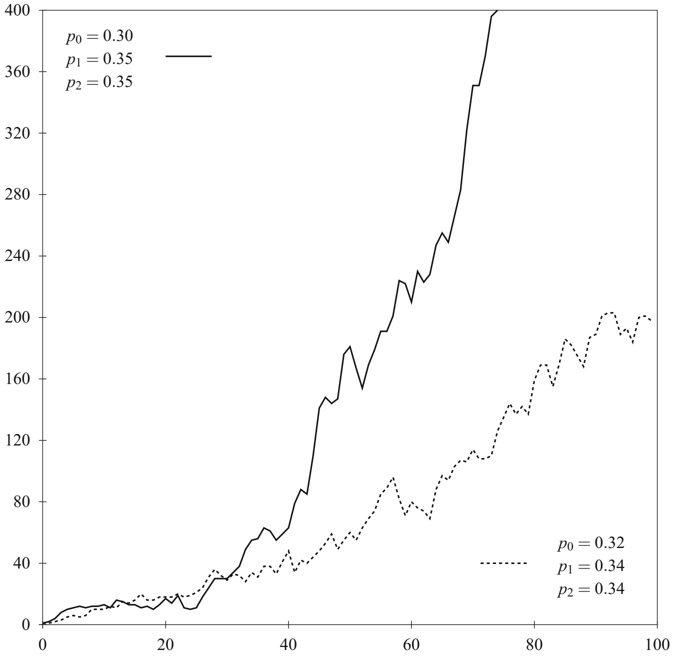


Fig. 6.1 Two realizations of branching processes with different offspring distribution.

In the subcritical regime $\mu < 1$, starting with one individual,

$$P(X_n > 0) = \sum_{k=1}^{\infty} P(X_n = k) \leq \sum_{k=1}^{\infty} kP(X_n = k) = E(X_n) = \mu^n$$

so the population/family name goes extinct exponentially fast.

In the critical case, we have the following result.

Theorem 6.1. *Let* $p_1 \neq 1$ *and* $\mu = 1$. *Then* $\lim_{n\to\infty} P(X_n = 0) = 1$.

Proof. Since the process (X_n) is a martingale with $E|X_n| = E(X_0) = 1$, the martingale convergence theorem implies that

$$X_n \xrightarrow{a.s.} X_\infty \quad \text{for some} \quad X_\infty \in L^1(\Omega, \mathscr{F}, P).$$

In addition, the assumption $p_1 \neq 1$ implies

$$P(\lim_{n\to\infty} X_n = k) = 0 \quad \text{for all} \quad k > 0.$$

This implies that $X_\infty = 0$ almost surely and proves the theorem. □

Note that the case $p_1 = \mu = 1$, which is excluded in the statement of the theorem, is in fact trivial: Each individual produces exactly one offspring so the process is

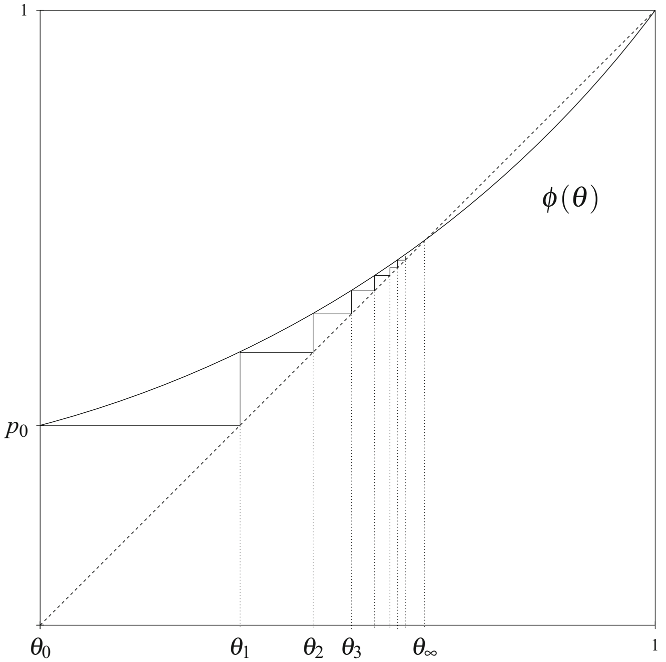


Fig. 6.2 Graph of the probability generating function when $\mu > 1$.

almost surely constant. Excluding this exceptional case, the fact that the population/family name dies out at criticality is common for stochastic processes, but this is sometimes difficult to prove. As we will see later, this is also the case for simple birth and death processes, bond percolation, and the contact process.

The supercritical regime is a little more difficult to study. This is done in the next section using the probability generating function of the offspring distribution.

6.3 Probability of survival

In both the subcritical and critical cases, the population/family name goes extinct almost surely. We now prove that the survival probability starting with one individual is strictly positive in the supercritical case $\mu > 1$ and express its value as a function of the offspring distribution using the probability generating function.

Theorem 6.2. *Let $\mu > 1$. Then $P(X_n > 0$ for all $n > 0) > 0$.*

Proof. The probability generating function of $Y_{n,i}$ is given by

$$\phi(\theta) = E(\theta^{Y_{n,i}}) = \sum_{k=0}^{\infty} p_k \theta^k \quad \text{for all} \quad \theta \in [0,1].$$

Letting $\theta_n = P(X_n = 0 \,|\, X_0 = 1)$, conditioning on the number of individuals at the first generation, and using independence, we get

$$\begin{aligned} \theta_{n+1} &= \textstyle\sum_{k \geq 0} P(X_{n+1} = 0 \,|\, X_1 = k) \, P(X_1 = k \,|\, X_0 = 1) \\ &= \textstyle\sum_{k \geq 0} p_k \, [P(X_{n+1} = 0 \,|\, X_1 = 1)]^k \\ &= \textstyle\sum_{k \geq 0} p_k \, \theta_n^k = \phi(\theta_n). \end{aligned}$$

In particular, observing that, because of monotonicity,

$$\theta_\infty = P(X_n = 0 \text{ for some } n \geq 0) = \lim_{n \to \infty} P(X_n = 0) = \lim_{n \to \infty} \theta_n$$

we deduce that $\phi(\theta_\infty) = \theta_\infty$, i.e., the probability of extinction is a fixed point of the probability generating function. Now, since $\mu > 1$, we must have $p_k > 0$ for some integer $k \geq 2$, from which it follows that

$$\phi'(\theta) = \textstyle\sum_{k \geq 0} k \, p_k \, \theta^{k-1} > 0 \quad \text{and} \quad \phi''(\theta) = \textstyle\sum_{k \geq 0} k \, (k-1) \, p_k \, \theta^{k-2} > 0$$

for all $\theta > 0$. Observing in addition that

$$\phi(0) = p_0 < 1 \quad \text{and} \quad \phi'(1) = \textstyle\sum_{k \geq 0} k \, p_k = E(Y_{n,i}) = \mu > 1,$$

we deduce that the probability generating function is strictly increasing and convex and has a derivative strictly larger than one at one. In particular, we obtain the graph of Figure 6.2, from which it follows that the probability generating function has a unique fixed point in the open interval $(0,1)$. Since $\theta_0 = 0$, the sequence (θ_n) converges to this fixed point, showing that the probability of extinction is less than one and the probability of survival strictly positive. □

The proof of Theorem 6.2 shows more generally that

$$P(X_n > 0 \text{ for all } n > 0) = 1 - \theta_\infty > 0 \quad \text{when} \quad \mu > 1$$

where θ_∞ is the unique fixed point in the open unit interval of the probability generating function. This gives an implicit expression of the survival probability as a function of the coefficients p_k.

6.4 Mean and variance of the number of individuals

To conclude this chapter, we study the expected value and the variance of the number of individuals for the process starting with one individual. These two quantities depend on their counterpart for the offspring distribution, i.e., the mean and variance of the offspring distribution

$$\mu = E(Y_{n,i}) \quad \text{and} \quad \sigma^2 = \mathrm{Var}(Y_{n,i}). \tag{6.2}$$

Recalling that the process $((1/\mu)^n X_n)$ is a martingale, we have

$$E(X_n) = \mu\, E(X_{n-1}) = \mu^2\, E(X_{n-2}) = \cdots = \mu^n\, E(X_0) = \mu^n$$

showing an exponential increase, then respectively, decrease of the mean number of individuals at generation n in the supercritical, then respectively, subcritical case, while the mean number of individuals stays constant in the critical case.

The variance of the number of individuals at generation n is more difficult to compute. To obtain the variance, the first step is to prove the following result known as the **law of total variance**.

Lemma 6.1 (Law of total variance). *Let X,Y be two random variables and assume that X has finite variance. Then,*

$$\mathrm{Var}\,(X) = E(\mathrm{Var}\,(X\,|\,Y)) + \mathrm{Var}\,(E(X\,|\,Y))$$

where by definition $\mathrm{Var}\,(X\,|\,Y) = E((X - E(X\,|\,Y))^2\,|\,Y)$.

Proof. To begin with, we observe that

$$\begin{aligned} E\,(\mathrm{Var}\,(X\,|\,Y)) &= E\,(E(X^2 - 2XE(X\,|\,Y) + E(X\,|\,Y)^2\,|\,Y)) \\ &= E\,(E(X^2\,|\,Y) - E\,(X\,|\,Y)^2) \\ &= E(X^2) - E\,(E\,(X\,|\,Y)^2). \end{aligned} \tag{6.3}$$

Also, by definition of the unconditional variance,

$$\begin{aligned} \mathrm{Var}\,(E(X\,|\,Y)) &= E\,((E(X\,|\,Y) - E(X))^2) \\ &= E\,(E(X\,|\,Y)^2 - 2E(X\,|\,Y)\,E(X) + E(X)^2) \\ &= E\,(E(X\,|\,Y)^2) - E(X)^2. \end{aligned} \tag{6.4}$$

Adding (6.3)–(6.4) gives $E(X^2) - E(X)^2 = \mathrm{Var}\,(X)$. □

We can now compute the variance of the number of individuals.

Lemma 6.2. *Assume* (6.2). *Then,*

$$\mathrm{Var}\,(X_n) = \begin{cases} n\,\sigma^2 & \text{when} \quad \mu = 1 \\ \sigma^2\,\mu^{n-1}\,(1-\mu^n)/(1-\mu) & \text{when} \quad \mu \neq 1. \end{cases}$$

Proof. By independence, we have

$$\begin{aligned} \mathrm{Var}\,(X_n\,|\,X_{n-1}) &= \mathrm{Var}\,(Y_{n-1,1} + \cdots + Y_{n-1,X_{n-1}}\,|\,X_{n-1}) \\ &= X_{n-1}\ \mathrm{Var}\,(Y_{n-1,i}) = \sigma^2 X_{n-1}. \end{aligned}$$

This and the law of total variance give

$$\begin{aligned} \mathrm{Var}\,(X_n) &= E(\mathrm{Var}\,(X_n\,|\,X_{n-1})) + \mathrm{Var}\,(E(X_n\,|\,X_{n-1})) \\ &= E(\sigma^2 X_{n-1}) + \mathrm{Var}\,(\mu\, X_{n-1}) = \sigma^2\,\mu^{n-1} + \mu^2\ \mathrm{Var}\,(X_{n-1}) \end{aligned}$$

and using a simple induction,

$$\begin{aligned}\text{Var}(X_n) &= \sigma^2 \mu^{n-1} + \mu^2 \text{Var}(X_{n-1}) = \sigma^2 (\mu^{n-1} + \mu^n) + \mu^4 \text{Var}(X_{n-2}) \\ &= \sigma^2 (\mu^{n-1} + \mu^n + \mu^{n+1} + \cdots + \mu^{2n-2}) + \mu^{2n} \text{Var}(X_0) \\ &= \sigma^2 \mu^{n-1} (1 + \mu + \cdots + \mu^{n-1}).\end{aligned}$$

Distinguishing whether $\mu = 1$ or $\mu \neq 1$ gives the result. □

Note that for supercritical branching processes the expected value increases exponentially fast, but the variance increases even faster. This is due to the fact that small variations at some generation intensify as time evolves. For example, only one more individual at generation one results on average in μ^{n-1} additional individuals at generation n. In contrast, for subcritical branching processes, both the expected value and the variance decrease to zero exponentially fast, which is consistent with the fact that the population dies out quickly.

6.5 Exercises

Exercise 6.1. Consider the branching process starting with one individual for which the offspring distribution is given by

$$p_0 = 1/6, \quad p_1 = 1/3, \quad p_2 = 1/2 \quad \text{where} \quad p_k = P(Y_{n,i} = k).$$

Find the probability of survival of this process. For numerical simulations of this process, see Program 2 at the end of this book.

Exercise 6.2. Consider the branching process starting with one individual for which the offspring distribution follows the shifted geometric distribution

$$p_k = (1-p)^k p \quad \text{for all} \quad k \geq 0.$$

Find the probability of survival of this process.

Exercise 6.3. Let $N \in \mathbb{N}^*$ and consider the branching process starting with one individual for which the offspring distribution is given by

$$p_0 = 1 - 1/\sqrt{N} \quad \text{and} \quad p_N = 1/\sqrt{N}.$$

1. Give the expected number of offspring per individual.
2. Find an upper bound for the survival probability of the process.
3. Is the survival probability of a branching process always nondecreasing with respect to the expected number of offspring per individual?

Exercise 6.4. Find the mean number of individuals that ever existed in a subcritical branching process, i.e., $\mu < 1$, starting with a single individual.

Chapter 7
Discrete-time Markov chains

The first mathematical results for discrete-time Markov chains with a finite state space were generated by Andrey Markov [71] whose motivation was to extend the law of large numbers to sequences of nonindependent random variables. The first important results in the more difficult context of countably infinite Markov chains were established 30 years later by Andrey Kolmogorov [57].

Recall that discrete-time stochastic processes are Markovian if the future of the process depends on the past only through the present. All the discrete-time processes discussed in this textbook and a number of models that arise from physics, biology, and sociology have this property. This chapter gives the main techniques to study discrete-time Markov chains.

To begin with, we show that a Markov chain can be represented as a matrix, the so-called transition matrix, or viewed as a directed graph. The latter is more suitable to study qualitative aspects of the process, whereas the former is used for computational purposes. In particular, the probability that the process goes from one state to another state in n time steps can be determined by computing the nth power of the transition matrix.

The main objective of this chapter is to use the transition matrix and the directed graph representation and introduce some theoretical tools to also answer the following two important questions.

Question 1 — Does the fraction of time spent in a given state converge to a limit as time goes to infinity, and if this is the case, what is the limit?

Question 2 — Does the probability of being in a given state converge to a limit as time goes to infinity, and if this is the case, what is the limit?

The answer to these questions requires a couple of steps.

First, we show that the state space of the process can be partitioned into so-called communication classes. This breaks down the process into elementary pieces so we can focus our attention on irreducible Markov chains, the ones with only one communication class. We can also classify the states as recurrent, meaning that the state is visited infinitely often, or transient, meaning that the state is visited only a

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_7

finite number of times. States in the same communication class, and therefore states of an irreducible Markov chain, are either all recurrent or all transient. When all the states are transient, the two limits above exist and are trivially equal to zero while the case of recurrent irreducible Markov chains is more complicated.

To study recurrent processes, we introduce the concept of stationary distributions, which are distributions on the state space that are invariant under the dynamics. We also explain how they can be computed in theory using the transition matrix, and in practice through a couple of examples including the case of doubly stochastic matrices and time-reversible processes. Stationary distributions are important because they give the value of the two limits above when they exist.

Recurrent states are called null recurrent states if the expected time between consecutive visits in the state is infinite, otherwise they are called positive recurrent states. Again, the states of a recurrent irreducible Markov chain are either all null recurrent or all positive recurrent. For null recurrent processes, the conclusion is the same as for transient processes: the two limits above exist and are equal to zero. In contrast, we show that an irreducible Markov chain has a unique stationary distribution if and only if it is positive recurrent, in which case the fraction of time spent in a given state is equal to the mass of this state under the stationary distribution which is also equal to the reciprocal of the expected time of return to this state.

Finally, we prove that, to also have convergence of the probability of being in a given state, we need one more ingredient called aperiodicity, which can be more easily checked using the directed graph representation. This property basically prevents the probability of being in a given state from being periodic.

Further reading

Of all the topics covered in this textbook, discrete-time Markov chain is undoubtedly the most common, so the literature is copious.

- Some of the early references in this topic are [12, 34].
- We also refer the reader to [28, 31, 79] for a presentation at the graduate level.
- For a presentation at the undergraduate level, see [29, 49, 85, 87, 89].
- For a brief review with a number of exercises with solutions, see [3].

7.1 Multi-step transition probabilities

To begin, we recall that the process (X_n) is a discrete-time Markov chain with finite or countable state space S whenever

$$P(X_{n+1} \in B \,|\, \mathscr{F}_n) = P(X_{n+1} \in B \,|\, X_n) \quad \text{for all} \quad B \subset S. \tag{7.1}$$

In particular, the evolution is defined by the **transition probabilities**

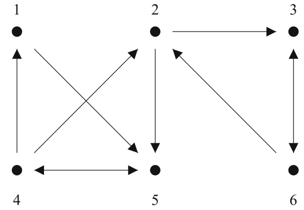
$$p(x,y) = P(X_{n+1} = y \,|\, X_n = x) \quad \text{for all} \quad x,y \in S$$

that we assume to be the same for all n, which is referred to as a time homogeneous Markov chain.

1. **Matrix representation** — One can think of $p(x,y)$ as the coefficient on row x and column y of a matrix P, called the **transition matrix**. The representation in the matrix form is a key tool to compute the so-called multistep transition probabilities and stationary distributions.
2. **Graph representation** — One can also think of the state space as the vertex set of a directed graph and draw an arrow $x \to y$ if and only if $p(x,y) > 0$. Even though this representation ignores the exact value of the transition probabilities, it is more convenient to find the so-called communication classes and the period of each class, both of which are introduced below.

Here is an example showing how to construct the oriented graph from the transition matrix of a discrete-time Markov chain:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & .3 & 0 & .7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ .2 & .5 & 0 & 0 & .3 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & .4 & .6 & 0 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 0 & 0 & \bullet & 0 \\ 0 & 0 & \bullet & 0 & \bullet & 0 \\ 0 & 0 & 0 & 0 & 0 & \bullet \\ \bullet & \bullet & 0 & 0 & \bullet & 0 \\ 0 & 0 & 0 & \bullet & 0 & 0 \\ 0 & \bullet & \bullet & 0 & 0 & 0 \end{pmatrix}$$



Probabilities of interest are the **multistep transition probabilities**

$$p_n(x,y) = P(X_n = y \,|\, X_0 = x) \quad \text{for all} \quad n \in \mathbb{N} \text{ and } x,y \in S.$$

By conditioning on the state at time n to beak down the $(n+m)$-step transition probabilities, one obtains the **Chapman–Kolmogorov's equations**

$$\begin{aligned} p_{n+m}(x,y) &= \textstyle\sum_{z \in S} P(X_{n+m} = y \,|\, X_n = z) P(X_n = z \,|\, X_0 = x) \\ &= \textstyle\sum_{z \in S} p_n(x,z)\, p_m(z,y). \end{aligned}$$

More generally, the multistep transition probabilities can be computed by conditioning on all the possible states visited at intermediate times, i.e., by adding the probabilities of all the paths going from state x to state y. In particular, the matrix representation is powerful enough to compute the multistep transition probabilities because it reduces lengthy calculations to simple matrix multiplications. More precisely, writing Chapman–Kolmogorov's equations in matrix form and iterating the procedure, one can prove that the coefficients of the nth power of the transition matrix are exactly the n-step transition probabilities, as shown in the next theorem.

Theorem 7.1. *For all states $x,y \in S$,*

$$p_n(x,y) = P^n(x,y) = \textit{coefficient } (x,y) \textit{ of the matrix } P^n. \tag{7.2}$$

Proof. We prove the result by induction. For $n = 1$, this follows from the definition of the transition matrix. Now, assume that the result holds for some $n \geq 1$. Then, applying the Chapman–Kolmogorov's equations with $m = 1$, we get

$$p_{n+1}(x,y) = \textstyle\sum_{z \in S} p_n(x,z)\, p(z,y) = \sum_{z \in S} P^n(x,z)\, P(z,y) = P^{n+1}(x,y)$$

from which (7.2) follows by induction. □

The main problem about discrete-time Markov chains is to study the convergence in distribution of the process, i.e., the asymptotic behavior of the n-step transition probabilities as well as the fraction of time the process spends in a given state in the long run. More precisely, letting

$$V_n(y) = \text{card}\,\{k = 0, 1, \ldots, n-1 : X_k = y\}$$

be the number of visits in state y by time n, do the limits

$$\lim_{n \to \infty} (1/n)\, V_n(y) \quad \text{and} \quad \lim_{n \to \infty} p_n(x,y) \tag{7.3}$$

exist or not? The previous theorem is rarely helpful to answer this question. It explains how to compute the multistep transition probabilities in theory but is typically used only for small time increments. To study the limit, one would need to use some tedious linear algebra: Find a new basis in which the transition matrix is diagonal or triangular, take the nth power of this matrix, and then the limit as $n \to \infty$. But things get quickly complicated using this approach. To understand the asymptotic behavior, we use instead a more theoretical approach to prove the existence or nonexistence of the limits of interest and compute them, when they exist, without using the multistep transition probabilities. The rest of this chapter presents the techniques that are used to study this limiting behavior.

7.2 Classification of states

The first step to study the limiting behavior is to classify the elements of the state space into so-called communication classes, which leads to a decomposition of the process into elementary pieces. To define this concept as well as recurrence and transience later, we need to introduce the **time of the *k*th return** to state y defined recursively as follows:

$$T_y^0 = 0 \quad \text{and} \quad T_y^k = \inf\{n > T_y^{k-1} : X_n = y\} \quad \text{for all} \quad k > 0. \tag{7.4}$$

Since the time of the first return plays a key role and will be used often, to avoid cumbersome notation, we write $T_y^1 = T_y$ from now on. Then, we define the probability that the process starting from state x ever returns to state y as

$$\rho_{xy} = P_x(T_y < \infty) = P(T_y < \infty \,|\, X_0 = x).$$

Definition 7.1 (Communication classes). Two (possibly identical) states $x, y \in S$ are said to communicate, which we write $x \leftrightarrow y$, whenever

$$p_n(x,y) > 0 \quad \text{and} \quad p_m(y,x) > 0 \quad \text{for some} \quad n,m \geq 0.$$

In words, two states communicate when the probability of going from one state to the other is positive is both directions. Using the directed graph representation, this means that there is a directed path from state x to state y and another one from state y to state x. From this observation, it is easy to see that the relation $\leftrightarrow$ defines a partition of the state space, which is made rigorous in the next lemma.

Lemma 7.1. *The binary relation $\leftrightarrow$ is an equivalence relation, thus inducing a partition of the state space S into communication classes.*

Proof. We must prove that $\leftrightarrow$ is reflexive, symmetric and transitive.

- Since $p_0(x,x) = 1$ for all $x \in S$, the relation is reflexive.
- The symmetry follows from the definition.
- The idea to prove transitivity is that if, in the directed graph representation, there is a path from state x to state y and a path from state y to state z then the concatenation of both paths is a path from state x to state z. To make this more rigorous, assume that $x \leftrightarrow y$ and $y \leftrightarrow z$. Then,

  $$p_n(x,y) > 0 \quad \text{and} \quad p_m(y,z) > 0 \quad \text{for some} \quad n,m \geq 0.$$

  Using Chapman–Kolmogorov's equations, we get

  $$p_{n+m}(x,z) = \textstyle\sum_{w \in S} p_n(x,w)\, p_m(w,z) \geq p_n(x,y)\, p_m(y,z) > 0.$$

  By symmetry, we also have $p_k(z,x) > 0$ for some $k \in \mathbb{N}$.

This completes the proof. □

Note that the process can move from one communication class to another but it also follows from the definition that when the process leaves a class it cannot come back to this class. A communication class C is said to be **closed** when

$$\rho_{xy} = 0 \quad \text{for all} \quad x \in C \quad \text{and} \quad y \notin C.$$

In words, once the process enters a closed class, it cannot leave this class, so the analysis often reduces to understanding Markov chains that have only one (necessarily closed) communication class. Such processes are called **irreducible**. Another reason why identifying the partition into communication classes helps us study a Markov chain is that certain properties are invariant across states within the same class. Such properties are called **class properties**.

A natural question about the states of a Markov chain is whether they are recurrent or transient, which turns out to be a class property.

Definition 7.2. State x is said to be

$$\textbf{recurrent} \text{ when } \rho_{xx} = 1 \quad \text{and} \quad \textbf{transient} \text{ when } \rho_{xx} < 1.$$

To better understand the dichotomy between recurrence and transience, and also explain the terminology, we need the strong Markov property. In the same spirit as the optional stopping theorem for martingales, this property basically states that the Markov property (7.1) still holds when one replaces the deterministic time n by an almost surely finite stopping time.

Theorem 7.2 (Strong Markov property). *Let T be a stopping time for the natural filtration of the process (X_n). Then,*

$$P(X_{T+n} = y \,|\, X_T = x, T < \infty) = p_n(x,y) \quad \textit{for all} \quad x,y \in S.$$

Proof. According to (7.1), we have

$$\begin{aligned} P(X_{T+n} = y, X_T = x, T = m) &= P(X_{m+n} = y, X_m = x, T = m) \\ &= P(X_{m+n} = y \,|\, X_m = x, T = m)\, P(X_m = x, T = m) \\ &= p_n(x,y)\, P(X_m = x, T = m) = p_n(x,y)\, P(X_T = x, T = m). \end{aligned}$$

Summing over all $m \in \mathbb{N}$, we deduce that

$$P(X_{T+n} = y, X_T = x, T < \infty) = p_n(x,y)\, P(X_T = x, T < \infty).$$

The result follows by dividing the last expression by $P(X_T = x, T < \infty)$. □

Using the strong Markov property, we obtain

$$\begin{aligned} P_x(T_x^k < \infty) &= P_x(T_x^k < \infty \,|\, T_x^{k-1} < \infty)\, P(T_x^{k-1} < \infty) \\ &= \rho_{xx} P_x(T_x^{k-1} < \infty) = \rho_{xx}^2 P_x(T_x^{k-2} < \infty) = \cdots = \rho_{xx}^k. \end{aligned} \tag{7.5}$$

Then, using (7.5), we can explain the terminology.

- When x is recurrent ($\rho_{xx} = 1$), we have $\rho_{xx}^k = 1$ for all k so, at least when starting from state x, the process returns to state x infinitely often.
- When x is transient ($\rho_{xx} < 1$), we have $\lim_{k \to \infty} \rho_{xx}^k = 0$ so, regardless of the initial state, the process returns to state x only a finite number of times.

To prove that recurrence and transience are class properties, we need the following characterization of recurrence and transience.

Lemma 7.2. *State x is recurrent if and only if $\sum_n p_n(x,x) = \infty$.*

Proof. Note that state x is recurrent if and only if the expected number of visits in x starting from x is infinite. Indeed, if x is recurrent then (7.5) implies that the number of visits in state x is almost surely infinite therefore

$$E_x(\textstyle\sum_n \mathbf{1}\{X_n = x\}) = \infty$$

whereas if x is transient then (7.5) implies that

$$E_x(\sum_n \mathbf{1}\{X_n = x\}) = 1 + \sum_k k\rho_{xx}^k(1-\rho_{xx}) = (1-\rho_{xx})^{-1} < \infty.$$

Since in addition the monotone convergence theorem implies that

$$E_x(\sum_n \mathbf{1}\{X_n = x\}) = \sum_n E_x(\mathbf{1}\{X_n = x\}) = \sum_n p_n(x,x),$$

indicating that the expected number of visits in state x starting from state x is equal to the sum in the statement of the lemma, then the proof is complete. □

Lemma 7.3. *Recurrence and transience are class properties.*

Proof. Since $\leftrightarrow$ is symmetric, it suffices to prove that if states x and y communicate and state x is recurrent, then y is recurrent. Now, if $x \leftrightarrow y$, then

$$p_n(x,y) > 0 \quad \text{and} \quad p_m(y,x) > 0 \quad \text{for some} \quad n,m \geq 0.$$

Observing also that

$$p_{k+n+m}(y,y) \geq p_m(y,x)\,p_k(x,x)\,p_n(x,y)$$

we deduce that, if x is recurrent then

$$\sum_k p_k(y,y) \geq \sum_k p_{k+n+m}(y,y) \geq p_m(y,x)\,p_n(x,y)\sum_k p_k(x,x) = \infty$$

therefore, y is recurrent according to Lemma 7.2. □

We can now combine all the results of this section to explain how to classify the states of a discrete-time Markov chain. It follows from Lemma 7.3 that, within the same communication class, all the states are recurrence or all the states are transient, so the class itself is said to be recurrent or transient. This, together with the definitions, implies that

- a communication class which is not closed is always transient

since in this case it is easy to identify at least one state that is transient. The converse is not true: a closed class can be transient as well. However, when the class is finite, at least one state must be visited infinitely often and is therefore recurrent, so Lemma 7.3 implies that

- a closed finite communication class is always recurrent.

The remaining classes, which are closed and infinite, are more difficult to study because both recurrence and transience are possible. In this case:

- for closed infinite communication classes, use Lemma 7.2.

For example, we will show in Chapter 8 that symmetric random walks on regular lattices, which are irreducible Markov chains with an infinite state space, are recurrent in one and two dimensions but transient in higher dimensions. The proof of this result about random walks relies on Lemma 7.2.

7.3 Stationary distribution

Before giving general conditions under which the limits in (7.3) exist, we start with some tools to compute these limits when they exist. More precisely, we first introduce the concept of stationary distribution, which can be viewed as a potential limit in distribution of the process, and show how this stationary distribution can be computed using the transition matrix.

Definition 7.3 (Stationary distribution). A distribution π on the state space S is a stationary distribution for the Markov chain (X_n) whenever

$$P_\pi(X_n = x) = \pi(x) \quad \text{for all} \quad (x,n) \in S \times \mathbb{N}$$

where P_π is the law of the process starting from the distribution π.

In other words, whenever the process is distributed according to a stationary distribution π, it remains so forever. The main tool to find stationary distributions is the transition matrix, as shown in the next lemma.

Theorem 7.3. *The distribution π is stationary if and only if $\pi P = \pi$.*

Proof. For any distribution π,

$$\begin{aligned} P_\pi(X_n = x) &= \textstyle\sum_{y \in S} P_\pi(X_n = x \,|\, X_0 = y)\, P_\pi(X_0 = y) \\ &= \textstyle\sum_{y \in S} \pi(y)\, p_n(y,x) = \sum_{y \in S} \pi(y)\, (P^n)(y,x) = (\pi P^n)(x). \end{aligned}$$

Taking $n = 1$ and assuming that π is stationary, we get

$$\pi P(x) = P_\pi(X_1 = x) = \pi(x) \quad \text{for all} \quad x \in S$$

showing that $\pi P = \pi$. Conversely, assuming that $\pi P = \pi$,

$$\begin{aligned} P_\pi(X_n = x) &= (\pi P^n)(x) = (\pi P P^{n-1})(x) \\ &= (\pi P^{n-1})(x) = \cdots = (\pi P^0)(x) = \pi(x) \end{aligned}$$

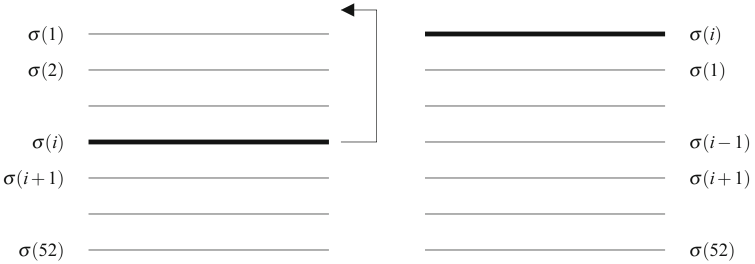
for all $(x,n) \in S \times \mathbb{N}$, therefore π is a stationary distribution. □

Theorem 7.3 is generally useful in practice, but it might sometimes lead to lengthy calculations. There are two situations in which determining a stationary distribution is simplified—when the transition matrix is doubly stochastic and when the Markov chain is time-reversible.

Doubly stochastic transition matrix — Note that a transition matrix is always stochastic: The coefficients are nonnegative and the rows sum to one. In particular, vectors with constant coordinates are always right eigenvectors corresponding to the eigenvalue one. If the columns also sum to one, the matrix is called doubly stochastic and the constant vectors are also left eigenvectors corresponding to the eigenvalue one. In this case, Theorem 7.3 implies that, at least when the state space is finite, the uniform distribution on S is stationary.

Example 7.1 (Card shuffling). From an ordinary deck of 52 cards, select one card uniformly at random and place it on top of the deck. We are interested in whether this procedure preserves a perfect mixing or not, i.e., assuming that all 52! possible permutations of the cards are initially equally likely, does this remain true after repeating the above procedure? This problem can be easily solved using the theory of discrete-time Markov chains.

To begin with, we define a Markov chain (X_n) that keeps track of the order of the cards: we number the cards and let X_n be the permutation of the cards after n selections. This defines a discrete-time Markov chain on the symmetric group $\mathfrak{S}_{52}$ with the following picture showing one time step:



In particular, using Theorem 7.3 in order to find the stationary distributions of the process leads to an unwelcoming linear system with 52! equations! Instead, observe that, for all $\sigma \in \mathfrak{S}_{52}$ and defining

$$\sigma_i = (\sigma(2),\dots,\sigma(i),\sigma(1),\sigma(i+1),\dots,\sigma(52)) \quad \text{for} \quad i = 1,2,\dots,52,$$

the transition probabilities satisfy

$$\begin{aligned} \textstyle\sum_{\tau\in\mathfrak{S}_{52}} p(\tau,\sigma) &= \textstyle\sum_{\tau\in\mathfrak{S}_{52}} P(X_{n+1} = \sigma \,|\, X_n = \tau) \\ &= \textstyle\sum_{i=1,2,\dots,52} P(X_{n+1} = \sigma \,|\, X_n = \sigma_i) = 52\,(1/52) = 1. \end{aligned}$$

This shows that the transition matrix is doubly stochastic. In particular, we obtain without calculation that the uniform distribution π on the symmetric group

$$\pi(\sigma) = 1/\operatorname{card}(\mathfrak{S}_{52}) = 1/52! \quad \text{for all} \quad \sigma \in \mathfrak{S}_{52}$$

is a stationary distribution. This means that, starting from a perfect mixing in which all permutations of the cards are equally likely, the procedure described above preserves this property. □

Time-reversible Markov chain — A distribution π on the state space is said to be reversible if it satisfies the so-called detailed balance equation

$$\pi(x)\,p(x,y) = \pi(y)\,p(y,x) \quad \text{for all} \quad x,y \in S. \tag{7.6}$$

Note that a reversible distribution satisfies

$$(\pi P)(x) = \textstyle\sum_{y\in S} \pi(y)\,p(y,x) = \sum_{y\in S} \pi(x)\,p(x,y) = \pi(x)$$

therefore, it is also a stationary distribution according to Theorem 7.3. This is useful in practice because, when there is indeed a reversible distribution, the system of equations (7.6) is in general easier to solve than the system $\pi P = \pi$. To gain some intuition, it is also worth mentioning that, starting from a reversible distribution, one can prove that the evolution forward in time and backward in time have the same distribution, which explains the terminology. Along these lines, we have the following result which gives a characterization of reversibility.

Lemma 7.4. *Assume that a Markov chain converges in distribution to π starting from any state. Then the distribution π is reversible if and only if for all*

$$x_0, x_1, \ldots, x_n = x_0 \quad \textit{such that} \quad p(x_{i-1}, x_i) > 0 \quad \textit{for} \quad 0 < i \leq n,$$

the following equation holds:

$$p(x_0,x_1)\,p(x_1,x_2)\cdots p(x_{n-1},x_n) = p(x_n,x_{n-1})\cdots p(x_2,x_1)\,p(x_1,x_0). \tag{7.7}$$

In other words, the probability that the process follows a given admissible path from state x_0 back to state x_0 is also the probability that it follows the reverse path.

Proof. Assume that π is reversible. Then,

$$\pi(x_{i-1})\,p(x_{i-1},x_i) = \pi(x_i)\,p(x_i,x_{i-1}) \quad \text{for all} \quad 0 < i \leq n$$

which, together with $x_0 = x_n$, implies that

$$\prod_{i=1}^{n} \frac{p(x_{i-1},x_i)}{p(x_i,x_{i-1})} = \prod_{i=1}^{n} \frac{\pi(x_i)}{\pi(x_{i-1})} = \frac{\pi(x_n)}{\pi(x_0)} = 1.$$

In particular, equation (7.7) holds. Conversely, assume that any cycle has the same probability as the reverse cycle. Taking $x_0 = x_n = x$ and $x_1 = y$, we get

$$\frac{p(x,y)}{p(y,x)} \prod_{i=2}^{n} \frac{p(x_{i-1},x_i)}{p(x_i,x_{i-1})} = \prod_{i=1}^{n} \frac{p(x_{i-1},x_i)}{p(x_i,x_{i-1})} = 1.$$

Summing over all $x_2, x_3, \ldots, x_{n-1}$, taking the limit as $n \to \infty$, and using that the n-step transition probabilities converge to the stationary distribution, we get

$$1 = \frac{p(x,y)}{p(y,x)}\,\frac{p_{n-1}(y,x)}{p_{n-1}(x,y)} = \frac{p(x,y)}{p(y,x)}\,\frac{\lim_{n\to\infty} p_{n-1}(y,x)}{\lim_{n\to\infty} p_{n-1}(x,y)} = \frac{p(x,y)}{p(y,x)}\,\frac{\pi(x)}{\pi(y)}$$

therefore $\pi(x)\,p(x,y) = \pi(y)\,p(y,x)$ and π is reversible. □

Example 7.2 (Symmetric random walk). Let $G = (V,E)$ be a finite connected graph, i.e., a collection of vertices along with a set of edges connecting some of the vertices. The fact that the graph is connected means that there is a path of edges connecting any two vertices. Now, assume that a particle moves on the set of vertices as follows: if the particle is located at x, then it crosses one of the edges incident to x chosen uniformly at random. The symmetric random walk is the process (X_n) that keeps track of the location of the particle at time n. Note that this is a Markov chain with a state space that is the vertex set and with transition probabilities

$$p(x,y) = (\deg(x))^{-1}\,\mathbf{1}\{(x,y)\in E\} \quad \text{for all} \quad x,y\in V.$$

The transition probabilities result in a quite complicated transition matrix; therefore, collecting the possible stationary distributions using Theorem 7.3 seems to be very difficult. However, it is clear that

$$\begin{aligned}\deg(x)\,p(x,y) &= \deg(x)\,(\deg(x))^{-1}\,\mathbf{1}\{(x,y)\in E\}\\ &= \mathbf{1}\{(x,y)\in E\} = \mathbf{1}\{(y,x)\in E\} = \deg(y)\,p(x,y)\end{aligned}$$

showing that the detailed balance equation (7.6) holds for

$$\pi(x) = \deg(x)/\textstyle\sum_y \deg(y) \quad \text{for all} \quad x\in V$$

which is therefore a reversible (and stationary) distribution. Note that convergence to this stationary distribution starting from any vertex is not guaranteed so Lemma 7.4 is not applicable. However, the existence of a reversible distribution can be anticipated from the lemma. Indeed, the probability that the particle follows a cycle in one direction only depends on the degree of the vertices in the cycle and is therefore equal to the probability of following the cycle in the opposite direction. This shows that equation (7.7) holds. □

One shortcoming of Theorem 7.3 is that it does not give any general condition for the existence and uniqueness of the stationary distribution, or the existence of the two limits in (7.3). In the next two sections, we introduce a little more of theory to better understand these aspects.

7.4 Number of visits

In this section, we study the fraction of time a Markov chain spends in a given state and give general conditions under which the first limit in (7.3) exists. Intuitively, if there is a unique stationary distribution, one expects the fraction of time spent in a state x to converge to $\pi(x)$. But since the number of visits in a transient state is finite and therefore the fraction of time spent in a transient state goes to zero, this suggests that transient Markov chains do not have stationary distributions. In fact, even recurrence is not sufficient to guarantee the existence of a stationary distribution: we need the time of the first return to be finite in expected value. This motivates the following definition.

Definition 7.4. A recurrent state x is said to be

$$\begin{aligned}\textbf{positive recurrent} &\quad \text{when} \quad E_x(T_x) < \infty\\ \textbf{null recurrent} &\quad \text{when} \quad E_x(T_x) = \infty.\end{aligned}$$

The expected value $E_x(T_x)$ is called the **mean recurrence time** at state x. The next theorem shows that the first limit in (7.3), i.e., the fraction of time spent in a given state, exists when the process is irreducible and all states are positive recurrent. In fact, we will see in a moment that positive recurrence is equivalent to the existence of a stationary distribution.

Theorem 7.4. *Assume that (X_n) is irreducible and that all states are positive recurrent. Then there is a unique stationary distribution π and*

$$(1/n)\,V_n(y) \xrightarrow{a.s.} \pi(y) = 1/E_y(T_y).$$

Proof. The proof is divided into four steps.

Step 1 — Existence of a stationary measure.

To begin with, we prove by construction the existence of a stationary measure, i.e., a measure μ on the state space that satisfies the equation

$$\textstyle\sum_{y\in S}\mu(y)\,p(y,z) = \mu(z) \quad \text{for all} \quad z \in S \tag{7.8}$$

or equivalently $\mu P = \mu$, but is not necessarily a probability measure. One can think of a natural candidate as follows: Fix a state x and assume that the measure of a state y is the expected number of visits in state y between two consecutive visits in state x. More precisely, we let

$$\mu_x(y) = \textstyle\sum_n P_x(X_n = y, T_x > n) \quad \text{for all} \quad y \in S.$$

To check (7.8) when $z \neq x$, we use Fubini's theorem to get

$$\begin{aligned}\textstyle\sum_{y\in S}\mu_x(y)\,p(y,z) &= \textstyle\sum_n\sum_{y\in S}P_x(X_n = y, T_x > n)\,p(y,z)\\ &= \textstyle\sum_n\sum_{y\in S}P_x(X_n = y, X_{n+1} = z, T_x > n) = \sum_n P_x(X_{n+1} = z, T_x > n+1)\\ &= \textstyle\sum_n P_x(X_n = z, T_x > n) = \mu_x(z).\end{aligned}$$

To check (7.8) when $z = x$, we again use Fubini's theorem to get

$$\begin{aligned}\textstyle\sum_{y\in S}\mu_x(y)\,p(y,x) &= \textstyle\sum_n\sum_{y\in S}P_x(X_n = y, X_{n+1} = x, T_x > n)\\ &= \textstyle\sum_n P_x(X_{n+1} = x, T_x = n+1) = \sum_n P_x(T_x = n+1) = 1 = \mu_x(z)\end{aligned}$$

where the last equality follows from the fact that, starting from x, there is exactly one visit in x before time T_x. In conclusion, μ_x is indeed stationary.

Step 2 — Existence of a stationary distribution.

Since $\mu_x P = \mu_x$, we obtain a stationary distribution by simply dividing the measure by its total mass so we only need to check that this total mass is neither zero nor infinity. Clearly the total mass is positive because $\mu_x(x) = 1$. Moreover, using that x is positive recurrent, one more application of Fubini's theorem gives

$$\begin{aligned}\mu_x(S) &= \textstyle\sum_{y\in S}\mu_x(y) = \sum_y\sum_n P_x(X_n = y, T_x > n)\\ &= \textstyle\sum_{y\in S} E_x(\sum_{n<T_x}\mathbf{1}\{X_n = y\}) = E_x(\sum_{n<T_x}\sum_{y\in S}\mathbf{1}\{X_n = y\})\\ &= E_x(T_x) < \infty.\end{aligned}$$

Step 3 — Convergence of the fraction of time spent in a state.

Note that, for the process starting from state y, the number of visits in that state and the return times to this state are related by the following equivalence:

$$V_n(y) = m \quad \text{if and only if} \quad T_y^m < n \leq T_y^{m+1}. \tag{7.9}$$

In addition, it follows from the strong Markov property that, starting from state y, the times between consecutive visits in that state are independent and identically distributed, therefore the strong law of large numbers implies that

$$T_y^m/m \xrightarrow{a.s.} E_y(T_y) < \infty. \tag{7.10}$$

The expected value is finite because state y is positive recurrent. Recurrence also implies that, regardless of the initial distribution, the first time the process visits y is almost surely finite which, together with (7.9)–(7.10), implies that

$$\lim_{n\to\infty}(1/n)\,V_n(y) = \lim_{m\to\infty} m/T_y^m = 1/E_y(T_y) \tag{7.11}$$

almost surely.

Step 4 — Expression and uniqueness of the stationary distribution.

To find the expression of π, note that

$$\begin{aligned}E_\pi((1/n)\,V_n(y)) &= E_\pi((1/n)\,\text{card}\,\{k = 0,1,\ldots,n-1 : X_k = y\})\\ &= (1/n)\,\textstyle\sum_{k=0,1,\ldots,n-1} E_\pi(\mathbf{1}\{X_k = y\})\\ &= (1/n)\,\textstyle\sum_{k=0,1,\ldots,n-1} P_\pi(X_k = y)\\ &= (1/n)\,\textstyle\sum_{k=0,1,\ldots,n-1} \pi(y) = \pi(y).\end{aligned}$$

In particular, (7.11) and the dominated convergence theorem give

$$\begin{aligned}\pi(y) &= \lim_{n\to\infty} E_\pi((1/n)\,V_n(y)) = E_\pi\,(\lim_{n\to\infty}(1/n)\,V_n(y))\\ &= E_\pi(1/E_y(T_y)) = 1/E_y(T_y)\end{aligned}$$

which also proves that the stationary distribution is unique. □

Using the proof of Theorem 7.4, we can now establish the equivalence between the existence of a stationary distribution and positive recurrence.

Theorem 7.5. *For an irreducible Markov chain, there exists a stationary distribution if and only if all states are positive recurrent.*

Proof. There are two implications to be proved.

Necessary condition — The first two steps of the proof of Theorem 7.4 show that there is a stationary distribution when at least one state is positive recurrent.

Sufficient condition — Assume that there is a stationary distribution π and let $y \in S$. Then it follows from the fourth step of the proof of Theorem 7.4 that

$$\pi(y) = 1/E_y(T_y). \tag{7.12}$$

Now, fix a state x such that $\pi(x) > 0$ and an integer $n > 0$ such that $p_n(x,y) > 0$, which is possible because the process is irreducible. Then,

$$\pi(y) = (\pi P^n)(y) = \textstyle\sum_{z \in S} \pi(z)\, p_n(z,y) \geq \pi(x)\, p_n(x,y) > 0. \tag{7.13}$$

Combining (7.12)–(7.13), we get $E_y(T_y) < \infty$ so y is positive recurrent. □

Note from the proof that one positive recurrent state is enough to have a stationary distribution, which in turns implies that all states are positive recurrent. This shows that positive recurrence is a class property. Since transience also is a class property, being neither positive recurrent nor transient must be a class property. In particular, we deduce the following result.

Lemma 7.5. *Positive recurrence and null recurrence are class properties.*

In practice, checking positive recurrence by computing the expected value of the time of the first return might be tedious. However, things are simple for finite irreducible Markov chains. In this case, since the state space is finite, at least one state must be positive recurrent, so all states are positive recurrent, which gives the following result that is often useful in practice.

Theorem 7.6. *Finite irreducible Markov chains are positive recurrent.*

To conclude, we give two examples showing how to use the main results of this section and how to compute the fraction of time spent in a given state and the expected time between consecutive visits in a given state.

Example 7.3 (Symmetric random walk). Let $G = (V,E)$ be a finite connected graph and consider the associated random walk with transition probabilities

$$p(x,y) = (\deg(x))^{-1}\, \mathbf{1}\{(x,y) \in E\} \quad \text{for all} \quad x,y \in V.$$

In Example 7.2 above, we proved that

$$\pi(x) = \deg(x)/\textstyle\sum_y \deg(y) \quad \text{for all} \quad x \in V$$

is a reversible, and thus stationary distribution. Since in addition the graph is finite and connected, the process is finite and irreducible, so it follows from Theorem 7.6 that the process is also positive recurrent. In particular, by Theorem 7.4, the stationary distribution π is unique and, starting from any vertex,

$$(1/n)\, V_n(x) \xrightarrow{a.s.} \pi(x) = \deg(x)/\textstyle\sum_y \deg(y).$$

This shows that the fraction of time the random walk spends on each vertex is proportional to the degree of this vertex. □

Example 7.4 (Time between homeworks). Assume that each evening a student in probability theory works on her homework with probability

$$p_m = 1 - 1/(m+1)$$

if the last time she worked was m days ago. The monotonicity of $m \mapsto p_m$ models the fact that the student's guilt increases with the number of consecutive days she has spent without working on her probability class.

We are interested in the expected value of the time between the student's consecutive work sessions on probability theory. To compute the expected time, we let X_n be the number of days before day n and including day n since the student last worked on her probability class. In particular, she works on day n if and only if $X_n = 0$ so the expected time between consecutive homeworks is also the expected time between consecutive visits in state 0. Now, observe that (X_n) defines a discrete-time Markov chain with transition matrix

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

The process is clearly irreducible but positive recurrence is not obvious because the state space is infinite. However, one can solve $\pi P = \pi$ explicitly to prove that there is a unique stationary distribution. Indeed, we obtain by induction

$$\pi(0) = 2\,\pi(1) = 6\,\pi(2) = 24\,\pi(3) = \cdots = (m+1)!\;\pi(m)$$

for all $m \geq 0$. To extract the value of $\pi(0)$, we solve

$$1 = \sum_{m=0}^{\infty} \pi(m) = \sum_{m=0}^{\infty} \frac{\pi(0)}{(m+1)!} = \pi(0)\left(\sum_{m=0}^{\infty} \frac{1}{m!} - 1\right) = (e-1)\,\pi(0)$$

which gives existence and uniqueness of the stationary distribution

$$\pi(m) = \frac{\pi(0)}{(m+1)!} = \frac{1}{e-1}\,\frac{1}{(m+1)!} \quad \text{for all} \quad m \geq 0.$$

From Theorem 7.5, we deduce positive recurrence while Theorem 7.4 implies that the expected time between two consecutive visits in state 0 is

$$E_0(T_0) = 1/\pi(0) = e - 1 \approx 1.718.$$

Recalling that $X_n = 0$ if and only if the student works on day n, we conclude that the expected value of the time between homeworks is about 1.718 days. □

7.5 Convergence to the stationary distribution

In this last section, we study the existence of the second limit in (7.3) for irreducible Markov chains. Note that the existence of this limit is slightly stronger than the convergence of the fraction of time spent in a given state, and in fact positive recurrence is not sufficient to guarantee convergence of the multistep transition probabilities. The typical bad scenario occurs when an irreducible positive recurrent Markov chain gets trapped in some deterministic cyclic pattern, making the visit in some states only possible at some deterministic times, which is determined by the topology of the directed graph representing the process. In this case, the multistep transition probabilities form a periodic sequence: the Cesàro mean converges but not the sequence itself. To prevent this bad event from happening, we need one more ingredient called *aperiodicity*, which we now define.

Definition 7.5 (Period). The period of state x is

$$d(x) = \gcd\{n \geq 1 : p_n(x,x) > 0\}.$$

Note that the period of a state should not be confused with the minimum number of time steps for the process to return to this state. In particular, even though several steps are needed to return to a state, its period can still be equal to one as illustrated by the following example.

Example 7.5 (Symmetric random walk). Consider the symmetric random walk as defined in Example 7.2. The state space consists of the vertex set of the graph and the period of each vertex depends on the topology of the graph.

To begin with, assume that the random walk evolves on the cycle with m vertices depicted in the top-left corner of Figure 7.1 for $m = 8$. For this graph and any other graph, there is a positive probability that the random walk crosses an edge in one direction and then the same edge in the opposite direction therefore

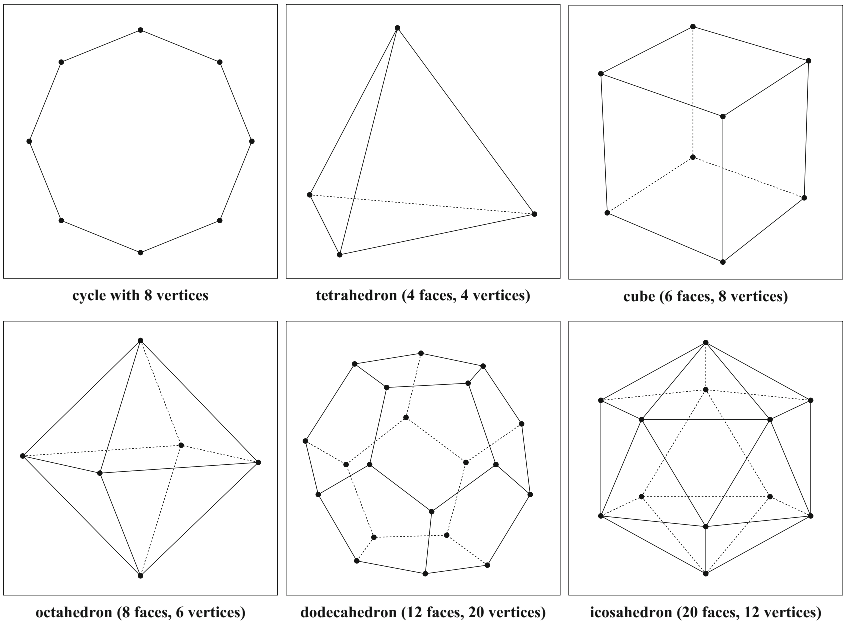
$$p_2(x,x) > 0 \quad \text{for all} \quad x \in V.$$

More generally, the process can return to its initial position in any even number of steps. But the random walk can also return to its initial position by turning around the cycle which takes a number of steps equal to m plus an even number so

$$p_n(x,x) > 0 \quad \text{if and only if} \quad n \in \{2k, m+2k\} \quad \text{for some} \quad k \in \mathbb{N}.$$

In conclusion, from the definition of period, we get

$$d(x) = \gcd(2,m) = \begin{cases} 1 & \text{when } m \text{ is odd} \\ 2 & \text{when } m \text{ is even.} \end{cases}$$



cycle with 8 vertices

tetrahedron (4 faces, 4 vertices)

cube (6 faces, 8 vertices)

octahedron (8 faces, 6 vertices)

dodecahedron (12 faces, 20 vertices)

icosahedron (20 faces, 12 vertices)

Fig. 7.1 Cycle with 8 vertices and the five Platonic solids.

Note in particular that, even though $p(x,x) = 0$, the period of each of the vertices of a triangle or a pentagon is equal to one.

We now consider the symmetric random walk on the five Platonic solids in Figure 7.1. Since the faces of the tetrahedron, octahedron, and icosahedron are triangles, the random walk can return to its initial position in two or three steps. Similarly, since the faces of the dodecahedron are pentagons, the random walk on this graph can return to its initial position in two or five steps. It follows that for these four graphs, the period of each vertex is given by

$$d(x) = \gcd(2,3) = \gcd(2,5) = 1.$$

In contrast, for the cube, the random walk can only return to its initial position in an even number of steps so the period is 2. To see that indeed an even number of steps is required, one can think of the coordinates of the eight corners of the cube as zeros and ones and notice that each jump of the random walk will change the parity of the sum of the coordinates. Note that the exact same argument implies that, for the random walk on the integers, each vertex also has period 2. $\square$

As for recurrence and transience, the period is a class property, i.e., all states within the same communication class have the same period. Also, an irreducible Markov chain is said to be **aperiodic** when all states have period one.

Lemma 7.6. *The period is a class property.*

Proof. Assume that $x \leftrightarrow y$. Then,

$$p_n(x,y) > 0 \quad \text{and} \quad p_m(y,x) > 0 \quad \text{for some } n,m \geq 0.$$

For all $k > 0$ such that $p_k(y,y) > 0$, we have

$$\begin{aligned} p_{n+m}(x,x) &\geq p_n(x,y)\ p_m(y,x) > 0 \\ p_{n+k+m}(x,x) &\geq p_n(x,y)\ p_k(y,y)\ p_m(y,x) > 0 \end{aligned}$$

therefore $d(x)$ divides both $n+m$ and $n+k+m$ so it divides k. Since $d(y)$ is the greatest integer that divides all such k, we deduce that $d(x) \leq d(y)$. By symmetry, the reverse inequality is also true so both periods are equal. □

The next theorem shows that, if in addition to being irreducible and positive recurrent, the process is also aperiodic, then the second limit in (7.3) exists and can be computed from the stationary distribution.

Theorem 7.7. *Assume that the discrete-time Markov chain (X_n) is irreducible, positive recurrent and aperiodic. Then,*

$$\lim_{n\to\infty} p_n(x,y) = \pi(y) \quad \textit{for all} \quad x,y \in S.$$

Proof. The theorem can be proved using the Perron–Frobenius theorem in the field of linear algebra. Instead, we follow a probabilistic approach based on an argument due to Wolfgang Doeblin whose key is to study the process

$$(Z_n) = ((X_n),(Y_n))$$

where (X_n) and (Y_n) are two independent copies of the Markov chain starting from different distributions. Note that this process is again a discrete-time Markov chain and that, due to independence, its transition probabilities are given by

$$\bar{p}((x_1,y_1),(x_2,y_2)) = p(x_1,x_2)\ p(y_1,y_2) \quad \text{for all} \quad (x_1,x_2),(y_1,y_2) \in S \times S.$$

To study this process and deduce the theorem, we proceed in four steps.

Step 1 — There exists $K < \infty$ such that $p_k(x,x) > 0$ for all $k \geq K$.

Since state x has period one, there exist

$$n,m \in I_x = \{k \geq 1 : p_k(x,x) > 0\}$$

relatively prime. In particular, by Bézout's identity, one can find

$$a,b \in \mathbb{Z} \quad \text{such that} \quad an+bm = 1 \ \text{ and } \ a < 0 < b.$$

Since in addition the set I_x is closed under addition, which follows from the fact that the concatenation of two directed paths from state x to state x is again a directed path from state x to state x, we deduce that

$$c = -an + (b-2a)\,m \in I_x \quad \text{and} \quad c+1 = (2b-2a)\,m \in I_x.$$

Let $k \geq c^2$. Then, there exist $q, r \in \mathbb{N}$ with $r < c$ such that

$$k = c^2 + qc + r = (c+q-r)\,c + r\,(c+1).$$

Invoking again that I_x is closed under addition, this shows that $k \in I_x$.

Step 2 — The process (Z_n) is irreducible.

Fix two pairs $(x_1, x_2), (y_1, y_2) \in S \times S$. By irreducibility,

$$p_n(x_1, x_2) > 0 \quad \text{and} \quad p_m(y_1, y_2) > 0 \quad \text{for some} \quad n, m > 0.$$

This, together with the first step, implies that, for k large,

$$\begin{aligned} \bar{p}_{n+m+k}((x_1, y_1), (x_2, y_2)) &= p_{n+m+k}(x_1, x_2)\; p_{n+m+k}(y_1, y_2) \\ &\geq p_{m+k}(x_1, x_1)\; p_n(x_1, x_2)\; p_{n+k}(y_1, y_1)\; p_m(y_1, y_2) > 0 \end{aligned}$$

showing that (Z_n) is irreducible.

Step 3 — The process (Z_n) is recurrent.

The measure $\bar{\pi}(x, y) = \pi(x)\,\pi(y)$ is stationary for the process (Z_n). Since each of the two coordinates of the process is positive recurrent,

$$\bar{\pi}(x, y) = \pi(x)\,\pi(y) = (1/E_x(T_x))(1/E_y(T_y)) > 0 \quad \text{for all} \quad (x, y) \in S \times S.$$

This implies that (Z_n) is recurrent.

Step 4 — Time to hit the diagonal $\Delta = \{(x, x) : x \in S\}$.

Let T be the first time the process (Z_n) hits the diagonal:

$$T = \inf\{n \geq 0 : Z_n \in \Delta\} = \inf\{n \geq 0 : X_n = Y_n\}.$$

Since (Z_n) is irreducible and recurrent,

$$\lim_{n \to \infty} P(T > n) = 0. \tag{7.14}$$

Moreover, after time T, both coordinates have the same distribution so

$$\begin{aligned} P(X_n = y) &= P(X_n = y \text{ and } T \leq n) + P(X_n = y \text{ and } T > n) \\ &= P(Y_n = y \text{ and } T \leq n) + P(X_n = y \text{ and } T > n) \\ &\leq P(Y_n = y) + P(X_n = y \text{ and } T > n). \end{aligned} \tag{7.15}$$

By symmetry, we also have

$$P(Y_n = y) \leq P(X_n = y) + P(Y_n = y \text{ and } T > n). \tag{7.16}$$

Combining (7.15)–(7.16), we get

$$\begin{aligned} \textstyle\sum_{y \in S} |P(X_n = y) - P(Y_n = y)| &\leq \textstyle\sum_{y \in S} 2P(X_n = y \text{ and } T > n) \\ &\leq 2P(T > n). \end{aligned} \tag{7.17}$$

To conclude, we let $X_0 = x$ and let Y_0 have distribution π so that the second coordinate is always at stationarity, in which case (7.17) becomes

$$\begin{aligned} \textstyle\sum_{y \in S} |P_x(X_n = y) - P_\pi(Y_n = y)| &\leq \textstyle\sum_{y \in S} |p_n(x, y) - \pi(y)| \\ &\leq 2P(T > n). \end{aligned}$$

Since this goes to zero by (7.14), the proof is complete. □

The process (Z_n) in the previous proof is called a coupling of (X_n) and (Y_n). More generally, a coupling of two stochastic processes is a pair of stochastic processes defined on the same probability space such that the first and second coordinates have the same distribution as the first and second processes, respectively. In the previous proof, the two coordinates are independent but we will see later couplings where both coordinates are strongly correlated. This will be used to prove various monotonicity properties for spatially explicit models.

Example 7.6 (Card shuffling). We return to one of our previous examples: from an ordinary deck of 52 cards, select one card uniformly at random and place it on top of the deck. A natural question is to determine whether this leads to a perfect mixing where all 52! possible permutations of the cards are equally likely.

In Example 7.1, we proved that the transition matrix of the process (X_n) that keeps track of the order of the cards is doubly stochastic and that

$$\pi(\sigma) = 1/\operatorname{card}(\mathfrak{S}_{52}) = 1/52! \quad \text{for all} \quad \sigma \in \mathfrak{S}_{52}$$

is a stationary distribution. In words, the perfect mixing is stationary. To determine whether our shuffling technique leads to a perfect mixing starting from any possible ordering of the cards, it suffices to prove convergence to the distribution π. Since our procedure allows us to select the card at the top of the deck,

$$p(\sigma, \sigma) = 1/52 > 0 \quad \text{for all} \quad \sigma \in \mathfrak{S}_{52},$$

from which it follows that the process is aperiodic. Since in addition the process is finite and irreducible, and therefore positive recurrent, we directly deduce from Theorem 7.7 that, regardless of the initial order of the cards,

$$\textstyle\lim_{n \to \infty} P(X_n = \sigma) = \pi(\sigma) = 1/52! \quad \text{for all} \quad \sigma \in \mathfrak{S}_{52}.$$

This shows that the shuffling technique indeed leads to a perfect mixing. □

Example 7.7. Assume that, in a certain city, it rains independently each morning and each evening with the same probability p. A distracted woman living in this city walks to her office every morning and walks back home every evening. While going one way or another, she takes an umbrella if and only if it is raining (and she has an umbrella with her). We are interested in the probability that the woman gets wet in the long run if she has a total of r umbrellas.

The first step to solve this problem is to define a Markov chain that keeps track of the location of the umbrellas: for all $n \in \mathbb{N}$, we let

$$X_n = \text{number of umbrellas at home in the morning of day } n.$$

Since the number of umbrellas at home does not change if and only if the weather is the same morning and evening, we obtain the transition probabilities

$$p(x,x) = p^2 + (1-p)^2 \quad \text{and} \quad p(x,x+1) = p(x,x-1) = p(1-p)$$

for all $x = 1,2,\ldots,r-1$, with the boundary conditions

$$p(0,0) = 1-p \quad \text{and} \quad p(r,r) = 1-p(1-p).$$

Letting $q = 1-p$, the transition matrix is the tridiagonal matrix

$$P = \begin{pmatrix} q & p & & & \\ pq & 1-2pq & pq & & \\ & \ddots & \ddots & \ddots & \\ & & pq & 1-2pq & pq \\ & & & pq & 1-pq \end{pmatrix}.$$

The process is finite and, at least when $p \in (0,1)$, irreducible, so there exists a unique stationary distribution π. Observing that

$$\begin{aligned} (q,1,\ldots,1)\cdot P &= (q^2+pq,1,1,\ldots,1) \\ &= (q(p+q),1,1,\ldots,1) = (q,1,1,\ldots,1), \end{aligned}$$

we deduce that π is given by

$$\pi(0) = q/(r+q) \quad \text{and} \quad \pi(x) = 1/(r+q) \quad \text{for} \quad 0 < x \leq r. \tag{7.18}$$

Since $p(x,x) > 0$ for all x, the process is also aperiodic so

$$\lim_{n\to\infty} P(X_n = x) = \pi(x) \quad \text{for} \quad 0 \leq x \leq r, \tag{7.19}$$

according to Theorem 7.7. To conclude, let W_n be the event that the woman gets wet in day n and observe that this event occurs in the following two cases:

- $X_n = 0$ and it rains in the morning of day n,
- $X_n = r$ and it rains in the evening but not in the morning of day n.

Combining this together with (7.18)–(7.19), we deduce that the probability that the woman gets wet a given day in the long run is

$$\begin{aligned} \lim_{n\to\infty} P(W_n) &= \lim_{n\to\infty} P(W_n \,|\, X_n = 0)\, P(X_n = 0) \\ &\qquad + \lim_{n\to\infty} P(W_n \,|\, X_n = r)\, P(X_n = r) \\ &= p\,\pi(0) + pq\,\pi(r) = 2pq/(r+q). \end{aligned}$$

This probability is small when the weather is somewhat predictable, meaning that p is close to either zero or one, but larger when the weather changes often. □

To summarize, irreducibility and positive recurrence guarantee the existence and uniqueness of the stationary distribution and the convergence of the fraction of time spent in a given state. Adding to that aperiodicity gives also convergence of the multistep transition probabilities. Putting all the pieces together, we obtain the following picture for discrete-time Markov chains.

For an irreducible Markov chain (X_n), we have the following alternative:

1. All states are transient or all states are null recurrent in which case there is no stationary distribution and, for all states $x, y \in S$,

$$(1/n)\, V_n(y) \xrightarrow{a.s.} \lim_{n\to\infty} p_n(x,y) = 0.$$

2. All states are positive recurrent in which case there is a unique stationary distribution π and, for all states $y \in S$,

$$(1/n)\, V_n(y) \xrightarrow{a.s.} \pi(y) = 1/E_y(T_y) > 0.$$

   If in addition the process is aperiodic then $\lim_{n\to\infty} p_n(x,y) = \pi(y)$.

To conclude, we show that none of the three ingredients—irreducibility, positive recurrence, or aperiodicity—can be omitted in general.

Irreducibility — Consider a Markov chain (X_n) with m communication classes all positive recurrent. Positive recurrence implies that each of the classes must be closed so the process restricted to a given class is irreducible, which results in m distinct stationary distributions, one for each class. More precisely, calling the classes C_i, there is a unique stationary distribution π_i such that

$$\pi_i(x) > 0 \quad \text{if and only if} \quad x \in C_i.$$

In fact, there are infinitely many stationary distributions because any convex combination of stationary distribution is again a stationary distribution. Indeed, letting p_i be nonnegative real numbers that sum up to one,

$$\left(\sum_{i=1}^{m} p_i\, \pi_i\right) P = \sum_{i=1}^{m} p_i\, (\pi_i P) = \sum_{i=1}^{m} p_i\, \pi_i$$

therefore $p_1\,\pi_1 + \cdots + p_m\,\pi_m$ is a stationary distribution, so the lack of irreducibility can lead to multiple stationary distributions. □

Positive recurrence — Theorem 7.5 already shows that positive recurrence is anecessary condition for the existence of a stationary distribution. However, we

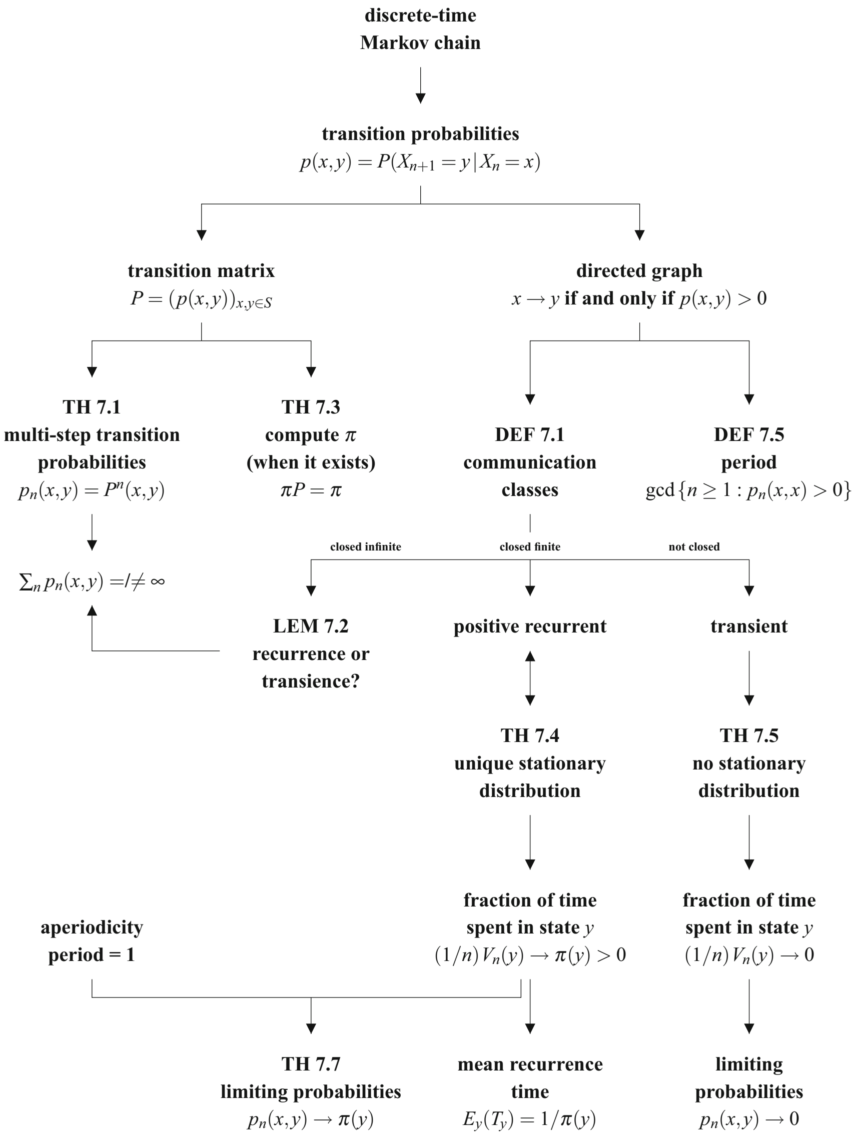


Fig. 7.2 Summary of Chapter 7.

point out that, when the process is irreducible but not positive recurrent, the fraction of time spent in a given state still converges but the limit is zero. □

Aperiodicity — The symmetric random walk (X_n) on the polygon with $2m$ sides is irreducible and positive recurrent, and we easily check that the uniform distribution on the vertex set is stationary. In particular, the fraction of time spent in a given vertex converges to $1/2m$. The process, however, has period 2 and

$$\lim_{n\to\infty} p_n(x,x) = \begin{cases} 1/m & \text{when} \quad n \text{ is even} \\ 0 & \text{when} \quad n \text{ is odd} \end{cases}$$

so the sequence $p_n(x,x)$ does not converge. □

Most of the results of this chapter are summarized in the diagram of Figure 7.2, which also describes the strategy to study irreducible Markov chains. If the process is not irreducible but hits eventually a closed communication class, the same approach can be used looking at the restriction of the process to this class. In fact, using communication classes, recurrence, and transience, we distinguish mostly two types of Markov chains that are common in application.

1. There is only one communication class: the process is irreducible.
2. There is one class that is not closed, hence transient, and at least one absorbing state, i.e., a state x where the process can get stuck: $p(x,x) = 1$.

This chapter gives the tools to study the first type of Markov chains. For the second type, the main question is to determine the probability that the process stays in the transient class forever and the probability that it hits each of the absorbing states. This can be a difficult problem because the approach to be used is generally model-specific. Examples of such processes are the gambler's ruin chain for which we used the optional stopping theorem, and branching processes for which we used the probability generating function. A first-step analysis, which consists in conditioning on the outcome of the first jump, is also useful in this context.

7.6 Exercises

General Markov chains

Exercise 7.1. Prove that if a finite Markov chain can go from any state to any other state in m steps then it can do so in n steps for all $n \geq m$.

Exercise 7.2. Prove that if a Markov chain with m states can go from state x to state y then it can go from state x to state y in less than m steps.

Hint: Think of the directed graph representation.

Exercise 7.3. Consider a Markov chain with state space $S = \{0,1\}$. The evolution of such a process is only characterized by

$$p = p(0,1) \quad \text{and} \quad q = p(1,0)$$

and we assume that $p,q \in (0,1)$ to avoid trivialities.

1. Find all the stationary distributions of this process.
2. Prove that, for all $n \geq 0$, we have

$$P(X_{n+1} = 1) - p/(p+q) = (1-p-q)(P(X_n = 1) - p/(p+q)).$$

3. Deduce exponentially fast convergence to the stationary distribution.

Hint: For the second question, condition the left-hand side on X_n.

Exercise 7.4. Let (X_n) be a finite irreducible Markov chain. Fix two states $a,b \in S$ and assume that a function ϕ defined on the state space satisfies

$$\phi(a) = 1, \quad \phi(b) = 0 \quad \text{and} \quad \phi(x) = \textstyle\sum_{y \in S} \phi(y)\, p(x,y)$$

for all states $x \in S$, $x \neq a$ and $x \neq b$.

1. Prove that $(\phi(X_n))$ is a martingale.
2. Use the optional stopping theorem to deduce that

$$\phi(x) = P_x(T_a < T_b) \quad \text{for all} \quad x \in S$$

   where T_y is the first time the process visits state y.

More applied problems

Exercise 7.5 (Taxi driver). The city of Paris is partitioned into twenty districts with the international airport being located outside the city. Assume that the customers of a taxi driver behave according to the following random process:

1. Customers taking the taxi in the city go to the airport with probability 0.1 while all the customers taking the taxi at the airport go to the city.
2. Customers going to the city are equally likely to go to any of the districts.

Assuming that the taxi driver charges customers 10 euros if they stay in the same district, 20 euros if they go from one district to another, and 50 euros if they go or come from the airport, and assuming that the taxi driver is lucky enough to always have a customer in his car, find the taxi driver's expected profit per trip.

Hint: Construct a discrete-time Markov chain with only two states that keeps track of the location of the taxi driver at the end of trip n.

Exercise 7.6 (Weather forecast). Assume that each day is either sunny or rainy, which depends on the previous two days' weather as follows:

- Assume that if today is sunny and/or yesterday was sunny then tomorrow will again be sunny with a fixed probability 0.8.
- Assume that if today is rainy and yesterday was rainy then tomorrow will again be rainy with a fixed probability 0.5.

Let (X_n) be the process that keeps track of the weather in day n. This process is not a discrete-time Markov chain because each day's weather depends on the previous two days' weather but it is possible to analyze the process as follows.

1. Find the transition matrix of the process (Y_n) where $Y_n = (X_n, X_{n+1})$.
2. Deduce the fraction of days which are sunny in the long run.

Assume now that the number of car accidents is Poisson distributed with parameter two when it is sunny but three when it is rainy.

3. Find the fraction of days with no car accident in the long run.

Exercise 7.7. Assume that each day is either sunny or rainy and that today's weather is the same as yesterday's weather with probability $p \in (0,1)$.

1. Compute the fraction of days that in the long run are sunny.
2. Prove more generally that the probability that the weather in n days will be the same as today's weather is equal to

$$1/2 + (1/2)(p-q)^n \quad \text{where} \quad q = 1 - p \in (0,1).$$

Exercise 7.8. On a certain road, one of every ten cars is followed by a truck, while four of every five trucks is followed by a car.

1. What fraction of vehicles on the road are trucks?
2. Compute the probability that the third vehicle following a car is a truck.
3. Find the probability that all ten vehicles following a truck are cars.

Exercise 7.9. A teenager plays a video game that has five levels that increase in difficulty. Let $p < 1$ and assume that the probability of completing level i is equal to p^i and that each level lasts for ten minutes regardless of whether it is successfully completed or not. Each time the teenager fails, he has to restart at level 1.

1. Find the expected number of games the teednager should play until he completes successfully all five levels of the video game.
2. Compute the expected amount of time a game lasts.
3. Compute these two quantities explicitly when $p = 1/2$.

Hint: For the second question, let X_n be the level the teenager is playing at time n where one time unit equals ten minutes.

Exercise 7.10. Assume that the numbers of new members of a chess club on successive years are independent Poisson random variables with mean μ and that the number of years members stay in the club are independent geometric random variables with parameter p. Let X_n be the number of members in year n.

1. Prove that (X_n) is a discrete-time Markov chain.
2. Letting $\mu_n = E(X_n)$, prove that

$$\mu_n = (\mu/p)(1-q^n) + q^n \mu_0 \quad \text{where} \quad q = 1-p.$$

3. Prove that the stationary distribution is a Poisson random variable.

Exercise 7.11. The following problem is from [62]. Consider a population of N agents who are either leftist, centrist, or rightist. At each time step, two agents called A and B, possibly the same, are chosen uniformly at random then agent A adopts the opinion of agent B unless the two agents are leftist and rightist, in which case they disagree too much to interact. Let (X_n) be the process that keeps track of the number of centrist agents in the population.

1. Prove that (X_n) is a discrete-time Markov chain.
2. Specify the communication classes, recurrence, and transience.
3. Using the optional stopping theorem, show that all the agents are centrist eventually with probability the initial fraction of centrist agents.

Doubly stochastic Markov chains

Exercise 7.12. A man has m identical hats that he keeps in two drawers, one fair coin in his pocket, and the following strange ritual. Each morning, he flips the coin to choose a drawer at random and take one hat from this drawer, if there is one, to wear all the day. In the evening of the days when he wears a hat, he flips again the coin to choose a drawer at random where to put the hat back. Find the fraction of days the man does not wear a hat.

Hint: Introduce a Markov chain that keeps track of the hats and use the fact that the transition matrix is doubly stochastic.

Exercise 7.13. Let X_n be the sum of n independent rolls of a fair die. Find

$$\lim_{n\to\infty} P(X_n \text{ is a multiple of } m)$$

where $m \geq 6$ is an integer.

Hint: Use again doubly stochasticity.

Time-reversible Markov chains

Exercise 7.14 (Ehrenfest chain). In the Ehrenfest chain, M particles are distributed among two compartments. At each time step, one particle is chosen uniformly at random and moved to the other compartment. Prove that the expected number μ_n of particles in the first compartment at time n is given by

$$\mu_n = \frac{M}{2} + \left(\mu_0 - \frac{M}{2}\right)\left(1 - \frac{2}{M}\right)^n.$$

This shows in particular that $\mu_n \to M/2$ as $n \to \infty$.

Hint: Express μ_{n+1} using μ_n by conditioning on X_n and use induction.

Exercise 7.15. For the Ehrenfest chain with M particles starting with x particles in the first compartment, find the probability that the second compartment becomes empty before the first compartment.

Hint: Condition on the outcome of the first update.

Exercise 7.16. Consider a Markov chain with state space $S = \{0, 1, \ldots, M\}$ for which the transition probabilities satisfy the condition

$$p(x,y) > 0 \quad \text{if and only if} \quad |x - y| = 1.$$

1. Use reversibility to find the stationary distribution of the process.
2. Use this to compute the stationary distribution of the Ehrenfest chain.

Exercise 7.17. Assume that b blue and r red balls with $b + r = 2m$ are equally divided into two urns. At each time step, one ball is chosen uniformly at random from each urn and the two balls are exchanged. Letting X_n be the number of blue balls in the first urn at time n, use reversibility to prove that

$$\lim_{n\to\infty} P(X_n = x) = \binom{b}{x}\binom{2m-b}{m-x}\bigg/\binom{2m}{m} \quad \text{for all} \quad 0 \le x \le b.$$

Chapter 8
Symmetric simple random walks

Random walks denote a general class of stochastic processes for which the definition significantly varies across the literature. Since the ultimate target of this textbook is spatial stochastic processes, the random walks we are interested in are the symmetric simple random walks on graphs as defined in Example 7.2. As previously mentioned, the random walk on the integers can be generated from the consecutive flips of a fair coin and is therefore a natural mathematical object that was already studied in the 17th century by Bernoulli and de Moivre. However, it was only as recently as 1905 that the term *random walk* was introduced in [81].

Consider a graph $G = (V, E)$, which is characterized by its vertex set and its edge set, and recall that the symmetric simple random walk on this graph is the discrete-time Markov chain (X_n) that keeps track of the location of a particle that jumps from one vertex to any of its neighbors with equal probability. Formally, this is the process with state space V and transition probabilities

$$p(x,y) = (\deg(x))^{-1}\,\mathbf{1}\{(x,y) \in E\} \quad \text{for all} \quad x, y \in V$$

where $\deg(x)$ is the number of vertices connected to x by an edge. Figure 8.1 gives a schematic picture of such a process. Note that each connected component of the graph corresponds to a closed communication class, where two vertices belong to the same connected component if they are connected by a path of edges. In particular, it suffices to focus our attention on connected graphs, in which case the random walk is irreducible and states are either all recurrent or all transient since these properties are class properties. Also, a connected graph itself is said to be recurrent or transient if the symmetric random walk on this graph is.

Examples 7.2–7.3 show that, if in addition to be connected the graph is finite, then the process has a unique stationary distribution π which is reversible and the fraction of time spent in vertex x satisfies

$$(1/n)\,V_n(x) \xrightarrow{a.s.} \pi(x) = \deg(x)/\textstyle\sum_y \deg(y) \quad \text{for all} \quad x \in V.$$

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_8

This indicates that the fraction of time the random walk spends on a given vertex in the long run is proportional to the degree of this vertex so vertices with many connections are visited more often. In the special case of a regular graph, i.e., all the vertices have the same degree, the stationary distribution is the uniform distribution on the set of vertices. The behavior of the process on infinite graphs, and in particular whether these graphs are recurrent or transient, is more complicated. The main objective of this chapter is to explore this aspect.

In the first section, we focus on the d-dimensional integer lattices. In this case, since the graph is regular, a measure is stationary if and only if it puts the same mass on each vertex. Since the number of vertices is infinite, this shows that there is no stationary distribution so the process is either transient or null recurrent. The first section shows that the infinite integer lattice is recurrent and the process null recurrent when $d \leq 2$ but transient when $d > 2$. This will be used later to study the long-term behavior of the voter model.

In the second section, we describe a useful connection between a certain class of discrete-time Markov chains and electrical networks. This connection shows in particular that whether a graph is recurrent or not is related to whether the effective resistance of the associated electrical network is infinite or not. Using this connection in combination with the results of the first section, we deduce that any subgraph of the two-dimensional integer lattice is recurrent, whereas any graph that contains the three-dimensional integer lattice as a subgraph is transient.

The last section explores a notion related to recurrence: the so-called infinite collision property. A connected graph has this property if the number of collisions between two independent random walks starting from the same vertex is almost surely infinite. We prove that, at least when there is a uniform bound on the degree, a graph is recurrent if and only if the the expected number of collisions is infinite. However, there are recurrent graphs with uniformly bounded degree that do not have the infinite collision property. We also give an example of a graph for which the degree is not bounded and that also is transient and has the infinite collision property.

Further reading

- A classical reference on random walks is Spitzer [91], whose study relies on harmonic analysis and potential theory.
- Another good reference more in the spirit of our presentation is Woess [96] which looks in particular at random walks on graphs.
- Doyle and Snell [24] is the first manuscript that explores the connections between random walks and electrical networks.
- See also Lawler [65] for more exotic results about random walks.

8.1 Infinite lattices

As previously mentioned, when the underlying graph on which the symmetric random walk evolves is infinite, things become more complicated. In particular, there is no longer a stationary distribution. But whether the graph is recurrent or transientcan still be determined for the special case of the d-dimensional integer lattices. In this case, at each time step, the walk moves in one of the $2d$ possible directions chosen uniformly at random and exact calculations or at least good estimates are possible. Because integer lattices are translation invariant, the process can be conveniently written as the sum of independent random variables:

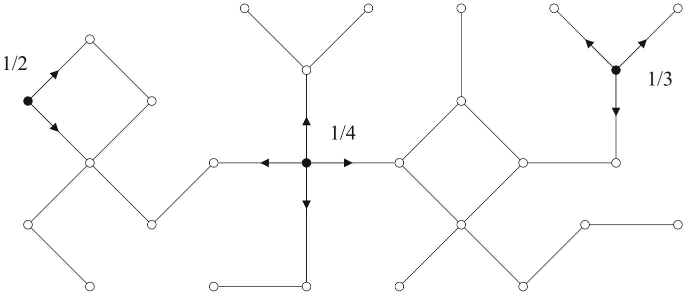


Fig. 8.1 An example of symmetric simple random walk.

$$X_n = Y_1 + \cdots + Y_n \quad \text{where} \quad Y_i \sim \text{Uniform}\{-e_1, e_1, \ldots, -e_d, e_d\} \tag{8.1}$$

are independent. Here, e_i is the ith unit vector in d dimensions. It turns out that the behavior of these processes strongly depends on the spatial dimension d, which will be used later to study the voter model.

To begin with, we look at the symmetric random walk on the one- and two-dimensional integer lattices, i.e., the process (8.1) with $d = 1, 2$. As previously mentioned, since these graphs are connected and infinite, the key to studying recurrence and transience is to use Lemma 7.2.

Theorem 8.1. *The one-dimensional integer lattice is recurrent.*

Proof. Since the period is two,

$$\sum_{n=0}^{\infty} p_n(0,0) = \sum_{n=0}^{\infty} P(X_{2n} = 0 \,|\, X_0 = 0) = \sum_{n=0}^{\infty} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}.$$

Using Stirling's formula

$$n! \sim \sqrt{2\pi n}\,(n/e)^n \quad \text{as} \quad n \to \infty \tag{8.2}$$

to estimate the previous sum, we get

$$\binom{2n}{n}\left(\frac{1}{2}\right)^{2n} = \frac{(2n)!}{n!n!}\left(\frac{1}{2}\right)^{2n} \sim \frac{\sqrt{4\pi n}}{2\pi n}\,\frac{(2n/e)^{2n}}{(n/e)^{2n}}\left(\frac{1}{2}\right)^{2n} = \frac{1}{\sqrt{\pi n}} \tag{8.3}$$

which implies recurrence according to Lemma 7.2. □

For a probabilistic proof of Stirling's formula (8.2) based on the central limit theorem with minimal calculations, we refer the reader to Exercise 3.5. Here is another more elegant and less technical proof based on a symmetry argument.

Proof of Theorem 8.1. Assume without loss of generality that the random walk first jumps one unit to the right. Then, the event that the walk never returns to state zero, that we call B, is the event that

- it reaches state 2 before state 0 starting from 1, then
- state 4 before state 0 starting from 2, then
- state 8 before state 0 starting from 4, and so on.

Since, by obvious symmetry, all the events above have probability one-half, this must imply that the event B has probability zero. More precisely, letting

$$T_x = \inf\{n > 0 : X_n = x\} \quad \text{for all} \quad x \in \mathbb{Z}.$$

and using a first-step analysis and symmetry, we get

$$\begin{aligned} P_0(B) &= P_{-1}(B)P_0(X_1 = -1) + P_1(B)P_0(X_1 = 1) = P_1(B) \\ &= P_1(T_2 < T_0)P_2(T_4 < T_0)P_4(T_8 < T_0) \cdots P_{2^{n-1}}(T_{2^n} < T_0) \cdots \\ &= \lim_{n\to\infty}(1/2)^n = 0. \end{aligned}$$

This complete the proof. □

Following the same idea as in the previous proof, we now show that the two-dimensional integer lattice also is a recurrent graph.

Theorem 8.2. *The two-dimensional integer lattice is recurrent.*

Proof. To begin with, we observe that directed loops in two dimensions are composed of as many right and up arrows as left and down arrows. In addition, the number of right and left arrows must be the same and the number of up and down arrows must be the same. This implies that

$$p_{2n}(0,0) = \sum_{k=0}^{n}\binom{2n}{n}\binom{n}{k}^2\left(\frac{1}{4}\right)^{2n}.$$

Observe also that

$$\sum_{k=0}^{n}\binom{n}{k}^2 = \sum_{k=0}^{n}\binom{n}{k}\binom{n}{n-k} = \binom{2n}{n}$$

which follows from the fact that the last two terms both represent the number of subsets with cardinal n of a set with cardinal $2n$. In particular, using again the estimate proved in (8.3), we deduce that, for all n large,

$$p_{2n}(0,0) = \binom{2n}{n}^2 \left(\frac{1}{4}\right)^{2n} \sim \frac{1}{\pi n} \quad \text{and} \quad \sum_{n=0}^{\infty} p_n(0,0) = \infty$$

showing that the two-dimensional lattice is recurrent. □

As previously mentioned, the one- and two-dimensional random walks are in fact null recurrent, meaning in particular that there is no stationary distribution. This can be proved using the optional stopping theorem: the one-dimensional random walk is a martingale with bounded increments so, assuming by contradiction that

$$E_0(T_1) < \infty \quad \text{where} \quad T_1 = \inf\{n > 0 : X_n = 1\},$$

we can apply the optional stopping theorem to get

$$E_0(X_{T_1}) = E_0(X_0) = 0$$

which is not possible because the left-hand side is obviously equal to one. In particular, the mean time to go from zero to one must be infinite, which also implies that the mean recurrence time at state zero is infinite so the one-dimensional random walk is null recurrent. Applying this to each of the coordinates of higher-dimensional random walks, we deduce that these processes cannot be positive recurrent. This implies that the two-dimensional random walk is null recurrent and we now prove that, in higher dimensions, the process becomes in fact transient.

We focus on the three-dimensional case only and will use later the connection with electrical networks to deduce results about lattices in higher dimensions and more general graphs. The next theorem shows that symmetric random walks in $d = 3$ become transient, indicating that the spatial dimension plays the role of a critical parameter in the behavior of the process.

Theorem 8.3. *The three-dimensional integer lattice is transient.*

Proof. Observing that in three dimensions directed loops have as many moves going right, up, or forward as moves going left, down, or backward, and that the number of moves in opposite direction must be equal, we get

$$\begin{aligned} p_{2n}(0,0) &= \binom{2n}{n} \sum_{i+j+k=n} \binom{n}{i,j,k}^2 \left(\frac{1}{6}\right)^{2n} \\ &= \binom{2n}{n} \left(\frac{1}{12}\right)^n \sum_{i+j+k=n} \binom{n}{i,j,k}^2 \left(\frac{1}{3}\right)^n. \end{aligned}$$

Since in addition $i!\,j!\,k! \geq ((n/3)!)^3$ when $i+j+k=n$ and

$$\sum_{i+j+k=n} \binom{n}{i,j,k} \left(\frac{1}{3}\right)^n = \left(\frac{1}{3}+\frac{1}{3}+\frac{1}{3}\right)^n = 1$$

we deduce from (8.2) that, for all n large,

$$\begin{aligned} p_{2n}(0,0) &\leq \binom{2n}{n} \left(\frac{1}{12}\right)^n \frac{n!}{((n/3)!)^3} = \left(\frac{1}{12}\right)^n \frac{(2n)!}{n!\,((n/3)!)^3} \\ &\sim \left(\frac{1}{12}\right)^n \frac{\sqrt{4\pi n}}{\sqrt{2\pi n}\,\sqrt{(2/3)\pi n}^{\,3}}\ \frac{(2n/e)^{2n}}{(n/e)^n\,(n/3e)^n} = O(n^{-3/2}). \end{aligned}$$

This implies that the three-dimensional lattice is transient. □

8.2 Electrical networks

For arbitrary infinite connected graphs, there is no general theory to determine whether the graph is recurrent or transient. However, different graphs can be compared using a connection between random walks and electrical networks. This connection and the results collected above will give, with no calculation, additional interesting results. The idea is to think of the graph as an electrical network in which each edge has a certain conductance. For a complete picture of the connections between Markov chains and electrical networks, we refer to Doyle and Snell [24]. Here, we only focus on the relationship between recurrence and transience of symmetric random walks and electrical resistance. The general Markov chain associated to an electrical network is the process with transition probabilities

$$p(x,y) = P(X_{n+1}=y\,|\,X_n=x) = \frac{c(x,y)}{\sum_{z\sim x} c(x,z)}\ \mathbf{1}\{(x,y)\in E\} \tag{8.4}$$

where $c(x,y)$ is the conductance along the edge (x,y) and where $z \sim x$ means that both vertices are connected by an edge. The symmetric random walk is simply obtained by assuming that all the edges have the same conductance since in this case the transition probabilities in (8.4) reduce to

$$\begin{aligned} p(x,y) &= \mathbf{1}\{(x,y)\in E\}/\,\mathrm{card}\,\{z : (x,z)\in E\} \\ &= (\deg(x))^{-1}\,\mathbf{1}\{(x,y)\in E\}. \end{aligned}$$

We have the following result from [24] that we state without proof.

Theorem 8.4. *The symmetric random walk on a connected graph is recurrent if and only if the effective resistance of the associated electrical network between a nominated vertex and vertices at infinity is infinite.*

To explain the statement of the theorem, we note that the resistance of an edge is defined as the inverse of its conductance. Having a fixed vertex, the resistance between this vertex and vertices at infinity is defined as the limit as $n \to \infty$ of the resistance between the vertex and vertices which are at graph distance n. At least intheory, this resistance can be computed by applying successively the two basic rules illustrated in Figure 8.2, but in practice, the resistance between vertices of a complex network can be difficult, if not impossible, to compute. However, interesting qualitative results can be obtained using the following two observations.

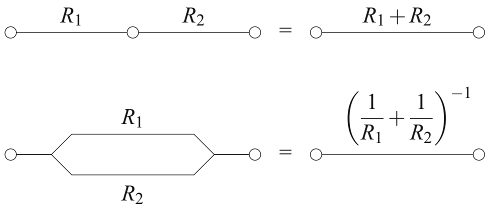


Fig. 8.2 The basic rules to compute electrical resistance.

- **Rayleigh's monotonicity law** — If the resistance of an edge is increased/decreased, the resistance between any two vertices can only increase/decrease.
- **Edge removal** — Removing an edge from an electrical network is equivalent to setting its conductance equal to zero, i.e., its resistance equal to infinity.

In particular, removing an edge from an electrical network can only increase the resistance between a nominated vertex and vertices at infinity. This and Theorem 8.4 further imply that any subgraph of a recurrent graph is recurrent and any supergraph of a transient graph is transient. In view of our analysis of lattices, subgraphs of the two-dimensional lattice are recurrent, whereas supergraphs of the three-dimensional lattice, which includes lattices in higher dimensions, are transient. To give another example, the connection with electrical networks has also been used in [41] to prove that the infinite percolation cluster of supercritical bond percolation in three dimensions, which is defined later in this textbook, is transient.

8.3 The infinite collision property

We now study another notion related to recurrence and transience called the **infinite collision property**. A connected graph is said to have this property if the number of collisions between two independent symmetric random walks (X_n) and (Y_n) starting from the same vertex is almost surely infinite, i.e.,

$$\operatorname{card}\{n : X_n = Y_n\} = \infty \quad \text{with probability one.}$$

One intuitively expects that a graph has the infinite collision property if and only if it is recurrent. Things, however, are more subtle even though both notions are indeed related. To begin with, we prove that, at least for graphs with bounded degree, the graph is recurrent if and only if the expected number of collisions is infinite.

Theorem 8.5 (Number of collisions). *Let $G = (V,E)$ be an infinite connected graph with bounded degree and let (X_n) and (Y_n) be two independent copies of the symmetric random walk starting at the same vertex x. Then,*

$$G \text{ is recurrent} \quad \text{if and only if} \quad E_x(\textstyle\sum_n \mathbf{1}\{X_n = Y_n\}) = \infty.$$

Proof. Comparing the probability that the random walk follows a specific path going in one direction and the opposite direction, we get

$$\begin{aligned} &\deg(z_0)\,P(X_i = z_i \text{ for } i = 0,1,\ldots,n) \\ &\qquad = \deg(z_0)\,(\deg(z_0)\cdots\deg(z_{n-1}))^{-1} \\ &\qquad = \deg(z_n)\,(\deg(z_n)\cdots\deg(z_1))^{-1} \\ &\qquad = \deg(z_n)\,P(X_i = z_{n-i} \text{ for } i = 0,1,\ldots,n). \end{aligned}$$

In particular, taking $(z_0, z_n) = (x,y)$ and summing over all the possible paths of length n connecting vertex x and vertex y, we deduce that

$$\deg(x)\,p_n(x,y) = \deg(y)\,p_n(y,x).$$

Then, by conditioning on the position at time n,

$$\begin{aligned} \deg(y)\,p_{2n}(x,x) &= \deg(y)\,\textstyle\sum_y P(X_{2n} = x \,|\, X_n = y)\,P(X_n = y \,|\, X_0 = x) \\ &= \deg(y)\,\textstyle\sum_y p_n(x,y)\,p_n(y,x) \\ &= \deg(x)\,\textstyle\sum_y p_n(x,y)\,p_n(x,y) \\ &= \deg(x)\,\textstyle\sum_y (p_n(x,y))^2. \end{aligned} \tag{8.5}$$

Finally, let $m = \sup_y \deg(y) < \infty$. According to (8.5) and the monotone convergence theorem, the expected number of collisions is

$$\begin{aligned} &E_x(\textstyle\sum_n \mathbf{1}\{X_n = Y_n\}) = \sum_n P_x(X_n = Y_n) \\ &\qquad = \textstyle\sum_n \sum_y P_x(X_n = Y_n = y) = \sum_n \sum_y (p_n(x,y))^2 \\ &\qquad \leq (m/\deg(y))\,\textstyle\sum_n \sum_y (p_n(x,y))^2 = (m/\deg(x))\,\sum_n p_{2n}(x,x) \end{aligned}$$

hence, if the expected number of collisions is infinite then the sum on the right-hand side is infinite and so the graph is recurrent by Lemma 7.2. Similarly,

$$\begin{aligned} E_x(\textstyle\sum_n \mathbf{1}\{X_n = Y_n\}) &\geq (1/\deg(y))\,\textstyle\sum_n \sum_y (p_n(x,y))^2 \\ &= (1/\deg(x))\,\textstyle\sum_n p_{2n}(x,x) \end{aligned}$$

which shows the converse. □

The condition on the expected number of collisions is somewhat weaker than the infinite collision property. In fact, there exist infinite connected graphs with bounded degree which are recurrent but for which the number of collisions is almost surely finite. For instance, the graph comb $(\mathbb{Z})$ depicted on the left-hand side of Figure 8.3 is a subgraph of the two-dimensional regular lattice and is therefore recurrent. Krishnapur and Peres [59] have proved however that this graph has the finite collision property, i.e., the number of collisions between two independent random walks starting at the same vertex is almost surely finite. This means in particular that, for this graph, the number of collisions is finite but has an infinite mean.

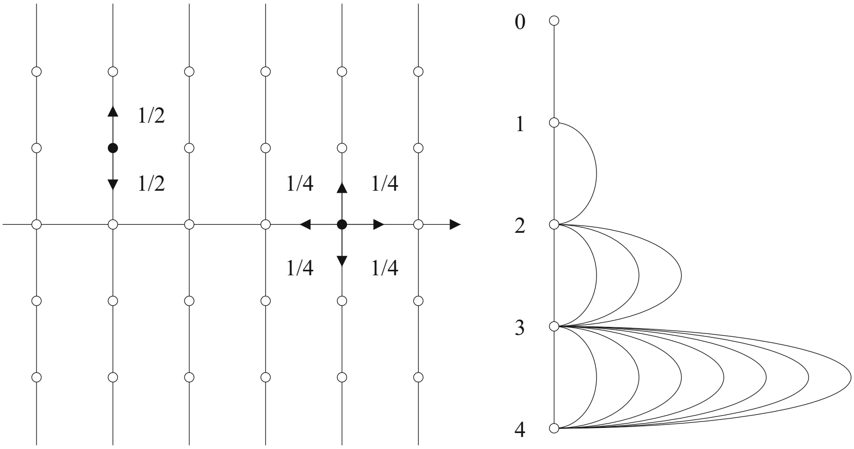


Fig. 8.3 Picture of comb $(\mathbb{Z})$ and the graph of Exercise 8.5.

According to Theorem 8.5, infinite connected graphs with bounded degree that are transient cannot have the infinite collision property. However, dropping the assumption on the boundedness, there exist connected graphs that are transient and have the infinite collision property (see Exercise 8.5).

In conclusion, focusing on connected graphs, finiteness implies positive recurrence of the symmetric random walk, while the main results collected in this chapter for infinite graphs are summarized in Figure 8.4.

8.4 Exercises

Exercise 8.1. For the random walk on the polygon with m sides, find

1. the expected number of steps to return to the initial position,
2. the probability that the random walk visits all the other states before returning to its initial position.

Hint: For the second part, think of the gambler's ruin chain.

Exercise 8.2. Recall that a chessboard is an 8×8 grid of squares and that a king can move one square horizontally, vertically, or diagonally.

1. Assuming that each of the legal moves is chosen uniformly at random at each time step, find the expected number of steps for the king to go from the top left corner of the chessboard back to this square.
2. Repeat the first question for a queen that can move any number of squares horizontally, vertically, or diagonally.

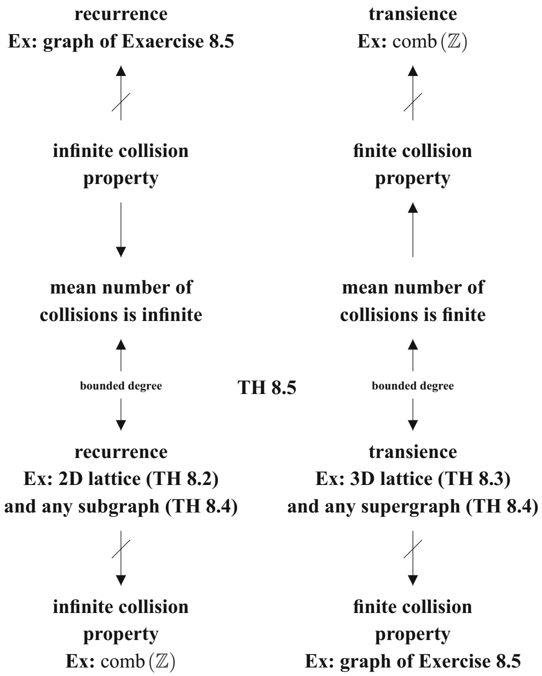

recurrence
Ex: graph of Exaercise 8.5
transience
Ex: comb $(\mathbb{Z})$
infinite collision property
finite collision property
mean number of collisions is infinite
mean number of collisions is finite
bounded degree
TH 8.5
bounded degree
recurrence
Ex: 2D lattice (TH 8.2)
and any subgraph (TH 8.4)
transience
Ex: 3D lattice (TH 8.3)
and any supergraph (TH 8.4)
infinite collision property
Ex: comb $(\mathbb{Z})$
finite collision property
Ex: graph of Exercise 8.5

Fig. 8.4 Summary of Chapter 8 for infinite graphs.

3. Repeat the first question for a knight that can move two squares horizontally or vertically and then one square in a perpendicular direction.
4. Finally, repeat the first question for a bishop using the fact that it can move any number of squares but only diagonally.

Hint: Use Example 7.3.

Exercise 8.3 (Asymmetric random walk). Consider the process on $\mathbb{Z}$ that jumps to the right with probability p and to the left with probability $q = 1 - p$. Use Stirling's formula to check whether the process is recurrent or transient.

Exercise 8.4. Consider the random walk on $\mathbb{Z}$ that jumps to the right with probability p and to the left with probability q where $p+q=1$, and let

$$T_x = \inf\{n > 0 : X_n = x\}$$

denote the first time the walk returns at site x.

1. Letting $\sigma_{-1} = P_1(T_0 < \infty)$, prove that either $\sigma_{-1} = 1$ or $\sigma_{-1} = q/p$.
2. Deduce that the random walk is recurrent when $p = q$.
3. Using $\sigma_1 = P_{-1}(T_0 < \infty) = 1$ and that the process is transient when $p > q$, deduce that the probability that the walk returns to its initial position is

$$\rho_{00} = P_0(T_0 < \infty) = 2\min(p,q) \quad \text{for all} \quad 0 \le p \le 1.$$

Exercise 8.5. Consider the infinite connected graph with unbounded degree depicted on the right-hand side of Figure 8.3 with vertex set the set of the natural integers and in which vertices x and $x+1$ are connected by 2^x edges.

1. Prove that this graph is transient.
2. Prove also that, though it is transient, it has the infinite collision property.

Exercise 8.6 (Simple exclusion process). Let $G = (V,E)$ be a connected graph with M vertices and let $m \le M$ be a positive integer. The simple exclusion process is the system of m symmetric random walks on the graph conditioned to be at different vertices. More precisely, the process starts with at most one particle per vertex and evolves as follows: At each time step, an edge (x,y) is chosen uniformly at random, and if there is a particle at x but not at y then the particle crosses the edge. Find the fraction of time each vertex is occupied in the long run.

Hint: Define a process (ξ_n) that keeps track of the configuration of the particles and use reversibility to find its stationary distribution.

Chapter 9
Poisson point and Poisson processes

This chapter starts with a general description of Poisson point processes. These processes are defined from four natural axioms describing the spatial distribution of so-called Poisson points scattered homogeneously in a random manner across the d-dimensional Euclidean space. They are used for instance as a model for the distribution of stars and galaxies in three dimensions or the distribution of plants and trees in two dimensions. These processes are characterized by a single parameter called the intensity that can be seen as a measure of the density of Poisson points. It is proved that the number of Poisson points in every bounded Borel set is Poisson distributed with parameter proportional to the Lebesgue measure of this set, which explains the name of these processes.

Specializing in the $d = 1$ case and thinking of the one-dimensional space as the time axis, the process that counts the number of Poisson points since time zero is a continuous-time process called Poisson process. In this one-dimensional context, these processes are key to defining a temporal structure and are employed to construct general continuous-time Markov chains with countable state space but also interacting particle systems. In one dimension, it is proved that the distance or time between consecutive Poisson points are independent and exponentially distributed with parameter the intensity of the process. In particular, a Poisson process can be viewed as a stack of infinitely many independent exponential random variables. Motivated by this result, we prove a couple of properties of the exponential distribution that are useful to better understand Poisson processes, including the fact that this distribution is characterized by its lack of memory.

To conclude, we state and prove the main three properties of Poisson (point) processes, which can be loosely described as follows.

Superposition — Having n independent Poisson point processes, the superposition of all their Poisson points is again a Poisson point process. The intensity of this new process is equal to the sum of the n intensities.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_9

Thinning — Having one Poisson point process and a palette of n colors, painting each Poisson point independently color i with a fixed probability results in n independent Poisson point processes.

Conditioning — Having a bounded Borel set and a Poisson point process, given that the set contains n Poisson points, the conditional distribution of these points is uniform over the set. For simplicity, we only prove this property in one dimension using the connection with the exponential distribution.

These properties are also illustrated by a couple of applications.

Further reading

- One of the few books exclusively devoted to Poisson processes and Poisson point processes is Kingman [55].
- For general textbooks on stochastic processes covering the theory of Poisson processes with a number of examples, see [29, 50, 85, 87].

9.1 Poisson point process and the Poisson distribution

Throughout this section, we think of the d-dimensional Euclidean space as a universe where countably many so-called **Poisson points** are scattered in a random manner. There are multiple ways to define the spatial distribution of these Poisson points, but natural assumptions include that:

1. The number of Poisson points in nonoverlapping bounded regions of the universe are independent random variables.
2. The number of Poisson points in a given bounded region only depends upon the volume or more generally the Lebesgue measure of this region.
3. The probability that a bounded region with Lebesgue measure v contains one Poisson point is given by $\mu v + o(v)$ as $v \to 0$.
4. The probability that a bounded region with Lebesgue measure v contains at least two Poisson points is given by $o(v)$ as $v \to 0$.

The process defined from the four assumptions above is called the **Poisson point process** with parameter or intensity μ. Formally, this is a process

$$\begin{aligned} X &= \{X(B) : B \in \mathscr{B}(\mathbb{R}^d) \text{ bounded}\} \text{ where} \\ \mathscr{B}(\mathbb{R}^d) &= \text{the Borel } \sigma\text{-algebra on the } d\text{-dimensional Euclidean space} \\ X(B) &= \text{the number of Poisson points in the Borel set } B. \end{aligned}$$

From now on, we write $X = \mathscr{P}(\mu)$ for short to indicate that X is the Poisson point process with intensity μ. The four axioms above can be rewritten formally in terms of this process, which gives the following more rigorous definition.

Definition 9.1. The process $X = \mathscr{P}(\mu)$ if

(1) $X(B_1),\ldots,X(B_n)$ are independent whenever $B_i \cap B_j = \varnothing$ for $i \neq j$,

(2) the distribution of $X(B)$ only depends on the Lebesgue measure $\lambda(B)$,

(3) $P(X(B)=1) = \mu v + o(v)$ as $v = \lambda(B) \to 0$,

(4) $P(X(B) \geq 2) = o(v)$ as $v = \lambda(B) \to 0$.

The following theorem gives an alternative definition. This definition is often the one given in the literature on stochastic processes. The theorem gives in particular the probability mass function of $X(B)$ and shows that it is related to the Poisson distribution, which explains the name of the process.

Theorem 9.1. *The process $X = \mathscr{P}(\mu)$ if and only if*

(a) $X(B_1),\ldots,X(B_n)$ are independent whenever $B_i \cap B_j = \varnothing$ for $i \neq j$ and

(b) for every bounded Borel set B, we have $X(B) \sim \text{Poisson}\,(\mu\,\lambda(B))$.

Proof. We need to prove two implications.

Sufficient condition — Assume that X satisfies (a) and (b). Then, it is clear that conditions (1) and (2) in Definition 9.1 are satisfied. Moreover,

$$\begin{aligned} P(X(B)=1) &= \mu v e^{-\mu v} = \mu v\,(1-\mu v + o(v)) = \mu v + o(v) \\ P(X(B) \geq 2) &= 1-(1+\mu v)\,e^{-\mu v} = 1-(1+\mu v)(1-\mu v + o(v)) = o(v) \end{aligned}$$

as $v = \lambda(B) \to 0$, which shows (3) and (4).

Necessary condition — Let $X = \mathscr{P}(\mu)$. Condition (a) is obvious. To check (b), fix a bounded Borel set B and a sequence of partitions of B

$$\{B_{1,n}, B_{2,n}, \ldots, B_{n,n}\} \ \text{ for all } \ n \geq 1 \quad \text{with} \quad \lambda(B_{i,n}) = v/n = \lambda(B)/n.$$

Let Ω_n be the event that $X(B_{i,n}) \geq 2$ for some $i = 1,2,\ldots,n$. Then,

$$\lim_{n\to\infty} P(\Omega_n) \leq \lim_{n\to\infty} nP(X(B_{1,n}) \geq 2) = \lim_{n\to\infty} n\,o(v/n) = 0.$$

Since in addition the variables $X(B_{i,n})$ are independent,

$$\begin{aligned} P(X(B)=k) &= \lim_{n\to\infty} P(X(B)=k \,|\, \Omega_n^c) \\ &= \lim_{n\to\infty} P(X(B_{1,n})+\cdots+X(B_{n,n}) = k \,|\, \Omega_n^c) \\ &= \lim_{n\to\infty} P(\text{card}\,\{i : X(B_{i,n}) = 1\} = k \,|\, \Omega_n^c) \\ &= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\mu v}{n}\right)^k \left(1-\frac{\mu v}{n}\right)^{n-k} = \frac{(\mu v)^k}{k!}\,e^{-\mu v} \end{aligned}$$

which, as desired, implies that $X(B) \sim \text{Poisson}\,(\mu\,\lambda(B))$. □

9.2 Poisson process and the exponential distribution

In this section, we define closely related cousins of Poisson point processes called Poisson processes. The latter can be easily constructed from the former by specializing in the one-dimensional case $d = 1$ and by thinking of the one-dimensional space as the time axis. The process that counts the number of Poisson points between time zero and time t is a continuous-time process, called Poisson process, which turns out to be a continuous-time Markov chain.

Definition 9.2 (Rate). We say that something occurs at rate μ if the number of occurrences in any Borel set $B \subset \mathbb{R}$ satisfies the axioms of Definition 9.1.

Definition 9.3 (Poisson process). The continuous-time process (X_t) is a Poisson process with rate μ whenever X_t is equal to the number of Poisson points in the interval $[0,t]$ for a Poisson point process with intensity μ, i.e.,

$$X_t = X([0,t]) \quad \text{where} \quad X = \mathscr{P}(\mu).$$

It directly follows from Theorem 9.1 that the process (X_t) is a Poisson process with intensity μ if and only if

(a) **Independent increments** — For all $t_1 < t_2 < \cdots < t_{2n}$,

$$X_{t_2} - X_{t_1}, X_{t_4} - X_{t_3}, \ldots, X_{t_{2n}} - X_{t_{2n-1}} \ \text{ are independent.}$$

(b) **Poisson increments** — For all $s < t$,

$$X_t - X_s \sim \text{Poisson}\,(\mu(t-s)).$$

In one dimension, Poisson points can be ordered and a natural question is: What is the distribution of the distance between two consecutive points? The next theorem shows that this distribution is exponential with the parameter the intensity.

Theorem 9.2. *Let (X_t) be a rate μ Poisson process. Then, times between consecutive jumps are independent exponential random variables with parameter μ.*

Proof. Let $\tau_0 = 0$ and let

$$\tau_i = \inf\{t : X_t = i\} \quad \text{and} \quad T_i = \tau_i - \tau_{i-1} \quad \text{for} \quad i \in \mathbb{N}^*$$

be respectively the times at which the process jumps and the corresponding interarrival times. Because the first jump occurs after time t if and only if there is no jump by time t, the distribution function of the time to the first jump is

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - P(X_t = 0) = 1 - e^{-\mu t}$$

which implies that $T_1 \sim \text{Exponential}\,(\mu)$. More generally, since the Poisson process has independent and stationary increments, we have

$$\begin{aligned} P(T_i > t \mid \tau_{i-1} = s) &= P(\text{no jump between } s \text{ and } s+t \mid \tau_{i-1} = s) \\ &= P(\text{no jump between } s \text{ and } s+t) = e^{-\mu t} \end{aligned}$$

for all $i > 1$. In particular,

$$F_{T_i}(t) = 1 - E(P(T_i > t \mid \tau_{i-1})) = 1 - e^{-\mu t}$$

and the proof is complete. Here is another approach that also emphasizes the connection between discrete and continuous time: let

$$Y_{i,n} = \text{number of jumps between times } (i-1)(t/n) \text{ and } i\,(t/n).$$

Then, it follows from Definition 9.1 that

$$\begin{aligned} P(T_1 > t) &= \lim_{n\to\infty} P(Y_{i,n} = 0 \text{ for all } i = 1,2,\ldots,n) \\ &= \lim_{n\to\infty} (1 - \mu t/n + o(t/n))^n = e^{-\mu t} \end{aligned}$$

which, as previously, gives $T_1 \sim \text{Exponential}\,(\mu)$. □

Motivated by the previous theorem, we now state properties of the exponential random variables that also help to understand Poisson processes.

Theorem 9.3. *Let T be a continuous random variable. Then, T is exponentially distributed with some parameter $\mu > 0$ if and only if*

$$T \textit{ is } \textbf{memoryless}: \ P(T > t+s \,|\, T > t) = P(T > s) \ \textit{ for all } \ s,t > 0. \qquad (9.1)$$

Proof. Assume that $T \sim \text{Exponential}\,(\mu)$. Then,

$$P(T > t+s \,|\, T > t) = \frac{P(T > t+s)}{P(T > t)} = \frac{e^{-\mu(t+s)}}{e^{-\mu t}} = e^{-\mu s} = P(T > s).$$

Now, assume (9.1) and let $\phi(t) = P(T > t)$ for all $t \geq 0$. Then,

$$\phi(t+s) = \phi(t)\,\phi(s) \quad \text{for all} \quad s,t > 0.$$

Because $\phi(1) \in (0,1)$, there exists a unique $\mu > 0$ such that $\phi(1) = e^{-\mu}$. The goal is to prove that T is the exponential random variable with parameter μ which, in turn, can be proved by showing that $\phi(t) = e^{-\mu t}$. We prove the result on the set of integers, then positive rational numbers, then positive real numbers.

Integers — Let $n \in \mathbb{N}^*$. By induction, we get

$$\phi(n) = \phi(n-1)\,\phi(1) = \cdots = (\phi(1))^n = e^{-\mu n}.$$

Rational numbers — Let $r = p/q \in \mathbb{Q}_+$. Since $p = qr \in \mathbb{N}^*$,

$$e^{-\mu p} = \phi(p) = \phi(qr) = (\phi(r))^q \quad \text{and} \quad \phi(r) = (e^{-\mu p})^{1/q} = e^{-\mu r}.$$

Real numbers — Let $t \in \mathbb{R}_+$ and fix a decreasing sequence of positive rational numbers (t_n) that converges to t. Since the function ϕ is right-continuous,

$$\phi(t) = \phi(\lim_{n\to\infty} t_n) = \lim_{n\to\infty} \phi(t_n) = \lim_{n\to\infty} e^{-\mu t_n} = e^{-\mu t}$$

therefore $F_T(t) = 1 - \phi(t) = 1 - e^{-\mu t}$ and $T \sim \text{Exponential}(\mu)$. □

Example 9.1 (Memoryless). In this example, we assume that a turtle needs s units of time to cross a street where cars pass at the times of a rate μ Poisson process. Accordingly, before crossing the street, the turtle waits until it sees that no car will come in the next s units of time. We are interested in the expected value of the time at which the turtle will be on the other side of the street.

To solve this problem, we first let W be the time the turtle waits before crossing the street and condition on whether the time X at which the next car passes is less or more than s. Note that the waiting time is zero if the next car passes in more than s units of time. Otherwise, by memoryless of the exponential distribution, we can use a restart argument to obtain that the conditional expected waiting time is equal to the sum of its unconditional counterpart and the time at which the next car passes, which can be expressed in equations as follows:

$$\begin{aligned} E(W) &= E(W \,|\, X < s)P(X < s) + E(W \,|\, X > s)P(X > s) \\ &= E(W \,|\, X < s)P(X < s) = (E(X \,|\, X < s) + E(W))P(X < s). \end{aligned}$$

Solving for $E(W)$, we get the expected value

$$E(W) = E(X \,|\, X < s)P(X < s)/P(X > s). \tag{9.2}$$

Using again memoryless and the fact that

$$E(X) = E(X \,|\, X < s)P(X < s) + E(X \,|\, X > s)P(X > s)$$

we also have

$$\begin{aligned} E(X \,|\, X < s) &= (E(X) - E(X \,|\, X > s)P(X > s))/P(X < s) \\ &= (E(X) - E(X + s)P(X > s))/P(X < s). \end{aligned} \tag{9.3}$$

Combining (9.2)–(9.3), we conclude that the expected value of the time at which the turtle is on the other side of the street is given by

$$\begin{aligned} E(W + s) &= (E(X) - E(X + s)P(X > s))/P(X > s) + s \\ &= E(X)/P(X > s) - E(X + s) + s \\ &= (1/\mu)(e^{\mu s} - 1) - s + s = (1/\mu)(e^{\mu s} - 1). \end{aligned}$$

Note that, if the turtle is very fast or if there are almost no cars, meaning that s is small or μ is small, then the expected time is

$$E(W + s) = (1/\mu)(e^{\mu s} - 1) \approx (1/\mu)(1 + \mu s - 1) = s$$

in accordance with our intuition. □

The next lemma gives one more property of the exponential random variable. This property can be viewed as a sort of dual version of the superposition property that will be stated later. In the lemma, the random variables T_i are independent exponential random variables with parameters μ_i.

Lemma 9.1. *Let* $Y_n = \min(T_1, T_2, \ldots, T_n)$. *Then,*

$$Y_n \sim \text{Exponential}\,(\textstyle\sum_k \mu_k) \quad \textit{and} \quad P(T_i = Y_n) = \mu_i / \textstyle\sum_k \mu_k.$$

Proof. First, we assume that $n = 2$. By independence,

$$P(Y_2 > t) = P(T_1 > t \text{ and } T_2 > t) = P(T_1 > t)\,P(T_2 > t) = e^{-(\mu_1 + \mu_2)t}$$

which implies that $Y_2 \sim \text{Exponential}\,(\mu_1 + \mu_2)$. In addition,

$$\begin{aligned} P(Y_2 = T_1) &= P(T_1 < T_2) = \int_0^\infty P(T_1 < T_2 \,|\, T_1 = t)\, f_{T_1}(t)\, dt \\ &= \int_0^\infty P(T_2 > t)\, \mu_1\, e^{-\mu_1 t}\, dt = \int_0^\infty \mu_1\, e^{-(\mu_1+\mu_2)t}\, dt = \frac{\mu_1}{\mu_1 + \mu_2}. \end{aligned}$$

This shows the result for two random variables. Now, assume the result for $n-1$ random variables, fix $1 \le i \le n$ and let

$$Y_n^i = \min(T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_n).$$

Since the result holds for $n-1$ random variables,

$$Y_n^i \sim \text{Exponential}\,(\mu) \quad \text{where} \quad \mu = \mu_1 + \cdots + \mu_{i-1} + \mu_{i+1} + \cdots + \mu_n.$$

Since also $Y_n = \min(T_i, Y_n^i)$ and that T_i and Y_n^i are independent, applying the result for two random variables, we conclude that

$$\begin{aligned} Y_n &\sim \text{Exponential}\,(\mu_i + \mu) = \text{Exponential}\,(\textstyle\sum_k \mu_k) \\ P(Y_n = T_i) &= P(T_i < Y_n^i) = \mu_i/(\mu_i + \mu) = \mu_i/(\textstyle\sum_k \mu_k). \end{aligned}$$

This shows the lemma by induction. □

Example 9.2 (Minimum spanning tree). Letting $G = (V, E)$ be a connected graph, a spanning tree of this graph is a subgraph with no cycle and the same vertex as the original graph G. Putting a weight on each edge of the graph G, the minimum spanning tree problem in graph theory consists in finding the spanning tree with the minimum weight. Note that, when weights assigned to different edges can be equal, the minimum spanning tree might not be unique.

To put this problem into a more concrete context, assume that roads must be constructed in order to connect n cities, i.e., in such a way that there is a path connecting any two cities. Assume in addition that the $n(n-1)/2$ possible roads connecting two cities have independent costs that are exponentially distributed with parameter one. We are interested in the expected value of the minimal cost c_n to connect all n

cities. Thinking of the cities as vertices, the possible roads as the edges of a complete graph, and the cost of each road as the weight of the corresponding edge, the problem is to find the minimum spanning tree of some randomly weighted complete graph. The general problem is difficult, but using the previous lemma and the memoryless property of the exponential distribution, we can answer this question in the case of three and four cities.

Because independent exponential random variables are almost surely different, the minimum spanning tree is unique with probability one. In the case of three cities, the minimum spanning tree is obtained by choosing the two edges with the smallest weight. Also, letting X_i be the cost of the ith cheapest road, and using memoryless as well as the previous lemma, we obtain

$$E(X_1) = 1/3 \quad \text{and} \quad E(X_2) = E(X_1) + 1/2 = 5/6$$

since, by memoryless, X_2 is equal to X_1 plus the minimum of two independent exponential random variables with parameter one. It follows that

$$E(c_3) = E(X_1 + X_2) = 1/3 + 5/6 = 7/6.$$

In the case of four cities, of the six possible roads, exactly three must be constructed to connect the cities. In addition, looking at the subgraphs with three edges, there are exactly four triangles for a total of 6 choose 3 possible subgraphs, so the fraction of such subgraphs that are triangles is

$$\binom{4}{3} \Big/ \binom{6}{3} = 4/20 = 1/5.$$

As previously, let X_i be the cost of the ith cheapest road. Because each of the six possible roads is equally likely to be the one with cost X_i, the probability that the three cheapest roads do not form a triangle is 4/5, in which case these three roads connect all four cities. In the case where the three cheapest roads form a triangle, the minimum spanning tree is obtained by choosing the two cheapest roads and the fourth cheapest one. In particular, the expected minimal cost is

$$E(c_4) = (4/5)\,E(X_1 + X_2 + X_3) + (1/5)\,E(X_1 + X_2 + X_4). \tag{9.4}$$

By memoryless and the previous lemma, we also have

$$E(X_i) = E(X_{i-1}) + \frac{1}{6-i+1} = \cdots = \frac{1}{6} + \frac{1}{5} + \cdots + \frac{1}{6-i+1}. \tag{9.5}$$

Combining (9.4)–(9.5), we conclude that

$$E(c_4) = \left(\frac{1}{6}\right) + \left(\frac{1}{6} + \frac{1}{5}\right) + \left(\frac{1}{6} + \frac{1}{5} + \frac{1}{4}\right) + \left(\frac{1}{5}\right)\left(\frac{1}{3}\right) = \frac{73}{60}$$

which gives an expected cost of about 1.2167. □

9.3 Superposition and thinning

After exploring the connections between Poisson processes and independent exponential random variables, we now return to general Poisson point processes. These processes in any dimension, as well as the special case of Poisson processes in one dimension, satisfy three important properties: superposition, thinning, and conditioning. In this section, we prove the first two properties in any spatial dimension. These properties are useful both in practice and to understand the theory of continuous-time Markov chains. Conditioning appears in more exotic examples and we only give a proof in one dimension in the next section using the connection between Poisson processes and the exponential random variable.

The superposition property states that, having n independent Poisson point processes with possibly different intensities $\mu_1, \mu_2, \ldots, \mu_n$, the set of points obtained by superposing all the Poisson points is again a Poisson point process. In addition, its intensity μ and the probability p_i that a Poisson point in this process comes from the ith process are given, respectively, by

$$\mu = \mu_1 + \mu_2 + \cdots + \mu_n \quad \text{and} \quad p_i = \mu_i / (\mu_1 + \mu_2 + \cdots + \mu_n).$$

In one dimension, in which case summing Poisson processes is the analog of superposing Poisson points, this can be directly proved from the connection between the process and the exponential distribution using also memoryless and Lemma 9.1. Indeed, by using memoryless, Poisson processes are characterized by the fact that the time to the next jump is exponentially distributed with the parameter the intensity. But the time to the next jump for a sum of independent Poisson processes is the next time one of the processes jumps, which is also the minimum of n independent exponential random variables. Since this time is exponentially distributed with the parameter the sum of the intensities according to Lemma 9.1, it follows that the sum of the processes is indeed the Poisson process for which the intensity is the sum μ of the intensity. In addition, according to the same lemma, the probability that the next jump results from a jump in the ith process is p_i. In short, the key is to use the connections among three different representations: superposing the points of Poisson point processes is the analog of summing Poisson processes which is also the analog of taking the minimum of exponential random variables. These connections are illustrated in Figure 9.1 in the case of two processes.

We now state and prove the superposition property for general Poisson point processes not restricted to the one-dimensional case. We focus on two processes only because the result for more processes easily follows by using induction. The result is illustrated in Figure 9.2. The key to the proof for general Poisson point processes is simply that the sum of independent Poisson random variables is again a Poisson random variable with parameter the sum of the parameters.

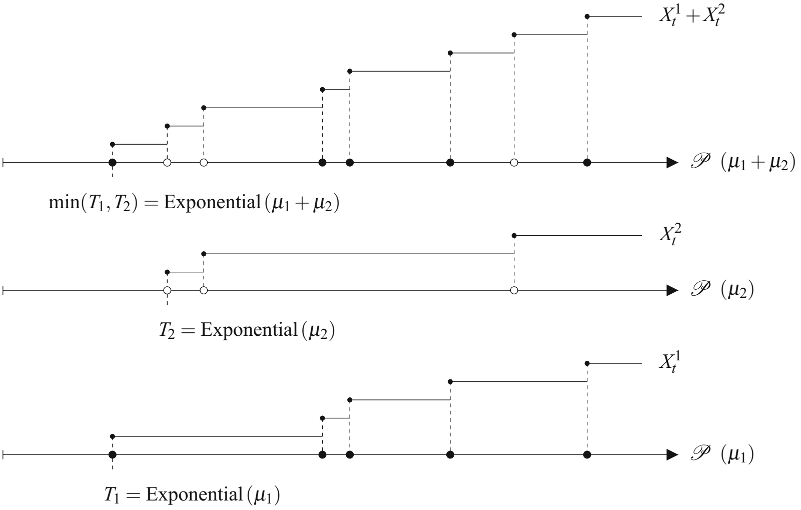

$X_t^1 + X_t^2$
$\mathscr{P}(\mu_1 + \mu_2)$
$\min(T_1, T_2) = \text{Exponential}(\mu_1 + \mu_2)$
X_t^2
$\mathscr{P}(\mu_2)$
$T_2 = \text{Exponential}(\mu_2)$
X_t^1
$\mathscr{P}(\mu_1)$
$T_1 = \text{Exponential}(\mu_1)$

Fig. 9.1 Relationship among superposition of independent Poisson point processes, sum of independent Poisson processes, and minimum of independent exponential random variables.

Lemma 9.2 (Superposition). *Assume that black and white points are distributed in space according to two independent Poisson point processes X_1 and X_2 that have respective intensities μ_1 and μ_2. Then,*

- *The resulting set of Poisson points is $X = \mathscr{P}(\mu_1 + \mu_2)$.*
- *Each point is independently black with probability $\mu_1/(\mu_1 + \mu_2)$.*

Proof. Condition (a) in Theorem 9.1 is inherited from its analog for the two processes X_1 and X_2 while condition (b) follows from the fact that if

$$Y_1 \sim \text{Poisson}(\mu_1) \quad \text{and} \quad Y_2 \sim \text{Poisson}(\mu_2)$$

are independent then

$$\begin{aligned} P(Y_1 + Y_2 = k) &= \sum_{i+j=k} P(Y_1 = i)\,P(Y_2 = j) = \sum_{i+j=k} \left(\frac{\mu_1^i}{i!}\,e^{-\mu_1}\right)\left(\frac{\mu_2^j}{j!}\,e^{-\mu_2}\right) \\ &= \sum_{i+j=k} \binom{i+j}{i} \frac{\mu_1^i\,\mu_2^j}{k!}\,e^{-(\mu_1+\mu_2)} = \frac{(\mu_1 + \mu_2)^k}{k!}\,e^{-(\mu_1+\mu_2)} \end{aligned}$$

showing that $Y_1 + Y_2 \sim \text{Poisson}(\mu_1 + \mu_2)$. To prove the second assertion, we observe that the conditional probability of one black Poisson point in the Borel set B given that there is exactly one Poisson point in this set is given by

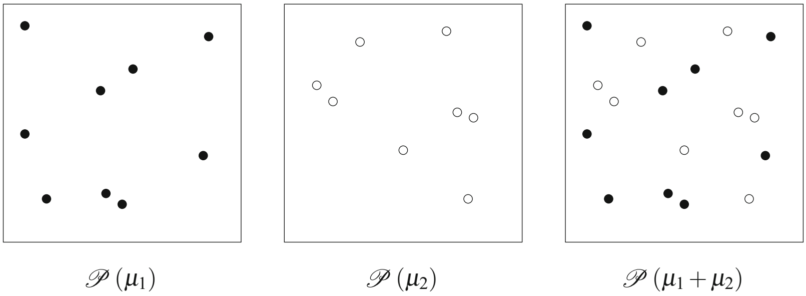


Fig. 9.2 Illustration of the superposition property.

$$\begin{aligned} P(X_1(B)=1\,|\,X(B)=1) &= P(X_1(B)=1)\,P(X_2(B)=0)/P(X(B)=1) \\ &= \mu_1\,e^{-\mu_1 v}\,e^{-\mu_2 v}/(\mu_1+\mu_2)\,e^{-(\mu_1+\mu_2)v} \\ &= \mu_1/(\mu_1+\mu_2). \end{aligned}$$

This completes the proof of the lemma. □

Example 9.3 (Superposition). Fix two positive integers a and b, and assume that a particle starting from the origin on the two-dimensional integer lattice jumps one step to the right or one step up at the times of independent Poisson processes with respective intensities μ_1 and μ_2. The objective is to find the value of the ratio μ_1/μ_2 that maximizes the probability that the particle crosses point (a,b). This event is depicted in the left picture of Figure 9.3.

Intuitively, the particle is more likely to visit (a,b) when the expected value of the time it takes to move a times to the right and the expected value of the time it takes to move b times up are the same. Writing down the expected values to guess the answer shows that our best candidate satisfies

$$a/\mu_1 = b/\mu_2 \quad \text{so} \quad \mu_1/\mu_2 = a/b.$$

This can be proved rigorously using the superposition property. Indeed, the superposition property implies that the particle jumps at rate $\mu_1+\mu_2$ and that, at each jump, it moves independently right or up with respective probabilities

$$p = \frac{\mu_1}{\mu_1+\mu_2} \quad \text{and} \quad q = \frac{\mu_2}{\mu_1+\mu_2}$$

so each given path going from the origin to (a,b) has probability $p^a\,q^b$ to be followed by the particle. In particular, the probability that the particle visits that point is simply the number of such paths times their common probability:

$$p(\mu_1,\mu_2) = \binom{a+b}{a}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^a\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^b.$$

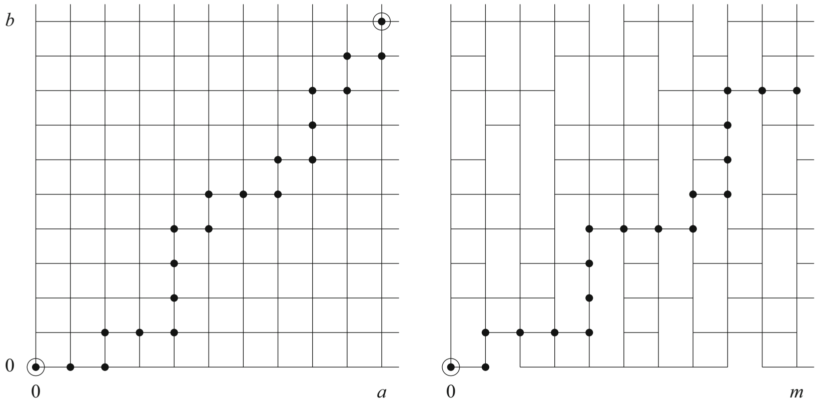


Fig. 9.3 Pictures of the particle's trajectory in Examples 9.3 and 9.4.

To check our guess, let $\phi(x) = x^a\,(1-x)^b$ and note that

$$\begin{aligned}\phi'(x) &= ax^{a-1}\,(1-x)^b - bx^a\,(1-x)^{b-1} \\ &= x^{a-1}\,(1-x)^{b-1}\,(a(1-x)-bx) \\ &= x^{a-1}\,(1-x)^{b-1}\,(a-(a+b)x) = 0\end{aligned}$$

when $x = a/(a+b)$. In conclusion, for all $\mu_1, \mu_2 > 0$,

$$p(\mu_1,\mu_2) = \binom{a+b}{a}\,\phi\left(\frac{\mu_1}{\mu_1+\mu_2}\right) \leq \binom{a+b}{a}\,\phi\left(\frac{a}{a+b}\right) = p(a,b)$$

so the probability is indeed maximal for $\mu_1/\mu_2 = a/b$. □

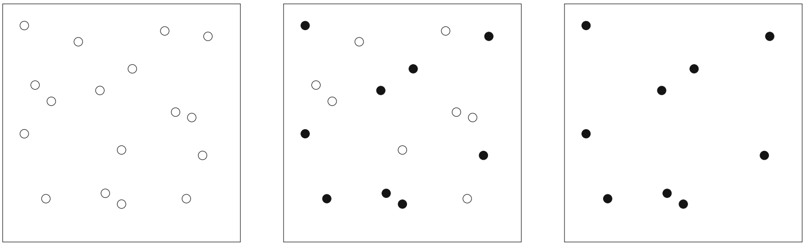
We now turn our attention to the property of thinning. Thinning can be viewed as the converse of superposition: We start with one Poisson point process and a palette of n colors, and independently paint each point color i with probability p_i. The result is a set of n independent Poisson point processes, the intensity of which has been reduced by the factor p_i. As for the superposition property, we only prove the result for two colors because the general result easily follows by induction. For an illustration of the thinning property, we refer the reader to Figure 9.4.

Lemma 9.3 (Thinning). *Assume that $X = \mathscr{P}(\mu)$ and paint each Poisson point independently black or white with respective probabilities p and q. Then,*

- *The set of black Poisson points is distributed according to $X_1 = \mathscr{P}(\mu p)$.*
- *The set of white Poisson points is distributed according to $X_2 = \mathscr{P}(\mu q)$.*

In addition, the Poisson point processes X_1 and X_2 are independent.

Proof. The independence follows from the construction. By symmetry, it suffices to study the distribution of black points only. Fix a bounded Borel set B and to simplify the notation let $\alpha = \mu\,\lambda(B)$. Letting $X_1(B)$ be the number of black Poisson points in the set B, by conditioning on the total number of points, we get



$\mathscr{P}(\mu)$ black with probability p $\mathscr{P}(\mu p)$

Fig. 9.4 Illustration of the thinning property.

$$P(X_1(B)=k\,|\,X(B)=n)=P(\text{Binomial}\,(n,p)=k)=\binom{n}{k}\,p^k\,(1-p)^{n-k}.$$

Using in addition the decomposition

$$P(X_1(B)=k)=\textstyle\sum_{n\geq k}P(X_1(B)=k\,|\,X(B)=n)\,P(X(B)=n)$$

we deduce that

$$\begin{aligned}P(X_1(B)=k)&=\sum_{n=k}^{\infty}\binom{n}{k}\,p^k\,(1-p)^{n-k}\,\frac{\alpha^n}{n!}\,e^{-\alpha}=\frac{p^k}{k!}\,e^{-\alpha}\sum_{n=k}^{\infty}\frac{\alpha^n\,(1-p)^{n-k}}{(n-k)!}\\&=\frac{(\alpha p)^k}{k!}\,e^{-\alpha}\sum_{n=k}^{\infty}\frac{(\alpha(1-p))^{n-k}}{(n-k)!}=\frac{(\alpha p)^k}{k!}\,e^{-\alpha}\sum_{n=0}^{\infty}\frac{(\alpha(1-p))^n}{n!}\\&=\frac{(\alpha p)^k}{k!}\,e^{-\alpha}\,e^{\alpha(1-p)}=\frac{(\alpha p)^k}{k!}\,e^{-\alpha p}.\end{aligned}$$

The lemma then follows from Theorem 9.1. □

Example 9.4 (Thinning). Fix $p\in(0,1)$ and consider the subgraph of the lattice in two dimensions obtained by independently removing each horizontal edge with probability $q=1-p$. Assume that a particle on this graph jumps at rate μ, going one step to the right if it can cross an edge and going up otherwise. To give a concrete application of this model, one can think of the subgraph as a street network and the particle as a car with a driver who wants to go east. However, each street oriented east–west is a one-way street going west with probability q. When the driver meets such a street, he decides to go north. We are interested in the expected value of the time T_m and the distance D_m the driver has to travel to move m units east.

Returning to the particle interpretation, thinning implies that the particle jumps to the right or up at the times of independent Poisson processes with respective intensities μp and μq. In particular, the time T_n it takes to move n steps to the right is the nth arrival time of a Poisson process with intensity μp therefore

$$E(T_m)=mE(T_1)=m/\mu p\quad\text{for all}\quad n\in\mathbb{N}^*.$$

Letting (X_t) be the Poisson process with intensity μq that keeps track of the number of vertical jumps of the particle and conditioning on T_m, we also get

$$\begin{aligned} E(D_m) &= m + E(X_{T_m}) = m + \int_0^\infty E(X_{T_m} \,|\, T_m = t)\, \phi_{T_m}(t)\, dt \\ &= m + \int_0^\infty \mu q t\, \phi_{T_m}(t)\, dt = m + \mu q E(T_m) = m + mq/p = m/p \end{aligned}$$

where ϕ_{T_m} is the density function of T_m.

The problem can also be solved backwards by finding first the expected distance and then deducing the expected time, without using Poisson processes. To do this, let Y_n be the number of up moves before the nth right move and note that Y_1 is a shifted geometric random variable with parameter p. In particular,

$$E(Y_m) = mE(Y_1) = m(-1 + E(\text{Geometric}(p))) = m(-1 + 1/p).$$

from which it follows that

$$E(D_m) = m + E(Y_m) = m/p \quad \text{and} \quad E(T_m) = (1/\mu)E(D_m) = m/\mu p$$

as we previously proved. □

9.4 The conditioning property

To conclude this chapter, we specialize in the one-dimensional case and prove the conditioning property for Poisson processes. This property states that, given that an interval contains n Poisson points, these points are uniformly distributed across the interval. More precisely, we have the following lemma.

Lemma 9.4 (Conditioning). *Let (X_t) be a rate μ Poisson process,*

$$S_i = \inf\{s : X_s = i\} \quad \textit{and} \quad U_i \sim \text{Uniform}(0,t) \textit{ be independent.}$$

Then, given the event that $X_t = n$,

$$\{S_1, S_2, \ldots, S_n\} = \{U_1, U_2, \ldots, U_n\} \quad \textit{in distribution.}$$

Proof. Let V_i be the ith smallest value among the n uniform random variables. To prove the lemma, it suffices to prove that

$$(S_1, S_2, \ldots, S_n) = (V_1, V_2, \ldots, V_n) \quad \text{in distribution.}$$

To do this, we define

$$\begin{aligned} f(s_1, s_2, \ldots, s_n) &= \text{the joint density function of } (V_1, V_2, \ldots, V_n) \\ h(s_1, s_2, \ldots, s_n) &= \text{the joint density function of } (S_1, S_2, \ldots, S_n). \end{aligned}$$

Then, we observe that, since

$$P((U_1,U_2,\dots,U_n)=(V_1,V_2,\dots,V_n))=1/n!$$

the joint density function $f(s_1,s_2,\dots,s_n)$ is given by

$$n!\prod_{i=1}^{n} f_{U_i}(s_i)=n!\prod_{i=1}^{n}\left(\frac{1}{t}\right)=n!\left(\frac{1}{t}\right)^n \quad \text{for all} \quad 0<s_1<\cdots<s_n<t.$$

In other respects, since the interarrival times $T_i=S_i-S_{i-1}$ are independent exponential random variables with rate μ, the conditional density function of the arrival times given that there are exactly n arrivals by time t is equal to

$$\begin{aligned} h(s_1,s_2,\dots,s_n\,|\,X_t=n) &= h(s_1,s_2,\dots,s_n \text{ and } X_t=n)/P(X_t=n)\\ &= f_{T_1}(s_1)\,f_{T_2}(s_2-s_1)\cdots f_{T_n}(s_n-s_{n-1})\,P(T_{n+1}>t-s_n)/P(X_t=n)\\ &= \mu\,e^{-\mu s_1}\,\mu\,e^{-\mu(s_2-s_1)}\cdots\mu\,e^{-\mu(s_n-s_{n-1})}\,e^{-\mu(t-s_n)}/P(X_t=n)\\ &= \mu^n\,e^{-\mu t}/P(X_t=n)=n!\,(1/t)^n=f(s_1,s_2,\dots,s_n)\end{aligned}$$

for all $0<s_1<\cdots<s_n<t$. This completes the proof. □

Example 9.5 (Conditioning). Let $0<s<t$ be two times and assume that, between time zero and time t, customers arrive at a waiting room in accordance with a Poisson process with intensity μ. All the customers arriving by time s are simultaneously served at time s while all the ones arriving after that time are simultaneously served at time t. Time t being fixed, how should we choose s in order to minimize the expected value of the overall waiting time of all the customers?

The optimal time s can be determined using the conditioning property of the underlying Poisson process. Let X_s and W_- be, respectively, the number of customers and the overall waiting time of the customers arriving before time s. The number of such customers is Poisson distributed with mean μs. In addition, given that there are n customers arriving in this time interval, the n arrival times are uniformly distributed in $(0,s)$. That is, labeling these customers uniformly at random $1,2,\dots,n$, and letting W_i be the waiting time of customer i, we have

$$W_i\sim\text{Uniform}\,(0,s)\quad\text{for all}\quad i=1,2,\dots,n.$$

Putting things together, we deduce that

$$\begin{aligned} E(W_-) &= E\,(E(W_-\,|X_s))=\textstyle\sum_n E(W_-\,|X_s=n)\,P(X_s=n)\\ &= \textstyle\sum_n E(W_1+W_2+W_3+\cdots+W_n)\,P(X_s=n)\\ &= \textstyle\sum_n n\,E(W_1)\,P(X_s=n)=E(X_s)\,E(W_1)\\ &= (\mu s)(s/2)=\mu s^2/2.\end{aligned}$$

Reasoning similarly for the second time interval (s,t), and letting W be the overall waiting time of all the customers up to time t, we deduce that

$$E(W)=\mu s^2/2+\mu(t-s)^2/2.$$

Differentiating with respect to s, we get

$$\frac{d}{ds}\left(\frac{\mu s^2}{2}+\frac{\mu(t-s)^2}{2}\right)=\frac{2\mu s}{2}-\frac{2\mu(t-s)}{2}=\mu(2s-t)$$

which is zero if and only if $s=t/2$. We easily check that this is a global minimum so the overall expected waiting time is minimized at $s=t/2$. □

9.5 Exercises

Poisson point processes

Exercise 9.1 (Distance to the closest Poisson point). Prove that, for a Poisson point process with intensity μ in two dimensions, the expected distance between the origin and the closest Poisson point is $1/(2\sqrt{\mu})$.

Exercise 9.2 (Forest fire model). Think of the Poisson points of a two-dimensional Poisson point process with intensity μ as the positions of trees that can be either alive or on fire. Initially, only one tree is on fire. At each time step, all the trees within distance one of a burning tree start burning forever.

1. Assume that each $x\in\mathbb{Z}^2$ is independently open with probability p. Prove that there is no infinite path of open sites such that each site is one of the eight nearest neighbors of the next site whenever $p<1/7$.
2. Deduce that, for $\mu<\ln(7/6)$, the ultimate number of burning trees is finite.

Memoryless property

Exercise 9.3. Assume that the times students 1 and 2 need to complete their homework assignment are independent exponential random variables with respective parameters μ_1 and μ_2. Find the probability that student 1 completes her assignment last if she starts t units of time before student 2.

Hint: Use the fact that the exponential distribution is memoryless.

Exercise 9.4. A doctor has scheduled two appointments 30 minutes apart. The amount of time appointments last are independent exponential random variables with a mean of 30 minutes. Assuming that both patients are right on time, find the expected amount of time the second patient spends in the doctor's office.

Hint: Use again memoryless.

Exercise 9.5. A child has a toy robot that needs two batteries to work. Assume that the child has a total of n such batteries, each of which has an independent lifetime that is exponentially distributed with the same mean m hours.

1. Compute the expected value of the time T the child can play with the robot.
2. Repeat this question when the toy robot needs k batteries to work.
3. Returning to the case where the toy robot needs $k = 2$ batteries, compute also the probability p_i that the ith battery put in use is the last one to fail.

Exercise 9.6. Assume that n batteries are simultaneously put in use and that battery i has an exponentially distributed lifetime with rate μ_i. Compute the expected value of the time to the first and to the second failure.

Hint: For the second failure, condition on the first battery that fails.

Exercise 9.7. Each day, a store buys strawberries for b dollars the kilogram and sell them for s dollars the kilogram with $b < s$. Assume that the daily demand X from customers is exponentially distributed with parameter μ.

1. For all $x > 0$, compute $E((X-x)^+)$.
2. Use the previous question to deduce the quantity of strawberries the store should buy to maximize the expected value of its profit.

Exercise 9.8. Consider a planetary system with n planets in orbit around a star. Assuming that the distance from the star to each of the planets are independent exponential random variables with rate μ, find the expected value of the distance between the star and the farthest planet.

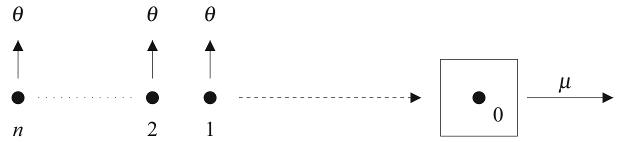
Hint: Use the fact that the exponential distribution is memoryless.

Exercise 9.9. A group of n people is to be assigned to n jobs, with one person assigned to each job. Assume that assigning person i to job j has a cost $X(i,j)$ and that all the costs are independent and exponential with rate one.

1. Compute the expected value of the total cost C, i.e., the sum of the n costs, when all $n!$ possible assignments are equally likely.
2. Same question when we first give job j to person i where $X(i,j)$ is the minimum cost, and repeat this procedure with the remaining pairs people/jobs.

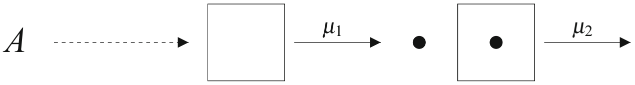
Superposition property

Exercise 9.10. Customers are in line to receive service provided by one server. The service times are independent exponential random variables with rate μ and each waiting customer will only wait an exponentially distributed amount of time with rate θ. The following picture gives a sketch of the system.



1. Find the probability p_n that the nth person in line is eventually served.
2. Compute the expected time μ_n until the nth person in line leaves the line either by entering service or departing without service.

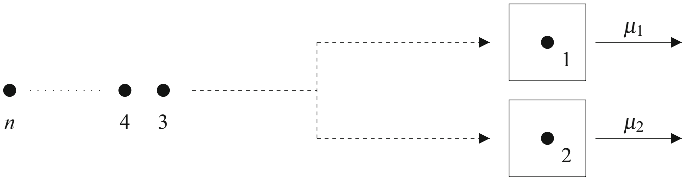
Exercise 9.11. Consider a two-server system where customers must go through server 1 and then server 2, and where the service time at server i is exponential with parameter μ_i. Customer A arrives and finds server 1 free and two customers at server 2, one in service and one waiting. The following picture gives a schematic representation of the system.



Find the expected value of the time T customer A spends in the system.

Hint: Condition on whether zero, one or two customers leave the system before customer A moves to server 2.

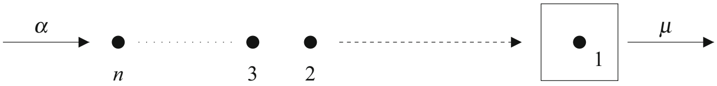
Exercise 9.12. Two cashiers (cashiers 1 and 2) must take care of n customers, with the service time for cashier i being exponentially distributed with rate μ_i. Whenever a cashier has completed her job with a customer, she begins taking care of the next one in line. The following picture gives a sketch of the system.



1. Compute the expected value of the time to take care of all n customers.
2. Find the probability that the last customer in line is also the last customer who leaves the system.

Hint: For the first part, condition on the last cashier who is working, and for the second part, condition on the cashier taking care of the last customer in line.

Exercise 9.13. Consider a single-server queuing system where customers arrive at rate α and service times are independent and exponential with rate μ, and suppose that a customer arrives and finds $n-1$ other customers in the system. The following picture gives a schematic representation of the system.



1. Compute the expected value of the number X of customers in the system at the time that customer leaves the system.
2. Find more generally the probability mass function of X.

Thinning and conditioning

Exercise 9.14. Let $a < 0 < b$ and assume that Poisson points are distributed on the real line according to a Poisson point process with intensity μ. Find the conditional probability that there are n Poisson points in the interval $(a, 0)$ given that there are exactly $2n$ Poisson points in the interval (a, b).

Exercise 9.15. Let X be a one-dimensional Poisson point process with intensity μ and assume that, if a Poisson point occurs at position s, it is painted independently of everything else black with probability $p(s)$. Use Lemma 9.4 to prove that the number of black Poisson points between zero and t is

$$Y_t \sim \text{Poisson}\left(\mu \int_0^t p(s)\, ds\right).$$

Exercise 9.16 (Number of encounters). Suppose that cars enter a one-way road with length d at the times of a rate one Poisson process and that all cars travel at a constant speed X chosen independently from a fixed distribution F_X. Then a red truck enters the road going at constant speed X_0.

1. Let F_T be the distribution function of the time T a car spends on the road. Assuming that the red truck enters the road at time 0, express the probability $p(s)$ that a car entering the road at time s either passes or is passed by the truck, an event that we call an encounter, using the function F_T.
2. Use Exercise 9.15 to deduce the mean number of encounters using F_T.
3. Conclude that the mean number of encounters is minimized when the speed X_0 of the red truck is equal to the median of the distribution F_X.

Additional problems

Exercise 9.17. A student leaves her apartment at 8:00am to take the bus. Compute the expected amount of time the student will need to wait at the bus stop in each of the following two situations.

1. The student needs exactly m_1 minutes to get from her apartment to the bus stop and buses arrive at rate $1/m_2$ per minute.
2. The student needs an exponential time with mean m_1 minutes to get to the bus stop and there is exactly one bus every m_2 minutes starting from 8:00am.

Exercise 9.18. Consider the following guessing game. Poisson events occur with intensity μ until a fixed deterministic time $T > 1/\mu$ and, each time such an event occurs, the player must decide whether or not to stop. The player wins if she stops at the last event. In case no event occurs, we assume that she always loses. A natural strategy is to stop at the first event to occur after some fixed time s. Prove that, using this strategy, the best winning probability is e^{-1}.

Exercise 9.19 (Coupon collector's problem). There are n different types of coupons and each time someone collects a coupon, it is independently of type i with probability p_i. We want to compute the expected value of the number N of coupons one needs to collect to have at least one coupon of each type. Assuming that coupons are collected at rate one and letting T be the time it takes to have the complete collection in this continuous-time setting, prove that

$$E(N) = E(T) = \int_0^\infty \left(1 - \prod_{i=1}^n (1 - e^{-p_i t}) \right) dt.$$

Chapter 10
Continuous-time Markov chains

As in discrete time, continuous-time Markov chains are stochastic processes in which the future depends on the past only through the present, or equivalently, given the present, past, and future are independent. Since there is no *next time* when time is continuous, the process is now characterized by transition rates instead of transition probabilities. Also, the analog of the transition matrix in continuous time is the so-called intensity matrix.

This chapter starts by exploring the connections between continuous-time and discrete-time Markov chains to understand how continuous-time processes can be constructed. The Markov property implies that, not only should the next state the process visits depend on the current state, but also that the time to the next jump should not depend on the time since the last jump, suggesting that the time between consecutive jumps are memoryless. In particular, we first prove that the sequence of states visited by the process evolves according to a discrete-time Markov chain, called the embedded chain, and that the time the process stays in a given state is exponentially distributed. The transition probabilities of the embedded chain as well as the parameter of the exponential holding times can be computed explicitly from the transition rates. This characterization gives an algorithm to construct a continuous-time Markov chain at all times except for pathological cases, called explosive processes, in which there are infinitely many jumps in a finite time.

Then, we study the convergence of continuous-time Markov chains. As in discrete time, the objective is to answer the following two questions.

Question 1 — Does the fraction of time spent in a given state converge to a limit as time goes to infinity, and if this is the case, what is the limit?

Question 2 — Does the probability of being in a given state converge to a limit as time goes to infinity, and if this is the case, what is the limit?

Following the same strategy as for discrete-time Markov chains, we introduce the concept of stationary distributions since these distributions again give the value of the two limits above. The transition probabilities are a solution to differential equations involving the intensity matrix called Kolmogorov's backward and forward equations. Using these equations, we prove that a distribution on the state space

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_10

is stationary if and only if it belongs to the kernel of the intensity matrix. As for discrete-time Markov chains, when the process is time-reversible, the stationary distribution is easier to compute.

After redefining communication classes, irreducibility, transience, positive and null recurrence in the continuous-time setting, we get a complete picture of the convergence for irreducible Markov chains, which answers the two questions above. The convergence theory in continuous time is in fact simpler. If all the states are transient or all the states are null recurrent, then the fraction of time spent and the probability of being in a given state both converge to zero, as in discrete time. If all the states are positive recurrent, then there is a unique stationary distribution and the fraction of time spent and the probability of being in a given state both converge to the mass of this state under the stationary distribution. In particular, the main difference with discrete-time Markov chains is that aperiodicity is no longer required to insure convergence of the transition probabilities. In fact, because time is continuous, the concept of period is no longer defined, but the fact that the process now jumps at times that are continuous random variables implies a well-mixing property that can be seen as the analog of aperiodicity in discrete time.

To conclude the chapter, we study a class of models known as continuous-time birth and death processes used in biology to model population dynamics. More precisely, we study the recurrence, transience, and convergence of general irreducible birth and death processes. Then, we look at the probability of survival of nonirreducible processes for which zero is an absorbing state. An example of such processes is the so-called simple birth and death process that will again appear at the end of this textbook because of its connection with the contact process.

Further reading

- Early references on continuous-time Markov chains are [12, 31, 34].
- For more recent textbooks with exercises, see [29, 50, 79, 85, 87].

10.1 Definition and main assumptions

This section focuses on the connections between continuous-time Markov chains and discrete-time Markov chains, showing in particular that a continuous-time Markov chain can be seen as a discrete-time Markov chain together with a collection of exponentially distributed holding times attached to each state. The main objective is to give an explicit way for constructing continuous-time Markov chains. As for their discrete-time analogs, continuous-time Markov chains assume that the future depends on the past only through the present. More precisely, the continuous-time process (X_t) where $t \in \mathbb{R}_+$ is a time homogeneous Markov chain with finite or countable state space S whenever

$$P(X_t \in B \,|\, \mathscr{F}_s) = P(X_t \in B \,|\, X_s) \quad \text{for all} \quad s < t \quad \text{and} \quad B \subset S$$

where $(\mathscr{F}_t)$ is the natural filtration of the process. We also assume throughout this chapter that the processes are right-continuous, meaning that

$$t \mapsto X_t(\omega) \text{ is a right-continuous step function for almost all } \omega \in \Omega. \tag{10.1}$$

Markov chains are more difficult to conceptualize in continuous time than discrete time because there is no successor (i.e., no next time) on the real line, so the evolution cannot be specified by one-step transition probabilities. Instead, we define the **transition probabilities** more generally for all time intervals by setting

$$p_t(x,y) = P(X_t = y \,|\, X_0 = x) \quad \text{for all} \quad x,y \in S \ \text{ and } \ t \in \mathbb{R}_+.$$

In discrete time, Theorem 7.1 shows that the transition probabilities can be expressed in terms of the one-step transition probabilities. In continuous time, the transition probabilities satisfy instead a certain differential equation that involves the so-called **transition rates**

$$c(x,y) = \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} p_\varepsilon(x,y) \quad \text{for all} \quad x \neq y \tag{10.2}$$

provided these limits exist and are finite, which we assume from now on. This condition basically states that for $x \neq y$ the transition probabilities are differentiable at time $t = 0$. It also implies that

$$\lim_{\varepsilon \downarrow 0} p_\varepsilon(x,x) = \lim_{\varepsilon \downarrow 0} (1 - \textstyle\sum_{z \neq x} p_\varepsilon(x,z)) = 1$$

whenever the state space is finite, but not necessarily in the context of continuous-time Markov chains with an infinite state space. Throughout this chapter, we assume more generally that the transition probabilities are also differentiable at time $t = 0$ when $x = y$ and that

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-1} (1 - p_\varepsilon(x,x)) = \textstyle\sum_{z \neq x} c(x,z) < \infty \quad \text{for all} \quad x \in S. \tag{10.3}$$

A continuous-time Markov chain with this property is called **conservative**. All the processes with finite or countable state space considered in this textbook and most of the models of interest in physics, biology, and sociology have this property. Condition (10.3) basically says that the process does not jump instantaneously from state to state and allows us to define the matrix

$$Q = (q(x,y))_{x,y \in S} \quad \text{where} \quad q(x,y) = \begin{cases} c(x,y) & \text{for} \quad x \neq y \\ -\sum_{z \neq x} c(x,z) & \text{for} \quad x = y. \end{cases} \tag{10.4}$$

This matrix is called the **intensity matrix** and can be seen as the analog of the transition matrix for discrete-time Markov chains. The motivation for the form of the coefficients on the diagonal will become clear soon after we prove a couple of results. We now exhibit the connections between discrete-time and continuous-time Markov chains, starting with an example.

Example 10.1. The Poisson process (X_t) with rate μ is an example of continuous-time Markov chain since, for all $s < t$, we have

$$P(X_t = x \mid \mathscr{F}_s) = \frac{\mu(t-s)^{x-X_s}}{(x-X_s)!}\, e^{-\mu(t-s)}$$

which is a function of X_s and is therefore $\sigma(X_s)$-measurable. More generally, letting (Y_n) be a discrete-time Markov chain with transition probability $\bar{p}$ and forcing the process to jump at the times of the Poisson process results again in a continuous-time Markov chain. Indeed,

$$P(Y_{X_t} = x \mid \mathscr{F}_s) = \sum_{n=0}^{\infty} \frac{\mu(t-s)^n}{n!}\, e^{-\mu(t-s)}\, \bar{p}_n(Y_{X_s}, x)$$

where the summand is the probability of n jumps in $t-s$ time units times the probability that the discrete-time Markov chain goes from state Y_{X_s} to state x in exactly n steps. This is again $\sigma(X_s)$-measurable therefore the process (Y_{X_t}) is indeed a continuous-time Markov chain. □

The reason why the process in the example is Markovian is somewhat hidden. This follows from the fact that the number of jumps of the Poisson process in a given time interval is independent of the number of jumps before that time interval, which is intrinsically related to the fact that the times between consecutive jumps are exponentially distributed and therefore are memoryless. Following along these lines, one can prove that turning a discrete-time Markov chain into a continuous-time process in such a way that the time the process stays in state x is exponentially distributed with parameter μ_x which may depend on state x is again Markovian. The next section is mostly devoted to proving that the converse also holds for conservative continuous-time Markov chains, i.e., when (10.3) is satisfied.

10.2 Connection with discrete-time Markov chains

To first understand continuous-time Markov chains heuristically, we let τ_n be the time of the nth jump which can be defined recursively as

$$\tau_0 = 0 \quad \text{and} \quad \tau_n = \inf\{t > \tau_{n-1} : X_t \neq X_{t-}\} \quad \text{for all} \quad n > 0.$$

To study the time between consecutive jumps, let

$$\phi(s,t) = P(\tau_{n+1} - \tau_n > t + s \mid \tau_{n+1} - \tau_n > t) \quad \text{for all} \quad s, t \geq 0.$$

Since, given the present, past, and future are independent, the time to the next jump (future) must be independent of the time since the last jump (past). Thinking of time $\tau_n + t$ as the present, this means that $t \mapsto \phi(s,t)$ must be constant so

$$\begin{aligned} P(\tau_{n+1} - \tau_n > t + s \,|\, \tau_{n+1} - \tau_n > t) &= \phi(s,t) = \phi(s,0) \\ &= P(\tau_{n+1} - \tau_n > s \,|\, \tau_{n+1} - \tau_n > 0) = P(\tau_{n+1} - \tau_n > s) \end{aligned}$$

showing that the time between consecutive jumps must be memoryless and so exponentially distributed by Theorem 9.3. Similarly, the next state the process jumps to must only depend on the current state of the process. In equations,

$$P(X_{\tau_{n+1}} = x \,|\, \mathscr{F}_{\tau_n}) = P(X_{\tau_{n+1}} = x \,|\, X_{\tau_n})$$

suggesting that the sequence of states visited by the continuous-time process evolves according to a discrete-time Markov chain. In conclusion, a continuous-time Markov chain can be viewed as a modification of a discrete-time Markov chain in which times between consecutive jumps are exponentially distributed. The underlying discrete-time process is called the *embedded Markov chain*.

We now turn this heuristics into a proof giving also an explicit expression for the rate of the exponential holding times and the transition probabilities of the embedded Markov chain using the coefficients of the intensity matrix.

Theorem 10.1. *Let* $T = \inf\{s > t : X_s \neq X_t\}$. *Then,*

$$\begin{aligned} P(T - t > u \,|\, X_t = x) &= \exp(q(x,x)u) \\ P(X_T = y \,|\, X_t = x) &= -q(x,y)/q(x,x) \end{aligned} \tag{10.5}$$

provided x is not an absorbing state, i.e., $q(x,x) \neq 0$.

Proof. Since the process is Markovian, it is enough to prove the result for $t = 0$. The main key to the proof is to study the discrete-time process

$$Y_n^\varepsilon = X_{n\varepsilon} \quad \text{for all} \quad n \in \mathbb{N} \quad \text{where} \quad \varepsilon > 0 \text{ is small.}$$

Note that the discrete-time process (Y_n^ε) is a discrete-time Markov chain, which we think of as what we would know about the continuous-time process if we were only able to see it at a times multiple of the small parameter ε. This process is commonly called the **ε-skeleton**. Referring to the quote of French filmmaker Jean-Luc Godard: *The cinema is truth twenty-four times per second*, one can think of the continuous-time process as the *truth* and its skeleton as the sequence of the *film frames*. Now, define

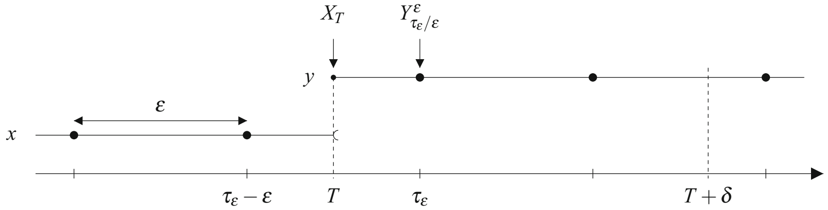
$$\tau_\varepsilon = \inf\{t \in \varepsilon\mathbb{N} : X_t \neq X_0 = x\} = \inf\{t \in \varepsilon\mathbb{N} : Y_t^\varepsilon \neq Y_0^\varepsilon = x\}.$$

Since the process is right-continuous, this time converges almost surely to time T. Indeed, it follows from (10.1) that, for almost all ω,

$$X_t = X_T \quad \text{for all} \quad t \in [T, T + \delta] \quad \text{for some} \quad \delta = \delta(\omega) > 0.$$

In particular, for all $\varepsilon \in (0, \delta)$,

$$\tau_\varepsilon - \varepsilon \leq T \leq \tau_\varepsilon \quad \text{and} \quad Y_{\tau_\varepsilon/\varepsilon}^\varepsilon = X_{\tau_\varepsilon} = X_T$$



X_T
$Y^\varepsilon_{\tau_\varepsilon/\varepsilon}$
y
ε
x
$\tau_\varepsilon - \varepsilon$
T
τ_ε
$T+\delta$

Fig. 10.1 Sample paths of the process (X_t) and its ε-skeleton (Y^ε_n).

as illustrated in Figure 10.1. We deduce that

$$\tau_\varepsilon \xrightarrow{a.s.} T \quad \text{and} \quad Y^\varepsilon_{\tau_\varepsilon/\varepsilon} \xrightarrow{a.s.} X_T \quad \text{as} \quad \varepsilon \downarrow 0. \tag{10.6}$$

Also, by the Markov property,

$$P_x(\tau_\varepsilon = \varepsilon n \text{ and } Y^\varepsilon_{\tau_\varepsilon/\varepsilon} = y) = p_\varepsilon(x,x)^{n-1}\, p_\varepsilon(x,y). \tag{10.7}$$

Both assertions follow from this equation by summing over all states or all times, also observing that, under our assumptions (10.2)–(10.3) and by definition of the intensity matrix (10.4), we have

$$\begin{aligned} p_\varepsilon(x,x) &= 1 - \textstyle\sum_{z\neq x} p_\varepsilon(x,z) = 1 + q(x,x)\,\varepsilon + o(\varepsilon) \\ p_\varepsilon(x,y) &= q(x,y)\,\varepsilon + o(\varepsilon) \end{aligned} \tag{10.8}$$

for all $x \neq y$ and when time $\varepsilon > 0$ is small.

Proof of the first assertion — Sum (10.7) over all states $y \neq x$ to obtain

$$P_x(\tau_\varepsilon = \varepsilon n) = p_\varepsilon(x,x)^{n-1}\,(1 - p_\varepsilon(x,x))$$

then take the limit as $\varepsilon \downarrow 0$ and use (10.6) and (10.8) to get

$$\begin{aligned} P_x(T > u) &= \lim_{\varepsilon\downarrow 0} P_x(\tau_\varepsilon > u) = \lim_{\varepsilon\downarrow 0} P_x(\tau_\varepsilon > \varepsilon(u/\varepsilon)) \\ &= \lim_{\varepsilon\downarrow 0} \textstyle\sum_{n>u/\varepsilon} p_\varepsilon(x,x)^{n-1}\,(1 - p_\varepsilon(x,x)) \\ &= \lim_{\varepsilon\downarrow 0} p_\varepsilon(x,x)^{u/\varepsilon} \textstyle\sum_n p_\varepsilon(x,x)^n\,(1 - p_\varepsilon(x,x)) = \lim_{\varepsilon\downarrow 0} p_\varepsilon(x,x)^{u/\varepsilon} \\ &= \lim_{\varepsilon\downarrow 0} (1 + q(x,x)\,\varepsilon)^{u/\varepsilon} = \exp(q(x,x)u) \end{aligned}$$

showing that $T \sim \text{Exponential}\,(-q(x,x))$.

Proof of the second assertion — Sum (10.7) over all times $n \geq 1$ to obtain

$$P_x(Y^\varepsilon_{\tau_\varepsilon/\varepsilon} = y) = \textstyle\sum_{n\geq 1} p_\varepsilon(x,x)^{n-1}\, p_\varepsilon(x,y) = p_\varepsilon(x,y)(1 - p_\varepsilon(x,x))^{-1}$$

then take the limit as $\varepsilon \downarrow 0$ and use (10.6) and (10.8) to get

$$P_x(X_T = y) = \lim_{\varepsilon\downarrow 0} P_x(Y^\varepsilon_{\tau_\varepsilon/\varepsilon} = y) = -q(x,y)/q(x,x)$$

which is the second assertion in (10.5). □

In words, the theorem says that, given that the process is in state x,

- the time to the next jump is exponential with rate $-q(x,x)$,
- the next state visited is state y with probability $-q(x,y)/q(x,x)$.

In particular, the sequence of states visited by the process is described by the discrete-time Markov chain with transition probabilities

$$\begin{aligned} p(x,y) &= -q(x,y)/q(x,x) \quad \text{when} \quad q(x,x) \neq 0 \\ &= \qquad 0 \qquad\qquad\ \text{when} \quad q(x,x) = 0. \end{aligned}$$

As previously mentioned, this discrete-time Markov chain is called the **embedded chain**. The properties above can also be understood in terms of competing exponential random variables that determine the time and target of the next jump. To explain this connection, for all $x \neq y$, let

$$T(x,y) \sim \text{Exponential}\,(q(x,y)) \ \text{ be independent.}$$

Then, given that the process is in state x at time $T(x,y)$, it jumps to state y, which, by the properties of the exponential distribution, implies that

- The time to the next jump is

$$T = \min_{z \neq x} T(x,z) \sim \text{Exponential}\,(\textstyle\sum_{z \neq x} q(x,z)) = \text{Exponential}\,(-q(x,x)).$$

- The probability that y is the next state visited is

$$\begin{aligned} P(X_T = y) &= P(T(x,y) = T) \\ &= P(T(x,y) = \min_{z \neq x} T(x,z)) = -q(x,y)/q(x,x). \end{aligned}$$

The previous properties show how to construct a continuous-time Markov chain from collections of independent Poisson processes at any time provided

$$\tau_\infty = \infty \quad \text{where} \quad \tau_\infty = \lim_{n \to \infty} \tau_n.$$

Note that $\tau_\infty = \infty$ when the transition rates are uniformly bounded. This is the case in particular for continuous-time Markov chains with finite state space. However, when the state space is infinite and the transition rates increase too fast, it is possible for time τ_∞ to be almost surely finite. Such processes are called **explosive** or continuous-time Markov chains with **explosion**.

Example 10.2 (Explosive process). Consider the continuous-time pure birth process with infinite state space $S = \mathbb{N}$ described by the transition rates

$$c(x,x+1) = \mu\, x^\alpha \quad \text{where} \quad \mu, \alpha > 0. \tag{10.9}$$

The special case $\alpha = 1$ is known as the **Yule process**. In this context and in view of the basic properties of Poisson processes, one can think of the process as keeping track of the number of individuals in a population where each individual independently gives birth to one offspring at the constant rate μ. Letting T_n be the time for the population to reach n individuals, which only differs from the time to the nth jump by a finite random time, we have

$$E(T_n \,|\, X_0 = 1) = \sum_{x=1}^{n-1} E(T_{x+1} \,|\, X_0 = x) = \sum_{x=1}^{n-1} \frac{1}{\mu x} \sim \frac{\ln(n)}{\mu}$$

as $n \to \infty$. Similarly, we have

$$\operatorname{Var}(T_n \,|\, X_0 = 1) = \sum_{x=1}^{n-1} \operatorname{Var}(T_{x+1} \,|\, X_0 = x) = \sum_{x=1}^{n-1} \left(\frac{1}{\mu x}\right)^2 \leq \frac{\pi^2}{6\mu^2}$$

for all $n > 1$. Since the mean goes to infinity while the variance is bounded, T_n converges almost surely to infinity. Indeed, by Chebyshev's inequality,

$$\begin{aligned} &\lim_{n\to\infty} P(T_n \leq (1-\varepsilon)E(T_n)) \\ &\quad \leq \lim_{n\to\infty} P(|T_n - E(T_n)| \leq \varepsilon E(T_n)) \\ &\quad \leq \lim_{n\to\infty} \operatorname{Var}(T_n)\,\varepsilon^{-2}\,(ET_n)^{-2} = 0 \end{aligned}$$

for all $\varepsilon > 0$. Since also T_n is increasing, it converges almost surely to infinity, so the Yule process is not explosive. Following the same reasoning, we prove more generally that, for all $\alpha \leq 1$, the process (10.9) is not explosive. In contrast, by monotone convergence, for all $\alpha > 1$,

$$E(\tau_\infty \,|\, X_0 = 1) = \sum_{x=1}^{\infty} E(T_{x+1} \,|\, X_0 = x) = \left(\frac{1}{\mu}\right) \sum_{x=1}^{\infty} \left(\frac{1}{x}\right)^{\alpha} < \infty$$

from which it follows that $T_\infty < \infty$ so the process is explosive. □

10.3 Stationary distribution

After exploring the connections between discrete-time and continuous-time Markov chains and now that we know how to construct a continuous-time Markov chain from its transition rates, we can start answering the two questions raised at the beginning of this chapter. More precisely, letting

$$V_t(y) = \int_0^t \mathbf{1}\{X_s = y\}\,ds \quad \text{for all} \quad (y,t) \in S \times \mathbb{R}_+$$

be the amount of time spent in state y by time t, do the limits

$$\lim_{t\to\infty} (1/t)\, V_t(y) \quad \text{and} \quad \lim_{t\to\infty} p_t(x,y)$$

exist or not? For explosive processes, it is clear that these limits exist and are equal to zero so we focus from now on on nonexplosive continuous-time Markov chains. Following the same strategy as for discrete-time Markov chains, the first step is to introduce stationary distributions and show how to compute them in practice since they give the possible values of the limits. The convergence theory will be explained in the next section, relying again on the concepts of communication classes, transience, and positive and null recurrence. The definition of a stationary distribution is essentially the same as for discrete-time Markov chains.

Definition 10.1 (Stationary distribution). A distribution π on the state space S is called a stationary distribution whenever

$$P_\pi(X_t = x) = \pi(x) \quad \text{for all} \quad (x,t) \in S \times \mathbb{R}_+$$

where P_π is the law of the process starting from the distribution π.

As in the discrete-time setting, one of the main tools to study the process, including the stationary distributions, is **Chapman–Kolmogorov's equations**. These equations are again proved by conditioning on the possible states at time s to break down the time interval into two pieces:

$$\begin{aligned} p_{s+t}(x,y) &= \textstyle\sum_{z\in S} P(X_{s+t} = y \,|\, X_s = z)\, P(X_s = z \,|\, X_0 = x) \\ &= \textstyle\sum_{z\in S} p_s(x,z)\, p_t(z,y). \end{aligned} \tag{10.10}$$

The stationary distribution is the solution to a linear system involving the intensity matrix instead of the transition matrix. This can be proved using the Kolmogorov's backward and forward equations below. These equations show that the transition probabilities are solutions to a system of ordinary differential equations again involving the intensity matrix.

Theorem 10.2 (Backward equations). *For a conservative continuous-time Markov chain, i.e., condition* (10.3) *holds, we have*

$$\frac{dp_t(x,y)}{dt} = \sum_{z\in S} q(x,z)\, p_t(z,y) = (Qp_t)(x,y) \quad \textit{for all} \quad x,y \in S. \tag{10.11}$$

Proof. Using Chapman–Kolmogorov's equations, we first get

$$\begin{aligned} p_{t+\varepsilon}(x,y) - p_t(x,y) &= \textstyle\sum_{z\in S} p_\varepsilon(x,z)\, p_t(z,y) - p_t(x,y) \\ &= \textstyle\sum_{z\neq x} p_\varepsilon(x,z)\, p_t(z,y) - p_t(x,y)\,(1 - p_\varepsilon(x,x)). \end{aligned} \tag{10.12}$$

Writting $S = \{x, x_1, x_2, \ldots\}$, for all $n \in \mathbb{N}^*$, the three sets $\{x\}$,

$$A_n = \{x_1, x_2, \ldots, x_n\} \quad \text{and} \quad B_n = \{x_{n+1}, x_{n+2}, \ldots\}$$

form a partition of the state space. On the one hand,

$$\begin{aligned}&\liminf_{\varepsilon\downarrow 0}\, \varepsilon^{-1} \textstyle\sum_{z\neq x} p_\varepsilon(x,z)\, p_t(z,y)\\ &\quad\geq \lim_{n\to\infty} \liminf_{\varepsilon\downarrow 0}\, \varepsilon^{-1} \textstyle\sum_{z\in A_n} p_\varepsilon(x,z)\, p_t(z,y)\\ &\quad= \lim_{n\to\infty} \textstyle\sum_{z\in A_n} q(x,z)\, p_t(z,y) = \sum_{z\neq x} q(x,z)\, p_t(z,y)\end{aligned} \tag{10.13}$$

where we use that A_n is finite. On the other hand,

$$\begin{aligned}&\limsup_{\varepsilon\downarrow 0}\, \varepsilon^{-1} \textstyle\sum_{z\neq x} p_\varepsilon(x,z)\, p_t(z,y)\\ &\quad\leq \limsup_{\varepsilon\downarrow 0}\, \varepsilon^{-1} \left(\textstyle\sum_{z\in A_n} p_\varepsilon(x,z)\, p_t(z,y) + \sum_{z\in B_n} p_\varepsilon(x,z)\right)\\ &\quad= \limsup_{\varepsilon\downarrow 0}\, \varepsilon^{-1} \left(\textstyle\sum_{z\in A_n} p_\varepsilon(x,z)\, p_t(z,y) + 1 - p_\varepsilon(x,x) - \sum_{z\in A_n} p_\varepsilon(x,z)\right)\\ &\quad= \textstyle\sum_{z\in A_n} q(x,z)\, p_t(z,y) - q(x,x) - \sum_{z\in A_n} q(x,z)\end{aligned}$$

where we again use that A_n is finite. Taking the limit as $n\to\infty$ and using that the process is conservative, we deduce that

$$\limsup_{\varepsilon\downarrow 0}\, \varepsilon^{-1} \textstyle\sum_{z\neq x} p_\varepsilon(x,z)\, p_t(z,y) \leq \sum_{z\neq x} q(x,z)\, p_t(z,y). \tag{10.14}$$

Combining (10.13)–(10.14), we get

$$\lim_{\varepsilon\downarrow 0}\, \varepsilon^{-1} \textstyle\sum_{z\neq x} p_\varepsilon(x,z)\, p_t(z,y) = \sum_{z\neq x} q(x,z)\, p_t(z,y)$$

which, together with (10.12), gives

$$\begin{aligned}&\lim_{\varepsilon\downarrow 0}\, \varepsilon^{-1} (p_{t+\varepsilon}(x,y) - p_t(x,y))\\ &\quad= \textstyle\sum_{z\neq x} q(x,z)\, p_t(z,y) + q(x,x)\, p_t(x,y) = (Qp_t)(x,y)\end{aligned}$$

for all $x,y\in S$ and proves (10.11). □

For Kolmogorov's forward equations, we need a stronger assumption.

Theorem 10.3 (Forward equations). *For a conservative continuous-time Markov chain such that the convergence in* (10.2) *is uniform in x,*

$$\frac{dp_t(x,y)}{dt} = \sum_{z\in S} p_t(x,z)\, q(z,y) = (p_t Q)(x,y) \quad \textit{for all} \quad x,y\in S. \tag{10.15}$$

Proof. This can be proved heuristically as for the backward equations but with conditioning on the state at time t instead of time ε. Ignoring the justification of exchanging limit and summation, we get formally

$$\begin{aligned}p_{t+\varepsilon}(x,y) - p_t(x,y) &= \textstyle\sum_{z\in S} p_t(x,z)\, p_\varepsilon(z,y) - p_t(x,y)\\ &= \textstyle\sum_{z\neq y} p_t(x,z)\, p_\varepsilon(z,y) - p_t(x,y)\,(1 - p_\varepsilon(y,y)).\end{aligned}$$

Then, recalling (10.8), dividing by ε and taking $\varepsilon\downarrow 0$,

$$\begin{aligned}\lim_{\varepsilon\downarrow 0}\, \varepsilon^{-1} (p_{t+\varepsilon}(x,y) - p_t(x,y)) &= \textstyle\sum_{z\neq y} p_t(x,z)\, q(z,y) + p_t(x,y)\, q(y,y)\\ &= \textstyle\sum_{z\in S} p_t(x,z)\, q(z,y) = (p_t Q)(x,y)\end{aligned}$$

for all $x, y \in S$. The limit and summation obviously commute when the state space is finite. For a countable state space, in addition to the arguments in the previous proof, one also needs to use the fact that the convergence in (10.2) is uniform in x, and we refer to [34, chapter 17] for a discussion on this aspect. □

Kolmogorov's equations can be seen as the analog of Theorem 7.1 since they give an implicit expression of the transition probabilities at any time using the transition rates. To solve these differential equations, one needs to compute all the powers of the intensity matrix. Indeed, solving equation (10.11) gives

$$p_t = e^{Qt} = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!} = \sum_{n=0}^{\infty} Q^n \, \frac{t^n}{n!}.$$

For this reason, Kolmogorov's equations can only be solved explicitly in a small number of cases, but they are useful in theory to show that the stationary distributions are solutions to a linear system. This is proved in the next theorem.

Theorem 10.4. *The distribution π is stationary if and only if $\pi Q = 0$.*

Proof. Note that π is stationary if and only if

$$\begin{aligned} \pi p_t(y) &= \textstyle\sum_x \pi(x)\, p_t(x,y) = \sum_x P(X_t = y \,|\, X_0 = x)\, P_\pi(X_0 = x) \\ &= P_\pi(X_t = y) = \pi(y) \quad \text{for all} \quad (y,t) \in S \times \mathbb{R}_+. \end{aligned} \tag{10.16}$$

In fact (10.16) shows more generally that πp_t is the distribution at time t starting from π. Now, assume that $\pi Q = 0$. Then, (10.11) implies that

$$(\pi p_t)' = \pi p_t' = \pi\,(Q p_t) = (\pi Q)\, p_t = 0 \quad \text{therefore} \quad \pi p_t = \pi p_0 = \pi$$

and it follows from (10.16) that the distribution π is stationary. Conversely, assume that the distribution π is stationary. Then, (10.15) and (10.16) imply that

$$\pi Q = (\pi p_t)\, Q = \pi\,(p_t Q) = \pi p_t' = (\pi p_t)' = \pi' = 0$$

which completes the proof. □

The theorem shows that finding the stationary distributions simply consists in finding the kernel of the intensity matrix. As in discrete-time, the algebra is simplified if we are lucky enough to have a time-reversible Markov chain. In continuous time, the detailed balance equation can be expressed by using the transition rates rather than the transition probabilities:

$$\pi(x)\, q(x,y) = \pi(y)\, q(y,x) \quad \text{for all} \quad x \neq y. \tag{10.17}$$

Again, any distribution π that satisfies the detailed balance equation is stationary since summing the right-hand side of (10.17) over all y gives

$$(\pi Q)(x) = \textstyle\sum_y \pi(y)\, q(y,x) = \sum_y \pi(x)\, q(x,y) = \pi(x) \sum_y q(x,y) = 0$$

so the previous theorem implies that π is stationary. When there is no such stationary distribution, Theorem 10.4 is the main tool to find one.

Example 10.3 (Symmetric random walk). Letting $G = (V, E)$ be a finite connected graph with m vertices, there are two natural ways in continuous time to define the symmetric random walk on this graph. One can define the dynamics from independent rate one Poisson processes positioned on either each edge or each vertex. More precisely, assume that a particle moves on the vertices and that

1. **Edge activation** — edges become active at rate one and if the particle is at x at the time edge (x, y) becomes active then it jumps to vertex y, or that
2. **Vertex activation** — vertices become active at rate one and if the particle is at x at the time vertex x becomes active then it jumps to a random neighbor.

The continuous-time process (X_t) that keeps track of the position of the particle induces in both contexts a continuous-time Markov chain and we are interested in the stationary distributions of this process.

The intensity matrix of the process looks too complicated so that we can collect all the stationary distributions using Theorem 10.4. However, with a little bit more of theory, we will soon be able to prove that the process has a unique stationary distribution π, and we can easily find this distribution by solving the detailed balance equation. Indeed, in the first context (edge activation), the particle jumps from vertex x to vertex y at rate one if the two vertices are connected so

$$q_1(x,y) = \mathbf{1}\{(x,y) \in E\} \quad \text{for all} \quad x, y \in V.$$

The detailed balance equation in this case is simply

$$\pi_1(x) = \pi_1(x)\, q_1(x,y) = \pi_1(y)\, q_1(y,x) = \pi_1(y) \quad \text{for all} \quad (x,y) \in E$$

which has solution $\pi_1 \equiv 1/m$, so the uniform distribution is a stationary distribution. In the second context (vertex activation), the particle jumps from vertex x to vertex y at rate $1/\deg(x)$ if the two vertices are connected so

$$q_2(x,y) = (\deg(x))^{-1}\, \mathbf{1}\{(x,y) \in E\} \quad \text{for all} \quad x, y \in V.$$

In this case, the detailed balance equation becomes

$$\pi_2(x)\, (\deg(x))^{-1} = \pi_2(x)\, q_2(x,y) = \pi_2(y)\, q_2(y,x) = \pi_2(y)\, (\deg(y))^{-1}$$

for all $(x,y) \in E$. In particular,

$$\pi_2(x) = \deg(x) / \textstyle\sum_y \deg(y) \quad \text{for all} \quad x \in V$$

is stationary. Note that $\pi_1 = \pi_2$ only for regular graphs. However, regardless of the degree distribution, both contexts lead to the same

$$-q(x,y)/q(x,x) = (\deg(x))^{-1}\, \mathbf{1}\{(x,y) \in E\} \quad \text{for all} \quad x, y \in V$$

so both processes have the same embedded chain, which is the discrete-time symmetric random walk introduced in Example 7.2. The fact that $\pi_1 \neq \pi_2$ in general is due to the fact that the first walk jumps from high-degree vertices at a larger rate, whereas the second walk always jumps at rate one. □

10.4 Limiting behavior

This section focuses on the convergence theory and gives general necessary and sufficient conditions for the existence and uniqueness of the stationary distribution and for convergence of the process to this distribution. As for discrete-time Markov chains, irreducibility and positive recurrence are the key ingredients. Since the arithmetic on the real line is quite different, the notion of period is no longer defined, but the fact that the process now jumps at times that are continuous random variables implies a well-mixing property that can be seen as the analog of aperiodicity in discrete time. To this extent, the convergence theory is somewhat simplified in continuous time. We now redefine the main ingredients in continuous time by simply imitating the analogous definitions for discrete-time Markov chains.

Definition 10.2 (Communication classes). Two (possibly identical) states $x, y \in S$ are said to communicate, which we write $x \leftrightarrow y$, whenever

$$p_s(x,y) > 0 \quad \text{and} \quad p_t(y,x) > 0 \quad \text{for some} \quad s,t \geq 0.$$

It follows from the definition that there exist

$$x_0 = x, x_1, \ldots, x_n = y \quad \text{such that} \quad c(x_{i-1}, x_i) > 0 \quad \text{for} \quad i = 1,2,\ldots,n.$$

In particular, two states communicate for the continuous-time Markov chain if and only if they communicate for the embedded chain so the results for discrete-time Markov chains imply that the relation $\leftrightarrow$ again induces a partition of the state space into communication classes and that this partition coincides with its counterpart for the embedded chain. As for discrete-time Markov chains, the process is said to be **irreducible** if it has only one communication class. An interesting consequence of the fact that the exponential random variable can be arbitrarily small/large is that the condition in Definition 10.2 holds in fact for all $s,t > 0$.

Lemma 10.1. *Assume that $x \leftrightarrow y$. Then $p_t(x,y) > 0$ for all $t > 0$.*

Proof. Fix a collection of states

$$x_0 = x, x_1, \ldots, x_n = y \quad \text{such that} \quad c(x_{i-1}, x_i) > 0 \quad \text{for} \quad i = 1,2,\ldots,n.$$

Then, for all $\varepsilon > 0$ small enough,

$$p_\varepsilon(x_{i-1}, x_i) = c(x_{i-1}, x_i)\,\varepsilon + o(\varepsilon) > 0 \quad \text{for} \quad i = 1,2,\ldots,n.$$

In particular, for $0 < \varepsilon < t/n$ small enough,

$$\begin{aligned} p_t(x,y) &\geq p_{t-n\varepsilon}(x,x)\, p_\varepsilon(x_0,x_1)\, p_\varepsilon(x_1,x_2) \cdots p_\varepsilon(x_{n-1},x_n) \\ &\geq \exp(q(x,x)(t-n\varepsilon))\, p_\varepsilon(x_0,x_1)\, p_\varepsilon(x_1,x_2) \cdots p_\varepsilon(x_{n-1},x_n) > 0 \end{aligned}$$

which completes the proof. □

The previous lemma is the key ingredient to compare the continuous-time Markov chain with a certain discrete-time Markov chain and deduce from the discrete-time results that we have already proved the main convergence result of this chapter. As previously explained, aperiodicity is out of the game in the continuous-time setting so we can answer the two questions raised at the beginning of this chapter at once. In particular, the following theorem can be seen as the continuous-time analog of both Theorems 7.4 and 7.7.

Theorem 10.5. *Let (X_t) be an irreducible continuous-time Markov chain. Then, we have the following alternative:*

1. *There is no stationary distribution and*

$$(1/t)\, V_t(y) \xrightarrow{a.s.} \lim_{t\to\infty} p_t(x,y) = 0 \quad \textit{for all} \quad x,y \in S.$$

2. *There is a unique stationary distribution π and*

$$(1/t)\, V_t(y) \xrightarrow{a.s.} \lim_{t\to\infty} p_t(x,y) = \pi(y) > 0 \quad \textit{for all} \quad x,y \in S.$$

Proof. Fix $\varepsilon > 0$. The key is to study the ε-skeleton of the continuous-time process using the results we have already collected for discrete-time Markov chains. The proof is divided into four steps.

Step 1 — The limit $\lim_{t\to\infty} p_t(x,y)$ always exists.

Let (Y_n^ε) be the ε-skeleton. Recall that this process is the discrete-time Markov chain with transition probabilities

$$\bar{p}(x,y) = p_\varepsilon(x,y) = P(X_\varepsilon = y \,|\, X_0 = x) \quad \text{for all} \quad x,y \in S.$$

Since (X_t) is irreducible, it follows from Lemma 10.1 that

$$\bar{p}(x,y) = p_\varepsilon(x,y) > 0 \quad \text{for all} \quad x,y \in S$$

so the ε-skeleton is irreducible and aperiodic. Theorem 7.7 and our analysis of discrete-time Markov chains then give the following alternative:

1. The ε-skeleton is transient or null recurrent and

$$\lim_{t\to\infty} p_t(x,y) = \lim_{n\to\infty} p_{n\varepsilon}(x,y) = 0 \quad \text{for all} \quad x,y \in S.$$

2. The ε-skeleton is positive recurrent. In this case, the process (Y_n^ε) has a unique stationary distribution π and we have

$$\lim_{t\to\infty} p_t(x,y) = \lim_{n\to\infty} p_{n\varepsilon}(x,y) = \pi(y) > 0 \quad \text{for all} \quad x,y \in S.$$

In either case, the limit $\lim_{t\to\infty} p_t(x,y)$ exists. The next two steps focus in detail on the case where the ε-skeleton is positive recurrent.

Step 2 — The distribution π is stationary for (X_t).

Recalling Chapman–Kolmogorov's equations

$$p_{s+t}(x,y) = \sum_{z\in S} p_s(x,z)\, p_t(z,y) \quad \text{for all} \quad x,y \in S$$

and taking the limit as $s \to \infty$, for all $(y,t) \in S \times \mathbb{R}_+$,

$$\begin{aligned} \pi p_t(y) &= \sum_{z\in S} \pi(z)\, p_t(z,y) = \sum_{z\in S} \lim_{s\to\infty} p_s(x,z)\, p_t(z,y) \\ &= \lim_{s\to\infty} \sum_{z\in S} p_s(x,z)\, p_t(z,y) = \lim_{s\to\infty} p_{s+t}(x,y) = \pi(y) \end{aligned}$$

The sum and limit indeed commute by the monotone convergence theorem since the summands are nonnegative. This implies that $\pi p_t = \pi$ so, by (10.16), the distribution π is stationary for the continuous-time process.

Step 3 — The distribution π is unique.

Let μ be a stationary distribution for the continuous-time process (X_t). Then $\mu p_t = \mu$ for all t again according to (10.16). In particular,

$$\begin{aligned} \mu(y) &= \lim_{t\to\infty} \mu p_t(y) = \lim_{t\to\infty} \sum_{z\in S} \mu(z)\, p_t(z,y) \\ &= \sum_{z\in S} \mu(z) \lim_{t\to\infty} p_t(z,y) = \sum_{z\in S} \mu(z)\, \pi(y) = \pi(y) \end{aligned}$$

for all $y \in S$, where the sum and limit again commute by monotone convergence. In conclusion, we have $\mu = \pi$, which shows uniqueness.

Step 4 — Convergence of the fraction of time spent in state y.

A remarkable consequence of the previous two steps is that both the fact that the ε-skeleton is positive recurrent and the expression of its stationary distribution are not sensitive to the choice of ε. In particular, using also Theorem 7.4, we get

$$(1/n)\,\mathrm{card}\{k = 0,1,\ldots,n-1 : X_{\varepsilon k} = y\} \xrightarrow{a.s.} \pi(y)$$

for all $\varepsilon > 0$ when the skeleton is positive recurrent. Therefore, the fraction of time the continuous-time process spends in state y converges almost surely with

$$\begin{aligned} \lim_{t\to\infty} (1/t)\, V_t(y) &= \lim_{t\to\infty,\varepsilon\downarrow 0} (\varepsilon/t)\,\mathrm{card}\{0 \le k < t/\varepsilon : X_{k\varepsilon} = y\} \\ &= \lim_{t\to\infty} (1/t)\,\mathrm{card}\{0 \le k < t : X_k = y\} = \pi(y). \end{aligned}$$

The fact that the fraction of time (X_t) spends in state y converges to zero when the ε-skeleton is not positive recurrent follows from the same argument.

Combining all four steps shows the theorem. □

Following the same approach as for discrete-time Markov chains, we now define recurrence and transience as well as the distinction between positive and null recurrence and explain how these concepts are related to the existence of a stationary distribution. In continuous time, to define the time of first return to state y, one should first let the process leave its initial state by setting

$$T_y = \inf\{t > 0 : X_{t-} \neq X_t = y\}.$$

Then, the probability that the continuous-time process starting from x ever returns to state y is again defined from the time of first return to state y as

$$\rho_{xy} = P_x(T_y < \infty) = P(T_y < \infty \,|\, X_0 = x).$$

As in discrete time, state x is said to be

$$\begin{array}{rll} \textbf{positive recurrent} & \text{when} & E_x(T_x) < \infty \\ \textbf{null recurrent} & \text{when} & \rho_{xx} = 1 \text{ but } E_x(T_x) = \infty \\ \textbf{transient} & \text{when} & \rho_{xx} < 1. \end{array}$$

We can again prove that a state x is recurrent if and only if the expected number of visits in state x starting from state x is infinite. In fact, recalling that τ_n refers to the time of the nth jump, at least for irreducible continuous-time Markov chains with no explosion, we have

$$\tau_n < \infty \quad \text{for all} \quad n \in \mathbb{N}^* \quad \text{and} \quad \tau_\infty = \lim_{n\to\infty} \tau_n = \infty$$

so a state is recurrent if and only if it is recurrent for its embedded chain. This, together with the fact that the communication classes of both processes also coincide, implies that recurrence and transience are again class properties. In addition, it follows from the definition that the fraction of time spent in a transient state converges to zero so, according to Theorem 10.5, irreducible transient Markov chains do not have any stationary distribution.

How the mean recurrence time $E_y(T_y)$ relates to the fraction of time spent in state y is a little more subtle since there is an obvious discrepancy between the continuous-time process and its embedded chain, and in fact one process can be positive recurrent and the other null recurrent as shown in Example 10.6 below. Indeed, in discrete time, it is implicitly assumed that the process spends one unit of time in a state each time it visits that state, whereas the expected amount of time the continuous-time Markov chain spends in state y each time it visits that state is given by $-1/q(y,y) > 0$. Using this observation, it can be proved that

$$\frac{1}{t}\int_0^t \mathbf{1}\{X_s = y\}\, ds \xrightarrow{a.s.} \lim_{t\to\infty} p_t(x,y) = -\frac{1}{q(y,y)\, E_y(T_y)}$$

which, together with Theorem 10.5, implies that

$$E_y(T_y) = -\frac{1}{q(y,y)\,\pi(y)} \neq \frac{1}{\pi(y)}$$

whenever there is a stationary distribution π. Invoking again Theorem 10.5, and more precisely the fact that a stationary distribution puts a positive mass on each state, we deduce that if one state is positive recurrent then all states are positive recurrent which is also equivalent to the existence of a unique stationary distribution. Putting all the pieces together, we obtain the following picture for irreducible continuous-time Markov chains.

For an irreducible Markov chain (X_t), we have the alternative:

1. All states are transient or all states are null recurrent in which case there is no stationary distribution and, for all $x, y \in S$,
$$(1/t)\,V_t(y) \xrightarrow{a.s.} \lim_{t\to\infty} p_t(x,y) = 0.$$
2. All states are positive recurrent in which case there is a unique stationary distribution π and, with probability one for all $x, y \in S$,
$$(1/t)\,V_t(y) \xrightarrow{a.s.} \lim_{t\to\infty} p_t(x,y) = \pi(y) = -(q(y,y)\,E_y(T_y))^{-1} > 0.$$

As for discrete-time Markov chains, a finite and irreducible continuous-time Markov chain is positive recurrent, which is useful in practice to check the existence and uniqueness of the stationary distribution. In addition, in the case where the state space is infinite, it is typically easier to check first the existence and uniqueness of the stationary distribution by using Theorem 10.4 and then deduce positive recurrence and the value of the mean recurrence time. Most of the results of this chapter are summarized in the diagram of Figure 10.2, which also describes the strategy to study an irreducible continuous-time Markov chain.

Example 10.4 (Symmetric random walks). We study the convergence of the two symmetric random walks introduced in the previous Example 10.3. Recall that these two processes keep track of the location of a particle moving on the vertex set of a finite connected graph and that the dynamics is defined from a collection of Poisson processes attached to the edges for the first process while these processes are attached to the vertices for the second process.

Since the graph is connected, there is a positive probability that the particle goes from any vertex to any another vertex in a finite time so both processes are irreducible. Since the graph is also finite, both processes are finite and therefore positive recurrent. In particular, there is a unique stationary distribution π which is the limit of the transition probabilities. Recalling the expressions of the stationary distributions obtained using reversibility, we get

$$\lim_{t\to\infty} p_t(x,y) = \begin{cases} \pi_1(y) = 1/\operatorname{card}(V) & \text{(edge activation)} \\ \pi_2(y) = \deg(y)/\sum_z \deg(z) & \text{(vertex activation)} \end{cases}$$

regardless of the initial position x of the particle. □

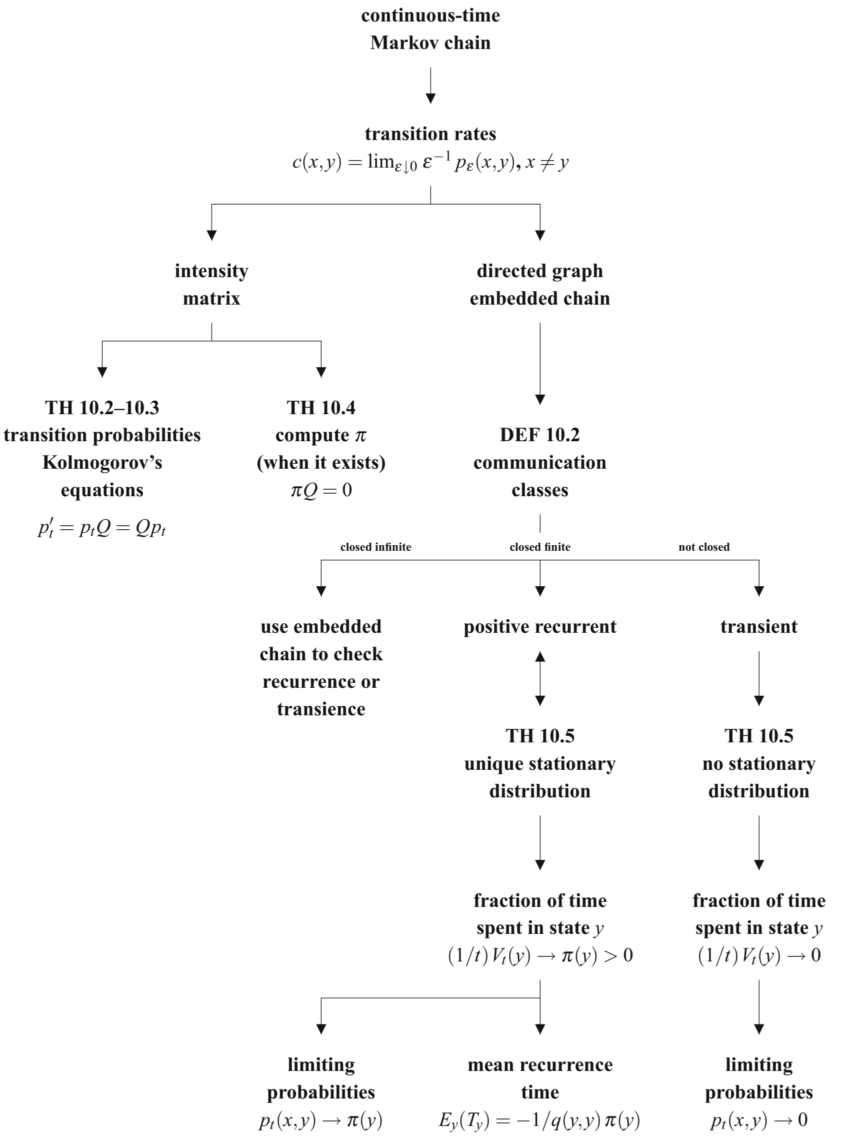

continuous-time Markov chain
transition rates
$c(x,y) = \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} p_\varepsilon(x,y), x \neq y$
intensity matrix
directed graph embedded chain
TH 10.2–10.3 transition probabilities Kolmogorov's equations
$p_t' = p_t Q = Q p_t$
TH 10.4 compute π (when it exists)
$\pi Q = 0$
DEF 10.2 communication classes
closed infinite
closed finite
not closed
use embedded chain to check recurrence or transience
positive recurrent
transient
TH 10.5 unique stationary distribution
TH 10.5 no stationary distribution
fraction of time spent in state y
$(1/t)\, V_t(y) \to \pi(y) > 0$
fraction of time spent in state y
$(1/t)\, V_t(y) \to 0$
limiting probabilities
$p_t(x,y) \to \pi(y)$
mean recurrence time
$E_y(T_y) = -1/q(y,y)\, \pi(y)$
limiting probabilities
$p_t(x,y) \to 0$

Fig. 10.2 Summary of Chapter 10.

Example 10.5 (The Jewish barber). The amount of time this barber takes to shave each of his customers are independent exponential random variables with mean three minutes, about the length of Brahms Hungarian Dance No. 5. Assume in addition that potential customers arrive at rate 30 per hour and that the barbershop has room for at most two customers. We are interested in the increase in the barber's business in the following two contexts:

1. The barber works twice as fast, i.e., the mean service time is 90 seconds.
2. Potential customers arrive twice as fast, i.e., at a rate of 60 per hour.

The key is to observe that the number of customers the barber can take care of is equal to the number of potential customers that indeed enter the barbershop which, in turn, is proportional to the rate at which potential customers arrive times the probability that a customer enters the barbershop. Since in addition a potential customer enters the shop if and only if there are fewer than two customers in the shop, the first step is to prove convergence of the fraction of time there are fewer than two customers and compute its limit.

The most natural Markov chain to consider keeps track of the number X_t of customers in the barbershop at time t. Since the customers arrive at the times of a Poisson process and since the service times are exponentially distributed, the process (X_t) is indeed a continuous-time Markov chain. In all three contexts, the process is finite and irreducible therefore there is a unique stationary distribution which is the limit of the fraction of time spent in each state. In the first context, the intensity matrix is given by

$$Q_1 = \begin{pmatrix} -30 & 30 & 0 \\ 20 & -50 & 30 \\ 0 & 20 & -20 \end{pmatrix}.$$

By Theorem 10.4, the stationary distribution is obtained by solving $\pi_1 Q_1 = 0$. Basic algebra gives the normalized vector $\pi_1 = (1/19)(4,6,9)$ from which we deduce that the fraction of potential customers entering the shop is

$$f_1 = \pi_1(0) + \pi_1(1) = 10/19.$$

In the other two scenarios, the intensity matrices are

$$Q_2 = \begin{pmatrix} -30 & 30 & 0 \\ 40 & -70 & 30 \\ 0 & 40 & -40 \end{pmatrix} \qquad Q_3 = \begin{pmatrix} -60 & 60 & 0 \\ 20 & -80 & 60 \\ 0 & 20 & -20 \end{pmatrix}.$$

Repeating the exact same reasoning, we find the following stationary distributions and corresponding fractions of potential customers that enter the shop:

$$\begin{aligned} \pi_2 = (1/37)(16,12,9) \quad &\text{and} \quad f_2 = \pi_2(0) + \pi_2(1) = 28/37 \\ \pi_3 = (1/13)(1,3,9) \quad &\text{and} \quad f_3 = \pi_3(0) + \pi_3(1) = 4/13. \end{aligned}$$

Accounting also for the rate at which potential customers arrive, we obtain the following values for the increase in the barber's business:

$$\begin{aligned} 30f_2/30f_1 &= f_2/f_1 = 266/185 \approx 1.438 \\ 60f_3/30f_1 &= 2f_3/f_1 = 76/65 \approx 1.169 \end{aligned}$$

in scenarios 2 and 3, respectively. □

10.5 Birth and death processes

Birth and death processes refer to a special class of continuous-time Markov chains that are natural to model the dynamics of populations in biology, though their applications are not restricted to this area. For this class of models, the state space is the set (or a subset) of all nonnegative integers and the transition rates satisfy

$$c(x,y) \neq 0 \quad \text{only if} \quad |x-y| = 1. \tag{10.18}$$

By convention, we write

$$c(x,x+1) = \beta_x \;\text{ for }\; x \geq 0 \quad \text{and} \quad c(x,x-1) = \delta_x \;\text{ for }\; x \geq 1 \tag{10.19}$$

which fully characterizes the process. As the name suggests, one can think of the state of the process as the number of individuals in a population, the rates β_x as birth parameters and the rates δ_x as death parameters, but the context can be more general, in which case birth and death events are no longer interpreted as actual births and deaths. For instance, the state can be thought of as the number of individuals in a store, in which case there is a birth each time someone enters the store and a death each time someone leaves the store.

Returning to the biological context for concreteness, the first natural question is whether state zero is recurrent, thus preventing the population from increasing too much, or transient, in which case one expects the population to grow indefinitely. If one of the parameters is equal to zero for some $x > 0$, then this can be solved by simply looking at the communication classes. State zero is clearly recurrent when $\beta_0 = 0$. Otherwise, we define

$$x_* = \inf\{x > 0 : \beta_x \, \delta_x = 0\}$$

which leads to the following two cases:

- when $\delta_{x_*} = 0$, regardless of β_*, the set $\{0, 1, \ldots, x_* - 1\}$ is a communication class which is not closed so state zero is transient,
- when $\delta_{x_*} \neq 0$, we must have $\beta_{x_*} = 0$ in which case $\{0, 1, \ldots, x_*\}$ is a finite communication class which is closed so state zero is recurrent.

We now assume that all the birth and death parameters in (10.19) are positive, which makes the process irreducible. In this case, the key quantities to check recurrence or transience, and the existence and uniqueness of a stationary distribution are

$$g(N) = \sum_{x=0}^{N-1} \frac{\delta_1 \, \delta_2 \cdots \delta_x}{\beta_1 \, \beta_2 \cdots \beta_x} \quad \text{and} \quad h(N) = \sum_{x=1}^{N} \frac{\beta_0 \, \beta_1 \cdots \beta_{x-1}}{\delta_1 \, \delta_2 \cdots \delta_x} \tag{10.20}$$

as shown in the following two results.

Theorem 10.6. *The process is recurrent if and only if*

$$\lim_{N\to\infty} g(N) = \sum_{x=0}^{\infty} \frac{\delta_1\,\delta_2\cdots\delta_x}{\beta_1\,\beta_2\cdots\beta_x} = \infty.$$

Proof. Fix an integer $N > 0$ and, for $x = 0, 1, \ldots, N$, let

$$p_N(x) = P(T_N < T_0 \,|\, X_0 = x) \quad \text{where} \quad T_y = \inf\{t : X_t = y\}.$$

Using a first-step analysis, i.e., conditioning on whether the first update is a birth or a death, and the basic properties of the exponential distribution, we get

$$(\beta_x + \delta_x)\, p_N(x) = \beta_x\, p_N(x+1) + \delta_x\, p_N(x-1) \quad \text{for all} \quad 0 < x < N.$$

Rearranging the terms, we deduce that

$$\begin{aligned} p_N(x+1) - p_N(x) &= \frac{\delta_x}{\beta_x}\,(p_N(x) - p_N(x-1)) \\ &= \cdots = \frac{\delta_1\delta_2\cdots\delta_x}{\beta_1\beta_2\cdots\beta_x}\,(p_N(1) - p_N(0)) \end{aligned}$$

Now, using the obvious boundary conditions $p_N(0) = 0$ and $p_N(N) = 1$, and writing $p_N(N)$ as the sum of a telescopic series, we obtain

$$\begin{aligned} 1 = p_N(N) &= \sum_{x=0}^{N-1} (p_N(x+1) - p_N(x)) \\ &= \sum_{x=0}^{N-1} \frac{\delta_1\delta_2\cdots\delta_x}{\beta_1\beta_2\cdots\beta_x}\, p_N(1) = g(N)\, p_N(1) \end{aligned}$$

so $p_N(1) = 1/g(N)$. In particular, the process is recurrent if and only if

$$\begin{aligned} 0 \text{ is recurrent} \quad &\text{if and only if} \quad \lim_{N\to\infty} P_1(T_N < T_0) = 0 \\ &\text{if and only if} \quad \lim_{N\to\infty} p_N(1) = 0 \\ &\text{if and only if} \quad \lim_{N\to\infty} g(N) = \infty \end{aligned}$$

which completes the proof. □

Theorem 10.7. *There is a unique stationary distribution if and only if*

$$\lim_{N\to\infty} h(N) = \sum_{x=1}^{\infty} \frac{\beta_0\,\beta_1\cdots\beta_{x-1}}{\delta_1\,\delta_2\cdots\delta_x} < \infty.$$

Proof. The idea is simply to find all the solutions of $\pi Q = 0$ where Q is the intensity matrix of the process. Note that condition (10.18) implies that the intensity matrix is a tridiagonal matrix; therefore, the system to be solved reduces to

$$\begin{aligned} \beta_0\,\pi(0) - \delta_1\,\pi(1) &= 0 \\ \beta_{x-1}\,\pi(x-1) - (\beta_x + \delta_x)\,\pi(x) + \delta_{x+1}\,\pi_{x+1} &= 0 \quad \text{for all } x > 0. \end{aligned}$$

The form of the system implies that the set of solutions is a one-dimensional vector space, and we easily prove by induction that

$$\pi(x) = \frac{\beta_0\,\beta_1 \cdots \beta_{x-1}}{\delta_1\,\delta_2 \cdots \delta_x}\,\pi(0) \quad \text{for all} \quad x > 0$$

so there is a unique stationary distribution if and only if

$$\lim_{N\to\infty} h(N) = \sum_{x=1}^{\infty} \frac{\beta_0\,\beta_1 \cdots \beta_{x-1}}{\delta_1\,\delta_2 \cdots \delta_x} = \sum_{x=1}^{\infty} \frac{\pi(x)}{\pi(0)} < \infty.$$

This completes the proof. □

From the theorem and the results of the previous section, it follows that the birth and death process is positive recurrent if and only if $\lim_{N\to\infty} h(N) < \infty$.

Example 10.6. As previously mentioned, we now give an example of a continuous-time Markov chain which is positive recurrent but whose embedded chain is null recurrent. Let (X_t) be the birth and death process with $\beta_0 = 1$, $\delta_0 = 0$,

$$\beta_x = x \quad \text{and} \quad \delta_x = x+1 \quad \text{for all} \quad x > 0.$$

In this case, we have

$$\sum_{x=1}^{\infty} \frac{\beta_0\,\beta_1 \cdots \beta_{x-1}}{\delta_1\,\delta_2 \cdots \delta_x} = \sum_{x=1}^{\infty} \frac{(x-1)!}{(x+1)!} = \sum_{x=1}^{\infty} \frac{1}{x(x+1)} < \infty$$

so, according to Theorem 10.7, the process is positive recurrent and its embedded chain must be recurrent. To prove that the latter is null recurrent, it suffices to prove that it does not have a stationary distribution. Observe that the transition probabilities of the embedded chain are given by $p(0,1) = 1$ and

$$\begin{aligned} p(x,x+1) &= \beta_x/(\beta_x+\delta_x) = x/(2x+1) \\ p(x,x-1) &= \delta_x/(\beta_x+\delta_x) = (x+1)/(2x+1) \end{aligned} \tag{10.21}$$

for all $x > 0$. In particular, the linear system $\pi P = \pi$ where P is the transition matrix of the embedded chain can be written as

$$\begin{aligned} \pi(0) &= (2/3)\,\pi(1) \\ \pi(x) &= p(x-1,x)\,\pi(x-1) + p(x+1,x)\,\pi(x+1) \quad \text{for all} \quad x > 0. \end{aligned}$$

The form of the system shows that the set of solutions is a one-dimensional vector space and we find a particular solution using reversibility, which gives

$$\pi(x) = \frac{p(x-1,x)}{p(x,x-1)}\,\pi(x-1) = \cdots = \prod_{z=1}^{x} \frac{p(z-1,z)}{p(z,z-1)}\,\pi(0).$$

Recalling (10.21), we also have

$$\prod_{z=1}^{x} \frac{p(z-1,z)}{p(z,z-1)} = \left(\frac{3}{2}\right) \prod_{z=2}^{x} \left(\frac{z-1}{z+1}\right)\left(\frac{2z+1}{2z-1}\right) = \frac{2x+1}{x(x+1)} > \frac{2}{x+1}$$

which shows that the sum of the $\pi(x)$ is either zero or infinite so the embedded chain does not have a stationary distribution. In conclusion, the process (X_t) is positive recurrent whereas its embedded chain is null recurrent. □

In population dynamics, it is common to assume that new arrivals are due exclusively to births from individuals. To include this aspect, we set $\beta_0 = 0$, meaning that state zero is an absorbing state. Assuming that all the other birth and death parameters are positive, we have exactly two communication classes: the absorbing state zero, and the class of all the positive integers which is not closed and therefore transient. In this case, state zero is obviously recurrent and there is no stationary distribution except the point mass at zero. The main problem for such models is to find the probability that the population survives:

$$P_x(X_t \neq 0 \text{ for all } t) = P(X_t \neq 0 \text{ for all } t \,|\, X_0 = x).$$

Repeating the exact same argument as in the proof of Theorem 10.6, we easily show that the process dies out almost surely if $\lim_{N\to\infty} g(N) = \infty$. Otherwise, the probability of survival is given by

$$\begin{aligned} P_x(X_t \neq 0 \text{ for all } t) &= \lim_{N\to\infty} P_x(T_N < T_0) = \lim_{N\to\infty} p_N(x) \\ &= \lim_{N\to\infty} g(x)\, p_N(1) = \lim_{N\to\infty} g(x)/g(N) \end{aligned} \qquad (10.22)$$

where the function g has been defined in (10.20).

Example 10.7 (Simple birth and death process). The simplest example, still biologically relevant, of birth and death process assumes that, independently of each other, individuals give birth at rate β to one offspring and die at rate one. In this simple case, the superposition rule for independent Poisson point processes implies that the transition rates are given by

$$\beta_x = \beta x \;\text{ for }\; x \geq 0 \quad\text{ and }\quad \delta_x = x \;\text{ for }\; x \geq 1.$$

This process can be constructed and simulated numerically as follows: given that the number of individuals at the current time is x,

- the time of the next event is exponential with rate $(\beta + 1)x$ and
- this is a birth with probability $\beta/(\beta + 1)$ and a death otherwise.

For realizations of the simple birth and death process, we refer to Figure 10.3, and for actual Matlab programs that simulate the process using the algorithm above, we refer to the simulation chapter at the end of the textbook. Note that zero is an absorbing state while the positive integers form a transient communication class, and it follows from (10.22) that the survival probability is

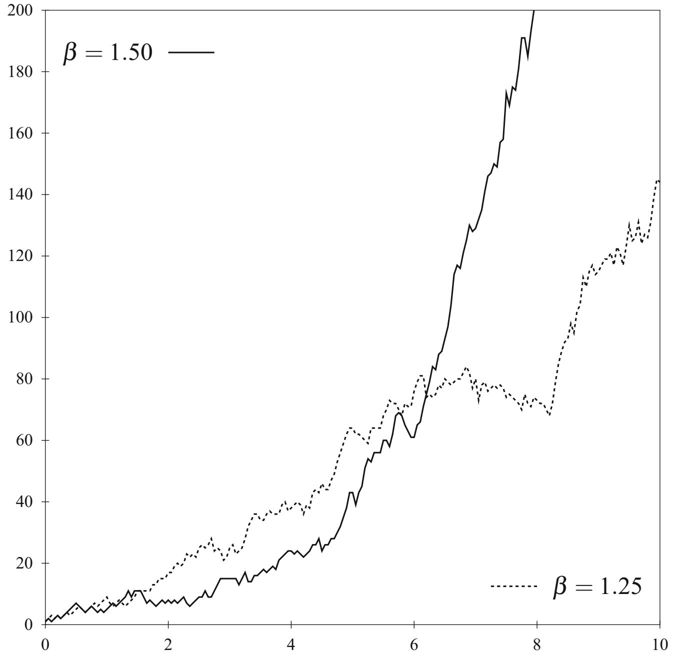


Fig. 10.3 Realizations of the simple birth and death process with $\beta = 1.50$ and $\beta = 1.25$.

$$P_x(X_t \neq 0 \text{ for all } t) = \begin{cases} 0 & \text{when} \quad \beta \leq 1 \\ 1 - \beta^{-x} & \text{when} \quad \beta > 1 \end{cases} \tag{10.23}$$

since $g(x) = \sum_{z<x} \beta^{-z} = (1-\beta^{-x})/(1-\beta^{-1})$ when $\beta > 1$. □

We conclude this chapter with a popular example of birth and death process outside the context of biology. This is a model in queueing theory where birth and death respectively correspond to customers entering and leaving a service facility.

Example 10.8 ($M/M/1$ queue). Assume that customers arrive independently at rate α at a single-server queue where service times are independent exponential random variables with the same parameter μ. A natural question about this model is to determine whether the system is stable or not, i.e., whether the number of customers in the system converges to a certain distribution or blows up to infinity. We are also interested in the expected amount of time a customer spends waiting in the queue when the system is stable.

The process (X_t) that keeps track of the number of customers in the system is a continuous-time Markov chain. The process is clearly irreducible and, according to the convergence theory for continuous-time Markov chains, the system is stable if and only if the process is positive recurrent. Since the state space is infinite, it is easier to directly check the existence of a stationary distribution. One key ingredient to study this queue and other queueing systems is to observe that the number of customers in the system can only increase or decrease by one at each update, making the process a birth and death process. For the $M/M/1$ queue, the birth and death parameters are simply given by

$$\beta_x = \alpha \ \text{ for } \ x \geq 0 \quad \text{and} \quad \delta_x = \mu \ \text{ for } \ x \geq 1.$$

In particular, if there is one, the stationary distribution must satisfy

$$\pi(x) = \frac{\beta_0 \, \beta_1 \cdots \beta_{x-1}}{\delta_1 \, \delta_2 \cdots \delta_x} \, \pi(0) = \left(\frac{\alpha}{\mu}\right)^x \pi(0) \quad \text{for all} \quad x \geq 0.$$

This leads to three cases:

1. when $\alpha < \mu$, the system is positive recurrent: there is a unique stationary distribution and the process converges to this distribution,
2. when $\alpha = \mu$, there is no stationary distribution,
3. when $\alpha > \mu$, the system is transient and $\lim_{t \to \infty} X_t = \infty$ almost surely.

In particular, the system is stable if and only if $\alpha < \mu$, i.e., the rate at which customers enter is less than the rate at which they leave. In this case,

$$1 = \sum_x \pi(x) = \sum_x \left(\frac{\alpha}{\mu}\right)^x \pi(0) = \frac{\pi(0)}{1 - \alpha/\mu} \quad \text{so} \quad \pi(x) = \left(1 - \frac{\alpha}{\mu}\right)\left(\frac{\alpha}{\mu}\right)^x$$

so π is the shifted geometric distribution with parameter $1 - \alpha/\mu$.

To also compute the expected amount of time T a customer spends waiting in the queue in the long run, the key is to condition on the number of customers in the system and use the expression of the stationary distribution. Since the mean waiting time given that there are already x customers in the system is x/μ,

$$E(T) = (1/\mu)E(\pi) = (1/\mu)E(\text{Geometric}\,(1 - \alpha/\mu) - 1)$$

therefore the mean waiting time is $(\alpha/\mu)/(\mu - \alpha)$. □

The exercises below give other examples of queueing systems. For an overview of queueing theory with a lot of examples including queueing networks, we refer the reader to the book of Asmussen [1].

10.6 Exercises

General Markov chains

Exercise 10.1. A frog climbs a ladder with n bars, jumping to the next bar at rate ρ but falling at the bottom of the ladder independently at rate μ.

1. Find the fraction of time the frog spends at the top of the ladder, which can be conveniently expressed using $a = \rho/(\rho+\mu)$.
2. What does this fraction become when μ is very small, very large?

Exercise 10.2. A particle moves clockwise on the vertices on a polygon with m vertices, going from vertex x to the next vertex at rate $\mu_x > 0$. Find the fraction of time the particle spends on each of the vertices.

Exercise 10.3. Consider a Markov chain with state space $S = \{0,1\}$. The evolution is characterized by two parameters, namely

$$\mu_1 = c(0,1) \quad \text{and} \quad \mu_0 = c(1,0)$$

which we assume to be positive to avoid trivialities.

1. Use Kolmogorov's backward equations to prove that

$$p_t'(0,0) = \mu_0 - (\mu_0+\mu_1)\, p_t(0,0).$$

2. Use the function $\phi(t) = p_t(0,0) - \mu_0/(\mu_0+\mu_1)$ to conclude that

$$p_t(0,0) = \frac{\mu_1}{\mu_0+\mu_1}\, e^{-(\mu_0+\mu_1)t} + \frac{\mu_0}{\mu_0+\mu_1}.$$

Exercise 10.4 (Yule process). Consider a population of cells where each cell divides into two daughter cells at rate μ and let (X_t) be the continuous-time Markov chain that keeps track of the number of cells.

1. Consider the independent random variables

$$Y_i \sim \text{Exponential}\,(\mu) \;\text{ and }\; Z_i \sim \text{Exponential}\,(i\mu) \;\text{ for }\; i = 1,2,\ldots,n.$$

Interpreting Y_i as the lifetime of a component, show that

$$\max\,(Y_1,Y_2,\ldots,Y_n) = Z_1 + Z_2 + \cdots Z_n \quad \text{in distribution.}$$

2. Using the previous question, compute $P(X_t \geq n \,|\, X_0 = 1)$.
3. Conclude that, starting with a single cell, $X_t \sim \text{Geometric}\,(e^{-\mu t})$.
4. Prove more generally that, for all $m \leq n$,

$$P(X_t = n \,|\, X_0 = m) = \binom{n-1}{m-1}(1-e^{-\mu t})^{n-m}\, e^{-\mu t m}.$$

Reversible Markov chains

Exercise 10.5. Assume that M particles are distributed among two compartments and that particles independently move from compartment i to compartment j at the same rate μ_j for $i, j = 1, 2$ with $i \neq j$.

1. Compute the fraction of time compartment 2 is empty.
2. Compute also the expected number of particles in compartment 2.

Hint: Use reversibility to prove that the stationary distribution is binomial.

Exercise 10.6. This problem is from [60]. Consider a population of N individuals possessing all together M one-dollar bills. Each individual with at least one dollar gives one dollar at rate one to another individual chosen uniformly at random.

1. Define a Markov chain that describes the behavior of this system.
2. Find the fraction of time a given individual has d dollars.

Hint: For the second question, use reversibility.

Exercise 10.7. Consider a system with n components where components fail independently at the same rate μ. The repair time for each component is exponential with rate ρ divided by the number of components which are down, modeling the fact that the effort of the repair team is equally distributed among all the components that need attention. Use reversibility to find the fraction of time all the components are simultaneously working in the long run.

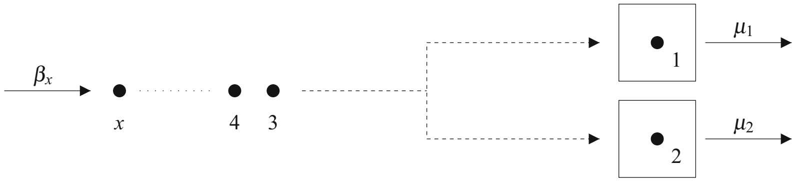
Exercise 10.8. In the context of Exercise 10.7, assume that components are no longer identical: component x fails at rate μ_x and, independently of the rest of the system, the repair time for this component is exponential with rate ρ_x.

1. Find the fraction of time all the components are simultaneously working.
2. Simplify the answer when $\mu_x \equiv \mu$ and $\rho_x \equiv \rho$.

Hint: For the first question, consider a Markov chain with 2^n states.

Queueing systems

The following exercises study various queueing systems. Such systems consist of one or several servers. Typically, customers enter the system at a rate that depends on the number of customers already in the system while they leave at a rate that depends on the number of servers and how fast they complete each service. The following gives an illustration when there are two servers.



In this context, since the arrival and departure times are independent exponential random variables, the process (X_t) that keeps track of the number of customers in the system at time t is a continuous-time Markov chain. The main objective is to check the existence and uniqueness of a stationary distribution, in which case the system reaches an equilibrium and is said to be stable.

Exercise 10.9. Consider a taxi station where taxis and customers arrive respectively at rate one per minute and two per minute. Assume also that taxis always wait for customers but that customers never wait for taxis.

1. Find the fraction of arriving customers that get a taxi.
2. Find the expected number of taxis waiting.

Hint: Use Example 10.8.

Exercise 10.10. Assuming in Exercise 10.9 that there is no taxi waiting initially, find the expected value of the first time at which five taxis will be waiting.

Hint: Compute first the expected time to go from n to $n+1$ taxis by conditioning on whether the next taxi arrives before or after the next customer.

Exercise 10.11. Assume that cars arrive at a stop sign at rate five per minute and that the first car in line will wait an exponential amount of time with mean 5 seconds before leaving. If a pedestrian finds three cars at the stop sign and will only cross the street when there are no more cars, find the expected amount of time the pedestrian will have to wait before crossing the street.

Hint: Use that the expected value of the time to go from n cars to $n-1$ cars waiting at the stop sign does not depend on n.

Exercise 10.12. Customers arrive at rate α at a single-server queue where service times are independent exponential random variables with rate μ and customers not in service yet leave at rate ρ regardless of their position.

1. Study the existence and uniqueness of a stationary distribution.
2. Find the stationary distribution when $\rho = \mu$.

Exercise 10.13. Customers arrive at rate α at a single-server queue but only join the system with probability p_x where x is the number of people already in the system. Assume in addition that the service times are independent exponential random variables with parameter μ.

1. Prove that there is a stationary distribution if $\lim_{x\to\infty} p_x < \mu/\alpha$.

2. Study the special case $p_x = (x+1)^{-1}$.
3. Compute also the stationary distribution when there is a waiting room with limited capacity: customers arrive at rate α but enter the system if and only if there are less than N people in the system.

Exercise 10.14 (**$M/M/s$ queue).** In this queue, customers arrive at rate α at a multi-server queueing system where there are s servers and service times are independent and exponentially distributed with parameter μ.

1. Prove that the queue has a stationary distribution if $\alpha < s\mu$.
2. Find the stationary distribution of the process when $s = \infty$ in which case every arrival experiences immediate service at rate μ.
3. When customers arrive at rate $\alpha = 1$, compare the expected number of services completed per unit of time to check if a single server working at rate two is more or less efficient than three servers working each at rate one.

Exercise 10.15. Customers arrive at a two-server station at rate α with the service times for the two servers being exponentially distributed with two different parameters given by μ_A and μ_B. Upon arrival, customers either wait in line and go to the first server available, or go to each of the two servers with equal probability if there is no other customer in the system.

1. Prove that this system is time-reversible.
2. Find the fraction of time each server is busy when

$$\alpha = \mu_A = 1 \quad \text{and} \quad \mu_B = 2.$$

Exercise 10.16. Consider a system with two servers where the service times at both servers are exponentially distributed with rate μ. Customers arrive at rate α and enter server 1 if it is free and leave otherwise. Similarly, the customers who complete service with server 1 enter server 2 if it is free and leave otherwise.

1. Find the fraction of customers who enter the system.
2. Compute also the fraction of customers in the long run who are lucky enough to visit both server 1 and server 2.

Part III
Special models

Chapter 11
Logistic growth process

Starting from this chapter, we focus on special stochastic processes that are important in physics, biology, and sociology. The models are not chosen at random. They form a coherent structure and tell a story. All the models are minimal in term of their number of parameters and can be classified as either invasion models or as competition models. They mostly differ in their level of realism from the point of view of the inclusion or not of a temporal structure and/or an explicit spatial structure, but are otherwise closely related.

Logistic growth refers to a decrease of population expansion as resources become scarce. The use of the logistic function with graph that is a sigmoid curve to model the growth of such a population was first proposed by Verhulst [93]. This chapter is concerned with the logistic growth process, a stochastic version of the model he introduced. This is the simplest example of a spatially implicit invasion model. Here and after, the term *spatially implicit* means that individuals interact, but the interactions are global in the sense that each individual is equally likely to interact with any other individual in the population. In particular, the actual spatial distance between individuals is ignored.

The first section gives a precise description and an algorithm to simulate the model. The logistic growth process is another example of a continuous-time birth and death process. It is similar to the simple birth and death process in the sense that, independently of one another, individuals give birth to a single offspring at rate β and die at rate one. The logistic growth process, however, also includes competition for space by assuming that new offspring indeed take place in the system only if there is enough space available. In particular, the process can be viewed as a spatially implicit version of the simple birth and death process.

The second section supports the idea that the process can be seen as the stochastic counterpart of the logistic equation [93]. A look at the communication classes of the stochastic process reveals, however, that the long-term behavior of the two models strongly differ: While the logistic equation converges to an interior fixed point corresponding to the so-called carrying capacity, due to random fluctuations, the stochastic process always dies out eventually.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_11

The third section focuses on the transient behavior of the process before it goes extinct. This is achieved by looking at the so-called quasi-stationary distribution. When the individual birth rate exceeds the individual death rate, the number of individuals before extinction fluctuates randomly around the interior fixed point of its deterministic counterpart: the carrying capacity.

To conclude, we show that the mean time to extinction when $\beta > 1$ grows exponentially with the size of the system. In addition, starting from a single individual, the population either dies out quickly or lives for a very long time with respective probabilities corresponding to the probability of extinction and the probability of survival of the simple birth and death process.

11.1 Model description

To include competition for space, we start with a set of N sites that we think of as spatial locations that can be either empty or occupied by one individual. The process is then described by the following simple rules:

- Independently of one another, individuals give birth to a single offspring at rate β and die at the normalized rate one.
- At birth, the offspring is sent to a site chosen uniformly at random. If the site is empty, it becomes occupied; otherwise, the birth is suppressed.

The first item corresponds to the evolution rule of the simple birth and death process while the second item models competition for space. The logistic growth process (X_t) is the continuous-time Markov chain that keeps track of the number of individuals in the population at time t. Due to space limitations, the state space of the process is now finite and reduces to $N+1$ states. According to the superposition property for independent Poisson processes, the rate at which new offspring are produced is β times the number of individuals, as in the simple birth and death process. However, to obtain the birth parameter, i.e., the rate at which the population increases by one, according to the thinning property for independent Poisson processes, this quantity must be multiplied by the fraction of empty sites, which is the probability that the offspring indeed takes place in the system. In particular, we obtain the following birth and death parameters:

$$c(x,x+1) = \beta_x = \beta x \left(1 - \frac{x}{N}\right) \quad \text{and} \quad c(x,x-1) = \delta_x = x \tag{11.1}$$

for all $x = 0, 1, \ldots, N$. Numerical simulations of the process can be obtained by using these transition rates or alternatively going back to the interpretation of the process using an approach common in the field of interacting particle systems. Following the second approach, we keep track of the state at each site rather than the number of individuals. In addition, using the superposition and thinning properties for independent Poisson processes gives the following construction:

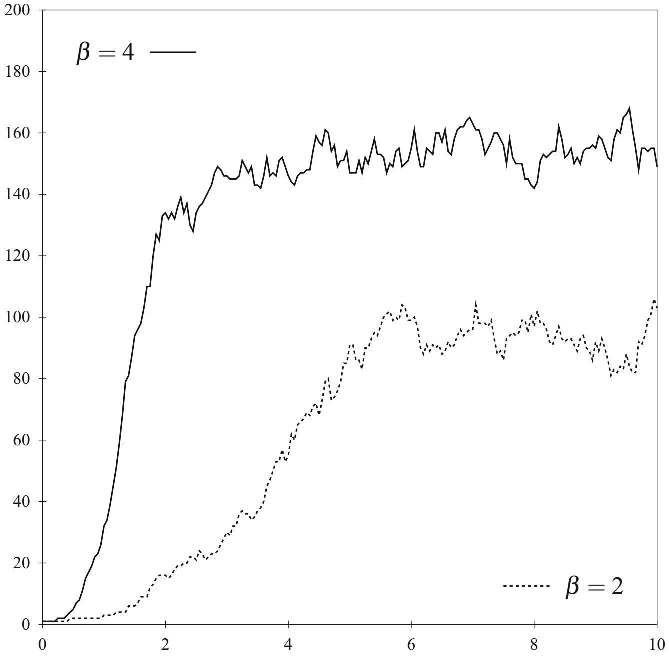


Fig. 11.1 Realizations of the logistic growth process with $N = 200$ sites.

- Regardless of the number of individuals, the time to the next (potential) event is exponentially distributed with parameter $(\beta + 1)N$.
- At that time, we choose a site uniformly at random, say site i.
- If site i is empty, we do nothing. Otherwise,
  - ○ with probability $\beta/(\beta + 1)$, we choose another site uniformly at random and place an individual at this site if it is empty while
  - ○ with probability $1/(\beta + 1)$, we kill the individual at site i.

As for the simple birth and death process, C and Matlab codes for the process are given at the end of this textbook in the simulation chapter. The C program relies on the algorithm above, whereas for computational efficiency, the Matlab program uses the transition rates of the process.

11.2 Long-term behavior

Though it is somewhat hidden behind the equations, the process is closely related to the popular (deterministic) logistic equation [93]. The latter depends on two parameters—the intrinsic rate of growth r and the carrying capacity K—and is described by the following differential equation for the density of individuals:

$$x'(t) = rx\left(1 - \frac{x}{K}\right). \tag{11.2}$$

Starting with a positive density of individuals, the solution to this differential equation converges to zero when the intrinsic rate of growth is negative and to the carrying capacity K when the intrinsic rate of growth is positive. Now, letting

$$r = \beta - 1 \quad \text{and} \quad K = N\left(1 - \frac{1}{\beta}\right) = N\left(\frac{\beta - 1}{\beta}\right), \tag{11.3}$$

the right-hand side of (11.2) becomes

$$\begin{aligned} rx\left(1 - \frac{x}{K}\right) &= (\beta - 1)\left(1 - \frac{\beta x}{N(\beta - 1)}\right)x \\ &= \left(\beta - 1 - \frac{\beta x}{N}\right)x = \beta x\left(1 - \frac{x}{N}\right) - x = \beta_x - \delta_x \end{aligned}$$

showing that (11.2) is the deterministic analog of the logistic growth process when the parameters satisfy the two conditions in (11.3). The first condition is natural. It says that the intrinsic rate of growth is the difference between the intrinsic birth rate and the intrinsic death rate. The second condition suggests that, when individuals compete for space, the fraction of space that is occupied approaches $1 - 1/\beta$, which turns out to be the survival probability of the simple birth and death process starting with one individual. This result is proved rigorously in the next section looking at the so-called quasi-stationary distribution of the process. The fact that the fraction of occupied sites approaches the survival probability of the simple birth and death process is not a coincidence and can be understood from the self-duality property of the contact process that we will see later.

Though the deterministic and stochastic logistic growth models have some similarities, they also have one major disagreement: Because the population is uniformly bounded, random fluctuations will eventually drive the stochastic model to extinction. This can be easily proved by simply looking at the communication classes. There are two communication classes, namely

$$C_0 = \{0\} \quad \text{and} \quad C_+ = \{1, 2, \ldots, N\}$$

with state zero being absorbing and all states in C_+ being transient because this class is not closed. In particular, the process hits state zero after an almost surely finite time. However, when $\beta > 1$, the expected time to extinction increases exponentially with the size of the system, and the behavior before extinction, which is characterized by a so-called quasi-stationary distribution, is somewhat reminiscent of the limiting behavior of the ordinary differential equation (11.2). The quasi-stationary distribution and the time to extinction are studied in the next two sections.

11.3 Quasi-stationary distribution

The quasi-stationary distribution of the logistic growth process is the stationary distribution of the process conditioned on nonextinction. It is not equal but it is well-approximated by the stationary distribution of the process (Y_t) obtained by removing state zero from the state space, so all states become recurrent, i.e., this new process is obtained by setting $\delta_1 = 0$ and leaving all the other parameters in (11.1) unchanged. For a complete picture of the quasi-stationary distribution as well as the mean extinction time discussed in the next section for the logistic growth process, we refer the reader to the book of Nåsell [74].

Theorem 11.1. *The stationary distribution π of the process (Y_t) satisfies*

$$\pi(x) = \frac{\pi(1)}{x} \left(\frac{\beta}{N}\right)^{x-1} \frac{(N-1)\,!}{(N-x)\,!} \quad \textit{for} \quad x = 1, 2, \ldots, N. \tag{11.4}$$

Proof. Note that there is indeed a unique stationary distribution π because the process is finite and irreducible. Following the proof of Theorem 10.7,

$$\pi(x) = \frac{\beta_1 \beta_2 \cdots \beta_{x-1}}{\delta_2 \delta_3 \cdots \delta_x}\, \pi(1) \quad \text{for} \quad x = 2, 3, \ldots, N. \tag{11.5}$$

Plugging (11.1) into (11.5), we obtain

$$\begin{aligned} \pi(x) &= \frac{\pi(1)}{\delta_x} \prod_{y=1}^{x-1} \frac{\beta_y}{\delta_y} = \frac{\pi(1)}{x} \prod_{y=1}^{x-1} \beta \left(1 - \frac{y}{N}\right) \\ &= \frac{\pi(1)}{x} \prod_{y=1}^{x-1} \left(\frac{\beta}{N}\right)(N-y) = \frac{\pi(1)}{x} \left(\frac{\beta}{N}\right)^{x-1} \frac{(N-1)\,!}{(N-x)\,!} \end{aligned}$$

which completes the proof. □

We also point out that

$$\frac{\pi(x+1)}{\pi(x)} = \frac{\beta_x}{\delta_{x+1}} = \frac{x}{x+1}\,\frac{\beta_x}{\delta_x} \approx 1 \quad \text{when} \quad \beta_x \approx \delta_x$$

indicating that the most visited states are near the carrying capacity K. In particular, before extinction, the process indeed oscillates randomly around the carrying capacity, as suggested by the simulation curves in Figure 11.1. We now prove that the expected amount of time this transient behavior lasts increases exponentially with the size of the system, i.e., the number of spatial locations.

11.4 Time to extinction

In this section, we return to the logistic growth process, which has zero as an absorbing state. The time to extinction is defined as

$$T = \inf\{t > 0 : X_t = 0\}$$

which is a random variable with a distribution that depends on the initial state. To compute the expected time to extinction, we also introduce

$$T_x = E_x(T) = E(T \mid X_0 = x) \quad \text{and} \quad \tau_x = T_{x+1} - T_x.$$

The following theorem gives an expression for τ_x which is not restricted to the logistic growth process and holds for birth and death processes in general.

Theorem 11.2. *For all* $x = 0, 1, 2, \ldots, N-1$,

$$\tau_x = \prod_{z=1}^{x} \frac{\delta_z}{\beta_z} \left[\tau_0 - \sum_{y=1}^{x} \frac{1}{\beta_y} \prod_{z=1}^{y} \frac{\beta_z}{\delta_z}\right].$$

Proof. Following the same approach as in the proof of Theorem 10.6, we use a first-step analysis, i.e., we condition on whether the first update is a birth or a death, and the basic property of the exponential random variable to get

$$(\beta_x + \delta_x)\, T_x = \beta_x\, T_{x+1} + \delta_x\, T_{x-1} + 1 \quad \text{for all} \quad x = 1, 2, \ldots, N-1.$$

Then, rearranging the terms, we deduce that

$$\beta_x\, (T_{x+1} - T_x) = \delta_x\, (T_x - T_{x-1}) - 1.$$

Finally, recalling that $\tau_x = T_{x+1} - T_x$, we conclude that

$$\begin{aligned} \tau_x &= \frac{\delta_x}{\beta_x}\, \tau_{x-1} - \frac{1}{\beta_x} = \frac{\delta_x}{\beta_x} \frac{\delta_{x-1}}{\beta_{x-1}}\, \tau_{x-2} - \frac{\delta_x}{\beta_x} \frac{1}{\beta_{x-1}} - \frac{1}{\beta_x} \\ &= \cdots = \prod_{z=1}^{x} \frac{\delta_z}{\beta_z}\, \tau_0 - \sum_{y=1}^{x} \frac{1}{\beta_y} \prod_{z=y+1}^{x} \frac{\delta_z}{\beta_z} = \prod_{z=1}^{x} \frac{\delta_z}{\beta_z} \left[\tau_0 - \sum_{y=1}^{x} \frac{1}{\beta_y} \prod_{z=1}^{y} \frac{\beta_z}{\delta_z}\right]. \end{aligned}$$

This completes the proof. □

Since $\tau_x \geq 0$ and $\tau_0 = T_1$, the theorem implies that

$$\begin{aligned} T_1 &\geq \sum_{y=1}^{x} \frac{1}{\beta_y} \prod_{z=1}^{y} \frac{\beta_z}{\delta_z} \geq \frac{1}{\beta_{K/2}} \prod_{z=1}^{K/2} \frac{\beta_z}{\delta_z} \geq \frac{1}{\beta_{K/2}} \left[\beta \left(1 - \frac{K}{2N}\right)\right]^{K/2} \\ &\geq \frac{1}{\beta_{K/2}} \left[\beta \left(1 - \frac{1 - \beta^{-1}}{2}\right)\right]^{K/2} = \frac{1}{\beta_{K/2}} \left(\frac{\beta + 1}{2}\right)^{K/2} \end{aligned}$$

so the expected time to extinction starting with one individual grows exponentially with the size of the system when $\beta > 1$, or equivalently $r > 0$.

To conclude, we note that, when the size of the system is large, the logistic growth process behaves almost like the simple birth and death process as long as the number of individuals is not too large because in this case births are successful with probability close to one. This and the previous results give the following description of the process starting with a single individual when $\beta > 1$.

- With probability $1/\beta$ = probability of extinction of the simple birth and death process starting with one individual, the population goes extinct quickly.
- With probability $p_s = 1 - 1/\beta$ = probability of survival of the simple birth and death process starting with one individual, the population first grows until the fraction of occupied sites approaches p_s, then fluctuates randomly around this value, and finally collapses after a time with an expected value that is large.

11.5 Exercises

Exercise 11.1 (Sexual reproduction). In this process, there are again N sites that are either empty or occupied, but the offspring now have two parents. To model this aspect, we assume that each pair of (possibly identical) occupied sites gives birth independently at rate β/N to an offspring that is then sent to a site chosen uniformly at random. The birth rate is divided by N so that births and deaths occur at the same time scale. Again, the birth is suppressed if the target site is already occupied and individuals independently die at rate one.

1. Write the transition rates of the continuous-time Markov chain that keeps track of the number of individuals.

This stochastic process is closely related to the following ordinary differential equation for the density u of occupied sites:

$$u' = \beta u^2 (1-u) - u.$$

2. Find the fixed points and determine their stability.
3. Modify Program 5 to simulate the stochastic process and estimate the probability that the process survives for a reasonably long time for various values of the birth rate and the initial state.

The following numbers give the probability of survival $p(\beta, x)$ for the process with parameter β starting with x individuals obtained from the average of 1000 realizations. The cutoff time and population size are $T = 200$ and $N = 1000$, respectivegly.

```
p(5,266)=0.321     p(6,201)=0.257     p(7,163)=0.269
p(5,276)=0.468     p(6,211)=0.465     p(7,173)=0.491
p(5,286)=0.641     p(6,221)=0.684     p(7,183)=0.714
```

4. Comment on the main differences between these simulation results for the stochastic process and the differential equation.

Exercise 11.2 (Multi-type process). Each of the N sites is now empty or occupied by a type 1 or a type 2 individual. Individuals give birth at rate β to an offspring of their own type and die at rate one like in the logistic growth process.

1. Specify the state space and the transition rates of the continuous-time Markov chain that keeps track of the number of type 1 and type 2 individuals.

This model is closely related to the following system of ordinary differential equations for the densities u_1 and u_2 of type 1 and type 2 individuals:

$$u_1' = \beta u_1 (1 - u_1 - u_2) - u_1 \quad \text{and} \quad u_2' = \beta u_2 (1 - u_1 - u_2) - u_2.$$

2. Find all the fixed points of this system.
3. Modify Program 5 to simulate the stochastic process starting from the configuration in which half of the sites are occupied by a type 1 individual and half of the sites are occupied by a type 2 individual.
4. Compare the behavior shown by the simulation with the limiting behavior of the deterministic counterpart when $u_1 = u_2$ at time zero.

Exercise 11.3 (Epidemic model). Assume that each of the N sites is occupied by an individual who is either healthy or infected. Each infected individual makes contact at rate β with a randomly chosen individual, and if the latter is healthy, then she becomes infected and remains so forever.

1. Write the transition rates of the continuous-time Markov chain that keeps track of the number of infected individuals.
2. Compute the expected value of the time T_N until all the individuals are infected for the system starting with a single infected individual.
3. Prove that $E(T_N) \sim (2/\beta)\ln(N)$ when N is large.

Exercise 11.4 (Epidemics with recovery). Assume that each of the N sites is occupied by an individual who is either healthy, infected, or immune. Individuals get infected as in the model of Exercise 11.3, but now infected individuals become immune at rate one, and remain immune forever.

1. Specify the state space and the transition rates of the continuous-time Markov chain that keeps track of the number of infected individuals and the number of immune individuals.
2. Enumerate all the possible absorbing states and limiting behaviors.
3. Explain heuristically how the infection rate β should affect the number of individuals that have ever been infected.
4. Modify Program 5 to simulate the process and check your guess.

Exercise 11.5 (Host-symbiont system). Each of the N sites is now either empty or occupied by a host, and each host is either healthy or infected. Both healthy and infected hosts give birth at rate β and die at rate one as in the logistic growth process, and we also assume vertical transmission: Offspring of infected hosts are infected at birth. In addition, infected hosts recover at rate one and make contact at rate α with a site chosen uniformly at random, which results in a new infection in case this site is occupied by a healthy host.

1. Specify the state space and the transition rates of the continuous-time Markov chain that keeps track of the number of healthy and infected hosts.

This model is closely related to the following system of ordinary differential equations for the densities u_1 and u_2 of healthy and infected hosts:

$$\begin{aligned} u_1' &= \beta u_1 (1 - u_1 - u_2) - u_1 - \alpha u_1 u_2 + u_2 \\ u_2' &= \beta u_2 (1 - u_1 - u_2) - u_2 + \alpha u_1 u_2 - u_2. \end{aligned}$$

2. Find the fixed point of this system.
3. Modify Program 5 to simulate the stochastic process and use this simulation to check if the process starting from the all-infected configuration behaves for a reasonably long time as indicated by its deterministic counterpart.

Exercise 11.6 (Predator-prey system). Each of the N sites is either empty or occupied by a prey or a predator. Preys and predators die at rate one and give birth at respective rates β_1 and β_2 to an offspring which is then sent to a site chosen uniformly at random. Births of preys are suppressed if the target site is not empty while births of predators are suppressed if the target site is not occupied by a prey, which models the fact that the predator needs to feed on the prey to survive.

1. Specify the state space and the transition rates of the continuous-time Markov chain that keeps track of the number of preys and predators.
2. Prove that there can be only three regimes: extinction of preys and predators, survival of the preys for a long time, and coexistence for a long time.
3. Modify Program 5 to simulate the process and check that there is a parameter region in which there is coexistence for a long time.

Chapter 12
Wright–Fisher and Moran models

The first references introducing the Wright–Fisher model and the Moran model discussed in this chapter are [97] and [73], respectively. Contrary to the logistic growth process, these stochastic processes are competition models. In particular, the relevant mathematical tools to study these models are quite different. However, all three models are closely related from the point of view of the level of complexity they take into account, in that all three stochastic models are spatially implicit dynamical systems. In particular, as in the logistic growth process, the models can be described starting from a set of sites that we think of as spatial locations and the dynamics is dictated by global interactions. Rather than being either empty or occupied, these sites are now occupied by one of two possible types of individuals that we call type 0 and type 1.

The first section gives a description and algorithms to simulate the models. Both models were originally introduced in the context of population genetics to understand the evolution of the gene frequency in a constant size population of haploid individuals, but can be seen as the simplest spatially implicit competition models in the same way the logistic growth process can be seen as the simplest spatially implicit invasion model. The two models only differ in their time scale: the Wright–Fisher model counts time in number of generations, whereas the Moran model is much slower and counts time in number of births.

For both models, all the individuals have eventually the same type so the first natural question in this context is: Who wins? For both models, the answer is the same as for the gambler's ruin chain, namely, the probability that a given type wins is equal to the initial fraction of individuals of this type. The second section shows this result for the Wright–Fisher model using the optional stopping theorem. We also give a more intuitive and more constructive alternative proof of this result looking at the genealogy of the process. The proof for the Moran model is in fact easier because this model is only a time-change of the gambler's ruin chain.

The second natural question about the Wright–Fisher model is: How long does it take for the population to fixate? This problem is more difficult but, using the diffusion approximation of the process, we will be able, in the third section, to

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_12

compute the expected value of the time to fixation rescaled by the population size in the limit as the number of individuals goes to infinity. This result gives in particular good approximations in the context of large populations.

The fourth section is concerned with an aspect that is less natural mathematically but important in population genetics: Given a small sample of individuals at the present time, how far do we need to go back in time to find their most recent common ancestor? We can again answer this question in the large population limit, which gives rise to a new process called Kingman's coalescent. This process has some interesting similarities with the dual process of the voter model that will be studied at the end of this textbook.

Finally, the last section looks at a natural variant of the Moran model that includes selection, meaning that individuals no longer have the same fitness or reproduction success. For this process, we prove that the probability that a given type wins is again the same as the probability of winning or losing in the gambler's ruin chain but in the context of unfair games.

Further reading

- For a review of the Wright–Fisher model, variants of this model and other results in coalescent theory, we refer the reader to Neuhauser [76].
- See also Ewens [33] and Nowak [80] for additional mathematical models in the field of population genetics.
- The early pioneer works on diffusion processes that will be briefly mentioned in this chapter are Itô and McKean [47] and Lévy [68].
- Ethier and Kurtz [31] and Ewens [33] also give applications of diffusion processes in the context of population genetics. For an elementary introduction to diffusion processes at the undergraduate level, see [50].

12.1 The binomial distribution

As previously mentioned, we start with a set of N sites that are interpreted as spatial locations. Each site is occupied by either a type 0 or a type 1 individual. The Wright–Fisher model describes populations with nonoverlapping generations that evolve in discrete time according to the following simple rules.

- The parent of each individual at each generation is chosen independently and uniformly at random from the population at the previous generation.
- Each individual is of the same type as its parent.

The process itself is the discrete-time Markov chain (X_n) that keeps track of the number of type 1 individuals at generation n. In view of the evolution rules, each individual at each generation is independently of type 1 with probability the fraction of type 1 individuals at the previous generation, which gives the following transition probabilities that are binomial:

$$p(x,y) = \binom{N}{y} p^y (1-p)^{N-y} \quad \text{where} \quad p = x/N. \tag{12.1}$$

To exhibit later important connections with the voter model, it is useful to also introduce the Moran model that describes similar dynamics but at a slower time scale: there is only one birth at each time step so that generations now overlap. More precisely, the process again keeps track of the number of type 1 individuals at time n but the population now evolves according to the following rules.

- At each time step, one individual chosen uniformly at random is replaced by the offspring of an individual chosen uniformly at random from the population at the previous time step.
- Each individual is again of the same type as its parent.

In particular, the number of type 1 individuals increases by one with probability the fraction of sites occupied by a type 1 individuals times the fraction of sites occupied by a type 0 individual, and similarly for the number of type 0 individuals, which gives the transition probabilities

$$p(x, x \pm 1) = \left(\frac{x}{N}\right)\left(1 - \frac{x}{N}\right) \quad \text{and} \quad p(x,x) = \left(\frac{x}{N}\right)^2 + \left(1 - \frac{x}{N}\right)^2 \tag{12.2}$$

for all $x = 0, 1, \ldots, N$. This implies that the Moran model is a martingale. In fact, using the same techniques as for the gambler's ruin chain for fair games, we can prove that, starting with x type 1 individuals, this type eventually wins with probability the initial fraction x/N of type 1 individuals. As we will see in a moment, the same result holds for the Wright–Fisher model.

As for the logistic growth process, both models can be simulated by either using the expression of their transition probabilities (12.1)–(12.2) or going back to their interpretation as spatially implicit models and keeping track of the sites which are occupied by a type 1 individual. Using the latter interpretation gives the following algorithm for the Wright–Fisher model.

- For each site, say i, choose a site uniformly at random, say j.
- Fix the type of the individual at site i at generation $n+1$ to be the same as the type of the individual at site j at generation n.

For realizations of the Wright–Fisher model, we refer to Figure 12.1. The algorithm for the Moran model is similar except that, instead of updating all the sites at once, we only choose one pair of sites (i, j) uniformly at random at each time step. The simulation chapter at the end of this textbook gives actual simulation programs for both models.

12.2 Probability of fixation

To study the limiting behavior of the processes, we first observe that there are two absorbing states: state 0 and state N, which correspond, respectively, to a population with only type 0 individuals and to a population with only type 1 individuals.

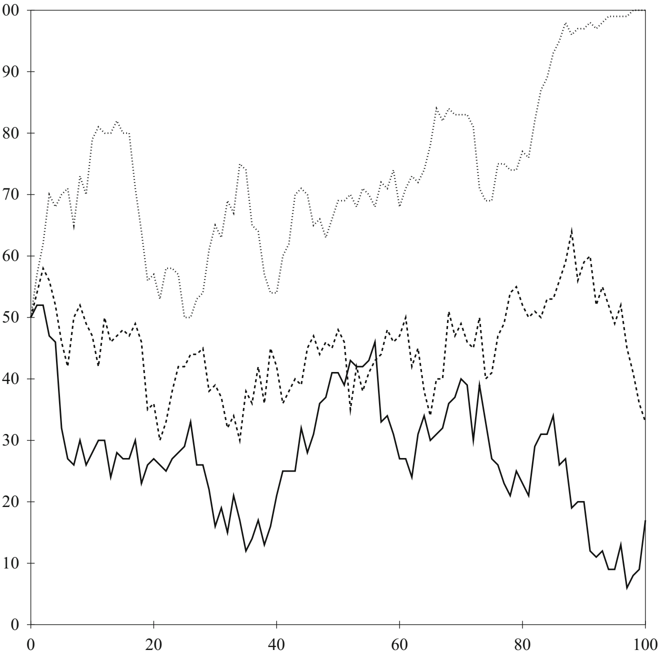


Fig. 12.1 Realizations of the Wright–Fisher model with population size $N = 100$.

These two states form two closed communication classes. All the other states constitute a third communication class which is not closed and is therefore transient. In particular, the system converges almost surely to one of its two absorbing states and the first natural question about the processes is: Who wins?

Note that the partition into communication classes is the same as for the gambler's ruin chain, and recall that, for the gambler's ruin chain, the probability that the gambler quits a winner can be computed using either a first-step analysis or the optional stopping theorem. For the Wright–Fisher model, the use of a first-step analysis looks complicated because the process can jump from any transient state to any state, which leads to tedious calculations. The optional stopping theorem, however, is equally easy to use in this context.

Theorem 12.1. *Starting with $X_0 = x$ individuals of type 1,*

$$P_x(\textit{type 1 wins}) = P_x(X_n = N \textit{ for some } n) = x/N.$$

Proof. The transition probabilities are given by

$$p(x,y) = P(\text{Binomial}\,(N,x/N) = y) \quad \text{for all} \quad x,y = 0,1,\ldots,N,$$

from which it follows that, for all x,

$$E(X_{n+1}\,|\,X_n = x) = \textstyle\sum_y y\,p(x,y) = E(\text{Binomial}\,(N,x/N)) = x.$$

This shows that (X_n) is a martingale. Now, let p_0 and p_N be, respectively, the probability of fixation to all type 0 and all type 1, and consider the time to fixation

$$T = \inf\{t : X_t \in \{0,N\}\}.$$

At each time step, the probability that the process jumps to one of its two absorbing states is at least $p(\lfloor N/2 \rfloor, 0)$, which is strictly positive. In particular,

$$P(T = \infty) \leq P(\text{Geometric}\,(p(\lfloor N/2 \rfloor, 0)) = \infty) = 0$$

so time T is almost surely finite. Since in addition (X_n) is bounded, the optional stopping theorem applies, and we get

$$x = E_x(X_0) = E_x(X_T) = N \times P(X_T = N) + 0 \times P(X_T = 0) = N\,p_N$$

showing that, starting with x type 1 individuals, $p_N = x/N$. □

Ewens [32] gave the following intuition behind this result: After a long enough time, all the individuals must have descended from just one of the individuals present at generation zero and the probability that this common ancestor was of type 1 is equal to the relative frequency x/N of type 1 at generation zero. The following is an alternative proof of the previous theorem using this argument.

Proof. The idea is to represent the process graphically, which is a common technique in the field of interacting particle systems that will be explained later. The following construction is inspired from the voter model. Let

$$U(i,n) \sim \text{Uniform}\,\{1,2,\ldots,N\} \quad \text{for all} \quad (i,n) \in \{1,2,\ldots,N\} \times \mathbb{N}^*$$

be independent and consider the process

$$Y_n : \{1,2,\ldots,N\} \to \{0,1\} \quad \text{with} \quad Y_n(i) = Y_{n-1}(U(i,n)).$$

In words, the process (Y_n) keeps track of the type of the individual at each site rather than the number of type 1 individuals. In particular, the Wright–Fisher model is simply obtained by setting

$$X_n = \sum_{i=1}^{N} Y_n(i) \quad \text{for all} \quad n \in \mathbb{N}.$$

Now, draw an edge between the points (i,n) and $(U(i,n), n-1)$, i.e., each edge connects an individual to its parent, and write

$$(j,0) \to (i,n) \quad \text{if there is a path of edges from } (j,0) \text{ to } (i,n). \tag{12.3}$$

Note that, going backward in time, paths coalesce when they intersect. In addition, there exists a random time τ almost surely finite such that all the paths had time to

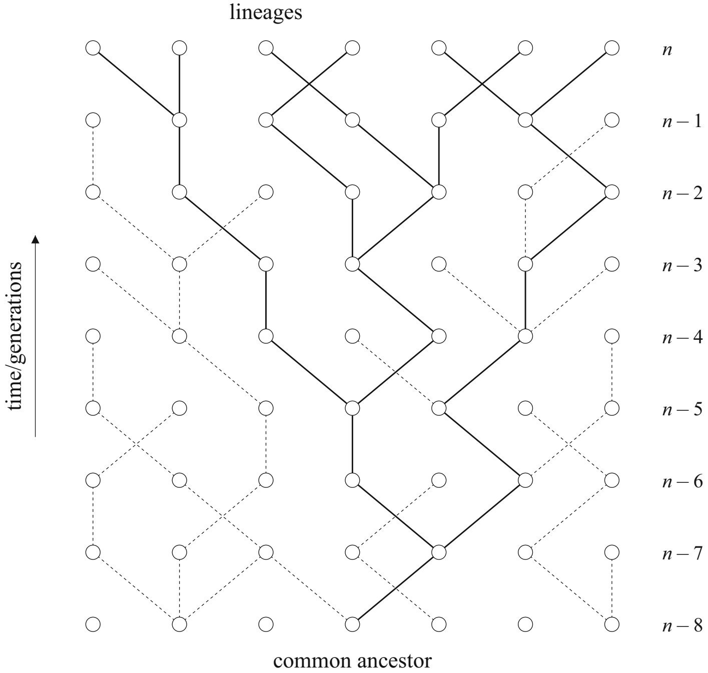


Fig. 12.2 Schematic illustration of Ewens' argument.

coalesce going τ units of time backward, as illustrated in Figure 12.2. This means that all the individuals at this random time and after this random time have a common ancestor, that is, there exists j such that

$$(j,0) \to (i,\tau) \quad \text{for all} \quad i = 1,2,\dots,N.$$

This, together with the construction of (Y_n) and the fact that j is chosen uniformly at random, implies that, starting with x individuals of type 1,

$$\begin{aligned} p_N &= P_x(X_n = N \text{ for some } n) = P_x(X_T = N) \\ &= P(Y_T(i) = 1 \text{ for all } i = 1,2,\dots,N \,|\, X_0 = x) \\ &= P(Y_0(j) = 1 \,|\, X_0 = x) = x/N. \end{aligned}$$

This completes the proof. □

This alternative proof looks more complicated than the original one but has two advantages. First of all, it does not rely on any sophisticated tool such as the optional stopping theorem, but instead relies on a more constructive argument. In addition, the construction given in the proof is more robust in the sense that similar constructions can be used to study Kingman's coalescent and the voter model.

12.3 Diffusion approximation and time to absorption

After studying the probability that type 1 outcompetes type 0, we now turn our attention to the second most natural question about the model: how long does it take for the population to fixate? Good approximations of the expected value of the time to absorption for large populations can be found by looking at the so-called diffusion approximation of the model.

Diffusion processes are continuous-time Markov processes with continuous sample paths, and so an uncountable state space. The diffusion approximation of a discrete-time Markov chain is obtained by rescaling the state space by its size and speeding up time by the same factor. Recalling that N denotes the population size, the diffusion approximation of the Wright–Fisher model is thus defined as

$$Y_t = \lim_{N \to \infty} N^{-1} X_{\lfloor Nt \rfloor} \quad \text{where} \quad \lfloor Nt \rfloor = \text{integer part of } Nt. \tag{12.4}$$

The state space of this process is the unit interval $[0,1]$, which represents the frequency, rather than the number, of type 1 individuals. It can be proved that diffusion processes are fully characterized by the two quantities

$$\begin{array}{rrl} \text{drift parameter} & \mu(x) & = \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} E(Y_{t+\varepsilon} - Y_t \,|\, Y_t = x) \\ \text{diffusion parameter} & \sigma^2(x) & = \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} E((Y_{t+\varepsilon} - Y_t)^2 \,|\, Y_t = x) \end{array} \tag{12.5}$$

while, for all $n \geq 3$, we have

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-1} E((Y_{t+\varepsilon} - Y_t)^n \,|\, Y_t = x) = 0. \tag{12.6}$$

We refer to [50, chapter 15] for more details on this aspect. To study the time it takes for the population to fixate, we let

$$T = \inf\{t : Y_t \in \{0,1\}\} \quad \text{and} \quad u(x) = E(T \,|\, Y_0 = x)$$

be the time to absorption and its expected value starting from frequency x. Then, the function u thus defined satisfies the following differential equation.

Theorem 12.2. *For all* $x \in [0,1]$*, we have*

$$\mu(x)\, u'(x) + (1/2)\, \sigma^2(x)\, u''(x) = -1.$$

Proof. First, we observe that, by the Markov property,

$$E(T \,|\, Y_\varepsilon = z) = E(T \,|\, Y_0 = z) + \varepsilon = u(z) + \varepsilon$$

for all $\varepsilon > 0$ small. In particular,

$$\begin{aligned} u(x) &= E(T \,|\, Y_0 = x) = E(T - \varepsilon \,|\, Y_0 = x) + \varepsilon \\ &= E(E(T - \varepsilon \,|\, Y_\varepsilon) \,|\, Y_0 = x) + \varepsilon = E(u(Y_\varepsilon) \,|\, Y_0 = x) + \varepsilon. \end{aligned} \tag{12.7}$$

Using a Taylor expansion and recalling (12.5)–(12.6), we also have

$$\begin{aligned} E(u(Y_\varepsilon)\,|\,Y_0 = x) &= E(u(x + (Y_\varepsilon - x))\,|\,Y_0 = x) \\ &= E(u(x) + (Y_\varepsilon - x)\,u'(x) \\ &\qquad + (1/2)(Y_\varepsilon - x)^2\,u''(x) + O(Y_\varepsilon - x)^3\,|\,Y_0 = x) \\ &= u(x) + \varepsilon\,\mu(x)\,u'(x) + \varepsilon\,(1/2)\,\sigma^2(x)\,u''(x) + o(\varepsilon). \end{aligned} \tag{12.8}$$

Combining (12.7)–(12.8), we deduce that

$$\varepsilon\,\mu(x)\,u'(x) + \varepsilon\,(1/2)\,\sigma^2(x)\,u''(x) + \varepsilon + o(\varepsilon) = 0.$$

The result follows by dividing by ε and taking the limit as $\varepsilon \downarrow 0$. □

In the special case of the Wright–Fisher model, the drift parameter and diffusion parameter in (12.5) are given, respectively, by

- $\mu(x) = 0$ because (X_n) is a martingale and
- $\sigma^2(x) = x(1-x)$ because

$$\begin{aligned} \sigma^2(x) &= \lim_{N\to\infty} N E((Y_{t+1/N} - Y_t)^2\,|\,Y_t = x) \\ &= \lim_{N\to\infty} N E\ ((N^{-1}X_{\lfloor Nt\rfloor+1} - N^{-1}X_{\lfloor Nt\rfloor})^2\,|\,X_{\lfloor Nt\rfloor} = Nx) \\ &= \lim_{N\to\infty} N^{-1} E((X_{\lfloor Nt\rfloor+1} - Nx)^2\,|\,X_{\lfloor Nt\rfloor} = Nx) \\ &= \lim_{N\to\infty} N^{-1}\ \mathrm{Var}\,(\mathrm{Binomial}\,(N,x)) = \lim_{N\to\infty} N^{-1}Nx\,(1-x). \end{aligned}$$

In particular, according to Theorem 12.2,

$$x\,(1-x)\,u''(x) = -2 \quad \text{with boundary conditions} \quad u(0) = u(1) = 1$$

which has the unique solution

$$u(x) = -2\,(x\ln(x) + (1-x)\ln(1-x)).$$

For example, taking $x = 1/2$, we have

$$u(1/2) = -2\ln(1/2) = 2\ln(2) \approx 1.386,$$

meaning that, when the population size is large and half of the individuals are initially of type 1, the expected value of the time to fixation counted in number of generations is approximately equal to 1.386 times the population size.

12.4 Kingman's coalescent

In this section, we answer a third question about the model: Given a sample of n individuals at the present time, how far do we need to go backward in time to find their most recent common ancestor? This question might not be natural from a

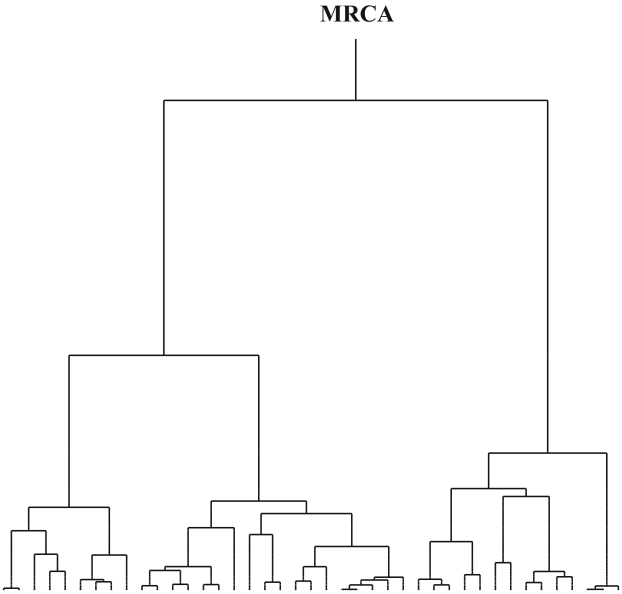


Fig. 12.3 Realization of Kingman's coalescent.

mathematical point of view but it is important for population geneticists. The answer to this question is somewhat related to the construction given above in the second proof of Theorem 12.1. The basic idea is to let the process evolve long enough so that all the individuals have a common ancestor, take a random sample of n individuals, and then find the amount of time we need to go backward until the n corresponding paths of edges, as defined in the proof, coalesce.

As for the expected time to absorption, we answer this question for the diffusion approximation of the Wright–Fisher model. The process that keeps track of the times of the coalescing events as well as the topological structure of the family tree of the sample of n individuals was introduced in [53, 54] and is known as Kingman's n-coalescent. It can be represented as a random tree with n leaves in which branches coalesce by pairs at some random times. For an example of realization of the process, we refer to Figure 12.3 where time goes down, from top to bottom.

We start by giving the formal definition of the n-coalescent and then argue that this process indeed keeps track of the genealogy of the sample of n individuals in the diffusion limit of the Wright–Fisher model. We also give a couple of results about the n-coalescent, including the explicit expression of the expected value and variance of the time to the most recent common ancestor. By definition, the n-coalescent is the continuous-time Markov chain $(\mathscr{K}_t)$ with state space

$$S_n = \text{set of all the possible partitions of } \{1, 2, \ldots, n\}$$

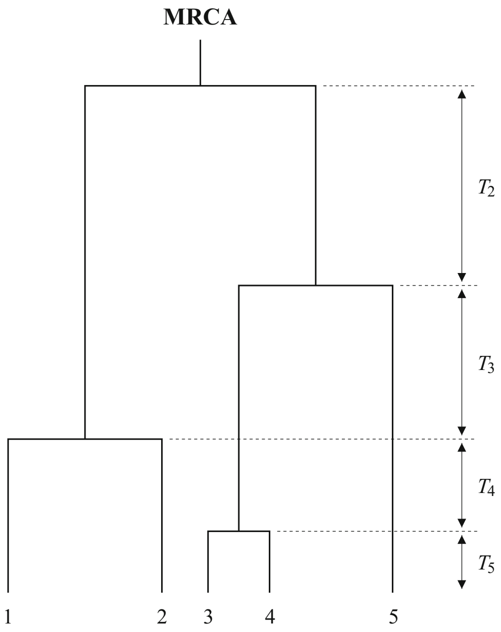


Fig. 12.4 Picture of the sequence T_j.

and transition rates

$$c(\xi,\eta) = \lim_{\varepsilon\downarrow 0} \varepsilon^{-1} P(\mathscr{K}_{t+\varepsilon} = \eta \,|\, \mathscr{K}_t = \xi) = \mathbf{1}\{\xi \prec \eta\} \tag{12.9}$$

where $\xi \prec \eta$ means that η is one of the partitions obtained from ξ by merging exactly two of the elements of the partition ξ. In addition,

$$\mathscr{K}_0 = \{\{1\},\{2\},\dots,\{n\}\},$$

the unique partition that consists of n singletons. To see that this process indeed described the genealogy of the sample of n individuals, assume that the Wright–Fisher model is defined for negative times, running from time $-\infty$, and think of time zero as the present time. Then, in the diffusion limit, we have the following connection between the two processes.

Lemma 12.1. *For all $i,j \in \{1,2,\dots,n\}$,*

$$\begin{aligned} &P(i,j \textit{ belong to the same element of the partition } \mathscr{K}_t) \\ &\quad = P(\textit{individuals } i,j \textit{ have a common ancestor at time } -t). \end{aligned}$$

Proof. As previously mentioned, it is convenient to think of the genealogy of the sample as a tree with n leaves corresponding to the n individuals and where branches coalesce whenever the corresponding individuals have the same parent, as shown in Figures 12.3 and 12.4. To prove the lemma, it suffices to show that

1. any two branches coalesce at rate one and
2. coalescing events involving at least three branches occur at rate zero,

in accordance with the transition rates (12.9) where the transition $\xi \to \eta$ means that exactly two branches coalesce. To show the first item, observe that, for a given pair of individuals, independently at each generation, these individuals have the same parent with probability $1/N$. In particular, the time T one needs to go backward in time to find their most recent common ancestor satisfies

$$\begin{aligned} P(T > t) &= \lim_{N \to \infty} P(\text{no common ancestor for } \lfloor Nt \rfloor \text{ generations}) \\ &= \lim_{N \to \infty} P(\text{Geometric}\,(1/N) > \lfloor Nt \rfloor) \\ &= \lim_{N \to \infty} (1 - 1/N)^{\lfloor Nt \rfloor} = \exp(-t) \end{aligned}$$

in the diffusion limit. Hence, pairs of branches indeed coalesce at rate one. In addition, the probability that at least three given individuals have the same parent is at most $1/N^2$ so coalescing events involving at least three branches occur at rate zero in the diffusion limit. This proves the second item. □

After showing the connection between the n-coalescent and the genealogy of the sample of size n in the Wright–Fisher model, we now study more closely the coalescent process. At each jump of the process, the number of elements in the partition decreases by one so the process must visit exactly n partitions:

$$\{\{1\},\{2\},\ldots,\{n\}\} = \xi_n \prec \xi_{n-1} \prec \cdots \xi_2 \prec \xi_1 = \{1,2,\ldots,n\} \tag{12.10}$$

with the number of elements in ξ_k being equal to k. For instance, in the realization shown in Figure 12.4, the intermediate partitions are given by

$$\begin{aligned} \xi_4 &= \{\{1\},\{2\},\{3,4\},\{5\}\} \\ \xi_3 &= \{\{1,2\},\{3,4\},\{5\}\} \\ \xi_2 &= \{\{1,2\},\{3,4,5\}\}. \end{aligned}$$

The coalescent process is thus characterized by two components: the sequence of partitions the process visits and the sequence of times spent in each partition. Using the same notation as in the figure, we let

$$T_k = \text{time spent in state/partition } \xi_k \quad \text{for} \quad k = 2,3,\ldots,n, \tag{12.11}$$

and observe that the time spent in ξ_1 is infinite since the last partition is minimal and therefore is an absorbing state. A nice property of the n-coalescent is that these two components are independent, as shown in the following result.

Theorem 12.3. *The two sequences* (12.10) *and* (12.11) *are independent. In addition, the distribution of the holding times is given by*

$$T_k \sim \text{Exponential}\left(\frac{k(k-1)}{2}\right) \quad \textit{for} \quad k = 2,3,\ldots,n.$$

Proof. For each $\xi \in S_n$, the number of partitions η such that $\xi \prec \eta$ is clearly equal to the number of pairs of elements in ξ. Because in addition ξ_k has always k elements and since $\xi_k \to \eta$ at rate one, the superposition property gives

$$T_k \sim \text{Exponential}\left(\binom{k}{2}\right) = \text{Exponential}\left(\frac{k(k-1)}{2}\right).$$

Because this distribution does not depend on the specific form of the partition ξ_k, we also have that the two sequences (12.10) and (12.11) are independent. □

Using the previous theorem, we can compute the expected value and the variance of the time to the most recent common ancestor $T_{\text{MRCA}}(n)$, which is the time at which the n-coalescent reaches the partition ξ_1.

Theorem 12.4. *For each fixed n,*

$$\begin{aligned} E(T_{\text{MRCA}}(n)) &= 2\left(1 - \frac{1}{n}\right) \\ \text{Var}(T_{\text{MRCA}}(n)) &= 8 \sum_{k=1}^{n-1} \left(\frac{1}{k}\right)^2 - 4\left(1 - \frac{1}{n}\right)\left(3 + \frac{1}{n}\right). \end{aligned}$$

Proof. By definition, we have

$$T_{\text{MRCA}}(n) = T_n + T_{n-1} + \cdots + T_2$$

which, together with Theorem 12.3, implies that

$$\begin{aligned} E(T_{\text{MRCA}}(n)) &= \sum_{k=2}^{n} E(T_k) = \sum_{k=2}^{n} \left(\frac{2}{k(k-1)}\right) \\ &= 2 \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right). \end{aligned}$$

For the variance, we use the same partial fraction decomposition and the fact that the exponential holding times are independent to obtain

$$\begin{aligned} \text{Var}(T_{\text{MRCA}}(n)) &= \sum_{k=2}^{n} \text{Var}(T_k) = \sum_{k=2}^{n} \left(\frac{2}{k(k-1)}\right)^2 = 4 \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k}\right)^2 \\ &= 4 \sum_{k=2}^{n} \left[\left(\frac{1}{k-1}\right)^2 + \left(\frac{1}{k}\right)^2\right] - 8 \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k}\right). \end{aligned}$$

The rest of the calculation is straightforward. □

The theorem implies that

$$E(T_{\text{MRCA}}(2)) = 1 \quad \text{and} \quad \lim_{n\to\infty} E(T_{\text{MRCA}}(n)) = 2,$$

indicating that the time to the most recent common ancestor for a large sample is only about twice as large as that for a sample of size two. Moreover,

$$\begin{aligned} \text{Var}(T_{\text{MRCA}}(2)) &= 1 \\ \lim_{n\to\infty} \text{Var}(T_{\text{MRCA}}(n)) &= 4\pi^2/3 - 12 \approx 1.16 \end{aligned}$$

showing that the last coalescence time T_2 contributes much more to the variance of the time to the most recent common ancestor than the other coalescence times.

12.5 Moran model with selection

There are several more realistic versions of the Wright–Fisher and Moran models. In this last section, we study the most natural of these variants: the model with selection. In this version, one type has a selective advantage over the other type, i.e., both types no longer have the same fitness. Selection is incorporated in the model by assuming that the individuals no longer have the same probability of being selected as parent. For simplicity, we focus on the Moran model rather than the Wright–Fisher model with selection.

In the biological context, we usually think of type 0 as a resident and type 1 as a mutant. Each resident has fitness one while each mutant has fitness $\phi > 1$. At each time step, an individual is chosen at random with a probability proportional to its fitness, meaning that the probability of choosing a given mutant is ϕ times the probability of choosing a given resident. This individual gives birth to an offspring of its own type, which results in the offspring replacing an individual chosen uniformly at random so that the population size again remains constant, equal to N. A natural question about this model is: Assuming that a mutation produces one mutant in a population of residents, what is the probability that the mutant population eventually outcompetes the resident population? Using some connection with the gambler's ruin chain gives the following answer.

Theorem 12.5. *Let X_n be the number of mutants at time n. Then,*

$$P(X_n = N \textit{ for some } n) = (1-\phi^{-1})/(1-\phi^{-N}) \quad \textit{for all} \quad \phi > 1.$$

Proof. The process (X_n) is again a discrete-time Markov chain. At each time step, the number of mutants either increases or decreases by one, or remains the same, and the probability of an increase is the probability that a mutant is selected for reproduction times the probability that a resident is chosen to be replaced by the offspring. This gives the transition probabilities

$$p(x,x+1) = \frac{\phi x}{\phi x + (N-x)}\left(1 - \frac{x}{N}\right) \quad \text{for all} \quad 0 \le x \le N.$$

Exchanging the roles of mutants and residents, we get

$$p(x,x-1) = \frac{N-x}{\phi x + (N-x)} \left(\frac{x}{N}\right) \quad \text{for all} \quad 0 \leq x \leq N.$$

Finally, the transition probability $p(x,x)$ is simply equal to one minus the sum of the two probabilities above, since there is no other transition from state x. The Markov chain clearly has three communication classes: (1) the state where all the individuals are resident, which is an absorbing state; (2) the state where all the individuals are mutants, which is another absorbing state; and (3) all the other states, which form a transient communication class. In particular, the process eventually hits one of the two absorbing states. To find the probability that the state where all the individuals are mutants is reached before the other absorbing state, note that

$$\frac{p(x,x+1)}{p(x,x-1)} = \left(\frac{\phi x}{N-x}\right)\left(\frac{1-x/N}{x/N}\right) = \left(\frac{\phi x}{N-x}\right)\left(\frac{N-x}{x}\right) = \phi.$$

This implies that the limiting behavior of the Moran model with selection is the same as the limiting behavior of the gambler's ruin chain starting at one dollar with target N where the probability of winning and the probability of losing each game are respectively given by

$$p = \frac{\phi}{\phi+1} \quad \text{and} \quad q = \frac{1}{\phi+1} \quad \text{so} \quad q/p = \phi^{-1}.$$

In particular, starting with a single mutant, the probability that the mutants outcompete the resident population is given by the winning probability

$$P(X_n = N \text{ for some } n) = (1-\phi^{-1})/(1-\phi^{-N})$$

found in Example 5.1. □

For a discussion of the Moran model in the context of evolutionary graph theory and additional results based on heuristic arguments, we refer to [80].

12.6 Exercises

Exercise 12.1 (Wright–Fisher model with selection). In the Wright–Fisher model with selection, each type 0 individual has fitness one, whereas each type 1 individual has fitness $\phi \geq 1$. At each time step, individuals are independently selected for reproduction with a probability proportional to their fitness.

1. Write the transition probabilities of the process (X_n) that keeps track of the number of type 1 individuals.
2. Modify Program 7 to simulate this process and estimate the probability that type 1 wins for the system starting with a single type 1 individual.

Exercise 12.2 (Moran model with mutation). In this model, the offspring is of the same type as its parent with probability $1-\theta/N$ and of the other type with probability θ/N. The parameter θ measures the probability of a mutation and is divided by the population size to reflect the fact that mutations are rare. The evolution rules are otherwise the same as for the model with no mutation.

1. Write the transition probabilities of the process.
2. Prove that, whenever $\theta > 0$, there is a unique stationary distribution.
3. Modify Program 8 to simulate the process and study the fraction of time spent in each state. Focus in particular on the different shapes of the stationary distribution depending on whether θ is larger or smaller than one.

Exercise 12.3 (Majority rule model). The Moran model can be turned into an opinion model by thinking of the two types as two opinions: at each time step, a randomly chosen individual updates its opinion by mimicking another randomly chosen individual. A closely related model is the majority rule model introduced in [38] and studied in [63]. Individuals are either leftist or rightist and, at each time step, a discussion group of three individuals is chosen uniformly at random, which causes all three individuals to adopt the majority opinion of the group.

1. Letting X_n be the number of leftists at time n, compute
$$q_j(x) = P(X_{n+1} = x+j \,|\, X_n = x) \quad \text{for} \quad j = -1, 1 \text{ and } 1 < x < N-1.$$
2. Express the probability p_x that all the individuals are eventually leftist when starting with x leftist individuals as a function of
$$\mu_z = q_{-1}(z)/q_1(z) \quad \text{for} \quad 1 < z < N-1.$$
3. Deduce from the previous two questions that
$$p_x = \left(\frac{1}{2}\right)^{N-3} \sum_{z=0}^{x-2} \binom{N-3}{z} \quad \text{for all} \quad 1 < x < N-1.$$

Exercise 12.4 (Kingman's coalescent with mutation). In the presence of mutations as in the context of Exercise 12.2, the genealogy of a sample of n individuals is again described by Kingman's coalescent except that mutations occur independently at rate θ along the branches of the tree structure.

1. Compute the expected value of the cumulative length L of all the (vertical) branches in Kingman's coalescent's tree structure from the current time back to the time to the most recent common ancestor.
2. Deduce the expected value of the number of mutations M that occur since the time to the most recent common ancestor.

Hint: For the second question, condition on L.

Exercise 12.5 (Isothermal theorem). Consider the Moran model with selection including a population structure in the form of a weighted directed graph. More precisely, label the individuals $1,2,\ldots,N$ and let W be a stochastic matrix, i.e.,

$$W(i,j) \geq 0 \quad \text{and} \quad \sum_j W(i,j) = 1 \quad \text{for all} \quad i,j = 1,2,\ldots,N.$$

Individuals are either residents (type 0) or mutants (type 1) and are selected for reproduction as in the Moran model with selection with parameter ϕ. However, if the individual at vertex i is selected, its offspring is now sent to vertex j with probability $W(i,j)$ rather than to a vertex chosen uniformly at random.

1. Letting Z_n be the set of vertices occupied by a mutant at time n, compute
$$P(X_{n+1} = X_n \pm 1 \,|\, Z_n) \quad \text{where} \quad X_n = \text{card}(Z_n).$$
2. Define the temperature of vertex j as
$$T(j) = \sum_i W(i,j) = \text{total weight directed to vertex } j.$$
Prove that, when the graph is isothermal, i.e., all the vertices have the same temperature, the matrix W is doubly stochastic.
3. Deduce that the limiting behavior of (X_n) is the same as for the gambler's ruin chain where the probability of winning is $\phi/(\phi+1)$ at each game.
4. Conclude that, as for the Moran model with no population structure,
$$P(X_n = N \text{ for some } n) = (1-\phi^{-1})/(1-\phi^{-N}) \quad \text{for all} \quad \phi > 1.$$

Exercise 12.6. Consider an isothermal graph with vertices labeled $1,2,\ldots,N-1$, described by a stochastic matrix W. Then, add a vertex N and set

$$\begin{array}{rclll} \bar{W}(i,j) &=& W(i,j) & \text{for all} & i,j \neq N \\ \bar{W}(N,j) &=& 1/N & \text{for all} & j = 1,2,\ldots,N \\ \bar{W}(i,N) &=& 0 & \text{for all} & i = 1,2,\ldots,N-1. \end{array}$$

This defines a new stochastic matrix $\bar{W}$.

1. Prove that the weighted directed graph described by $\bar{W}$ is not isothermal.
2. Prove that the conclusion of the isothermal theorem in Exercise 12.5 does not hold in general for this graph, i.e.,
$$P(X_n = N \text{ for some } n) \neq (1-\phi^{-1})/(1-\phi^{-N}).$$

Chapter 13
Percolation models

We now make a U-turn going from dynamical spatially implicit processes in the last two chapters to static spatially explicit percolation models in this chapter. In particular, the actual spatial distance between individuals is no longer ignored and is in fact the key to constructing the models. The processes, however, no longer evolve in time, though, as we will see in a moment, one can interpret each realization as the final set resulting from the invasion of a rumor or disease that spreads in space. More precisely, percolation models are random graphs and therefore static random structures. These processes are generally difficult to study because there are only a few universal techniques available and the analysis of a new model often requires the development of new mathematical tools. In addition, quantities such as survival probabilities and critical values are usually impossible to compute explicitly, at least with the tools currently available. In particular, most of the results in the field of spatial stochastic processes are qualitative rather than quantitative.

The article of Broadbent and Hammersley [13] that introduces the bond percolation model studied at the beginning of this chapter is the first paper on percolation theory in the mathematics literature. Their work was motivated by the simple question: What is the probability that the center of a porous stone immersed in water is wet? Percolation models now arise in a wide variety of applied sciences such as physics, biology, and sociology, and are also used to understand other processes such as the particle systems that will be introduced at the end of this textbook. Problems in this field are easy to formulate but mathematically challenging.

In this chapter, we focus on bond percolation and oriented site percolation in two dimensions. The first section gives a description of bond percolation along with various objects of interest. In a nutshell, imagine that each edge of the two-dimensional lattice is independently open with probability p and consider the random subgraph induced by the open edges. The most natural question is: Is the connected component containing a given site, say the origin, finite or infinite? We say that percolation occurs when this connected component is infinite.

In the second section, we use this percolation model to illustrate an important technique in probability called coupling that will again appear in the

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_13

subsequent chapters. The basic idea is to construct processes with different parameters conjointly on the same probability space to show some monotonicity properties. This is used to prove that the probability that percolation occurs is nondecreasing with respect to the parameter p, a result that is intuitively obvious but difficult to prove because the probability of percolation cannot be computed explicitly. This implies in particular that the system exhibits at most one phase transition at some critical value for the parameter p. Below this critical value, there is no percolation, whereas above this critical value, the probability of percolation is positive.

The third section focuses on estimating this critical value. As previously mentioned, critical values are usually impossible to compute explicitly for spatial stochastic processes. To this extent, bond percolation in two dimensions appears as an exception because the critical value can be computed using a beautiful argument due to Kesten [51] based on planar duality. The objective of this section is to present the main ideas of the proof and some technical details will be omitted. In particular, the argument relies on the uniqueness of the so-called infinite percolation cluster, the proof of which is out of the scope of this textbook.

In the fourth section, we introduce oriented site percolation. The open cluster containing the origin and the percolation event are defined as for bond percolation. To briefly explain the name of the process, the word *site* means that the vertices rather than the edges are now open or closed. In addition, the word *oriented* means that the underlying deterministic graph is now a directed graph so open paths can only move in one direction. The orientation of the edges can be interpreted as the direction of time so that even if the object of interest is a static random subgraph, this subgraph can again be seen as the final set resulting from the invasion of a rumor or disease. The orientation can also be interpreted as gravity, in which case the random subgraph can be viewed as the set of sites that can be reached by a liquid.

As for bond percolation, a standard coupling argument shows that the probability of percolation is monotone with respect to the probability p that a given vertex is open, thus showing that there is at most one phase transition at some critical value for the parameter p. However, unlike for bond percolation, this critical value is not known, though lower and upper bounds can be found. In particular, the fifth section shows that the critical value is less than $80/81 < 1$, which implies that percolation can occur when p is close enough to one. The proof we give is self-contained and relies on another beautiful technique called the contour argument.

Further reading

- Kesten [52] is the first rigorous book on percolation theory.
- For a detailed introduction to bond percolation with a lot of helpful illustrations, we also refer the reader to Grimmett [40].
- A good reference on oriented site percolation is the review paper of Durrett [26].

13.1 Bond percolation in two dimensions

For concreteness, we describe the model in the context of epidemics. Imagine that each site of the two-dimensional integer lattice is occupied by an individual and that, independently of each other, any two nearest neighbors are friends with probability p. If an infection can only spread from friend to friend, what is the probability that an infection introduced at the origin will result in an epidemic, meaning that infinitely many individuals get infected?

Bond percolation in two dimensions is simply the process that consists of the countable collection of independent Bernoulli random variables with success probability p that keep track of the friendship relationships. More precisely, let

$$\xi(e) \sim \text{Bernoulli}(p) \quad \text{for each} \quad e \in E = \{(x,y) \in \mathbb{Z}^2 \times \mathbb{Z}^2 : \|x-y\| = 1\}.$$

In the terminology of percolation theory, edge e is said to be

$$\textbf{open} \text{ when } \xi(e) = 1 \text{ and } \textbf{closed} \text{ when } \xi(e) = 0.$$

We say that there is an **open path** between x and y, which we write $x \leftrightarrow y$, whenever there exists a sequence of vertices $x = x_0, x_1, \ldots, x_n = y$ such that

$$(x_i, x_{i+1}) \in E \quad \text{is open} \quad \text{for all } i = 0, 1, \ldots, n-1. \tag{13.1}$$

Define the **open cluster** at vertex x as

$$C_x = \{y \in \mathbb{Z}^2 : x \leftrightarrow y\}.$$

Bond percolation induces a random subgraph of the lattice, namely, the subgraph of the two-dimensional lattice induced by the set of open edges. For realizations of this random subgraphs, we refer the reader to Figure 13.1. The cluster C_0 is the connected component of this subgraph containing the origin. In particular, the event that an epidemic occurs, i.e., infinitely many individuals are ultimately infected, is the event that this cluster is infinite. Thus, the probability of an epidemic is given by the so-called **percolation probability**

$$\theta(p) = P_p(|C_0| = \infty) \quad \text{where} \quad |C_0| = \text{card}(C_0). \tag{13.2}$$

13.2 Monotonicity of the percolation probability

It should be intuitively clear that the percolation probability (13.2) is a nondecreasing function of the density p of open edges. This can be made rigorous using a so-called **coupling argument**. Coupling is a very powerful technique in probability theory in general and is not limited to percolation models. The basic idea is to find a joint construction of two random variables or stochastic processes, which is called a **coupling**, such that one dominates the other one with probability one. This allows

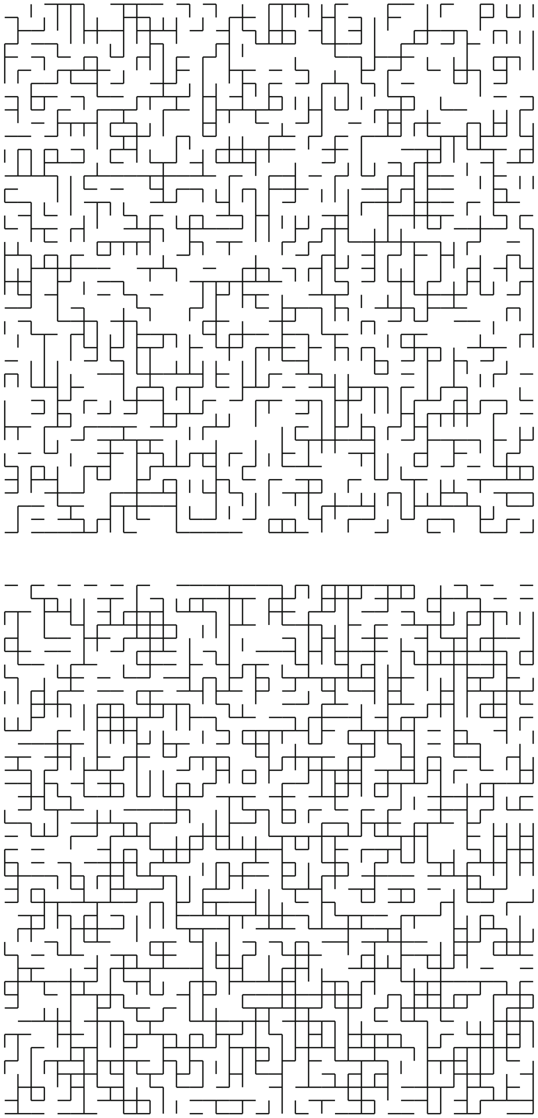


Fig. 13.1 Realizations of bond percolation with parameter $p = 0.45$ at the top and $p = 0.55$ at the bottom.

us to easily deduce a stochastic domination. In addition to establishing the monotonicity of the percolation probability, the following theorem gives an illustration of how coupling can be used in practice.

Theorem 13.1. *The function $p \mapsto \theta(p)$ is nondecreasing.*

Proof. Having two parameters $p_1 \le p_2$, the objective is to find a joint construction of bond percolation with parameter p_1 and bond percolation with parameter p_2 such that the latter dominates the former with probability one. To do this, we let

$$U(e) \sim \text{Uniform}\,(0,1) \;\; \text{for each edge } e \in E \tag{13.3}$$

be independent and define, for $i = 1,2$, the two collections

$$\xi_i(e) = 1 \;\; \text{if and only if} \;\; U(e) \leq p_i \;\; \text{for each edge } e \in E.$$

These random variables have the following three properties:

- The random variables $\xi_i(e)$, $e \in E$, are independent since they are constructed from (13.3) which are independent random variables.
- In addition, $P(\xi_i(e) = 1) = P(U(e) \leq p_i) = p_i$ hence $\{\xi_i(e) : e \in E\}$ indeed defines a realization of a bond percolation process with parameter p_i.
- Being constructed from the same collection of uniform random variables (13.3), the two percolation processes are not independent and we have

  $$\begin{aligned} \{\xi_1(e) = 1\} &= \{U(e) \leq p_1\} \\ &\subset \{U(e) \leq p_2\} = \{\xi_2(e) = 1\} \quad \text{for each } e \in E. \end{aligned}$$

  In particular, we have $(\xi_1) \leq (\xi_2)$, i.e., each of the coordinates of ξ_1 is, with probability one, at most equal to its counterpart for ξ_2.

In conclusion, we have the desired coupling. Now, define

$$f(\xi) = \mathbf{1}\{|C_0(\xi)| = \infty\} \quad \text{for all} \quad \xi : E \to \{0,1\}$$

where $C_0(\xi)$ is the cluster at the origin. Since $(\xi_1) \leq (\xi_2)$ and since f is clearly nondecreasing, we conclude that

$$\theta(p_1) = P(|C_0(\xi_1)| = \infty) = E(f(\xi_1)) \leq E(f(\xi_2)) = \theta(p_2).$$

This completes the proof. □

13.3 The critical phenomenon

The fact that the percolation probability is nondecreasing with respect to p implies the existence of a **critical value**: above this critical value, there is a positive probability of an epidemic, while below this critical value the probability of an epidemic is equal to zero. More precisely, we introduce

$$p_c = \sup\{p : \theta(p) = 0\}. \tag{13.4}$$

Looking more generally at the random subgraph of open edges rather than at just the cluster containing the origin, we note that whether this subgraph has an infinite component or not does not depend on the configuration of edges within any bounded Euclidean ball; therefore, this event is a tail event. In particular, Kolmogorov's zero-one law implies that this event has either probability zero or probability one. From this and the definition (13.4), we deduce

- For all $p > p_c$, the percolation probability $\theta(p) > 0$ and, with probability one, there exists at least one infinite cluster of open edges.
- For all $p < p_c$, the percolation probability $\theta(p) = 0$ and, with probability one, there is no infinite cluster of open edges.

Interesting but also challenging results that improve the previous observations state that, below the critical value, the size of a typical cluster decays exponentially, whereas above the critical value there exists exactly one infinite cluster of open edges. More precisely, we have the following two results.

Theorem 13.2. *Assume that $p < p_c$. Then,*

$$P(|C_0| > n) \le \exp(-\alpha n) \quad \textit{for some} \quad \alpha > 0.$$

Theorem 13.3. *Assume that $p > p_c$. Then, the subgraph induced by the open edges has a unique infinite connected component. This infinite connected component is called the* **infinite percolation cluster**.

The previous result on the uniqueness of the infinite percolation cluster in the supercritical regime is due to Burton and Keane [14] but the proof can also be found in the book of Grimmett [40, Chapter 8].

Critical values for spatially explicit models such as percolation processes and interacting particle systems are usually impossible to compute. However, this can be done for bond percolation in two dimensions relying on two ingredients: the existence and uniqueness of the infinite percolation cluster in the supercritical case and **planar duality**, which we now define.

Definition 13.1 (Planar graph). A planar graph is a graph that can be drawn on the plane in such a way that its edges do not intersect each other.

Definition 13.2 (Planar duality). The dual graph of a planar graph G is the graph

- that has a vertex corresponding to each plane region of G
- and an edge joining two neighboring regions for each edge in G.

Strictly speaking, the dual graph of a planar graph is not unique but different dual graphs are isomorphic. Figure 13.2 shows a picture of the complete graphs with three and four vertices, respectively, along with their dual graph. Note that the complete graph with four vertices is indeed a planar graph since it can be drawn without intersection of its edges, even though it is not the most natural way to draw it. Complete graphs with more vertices, however, are no longer planar graphs.

To use planar duality to study bond percolation in two dimensions, we first observe that the graph with vertex set and edge set given, respectively, by

$$\begin{aligned} \mathbb{D}^2 &= \{x + (1/2, 1/2) : x \in \mathbb{Z}^2\} \\ F &= \{(x,y) \in \mathbb{D}^2 \times \mathbb{D}^2 : \|x - y\| = 1\} \end{aligned} \tag{13.5}$$

is the dual graph of the two-dimensional integer lattice. We call this graph the dual lattice. Since it is isomorphic to the two-dimensional lattice itself, we say that the

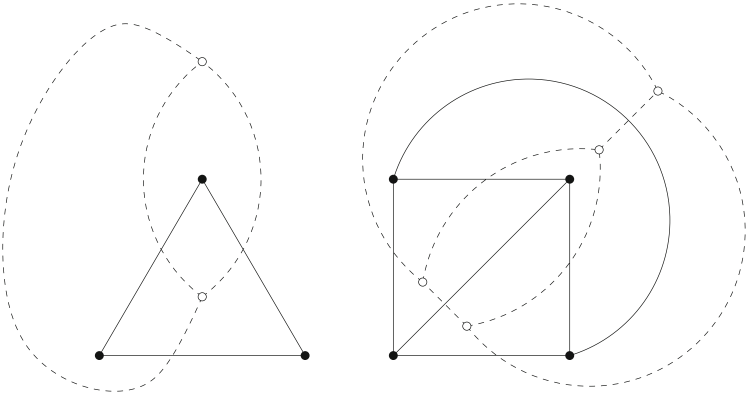


Fig. 13.2 Complete graphs with 3 and 4 vertices along with their dual graphs in dashed lines.

two-dimensional lattice is self-dual, which is one of the keys to computing the critical value of bond percolation. Since both graphs are dual of each other, each edge of the dual lattice intersects exactly one edge of the original lattice, and given a realization of bond percolation on $\mathbb{Z}^2$, we declare an edge of the dual lattice to be open if and only if the corresponding edge of the original lattice is open. This simple construction has two important consequences:

- Each realization of bond percolation on the original lattice induces a realization of bond percolation with the same parameter on the dual lattice (13.5).
- Through this coupling of bond percolation processes, open paths of the original lattice and closed paths of the dual lattice (13.5) cannot intersect.

With these preliminary results in hands, we can now obtain explicitly the critical value (13.4) of bond percolation, a result due to Kesten [51].

Theorem 13.4. *We have $p_c = 1/2$.*

Proof. The proof is divided into two steps.

Step 1 — First, we prove that $p_c \leq 1/2$. Referring to Figure 13.3, we consider the following two subgraphs of the original lattice and dual lattice:

- G_n is the $(n+1) \times n$ subgraph of the original lattice in solid lines.
- H_n is the $n \times (n+1)$ subgraph of the dual lattice in dashed lines.

We also introduce the corresponding events

$$\begin{aligned} A_n &= \text{there is an open path connecting the left and the right sides of the} \\ &\quad\ \text{subgraph } G_n \text{ of the original lattice} \\ B_n &= \text{there is a closed path connecting the top and the bottom sides of the} \\ &\quad\ \text{subgraph } H_n \text{ of the dual lattice.} \end{aligned}$$

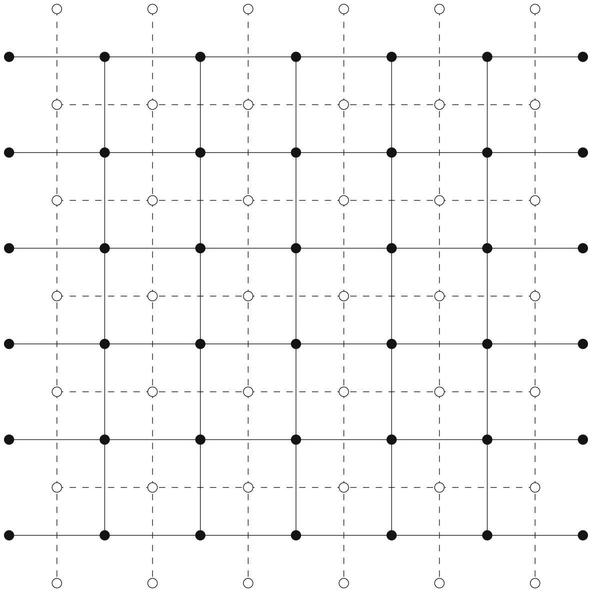


Fig. 13.3 Pictures related to the proof of Theorem 13.4.

For the coupling, the event A_n occurs if and only if B_n does not occur, which implies that both events form a partition of the sample space. Also, the probability of A_n when edges are open with probability $1-p$ is equal to the probability of B_n when edges are open with probability p. These two properties imply that

$$P_p(A_n) + P_p(B_n) = 1 \quad \text{and} \quad P_{1-p}(A_n) = P_p(B_n)$$

where the index refers to the parameter. In particular,

$$\begin{array}{lll} p < p_c & \text{implies} & \lim_{n\to\infty} P_p(A_n) = 0 \\ & \text{implies} & \lim_{n\to\infty} P_{1-p}(A_n) = \lim_{n\to\infty} P_p(B_n) = 1 \\ & \text{implies} & 1 - p \geq p_c \\ & \text{implies} & p \leq 1 - p_c \end{array}$$

from which it follows that $p_c \leq 1 - p_c$ therefore $p_c \leq 1/2$.

Step 2 — To establish the converse, we prove in fact a stronger result: the lack of percolation when the density of open edges is exactly one-half. To show this result, we proceed by contradiction and consider the new events

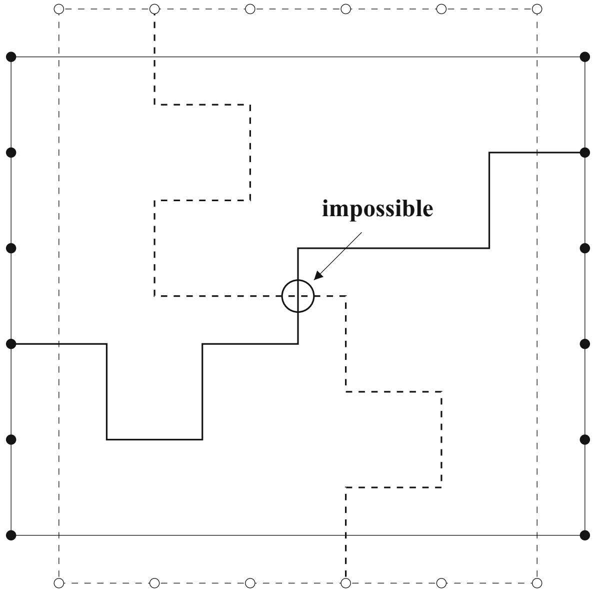


Fig. 13.4 Pictures related to the proof of Theorem 13.4.

A'_n = there are two infinite open paths outside the subgraph G_n one starting from its left side and one from its right side

B'_n = there are two infinite closed paths outside the subgraph H_n one starting from its bottom side and one from its top side.

Assuming by contradiction that $\theta(1/2) > 0$,

$$\lim_{n\to\infty} P_{1/2}(A'_n \cap B'_n) = 1.$$

The uniqueness of the infinite percolation cluster stated in Theorem 13.3 then implies that the two paths in the event A'_n are linked by an open path and the two paths in the event B'_n are linked by a closed path. As illustrated in Figure 13.4, this leads to the existence of an open path on the original lattice and a closed path on the dual lattice intersecting each other, which is not possible. In conclusion, percolation does not occur therefore we must have $p_c \geq 1/2$. □

Bond percolation can be defined in higher dimensions $d > 2$ as well. In this case, planar duality is no longer available because the integer lattice is not a planar graph, i.e., it cannot be drawn on the plane without intersection of its edges. However, it can be proved that the critical value is not degenerate, i.e., it belongs to $(0,1)$. More precisely, we have the following result.

Theorem 13.5. *In d dimensions, we have* $1/(2d-1) \leq p_c \leq 1/2$.

Proof. Bond percolation processes in different dimensions but with the same parameter can be coupled to show that the critical value (13.4) is nonincreasing with respect to the dimension. This, together with the previous theorem, gives the general upper bound $p_c \leq 1/2$.

Now, we observe that, to construct a self-avoiding path, i.e., a path that does not intersect itself, starting from the origin, there are $2d$ possible choices for the first edge and then at most $2d-1$ possible choices for each of the subsequent edges. In addition, by independence, the probability of a given self-avoiding path of length n being open is p^n, from which it follows that

$$\begin{aligned} P_p(|C_0| = \infty) &= \lim_{n\to\infty} P_p(C_0 \not\subset B(0,n)) \\ &\leq \lim_{n\to\infty} P_p(\text{there is an open path of length } n \text{ starting at } 0) \\ &\leq \lim_{n\to\infty} 2d\,(2d-1)^{n-1}\,p^n = 0 \end{aligned}$$

for all $p < (2d-1)^{-1}$. This completes the proof. □

This theorem will be used later to prove the existence of and construct graphically interacting particle systems on the infinite integer lattice.

13.4 Oriented site percolation in two dimensions

To define oriented site percolation in two dimensions, the first step is to introduce a deterministic directed graph and then attach a Bernoulli random variable to each vertex of the graph. Let $\mathscr{G}$ be the set of sites

$$\mathscr{G} = \{(x,n) \in \mathbb{Z} \times \mathbb{Z}_+ : x+n \text{ is even}\},$$

which we turn into a directed graph by putting an arrow

$$(x,n) \to (x',n') \quad \text{if and only if} \quad x' = x \pm 1 \text{ and } n' = n+1.$$

Site percolation consists of a collection of independent Bernoulli random variables with the same success probability p indexed by the set $\mathscr{G}$, i.e.,

$$\xi(x,n) \sim \text{Bernoulli}\,(p) \quad \text{for each site} \quad (x,n) \in \mathscr{G}.$$

The terminology is as before: Site (x,n) is said to be

$$\textbf{open} \text{ when } \xi(x,n) = 1 \;\text{ and }\; \textbf{closed} \text{ when } \xi(x,n) = 0.$$

There is an **open path** from $(0,0)$ to (x,n), which we write $(0,0) \to (x,n)$, whenever there exist $0 = x_0, x_1, \ldots, x_n = x$ such that

$$|x_{i+1} - x_i| = 1 \;\text{ for }\; 0 \leq i < n \quad \text{and} \quad (x_i, i) \;\text{ is open for }\; 0 \leq i \leq n.$$

Finally, we define the **open cluster** starting at the origin as

$$C_0 = \{(x,n) \in \mathscr{G} : (0,0) \to (x,n)\}.$$

As for bond percolation, the previous construction induces a random subgraph C_0 and we define the **percolation probability** as in (13.2) by setting

$$\theta(p) = P_p(|C_0| = \infty) \quad \text{where} \quad |C_0| = \text{card}\,(C_0), \tag{13.6}$$

which can be proved to be nondecreasing with respect to p relying again on a coupling argument. Putting the graph upside down, and thinking of the orientation of the graph as gravity, we define the set of **wet sites** at level n as

$$W_n = \{y \in \mathbb{Z} : (x,0) \to (y,n) \text{ for some } x \in 2\mathbb{Z}\}.$$

Note that wet sites are open, but open sites are not necessarily wet, as shown in the simulation pictures of Figure 13.5 in which open sites are represented by black dots, whereas wet sites are the one that can in addition be reached from level zero by a path of open sites. Using the terminology of epidemiology, one can think of open as susceptible and wet as infected: susceptible individuals become infected only if they can be reached by a path of infection.

13.5 Critical value and contour argument

The monotonicity of the percolation probability again implies the existence of a critical value, which is defined as in (13.4) by

$$p_c = \sup\{p : \theta(p) = 0\}.$$

Like for bond percolation, it can be proved that

- For all $p > p_c$, the percolation probability $\theta(p) > 0$ and, with probability one, there exists an infinite cluster of open sites.
- For all $p < p_c$, the percolation probability $\theta(p) = 0$ and, with probability one, there is no infinite cluster of open sites.

However, unlike the critical value of bond percolation in two dimensions, this critical value cannot be computed explicitly, but we can still prove that it is not degenerate in the sense that it is bounded away from zero and one. The fact that the critical value is strictly positive simply follows from the same argument as in the proof of Theorem 13.5. The proof that it is strictly smaller than one is more involved and relies on a so-called **contour argument**.

Theorem 13.6. *We have $p_c \leq 1 - 3^{-4} = 80/81$.*

Proof. To begin with, we let C_N denote the set of wet sites when starting from the following deterministic configuration of open sites at level zero:

$$W_0 = \{-2N, -2N+2, \dots, 0\}$$

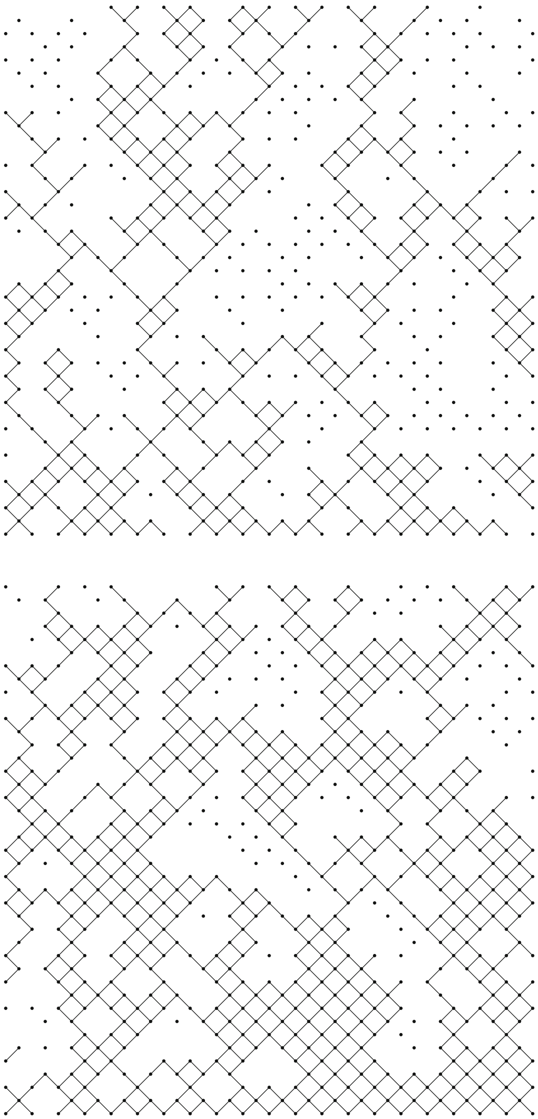


Fig. 13.5 Realizations of oriented site percolation with parameter $p = 0.70$ at the top and $p = 0.75$ at the bottom.

and observe that the percolation probability is bounded from below by

$$\begin{aligned} \theta(p) &= P_p(|C_0| = \infty) \\ &\geq P_p(|C_N| = \infty) \times P_p\,(W_N = \{-N,\ldots,N\}\,|\,W_0 = \{0\}) \\ &\geq P_p(|C_N| = \infty) \times p^{1+2+\cdots+N+(N+1)}. \end{aligned} \tag{13.7}$$

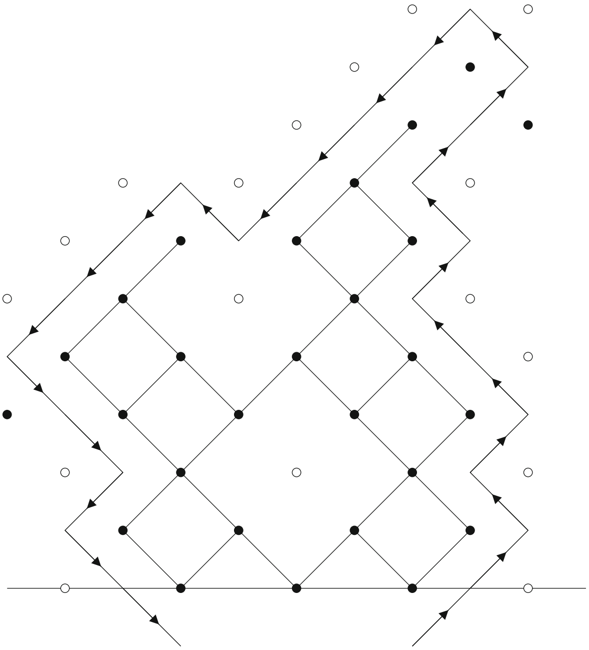


Fig. 13.6 Picture of the contour Γ.

In particular, it suffices to prove that

$$P_p(|C_N| < \infty) < 1 \quad \text{for all } p > 1 - 3^{-4} \text{ and all } N \text{ large.}$$

For each realization such that $|C_N| < \infty$, we define the oriented contour Γ shown in Figure 13.6. The basic idea is then to bound the probability

$$P_p(|C_N| < \infty) = P_p(\text{the contour } \Gamma \text{ exists}). \tag{13.8}$$

There are three keys to estimating this probability.

- The shortest possible contour has length $2N+4$.
- There are at most 3^m contours of length m.
- Sites to the right of an up-left or a down-left arrow must be closed. Moreover, sites to the right of two such arrows going up or two such arrows going down cannot be identical. Since in addition the contour ends on the left of where it starts, we must have the lower bound

$$\text{number of up-left arrows} + \text{number of down-left arrows} \geq m/2$$

  for any contour of length m, so at least $m/4$ sites must be closed.

It follows that, for all $p > 1 - 3^{-4}$ and all N large,

$$\begin{aligned} P_p(\Gamma \text{ exists}) &= \textstyle\sum_{m \geq 2N+4} P_p(\text{the contour } \Gamma \text{ has length } m) \\ &\leq \textstyle\sum_{m \geq 2N+4} 3^m (1-p)^{m/4} = \sum_{m \geq 2N+4} (3(1-p)^{1/4})^m < 1. \end{aligned}$$

This, together with (13.7)–(13.8), implies that $p_c \leq 1 - 3^{-4}$. □

13.6 Exercises

Exercise 13.1 (Percolation probability). Consider bond percolation in two dimensions with parameter p. Assume that the origin is a wet site and that each site connected to a wet site by an open edge is also wet. In particular, the resulting set of wet sites corresponds to the open cluster containing the origin, i.e., the set of sites that can be reached from the origin by a path of open edges.

1. Start from Program 10 to write a C program that displays the set of wet sites contained in the 100×100 square centered at the origin.
2. Use this program to also obtain an approximation of the percolation probability for various values of $p > 1/2$. Explain your approach.

Exercise 13.2 (Critical value). Consider the oriented site percolation process in two dimensions starting with all sites open at level zero.

1. Modify Program 11 to simulate the finite system stopped at a high level and check if there are wet sites at that level.
2. Use this program to estimate the critical value p_c of oriented site percolation in two dimensions. Explain your approach.

Exercise 13.3 (Density of occupied sites). Consider the oriented site percolation process with parameter p starting with all sites open at level zero. We are interested in the fraction of occupied sites at equilibrium:

$$\rho(p) = \lim_{n \to \infty} P(0 \in W_{2n}).$$

Modify Program 11 to simulate the finite system and estimate $\rho(p)$. The following numbers show an example of output where the (approximate) fractions of occupied sites are obtained from a single realization of the process:

```
r(0.75)=0.5924    r(0.80)=0.7266    r(0.85)=0.8164
```

Exercise 13.4 (Spatial correlations). Consider the oriented site percolation process in two dimensions starting with all sites open at level zero. To study the spatial correlations, we can look at the fraction of levels at which two given integers with the same parity are in the same state depending on their distance:

$$c(p,d) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}\{\text{card}(\{0,2d\} \cap W_{2k}) \neq 1\}$$

where p is the parameter of the model.

1. Modify Program 11 to simulate the process and study the effect of the distance d on the limit above.

The following table gives approximations of $c(p,d)$ obtained from one realization of the process on the torus with 100 vertices run until level 100,000.

```
c(.75,1)=.6052   c(.80,1)=.6368   c(.85,1)=.7103
c(.75,2)=.5533   c(.80,2)=.6104   c(.85,2)=.7033
c(.75,5)=.5220   c(.80,5)=.6063   c(.85,5)=.7030
c(.75,8)=.5193   c(.80,8)=.6034   c(.85,8)=.7018
```

For each parameter p, this fraction decreases with the distance.

2. Explain (heuristically) why this result is expected.
3. Explain how $c(p,d)$ when distance d is large can be approximated from the simulation results in Exercise 13.3.

Exercise 13.5 (Right edge). Consider the oriented site percolation process in two dimensions starting with only the origin open at level zero and let

$$r_n(p) = \sup\{x \in \mathbb{Z} : (x,n) \text{ is wet}\}$$

be the first coordinate of the rightmost wet site at level n.

1. Modify Program 11 to simulate this system and compute the position of the right edge $r_n(p)$ for various values of the level n and the parameter p.
2. Use this program to check that, conditional on the percolation event, $r_n(p)$ satisfies a law of large numbers, i.e.,

   $$(1/n)\, r_n(p) \xrightarrow{a.s.} \alpha(p)$$

   and estimate the limit $\alpha(p)$ for various values of p.

Exercise 13.6 (Forest fire model). Think of the Poisson points of a two-dimensional Poisson point process with intensity μ as the positions of trees that can be either alive or on fire. Initially, only one tree is on fire and, at each time step, all the trees within distance one of a burning tree start burning forever. Recall from Exercise 9.2 that the ultimate number of burning trees is almost surely finite when

$$\mu < \ln(7/6) \approx 0.154.$$

The objective of this problem is to show that, conversely, there is a positive probability that the fire spreads without bound if μ is larger than a specific value. To show this result, we first partition space into squares with area $1/5$ and declare a square to be open if it contains at least one Poisson point.

1. Compute the probability that a given square is open.
2. Give an upper bound for the Euclidean distance between two points that belong to two adjacent squares.
3. Deduce that the probability that the fire spreads without bound is positive whenever the intensity $\mu > 20\ln(3) \approx 21.972$.

Hint: For the last question, use Theorem 13.6.

Chapter 14
Interacting particle systems

Most mathematical models introduced in the life and social sciences literature that describe inherently spatial phenomena of interacting populations consist of systems of ordinary differential equations or stochastic counterparts assuming global interactions such as the logistic growth process and the Moran model. These models, however, leave out any spatial structure, while past research has identified the spatial component as an important factor in how communities are shaped, and spatial models can result in predictions that differ from nonspatial models.

In contrast, the framework of interacting particle systems that concludes this textbook is ideally suited to investigate the consequences of the inclusion of a spatial structure in the form of stochastic and local interactions. The framework was introduced in the early seventies independently by Spitzer [92] in the United States and Dobrushin [21, 22] in Soviet Union. In these models, members of the population (particles) such as cells, plants, agents, or players are traditionally located on the set of vertices of the d-dimensional integer lattice, but this can be extended to more general graphs. The dynamics is dictated by the geometry of the graph as particles can only interact with their neighbors, thus modeling an explicit spatial structure. These models include in particular the two key components of the models presented in the previous three chapters: a temporal structure like the logistic growth process and the Moran model and an explicit spatial structure like percolation models. The main objective of research in the field of interacting particle systems is to deduce the macroscopic behavior of the system and the emergence of spatial patterns from the microscopic rules described by the local interactions. Good references on this topic are Liggett [69, 70] and Durrett [27].

This chapter starts with a definition of the general framework of interacting particle systems. An important difference with the continuous-time Markov chains studied previously is that, at least when the underlying spatial structure is infinite, the state space is uncountable so the dynamics cannot be described using an intensity matrix. Instead, the processes are defined from rate functions describing local interactions taking place in a collection of finite neighborhoods.

The second and third sections focus on two of the early models: the contact process and the voter model that can be seen, respectively, as the simplest invasion

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_14

model and the simplest competition model including an explicit space. These processes are important from a mathematical perspective because they have been used as test models to develop the theory. From a modeling point of view, the reader will learn that the contact process can be seen as the spatial analog of the logistic growth process and the voter model as the spatial analog of the Moran model.

The rest of the chapter is concerned with graphical representations. The main objective is to prove that interacting particle systems on infinite graphs indeed exist. To see that the existence of these systems is not obvious, observe that, contrary to the continuous-time Markov chains studied earlier, when there are infinitely many sites, there are also infinitely many updates in finite time, so the time to the first jump does not exist. The key to the proof is the so-called graphical representation introduced by Harris [43]. The basic idea is to attach at each site a collection of independent Poisson processes describing the local interactions of the system and use the existence of a phase transition for bond percolation to show that there is a random partition of the graph into islands with the following properties: each island is almost surely finite and sites belonging to different islands do not influence each other before some small positive time. This allows us to construct the process on each island separately up to this time and, using the Markov property, at any later time. Though the graphical representation was originally introduced to prove the existence of a large class of interacting particle systems, the construction is also useful in practice together with the superposition and thinning properties of Poisson processes to simulate processes on finite graphs.

The fourth section shows how the contact process and the voter model can be constructed heuristically from a graphical representation. The fifth section gives algorithms to construct and simulate the contact process and the voter model, while the last section shows how the graphical representation can be used to prove the existence of these systems on infinite graphs. We will see in the next two chapters that the graphical representation can also be used to couple interacting particle systems with different parameters or starting from different initial configurations, and to construct so-called dual processes.

14.1 General framework

From a mathematical point of view, an interacting particle system is a continuous-time Markov chain with a state at time t that is the function

$$\xi_t : \mathbb{Z}^d \to \{0, 1, \dots, \kappa - 1\}.$$

In the terminology of interacting particle systems,

- Elements $x \in \mathbb{Z}^d$ are called **sites** or, using the terminology of graph theory, **vertices**, and have to be thought of as spatial locations.

- Elements $i \in \{0,1,\ldots,\kappa-1\}$ are called **types** or **colors**, e.g., in population dynamics, the types can be empty and occupied; in disease dynamics, healthy and infected; in opinion dynamics, Democrat and Republican, and so on.
- The state of the process at a given time is called a **spatial configuration**.
- We have $\xi_t(x) =$ type of vertex x at time t.

Note that there is a natural one-to-one correspondence

$$\phi : \{0,1\}^{\mathbb{N}} \to [0,1] \quad \text{with} \quad \phi((a_0,a_1,a_2,\ldots)) = \textstyle\sum_n a_n 2^{-(n+1)}$$

which implies that the cardinality of the state space is at least $2^{\aleph_0} = \aleph_1$ therefore the state space is not countable. In particular, the dynamics of an interacting particle system cannot be described by an intensity matrix (10.4). Instead, the dynamics is described by local interactions indicating the rate at which a vertex changes its type based on the state of the system in a neighborhood of the vertex. The fact that this rate only depends on a neighborhood models the presence of a spatial structure. To define the dynamics, we let

$$N_x = \{y \in \mathbb{Z}^d : \|x-y\| = 1\} \quad \text{for all} \quad x \in \mathbb{Z}^d \tag{14.1}$$

be the **interaction neighborhood** of vertex x. Then, it is assumed that the type at vertex x flips from type i to type $j \neq i$ at a rate

$$\begin{aligned} &\lim_{\varepsilon\downarrow 0} \varepsilon^{-1} P(\xi_{t+\varepsilon}(x) = j \,|\, \xi_t \text{ with } \xi_t(x) = i) \\ &\qquad = c_{i\to j}(x,\xi_t) = c_{i\to j}(\xi_t(y) : y \in N_x) \end{aligned} \tag{14.2}$$

that only depends on the configuration in the neighborhood (14.1) of vertex x, which is referred to as local interactions. Note that the framework can be easily extended to more general connected graphs $G = (V,E)$ by simply redefining

$$N_x = \{y \in V : (x,y) \in E\} \quad \text{for all} \quad x \in V. \tag{14.3}$$

The graph G needs to be thought of as a network of interactions: vertices x and y can only interact if they are connected by an edge. The inclusion of a spatial structure can thus be divided into three different levels:

1. **No space** means no interaction: the branching process and the simple birth and death process are examples of such processes because the number of offspring produced by a given individual does not depend on the state of the process.
2. **Implicit space** means global interactions: there are spatial locations, but all pairs of locations are equally likely to interact so there is no underlying geometrical structure. Examples of such models are the logistic growth process, the Wright–Fisher model and the Moran model.
3. **Explicit space** means local interactions: there are spatial locations along with a geometry described by a graph and pairs of locations can only interact if they

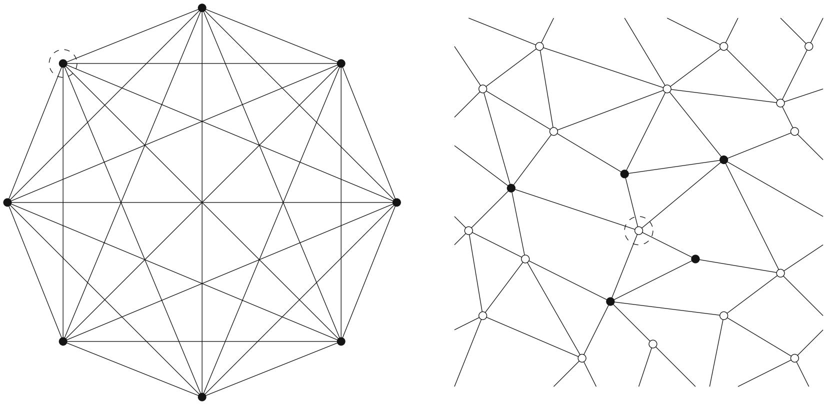


Fig. 14.1 Complete graph and planar graph. The vertices colored in black represent the neighbors of the vertex circled in dashed line.

are connected by an edge. Such a spatial component is taken into account by interacting particle systems.

To understand how the general local transition rates (14.2) defined above work in practice, we now focus on the contact process and the voter model, two of the simplest and most popular examples of interacting particle systems.

14.2 Invasion: the contact process

The contact process, introduced in the seventies by Harris [44], is the simplest example of invasion model based on the framework of particle systems. To fix the ideas, we describe the model in the context of population dynamics but it can be turned into an epidemic model by replacing empty by healthy and occupied by infected. Each site of the d-dimensional integer lattice is either empty or occupied by an individual and the evolution is described by the following rules:

- Independently of each other, individuals give birth to a single offspring at rate β and die at the normalized rate one.
- At birth, the offspring is sent to one of the $2d$ nearest neighbors of the parent's location chosen uniformly at random. If this target site is empty then it becomes occupied, otherwise the birth is suppressed.

The local transition rates of the contact process can be computed using the superposition and thinning properties of Poisson processes: since an empty site receives an offspring from each of its occupied neighbor at rate $\beta/2d$, the local transition rates for the contact process are given by

$$c_{0\to 1}(x,\xi_t) = \left(\frac{\beta}{2d}\right) \sum_{y\in N_x} \xi_t(y) \quad \text{and} \quad c_{1\to 0}(x,\xi_t) = 1. \tag{14.4}$$

As mentioned above, the contact process can as well be defined on more general graphs. An interesting particular case is the **complete graph** with N vertices. Recall that in such a graph all the vertices are connected to each other and for our purpose it is also convenient to assume that each vertex is connected to itself. Figure 14.1 gives a picture of the complete graph with eight vertices. In this particular case, the interaction neighborhood (14.3) of each vertex is equal to the entire vertex set therefore the description of the contact process on the complete graph reduces to the following two simple rules:

- Independently of each other, individuals give birth to a single offspring at rate β and die at the normalized rate one.
- At birth, the offspring is sent to a site chosen uniformly at random. If the site is empty it becomes occupied, otherwise the birth is suppressed.

The number of individuals in a population evolving according to these rules is exactly the logistic growth process. In particular, the number of individuals in the contact process on the complete graph reduces to the **logistic growth process** so the process can be viewed as the spatial analog of the logistic growth process.

14.3 Competition: the voter model

The voter model was introduced in the seventies independently by Clifford and Sudbury [18] and Holley and Liggett [45]. This is the simplest example of competition model based on the framework of interacting particle systems. In the voter model, each site of the d-dimensional integer lattice is occupied by an individual that can be either of type 0 or of type 1. As for the contact process, there are different possible interpretations. One can think of both types as two possible opinions, which explains the name of the model. In this context, the evolution is described by the following simple rule:

- Independently of each other, individuals update their opinion at rate one by mimicking one of their $2d$ nearest neighbors chosen uniformly at random.

One can also think of both types as two different types of allele in a population of haploid individuals. In this population genetics context,

- Individuals give birth independently at rate one to a single offspring that then replaces one of the parent's $2d$ neighbors chosen uniformly at random.
- Each individual is of the same type as its parent.

That we consider the voter model as a model for the dynamics of opinions or as a model for the dynamics of genes, the transition rates are the same and can be found by again using the superposition and thinning properties. Using the terminology of population genetics, note that each individual receives an offspring of the other type from each of its neighbors of the other type at rate $1/2d$, therefore the transition rates of the voter model are given by

$$\begin{aligned} c_{0\to 1}(x,\xi) &= \left(\frac{1}{2d}\right) \sum_{y\in N_x} \xi(y) \\ c_{1\to 0}(x,\xi) &= \left(\frac{1}{2d}\right) \sum_{y\in N_x} (1-\xi(y)). \end{aligned} \tag{14.5}$$

As previously, the model can be defined on more general graphs, and we look at the particular case of the complete graph. In this case, since the interaction neighborhood of each vertex is equal to the entire vertex set, the evolution rules become:

- Individuals give birth at rate one to a single offspring which then replaces an individual chosen uniformly at random from the population.
- Each individual is of the same type as its parent.

Note that the number of type 1 individuals in a population evolving according to these two rules defines a continuous-time Markov chain in which the sequence of states visited is the same as for the Moran model. In particular, the number of type 1 individuals in the voter model on the complete graph is closely related to the **Moran model** so the voter model can be viewed as the spatial analog of the Moran model and its cousin the **Wright–Fisher model**.

The advantage of thinking of spatially implicit stochastic models as interacting particle systems on the complete graph is that very powerful techniques developed to study interacting particle systems can also be used for spatially implicit models, which often gives interesting results.

14.4 Graphical representation

As for continuous-time Markov chains with a finite state space, interacting particle systems can be constructed from a collection of independent Poisson processes. However, since there are *a priori* infinitely many sites, there are also infinitely many collections of Poisson processes, but because the dynamics is described by local interactions, there are only finitely many Poisson processes at each site. Looking at the d-dimensional integer lattice and adding one dimension for time, the graphical representation can be seen as a random graph in dimension $d+1$, that we typically call a percolation structure.

For concreteness, we now deal with the contact process. There are two types of events for this process: births and deaths. To take care of the birth events, we attach $2d$ independent Poisson processes to each site, one for each neighbor, while to take care of the death events, which are spontaneous, we only attach one Poisson process to each site. More precisely, we let

- $\{T_n(x,y) : n \geq 1\}$ for $y \in N_x$ be the times of a rate $\beta/2d$ Poisson process,
- $\{U_n(x) : n \geq 1\}$ for $x \in \mathbb{Z}^d$ be the times of a rate one Poisson process.

To turn this into a percolation structure,

- we draw an arrow $x \to y$ at times $T_n(x,y)$ to indicate that a birth occurs if the tail of the arrow is occupied and the tip of the arrow is empty and
- we put a cross at site x at times $U_n(x)$ to indicate that a death occurs at vertex x in case it is occupied.

A realization of the resulting graph is represented in Figure 14.2 where it is used to construct the contact process starting with all sites occupied: space-time points which are occupied are drawn in thick lines.

The graphical representation of the voter model is simpler because it only involves one type of events at each vertex. In this case, we attach $2d$ independent Poisson processes to each site by letting

- $\{T_n(x,y) : n \geq 1\}$ for $y \in N_x$ be the times of a rate $1/2d$ Poisson process,

which we turn into a percolation structure by

- drawing an arrow $x \to y$ at times $T_n(x,y)$ to indicate that the individual at the tip of the arrow takes on the type of the one at the tail of the arrow.

Figure 14.3 shows how to construct the voter model given a particular initial configuration and a realization of the graphical representation. The thick lines represent the space-time region occupied by type 1 individuals.

14.5 Numerical simulations in finite volume

The graphical representation can be used to simulate numerically interacting particle systems. Here, we first give an algorithm to simulate the contact process on a finite grid with periodic boundary conditions. This algorithm relies on the superposition and thinning properties of Poisson processes.

(a) **When** — On the $N \times N$ lattice, the time of the first potential update, i.e., the first time a symbol appears in the graphical representation, is

$$\begin{aligned} T &= \min\{T_1(x,y) : (x,y) \in E\} \wedge \min\{U_1(x) : x \in V\} \\ &\sim \text{Exponential}\,((\beta+1)N^2). \end{aligned}$$

(b) **Where** — Since the total rate is the same at each vertex, each of the vertices is equally likely to be the one at which the first potential event occurs therefore we select a vertex uniformly at random:

$$X \sim \text{Uniform}\{1,2,\ldots,N\}^2.$$

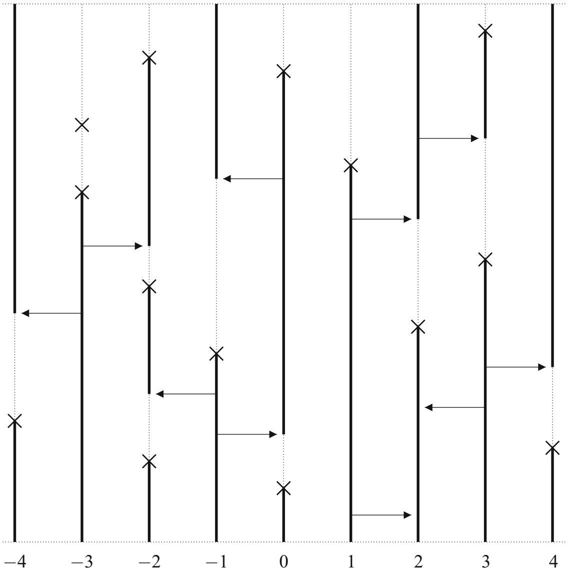


Fig. 14.2 Graphical representation of the contact process.

(c) **What** — If site X is empty, nothing happens and we go to (d). Otherwise, by the standard properties of the exponential distribution,

- With probability $1/(\beta+1)$, the first symbol that appears is a death mark, so with this probability we kill the individual at vertex X.
- With probability $\beta/(\beta+1)$, the first symbol that appears is an arrow, so with this probability we choose one of the four neighbors of vertex X uniformly at random and put a particle at this site if it is empty.

(d) By memoryless of the exponential random variable, the characteristics of the next update are the same as for the first update so we can go back to (a).

The algorithm to simulate the voter model is simpler.

(a) **When** — On the $N \times N$ lattice, the time of the first potential update, i.e., the first time a symbol appears in the graphical representation, is now

$$T = \min\{T_1(x,y) : (x,y) \in E\} \sim \text{Exponential}(N^2).$$

(b) **Where** — As previously, the first potential update is at vertex

$$X \sim \text{Uniform}\{1,2,\ldots,N\}^2.$$

(c) **What** — We choose one of the four neighbors of vertex X uniformly at random and set the type at X equal to the type of this neighbor.

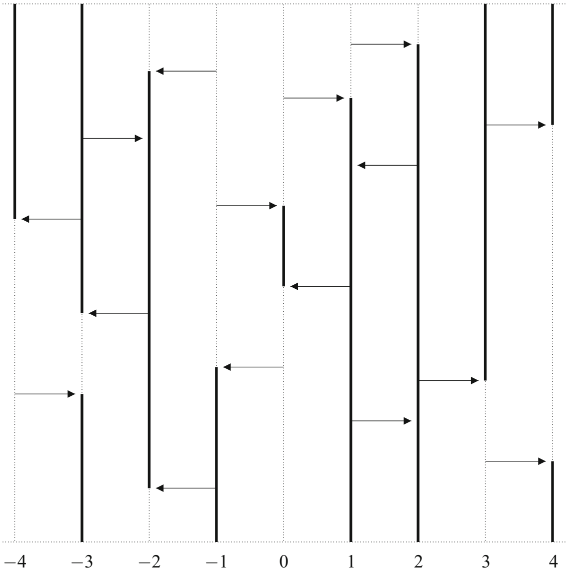


Fig. 14.3 Graphical representation of the voter model.

(d) Again, by memoryless of the exponential random variable, we can go back to (a) to determine the time and location of the next update.

For simulations of the contact process and voter model written in C using these two algorithms, see the simulation chapter at the end of the textbook.

14.6 Existence on infinite graphs

In this section, we use the graphical representation to prove that interacting particle systems on infinite graphs indeed exist. Since there are infinitely many Poisson processes, the time to the first update does not exist. Also, to prove the existence of the process, one must show that the state of any given space-time point can be deduced from finitely many interactions that can then be ordered in time, a result due to Harris [43]. For concreteness, we focus on the contact process but the proof easily extends to the general framework described above provided the degree of the graph is uniformly bounded. First, we let $\varepsilon > 0$ be small and, having a realization of the graphical representation, declare an edge (x,y) to be open if and only if there is an arrow that connects both vertices by time ε. Then, defining open paths as for bond percolation (13.1), we call island of influence of x the random cluster

$$C_x(\varepsilon) = \{y \in \mathbb{Z}^d : \text{there is an open path connecting } x \text{ and } y\}.$$

It follows that the graphical representation outside $C_x(\varepsilon)$ is irrelevant in determining the state of vertex x up to time ε. Moreover, with probability one, the island of influence of every vertex is finite provided ε is small enough.

Lemma 14.1. *For $\varepsilon > 0$ small enough,*

$$P(|C_x(\varepsilon)| = \infty \textit{ for some } x \in \mathbb{Z}^d) = 0.$$

Proof. From the graphical representation of the contact process, it is clear that edges are independently open with probability

$$\begin{aligned} p_\varepsilon &= P(\text{there is an arrow } x \to y \text{ or } y \to x \text{ by time } \varepsilon) \\ &= P(\text{Exponential}\,(\beta/d) < \varepsilon) \\ &= 1 - \exp(-\beta\varepsilon/d). \end{aligned}$$

In particular, by Theorem 13.5, which gives bounds for the critical value of bond percolation, all the islands of influence are almost surely finite when

$$1 - \exp(-\beta\varepsilon/d) < 1/(2d-1)$$

which holds for all $\varepsilon > 0$ small. □

The previous lemma breaks down the lattice into finite islands that do not interact with each other before time ε therefore the process can be constructed independently on each of those islands up to this time. Since the exponential random variable is memoryless, and so the graphical representation translation invariant in time, we can restart the procedure to construct the process up to time 2ε and so on. In conclusion, the process is well defined at all times.

Chapter 15
The contact process

In this chapter, we prove that, similarly to the other invasion models introduced in this textbook, the contact process exhibits a phase transition: There is a critical birth parameter above which the process starting with one individual survives with positive probability but below which it goes extinct eventually with probability one. Note that the inclusion of local interactions induces multiple notions of survival and we distinguish two levels: identifying the configuration of the process at time t with the set of occupied sites at that time, we say that

- The process **dies out** when $\xi_t = \varnothing$ for some t.
- The process **survives weakly** when $\xi_t \neq \varnothing$ for all t.
- The process **survives strongly** when $0 \in \xi_t$ infinitely often.

The starting point is to think of the process as being generated by a graphical representation, as defined above. Using the graphical representation, we can prove the following three important properties of the contact process.

Monotonicity — The probability of survival starting from a given configuration is nondecreasing with respect to the birth parameter. As for percolation models, this implies the existence of at most one phase transition at some critical value.

Attractiveness — The space-time set of occupied sites is also nondecreasing for the inclusion with respect to the initial configuration. A consequence of this property known as attractiveness is that the process starting from the all occupied configuration converges in distribution to an invariant measure called the upper invariant measure since this is the largest one.

Self-duality — Keeping track of the potential ancestors of a given space-time point going backward in time through the graphical representation defines a new process called the dual process. This process turns out to have the same distribution as the contact process itself, from which we deduce an interesting connection between the probability of survival and the upper invariant measure.

As for percolation models, we can also prove that the critical value is not degenerate so there is indeed a phase transition. In particular, we give a lower bound for the crit-

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_15

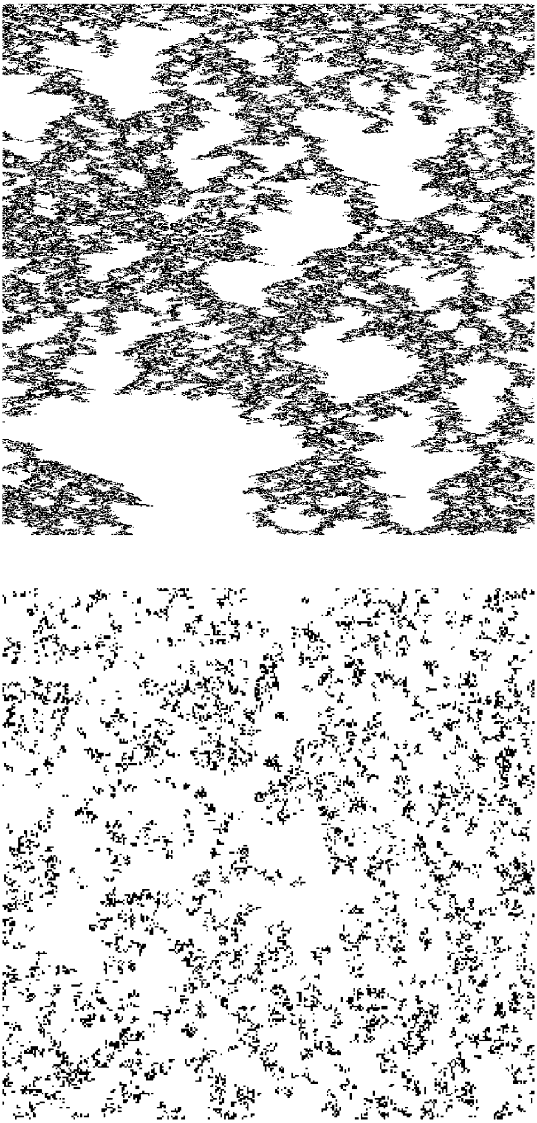


Fig. 15.1 Realization of the contact process with birth parameter $\beta = 3.30$ from time 0 at the top of the picture to time 10,000 at the bottom on the one-dimensional torus with 600 vertices, and snapshot at time 1000 of the contact process with birth parameter $\beta = 1.70$ on a 300×300 lattice with periodic boundary conditions.

ical value on the lattice to show an important feature of the contact process: due to the inclusion of local interactions, the critical value is strictly larger than one,which contrasts with nonspatial and spatially implicit invasion models. We conclude the chapter with an overview of more sophisticated results about the convergence and the shape of the process, including the surprising fact that it is possible for the system on homogeneous trees to survive weakly but not strongly. For realizations of the contact process on lattices, we refer the reader to Figure 15.1.

15.1 Monotonicity and attractiveness

In this section, we prove that the occupied space-time region of the contact process is stochastically nondecreasing with respect to both its birth parameter and its initial configuration. The proofs are based on coupling arguments to compare processes with different birth parameters or initial configurations. To state our results, we let (ξ_t^i) be the contact processes with parameter β_i starting from ξ_0^i.

Lemma 15.1 (Monotonicity). *For all* $(x,t) \in \mathbb{Z}^d \times \mathbb{R}_+$,

$$P(x \in \xi_t^1) \leq P(x \in \xi_t^2) \quad \textit{when} \quad \xi_0^1 = \xi_0^2 \quad \textit{and} \quad \beta_1 \leq \beta_2.$$

Proof. The basic idea is to couple the two processes in such a way that the occupied space-time region for the first process is included in its counterpart for the second process. To do this, we define a coupling of their graphical representations using the superposition property of Poisson processes as follows.

- For all $x, y \in \mathbb{Z}^d$ with $\|x-y\| = 1$, we draw a continuous arrow $x \to y$ at the arrival times of a Poisson process with parameter $\beta_1/2d$.
- For all $x, y \in \mathbb{Z}^d$ with $\|x-y\| = 1$, we draw a dashed arrow $x \to y$ at the arrival times of a Poisson process with parameter $(\beta_2 - \beta_1)/2d$.
- For all $x \in \mathbb{Z}^d$, we put a death mark × at vertex x at the arrival times of a Poisson process with parameter one.

The contact process (ξ_t^1) can be constructed from this graphical representation by ignoring the dashed arrows. In addition, because the superposition of the first two collections of Poisson processes above results in a collection of Poisson processes with intensity $\beta_2/2d$, the contact process (ξ_t^2) can be constructed from this graphical representation by using both types of arrows, meaning that births occur through both continuous and dashed arrows.

Since the set of arrows for the first process is included in the set of arrows for the second process, the occupied space-time region for the first process is included in its counterpart for the second process, as shown in Figure 15.2. Finally, let

$$f_{x,t}(\{\xi_s : s \in \mathbb{R}_+\}) = \mathbf{1}\{x \in \xi_t\} \quad \text{for all} \quad \xi : \mathbb{R}_+ \to \{0,1\}^{\mathbb{Z}^d}.$$

Since the function $f_{x,t}$ is nondecreasing, we conclude that

$$\begin{aligned} P(x \in \xi_t^1) &= E(f_{x,t}(\{\xi_s^1 : s \in \mathbb{R}_+\})) \\ &\leq E(f_{x,t}(\{\xi_s^2 : s \in \mathbb{R}_+\})) = P(x \in \xi_t^2). \end{aligned}$$

This completes the proof. □

It directly follows from the lemma that the probability of weak/strong survival of the process starting from a single individual is nondecreasing with respect to the birth parameter. More precisely, the two functions of β defined as

$$\begin{aligned} \theta_W(\beta) &= P_\beta(\xi_t \neq \varnothing \text{ for all } t \,|\, \xi_0 = \{0\}) \\ \theta_S(\beta) &= P_\beta(0 \in \xi_t \text{ infinitely often} \,|\, \xi_0 = \{0\}) \end{aligned} \tag{15.1}$$

are nondecreasing. The monotonicity of the probability of weak/strong survival with respect to the birth parameter motivates the introduction of

$$\beta_W = \sup\{\beta : \theta_W(\beta) = 0\} \quad \text{and} \quad \beta_S = \sup\{\beta : \theta_S(\beta) = 0\}. \tag{15.2}$$

Some important properties of these critical values will be given later.

Having proved monotonicity, we now study attractiveness. The process is said to be attractive if the occupied space-time region is stochastically nondecreasing with respect to the initial configuration, as opposed to the birth parameter.

Lemma 15.2 (Attractiveness). *For all* $(x,t) \in \mathbb{Z}^d \times \mathbb{R}_+$,

$$P(x \in \xi_t^1) \leq P(x \in \xi_t^2) \quad \textit{when} \quad \xi_0^1 \subset \xi_0^2 \quad \textit{and} \quad \beta_1 = \beta_2.$$

Proof. This is similar to the proof of Lemma 15.1. Because the birth parameters are now equal, both processes can be constructed from the same graphical representation. The key is to observe that, the graphical representation being fixed, the occupied space-time region is nondecreasing with respect to the initial configuration, and then we use the function $f_{x,t}$ to conclude. □

Letting μ_t be the distribution at time t of the contact process starting from the all-occupied configuration, it is clear that $\mu_t \leq \mu_0$. Then, using attractiveness and the fact that the contact process is a Markov process, we deduce that

$$\mu_{t+s} \leq \mu_s \quad \text{for all} \quad s \geq 0.$$

In particular, starting from the all-occupied configuration,

$$\bar{\mu} = \lim_{t \to \infty} \mu_t$$

exists. It is called the **upper invariant measure** of the contact process and, again due to attractiveness, this is the largest possible stationary distribution. There is also an interesting connection between the upper invariant measure and the survival probability of the process starting from a single individual that will be explained later using duality techniques.

15.2 Self-duality

In this section, we define the dual process of the contact process and explain the connection between the upper invariant measure and the survival probability. Duality is a powerful tool in the field of interacting particle systems, but it is also model specific and, to this extent, is not always available or mathematically tractable.

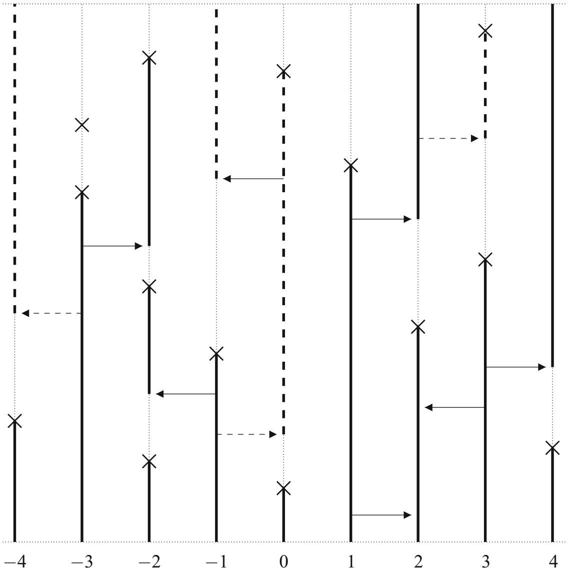


Fig. 15.2 Coupling of contact processes with different birth parameters.

The dual process is defined from a graphical representation of the process, which we assume to be fixed from now on.

Definition 15.1. We say that there is a **path** $(z,s) \to (x,t)$ if there exist

$$s = s_0 < s_1 < \cdots < s_{n+1} = t \quad \text{and} \quad z = x_0, x_1, x_2, \ldots, x_n = x \in \mathbb{Z}^d$$

such that the following two conditions hold:

- for $i = 1,2,\ldots,n$, there is an arrow $x_{i-1} \to x_i$ at time s_i and
- for $i = 0,1,\ldots,n$, there is no $\times$ on the segment $\{x_i\} \times (s_i, s_{i+1})$.

If this holds, we also say that there is a **dual path** $(x,t) \to (z,s)$. In other words, dual paths are defined as paths except that they evolve backward in time and follow the arrows of the graphical representation in the opposite direction.

The **dual process** starting at (x,t) is then defined as

$$\hat{\xi}_s(x,t) = \{z \in \mathbb{Z}^d : \text{there is a dual path } (x,t) \to (z,t-s)\}$$

for all $0 \le s \le t$. The dual process of the contact process has two important properties. First, since exchanging the direction of the arrows as well as the direction of time does not change the distribution of the graphical representation, the distribution of the process starting from a single individual at vertex x is equal to the distribution of its dual process starting at (x,t), i.e.,

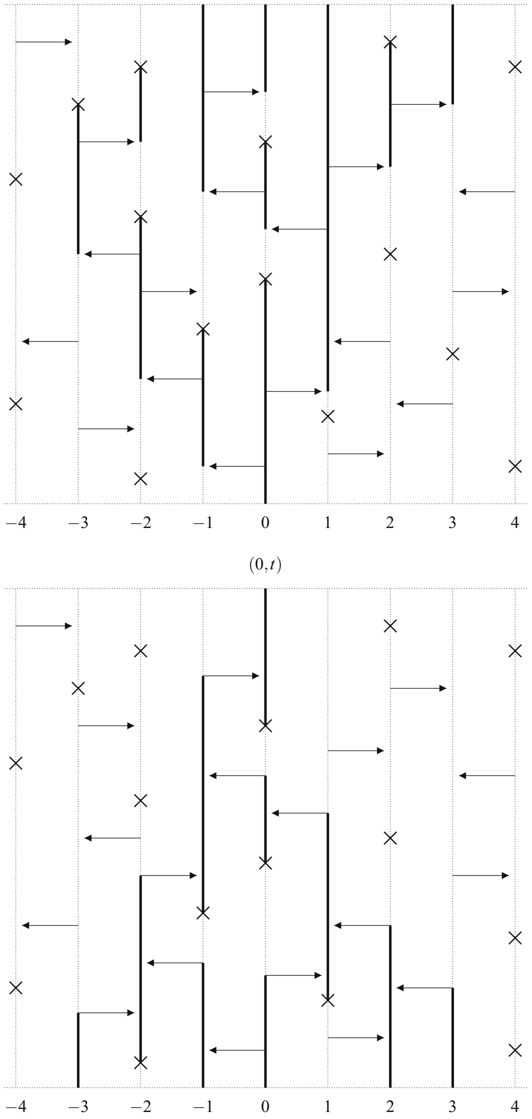


Fig. 15.3 Realization of the contact process starting from a single individual at the origin and realization of the dual process starting from $(0,t)$. The duality relationship (15.4) indicates that, for this realization, the origin is occupied at time t if and only if at least one of the five sites at the bottom of the picture is occupied.

$$\{\xi_s : 0 \leq s \leq t\} = \{\hat{\xi}_s(x,t) : 0 \leq s \leq t\} \quad \text{in distribution.} \tag{15.3}$$

Figure 15.3 shows a picture of these two random sets in thick lines. In addition, it follows from the definition of paths and dual paths and from the construction of the contact process from the graphical representation that we have the following important property called **duality relationship**:

$$\begin{array}{lll} x \in \xi_t & \text{if and only if} & \xi_s \cap \{z : \text{there is a path } (z,s) \to (x,t)\} \neq \varnothing \\ & \text{if and only if} & \xi_s \cap \{z : \text{there is a dual path } (x,t) \to (z,s)\} \neq \varnothing \\ & \text{if and only if} & \xi_s \cap \hat{\xi}_{t-s}(x,t) \neq \varnothing. \end{array} \qquad (15.4)$$

Setting $\xi_0 = \mathbb{Z}^d$ and $s = 0$ in (15.4) and taking the probability, we obtain that, for the contact process starting from the all-occupied configuration,

$$P(x \in \xi_t) = P(\hat{\xi}_t(x,t) \cap \xi_0 \neq \varnothing) = P(\hat{\xi}_t(x,t) \neq \varnothing).$$

Then, using (15.3) and taking the limit as $t \to \infty$, we get

$$\begin{array}{rl} \bar{\mu}\,(\{x \in \xi\}) & = \lim_{t\to\infty} P(x \in \xi_t) \\ & = \lim_{t\to\infty} P(\hat{\xi}_t(x,t) \neq \varnothing) = \lim_{t\to\infty} P_x(\xi_t \neq \varnothing) \\ & = P_x(\xi_t \neq \varnothing \text{ for all } t) = \theta_W(\beta) \end{array} \qquad (15.5)$$

where the subscript x indicates that the process starts with a single individual at vertex x. In other words, self-duality implies that the probability that a given vertex is occupied under the upper invariant measure is equal to the probability that the process starting with a single individual survives weakly. Therefore, an immediate consequence of (15.5) is that

$$\bar{\mu} \neq \delta_\varnothing \quad \text{if and only if} \quad \theta_W(\beta) > 0$$

where the measure $\delta_\varnothing$ is the measure that concentrates on the configuration in which all sites are empty, meaning that the process starting from the all-occupied configuration survives strongly if and only if the probability of weak survival starting with a single individual is positive.

Another interesting consequence of (15.5) is that it explains a result seen earlier about the logistic growth process. Indeed, self-duality still holds for the contact process on the complete graph, which is the logistic growth process. In this context, (15.5) says that the fraction of occupied sites under the quasi-stationary distribution of the logistic growth process should approach the probability that the process starting from a single individual survives for a long time, which itself is approximately the survival probability of the simple birth and death process.

15.3 The critical value

Having a better understanding of the contact process in the supercritical regime, we now look more carefully at the two critical values defined in (15.2). Because strong survival implies weak survival, we clearly have $\beta_W \leq \beta_S$. In fact, it is known that, at least for the contact process on the infinite integer lattice, these two critical values are equal. In particular, except maybe at the critical value, the process cannot survive weakly but not strongly, meaning that the process either dies out or expands in space. Throughout this section, we let β_c be the common value of the critical value for weak survival and the critical value for strong survival.

The contact process modified so that births onto sites that are already occupied are not suppressed is called the **branching random walk**. The number of individuals in this process evolves according to the simple birth and death process with birth rate β. In particular, we deduce from a standard coupling argument that, provided both processes start from a single individual, the number of individuals in the birth and death process dominates stochastically its counterpart in the contact process. This, together with (10.23), implies that $\beta_c \geq 1$.

As previously mentioned, an interesting aspect of the contact process is that, in contrast with branching processes and the simple birth and death process, the critical value is strictly larger than one, which is due to the inclusion of a spatial structure in the form of local interactions. To prove this result, we first define

$$\sigma(A) = P(\xi_t \neq \varnothing \text{ for all } t \,|\, \xi_0 = A) \quad \text{for all} \quad A \subset \mathbb{Z}^d \text{ finite}, \tag{15.6}$$

and prove the following lemma.

Lemma 15.3. *The probability* (15.6) *is* **submodular***, i.e.,*

$$\sigma(A \cup B) + \sigma(A \cap B) \leq \sigma(A) + \sigma(B) \quad \textit{for all} \quad A, B \subset \mathbb{Z}^d \textit{ finite}. \tag{15.7}$$

Proof. This again follows from a coupling argument. Having a graphical representation of the contact process for a given birth parameter, we simultaneously construct the four contact processes whose initial configurations consist of the four sets in the statement of the lemma using this same graphical representation. Writing the initial configuration as a superscript, for this coupling,

$$P(\xi_t^{A \cup B} = \xi_t^A \cup \xi_t^B) = 1 \quad \text{and} \quad P(\xi_t^{A \cap B} \subset \xi_t^A \cap \xi_t^B) = 1 \quad \text{for all} \quad t > 0.$$

In particular, defining the Bernoulli random variables

$$\phi(D) = \mathbf{1}\{\xi_t^D \neq \varnothing \text{ for all } t > 0\} \quad \text{for all} \quad D \subset \mathbb{Z}^d$$

we deduce that, with probability one,

$$\phi(A \cup B) = \phi(A) \vee \phi(B) \quad \text{and} \quad \phi(A \cap B) \leq \phi(A) \wedge \phi(B).$$

Taking the expected value, we conclude that

$$\begin{aligned} \sigma(A \cup B) + \sigma(A \cap B) &= E(\phi(A \cup B)) + E(\phi(A \cap B)) \\ &\leq E(\phi(A) \vee \phi(B)) + E(\phi(A) \wedge \phi(B)) \\ &= E(\phi(A)) + E(\phi(B)) = \sigma(A) + \sigma(B). \end{aligned}$$

This completes the proof. □

Using the previous lemma and a first-step analysis, we can now prove that the critical value of the contact process is strictly larger than one.

Theorem 15.1. *For the process on $\mathbb{Z}^d$, we have $\beta_c \geq 2d/(2d-1) > 1$.*

Proof. Let $A \subset \mathbb{Z}^d$ be finite and recall that

- an occupied site $x \in A$ becomes empty at rate one,
- an empty site z with an occupied neighbor x in the set A becomes occupied due to a birth that originates from x at rate $\beta/2d$.

In particular, a first-step analysis gives

$$\sigma(A) = \sum_{x \in A} \frac{\sigma(A - \{x\}) + (\beta/2d) \sum_{z \in N_x} \sigma(A \cup \{z\})}{(\beta+1)\,\mathrm{card}\,(A)}. \tag{15.8}$$

Now, consider the two probabilities

$$\sigma_1 = \sigma(\{x\}) \quad \text{and} \quad \sigma_2 = \sigma(\{x,y\}) \quad \text{where} \quad \|x-y\| = 1.$$

By translation invariance, these two probabilities do not depend on the choice of x and y. Using (15.8) with $A = \{x,y\}$ as well as (15.7), we get

$$\begin{aligned} 2(\beta+1)\,\sigma_2 &= 2\sigma_1 + (\beta/d) \textstyle\sum_{z \in N_x} \sigma(\{x,y,z\}) \\ &= 2\sigma_1 + (\beta/d)\,\sigma_2 + (\beta/d) \textstyle\sum_{z \in N_x, z \neq y} \sigma(\{x,y,z\}) \\ &\leq 2\sigma_1 + (\beta/d)\,\sigma_2 + (\beta/d)(2d-1)(2\sigma_2 - \sigma_1). \end{aligned}$$

Some basic algebra gives

$$\begin{aligned} &(\sigma_1 - \sigma_2)(2d - \beta\,(2d-1)) \\ &\quad = \; 2d\sigma_1 + \beta\sigma_2 + \beta\,(2d-1)(2\sigma_2 - \sigma_1) - 2d\,(\beta+1)\,\sigma_2 \geq 0. \end{aligned} \tag{15.9}$$

Now, observe that, when $\beta > \beta_c$, the probability of survival starting with two individuals is strictly larger than the probability of survival starting with one individual, from which it follows that

$$\begin{array}{lll} \sigma_2 > \sigma_1 > 0 & \text{when} & \beta > \beta_c \\ \sigma_2 = \sigma_1 = 0 & \text{when} & \beta < \beta_c. \end{array} \tag{15.10}$$

Combining (15.9)–(15.10), we deduce that

$$\begin{array}{lll} \beta < 2d/(2d-1) & \text{implies that} & \sigma_1 \geq \sigma_2 \\ & \text{implies that} & \sigma_1 = \sigma_2 = 0 \\ & \text{implies that} & \text{the process dies out.} \end{array}$$

In particular, we must have $\beta_c \geq 2d/(2d-1)$. □

To conclude, we give an overview of the contact process with the results we have proved so far and results that are significantly more difficult to establish.

15.4 Overview of the contact process

The contact process was only well understood in one dimension until the nineties when Bezuidenhout and Grimmett [5, 6] proved a number of open problems about the process in higher dimensions. The common background of their proofs is the use of a block construction to couple the contact process properly rescaled in space and time with oriented site percolation. Note indeed the similarities between the two models: Thinking of the set of sites that are wet at level n as the set of sites that are occupied at time n, one can view oriented site percolation as a discrete-time version of the contact process, even though the details behind the construction in [5, 6] are far from being that simple.

As previously explained, we have $\beta_c = \beta_W = \beta_S > 1$ for the contact process on infinite integer lattices, indicating that above the common critical value the process survives strongly, whereas below the critical value the process dies out. Even though this critical value is not known, it is proved in [5] that the process dies out at criticality, just like the branching process, the simple birth and death process and bond percolation in two dimensions.

Theorem 15.2. *For the process on $\mathbb{Z}^d$, we have $\theta_W(\beta_c) = \theta_S(\beta_c) = 0$.*

In addition to the coupling with oriented site percolation, the proof relies on a continuity argument to show that the parameter region in which the process survives is an open set and so, by the monotonicity result proved above, an open interval.

Recall that the contact process is attractive, which guarantees that the process starting from the all occupied configuration converges in distribution to a certain invariant measure $\bar{\mu}$ called the **upper invariant measure**. In addition, because the process is self-dual, the probability that a given vertex is occupied under the upper invariant measure is equal to the probability that the process starting with a single individual survives. An immediate consequence is that

$$\bar{\mu} \neq \delta_\varnothing \quad \text{if and only if} \quad \theta_W(\beta) = \theta_S(\beta) > 0$$

where the measure $\delta_\varnothing$ is the measure that concentrates on the configuration in which all sites are empty. This shows that the supercritical contact process has at least two stationary distributions: the smallest one is $\delta_\varnothing$ and the largest one is the upper invariant measure $\bar{\mu}$.

The next natural step is to determine whether the process can have intermediate stationary distributions. The following theorem, called a **complete convergence theorem**, shows that there is no additional stationary distribution. More precisely, the theorem states that if the contact process survives, then it must converge to its upper invariant measure.

Theorem 15.3. *For all $A \subset \mathbb{Z}^d$ finite,*

$$\xi_t \xrightarrow{d} \sigma(A)\,\bar{\mu} + (1 - \sigma(A))\,\delta_\varnothing.$$

Recall that $\sigma(A)$ in the statement of the theorem has been defined earlier as the survival probability of the contact process starting from $\xi_0 = A$.

Next, we look at the geometry of the set of occupied sites for the process starting with a single individual at the origin. According to the previous theorem, conditional on survival, the process converges in distribution to $\bar{\mu}$ and we let K_t be the random set in which the contact process starting with a single individual at the origin is distributed according $\bar{\mu}$ at time t. Our definition of this set is not rigorous, but can be made rigorous by using a coupling between the process starting with a single individual and the process starting with all sites occupied. The next theorem, called a **shape theorem**, states that this set grows linearly.

Theorem 15.4. *There is a deterministic convex set $U \subset \mathbb{R}^d$ such that*

$$(1-\varepsilon)(U \cap \mathbb{Z}^d) \subset (1/t)\,K_t \subset (1+\varepsilon)(U \cap \mathbb{Z}^d)$$

eventually for all $\varepsilon > 0$.

Numerical simulations of the supercritical contact process suggest that the convex set U is simply a Euclidean ball centered at the origin.

Bezuidenhout and Grimmett [6] also proved an important result in the subcritical case: The time to extinction of the subcritical contact process starting with a single individual decays exponentially, which can be seen as the analog of Theorem 13.2 regarding bond percolation.

Theorem 15.5. *Assume that $\beta < \beta_c$. Then,*

$$P(\xi_t \neq \varnothing \,|\, \xi_0 = \{0\}) \leq \exp(-\alpha t) \quad \textit{for some} \quad \alpha > 0.$$

In other words, the process dies out exponentially fast.

Right after the publication of [5, 6], research about the contact process on the infinite integer lattice obviously slowed down. However, researchers in the field of interacting particle systems started studying the contact process on other infinite connected graphs. The first remarkable result in this topic is due to Pemantle [82] who proved that $\beta_W < \beta_S$ for the contact process on homogeneous trees with degree larger than two. More precisely, we have the following theorem.

Theorem 15.6. *For the process on the homogeneous tree with degree d,*

$$\beta_W \leq 1/(d-1) \quad \textit{and} \quad \beta_S \geq 1/(2\sqrt{d}).$$

The theorem gives $\beta_W < \beta_S$ when $d \geq 6$ but these bounds are improved in [82] to also prove that the strict inequality holds more generally for all $d > 2$. This means that the contact process on regular trees exhibits two phase transitions: There is an intermediate phase in which the process survives weakly but not strongly. In this intermediate phase, each bounded region of the tree becomes and remains empty after some almost surely finite time but the population can survive globally by drifting off to infinity. See also [89, chapter 7] for additional results.

15.5 Exercises

Exercise 15.1 (Probability of survival). Consider the contact process on a large one-dimensional torus with parameter β starting from a single occupied site.

1. Modify Program 12 to simulate this process and estimate the probability of survival of the infinite counterpart. Justify your approach.

The following numbers are the (approximate) probabilities of survival obtained from 500 independent realizations of the process for different birth rates.

```
p(3)=0.000    p(4)=0.604    p(5)=0.738    p(6)=0.814
```

The simulation results suggest that the critical value of the one-dimensional contact process is between 3 and 4.

2. Use the simulation from question 1 to improve this estimate.

Exercise 15.2 (Density of occupied sites). Consider the contact process on a large one-dimensional torus with parameter β starting from the all-occupied configuration. We are interested in the fraction of occupied sites at equilibrium:

$$\rho(\beta) = \lim_{t\to\infty} \frac{1}{N} \sum_{x=1}^{N} \xi_t(x).$$

where N is the total number of sites.

1. Modify Program 12 to simulate this process and estimate the fraction $\rho(\beta)$ of sites that are occupied at equilibrium.

The following numbers are the (approximate) fractions of occupied sites obtained from one realization of the process for different birth rates.

```
r(4)=0.605    r(5)=0.731    r(6)=0.790    r(7)=0.828
```

2. For the same birth rate, these numbers are close to the survival probabilities shown in Exercise 15.1. Explain why this is expected.

Exercise 15.3 (Spatial correlations). Consider the contact process on a large one-dimensional torus starting from the all-occupied configuration. To study the spatial correlations, we can look at the fraction of time two sites are in the same state depending on the distance between these two sites:

$$c(\beta,n) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}\{\eta_s(0) = \eta_s(n)\}\, ds$$

where β is the birth rate.

1. Modify Program 12 to simulate the process and study the effect of the distance n on the fraction of time $c(\beta,n)$.

The following table gives the fraction $c(\beta,n)$ obtained from one realization of the process for various birth rates and distances.

```
c(4,1)=0.701      c(5,1)=0.706      c(6,1)=0.737
c(4,2)=0.616      c(5,2)=0.643      c(6,2)=0.688
c(4,5)=0.547      c(5,5)=0.609      c(6,5)=0.674
c(4,8)=0.527      c(5,8)=0.600      c(6,8)=0.664
```

For each birth rate, this fraction decreases with the distance.

2. Explain (heuristically) why this result is expected.
3. Explain how $c(\beta,n)$ when distance n is large can be approximated from the simulation results in Exercise 15.2.

Exercise 15.4 (Threshold contact process). In the threshold contact process, individuals die independently at rate one, while empty sites with at least τ occupied neighbors become occupied at rate β. The integer τ is called the threshold.

1. Write the transition rates of the process.
2. Use a coupling argument to show that the survival probability starting from a given configuration is nonincreasing with respect to the threshold.
3. It is known that the basic contact process dies out when $\beta = 1$. Modify Program 12 to simulate the threshold contact process in two dimensions and check if this process also dies out when $\beta = 1$.

Exercise 15.5 (Multitype contact process). In the process introduced in [75], each site of the integer lattice is either empty, occupied by a type 1 individual or occupied by a type 2 individual, so the state at time t is a function

$$\xi_t : \mathbb{Z}^d \to \{0,1,2\} \quad \text{where} \quad 0 = \text{empty},\ 1 = \text{type } 1,\ 2 = \text{type } 2.$$

Individuals give birth to an offspring of their own type at rate β and die at the normalized rate one. At birth, the offspring is sent to one of the nearest neighbors of the parent's location chosen uniformly at random. Like in the contact process, the birth is suppressed if the target site is already occupied.

1. Write the transition rates of the process.
2. Find the values of β for which the population dies out/survives.
3. Modify Program 12 to simulate the process in two dimensions and study numerically its long-term behavior.

Exercise 15.6 (Forest fire model). The forest fire model [30] is another natural variant of the contact process. Each site of the lattice is either occupied by a live tree, on fire or burnt, so the state at time t is a function

$$\xi_t : \mathbb{Z}^d \to \{0,1,2\} \quad \text{where} \quad 0 = \text{alive},\ 1 = \text{on fire},\ 2 = \text{burnt}.$$

Burning trees send out sparks at rate β, which causes one of the neighboring trees chosen uniformly at random to start burning in case it is not burnt yet. Trees burn for an exponentially distributed amount of time with parameter one and then burn out, while a new tree grows at each burnt site at rate α.

1. Write the transition rates of the process.
2. Explain how to construct the process from a graphical representation.
3. Describe the possible long-term behaviors the process can exhibit and for which values of α and β they are expected.

Chapter 16
The voter model

As in the Wright–Fisher and Moran models, the two configurations in which all the individuals have the same type or share the same opinion are absorbing states for the voter model. However, at least starting with infinitely many individuals of each type, the time to fixation to one of these absorbing states is almost surely infinite, which allows for the possibility of interesting transient or long-term behaviors. In fact, even though it does not depend on any parameter, the voter model exhibits very rich dynamics with various behaviors depending on the spatial dimension. This contrasts with the contact process that has behavior similar in any dimension.

The key to studying the process is again duality. In this chapter, we first rely on the graphical representation introduced previously to describe the dual process. Contrary to the contact process, the voter model is not self-dual. Instead, there is a duality relationship between the voter model and coalescing random walks, which is the main ingredient to determine the type of an individual at the current time based on the initial configuration of the system. This duality relationship is used to prove the first main result about the voter model: it clusters in one and two dimensions, meaning that any finite set of individuals share the same type with probability converging to one, whereas coexistence occurs in higher dimensions. This phase transition is due to the recurrence/transience property proved earlier for symmetric random walks on the infinite integer lattice.

As for the contact process, we conclude the chapter with an overview of the model with additional important results answering in particular the following questions. How fast do the clusters grow in one and two dimensions? How strong are the spatial correlations at equilibrium in higher dimensions when coexistence is possible? What is the fraction of time a given site is of a given type? Though the proofs are somewhat technical and omitted in this textbook, they all heavily rely on the duality between the voter model and coalescing random walks. For realizations of the voter model on lattices, we refer the reader to Figure 16.1.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_16

16.1 Duality with coalescing random walks

To exhibit the duality relationship between the voter model and coalescing random walks, we first think of the process as being constructed from the graphical representation described above. Having a realization of the graphical representation, and following the same approach as for the contact process, we first define paths and dual paths. In the case of the voter model, these are defined as follows.

Definition 16.1. We say that there is a **path** $(z,s) \to (x,t)$ if there exist

$$s = s_0 < s_1 < \cdots < s_{n+1} = t \quad \text{and} \quad z = x_0, x_1, x_2, \dots, x_n = x \in \mathbb{Z}^d$$

such that the following two conditions hold:

- for $i = 1,2,\dots,n$, there is an arrow $x_{i-1} \to x_i$ at time s_i and
- for $i = 0,1,\dots,n$, there is no arrow pointing at $\{x_i\} \times (s_i, s_{i+1})$.

If this holds, we also say that there is a **dual path** $(x,t) \to (z,s)$, i.e., as before, dual paths evolve backward in time and follow the arrows in the opposite of their direction. For each $A \subset \mathbb{Z}^d$ finite, the **dual process** starting at (A,t) is

$$\hat{\xi}_s(A,t) = \{z \in \mathbb{Z}^d : \text{there is a dual path } (x,t) \to (z,t-s) \text{ for some } x \in A\}.$$

This process is defined for all $0 \le s \le t$. The bottom part of Figure 16.2 shows a picture of the dual process in thick lines. Note that, contrary to the dual process of the contact process, the dual process of the voter model does not branch or die, so the dual process starting from a single space-time point consists of a single dual path. Each time it encounters a space-time point where a birth occurs, i.e., the tip of an arrow, this dual path jumps to the parent's location. In particular, dual paths follow the ancestors so, when $A = \{x\}$, we have the **duality relationship**

$$\xi_t(x) = \xi_{t-s}(\hat{\xi}_s(x,t)) \quad \text{for all} \quad 0 \le s \le t. \tag{16.1}$$

Now, becasuse the arrows in the graphical representation point at a given vertex at the times of a Poisson process with intensity one and are equally likely to originate from each of the $2d$ neighbors of that vertex, dual paths evolve according to continuous-time symmetric random walks run at rate one. Looking more generally at the dual process starting from a finite set, we note that, when two dual paths intersect they coalesce, as illustrated in Figure 16.2. This shows that the dual process of the voter model consists of a system of **coalescing random walks**: there is one symmetric random walk starting at each point in A moving backward in time and each time one random walk jumps onto another one, both random walks coalesce.

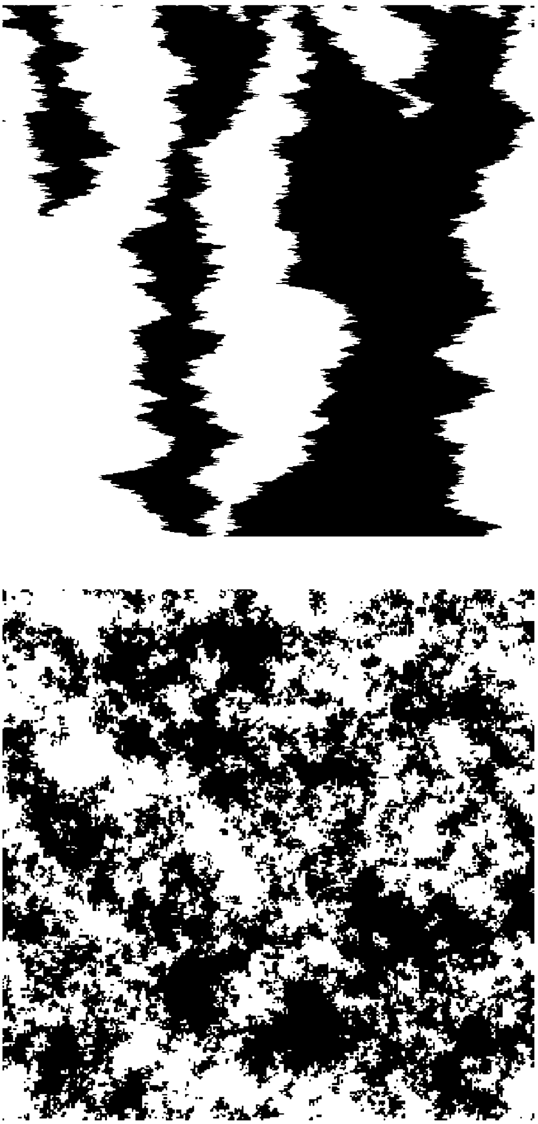


Fig. 16.1 Realization of the voter model from time 0 to time 10,000 on the one-dimensional torus with 600 vertices, and snapshot at time 1000 of the voter model on a 300 × 300 lattice with periodic boundary conditions.

16.2 Clustering versus coexistence

From now on, we assume that the process starts from the product measure with density $\rho \in (0,1)$ in which sites are of type 1 with probability ρ independently of each other. Using the duality between the voter model and coalescing random walks, we first prove that the system clusters in one and two dimensions.

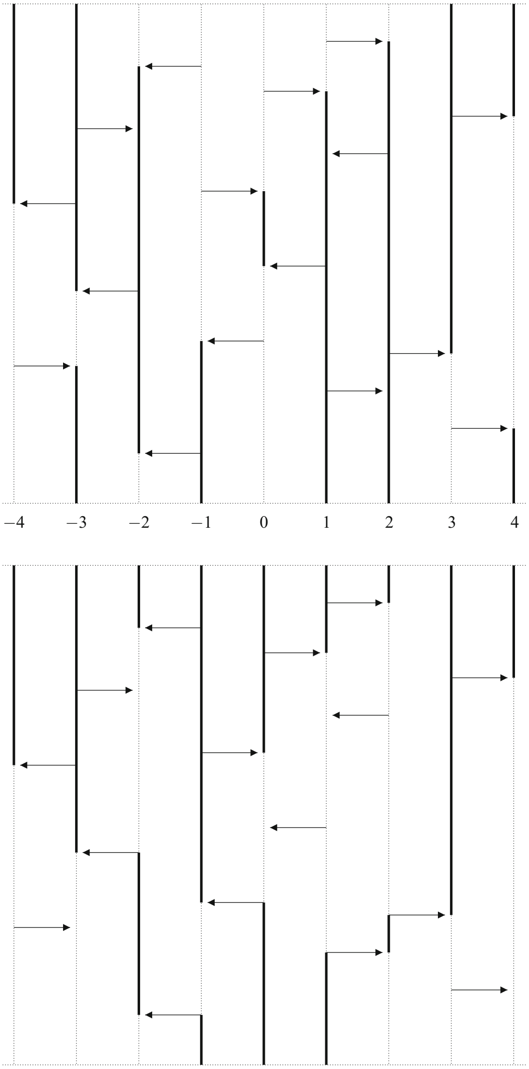


Fig. 16.2 Graphical representation and dual process of the voter model.

Theorem 16.1. *Assume that $d \leq 2$. Then,*

$$\lim_{t\to\infty} P(\xi_t(x) \neq \xi_t(y)) = 0 \quad \textit{for all} \quad x,y \in \mathbb{Z}^d.$$

Proof. Since the dual paths starting at (x,t) and (y,t) evolve according to independent random walks run at rate one until they coalesce, the difference between the random walks, which we denote by

$$X_s = \hat{\xi}_s(x,t) - \hat{\xi}_s(y,t) \quad \text{for all} \quad 0 \leq s \leq t$$

is a continuous-time random walk run at rate two and absorbed at site zero. In particular, using the duality relationship (16.1) and the fact that symmetric random walks in one and two dimensions are recurrent, we deduce that

$$\begin{aligned} P(\xi_t(x) \neq \xi_t(y)) &\leq P(\hat{\xi}_t(x,t) \neq \hat{\xi}_t(y,t)) \\ &= P(X_t \neq 0) = P(X_s \neq 0 \text{ for all } s < t) \to 0 \end{aligned}$$

as time $t \to \infty$. This completes the proof. □

In contrast, the process converges to a stationary distribution in which both types are present in higher dimensions: both types coexist at equilibrium.

Theorem 16.2. *Let $d > 2$. Then, the voter model converges in distribution to an invariant measure in which there is a positive density of both types.*

Proof. To prove convergence to a stationary distribution, we first observe that there is no type 1 individual in the set A at time t if and only if there is no type 1 individual in the corresponding dual process at dual time $s = t$, therefore

$$P(\xi_t \cap A = \varnothing) = E\left((1-\rho)^{|\hat{\xi}_t(A,t)|}\right) \tag{16.2}$$

where the process is again identified to the set of type 1 individuals. The dominated convergence theorem implies that both terms in (16.2) have a limit when time goes to infinity, which proves the existence of a stationary distribution.

Using again duality (16.1) and the fact that symmetric random walks on the d-dimensional integer lattice are transient when $d > 2$, we obtain that

$$P(\xi_t(x) \neq \xi_t(y)) = P(\xi_0(\hat{\xi}_t(x,t)) \neq \xi_0(\hat{\xi}_t(y,t))) = 2\rho\,(1-\rho)\,P(X_t \neq 0)$$

converges to a positive limit as time goes to infinity. This implies that both types coexist under the stationary distribution. The factor $2\rho\,(1-\rho)$ is due to the fact that, for two individuals to have different types, not only they must originate from two different ancestors but also the ancestors must have different types. □

To conclude, we give an overview with additional key results.

16.3 Overview of the voter model

The duality between the voter model and coalescing random walks is again the key to answering additional questions about the process.

To begin with, we observe that the density of each type, defined as the probability that a given vertex is of that type, is preserved by the voter model dynamics. The intuition behind this result is simply that, each time two neighbors interact, each neighbor is equally likely to be the one producing an offspring. This, together with the two results proved above, gives the following **complete convergence theorem** which is attributed to Holley and Liggett [45].

Theorem 16.3. *Starting with a density ρ of type 1 individuals,*

- *Consensus*: $\xi_t \xrightarrow{d} (1-\rho)\,\delta_0 + \rho\,\delta_1$ *when $d \le 2$, where δ_i is the measure that concentrates on the configuration in which all sites are of type i.*
- *Coexistence*: $\xi_t \xrightarrow{d} \xi_\infty$ *when $d > 2$, where $P(\xi_\infty(x) = 1) = \rho$.*

The first statement says that, any bounded region is, with probability close to one at large times, either occupied by only type 1 individuals with probability ρ or occupied by only type 0 individuals with probability $1-\rho$. This indicates that the process clusters. The second statement says that, in higher dimensions, both types coexist at equilibrium.

The next natural step is to determine how fast the clusters grow in low dimensions. To answer this question, the idea is to look more carefully at the dual process and to estimate how long it takes for two random walks to coalesce, which depends on their initial position. In one and two dimensions, two fixed vertices $x \neq y$ are eventually totally correlated, so to understand the size of the clusters, the idea is to look instead at the correlations between two evolving vertices with a distance that increases with time, say $t^\alpha x$ and $t^\alpha y$ which we identify with their nearest neighbor on the integer lattice. Bramson and Griffeath [11] proved that the size of the clusters scales asymptotically like the square root of time in one dimension.

Theorem 16.4 (Cluster size in $d = 1$). *As $t \to \infty$,*

- *Independence*:

$$\lim_{t\to\infty} P(\xi_t(t^\alpha x) \neq \xi_t(t^\alpha y)) = 2\rho\,(1-\rho) \;\; \textit{when} \;\; \alpha > 1/2.$$

- *Total correlations*:

$$\lim_{t\to\infty} P(\xi_t(t^\alpha x) \neq \xi_t(t^\alpha y)) = 0 \;\; \textit{when} \;\; \alpha < 1/2.$$

The behavior in two dimensions is more interesting: Though sites are again independent when $\alpha > 1/2$, Cox and Griffeath [20] proved the following result.

Theorem 16.5 (Cluster size in $d = 2$). *As $t \to \infty$,*

- *Independence*:

$$\lim_{t\to\infty} P(\xi_t(t^\alpha x) \neq \xi_t(t^\alpha y)) = 2\rho\,(1-\rho) \;\; \textit{when} \;\; \alpha > 1/2.$$

- *Some correlations*:

$$\lim_{t\to\infty} P(\xi_t(t^\alpha x) \neq \xi_t(t^\alpha y)) = 4\rho\,(1-\rho)\,\alpha \;\; \textit{when} \;\; \alpha < 1/2.$$

This indicates in particular that, in contrast with the one-dimensional case, there is no natural scale for the cluster size in two dimensions.

In higher dimensions, because coexistence occurs, two fixed vertices are never totally correlated, even at equilibrium. However, due to the presence of local interactions, sites are not independent either and a natural question is: How strong are the spatial correlations? To quantify the spatial correlations, the basic idea is to look at how much the number of vertices of type 1 in a large box deviates from its mean. Since initially sites are independently of type 1 with probability ρ, the central limit theorem implies that, as $n \to \infty$,

$$\left(\frac{1}{\sqrt{n}}\right)^{d} \sum_{\|x\|<n} (\xi_0(x) - \rho) \xrightarrow{d} \text{Normal}\,(0, \sigma^2).$$

The next result shows how spatial correlations build up.

Theorem 16.6 (Spatial correlations in $d > 2$). *As $n \to \infty$,*

$$\left(\frac{1}{\sqrt{n}}\right)^{d+2} \sum_{\|x\|<n} (\xi_\infty(x) - \rho) \xrightarrow{d} \text{Normal}\,(0, \sigma^2).$$

This was proved by Bramson and Griffeath [10] in three dimensions and extended to higher dimensions by Zähle [98]. In particular, to have a central limit type theorem, one has to further divide the sum by an extra n, showing that, even in high dimensions where both types coexist, spatial correlations due to local interactions are still quite strong.

Finally, we look at the fraction of time a given site x is of type 1 in the long run, a random variable called the **occupation time**. When coexistence occurs, sites change their type infinitely often so we expect the following law of large numbers: The fraction of time site x is of type 1 converges almost surely to ρ. In contrast, when clustering occurs, it is expected that this result does not hold. In fact, Cox and Griffeath [19] proved the following surprising result.

Theorem 16.7 (Occupation time). *Assume that $d \geq 2$. Then,*

$$\frac{1}{t}\int_0^t \xi_s(x)\,ds \xrightarrow{a.s.} \rho \quad \text{as} \quad t \to \infty.$$

This law of large numbers does not hold in one dimension. However, it is known that the system of annihilating random walks that describes the evolution of the boundaries of the voter model in one dimension is site recurrent, which implies that, even in one dimension, the type of any given site keeps changing indefinitely.

From the combination of the previous theorems, we obtain the following description of the voter model in one and two dimensions: Clusters form and appear to grow indefinitely so only one type can be present at equilibrium. However, any given site changes its type infinitely often, which indicates that clusters are not fixed in space but move around, and thus may give the impression of local transience though, strictly speaking, coexistence does not occur.

16.4 Exercises

Exercise 16.1. Consider the voter model (ξ_t) on the integer lattice $\mathbb{Z}^d$ starting with a finite number of individuals with opinion 1.

1. Prove that the process (X_t) that keeps track of the number of 1s is a martingale with respect to the natural filtration $(\mathscr{F}_t)$ of the voter model:

$$\lim_{s \downarrow 0} E(X_{t+s} \mid \mathscr{F}_t) = \lim_{s \downarrow 0} E(X_{t+s} \mid \xi_t) = X_t.$$

2. Deduce that opinion 0 outcompetes opinion 1, i.e.,

$$\lim_{t \to \infty} P(\xi_t(x) = 0) = 1 \quad \text{for all} \quad x \in \mathbb{Z}^d.$$

Hint: For the first question, express the conditional transition rates using the number of edges that connect neighbors with different opinions.

Exercise 16.2 (Cluster size). Consider the voter model on a large one-dimensional torus starting from the configuration in which individuals hold independently each of the two possible opinions with equal probability.

1. Modify Program 13 to simulate this voter model and compute the average size of the clusters at various times.

The following table gives values of the average cluster size at various times from time zero to time 1000 obtained from one realization of the voter model on the one-dimensional torus with 600 vertices:

```
  0.    2.027      1.    3.704      2.    4.762
  3.    5.769      4.    6.522      5.    7.317
  6.    7.895      7.    8.108      8.    8.571
  9.    8.824     16.   16.667     25.   20.000
 36.   23.077     49.   27.273     64.   33.333
 81.   33.333    100.   33.333    200.   60.000
300.  100.000    500.  150.000   1000.  150.000
```

2. Use the law of large numbers to explain why the average cluster size at time zero in this simulation is close to two.
3. Prove that, for each realization, the average cluster size can only increase.

Exercise 16.3 (Clustering in $d = 2$). Consider the voter model on a large two-dimensional torus starting from the configuration in which individuals hold independently each of the two possible opinions with equal probability.

1. Modify Program 13 to simulate this process and compute the fraction of edges that connect neighbors who disagree at various times. From now on, we call those edges dissonant edges.

The following table gives values of the fraction of dissonant edges at various times from time zero to time 1000 obtained from one realization of the voter model on the two-dimensional torus with 300 × 300 vertices:

```
   0. 0.50016     1. 0.38045      2. 0.33826
   3. 0.31492     4. 0.30205      5. 0.29125
   6. 0.28238     7. 0.27520      8. 0.27107
   9. 0.26585    10. 0.26171     25. 0.22976
  36. 0.21958    49. 0.20902     64. 0.20500
  81. 0.19982   100. 0.18953    200. 0.17937
 300. 0.16948   500. 0.16672   1000. 0.15852
```

2. Explain why the initial fraction of dissonant edges is close to one-half.
3. In contrast with the one-dimensional model, the fraction of dissonant edges fluctuates and may therefore increase. Explain why this is possible.

Exercise 16.4 (Biased voter model). In the biased voter model (ξ_t), like in the classical voter model, individuals give birth to offspring of their own type at rate one and the offspring replaces one of the $2d$ neighbors of the parent's site. However, individuals of type 0 give birth at rate one whereas individuals of type 1 give birth at rate $\mu > 1$. We consider the process starting with a single 1 at the origin.

1. Write the transition rates of the process.
2. Prove that, if X_t denotes the number of 1s at time t, then (μ^{-X_t}) is a martingale with respect to the natural filtration $(\mathscr{F}_t)$ of the biased voter model:

$$\lim_{s\downarrow 0} E(\mu^{-X_{t+s}} \,|\, \mathscr{F}_t) = \mu^{-X_t}.$$

3. Deduce the survival probability $P(\xi_t \not\equiv 0 \text{ for all } t)$.

Hint: For the second question, express the conditional transition rates using the number of edges that connect neighbors of different types.

Exercise 16.5 (Threshold voter model). In this process, individuals look at all their neighbors at rate one at which time they change their opinion if and only if they disagree with at least τ of their neighbors. The integer τ is called the threshold.

1. Write the transition rates of the process on $\mathbb{Z}^d$.
2. Modify Program 13 to simulate the process in two dimensions starting from the configuration in which individuals hold independently each of the opinions with equal probability and study the effect of the threshold on the behavior.

Exercise 16.6 (Axelrod model). This problem is inspired from [61]. In the two-feature two-state Axelrod model, each site of the d-dimensional integer lattice is occupied by an individual who is characterized by her opinion about two different cultural features, so the state at time t is a spatial configuration

$$\xi_t : \mathbb{Z}^d \to \{0,1\}^2.$$

Pairs of nearest neighbors interact at a rate equal to the number of cultural features they share, which results in a perfect agreement between the two neighbors.

1. Write the transition rates of the process.
2. Show that the process with two possible states per site obtained by identifying the opposite cultures is the voter model.
3. Use the fact that the voter model fluctuates, i.e., all the individuals change their opinion infinitely often, to deduce that the Axelrod model fluctuates as well.

Chapter 17
Numerical simulations in C and Matlab

This chapter is devoted to numerical simulations, focusing on the ten most popular stochastic models presented in this textbook. Following the chronology of the textbook, we start with the three classical models: the gambler's ruin chain, the branching process, and the simple birth and death process. Then, we look at the spatially implicit stochastic models and finally at the spatially explicit ones: percolation models and interacting particle systems.

As the title of the chapter indicates, the programs are in C and Matlab. The latter offers an easy-to-use graphical interface, which is convenient to plot sample paths of simple models, whereas the former is more efficient to deal with a large number of calculations. In particular, for the simple classical processes, we use Matlab to generate and draw sample paths. Depending on the model under consideration, quantities of interest are the probability of survival, the probability that a given type wins, and the time to fixation. Invoking the law of large numbers, these quantities can typically be estimated by looking at the average over a large number of independent sample paths. Both C and Matlab are appropriate to study this aspect for simple stochastic processes, whereas to simulate the more complex spatially explicit models, which requires many calculations, we only use C. We point out that our choice of using Matlab rather than similar programming languages offering an easy-to-use graphical interface such as Python, R, or Julia, is completely arbitrary and that the readers more familiar with any such languages should be able to translate our Matlab codes fairly easily.

Along the way, we also explain how to simulate general discrete and continuous random variables from the uniform random variable on the unit interval, which is the basic random variable that any programming language can simulate using a pseudo-random number generator.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_17

17.1 Classical models

This first section gives simulations of the gambler's ruin chain, the branching process, and the simple birth and death process in Matlab, which is used in particular to plot sample paths of the different processes.

Gambler's ruin chain — To simulate the gambler's ruin chain, one only needs to create a sequence of independent Bernoulli random variables to determine whether the gambler wins or loses at each step of the game. To generate this random variable from the uniform random variable U on the unit interval $(0,1)$, we use the same approach as in the proof of Theorem 13.1. Letting p be the success probability, or probability of winning, we have

$$P(U < p) = p \quad \text{so} \quad \mathbf{1}\{U < p\} \sim \text{Bernoulli}\,(p).$$

The process until the gambler either gets ruined or reaches her target can be simulated using this basic property and a simple `while` loop. The following Matlab program generates 1001 such sample paths, plots the first sample path and returns the number of times the gambler quits winner and the average number of games for the other 1000 sample paths. It also allows the user to choose both the probability p of winning a single game and the initial gambler's fortune. The reader can check that the outcomes give good approximations of (5.4).

Program 1 Gambler's ruin chain in Matlab.

```
clear all
N = 100;
success = 0;
time = 0;
prompt = 'winning probability = ';
p = input(prompt);
prompt = 'initial capital = ';
x = input(prompt);
for run = 1:1001
    clear money t
    n = 1;
    t(1) = 0;
    money(1) = x;
    while money(n) < N & money(n) > 0
        t(n+1) = t(n)+1;
        u = rand;
        if u < p
            money(n+1) = money(n)+1;
        else
            money(n+1) = money(n)-1;
        end
```

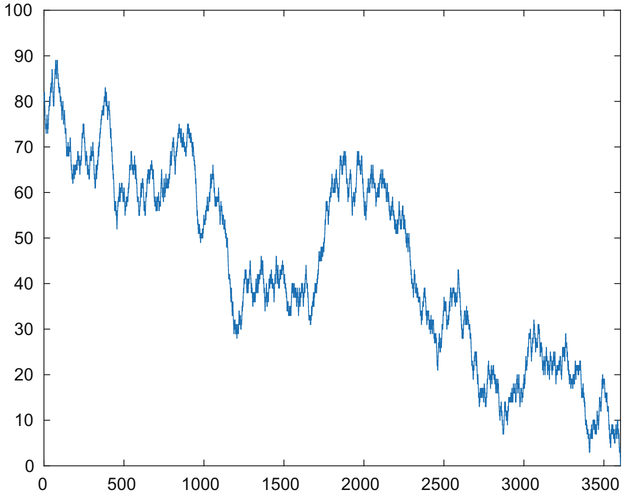


Fig. 17.1 Sample path of the gambler's ruin chain with winning probability $p = 0.49$ and initial fortune 80 dollars.

```
            n = n+1;
        end
        if run==1
            stairs (t, money)
            s = t(n);
            axis ([0 s 0 N])
            xlabel ('time');
            ylabel ('gambler fortune');
        else
            time = time+t(n);
            if money(n) > 0
                success = success+1;
            end
        end
    end
    success
    time = time/1000;
    time
```

Branching processes — We now turn our attention to the branching process. The process is generated by using the recursive formula (6.1) which involves the offspring distribution Y with probability mass function

$$P(Y = k) = p_k \quad \text{for all} \quad k \in \mathbb{N}.$$

Using the same trick as for the Bernoulli random variable, we can construct this random variable from the uniform random variable U by observing that

$$Y = \textstyle\sum_k k\,\mathbf{1}\{p_0 + \cdots + p_{k-1} < U < p_0 + \cdots + p_k\}$$

in distribution. As for the gambler's ruin chain, the Matlab program below generates 1001 sample paths for the process stopped when the population either goes extinct or exceeds 1000 individuals. The program plots the first sample path, which starts with five individuals, and returns the number of times the population exceeds 1000 before it goes extinct when starting with a single individual for the remaining sample paths.

Program 2 Branching process in Matlab.

```
clear all
N = 1000;
p0 = 1/6;
p1 = 1/3;
p2 = 1/2;
success = 0;
for run = 1:1001
    clear size t
    n = 1;
    t(1) = 0;
    if run==1
        size(1) = 5;
    else
        size(1) = 1;
    end
    while size(n) < N & size(n) > 0
        t(n+1) = t(n)+1;
        size(n+1) = 0;
        for i = 1:size(n)
            u = rand;
            if u > p0
                size(n+1) = size(n+1)+1;
            end
            if u > p0+p1
                size(n+1) = size(n+1)+1;
            end
        end
        n = n+1;
    end
    if run==1
        stairs (t, size)
        s = t(n);
```

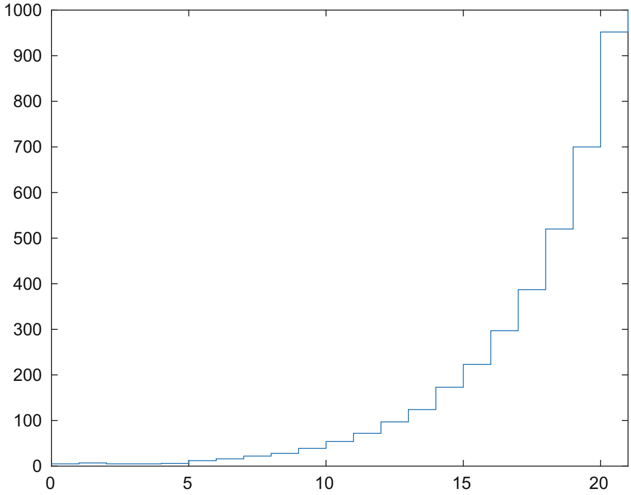


Fig. 17.2 Sample path of the branching process with offspring distribution as in Exercise 6.1.

```
            axis ([0 s 0 1000])
            xlabel ('time');
            ylabel ('population size');
        elseif run > 1 & size(n) > 0
            success = success+1;
        end
    end
    success
```

Simple birth and death process — For the simple birth and death process, each update corresponds to either a single birth or a single death with probabilities that do not depend on the population size so the process can be generated using Bernoulli random variables with the same success probability, as with the gambler's ruin chain. However, to keep track of the time, which is now continuous, one also needs to generate exponential random variables from the uniform distribution U. To do this, observe that, having a general random variable T with a continuous increasing distribution function F,

$$P(F^{-1}(U) \leq t) = P(U \leq F(t)) = F(t) = P(T \leq t) \quad \text{for all} \quad t \in \mathbb{R}$$

showing that $T = F^{-1}(U)$ in distribution. Applying this to the particular case of the exponential random variable with parameter μ, we obtain that

$$-(1/\mu)\ln(U) \sim \text{Exponential}(\mu).$$

The following Matlab program uses this property to generate the process with birth rate chosen by the user. Again, the program plots the first sample path starting with five individuals and returns the number of times the process starting with a single individual reaches 1000 individuals before going extinct for the other 1000 sample paths. The reader can check that this gives a good approximation of the probability of survival computed in (10.23).

Program 3 Simple birth and death process in Matlab.

```
clear all
N = 1000;
success = 0;
prompt = ’birth rate = ’;
beta = input(prompt);
for run = 1:1001
    clear size t
    n = 1;
    t(1) = 0;
    if run==1
        size(1) = 5;
    else
        size(1) = 1;
    end
    while size(n) < N & size(n) > 0
        s = beta+1;
        t(n+1) = t(n)-log(rand)/(s*size(n));
        u = rand;
        u = u*s;
        if u < 1
            size(n+1) = size(n)-1;
        else
            size(n+1) = size(n)+1;
        end
        n = n+1;
    end
    if run==1
        stairs (t, size)
        s = t(n);
        axis ([0 s 0 1000])
        xlabel (’time’);
        ylabel (’population size’);
    elseif run > 1 & size(n) > 0
        success = success+1;
    end
end
success
```

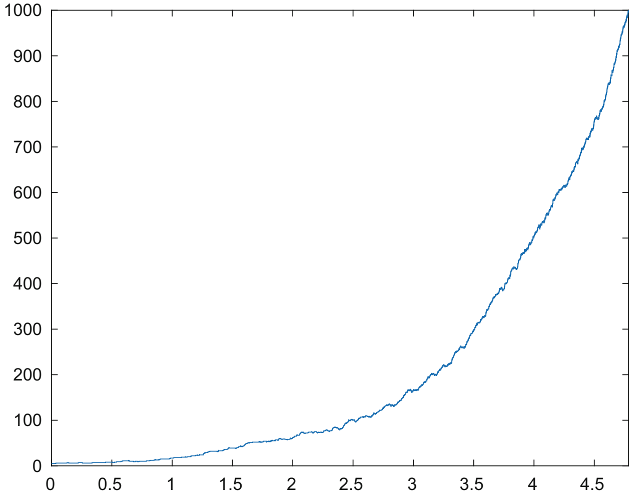


Fig. 17.3 Sample path of the simple birth and death process with birth rate $\beta = 2$.

Modifying the previous program, one can easily simulate other continuous-time birth and death processes. The following Matlab program shows, for example, how to deal with the $M/M/s$ queue where all the servers work at rate one and where both the rate at which customers enter the system and the number of servers are chosen by the user. The program displays only one sample path.

Program 4 $M/M/s$ queue in Matlab.

```
clear all
T = 100;
prompt = 'rate of arrival = ';
beta = input(prompt);
prompt = 'number of servers = ';
server = input(prompt);
n = 1;
t(1) = 0;
size(1) = 0;
while t(n) < T
    s = beta+min(server, size(n));
    t(n+1) = t(n)-log(rand)/s;
    u = rand;
    u = u*s;
    if u < beta
        size(n+1) = size(n)+1;
    else
        size(n+1) = size(n)-1;
    end
```

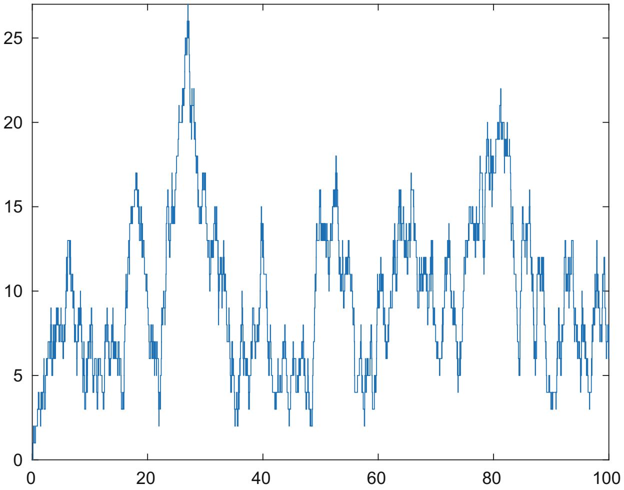


Fig. 17.4 Sample path of the $M/M/s$ queue with $s = 8$ servers working independently at rate one where customers enter the system at rate 7.

```
        n = n+1;
    end
    stairs (t, size)
    s = t(n);
    axis ([0 T 0 inf])
    xlabel ('time');
    ylabel ('number of customers');
```

17.2 Spatially implicit models

In this second section, we look at the three stochastic models where individuals are located on a finite set of sites and interact globally: the logistic growth process, the Moran model, and the Wright–Fisher model.

Logistic growth process — Our first approach to simulate the logistic growth process is to use the transition rates of the process. This is done in the following Matlab program which is very similar to the one generating the simple birth and death process. The code runs the process until time 20, displays the first sample path, and returns the number of times the population survives at least until time 20 for the other 1000 sample paths.

Program 5 Logistic growth process in Matlab.

```
clear all
N = 100;
```

```
T = 20;
success = 0;
prompt = 'birth rate = ';
beta = input(prompt);
for run = 1:1001
    clear size t
    n = 1;
    t(1) = 0;
    if run==1
        size(1) = 5;
    else
        size(1) = 1;
    end
    while t(n) < T & size(n) > 0
        s = beta*(1-size(n)/N)+1;
        t(n+1) = t(n)-log(rand)/(s*size(n));
        u = rand;
        u = u*s;
        if u < 1
            size(n+1) = size(n)-1;
        else
            size(n+1) = size(n)+1;
        end
        n = n+1;
    end
    if run==1
        stairs (t, size)
        axis ([0 T 0 N])
        xlabel ('time');
        ylabel ('population size');
    elseif run > 1 & size(n) > 0
        success = success+1;
    end
end
success
```

Our second approach is to keep track of the population vector that indicates the state of each of the spatial locations, empty or occupied, rather than the number of individuals. In other words, instead of using the transition rates, we choose a site at random and, if this site is occupied, either kill the individual at that site or place an offspring to another randomly chosen site. This approach is more demanding from a computational point of view so we use C language, which is much more efficient than Matlab in this context. The first advantage of this approach is that the algorithm is somewhat more transparent because it uses the original interpretation of the model as a spatially implicit model. Moreover, the code only needs to be slightly modified so that it generates the contact process, which is the spatial analog of the logistic growth process. Except for the way the dynamics is simulated, the program does the same as the previous Matlab code: it generates 1000 sample paths of the process, checks the number of individuals after ten units of time, and returns the number of times the population survives up to that time.

Program 6 Logistic growth process in C.

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>
#include <math.h>
#define T 10
#define N 200
int pop[N];
int i, j, size, runs, success;
float beta, s, t, u, mu;
// Take a number uniformly at random in (0,1)
float uniform() {
  u = (float) rand()/RAND_MAX;
  return u;
}
// Exponential random variable with parameter mu
float exponential(float u, float mu) {
  s = (float)-log(u)/mu;
  return s;
}
// Logistic growth process
int main (int argc, char*argv[]) {
  printf ("birth rate = ");
  scanf ("%f", &beta);
  time_t amorce;
  time (&amorce);
  srand (amorce);
  success = 0;
  for (runs = 0; runs < 1000; runs ++) {
    t = 0;
    size = 1;
    pop[0] = 1;
    for (i = 1; i < N; i ++) {
      pop[i] = 0;
    }
    while (t < T && size != 0) {
      u = uniform();
      t = t+exponential(u, (beta+1)*N);
      u = uniform();
      i = floor(N*u);
      i = i%N;
      if (pop[i]==1) {
        u = uniform();
        u = u*(beta+1);
        if (u < 1) {
          pop[i] = 0;
          size --;
        }
        else {
          u = uniform();
          j = floor(N*u);
          j = j%N;
          if (pop[j]==0) {
            pop[j] = 1;
            size ++;
          }
        }
      }
    }
    if (size != 0) {
      success ++;
    }
  }
  printf ("number of successes = %i\n", success);
}
```

Moran model — Following the same strategy as for the logistic growth process, we give two different algorithms to simulate the Moran model. The first one, written in Matlab, uses the transition probabilities of the process. The user can choose the initial number of type 1 individuals and the program displays one sample path as well as the number of times type 1 wins and the average time to fixation over 1000 additional independent realizations.

Program 7 Moran model in Matlab.

```
clear all
N = 100;
success = 0;
time = 0;
prompt = 'initial number of type 1 = ';
x = input(prompt);
for run = 1:1001
    clear size t
    n = 1;
    t(1) = 0;
```

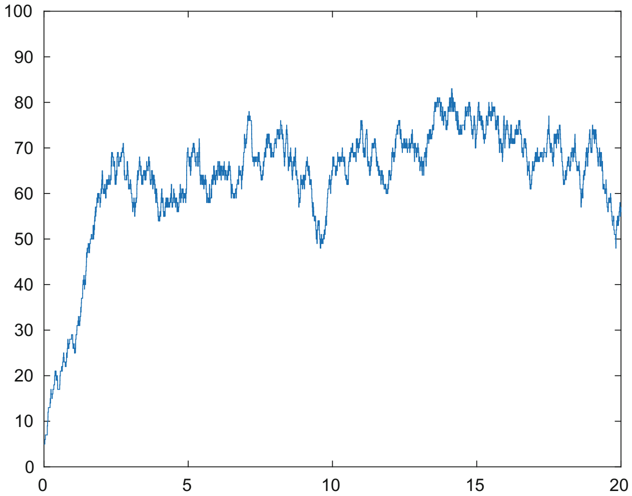


Fig. 17.5 Sample path of the logistic growth process with birth rate $\beta = 3$.

```
size(1) = x;
while size(n) < N
    t(n+1) = t(n)+1;
    s = (size(n)/N)*(1-size(n)/N);
    u = rand;
    if u < s
        size(n+1) = size(n)+1;
    elseif u > s & u < 2*s
        size(n+1) = size(n)-1;
    else
        size(n+1) = size(n);
    end
    n = n+1;
end
if run==1
    stairs (t, size)
    s = t(n);
    axis ([0 s 0 N])
    xlabel ('time');
    ylabel ('number of type 1');
else
    time = time+t(n);
    if size(n) > 0
        success = success+1;
    end
end
```

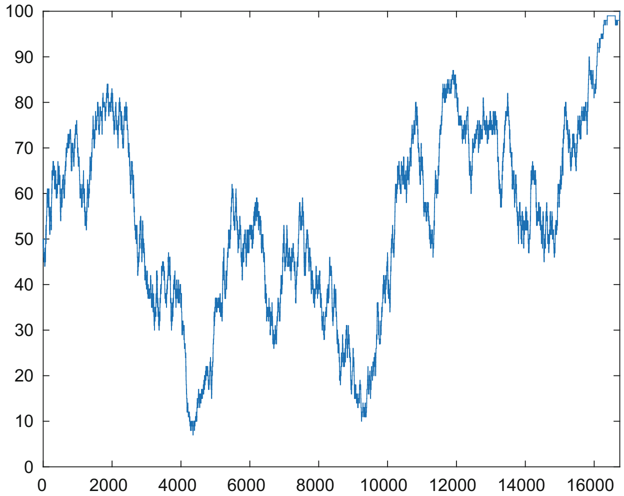


Fig. 17.6 Sample path of the Moran model starting with an equal number of type 0 individuals and type 1 individuals.

```
end
success
time = time/1000;
time
```

The second algorithm is written in C and returns to the interpretation of the model as a spatially implicit model by keeping track of the type of the individual at each spatial location. This algorithm is again important because a minor modification of the program simulates the more complicated voter model. Except for the algorithm, the program is similar to the previous Matlab code and again displays the number of times type 1 wins as well as the average time to fixation.

Program 8 Moran model in C.

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>
#include <math.h>
#define N 100
int pop[N];
int i, j, x, size, n, runs, success, T;
float u;
// Take a number uniformly at random in (0,1)
float uniform() {
  u = (float) rand()/RAND_MAX;
  return u;
}
// Moran model
int main (int argc, char*argv[]) {
  printf ("initial number of type 1 = ");
  scanf ("%i", &x);
  time_t amorce;
  time (&amorce);
  srand (amorce);
  success = 0;
  for (runs = 0; runs < 1000; runs ++) {
    n = 0;
    size = x;
    for (i = 0; i < x; i ++) {
      pop[i] = 1;
    }
    for (i = x; i < N; i ++) {
      pop[i] = 0;
    }
    while (size != 0 && size != N) {
      n ++;
      u = uniform();
      i = floor(N*u);
      i = i%N;
      u = uniform();
      j = floor(N*u);
      j = j%N;
      if (pop[i] > pop[j]) {
        size ++;
      }
      else if (pop[i] < pop[j]) {
        size --;
      }
      pop[j] = pop[i];
    }
    T = T+n;
    if (size != 0) {
      success ++;
    }
  }
  T = T/1000;
  printf ("sucesses for 1 = %3i\n", success);
  printf ("average time to fixation = %4i\n", T);
}
```

Wright–Fisher model — Recall that, except for the two absorbing states, the Wright–Fisher model can jump from any state to any other state, so using the transition probabilities to simulate the process is not convenient in this case. In particular, we only give the algorithm that keeps track of the type of the individual at each spatial location. The program again allows the user to choose the initial number of type 1 individuals, generates 1000 sample paths, and returns the number of times type 1 wins and the average time to fixation. The code is almost the same as for the Moran model except that, becasuse all the individuals are simultaneously updated based on the previous configuration, one needs to remember the previous configuration while constructing the current configuration.

Program 9 Wright–Fisher model in C.

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>
#include <math.h>
#define N 100
int pop[2][N];
int i, j, x, size, n, runs, success, T;
float u;
// Take a number uniformly at random in (0,1)
float uniform() {
  u = (float) rand()/RAND_MAX;
  return u;
}
// Wright--Fisher model
int main (int argc, char*argv[]) {
  printf ("initial number of type 1 = ");
  scanf ("%i", &x);
  time_t amorce;
  time (&amorce);
  srand (amorce);
  success = 0;
    for (runs = 0; runs < 1000; runs ++) {
    n = 0;
    size = x;
    for (i = 0; i < x; i ++) {
      pop[1][i] = 1;
    }
    for (i = x; i < N; i ++) {
      pop[1][i] = 0;
    }
    while (size != 0 && size != N) {
      n ++;
      size = 0;
      for (i = 0; i < N; i ++) {
        pop[0][i] = pop[1][i];
      }
      for (i = 0; i < N; i ++) {
        u = uniform();
        j = floor(N*u);
        j = j%N;
        pop[1][i] = pop[0][j];
        size = size+pop[1][i];
      }
    }
    T = T+n;
    if (size != 0) {
      success ++;
    }
  }
  T = T/1000;
  printf ("successes for 1 = %3i\n", success);
  printf ("average time to fixation = %4i\n", T);
}
```

17.3 Percolation models

For spatially explicit stochastic processes such as percolation models, the main objective is obviously to display a spatial configuration therefore the graphical interface seems to be essential in this case. Matlab, however, turns out to be slow not only to generate but also to display these spatial configurations. In particular, the rest of this chapter focuses on C programs that simulate almost instantaneously even large systems. To keep things as simple as possible, instead of using a sophisticated graphical interface, all the programs just display an array of characters indicating the state of the edges or vertices.

Bond percolation — Though percolation models are difficult to study analytically, since there is no time evolution, they are easy to simulate. Indeed, to generate a realization of bond or site percolation, we only need to create a collection of Bernoulli random variables to determine whether the edges or vertices are open or closed. The following C program allows the user to choose the density of open edges and generates a two-dimensional array with the characters | and _ indicating the orientation and the position of the open edges for bond percolation.

Program 10 Bond percolation in C.

```
#include <stdlib.h>
#include <time.h>
```

```
#include <math.h>
#include <stdio.h>
#define N 50
int x, y;
float p, u;
// Take a number uniformly at random in (0,1)
float uniform() {
  u = (float) rand()/RAND_MAX;
  return u;
}
// Bond percolation process
int main (int argc, char*argv[]) {
  printf ("p = ");
  scanf ("%f", &p);
  time_t amorce;
  time (&amorce);
  srand (amorce);
  printf (" ");
  for (x = 0; x < N; x ++) {
    u = uniform();
    if (u < p) {
      printf ("_ ");
    }
    else {
      printf ("  ");
    }
  }
  printf ("\n");
  for (y = 0; y < N; y ++) {
    for (x = 0; x < N; x ++) {
      u = uniform();
      if (u < p) {
        printf ("|");
      }
      else {
        printf (" ");
      }
      u = uniform();
      if (u < p) {
        printf ("_");
      }
      else {
        printf (" ");
      }
    }
    u = uniform();
    if (u < p) {
      printf ("|");
    }
    else {
      printf (" ");
    }
    printf ("\n");
  }
  return 0;
}
```

Oriented site percolation — To simulate oriented site percolation, we again generate a collection of independent Bernoulli random variables for which success probability is chosen by the user to determine whether sites are open or closed. Because the graph is now directed, after the configuration of open and closed sites has been determined, we also identify the wet sites recursively starting from level zero. The following program again displays the spatial configuration in the form of an array of characters where open sites are circles o when there are wet and dots . when they are not wet, while we leave a blank for the closed sites.

Program 11 Oriented site percolation in C.

```
#include <stdlib.h>
#include <time.h>
#include <math.h>
#include <stdio.h>
#define N 50
#define M 100
int site[M+1][N+1];
int wet[M+1][N+1];
int x, y;
float p, u;
// Take a number uniformly at random in (0,1)
float uniform() {
  u = (float) rand()/RAND_MAX;
  return u;
}
// Oriented site percolation process
int main (int argc, char*argv[]) {
  printf ("p = ");
  scanf ("%f", &p);
  time_t amorce;
  time (&amorce);
  srand (amorce);
  for (x = 0; x < M+1; x ++) {
    if (x%2==0) {
      site[x][0] = 1;
      wet[x][0] = 1;
    }
  }
  for (y = 1; y < N+1; y ++) {
    for (x = 0; x < M; x ++) {
      if ((x+y)%2==0) {
        u = uniform();
        if (u < p) {
          site[x][y] = 1;
        }
      }
    }
    site[M][y] = site[0][y];
  }
  for (x = 0; x < M+1; x ++) {
    for (y = 1; y < N+1; y ++) {
      wet[x][y] = 0;
    }
  }
  for (y = 1; y < N+1; y ++) {
    for (x = 0; x < M+1; x ++) {
      if (site[x][y]==1 &&
           wet[(x+M-1)%M][y-1]==1) {
        wet[x][y] = 1;
      }
      else if (site[x][y]==1 &&
                wet[(x+M+1)%M][y-1]==1) {
        wet[x][y] = 1;
      }
    }
  }
  for (y = N-1; y >= 0; y --) {
    for (x = 0; x < M; x ++) {
      if (wet[x][y]==1) {
        printf ("o");
      }
      else if (wet[x][y]==0 && site[x][y]==1) {
        printf (".");
      }
      else {
        printf (" ");
      }
    }
    printf ("\n");
  }
  return 0;
}
```

17.4 Interacting particle systems

Finally, we look at the more complicated models: the contact process and the voter model. For these two processes and interacting particle systems in general, sample paths cannot be represented as simple curves like the ones in Figures 17.1–17.6 because the states are not numbers but spatial configurations. The sample paths cannot either be represented by a single spatial configuration as is true for bond and oriented site percolation because there is now a time evolution. Combining explicit space and time evolution, the most natural way to represent sample paths is via movies that display the consecutive configurations. To create such movies and understand the emergence of spatial patterns, the author uses GTK+ libraries. But since the programs are quite long and it is not the purpose of this textbook to teach readers how to use a complex graphical interface, we simply run the process until a given time and then, following the same approach as for percolation models, display the spatial configuration at that time in the form of an array of characters, using `o` for sites in state 1 and leaving a blank for sites in state 0.

Contact process — Recall that the contact process can be seen as the spatial analog of the logistic growth process, which gives us some hint about how to simulate the process. Note, however, that because the spatial arrangement of the individuals now matters, the number of individuals in the contact process is not Markovian so the Matlab program for the logistic growth process above is not a good starting point to simulate the contact process. The C program, on the contrary, is a good starting point because it keeps track of the state of each of the spatial locations, empty or occupied, rather than the number of individuals. In fact, as explained in the previous chapters, to turn the program for the logistic growth process into a simulation for the contact process, we only need, at each birth, to send the offspring to one of the four nearest neighbors of the parent's location rather than a site chosen uniformly at random. The following program allows the user to choose the birth rate, runs the contact process starting from the all-occupied configuration until time 100 and then displays the configuration at that time which locally looks like the upper invariant measure of the contact process.

Program 12 Contact process in C.

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>
#include <math.h>
```

```
#define T 100
#define N 50
int pop[N*N];
int x, y, z;
float beta, s, t, u, mu;
// Take a number uniformly at random in (0,1)
float uniform() {
  u = (float) rand()/RAND_MAX;
  return u;
}
// Exponential random variable with parameter mu
float exponential (float u, float mu) {
  s = (float)-log(u)/mu;
  return s;
}
// Choose a neighbor of x on the 2D torus
int neighbor (int x, float u) {
  z = floor(4*u);
  switch (z) {
    case 0: y = x+1; break;
    case 1: y = x-1; break;
    case 2: y = x+N; break;
    case 3: y = x-N; break;
  }
  y = (y+N*N)%(N*N);
  return y;
}
// Contact process
int main (int argc, char*argv[]) {
  printf ("Birth rate = ");
  scanf ("%f", &beta);
  time_t amorce;
  time (&amorce);
  srand (amorce);
  t = 0;
  for (x = 0; x < N*N; x ++) {
    pop[x] = 1;
  }
  while (t < T) {
    u = uniform();
    t = t+exponential(u, (beta+1)*N*N);
    u = uniform();
    x = floor(N*N*u);
    x = x%(N*N);
    if (pop[x]==1) {
      u = uniform();
      u = u*(beta+1);
      if (u < 1) {
        pop[x] = 0;
      }
      else {
        u = uniform();
        y = neighbor (x, u);
        pop[y] = 1;
      }
    }
  }
// Display configuration: o are occupied sites
  for (y = 0; y < N; y ++) {
    for (x = 0; x < N; x ++) {
      if (pop[x+N*y]==1) {
        printf ("o ");
      }
      else {
        printf ("  ");
      }
    }
    printf ("\n");
  }
}
```

Voter model — The voter model can be simulated from the C program for the Moran model above in the same manner as the contact process is simulated from the logistic growth process, i.e., we only need to change the code so that, at each birth, the offspring is now sent to one of the four nearest neighbors of the parent's location. In contrast with the contact process, which converges in distribution to its upper invariant measure, the two-dimensional voter model clusters and how big the clusters are depends on the amount of time the process evolves. To explore this aspect via simulations, the following program allows the user to choose the time at which the process stops. The reader can try different times to check that, as time evolves, spatial correlations indeed build up and clusters indeed get larger.

Program 13 Voter model in C.

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>
#include <math.h>
#define N 50
int pop[N*N];
int T, x, y, z;
```

```
float s, t, u, mu;
// Take a number uniformly at random in (0,1)
float uniform() {
  u = (float) rand()/RAND_MAX;
  return u;
}
// Exponential random variable with parameter mu
float exponential(float u, float mu) {
  s = (float)-log(u)/mu;
  return s;
}
// Choose a neighbor of x on the 2D torus
int neighbor (int x, float u) {
  z = floor(4*u);
  switch (z) {
    case 0: y = x+1; break;
    case 1: y = x-1; break;
    case 2: y = x+N; break;
    case 3: y = x-N; break;
  }
  y = (y+N*N)%(N*N);
  return y;
}
// Voter model
int main (int argc, char*argv[]) {
  printf ("Stop the process at time = ");
  scanf ("%i", &T);
  time_t amorce;
  time (&amorce);
  srand (amorce);
  t = 0;
  for (x = 0; x < N*N; x ++) {
    u = uniform();
    if (u < 0.5) {
      pop[x] = 0;
    }
    else {
      pop[x] = 1;
    }
  }
  while (t < T) {
    u = uniform();
    t = t+exponential(u, N*N);
    u = uniform();
    x = floor(N*N*u);
```

```
   x = x%(N*N);
   u = uniform();
   y = neighbor (x, u);
   pop[x] = pop[y];
  }
 // Display configuration: o are type 1
  for (y = 0; y < N; y ++) {
   for (x = 0; x < N; x ++) {
    if (pop[x+N*y]==1) {
     printf ("o ");
    }
    else {
     printf ("  ");
    }
   }
   printf ("\n");
  }
 }
```

The diagram of Figure 17.7 gives an overview of the main stochastic models seen across this textbook, distinguishing in particular between invasion and competition models, whether the models evolve in time or not, and the level of details they take into account regarding the spatial component.

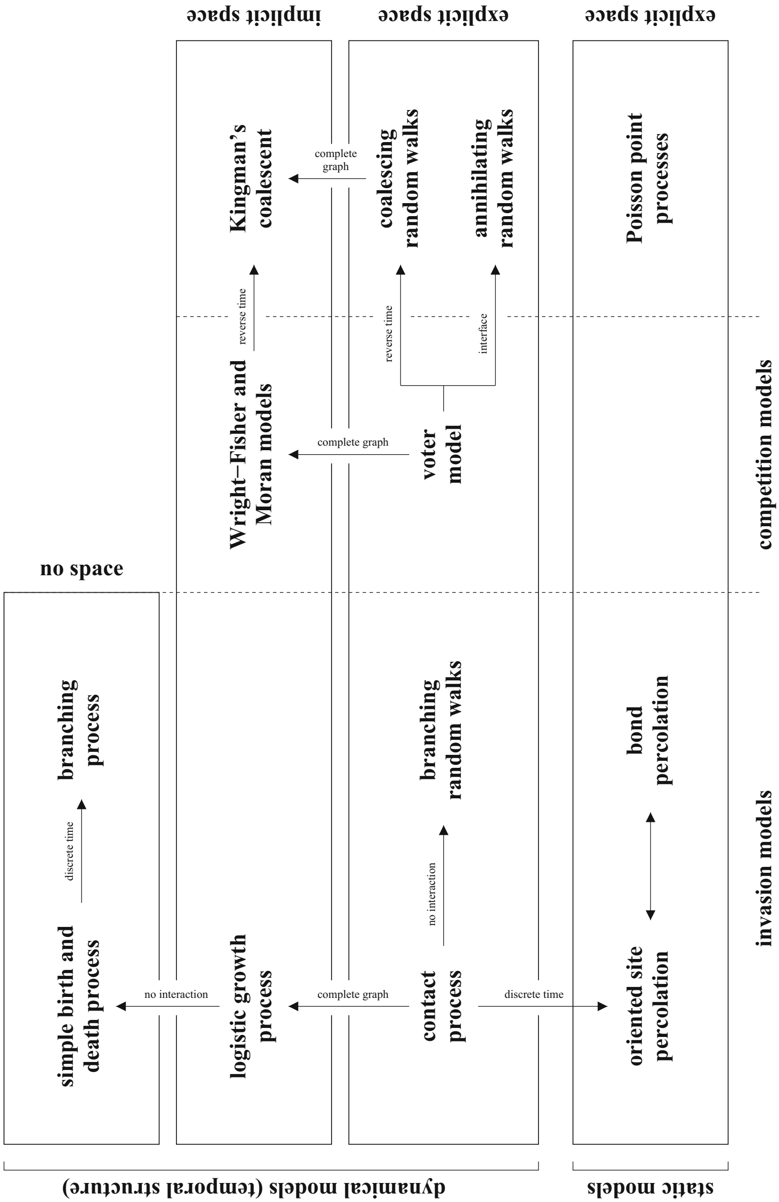


Fig. 17.7 Overview of the main stochastic models studied across the textbook.

References

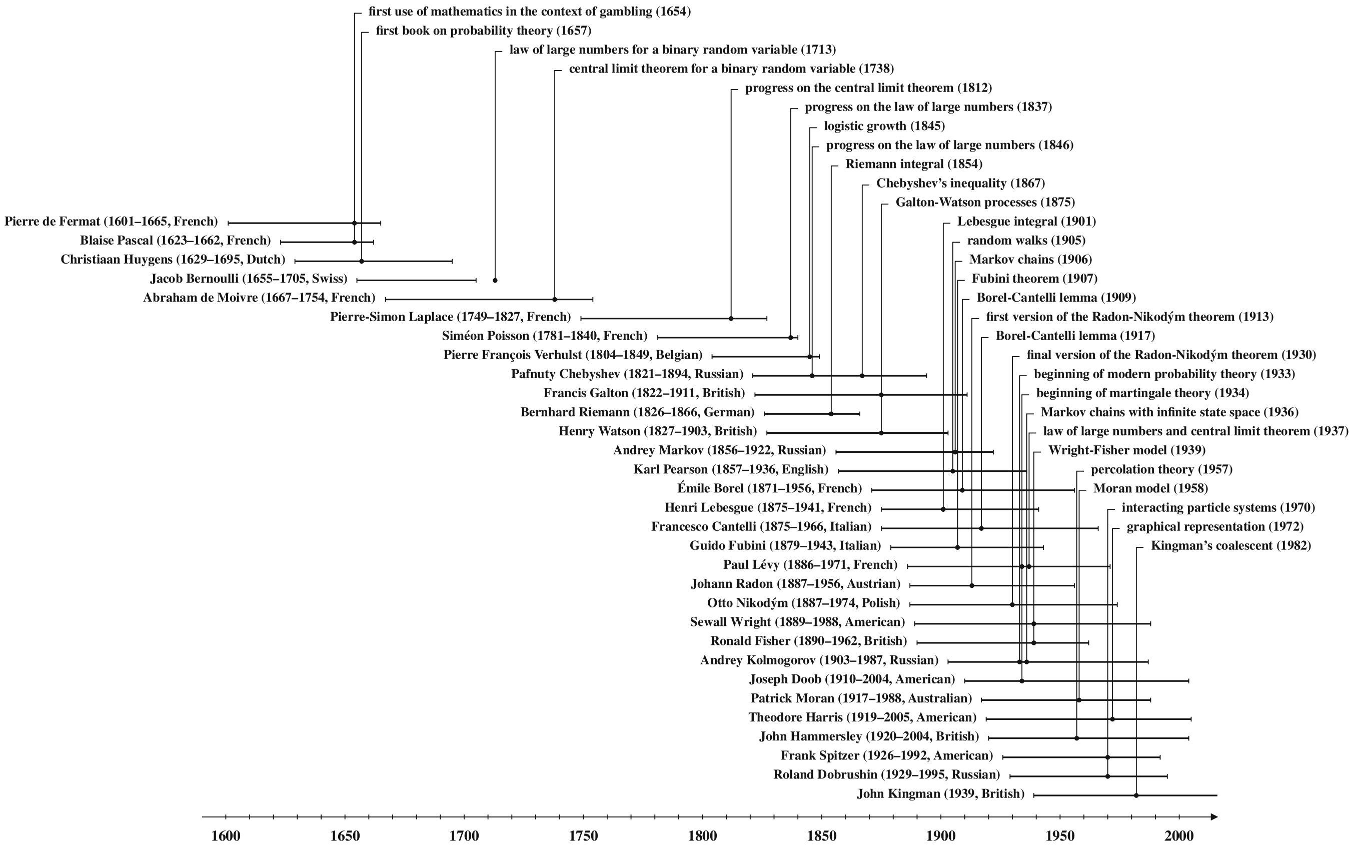
1. S. Asmussen. *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.
2. K. B. Athreya and P. E. Ney. *Branching processes*. Springer-Verlag, Heidelberg, 1972. Die Grundlehren der mathematischen Wissenschaften, Band 196.
3. P. Baldi, L. Mazliak, and P. Priouret. *Martingales and Markov chains*. Chapman & Hall/CRC, Boca Raton, FL, 2002. Solved exercises and elements of theory, Translated from the 1998 French original.
4. J. Bernoulli. *Ars Conjectandi: Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis*. 1713.
5. C. Bezuidenhout and G. Grimmett. The critical contact process dies out. *Ann. Probab.*, 18(4):1462–1482, 1990.
6. C. Bezuidenhout and G. Grimmett. Exponential decay for subcritical contact and percolation processes. *Ann. Probab.*, 19(3):984–1009, 1991.
7. P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, Inc., New York-London-Sydney, 1968.
8. P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.
9. E. Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rend. Circ. Mat. Palermo*, 27(2):247–271, 1909.
10. M. Bramson and D. Griffeath. Renormalizing the 3-dimensional voter model. *Ann. Probab.*, 7(3):418–432, 1979.
11. M. Bramson and D. Griffeath. Clustering and dispersion rates for some interacting particle systems on **Z**. *Ann. Probab.*, 8(2):183–213, 1980.
12. L. Breiman. *Probability*. Addison-Wesley, Reading, MA, 1968.
13. S. R. Broadbent and J. M. Hammersley. Percolation processes. I. Crystals and mazes. *Proc. Cambridge Philos. Soc.*, 53:629–641, 1957.
14. R. M. Burton and M. Keane. Density and uniqueness in percolation. *Comm. Math. Phys.*, 121(3):501–505, 1989.
15. F. P. Cantelli. Sulla probabilità come limite della frequenza. *Atti Accad. Naz. Lincei*, 26(1):39–45, 1917.
16. P. Chebyshev. Démonstration élémentaire d'une proposition générale de la théorie des probabilités. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 33:259–267, 1846.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6

17. P. Chebyshev. Des valeurs moyennes. *Journal de mathématiques pures et appliquées*, 12(2):177–184, 1867.
18. P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 60:581–588, 1973.
19. J. T. Cox and D. Griffeath. Occupation time limit theorems for the voter model. *Ann. Probab.*, 11(4):876–893, 1983.
20. J. T. Cox and D. Griffeath. Diffusive clustering in the two-dimensional voter model. *Ann. Probab.*, 14(2):347–370, 1986.
21. R. L. Dobrušin. Markov processes with a large number of locally interacting components: Existence of a limit process and its ergodicity. *Problemy Peredači Informacii*, 7(2):70–87, 1971.
22. R. L. Dobrušin. Markov processes with a large number of locally interacting components: The invertible case and certain generalizations. *Problemy Peredači Informacii*, 7(3):57–66, 1971.
23. J. L. Doob. *Stochastic processes*. John Wiley & Sons, Inc., New York; Chapman & Hall, Limited, London, 1953.
24. P. G. Doyle and J. L. Snell. *Random walks and electric networks*, volume 22 of *Carus Mathematical Monographs*. Mathematical Association of America, Washington, DC, 1984.
25. K. D. Duc and D. Delaunay. *Probabilité CPGE scientifiques 1re/2e année: Cours, exercices corrigés*. Collection Prépas Sciences. De Boeck Supérieur, 2015.
26. R. Durrett. Oriented percolation in two dimensions. *Ann. Probab.*, 12(4):999–1040, 1984.
27. R. Durrett. Ten lectures on particle systems. In *Lectures on probability theory (Saint-Flour, 1993)*, volume 1608 of *Lecture Notes in Math.*, pages 97–201. Springer-Verlag, Berlin, 1995.
28. R. Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition, 1996.
29. R. Durrett. *Essentials of stochastic processes*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2012.
30. R. Durrett and C. Neuhauser. Epidemics with recovery in $D = 2$. *Ann. Appl. Probab.*, 1(2):189–206, 1991.
31. 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.
32. W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, 3:87–112; erratum, ibid. 3 (1972), 240; erratum, ibid. 3 (1972), 376, 1972.
33. W. J. Ewens. *Mathematical population genetics. I*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, second edition, 2004. Theoretical introduction.
34. W. Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons, Inc., New York, 1968.
35. W. Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York, 1971.
36. E. Foxall and N. Lanchier. Rigorous results for the naming game in language dynamics. *In preparation*.
37. G. Fubini. Sugli integrali multipli. *Rom. Acc. L. Rend. (5)*, 16(1):608–614, 1907.
38. S. Galam. Sociophysics: a review of galam models. *Int. J. Mod. Phys. C*, 19:409–440, 2008.
39. F. Galton and H. W. Watson. On the probability of the extinction of families. *J. Roy. Anthropol. Inst.*, 4:138–144, 1875.
40. G. R. Grimmett. *Percolation*. Springer-Verlag, New York, 1989.
41. G. R. Grimmett, H. Kesten, and Y. Zhang. Random walk on the infinite cluster of the percolation model. *Probab. Theory Related Fields*, 96(1):33–44, 1993.
42. T. E. Harris. *The theory of branching processes*. Die Grundlehren der Mathematischen Wissenschaften, Bd. 119. Springer-Verlag, Berlin; Prentice-Hall, Inc., Englewood Cliffs, NJ, 1963.
43. T. E. Harris. Nearest-neighbor Markov interaction processes on multidimensional lattices. *Adv. Math.*, 9:66–89, 1972.
44. T. E. Harris. Contact interactions on a lattice. *Ann. Prob.*, 2:969–988, 1974.
45. R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann. Prob.*, 3(4):643–663, 1975.

46. C. Huygens. *De ratiociniis in ludo aleae*. 1657.
47. K. Itô and Jr. McKean, H. P. *Diffusion processes and their sample paths*. Die Grundlehren der Mathematischen Wissenschaften, Band 125. Academic Press, New York; Springer-Verlag, Berlin, 1965.
48. P. Jagers. *Branching processes with biological applications*. Wiley-Interscience [John Wiley & Sons], London, 1975. Wiley Series in Probability and Mathematical Statistics–Applied Probability and Statistics.
49. S. Karlin and H. M. Taylor. *A first course in stochastic processes*. Academic Press [A subsidiary of Harcourt Brace Jovanovich], New York, second edition, 1975.
50. S. Karlin and H. M. Taylor. *A second course in stochastic processes*. Academic Press [A subsidiary of Harcourt Brace Jovanovich], New York, 1981.
51. H. Kesten. The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$. *Comm. Math. Phys.*, 74(1):41–59, 1980.
52. H. Kesten. *Percolation theory for mathematicians*, volume 2 of *Progress in Probability and Statistics*. Birkhäuser, Boston, MA, 1982.
53. J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982.
54. J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.*, (Special Vol. 19A):27–43, 1982. Essays in statistical science.
55. J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1993. Oxford Science Publications.
56. A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. 1933.
57. A. N. Kolmogorov. Zur Theorie der Markoffschen Ketten. *Math. Ann.*, 112(1):155–160, 1936.
58. A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing, New York, 1956. Translation edited by Nathan Morrison, with an added bibliography by A. T. Bharucha-Reid.
59. M. Krishnapur and Y. Peres. Recurrent graphs where two independent random walks collide finitely often. *Electron. Comm. Probab.*, 9:72–81 (electronic), 2004.
60. N. Lanchier. Rigorous proof of the Boltzmann–Gibbs distribution of money on connected graphs. *In preparation*.
61. N. Lanchier. The Axelrod model for the dissemination of culture revisited. *Ann. Appl. Probab.*, 22(2):860–880, 2012.
62. N. Lanchier and S. Scarlatos. Limiting behavior for a general class of voter models with confidence threshold. *Preprint. Available as arXiv:1412.4142.*
63. N. Lanchier and N. Taylor. Galam's bottom-up hierarchical system and public debate model revisited. *Adv. in Appl. Probab.*, 47(3):668–692, 2015.
64. P. S. Laplace. *Théorie analytique des probabilités*. 1991.
65. G. F. Lawler. *Intersections of random walks*. Probability and Its Applications. Birkhäuser Boston, Boston, MA, 1991.
66. H. Lebesgue. *Intégrale, longueur, aire*. PhD thesis, Université de Nancy, France, 1902.
67. P. Lévy. *Théory de l'addition des variables aléatoires*. Gauthier-Villars, Paris, 1937.
68. P. Lévy. *Processus stochastiques et mouvement brownien*. Suivi d'une note de M. Loève. Deuxième édition revue et augmentée. Gauthier-Villars & Cie, Paris, 1965.
69. T. M. Liggett. *Interacting particle systems*, volume 276 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1985.
70. T. M. Liggett. *Stochastic interacting systems: contact, voter, and exclusion processes*, volume 324 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1999.
71. A. A. Markov. Extension of the law of large numbers to dependent quantities. *Izv. Fiz.-Matem. Obsch. Kazan Univ.*, 15:135–156, 1906.
72. A. de Moivre. *The doctrine of chances: or, a method for calculating the probabilities of events in play*. 1738.
73. P. A. P. Moran. Random processes in genetics. *Proc. Cambridge Philos. Soc.*, 54:60–71, 1958.

74. I. Nåsell. *Extinction and quasi-stationarity in the stochastic logistic SIS model*, volume 2022 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Mathematical Biosciences Subseries.
75. C. Neuhauser. Ergodic theorems for the multitype contact process. *Probab. Theory Related Fields*, 91(3–4):467–506, 1992.
76. C. Neuhauser. *Mathematical models in population genetics, In Handbook of statistical genetics. Vol. I, II.* John Wiley & Sons, Chichester, third edition, 2007.
77. J. Neveu. *Martingales à temps discret.* Masson et Cie, éditeurs, Paris, 1972.
78. O. Nikodým. Sur une généralisation des intégrales de M. J. Radon. *Fundamenta Mathematicae*, 15:131–179, 1930.
79. J. R. Norris. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.
80. M. A. Nowak. *Evolutionary dynamics*. The Belknap Press of Harvard University Press, Cambridge, MA, 2006. Exploring the equations of life.
81. K. Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.
82. R. Pemantle. The contact process on trees. *Ann. Probab.*, 20(4):2089–2116, 1992.
83. S. D. Poisson. *Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés*. Paris, France: Bachelier, 1837.
84. B. Riemann. *Über die Darstellbarkeit einer Function durch eine trigonometrische Reihe*. Habilitationsschrift, University of Göttingen, 1854.
85. S. M. Ross. *Stochastic processes*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, New York, second edition, 1996.
86. S. M. Ross. *A first course in probability*. Macmillan Co., New York; Collier Macmillan, London, ninth edition, 2014.
87. S. M. Ross. *Introduction to probability models*. Elsevier/Academic Press, Amsterdam, eleventh edition, 2014.
88. W. Rudin. *Real and complex analysis*. McGraw-Hill, New York, third edition, 1987.
89. R. B. Schinazi. *Classical and spatial stochastic processes*. Birkhäuser Boston, Boston, MA, 1999.
90. A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1996. Translated from the first (1980) Russian edition by R. P. Boas.
91. F. Spitzer. *Principles of random walk*. Springer-Verlag, second edition, 1976. Graduate Texts in Mathematics, Vol. 34.
92. Frank Spitzer. Interaction of Markov processes. *Adv. Math.*, 5:246–290 (1970), 1970.
93. P. F. Verhulst. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18:1–42, 1845.
94. J. Ville. *Étude critique de la notion de collectif.* Monographies des probabilités. Gauthier-Villars, 1939.
95. D. Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.
96. W. Woess. *Random walks on infinite graphs and groups*, volume 138 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2000.
97. S. Wright. *Statistical Genetics in Relation to Evolution*. Number 802–803 in Actualités scientifiques et industrielles. Hermann, 1939.
98. I. Zähle. Renormalization of the voter model in equilibrium. *Ann. Probab.*, 29(3):1262–1302, 2001.


first use of mathematics in the context of gambling (1654)
first book on probability theory (1657)
law of large numbers for a binary random variable (1713)
central limit theorem for a binary random variable (1738)
progress on the central limit theorem (1812)
progress on the law of large numbers (1837)
logistic growth (1845)
progress on the law of large numbers (1846)
Riemann integral (1854)
Chebyshev's inequality (1867)
Galton-Watson processes (1875)
Lebesgue integral (1901)
random walks (1905)
Markov chains (1906)
Fubini theorem (1907)
Borel-Cantelli lemma (1909)
first version of the Radon-Nikodým theorem (1913)
Borel-Cantelli lemma (1917)
final version of the Radon-Nikodým theorem (1930)
beginning of modern probability theory (1933)
beginning of martingale theory (1934)
Markov chains with infinite state space (1936)
law of large numbers and central limit theorem (1937)
Wright-Fisher model (1939)
percolation theory (1957)
Moran model (1958)
interacting particle systems (1970)
graphical representation (1972)
Kingman's coalescent (1982)
Pierre de Fermat (1601–1665, French)
Blaise Pascal (1623–1662, French)
Christiaan Huygens (1629–1695, Dutch)
Jacob Bernoulli (1655–1705, Swiss)
Abraham de Moivre (1667–1754, French)
Pierre-Simon Laplace (1749–1827, French)
Siméon Poisson (1781–1840, French)
Pierre François Verhulst (1804–1849, Belgian)
Pafnuty Chebyshev (1821–1894, Russian)
Francis Galton (1822–1911, British)
Bernhard Riemann (1826–1866, German)
Henry Watson (1827–1903, British)
Andrey Markov (1856–1922, Russian)
Karl Pearson (1857–1936, English)
Émile Borel (1871–1956, French)
Henri Lebesgue (1875–1941, French)
Francesco Cantelli (1875–1966, Italian)
Guido Fubini (1879–1943, Italian)
Paul Lévy (1886–1971, French)
Johann Radon (1887–1956, Austrian)
Otto Nikodým (1887–1974, Polish)
Sewall Wright (1889–1988, American)
Ronald Fisher (1890–1962, British)
Andrey Kolmogorov (1903–1987, Russian)
Joseph Doob (1910–2004, American)
Patrick Moran (1917–1988, Australian)
Theodore Harris (1919–2005, American)
John Hammersley (1920–2004, British)
Frank Spitzer (1926–1992, American)
Roland Dobrushin (1929–1995, Russian)
John Kingman (1939, British)
1600
1650
1700
1750
1800
1850
1900
1950
2000

Index

© Springer International Publishing AG 2017
N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6