SOR3012 Stochastic Processes and Risk

Andrea Munaro

Contents

Chapter 1

Probability

Uncertainty and randomness are unavoidable aspects of our experience: play cards, invest in shares, etc. Although probability has been around for several centuries, it wasn't until recently that the subject was made rigorous. In the thirties, Kolmogorov showed that it is full-fledged analysis, more precisely, measure theory.

The first part of the module is devoted to formally introducing the objects of probability. In other words, the goal is to abstract the common features arising in everyday examples in order to build a **probabilistic model** i.e., a mathematical description of an uncertain situation. The advantage of taking an abstract approach is that it allows to develop general tools that can be adapted to several specific situations. We start with the following definition.

Definition 1.0.1. Any well-defined procedure or chain of circumstances is called an **experiment**. The end result, or occurrence, is the **outcome** of the experiment, also known as **elementary event**. The set of all possible outcomes is the **sample space**, denoted by Ω.

Experiment	Possible outcomes
Roll a die	$\Omega = \{1, 2, 3, 4, 5, 6\}$
Toss a coin	$\Omega = \{\text{H, T}\}$
Infinite sequence of coin tosses	Ω is the set of all possible infinite sequences of H and T

In the first two experiments Ω is finite, whereas in the third it is infinite. Typically, rather than individual outcomes of the sample space, we are interested in collections of outcomes.

Experiment	Set of outcomes of interest
Roll a die	The outcome is an even number
Toss a coin	The outcome is either H or T
Infinite sequence of coin tosses	The outcome consists of finitely many H

These collections of outcomes are associated to the intuitive notion of event, which is then a subset of the sample space. If the result of the experiment belongs to this subset, we would say that the event occured. Thinking of events as subsets of the sample space, we can then perform on them the usual set-theoretic operations.

Notation	Set jargon	Probability jargon
Ω	Collection of objects	Sample space
ω	Member of Ω	Outcome (also called elementary event)
A	Subset of Ω	A occurs (more precisely, some outcome in A occurs)
A^c	Complement of A	A does not occur (more precisely, no outcome in A occurs)
$A \cap B$	Intersection	Both A and B occur
$A \cup B$	Union	At least one of A and B occurs
$A \setminus B$	Difference	A occurs but not B
$A \subseteq B$	Inclusion	If A occurs then B occurs

We would ultimately like to assign a probability to an event and so the natural question is: Is there any property events should satisfy? As already observed, in general, the sample space might be infinite and the events we are interested in might contain infinitely many outcomes:

Example 1.0.2. A coin is tossed until the first head turns up and we are concerned with the number of tosses before this happens. We let $\Omega = \{\omega_1, \omega_2, \dots\}$, where ω_i denotes the outcome "the first $i-1$ tosses are tails and the i-th is head". We might be interested in the following event A: "the first head occurs after an even number of tosses". Clearly, $A = \{\omega_2, \omega_4, \dots\} = \bigcup_{i=1}^{\infty} \{\omega_{2i}\}$ is a countable union of members of Ω i.e., elementary events.

Let's recall the important notion of countable set.

Definition 1.0.3. A set is **countable** if it is in bijection with a subset of the set of positive integers. A set is **uncountable** if it is not countable.

Example 1.0.4. The set of positive integers is obviously countable. Every finite set is countable. The set of rational numbers is countable. The set of all infinite $0,1$ sequences is uncountable: this can be shown using Cantor's diagonal argument. Other examples of uncountable sets are $\mathbb{R}$ and its subinterval $(0,1)$.

We need also to recall the notions of union and intersection of a collection of subsets:

Definition 1.0.5. Let $\mathcal{C}$ be a collection of subsets of a set X. The subset of X containing all elements that belong to at least one set of $\mathcal{C}$ is the **union** of the collection $\mathcal{C}$, denoted by $\bigcup \mathcal{C}$. If $\mathcal{C} = \{A_i : i = 1, \dots, n\}$ is finite, we usually write $\bigcup \mathcal{C} = \bigcup_{i=1}^{n} A_i$. If $\mathcal{C} = \{A_i : i \in \mathbb{N}\}$ is countable infinite, we usually write $\bigcup \mathcal{C} = \bigcup_{i=1}^{\infty} A_i$.

The subset of X containing all elements that belong to all sets of $\mathcal{C}$ is the **intersection** of the collection $\mathcal{C}$, denoted by $\bigcap \mathcal{C}$. If $\mathcal{C} = \{A_i : i = 1, \dots, n\}$ is finite, we let $\bigcap \mathcal{C} = \bigcap_{i=1}^{n} A_i$. If $\mathcal{C} = \{A_i : i \in \mathbb{N}\}$ is countable infinite, we let $\bigcap \mathcal{C} = \bigcap_{i=1}^{\infty} A_i$.

The following relations between unions and intersections will be used repeatedly.

Example 1.0.6 (De Morgan's laws). "It will not snow or rain" means "It will not snow and it will not rain". If S is event that it snows and R is event that it rains, then $(S \cup R)^c = S^c \cap R^c$. "It will not both snow and rain" means "Either it will not snow or it will not rain" i.e., $(S \cap R)^c = S^c \cup R^c$.

More generally, let $\{A_i : i \in I\}$ be a collection of sets with I countable. Then

$$\Big(\bigcup_i A_i\Big)^c = \bigcap_i A_i^c \quad \text{and} \quad \Big(\bigcap_i A_i\Big)^c = \bigcup_i A_i^c.$$

1.1 Sigma-fields and probability measures

Back to our question: which subsets of the sample space Ω are events? There are certain requirements that we wish the collection of events to satisfy:

- Ω is an event: this is the trivial event that something happened.
- If $A \subseteq \Omega$ is an event, so is A^c: if we are allowed to ask whether A has occurred, we should also be allowed to ask whether A has not occurred.
- If $A_1, A_2, \ldots \subseteq \Omega$ are events, so is their union $\bigcup_{i=1}^{\infty} A_i$: if we are allowed to ask whether each A_i has occurred, we should also be allowed to ask whether at least one of the A_i's has occurred, as seen in Example 1.0.2.

Definition 1.1.1. A collection $\mathcal{F}$ of subsets of Ω is called a **σ-field** (or **σ-algebra**) if it satisfies the following conditions:

1. $\Omega \in \mathcal{F}$;
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
3. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Our events will form a σ-field of the sample space Ω. Why don't we go further and allow uncountable unions? Well, our unions here will be closely tied to sums of probabilities and uncountable sums can be extremely messy.

Remark 1.1.2. Notice that 1. and 2. imply that $\varnothing \in \mathcal{F}$. Moreover, if $A_1, A_2, \ldots \in \mathcal{F}$ is a countable subfamily of $\mathcal{F}$ then, by De Morgan's laws, $\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^c)^c \in \mathcal{F}$. Notice also that, since $\varnothing \in \mathcal{F}$, we can extend any finite subfamily $A_1, \ldots, A_n$ of $\mathcal{F}$ to a countable family by setting $A_j = \varnothing$ for each $j > n$. Therefore, finite unions and intersections of members of $\mathcal{F}$ are still in $\mathcal{F}$. Moreover, if $A, B \in \mathcal{F}$, then $A \setminus B = A \cap B^c \in \mathcal{F}$.

In other words, we have verified that:

> A σ-field is stable (or closed) under countable set operations.

Example 1.1.3. The following are some examples of σ-fields:

- $\{\varnothing, \Omega\}$ is a σ-field of Ω.
- The power set of Ω is a σ-field of Ω.
- For any $A \subseteq \Omega$, $\{\varnothing, A, A^c, \Omega\}$ is a σ-field of Ω.
- The collection of all open intervals of $\mathbb{R}$ is not a σ-field. Indeed, $\bigcap_{n=1}^{\infty}(-\frac{1}{n}, \frac{1}{n}) = \{0\}$ is not an open interval.

Exercise 1.1.4. *Write down all σ-fields on $\{a, b\}$.*

Exercise 1.1.5. *Let Ω be an infinite set and let $\mathcal{A} = \{A \in \Omega : A \text{ is finite or } A^c \text{ is finite}\}$. Show that $\mathcal{A}$ is not a σ-field.*

A natural question might arise: Why not simply taking the power set of Ω all the time for our probabilistic interests? The reason is that, if Ω is uncountable, its power set is too rich and it turns out to be impossible to assign probabilities in a consistent fashion to all possible subsets. Luckily, in many situations, for example when Ω is countable, we can indeed simply consider the power set of Ω.

But let's deal with the general situation. If $\mathcal{C}$ is a collection of some basic events that we want to be able to discuss, we have seen that it is not necessarily a σ-field. What is typically done is to enlarge such a collection so that it in fact becomes one. This is done by considering the following notion.

Definition 1.1.6. Let $\mathcal{C}$ be a collection of subsets of Ω. The **σ-field generated by $\mathcal{C}$**, denoted $\sigma(\mathcal{C})$ is the smallest σ-field on Ω containing the collection $\mathcal{C}$.

Remark 1.1.7. Some cautionary words on language. We say that a σ-field $\mathcal{F}$ contains the collection $\mathcal{C}$ if each member of $\mathcal{C}$ belongs to $\mathcal{F}$ (i.e., every set in $\mathcal{C}$ is a set in $\mathcal{F}$). For two σ-fields $\mathcal{F}_1$ and $\mathcal{F}_2$, we say that $\mathcal{F}_1$ is smaller than $\mathcal{F}_2$ if $\mathcal{F}_1 \subseteq \mathcal{F}_2$.

Notice that we need to verify that this notion is well-defined!

Lemma 1.1.8. *Let $\mathcal{C}$ be a collection of subsets of Ω. Then the σ-field $\sigma(\mathcal{C})$ generated by $\mathcal{C}$ exists and is unique.*

Proof. Uniqueness is trivial: If we have two smallest σ-fields containing $\mathcal{C}$, say $\mathcal{F}_1$ and $\mathcal{F}_2$, then $\mathcal{F}_1 \subseteq \mathcal{F}_2$ and $\mathcal{F}_2 \subseteq \mathcal{F}_1$.

Consider now existence. Let $\mathscr{S}$ be the collection of all σ-fields on Ω containing $\mathcal{C}$. Notice that $\mathscr{S}$ is nonempty, as the power set of Ω belongs to $\mathscr{S}$. We claim that $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$ is the smallest σ-field on Ω containing $\mathcal{C}$. There are three things to be checked:

- $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$ is a σ-field.

  Here we need to check the three properties in Definition 1.1.6:

  1. Since each $\mathcal{F} \in \mathscr{S}$ is a σ-field on Ω, $\Omega \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$ and so $\Omega \in \bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$.
  2. Let $A \in \bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$. Then $A \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$ and so $A^c \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$. Therefore, $A^c \in \bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$.
  3. Let $A_1, A_2, \ldots \in \bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$. Then $A_1, A_2, \ldots \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$. But $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$ and so $\bigcup_{i=1}^{\infty} A_i \in \bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$.

- $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$ contains $\mathcal{C}$.

  This follows from the fact that, for each $\mathcal{F} \in \mathscr{S}$, $\mathcal{F}$ contains $\mathcal{C}$.

- $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$ is the smallest σ-field containing $\mathcal{C}$.

  Let $\mathcal{G}$ be any σ-field containing $\mathcal{C}$. Then $\mathcal{G} \in \mathscr{S}$ and so $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F} \subseteq \mathcal{G}$. □

In order to define what is arguably the most important σ-field on $\mathbb{R}^n$, we need a little bit of topology.

Definition 1.1.9. A subset $U \subseteq \mathbb{R}$ is **open** if, for each $x \in U$, there exists $\varepsilon > 0$ such that the open interval centered at x and with radius ε is contained in U. In other words, $(x - \varepsilon, x + \varepsilon) \subseteq U$.

More generally, a subset $U \subseteq \mathbb{R}^n$ is **open** if, for each $x \in U$, there exists $\varepsilon > 0$ such that the open ball centered at x and with radius ε is contained in U. In other words, $B_\varepsilon(x) = \{y \in \mathbb{R}^n : |x - y| < \varepsilon\} \subseteq U$, where $|x - y|$ denotes the Euclidean distance in $\mathbb{R}^n$ between x and y.

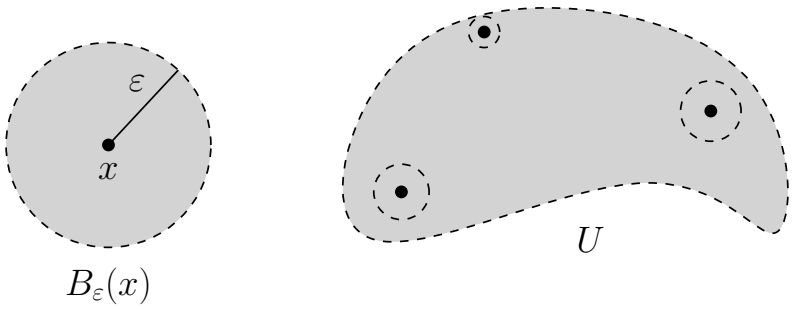


Figure 1.1: The open ball $B_\varepsilon(x)$ in $\mathbb{R}^2$ and an example of an arbitrary open set $U \subseteq \mathbb{R}^2$: for each of its points x there exists a sufficiently small open ball centered at x and contained in U.

Example 1.1.10. Every open interval in $\mathbb{R}$ is open. Similarly, every open disk in $\mathbb{R}^2$ is open. For each $x \in \mathbb{R}$, $\{x\}^c$ is an open set in $\mathbb{R}$ but $\{x\}$ is obviously not open. More generally, the closed interval $[x, y]$ is not open but $[x, y]^c$ is.

Exercise 1.1.11. *Let U_1 and U_2 be two open sets in $\mathbb{R}^n$. Is $U_1 \cap U_2$ open?*

Exercise 1.1.12. *Show that every open set in $\mathbb{R}^n$ is a union of open balls.*

Definition 1.1.13. The **Borel σ-field** $\mathcal{B}$ on $\mathbb{R}^n$ is the σ-field generated by the open sets in $\mathbb{R}^n$. The sets in $\mathcal{B}$ are called **Borel sets**.

The Borel σ-field on the real line is extremely important in probability theory and will appear again when we will talk about random variables. Rather than taking the whole family of open sets, it can in fact be equivalently generated by open intervals, closed intervals, half-lines, etc.

Proposition 1.1.14. *The Borel σ-field on the real line $\mathbb{R}$ is generated by any of the following collections of subsets of $\mathbb{R}$:*

- $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{R}\}$;
- $\mathcal{C} = \{(x, y) : x, y \in \mathbb{R},\ x < y\}$;
- $\mathcal{C} = \{[x, y] : x, y \in \mathbb{R},\ x \leq y\}$;
- $\mathcal{C} = \{(x, y] : x, y \in \mathbb{R},\ x < y\}$.

Remark 1.1.15. Proposition 1.1.14 says that the smallest σ-field on $\mathbb{R}$ containing all open sets can be generated by the family of closed intervals. This might appear odd, as closed intervals are not open. But recall that σ-fields are closed under complementation, and it is by taking complements of closed intervals (and their intersections and unions) that we manage to get the open sets.

Example 1.1.16. Since $\mathcal{B}$ is a σ-field containing all open sets, it contains all singletons $\{x\}$, as $\{x\}^c$ is an open set in $\mathbb{R}$.

To any experiment we will then associate the pair $(\Omega, \mathcal{F})$, where Ω is the set of all possible outcomes (elementary events) and $\mathcal{F}$ is a σ-field of subsets of Ω. We will try to assign a probability to each set in $\mathcal{F}$ and in order to do so, we will again be guided by intuition.

Suppose that an experiment has several possible outcomes that are not necessarily equally likely. How can we define the probability of a certain event A? One intuitive way is the following. We run the experiment a large number N of times, keeping the initial conditions as equal as possible. Denoting

by $N(A)$ the number of occurences of A after the first N trials, we would expect that when N becomes larger and larger, the ratio $N(A)/N$ converges to some finite limit. We may then define the probability $\mathbb{P}(A)$ that A occurs on a particular trial as this limit. In any case, for large N, $N(A)/N$ should be an approximation of $\mathbb{P}(A)$. Notice that

- $0 \leq N(A)/N \leq 1$;
- If $A = \varnothing$, then $N(A)/N = 0$. If $A = \Omega$, then $N(A)/N = 1$;
- If A and B are disjoint events, then $N(A \cup B) = N(A) + N(B)$ and so $N(A \cup B)/N = N(A)/N + N(B)/N$.

With the observations above in mind and recalling Example 1.0.2, we state the following:

Definition 1.1.17. A **probability measure** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P}\colon \mathcal{F} \to \mathbb{R}$ satisfying the following:

1. For each $A \in \mathcal{F}$, we have $0 \leq \mathbb{P}(A) \leq 1$;
2. $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(\Omega) = 1$;
3. For every countable infinite collection $A_1, A_2, \ldots$ of mutually disjoint members of $\mathcal{F}$ (i.e., $A_i \cap A_j = \varnothing$ for each $i \neq j$), we have
$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)^{1}.$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of a set Ω, a σ-field $\mathcal{F}$ of subsets of Ω and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is called a **probability space**. Any set in $\mathcal{F}$ is called **event**.

Observation 1.1.18. *The axioms $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(A) \leq 1$ in Definition 1.1.17 are in fact redundant i.e., they can be deduced from the others. Check it!*

Remark 1.1.19. The event Ω is the **sure event**: it contains all possible outcomes and $\mathbb{P}(\Omega) = 1$. It is worth noting that there may be also other events $E \in \mathcal{F}$ such that $\mathbb{P}(E) = 1$. Such events are called **almost sure**.

Countable additivity (the last condition in the definition) readily implies the following result:

Lemma 1.1.20 (Finite additivity). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For every finite collection $A_1, A_2, \ldots, A_n$ of mutually disjoint members of $\mathcal{F}$, we have*

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Proof. Define $A_m = \varnothing$ for each $m > n$. Since $\varnothing \in \mathcal{F}$ and the newly defined countable infinite collection $A_1, A_2, \ldots$ consists of mutually disjoint members of $\mathcal{F}$, we use countable additivity to conclude. □

Remark 1.1.21. As mentioned above, in many cases, we can take $\mathcal{F}$ as the power set of Ω. But sometimes it is simply not possible to assign a consistent probability to all subsets of Ω. Here is the rough intuition.

Choose a number at random from the interval $[0, 1]$ i.e., no number is more likely than any other. We expect that the probability that the number falls in any sub-interval $[a, a+h]$ is the same as the probability it falls in any other sub-interval of the same length. The probability must then be proportional to the length of the interval and, since the whole interval has length one, we conclude that the probability of

[1] Notice that we are requiring this series to be convergent.

falling in a sub-interval of $[0,1]$ is actually equal to its length. The problem is that we cannot define the length of every subset of $[0,1]$! One example of a set without a determinable length is the so-called Vitali set. The good news is that these sets are hard to construct and have no practical importance. However, should we not take into account these obstructions, our complete theoretical machinery would collapse!

The conceptual construction of a probability space has no absolute physical meaning, it is just guided by some intuitive physical interpretation. The properties which the measure $\mathbb{P}$ is required to satisfy are usually called the **probability axioms** and were introduced by Kolmogorov, though not exactly in the form above (see Observation 1.1.18). The first two axioms are just a matter of convention. The key one is countable additivity.

Think of a probability space as the mathematical description of an experiment. For example, tossing a coin, rolling dice, taking a number in a lottery, etc. In each case, there is a certain amount of randomness, or unpredictability in the experiment. To describe this mathematically, start with what we observe: the outcome. Ω is the set of all possible outcomes of the experiment: each point of Ω represents an outcome. An event is a set of outcomes belonging to the σ-field $\mathcal{F}$. The probability measure gives the probability of events. We can associate a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with any experiment. The informations allowing us to compute the actual value of $\mathbb{P}(A)$ are contained in the description of the experiment.

Example 1.1.22. Suppose the experiment is rolling a die i.e., a cube whose six faces are numbered 1 to 6. We can take $\Omega = \{1,2,3,4,5,6\}$ as the set of outcomes and, since Ω is countable, the power set of Ω as the σ-field $\mathcal{F}$. To get the probability measure, we note that if the die is well-made, the six sides are identical except for their label. No side can be more probable than another. Therefore,

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \cdots = \mathbb{P}(\{6\}).$$

The probabilities in the preceding example were derived from symmetry considerations: the possible outcomes were indistinguishable except by their labels. In fact, this is about the only situation in which we can confidently assign probabilities by inspection. But luckily, while nature is not always obliging enough to divide itself into equally-likely pieces, one can start with the equally-likely case and then determine the probabilities in more complex situations. Which is what the subject is about.

The idea of symmetry applies to events, not just outcomes. Consider a physical experiment with finitely or countably many outcomes, labeled in some convenient fashion.

Symmetry principle: If two events are indistinguishable except for the way the outcomes are labeled, they are equally likely.

For example, roll a die and consider the events "even" and "odd", i.e., $\{2,4,6\}$ and $\{1,3,5\}$. If we physically renumber the faces of the die, so that we interchange n and $7-n$ on each face, so that $1 \leftrightarrow 6$, $2 \leftrightarrow 5$ and $3 \leftrightarrow 4$, then the events "even" and "odd" are interchanged. The symmetry principle says that the two events must have the same probability.

Example 1.1.23. Suppose the experiment is tossing a coin. We can take $\Omega = \{H,T\}$, $\mathcal{F} = 2^{\Omega} = \{\varnothing, H, T, \Omega\}$ and $\mathbb{P}$ defined by

$$\mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(H) = p, \quad \mathbb{P}(T) = 1-p,$$

where p is a fixed real number in $[0,1]$. If $p = 1/2$, we say that the coin is **fair**.

Remark 1.1.24. The probability space for an experiment is not unique. This is useful in practice: it allows us to choose the probability space which works best in the particular circumstances, or to not choose one at all; we do not always have to specify the probability space. It is usually enough to know that it is there if we need it.

The simplest probability spaces are those whose sample space $\Omega = \{\omega_1, \omega_2, \dots\}$ contains countably many outcomes (we call such probability spaces, **countable probability spaces**). Recall that in such cases we may always take as σ-field $\mathcal{F}$ the power set 2^Ω. For countable probability spaces, a probability measure $\mathbb{P}$ on $\mathcal{F}$ is fully determined by the values assigned to the elementary events ω_i. Indeed, consider an event $A \subseteq \Omega$. Since A is countable (infinite or finite), then

$$A = \bigcup_{\omega \in A} \{\omega\}$$

can be expressed as a countable union of elementary events. Since these events are obviously mutually disjoint, it follows from countable additivity (or finite additivity) that

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}).$$

Suppose now that the sample space Ω is finite and that $\mathbb{P}(\{\omega\}) = p$, for each elementary event $\omega \in \Omega$. By the probability axioms, we have

$$1 = \mathbb{P}(\Omega) = \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = p|\Omega|,$$

from which $p = 1/|\Omega|$. Therefore, the probability of the event A is

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) = p|A| = \frac{|A|}{|\Omega|}.$$

Observation 1.1.25. *A probability space is a model for an experiment. A priori there is no reason why all outcomes should be equally probable. It is an assumption that should be made only when believed to be applicable.*

Example 1.1.26. Two fair dice are rolled. What is the probability that the sum is 7?

A convenient sample is constructed by viewing the two dice as distinguishable (say one blue and one red) and taking $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$. By symmetry, it is natural to assume that each of the $|\Omega| = 36$ outcomes is equally likely. The event "sum equals 7" is $A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ and so $\mathbb{P}(A) = |A|/|\Omega| = 1/6$.

Observation 1.1.27. *How do we know which probability space to assign to each experiment? Well, a model is a model: it may or may not relate to reality. In the previous example, we applied symmetry, as we believe that all outcomes of the experiment are equally likely.*

Example 1.1.28. A tea set has four cups and saucers with two cups and saucers in each of two different colors, say a and b. If the cups are placed at random on the saucers, what is the probability that no cup is on a saucer of the same color?

As a sample space, we consider the distinct ways of arranging the cups by color with the saucers fixed (suppose without loss of generality the saucers are listed as $aabb$). There are six possible ways of arranging the cups: $aabb, abba, abab, baab, baba, bbaa$. By symmetry, they are equally likely. Since only one of these arrangements has no cup on a saucer of the same color, the required probability is $1/6$.

How did we know there are exactly six possible ways? Well, that corresponds to the number of ways of placing the cups of color a. Indeed, for each such a choice, the positions of the cups of color b are forced. But then this is the general problem of counting the number of ways a subset of size k can be chosen from a set of size $n \geq k$ or, equivalently, the number of subsets of size k of a set of size n. Let us first count the number of *ordered* subsets of size k of a set of size n. We have n choices for the element in first position. For each such a choice, we have $n-1$ choices for the element in second position and so on up to $n-(k-1)$ choices for the last element in position k. Overall,

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

ordered subsets. Of course, if we are interested in *unordered* subsets, then we have overcounted: every subset was counted exactly $k!$ times (with every possible ordering of its elements). So we have to divide by $k!$.

Lemma 1.1.29. *The number of subsets of size k of a set of size n is*

$$\frac{n!}{k!(n-k)!},$$

*denoted by $\binom{n}{k}$. The numbers $\binom{n}{k}$ are called **binomial coefficients**.*

Exercise 1.1.30. *A bag contains 2021 red balls and 2021 black balls. We remove two balls at a time repeatedly and*

- *discard them if they are of the same color;*
- *discard the black ball and return to the bag the red ball if they are of different colors.*

What is the probability that this process will terminate with one red ball in the bag? *Hint: What are the possible outcomes?*

Back to our generic probability space $(\Omega, \mathcal{F}, \mathbb{P})$, our goal is now to derive several useful properties from the probability axioms.

Lemma 1.1.31. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ be events. The following are true:*

(a) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*;*

(b) If $A \subseteq B$*, then* $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$ *(monotonicity);*

(c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ *(inclusion-exclusion).*

Proof. Notice first that, as observed in Remark 1.1.2, all the sets considered belong to $\mathcal{F}$ and so we can indeed talk about their probabilities.

(a) Since $A \cap A^c = \varnothing$ and $A \cup A^c = \Omega$, finite additivity and the 2nd axiom imply that $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$.

(b) Since $A \cap (B \setminus A) = \varnothing$ and $A \cup (B \setminus A) = B$, finite additivity and the 2nd axiom imply that $\mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(B)$. Since $\mathbb{P}(B \setminus A) \geq 0$ (1st axiom), we then have that $\mathbb{P}(B) \geq \mathbb{P}(A)$.

(c) $A \cup B$ can be written as the disjoint union $A \cup (B \setminus A)$. We then have that,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B)) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

where in the first equality we use additivity, in the second the fact that $B \setminus A = B \setminus (A \cap B)$ and in the third (b). □

Example 1.1.32. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ be events. Although $\mathbb{P}(A) = \mathbb{P}(A \cap B)$ is obviously false in general, it is true if $\mathbb{P}(B) = 1$. Indeed, in this case

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) = \mathbb{P}(A) + (1 - \mathbb{P}(A \cup B)) \geq \mathbb{P}(A).$$

The reverse inequality always holds by monotonicity.

Recall the following fact from analysis: If $f\colon \mathbb{R} \to \mathbb{R}$ is a continuous function at x_0 and the sequence $x_1, x_2, \ldots$ converges to x_0, then the sequence $f(x_1), f(x_2), \ldots$ converges to $f(x_0)$ (if you haven't seen it, try to show it!). A similar statement holds for probability measures.

Lemma 1.1.33 (Continuity of probability). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For every increasing sequence of events $A_1, A_2, \ldots$ (i.e., $A_1 \subseteq A_2 \subseteq \cdots$), we have that*

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \lim_{i \to \infty} \mathbb{P}(A_i).$$

Similarly, for every decreasing sequence of events $B_1, B_2, \ldots$ (i.e., $B_1 \supseteq B_2 \supseteq \cdots$), we have that

$$\mathbb{P}\Big(\bigcap_{i=1}^{\infty} B_i\Big) = \lim_{i \to \infty} \mathbb{P}(B_i).$$

Proof. Consider the case of an increasing sequence $A_1 \subseteq A_2 \subseteq \cdots$. We write $\bigcup_{i=1}^{\infty} A_i$ as the disjoint union $A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots$. By countable additivity and recalling the definition of a sum of a series, we have

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \mathbb{P}(A_1) + \sum_{i=2}^{\infty} \mathbb{P}(A_i \setminus A_{i-1}) = \mathbb{P}(A_1) + \lim_{n \to \infty} \sum_{i=2}^{n} \mathbb{P}(A_i \setminus A_{i-1}).$$

Since $A_{i-1} \subseteq A_i$, monotonicity tells us that $\mathbb{P}(A_i \setminus A_{i-1}) = \mathbb{P}(A_i) - \mathbb{P}(A_{i-1})$ and so

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \mathbb{P}(A_1) + \lim_{n \to \infty} \sum_{i=2}^{n} (\mathbb{P}(A_i) - \mathbb{P}(A_{i-1})) = \lim_{n \to \infty} \mathbb{P}(A_n).$$

The second assertion follows by taking complements and is left as an exercise. □

Exercise 1.1.34. *Show the second assertion in Lemma 1.1.33.*

The following result, despite its simplicity, is extremely useful in probability theory. It asserts that the probability that at least one event in a sequence occurs can not exceed the sum of the probabilities of the events in the sequence.

Lemma 1.1.35 (Union bound). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, A_2, \ldots$ be a sequence of events. Then*

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Proof. We show first that, for each n, $\mathbb{P}(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$. We proceed by induction on n, the case $n = 1$ being trivial. Therefore, suppose that $A_1, \ldots, A_{n+1} \in \mathcal{F}$. We define $A = A_1 \cup \cdots \cup A_n$ and $B = A_{n+1} \setminus A$. Then $A_1 \cup \cdots \cup A_{n+1}$ can be written as the disjoint union $A \cup B$. But then

$$\mathbb{P}(A_1 \cup \cdots \cup A_{n+1}) = \mathbb{P}(A) + \mathbb{P}(B) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \mathbb{P}(B) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \mathbb{P}(A_{n+1}) = \sum_{i=1}^{n+1} \mathbb{P}(A_i),$$

where the first equality follows from finite additivity, the first inequality follows from the induction hypothesis and the last inequality follows from monotonicity.

We now show that $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. Define $C_n = A_1 \cup \cdots \cup A_n$, for each $n \in \mathbb{N}$. Clearly, $C_1 \subseteq C_2 \subseteq \cdots$ is an increasing sequence of events and so, by continuity of probability,

$$\lim_{i\to\infty} \mathbb{P}(C_i) = \mathbb{P}(\bigcup_{i=1}^{\infty} C_i) = \mathbb{P}(\bigcup_{i=1}^{\infty} A_i).$$

On the other hand, by the previous paragraph,

$$\mathbb{P}(C_j) = \mathbb{P}(A_1 \cup \cdots \cup A_j) \leq \sum_{i=1}^{j} \mathbb{P}(A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i),$$

where in the last inequality we use the 1st axiom of probability. Since $\mathbb{P}(C_j) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for each $j \in \mathbb{N}$, we have that $\lim_{i\to\infty} \mathbb{P}(C_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ (using a well-known property of the limit of a sequence) and so, $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. □

Exercise 1.1.36. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, A_2, \ldots$ be a countable family of events. Show that if $\mathbb{P}(A_i) = 1$, for each i, then $\mathbb{P}(\bigcap_{i=1}^{\infty} A_i) = 1$. Similarly, show that if $\mathbb{P}(A_i) = 0$, for each i, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = 0$.*

As mentioned, the union bound is a simple and yet extremely useful tool in probability. We now see it in action in the so-called **probabilistic method**.

Example 1.1.37. Our motivating question is the following: Will an arbitrary group of 6 members of a social network necessarily contain a subgroup of 3 mutual friends or a subgroup of 3 mutual strangers? Perhaps surprisingly, the answer is "Yes". A group of 5 individuals, however, does not necessarily have this property. We can model and generalize this problem via a graph, an ubiquitous object in computer science and operations research. So what is a graph? Informally speaking (which is enough for us), a **graph** is a set of points, called vertices, connected by lines, called edges. The complete graph K_n is the graph on n vertices such that any two vertices are connected by an edge. A two-coloring of the edges of K_n is an assignment of colors to its edges so that each edge is colored either red or blue.

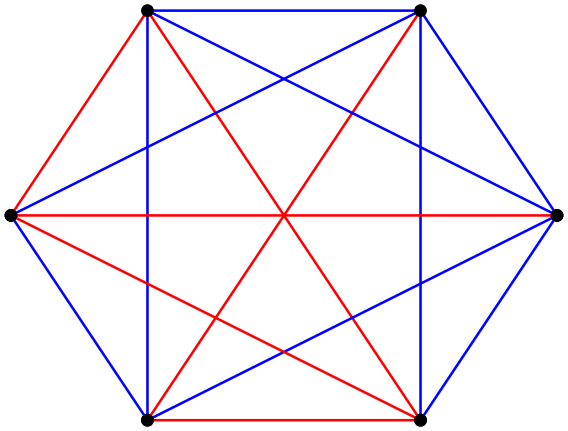


Figure 1.2: A two-coloring of K_6.

We can encode the fact of being friends by a red line and the fact of being strangers by a blue line. Therefore, generalizing our motivating question, we might ask: Does K_n always contain a monochromatic K_k i.e., a red K_k or a blue K_k, for any two-coloring? Frank Ramsey answered this question in his celebrated theorem:

Theorem 1.1.38 (Ramsey's theorem). *For any $k \geq 2$, there is a finite value of n for which any two-coloring of K_n contains a monochromatic K_k and so there is a smallest such value n, called the Ramsey number $R(k,k)$.*

We have remarked that $R(3,3) = 6$. This is in fact not difficult to show (try!) but as soon as the value of k increases, determining $R(k,k)$ has proved to be an extremely difficult problem. At the moment, we do not even know $R(5,5)$; we just know that it is between 43 and 48.

But can we say anything about how quickly Ramsey numbers grow with k? In a seminal paper from 1947 that gave birth to what is now called the probabilistic method, Erdős showed how a lower bound for $R(k,k)$ may be obtained almost effortlessly using a probabilistic argument.

Roughly speaking, the probabilistic method works as follows: Trying to prove that a structure with a certain desired property exists, one defines an appropriate probability space of structures and then shows that the desired property holds in this space with positive probability.

Consider a *random* two-coloring of K_n. The sample space is the set of all possible two-colorings of K_n. How many such colorings are there? Well, each edge can be colored either red or blue and since there are $\binom{n}{2}$ edges, we have $2^{\binom{n}{2}}$ possible colorings, where in *random* we assume that each has equal probability $2^{-\binom{n}{2}}$.

Let S be any fixed set of k vertices in K_n and let A_S be the event that S forms a monochromatic K_k. Then A_S is the union of the disjoint events $\{K_k \text{ is red}\}$ and $\{K_k \text{ is blue}\}$ and so

$$\mathbb{P}(A_S) = 2^{1-\binom{k}{2}}.$$

Let's now look at the event $\bigcup_{S:\ |S|=k} A_S$ that there is at least one monochromatic K_k. We can estimate its probability by the union bound:

$$\mathbb{P}\Big(\bigcup_{S:\ |S|=k} A_S\Big) \leq \sum_{S:\ |S|=k} \mathbb{P}(A_S) = \binom{n}{k} 2^{1-\binom{k}{2}}.$$

Therefore, if $r(k)$ denotes the largest integer n satisfying $\binom{n}{k}2^{1-\binom{k}{2}} < 1$, then

$$\mathbb{P}\Big(\bigcup_{S:\ |S|=k} A_S\Big) < 1$$

and so there must be *some* two-coloring of $K_{r(k)}$ without any monochromatic K_k i.e., $R(k,k) > r(k)$. One could in fact find an estimate for $r(k)$.

Given a sequence $A_1, A_2, \ldots \in \mathcal{F}$ of events, we are often interested in the event that infinitely many of them occur, or in its complement that only finitely many of them occur. The Borel-Cantelli lemmas (Lemma 1.1.39 and its partial converse Lemma 1.3.9 that we will address later) are examples of so-called **0-1 laws** in probability: they assert that, under some mild conditions, the two events we are interested in have probabilities either 0 or 1. We will be using them many times.

We first need some notation. Given our sequence $A_1, A_2, \ldots \in \mathcal{F}$ of events, for each m, we let $B_m = \bigcup_{n\geq m} A_n$ be the union of the events from the m-th on. Clearly, $B_1, B_2, \ldots$ is a decreasing sequence and we define

$$\limsup_n A_n = \bigcap_{m\geq 1} B_m = \bigcap_{m\geq 1} \bigcup_{n\geq m} A_n.$$

Similarly, $\liminf_n A_n$ is defined as follows. We first build the increasing sequence $C_1, C_2, \ldots$, where $C_m = \bigcap_{n\geq m} A_n$ and define

$$\liminf_n A_n = \bigcup_{m\geq 1} C_m = \bigcup_{m\geq 1} \bigcap_{n\geq m} A_n.$$

These two notions are related by De Morgan's laws: Indeed,

$$(\limsup_n A_n)^c = \Big(\bigcap_{m\geq 1}\bigcup_{n\geq m} A_n\Big)^c = \bigcup_{m\geq 1}\Big(\bigcup_{n\geq m} A_n\Big)^c = \bigcup_{m\geq 1}\Big(\bigcap_{n\geq m} A_n^c\Big) = \liminf_n A_n^c. \tag{1.1}$$

Observe that $\liminf_n A_n$ and $\limsup_n A_n$ are events. But which events? Well, an outcome $\omega \in \Omega$ lies in $\limsup_n A_n$ iff it lies in each of the sets B_m. But then ω lies in infinitely many A_n's, or else there exists an index M such that $\omega \notin A_n$ for each $n \geq M$, a contradiction. Therefore, $\limsup_n A_n$ is nothing but the event "infinitely many of the A_n's occur" ("A_n i.o." for short, where i.o. stands for infinitely often). On the other hand, an outcome $\omega \in \Omega$ lies in $\liminf_n A_n$ iff there exists an m such that $\omega \in A_n$ for each $n \geq m$ i.e., ω lies in all but finitely many A_n's. Therefore, $\liminf_n A_n$ is nothing but the event "all but finitely many of the A_n's occur" ("A_n a.a." for short, where a.a. stands for almost always).

The first Borel-Cantelli lemma says that whenever the probabilities of the events A_n decay fast enough, it is (almost surely) impossible for the events to occur infinitely often.

Lemma 1.1.39 (First Borel-Cantelli). *Let $A_1, A_2, \ldots \in \mathcal{F}$ be events. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ converges, then*

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}(\{A_n \text{ i.o.}\}) = 0.$$

In other words, with probability one only finitely many of the events A_n occur.

Proof. Let $B_m = \bigcup_{n\geq m} A_n$. By the union bound,

$$\mathbb{P}(B_m) = \mathbb{P}\Big(\bigcup_{n\geq m} A_n\Big) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n).$$

Since the series $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ converges, we have that $\sum_{n=m}^{\infty} \mathbb{P}(A_n)$ tends to 0 as $m \to \infty$ (tails of a convergent series vanish). But then, since $B_1, B_2, \ldots$ is a decreasing sequence, continuity of probability implies that

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}\Big(\bigcap_{m\geq 1} B_m\Big) = \lim_{m\to\infty} \mathbb{P}(B_m) = 0,$$

as claimed. □

Example 1.1.40. Consider an experiment in which a coin is tossed many times. Suppose that the probability of the event A_n of obtaining heads at the n-th toss is $1/n^2$. Then $\sum_n \mathbb{P}(A_n)$ converges and so the first Borel-Cantelli lemma implies that, almost surely, only finitely many heads will occur.

Exercise 1.1.41. *Let $A_1, A_2, \ldots \in \mathcal{F}$ be events. Show that $\liminf_n A_n \subseteq \limsup_n A_n$.*

1.2 Conditional probability

An experiment is repeated N times. On each trial we observe the occurences or non-occurences of two events A and B. Suppose we are interested only in the outcomes for which B occurs; all other trials are disregarded. The proportion of times that A occurs in this smaller collection of trials is $N(A\cap B)/N(B)$ and

$$\frac{N(A\cap B)}{N(B)} = \frac{N(A\cap B)/N}{N(B)/N}.$$

As these ratios can be thought as approximations for the probabilities, the probability that A occurs given that B occurs should be intuitively $\mathbb{P}(A\cap B)/\mathbb{P}(B)$.

Definition 1.2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The **conditional probability** that A occurs given that B occurs is the value

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We stress the fact that this is a definition. The next result justifies the term conditional probability:

Lemma 1.2.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let F be such that $\mathbb{P}(F) > 0$. The function $P\colon \mathcal{F} \to \mathbb{R}$ defined by $P(A) = \mathbb{P}(A|F)$ is a probability measure.*

Lemma 1.2.2 implies that we can apply all the tools developed so far to conditional probabilities. It is very instructive to prove it:

Exercise 1.2.3. *Prove Lemma 1.2.2.*

In many situations it is natural to assign values to some conditional probabilities and, from them, deduce the values of non-conditional probabilities.

Example 1.2.4. A student can't decide whether to study history or literature. If he takes literature, he will pass with probability $1/2$; if he takes history, he will pass with probability $1/3$. He made his decision based on a coin toss. What is the probability that he opted for history and passed the exam?

As a sample space we take $\{\text{history, literature}\} \times \{\text{pass, fail}\}$. If A is the event that he passed, then $A = \{\text{history, literature}\} \times \{\text{pass}\}$. If B denotes the event that he opted for history, then $\{\text{history}\} \times \{\text{pass, fail}\}$. We have

$$\mathbb{P}(B) = \mathbb{P}(B^c) = \frac{1}{2}, \qquad \mathbb{P}(A|B) = \frac{1}{3}, \qquad \mathbb{P}(A|B^c) = \frac{1}{2}$$

and so $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = 1/6$. Notice that making the sample space explicit was in fact not crucial in this case, as often happens with conditional probabilities.

Definition 1.2.5. Given a countable infinite collection of events $B_1, B_2, \ldots$, we say that the collection is a **partition** of Ω if $B_i \cap B_j = \varnothing$ for each $i \neq j$ and $\bigcup_{i=1}^{\infty} B_i = \Omega$. The same definition applies mutatis mutandis in the case of a finite collection.

Lemma 1.2.6 (Law of total probability). *Given a partition $B_1, B_2, \ldots$ of Ω such that $\mathbb{P}(B_i) > 0$ for each i, then*

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

A similar result holds in case the collection $B_1, B_2, \ldots, B_n$ is finite.

Proof. Exercise! □

The law of total probability is typically used as follows. Suppose we want to compute the probability that A occurs. Let B be another arbitrary event with $0 < \mathbb{P}(B) < 1$. There are two scenarios: either B or B^c occurs. If we know the probability of the two scenarios and the probability of A conditioned on each of them, then we can compute the probability of A.

Lemma 1.2.7 (Bayes' law). *Let A and B be two events such that $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)}{\mathbb{P}(B)} \cdot \mathbb{P}(A).$$

Proof. Exercise! □

Bayes' law tells how to update the estimate of the probability of A when new evidence restricts the sample space to B. The ratio $\mathbb{P}(B|A)/\mathbb{P}(B)$ determines "how compelling the new evidence is".

Combining Bayes' law with the law of total probability we obtain that, if $B_1, B_2, \dots$ is a partition of Ω such that $\mathbb{P}(B_i) > 0$ for each i, then

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{i=1}^{\infty}\mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

Example 1.2.8. Consider a lab screen for a certain virus. A person that carries the virus is screened positive in only 95% of the cases (5% chance of false negative). A person who does not carry the virus is screened positive in 1% of the cases (1% chance of false positive). Given that 0.5% of the population carries the virus, what is the probability that a person who has been screened positive is actually a carrier?

We take $\Omega = \{\text{carrier, not carrier}\} \times \{+, -\}$. Let A be the event "the person is a carrier" i.e., $A = \{\text{carrier}\}\times\{+, -\}$, and let B be the event "the person was screened positive" i.e., $B = \{\text{carrier, not carrier}\}\times\{+\}$. We are given the following information

$$\mathbb{P}(A) = 0.005 \quad \mathbb{P}(B|A) = 0.95 \quad \mathbb{P}(B|A^c) = 0.01.$$

Therefore, taking A, A^c as our partition of Ω, we have that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)} = \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.01 \cdot 0.995} \approx \frac{1}{3}.$$

Example 1.2.9. A random number N of dice is thrown. Let A_i be the event that $N = i$ and suppose that $\mathbb{P}(A_i) = 1/2^i$ $(i \geq 1)$. The sum of the scores is S. Compute $\mathbb{P}(N = 2|S = 4)$.

By Bayes' law,

$$\mathbb{P}(N = 2|S = 4) = \frac{\mathbb{P}(S = 4|N = 2)\mathbb{P}(N = 2)}{\mathbb{P}(S = 4)}.$$

The (countable) family of events $N = 1, N = 2, N = 3, \dots$ is a partition of Ω such that $\mathbb{P}(N = i) = 1/2^i > 0$. Therefore, by the law of total probability,

$$\mathbb{P}(S = 4) = \sum_{i=1}^{\infty}\mathbb{P}(S = 4|N = i)\mathbb{P}(N = i).$$

However, only the first four terms of the sum are non-zero. Indeed, if $i \geq 5$, then $\mathbb{P}(S = 4|N = i) = 0$. Therefore, the desired probability is

$$\mathbb{P}(N = 2|S = 4) = \frac{\mathbb{P}(S = 4|N = 2)\mathbb{P}(N = 2)}{\mathbb{P}(S = 4)} = \frac{\mathbb{P}(S = 4|N = 2)\mathbb{P}(N = 2)}{\sum_{i=1}^{4}\mathbb{P}(S = 4|N = i)\mathbb{P}(N = i)}.$$

We are then left to compute $\mathbb{P}(S = 4|N = i)$, for $i \in \{1, 2, 3, 4\}$. Let's consider $\mathbb{P}(S = 4|N = 3)$, the other cases being similar. This is the probability of getting a sum of 4 by throwing 3 dice. As usual, label the dice $1, 2, 3$ and let x_i be the number on die i. There are 6^3 possible outcomes and we need to count how many triples (x_1, x_2, x_3) are such that $x_1 + x_2 + x_3 = 4$. Since each x_i is at least 1, it is easy to see there are exactly 3 such triples, namely $(1, 1, 2), (1, 2, 1), (2, 1, 1)$. Therefore, $\mathbb{P}(S = 4|N = 3) = 3/6^3$.

More generally, we might be interested in the number of positive integer solutions to

$$x_1 + x_2 + \cdots + x_r = n$$

for some fixed $r \leq n$. Write n as $1 + 1 + \cdots + 1$, where there are n 1's and $n - 1$ pluses. To decompose n into r summands we only need to choose $r - 1$ pluses out of the $n - 1$ and this can be done in $\binom{n-1}{r-1}$ ways. What about if we are looking for positive solutions such that $x_i \leq 6$ for each i? How can we adapt the reasoning above?

Exercise 1.2.10. *What is the number of nonnegative integer solutions to $x_1 + x_2 + \cdots + x_m = n$?*

Exercise 1.2.11. *Consider n indistinguishable balls randomly distributed in m boxes. What is the probability that exactly k boxes remain empty?*

Exercise 1.2.12. *An urn contains b blue balls and c cyan balls. A ball is drawn at random, its color noted and it is returned to the urn together with d further balls of the same color. The process is repeated indefinitely.*

- *Compute the probability that the second ball drawn is cyan.*
- *Compute the probability that the first ball drawn is cyan given that the second ball drawn is cyan.*

1.3 Independence

In general, the occurence of some event B changes the probability that a certain event A occurs, the original $\mathbb{P}(A)$ being replaced by $\mathbb{P}(A|B)$. If the probability remains unchanged i.e., $\mathbb{P}(A|B) = \mathbb{P}(A)$, then we call A and B independent. Since in order to talk about $\mathbb{P}(A|B)$ we need $\mathbb{P}(B) > 0$, we give the following more general definition which agrees with this special case.

Definition 1.3.1. The events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. More generally, a family of events $\{A_i : i \in I\}$ is **independent** if $\mathbb{P}(\bigcap_{i\in J} A_i) = \prod_{i\in J} \mathbb{P}(A_i)$ for each finite subset J of I. A family $\{A_i : i \in I\}$ is **pairwise independent** if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for each $i \neq j$.

If the occurrence of two events is governed by distinct and noninteracting processes, such events will turn out to be independent (this will be our modelling assumption).

Independence is not easily visualized in terms of the sample space. A common first thought is that two events are independent if they are disjoint, but in fact the opposite is true: two disjoint events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ are never independent as $\mathbb{P}(A \cap B) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)$.

Example 1.3.2. Roll two dice and let A be the event "the first die is 4". Let B_1 be the event "the second die is 2". This satisfies our intuitive notion of independence since the outcome of the first dice roll has nothing to do with that of the second. To check independence, note that $\mathbb{P}(B_1) = 1/6 = \mathbb{P}(A)$ and $\mathbb{P}(A \cap B_1) = 1/36$.

Let B_2 be the event "the sum of the two dice is 3". Since $A \cap B_2 = \varnothing$, we have that $\mathbb{P}(A \cap B_2) = 0 < \mathbb{P}(A)\mathbb{P}(B_2)$ and so the events cannot be independent.

Let B_3 be the event "the sum of the two dice is 7". This time, A and B_3 are independent. Indeed, we have that $\mathbb{P}(B_3) = 6/36$ and $\mathbb{P}(A \cap B_3) = 1/36$.

Let B_4 be the event "the sum of the two dice is 9". We have that A and B_4 are not independent. Indeed, $\mathbb{P}(A \cap B_4) = 1/36$ but $\mathbb{P}(A)\mathbb{P}(B_4) = 1/6 \cdot 4/36$.

Remark 1.3.3. Independence is stronger than pairwise independence: Any independent family is clearly pairwise independent but the converse is not true. Indeed, toss two coins and consider the events "first coin gives H", "second coin gives H", "resulting number of heads is odd". They form a family which is pairwise independent but not independent.

Example 1.3.4 (Bernoulli trials). If an experiment involves a sequence of independent but identical stages, we say that we have a sequence of independent trials. If there are only two possible results at each stage, we say that we have a sequence of independent Bernoulli trials.

Consider an experiment that consists of n independent tosses of a biased coin, in which the probability of H is p, for some $p \in [0, 1]$. What is the probability of getting exactly k heads?

Let A_i be the event "the i-th toss is H". Independence means that the events $A_1, A_2, \ldots, A_n$ are independent (the occurrence of any of them is governed by distinct and noninteracting processes). Consider for example the outcome in which we have k heads followed by $n-k$ tails i.e., the elementary event $A_1 \cap A_2 \cap \cdots \cap A_k \cap A_{k+1}^c \cap \cdots \cap A_n^c$. Intuitively, the family $A_1, A_2, \ldots, A_k, A_{k+1}^c, \ldots, A_n^c$ is independent (see below for a formal proof) and so we have that

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_k \cap A_{k+1}^c \cap \cdots \cap A_n^c) = \mathbb{P}(A_1)\mathbb{P}(A_2)\cdots\mathbb{P}(A_k)\mathbb{P}(A_{k+1}^c)\cdots\mathbb{P}(A_n^c) = p^k(1-p)^{n-k}.$$

Moreover, any other elementary event consisting of k heads and $n-k$ tails will have the same probability. Therefore, by additivity, it is enough to count such elementary events. This is equivalent to counting the number of subsets of size k (the trials giving head) of a set of size n (the set of all trials). This number is $\binom{n}{k}$ and so the probability of getting exactly k heads is

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

As mentioned above, we now show that if $A_1, A_2 \ldots, A_n$ is an independent family then, replacing A_i by A_i^c for some i, still gives an independent family. By possibly repeating the argument, it is enough to show this for one value of i, say $i = n$. Therefore, we show the following: if $A_1, A_2 \ldots, A_n$ is an independent family, then $A_1, A_2 \ldots, A_{n-1}, A_n^c$ is an independent family. Consider a subset J of $A_1, A_2 \ldots, A_{n-1}, A_n^c$. If $A_n^c \notin J$, then $\mathbb{P}(\bigcap_{A \in J} A) = \prod_{A \in J} \mathbb{P}(A)$ by assumption. If $A_n^c \in J$ then, by possibly relabelling indices we have that J is of the form $J = \{A_1, \ldots, A_\ell, A_n^c\}$ for some $\ell \in \{1, \ldots, n-1\}$. Letting $B = A_1 \cap \cdots \cap A_\ell$, we have that

$$\begin{aligned}
\mathbb{P}(A_1 \cap \cdots \cap A_\ell \cap A_n^c) &= \mathbb{P}(B \cap A_n^c) \\
&= \mathbb{P}(B \setminus (B \cap A_n)) \\
&= \mathbb{P}(B) - \mathbb{P}(B \cap A_n) \\
&= \mathbb{P}(B) - \mathbb{P}(A_1 \cap \cdots \cap A_\ell \cap A_n) \\
&= \mathbb{P}(B) - \mathbb{P}(A_1)\cdots\mathbb{P}(A_\ell)\mathbb{P}(A_n) \\
&= \mathbb{P}(B) - \mathbb{P}(B)\mathbb{P}(A_n) \\
&= \mathbb{P}(B)(1 - \mathbb{P}(A_n)) \\
&= \mathbb{P}(B)\mathbb{P}(A_n^c) \\
&= \mathbb{P}(A_1)\cdots\mathbb{P}(A_\ell)\mathbb{P}(A_n^c),
\end{aligned}$$

which is what we wanted to show.

Exercise 1.3.5. *Suppose A and B are events and the probability of B is either zero or one. Show that A and B are independent.*

Example 1.3.6 (Gambler's ruin). A man wants to buy a car at a cost of N units of money. He starts with k units, for some $0 < k < N$ and tries to win the remainder by the following gamble with his bank manager. He tosses a fair coin repeatedly and independently. If H comes up, then the manager pays him one unit. If T comes up, then he pays the manager one unit. He plays the game until one of two events occurs: either he runs out of money and is bankrupted or he wins enough to buy the car. What is the probability that he is ultimately bankrupted?

We want to compute the probability of the event A_k "bankrupted if starting with k units". Notice that $\mathbb{P}(A_0) = 1$ and $\mathbb{P}(A_N) = 0$. Let B be the event "first toss is H". The law of total probability tells us that

$$\mathbb{P}(A_k) = \mathbb{P}(A_k|B)\mathbb{P}(B) + \mathbb{P}(A_k|B^c)\mathbb{P}(B^c).$$

But if the first toss is H, he has $k+1$ units and if the first toss is T, he has $k-1$ units. Since the tosses are independent, we have that $\mathbb{P}(A_k|B) = \mathbb{P}(A_{k+1})$ and $\mathbb{P}(A_k|B^c) = \mathbb{P}(A_{k-1})$. Therefore, letting $p_k = \mathbb{P}(A_k)$, we have that

$$p_k = \frac{1}{2}(p_{k+1} + p_{k-1}), \tag{1.2}$$

with $p_0 = 1$ and $p_N = 0$. We want to compute the value of p_k by using this recurrence relation together with the two "boundary conditions". Observe first that, by Equation (1.2), the difference between consecutive p_k's is always the same: $p_k - p_{k-1} = p_{k+1} - p_k$. Letting $b_k = p_k - p_{k-1}$ this common value, we have that $b_k = b_1$ and so

$$p_k = b_1 + p_{k-1} = b_1 + (b_1 + p_{k-2}) = \cdots = kb_1 + p_0.$$

Substituting N for k, we get $0 = p_N = Nb_1 + p_0 = Nb_1 + 1$, from which $b_1 = -1/N$ and so $p_k = 1 - k/N$.

Notice that, for each fixed k, the probability p_k he is bankrupted starting with k units tends to 1 as $N \to \infty$.

Exercise 1.3.7. *In this exercise we consider gambler's ruin in the case the coin has probability p of getting heads and probability q of getting tails, where $p + q = 1$ and $p \neq 1/2$. Using the previous notation, proceed as follows (each step is deduced from the previous):*

- *Show that $p_k = p \cdot p_{k+1} + q \cdot p_{k-1}$;*
- *Deduce that*
$$p_{k+1} - p_k = \frac{q}{p} \cdot (p_k - p_{k-1});$$
- *Deduce that*
$$b_k = \left(\frac{q}{p}\right)^{k-1} \cdot b_1;$$
- *Conclude that*
$$p_k = 1 - \frac{1 - \left(\frac{q}{p}\right)^k}{1 - \left(\frac{q}{p}\right)^N}.$$

Remark 1.3.8. Let's make a comment about the previous exercise in the realistic situation that our gambler plays against a gambling machine. Gambling machines in most countries permit by law a certain degree of "unfairness" by taking $p < 1/2$. This allows the house to make an income. Suppose that $p = 0.47$ and that the gambler starts with 10 units and aims at reaching 20 units. The probability he is bankrupted turns out to be roughly 77% (check it yourself!). Therefore, a "slightly unfair" game at each round can become devastatingly unfair in the long run.

We now state and prove the following partial converse to the first Borel-Cantelli lemma. It asserts that if the probabilities of the events A_n do not decay fast and if the events are in addition independent, then they must (almost surely) occur infinitely often.

Lemma 1.3.9 (Second Borel-Cantelli). *Let $A_1, A_2, \ldots \in \mathcal{F}$ be events. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ diverges and the events are independent, then*

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}(\{A_n \text{ i.o.}\}) = 1.$$

In other words, the events A_n occur infinitely often with probability one.

Proof. We have seen in Equation (1.1) that $(\limsup_n A_n)^c = \liminf_n A_n^c$. Therefore, $\limsup_n A_n = (\liminf_n A_n^c)^c$. Let now $C_m = \bigcap_{n\geq m} A_n^c$ so that $C_1, C_2, \ldots$ is an increasing sequence and, by definition, $\liminf_n A_n^c = \bigcup_{m\geq 1} C_m$. Continuity of probability implies that $\mathbb{P}(\liminf_n A_n^c) = \lim_{m\to\infty} \mathbb{P}(C_m)$. Combining these, we obtain

$$\mathbb{P}(\limsup_n A_n) = 1 - \mathbb{P}(\liminf_n A_n^c) = 1 - \lim_{m\to\infty} \mathbb{P}(\bigcap_{n\geq m} A_n^c). \tag{1.3}$$

To use independence, we need finite intersections: $\mathbb{P}(\bigcap_{n=m}^{\ell} A_n^c) = \prod_{n=m}^{\ell} \mathbb{P}(A_n^c)$. But notice that, $\bigcap_{n=m}^{m} A_n^c, \bigcap_{n=m}^{m+1} A_n^c, \bigcap_{n=m}^{m+2} A_n^c, \ldots$ is a decreasing sequence and so, by continuity of probability,

$$\mathbb{P}(\bigcap_{n\geq m} A_n^c) = \lim_{\ell\to\infty} \mathbb{P}(\bigcap_{n=m}^{\ell} A_n^c) = \lim_{\ell\to\infty} \prod_{n=m}^{\ell} \mathbb{P}(A_n^c) = \lim_{\ell\to\infty} \prod_{n=m}^{\ell} (1 - \mathbb{P}(A_n)). \tag{1.4}$$

We now use the inequality $1 - p \leq e^{-p}$ (Exercise 1.3.11) to estimate the last product:

$$\prod_{n=m}^{\ell} (1 - \mathbb{P}(A_n)) \leq \prod_{n=m}^{\ell} e^{-\mathbb{P}(A_n)} = \exp(-\sum_{n=m}^{\ell} \mathbb{P}(A_n)).$$

But by assumption, $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ diverges and so $\exp(-\sum_{n=m}^{\ell} \mathbb{P}(A_n)) \to 0$ as $\ell \to \infty$. By Equation (1.4), we obtain $\mathbb{P}(\bigcap_{n\geq m} A_n^c) = 0$ and plugging into Equation (1.3), $\mathbb{P}(\limsup_n A_n) = 1$, as desired. □

Observation 1.3.10. *Notice that, if we drop the independence assumption in Lemma 1.3.9, the statement fails to hold. Indeed, let A be an arbitrary event with $0 < \mathbb{P}(A) < 1$ and let $A_n = A$ for each $n \geq 1$. Clearly, the family $A_1, A_2, \ldots$ is not independent. We have that $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ diverges but $\mathbb{P}(\{A_n \text{ i.o.}\}) = \mathbb{P}(A) < 1$.*

We can combine the first and second Borel-Cantelli lemma in order to highlight the 0-1 property of the event $\{A_n \text{ i.o.}\}$:

Let $A_1, A_2, \ldots \in \mathcal{F}$ be a family of independent events. Then

$$\mathbb{P}(\{A_n \text{ i.o.}\}) = \begin{cases} 0 & \text{if } \sum_{n=1}^{\infty} \mathbb{P}(A_n) \text{ converges;} \\ 1 & \text{if } \sum_{n=1}^{\infty} \mathbb{P}(A_n) \text{ diverges.} \end{cases}$$

Exercise 1.3.11. *Show that $1 + x \leq e^x$, for each $x \in \mathbb{R}$.*

Example 1.3.12. Consider a sequence of independent tosses of a fair coin. What is the probability that there are infinitely many heads? Let A_n be the event that the n-th toss is heads. Then $\mathbb{P}(A_n) = 1/2$ for each n and so $\sum_n \mathbb{P}(A_n)$ diverges. Since the A_n's are independent, the second Borel-Cantelli lemma then implies that the desired probability is 1.

Example 1.3.13. Consider again a sequence of independent tosses of a fair coin. What is the probability that a run of 1000^{1000} heads occurs? Let A_n be the event that a run of 1000^{1000} heads occur starting from the n-th toss. We have that $\mathbb{P}(A_n) = \frac{1}{2^{1000^{1000}}}$. Moreover, the family $A_1, A_{1+1000^{1000}}, A_{1+2\cdot 1000^{1000}}, \ldots$ is independent. Since $\sum_n \mathbb{P}(A_n)$ diverges, the second Borel-Cantelli lemma implies that the probability that infinitely many of the A_n's occur is 1 and so the desired probability is 1.

Example 1.3.14. Similarly to the previous, a monkey hitting keys at random on a keyboard for an infinite amount of time will almost surely i.e., with probability one, type the seven volumes of Proust's *In search of lost time*.

Exercise 1.3.15. *Consider a sequence of rolling of a fair die. What is the probability that a run containing the numbers* $1, 6, 5, 4$ *occurs infinitely often?*

Exercise 1.3.16. *We perform infinitely many independent experiments. The* n*-th one is successful with probability* $n^{-\alpha}$ *and fails with probability* $1 - n^{-\alpha}$*, for some* $0 < \alpha < 1$*. What is the probability that we see* k *consecutive successes infinitely often?* *Hint: The answer depends on* k*.*

1.4 Random variables

Most of the times we are not interested in an experiment itself but rather in some consequence of its random outcome. A random variable can be thought of as a numerical "summary" of a certain aspect of the experiment. It is nothing but a function from the sample space Ω to the real line $\mathbb{R}$, where the "random" in the name comes from the experiment:

1. Chance determines the random outcome $\omega \in \Omega$;
2. The outcome ω determines a certain quantity of interest.

In other words, a random variable X represents an unknown quantity that varies with the outcome of a random event. Before the random event, we know which values X could possibly assume, but we do not know which one it will take until the random event happens. The terminology may appear confusing: a variable is a function? This is because the words "random variable" were in use long before the connection between probability and analysis was discovered.

Example 1.4.1. A fair coin is tossed twice. We can take $\Omega = \{HH, HT, TH, TT\}$. For any outcome $\omega \in \Omega$, we let $X(\omega)$ be the number of heads in the outcome. Therefore, $X(HH) = 2$, $X(HT) = X(TH) = 1$ and $X(TT) = 0$.

Crucially, the function $X \colon \Omega \to \mathbb{R}$ has to be sufficiently well-behaved so that we can talk about probabilities with which X assumes certain values:

Definition 1.4.2. A **random variable** is a function $X \colon \Omega \to \mathbb{R}$ such that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$.

Remark 1.4.3. Notice that whenever we talk about a random variable we implicitly assume an underlying probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

Example 1.4.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where $\mathcal{F}$ is the power set of Ω. Then obviously any function $X \colon \Omega \to \mathbb{R}$ is a random variable. Recall that if Ω is countable, we can always take its power set as our σ-field $\mathcal{F}$.

In general, the numerical value of a random variable is more likely to lie in certain subsets of $\mathbb{R}$, depending on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the function X itself. The following notion describes the distribution of the likelihoods of possible values of X. As mentioned above, it is the reason behind the technical requirement $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ in the definition of random variable.

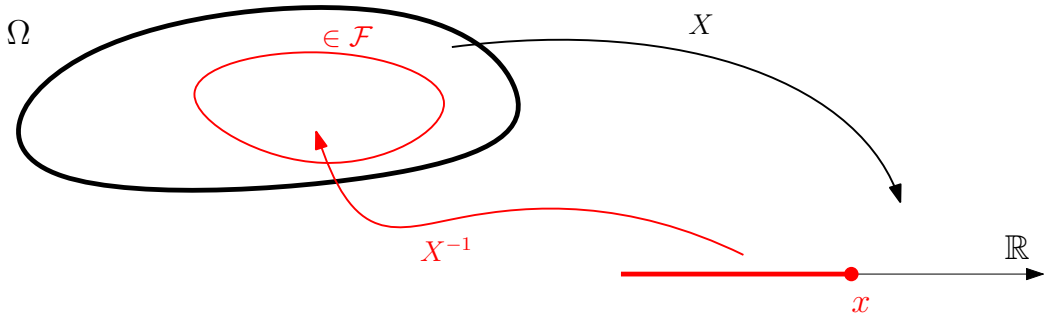


Figure 1.3: Visualization of a random variable X and the property $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$.

Definition 1.4.5. The **distribution function** of a random variable X is the function $F_X : \mathbb{R} \to [0,1]$ given by $F_X(x) = \mathbb{P}(X \leq x)$. Here and in the following we use the shorthands $X \leq x$ or $\{X \leq x\}$ for the event $\{\omega \in \Omega : X(\omega) \leq x\}$. We will also drop the subscript X in F_X when it is clear to which random variable we are referring.

Example 1.4.6. The distribution function of the random variable in Example 1.4.1 is given by:

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0; \\ 1/4 & \text{if } 0 \leq x < 1; \\ 3/4 & \text{if } 1 \leq x < 2; \\ 1 & \text{if } x \geq 2. \end{cases}$$

We are interested in two types of random variables:

Definition 1.4.7. The random variable X is **discrete** if it takes values in some countable subset of $\mathbb{R}$. The **probability mass function (pmf)** of a discrete random variable X is the function $f_X : \mathbb{R} \to [0,1]$ given by $f_X(x) = \mathbb{P}(X = x)$.

The random variable X is **continuous** if its distribution function can be expressed as

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\, \mathrm{d}u$$

for some integrable function $f_X : \mathbb{R} \to [0, \infty)$ called the **probability density function (pdf)** of X.

The name continuous comes from the fact (a generalization of the Fundamental theorem of calculus) that the function F_X is continuous. This is in sharp contrast to discrete random variables, whose distribution functions are never continuous (only right-continuous as we will see shortly).

Remark 1.4.8. The definition of random variable requires that $\{X \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. But what about $\{X = x\}$? It turns out that, since $\{x\}$ is a Borel set (Example 1.1.16), Theorem 1.4.28 will indeed imply that $\{X = x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$ and so it makes sense to write $\mathbb{P}(X = x)$.

Remark 1.4.9. If the distribution function $F_X : \mathbb{R} \to [0,1]$ of a continuous random variable X is differentiable at some $x \in \mathbb{R}$, the corresponding value $f_X(x)$ can be found by taking the derivative of F_X at x.

Remark 1.4.10. Observe that, knowing the probability mass function of a discrete random variable, we can immediately compute its distribution function using countable additivity. Indeed,

$$\{X \leq x\} = \bigcup_{k:\ k \leq x \text{ and } k \in \mathrm{Im}(X)} \{X = k\},$$

where the union is countable as $\mathrm{Im}(X)$ is countable (here $\mathrm{Im}(X)$ denotes the image of X i.e., the values taken by X).

Example 1.4.11 (Uniform random variable). The random variable X is uniform on $[a, b]$ if it has distribution function

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < a; \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b; \\ 1 & \text{if } x > b. \end{cases}$$

X is continuous, as it admits density function given by

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b; \\ 0 & \text{otherwise.} \end{cases}$$

The idea here is that we are picking a value "at random" from $[a, b]$ (values outside the interval are impossible, and all those inside have the same probability density). Therefore, the probability that $X \leq c$ for some $c \in [a, b]$ should intuitively be $\frac{c-a}{b-a}$.

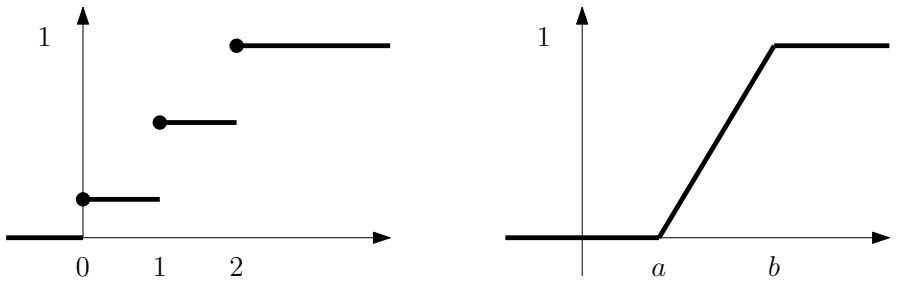


Figure 1.4: Distribution function of the discrete random variable in Example 1.4.1 (left) and of the uniform random variable on $[a, b]$ (right).

In the following we list several important discrete random variables.

Example 1.4.12 (Constant random variable). Let $c \in \mathbb{R}$ and let $X \colon \Omega \to \mathbb{R}$ be given by $X(\omega) = c$ for each $\omega \in \Omega$. This is a random variable with pmf

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 0 & \text{if } x \neq c; \\ 1 & \text{if } x = c. \end{cases}$$

and distribution function

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < c; \\ 1 & \text{if } x \geq c. \end{cases}$$

Example 1.4.13 (Bernoulli random variable). A coin is tossed once and let p be the probability of H. Let X be 1 if the toss gives H and 0 otherwise. This is a random variable with pmf

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 1-p & \text{if } x = 0; \\ p & \text{if } x = 1; \\ 0 & \text{otherwise.} \end{cases}$$

and distribution function

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0; \\ 1-p & \text{if } 0 \leq x < 1; \\ 1 & \text{if } x \geq 1. \end{cases}$$

Example 1.4.14 (Binomial random variable). A coin is tossed n times. At each toss, the coin gives H with probability p, independently of prior tosses. Let X be the number of heads in the n-toss sequence. We refer to X as a binomial random variable with parameters (n, p). We essentially already computed its pmf (see Example 1.3.4). It is given by

$$f_X(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

for $k \in \{0, 1, \ldots, n\}$.

Example 1.4.15 (Geometric random variable). Suppose that we repeatedly and independently toss a coin with probability of getting H being p. The geometric random variable is the number X of tosses needed for a head to come up for the first time. Its pmf is given by

$$f_X(k) = \mathbb{P}(X = k) = (1-p)^{k-1} p,$$

for $k = 1, 2, \ldots$

Example 1.4.16 (Poisson random variable). A random variable X is said to be Poisson if it has pmf given by

$$f_X(k) = \mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!},$$

for some parameter $\lambda > 0$ and $k = 0, 1, 2, \ldots$.

How does a Poisson random variable with parameter λ arise? It turns out that it is a limit of binomial random variables with parameters $(n, \lambda/n)$. Indeed,

$$\binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k}$$

and for any fixed k, taking the limit $n \to \infty$, we have that the quantity above tends to $e^{-\lambda} \frac{\lambda^k}{k!}$. In other words, a Poisson random variable with parameter λ approximates a binomial random variable with parameters (n, p) provided $\lambda = np$, n is large and p is small. It appears abundantly in life, for example, when counting the number of radio-active decays in a unit of time or the number of cars involved in accidents in a city on a given day.

Exercise 1.4.17. *An airplane engine breaks down during a flight with probability $1 - p$. An airplane lands safely only if at least half of its engines are functioning upon landing. What is preferable: a two-engine airplane or a four-engine airplane?*

Exercise 1.4.18. *There are n white balls and m black balls in an urn. Each time, we take out one ball (with replacement) until we have a black ball. What is the probability that we need at least k trials?*

One might worry that the condition in the definition of a random variable is too stringent. Luckily, this is not the case: almost all reasonable functions turn out to be random variables. We now provide several ways of construting new random variables.

Since real numbers have addition and multiplication, we can perform such operations on real-valued functions pointwise. If $f_1, f_2 \colon \Omega \to \mathbb{R}$ are two functions, then the pointwise sum $f_1 + f_2 \colon \Omega \to \mathbb{R}$ is defined by $(f_1 + f_2)(\omega) = f_1(\omega) + f_2(\omega)$ for each $\omega \in \Omega$. We define the pointwise product $f_1 f_2$ and the pointwise scalar λf_1 similarly.

Proposition 1.4.19. *Let X and Y be random variables and let $\lambda \in \mathbb{R}$. The following are random variables:*

(a) λX*;*

(b) $X + Y$*;*

(c) XY*;*

(d) $Z(\omega) = \begin{cases} Y(\omega)/X(\omega) & \text{if } X(\omega) \neq 0; \\ 0 & \text{if } X(\omega) = 0. \end{cases}$

(e) $\max\{X, Y\}$*;*

(f) $\min\{X, Y\}$.

Proof. We prove only (b) as the other proofs are similar. We need to show that $\{\omega : X(\omega) + Y(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. Since σ-fields are closed under complementation, it is then enough to show that $\{\omega : X(\omega) + Y(\omega) > x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. Observe that, since there exists a rational number between any two real numbers, we have

$$\{\omega : X(\omega) + Y(\omega) > x\} = \bigcup_{r \in \mathbb{Q}} \{\omega : X(\omega) > r, Y(\omega) > x - r\}.$$

But for fixed r and x, $\{\omega : X(\omega) > r\} \in \mathcal{F}$ and $\{\omega : Y(\omega) > x - r\} \in \mathcal{F}$, as X and Y are random variables. Therefore their intersection $\{\omega : X(\omega) > r, Y(\omega) > x - r\}$ belongs to $\mathcal{F}$ and hence the countable union $\bigcup_{r \in \mathbb{Q}} \{\omega : X(\omega) > r, Y(\omega) > x - r\}$ belongs to $\mathcal{F}$ as well. □

Example 1.4.20. A binomial random variable with parameters (n, p) is a sum of n Bernoulli random variables each with parameter p.

Exercise 1.4.21. *Show that if X and Y are random variables, then $\max\{X, Y\}$ and $\min\{X, Y\}$ are.*

But we can also compose functions. Given a random variable $X \colon \Omega \to \mathbb{R}$, and a function $g \colon \mathbb{R} \to \mathbb{R}$, we can consider $Y = g(X)$ i.e., the function $Y \colon \Omega \to \mathbb{R}$ mapping $\omega \in \Omega$ to $g(X(\omega)) \in \mathbb{R}$. It turns out that if g is continuous, we obtain another random variable:

Theorem 1.4.22. *Let X be a random variable and $g \colon \mathbb{R} \to \mathbb{R}$ a continuous function. Then $Y = g(X)$ is a random variable.*

Example 1.4.23. For a random variable X, all the following are random variables: $\sin(X), e^X, \log(X), X^n$.

We now observe some properties that the distribution function of a generic random variable satisfies. Hence the following result holds for both discrete and continuous random variables.

Lemma 1.4.24. *The distribution function F of a random variable X satisfies the following properties:*

(a) It is monotonically increasing i.e., if $x \leq y$, then $F(x) \leq F(y)$.

(b) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

(c) It is right-continuous i.e., $F(x+h) \to F(x)$ as h tends to 0 from the positive side.

Proof. (a) If $x \leq y$, then $\{\omega : X(\omega) \leq x\} \subseteq \{\omega : X(\omega) \leq y)\}$ and so

$$F(x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}) \leq \mathbb{P}(\{\omega : X(\omega) \leq y\}) = F(y)$$

by monotonicity of probability. Notice that this also implies that the limits in (b) exist, as $0 \leq F(x) \leq 1$ is bounded.

(b) Consider events of the form $B_{-n} = \{\omega : X(\omega) \leq -n\}$, for $n \in \mathbb{N}$. We have that $B_{-1} \supseteq B_{-2} \supseteq \cdots$ and $\bigcap_{n=1}^{\infty} B_{-n} = \varnothing$. Therefore, by continuity of probability, $F(-n) = \mathbb{P}(B_{-n}) \to \mathbb{P}(\varnothing) = 0$ as $n \to \infty$. The conclusion follows (why?). The other limit is left as an exercise.

(c) We use the analysis fact that a function $f\colon \mathbb{R} \to \mathbb{R}$ is right-continuous iff for each real sequence $\{y_n\}$ converging to x from the right (i.e. $\{y_n\}$ converges to x and $y_n \geq x$ for each n), we have that $\{f(y_n)\}$ converges to $f(x)$. Therefore, let $\{y_n\}$ be such a sequence. Borrowing notation from the previous point, we have that $\bigcap_n B_{y_n} = B_x$ and so, again by continuity of probability, $F(y_n) = \mathbb{P}(B_{y_n}) \to \mathbb{P}(B_x) = F(x)$ as $n \to \infty$. □

Exercise 1.4.25. *Fill the gaps in the proof of Lemma 1.4.24.*

A remarkable and reassuring fact is that the three properties in Lemma 1.4.24 in fact characterize distribution functions of random variables:

Theorem 1.4.26. *Let $F\colon \mathbb{R} \to \mathbb{R}$ be a function satisfying (a), (b) and (c) in Lemma 1.4.24. Then there exists a random variable X with distribution function F.*

The result above tells us that instead of directly providing a random variable, we can simply provide a function $F\colon \mathbb{R} \to \mathbb{R}$ satisfying (a), (b) and (c) in Lemma 1.4.24. Notice that it justifies the existence of the uniform random variable on $[a, b]$ (which was defined by providing its distribution function).

Exercise 1.4.27. *Determine whether the following functions $F\colon \mathbb{R} \to \mathbb{R}$ are distribution functions of a random variable:*

- $F(x) = \frac{x^2}{1+x^2}$;
- $F(x) = \frac{1}{\pi}(\arctan(x) + \frac{\pi}{2})$.

It turns out that, for a random variable X, not only it makes sense to compute $\mathbb{P}(X \leq x)$, which is the same as $\mathbb{P}(X \in (-\infty, x])$, but also $\mathbb{P}(X \in A)$ for all Borel sets $A \subseteq \mathbb{R}$. We first need to check that indeed $\{X \in A\}$ is an event whenever A is a Borel set. Recall that, in particular, every open set in $\mathbb{R}$ is a Borel set and the family $\mathcal{B}$ of Borel sets is extremely rich.

Theorem 1.4.28. *Let X be a random variable and let $A \in \mathcal{B}$ be a Borel set in $\mathbb{R}$. Then $\{X \in A\} \in \mathcal{F}$ i.e., $\{X \in A\}$ is an event.*

Proof. We proceed as follows. We let $\mathcal{G}$ be the family of all $A \subseteq \mathbb{R}$ such that $\{X \in A\} \in \mathcal{F}$ and show that $\mathcal{G}$ is a σ-field on $\mathbb{R}$ containing all open intervals in $\mathbb{R}$. Since we know $\mathcal{B}$ is generated by the open intervals (Proposition 1.1.14) and hence is the smallest σ-field on $\mathbb{R}$ containing the open intervals, we obtain that $\mathcal{B} \subseteq \mathcal{G}$, as desired.

Since $\{a < X < b\} \in \mathcal{F}$ for each real numbers $a < b$ (why?), $\mathcal{G}$ contains all open intervals. It remains to check that $\mathcal{G}$ is indeed a σ-field on $\mathbb{R}$. Clearly, $\mathbb{R} \in \mathcal{G}$, as $\{\omega : X(\omega) \in \mathbb{R}\} = \Omega \in \mathcal{F}$. Let now $A \in \mathcal{G}$. Then $\{X \in A\} \in \mathcal{F}$ and so $\{X \in A^c\} = \{X \in A\}^c \in \mathcal{F}$ as $\mathcal{F}$ is closed under complementation. Suppose finally that $A_1, A_2, \ldots \in \mathcal{G}$. Then $\{X \in \bigcup_n A_n\} = \bigcup_n \{X \in A_n\} \in \mathcal{F}$ as $\mathcal{F}$ is closed under countable unions. □

It is now easy to compute the probabilities of the events $\{X > x\}$ and $\{x \leq X \leq y\}$ knowing the distribution function of X:

Lemma 1.4.29. *Let F be the distribution function of the random variable X. Then*

(a) $\mathbb{P}(X > x) = 1 - F(x)$;

(b) $\mathbb{P}(x < X \leq y) = F(y) - F(x)$.

Proof. (a) Since $\{X > x\} = \Omega \setminus \{X \leq x\}$, we have that

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F(x).$$

(b) Ω can be written as the disjoint union $\{X \leq x\} \cup \{x < X \leq y\} \cup \{X > y\}$. Therefore, by finite additivity and (a),

$$1 = F(x) + \mathbb{P}(x < X \leq y) + (1 - F(y)),$$

as claimed. □

We somehow convinced ourselves that in the case of a discrete random variable, the probability mass function is more informative than the distribution function (see Remark 1.4.10). As the following result shows, the probability mass function indeed captures all the information in the probability space that is relevant to X: we can compute the probability of every event defined just in terms of X by simply knowing the pmf of X.

Lemma 1.4.30. *Let X be a discrete random variable with pmf f_X and let $A \subseteq \mathbb{R}$ be a Borel set. Then*

(a) The set $\{x \in \mathbb{R} : f_X(x) \neq 0\}$ is countable.

(b) $\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x)$.

Proof. (a) It follows from the fact that X takes countably many values.

(b) The event $\{X \in A\}$ can be written as the disjoint union $\bigcup_{x \in A} \{X = x\}$. Since the image of X is countable, at most countably many events of the form $\{X = x\}$ are not the empty set and so

$$\{X \in A\} = \bigcup_{x \in A \cap \mathrm{Im}(X)} \{X = x\}$$

is a countable union. But then countable additivity implies that

$$\mathbb{P}(X \in A) = \sum_{x \in A \cap \mathrm{Im}(X)} \mathbb{P}(X = x) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} f(x),$$

as claimed. □

Notice that, in view of (a), the sum in (b) is understood to be a countable sum (it contains only countably many non-zero terms). From (b) and by the second probability axiom, we have that

$$1 = \mathbb{P}(X \in \mathbb{R}) = \sum_x f_X(x).$$

For example, in the case of a binomial random variable X, we have

$$1 = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k},$$

in agreement with the Binomial Theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Remark 1.4.31. It follows from the definition that the probability mass function $f_X \colon \mathbb{R} \to [0,1]$ of a discrete random variable X satisfies:

- $f_X(x) \geq 0$ for each $x \in \mathbb{R}$;
- $\sum_x f_X(x) = 1$, where we stress again that this is a countable sum over the non-zero values of f_X.

It is easy to see that if a function f satisfies the two properties above, then it is a probability mass function for a discrete random variable.

Exercise 1.4.32. *Determine whether the following functions $f \colon \mathbb{N} \to [0,1]$ are probability mass functions of a discrete random variable:*

- $f(x) = \frac{1}{x(x+1)}$;
- $f(x) = \frac{4}{x(x+1)(x+2)}$.

The analogue of Lemma 1.4.30 and its consequences in the case of continuous random variables are given by the following.

Lemma 1.4.33. *If X is a continuous random variable with density function $f(x)$, then*

(1) $\mathbb{P}(x < X \leq y) = \int_x^y f(u)$.

(2) $\mathbb{P}(X = x) = 0$, *for each* $x \in \mathbb{R}$.

(3) $\mathbb{P}(x < X \leq y) = \mathbb{P}(x \leq X \leq y) = \mathbb{P}(x < X < y) = \mathbb{P}(x \leq X < y)$.

(4) $\int_{-\infty}^{\infty} f(u) = 1$.

Loosely speaking, the reason behind (2) is that there are uncountably many possible values for X and this number is so large that the probability of X taking any particular value is 0.

Proof. (1) By Lemma 1.4.29, definition of density function and additivity of integration, we have

$$\mathbb{P}(x < X \leq y) = F(y) - F(x) = \int_{-\infty}^{y} f(u) - \int_{-\infty}^{x} f(u) = \int_x^y f(u).$$

(2) For each $n \in \mathbb{N}$, we have that $\{X = x\} \subseteq \{x - 1/n < X \leq x\}$. Therefore, monotonicity and (1) imply that

$$\mathbb{P}(X = x) \leq \mathbb{P}\Big(x - \frac{1}{n} < X \leq x\Big) = \int_{x-\frac{1}{n}}^{x} f(u).$$

But $\int_{x-\frac{1}{n}}^{x} f(u)$ tends to 0 as $n \to \infty$.

(3) By finite additivity and (2),

$$\mathbb{P}(x \leq X \leq y) = \mathbb{P}(x < X \leq y) + \mathbb{P}(X = x) = \mathbb{P}(x < X \leq y).$$

The other equalities are proved similarly.

(4) By Lemma 1.4.24, $\lim_{x\to\infty} F(x) = 1$ and so $\int_{-\infty}^{\infty} f(u) = \lim_{x\to\infty} F(x) = 1$. □

Remark 1.4.34. It would be tempting to extend (b) in Lemma 1.4.30 to the continuous case and write $\mathbb{P}(X \in A) = \int_A f(u)$ for any Borel set A. The problem is that the expression doesn't make sense if we consider the Riemann integral (the integral you are familiar with), unless A is an interval. But Borel sets are much more rich than intervals! This is where one would replace the notion of Riemann integral with the more general notion of Lebesgue integral. Unfortunately, we have to content ourselves with Riemann integrals.

Exercise 1.4.35. *Let X be a random variable with distribution function*

$$F(x) = \begin{cases} 0 & \text{if } x < 0; \\ x^2 & \text{if } 0 \leq x \leq 1; \\ 1 & \text{if } x > 1. \end{cases}$$

Is X discrete or continuous? Compute $\mathbb{P}(1/4 < X < 5)$, $\mathbb{P}(0.2 < X < 0.8)$ and $\mathbb{P}(X = 1/2)$.

1.5 Expectation of discrete random variables

Suppose we have an experiment and a discrete random variable X arising from the experiment. We repeat the experiment a large number N of times and record the N values taken by X. Intuitively, we would expect that $\{X = x\}$ occurs approximately $\mathbb{P}(X = x)N$ many times. So the average of the values taken by X would approximately be

$$\frac{\sum_x x\mathbb{P}(X = x)N}{N} = \frac{\sum_x x f_X(x)N}{N} = \sum_x x f_X(x).$$

Definition 1.5.1. Let X be a discrete random variable with pmf $f_X(x)$. The **expected value** (or **expectation**, or **mean**) of X, denoted by $\mathbb{E}(X)$, is

$$\mathbb{E}(X) = \sum_x x f_X(x),$$

provided that $\sum_x |x f(x)|$ converges.

We will soon give a precise mathematical meaning to the notion of expectation (thanks to the Laws of large numbers). For the time being, the intuitive idea above is enough.

Remark 1.5.2. We use again the convention that $\sum_x x f_X(x)$ denotes the sum over the values of x for which $f_X(x) \neq 0$. Hence we are dealing with either a finite sum or the sum of a series, as X takes countably many values.

Remark 1.5.3. Since the expectation is defined only if $\sum_x x f_X(x)$ is absolutely convergent, the expectation is a real number. Indeed, recall that if a series is absolutely convergent, then it is also convergent.

Asking only for convergence would not be enough for our purposes, as the following undesirable property would hold: For each $x \in \mathbb{R}$, there exists a rearrangement of the series converging to x (this is the so-called Riemann's rearrangement theorem). Luckily, in the case of absolute convergence, all rearrangements converge to the same real number. As an example, think about the alternating harmonic series $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$: it converges to $\ln 2$ but we can rearrange the terms to make it convergent to any $x \in \mathbb{R}$!

Example 1.5.4. The expectation of the Bernoulli random variable in Example 1.4.13 is p. The expectation of the Binomial random variable in Example 1.4.14 is np. The expectation of the Poisson random variable in Example 1.4.16 is λ.

Exercise 1.5.5. *Let X and Y be discrete random variables with pmf's*

$$f_X(x) = \frac{4}{x(x+1)(x+2)} \text{ and } f_Y(x) = \frac{1}{x(x+1)},$$

respectively, where $x = 1, 2, \ldots$. Check whether X and Y admit an expectation and, if so, compute the value.

Given a discrete random variable X, how do we compute the expectation of the discrete random variable $Y = g(X)$? Well, according to the definition, we first have to compute the pmf of the newly defined $Y = g(X)$.

Lemma 1.5.6. *Let X be a discrete random variable and let $g\colon \mathbb{R} \to \mathbb{R}$. The pmf of $Y = g(X)$ is*

$$f_Y(y) = \sum_{x\colon\, g(x)=y} f_X(x).$$

Proof. We have that the composition function $Y = g \circ X$ acts as follows: $\omega \in \Omega \mapsto X(\omega) \in \mathbb{R} \mapsto g(X(\omega)) \in \mathbb{R}$. We rewrite the event $\{\omega : Y(\omega) = y\}$, in whose probability $f_Y(y)$ we are interested in, as a disjoint union:

$$\{\omega : Y(\omega) = y\} = \{\omega : g(X(\omega)) = y\} = \bigcup_{x\colon\, g(x)=y} \{\omega : X(\omega) = x\}.$$

By countable additivity,

$$\mathbb{P}(Y = y) = \sum_{x\colon\, g(x)=y} \mathbb{P}(X = x) = \sum_{x\colon\, g(x)=y} f_X(x),$$

as claimed. □

As doing the above procedure for each specific Y and then applying the definition of expectation becomes pretty tedious, the following settle once and for all the computation we need.

Theorem 1.5.7 (Law of the unconscious statistician, LOTUS). *Let X be a discrete random variable and $g\colon \mathbb{R} \to \mathbb{R}$. Then*

$$\mathbb{E}(g(X)) = \sum_x g(x) f_X(x),$$

provided the series is absolutely convergent.

Proof. Let $Y = g(X)$. We have seen in Lemma 1.5.6 that the pmf of Y is

$$f_Y(y) = \sum_{x:\ g(x)=y} f_X(x).$$

Therefore, by definition of expectation,

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_y y f_Y(y) \\
&= \sum_y y \sum_{x:\ g(x)=y} f_X(x) \\
&= \sum_y \sum_{x:\ g(x)=y} y f_X(x) \\
&= \sum_y \sum_{x:\ g(x)=y} g(x) f_X(x) \\
&= \sum_x g(x) f_X(x).
\end{aligned}$$

□

Example 1.5.8. Consider the following two discrete random variables X_1 and X_2, each taking three values and with pmf given by

$$\mathbb{P}(X_1 = 49) = \mathbb{P}(X_1 = 51) = \frac{1}{4} \quad \text{and} \quad \mathbb{P}(X_1 = 50) = \frac{1}{2};$$

$$\mathbb{P}(X_2 = 0) = \mathbb{P}(X_2 = 50) = \mathbb{P}(X_2 = 100) = \frac{1}{3}.$$

We have that

$$\mathbb{E}(X_1) = 49 \cdot \frac{1}{4} + 51 \cdot \frac{1}{4} + 50 \cdot \frac{1}{2} = 50 \quad \text{and} \quad \mathbb{E}(X_2) = 0 \cdot \frac{1}{3} + 50 \cdot \frac{1}{3} + 100 \cdot \frac{1}{3} = 50.$$

They have the same expected value but X_1 is much less "dispersed" than X_2.

In view of the previous example, we would like to introduce a measure of "dispersion". One way could be to measure how far things are from the expected value, on average. This leads to the notion of variance.

Definition 1.5.9. The **variance** of a discrete random variable X is the quantity

$$\operatorname{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

The **standard deviation** of X is the quantity $\sqrt{\operatorname{var}(X)}$. The **$k$-th moment** of X is $\mathbb{E}(X^k)$.

Notice that in the previous definition we ask that the expectations involved exist.

Remark 1.5.10. How do we compute the variance of a discrete random variable X? We can use the definition of expectation and first compute the pmf of the random variable $(X - \mathbb{E}(X))^2$. As already mentioned, a faster way is relying on LOTUS as shown in the following. Let $g(X)$ be the random variable $(X - \mathbb{E}(X))^2$ i.e., the function mapping $\omega \in \Omega$ to $g(X(\omega)) = (X(\omega) - \mathbb{E}(X))^2$. LOTUS allows us to write

$$\operatorname{var}(X) = \sum_x (x - \mathbb{E}(X))^2 f_X(x).$$

Clearly, $\mathrm{var}(X) \geq 0$, as the factors of each summand are non-negative. But when is that $\mathrm{var}(X) = 0$? Well, $\mathrm{var}(X) = 0$ if and only if $(x - \mathbb{E}(X))^2 f_X(x) = 0$ for each x. This means that, for each x such that $f_X(x) > 0$, we have $x - \mathbb{E}(X) = 0$. But then the random variable X is not really "random": its value is equal to $\mathbb{E}(X)$ with probability 1.

Exercise 1.5.11. *Show that $\mathrm{var}(X_1) \neq \mathrm{var}(X_2)$, where X_1 and X_2 are the random variables in Example 1.5.8.*

Proposition 1.5.12. *Let X be a discrete random variable and let $a, b \in \mathbb{R}$. Then*

(a) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.

(b) $\mathrm{var}(aX + b) = a^2\mathrm{var}(X)$.

(c) $\mathrm{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Proof. We repeatedly use LOTUS and Remark 1.5.10.

(a)
$$\mathbb{E}(aX + b) = \sum_x (ax + b) f_X(x) = a \sum_x x f_X(x) + b \sum_x f_X(x) = a\mathbb{E}(X) + b.$$

(b)
$$\begin{aligned}
\mathrm{var}(aX + b) &= \sum_x (ax + b - \mathbb{E}(aX + b))^2 f_X(x) \\
&= \sum_x (ax - a\mathbb{E}(X))^2 f_X(x) \\
&= a^2 \sum_x (x - \mathbb{E}(X))^2 f_X(x) \\
&= a^2 \mathrm{var}(X).
\end{aligned}$$

(c) Exercise. □

(a) and (b) show the behaviour of expectation and variance of $g(X)$, when g is a linear function. (c) provides an alternative way of computing the variance.

Example 1.5.13. Consider a Bernoulli random variable X with $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. Recall that $\mathbb{E}(X) = p$. LOTUS and (c) above then imply that $\mathrm{var}(X) = (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 = p(1-p)$.

Exercise 1.5.14. *Show that the variance of the Poisson random variable with parameter $\lambda > 0$ is λ.*

1.6 Multiple discrete random variables

It is often the case that each outcome of an experiment generates several real numbers of interest. We have seen how to treat these as individual random variables but it is often important to consider their "joint behaviour". For example, complicated systems are monitored by several computers that work together to run the system. If one fails or makes an error, the others can override it and the system fails only when a majority of computers fail. If X_i denotes the time until the i-th processor fails, then the time until the system fails depends jointly on the collection of random variables $X_1, \ldots, X_n$. As a concrete easy example, consider the following.

Example 1.6.1. We flip a fair coin twice and let X_1 be the number of heads on the first flip, X_2 be the number of heads on the second flip and $Y = 1 - X_1$. Clearly, all these random variables take values in $\{0, 1\}$ and have the same pmf (the constant function $1/2$). So we might think that the pair (X_1, X_2) "behaves" like the pair (X_1, Y). But they are in fact "different". For example, in (X_1, Y), the value of Y is completely determined by that of X_1, whereas the values of X_1 and X_2 are independent. This is not reflected by the pmf of the single variables and so we need to find a way to encode the information about their "collective behaviour".

Definition 1.6.2. Let $X_1, \ldots, X_n$ be discrete random variables. Their **joint pmf** is the function defined by

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(\{X_1 = x_1\} \cap \cdots \cap \{X_n = x_n\}).$$

We usually denote $\mathbb{P}(\{X_1 = x_1\} \cap \cdots \cap \{X_n = x_n\})$ by $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$.

Notice that $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ is a non-negative function from $\mathbb{R}^n$ to $\mathbb{R}$ which is non-zero only on a countable set of points of $\mathbb{R}^n$, namely the vectors whose i-th component is one of the countably many values X_i can take. Moreover, since Ω can be written as the union of pairwise disjoint events of the form $\{X_1 = x_1\} \cap \cdots \cap \{X_n = x_n\}$, we have that

$$\sum_{x_1,\ldots,x_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \sum_{x_1,\ldots,x_n} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(\Omega) = 1.$$

Back to our example, we have that f_{X_1,X_2} is the constant function $1/4$, whereas $f_{X_1,Y}$ is $1/2$ at the points $(0, 1)$ and $(1, 0)$ and 0 at the points $(0, 0)$ and $(1, 1)$.

As with the case of one random variable, the purpose of introducing the joint pmf is to extract all the information in the probability measure $\mathbb{P}$ that is relevant to the random variables we are considering. So we should be able to compute the probability of any event defined just in terms of the random variables by simply using their joint pmf. The following analogue of Lemma 1.4.30 in the case of multiple random variables shows that we can indeed do that.

Proposition 1.6.3. *Let $X_1, \ldots, X_n$ be discrete random variables and $A \subseteq \mathbb{R}^n$ be a Borel set. Then*

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \sum_{(x_1,\ldots,x_n)\in A} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

The following important result shows that if we know the joint mass function of two random variables, we can find all their separate mass functions. In this context, $f_X(x)$ and $f_Y(y)$ are called **marginal pmf**.

Corollary 1.6.4. *Let X and Y be discrete random variables. Then*

$$f_X(x) = \sum_y f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_x f_{X,Y}(x, y).$$

Proof. By countable additivity, we have

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y).$$

The expression for $f_Y(y)$ is obtained similarly. □

You can think of Corollary 1.6.4 as follows. The joint pmf is represented by a (countable) table, where the number in each square (x, y) is the value $f_{X,Y}(x, y)$. To calculate the marginal $f_X(x)$ for a given value of x, we simply add the numbers in the column corresponding to x. Similarly, to calculate the marginal pmf $f_Y(y)$ for a given value of y, we add the numbers in the row corresponding to y.

Given a pair (X, Y) of discrete random variables and a function $g\colon \mathbb{R}^2 \to \mathbb{R}$, we can build a new discrete random variable $Z = g(X, Y)$ defined by $Z(\omega) = g(X, Y)(\omega) = g(X(\omega), Y(\omega))$. Similarly to Lemma 1.5.6, the new random variable Z has pmf

$$f_Z(z) = \sum_{(x,y)\colon\ g(x,y)=z} f_{X,Y}(x, y).$$

We then have the following generalized version of the LOTUS whose proof we omit.

Theorem 1.6.5. $\mathbb{E}(g(X, Y)) = \sum_{x,y} g(x, y) f_{X,Y}(x, y)$.

Corollary 1.6.6. *Let X and Y be discrete random variables and $a, b \in \mathbb{R}$. Then*

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Proof. We have:

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_{x,y} (ax + by) f_{X,Y}(x, y) \\
&= a \sum_x \sum_y x f_{X,Y}(x, y) + b \sum_x \sum_y y f_{X,Y}(x, y) \\
&= a \sum_x \sum_y x f_{X,Y}(x, y) + b \sum_y \sum_x y f_{X,Y}(x, y) \\
&= a \sum_x x \sum_y f_{X,Y}(x, y) + b \sum_y y \sum_x f_{X,Y}(x, y) \\
&= a \sum_x x f_X(x) + b \sum_y y f_Y(y) \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y).
\end{aligned}$$

The first equality follows from Theorem 1.6.5. The exchange of the order of summation in the second equality is possible thanks to absolute convergence of the series $\sum_{x,y}(ax + by) f_{X,Y}(x, y)$. The fifth equality follows from Corollary 1.6.4. □

Corollary 1.6.6 shows that the expectation is linear and obviously generalizes to n random variables:

$$\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) = a_1\mathbb{E}(X_1) + \cdots + a_n\mathbb{E}(X_n).$$

Example 1.6.7. By using the definition of expectation, we have seen that the expectation of a Binomial random variable X with parameters n and p is np. A faster way to obtain this result is the following.

Let X_j be the random variable taking value 1 if the j-th flip resulted in heads and 0 otherwise (this is nothing but a Bernoulli random variable with parameter p). As X counts the number of heads in the n flips, we have that $X = X_1 + \cdots + X_n$ and so $\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np$.

Example 1.6.8 (Coupon collector). Each packet of a product is equally likely to contain any one of n different types of coupon, independently of every other packet. What is the expected number of packets you must buy to obtain at least one of each type of coupon?

Let R be the number of packets required to complete a set of n distinct coupons. We need to compute $\mathbb{E}(R)$. Let T_1 be the number of packets required to obtain the first coupon, T_2 the further number of packets required to obtain a second type of coupon, T_3 the further number required for a third type and so on. Then, $R = \sum_{i=1}^{n} T_i$. It is easy to see that

$$\mathbb{P}(T_k = r) = \left(\frac{k-1}{n}\right)^{r-1}\left(\frac{n-(k-1)}{n}\right).$$

Hence T_k is a geometric random variable with parameter $\frac{n-k+1}{n}$ and so with mean $\frac{n}{n-k+1}$. Since $R = \sum_{k=1}^{n} T_k$, we can then conclude by linearity of expectation that

$$\mathbb{E}(R) = n\sum_{k=1}^{n}\frac{1}{k},$$

which is roughly $n \log n$.

Exercise 1.6.9. *Let X be the number of fixed points in a random permutation of n items, say for example the number of students in a class of size n who receive their own homework after shuffling. Show that $\mathbb{E}(X) = \mathrm{var}(X) = 1$.*

Exercise 1.6.10. *Let X be a discrete random variable taking nonnegative integer values. Show that $\mathbb{E}(X) = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x)$. Use this to compute again the expectation of a geometric random variable with parameter p.*

1.7 Conditioning discrete random variables

Consider our usual setting $(\Omega, \mathcal{F}, \mathbb{P})$ of a probability space and let X be a discrete random variable. Suppose that we know that some event B occurs with $\mathbb{P}(B) > 0$. We have seen that this gives rise to a (conditional) probability measure, namely the function $P\colon \mathcal{F} \to \mathbb{R}$ defined by $P(A) = \mathbb{P}(A|B)$ (see Lemma 1.2.2). It makes therefore sense to consider the pmf of X with respect to the (conditional) measure P.

Definition 1.7.1. Let X be a discrete random variable and let B be an event with $\mathbb{P}(B) > 0$. The **conditional probability mass function of X given B** is the function $f_{X|B}(x) = \mathbb{P}(X = x|B)$.

Note that, by definition of conditional probability,

$$f_{X|B}(x) = \mathbb{P}(X = x|B) = \frac{\mathbb{P}(X = x, B)}{\mathbb{P}(B)}.$$

This function is clearly non-negative and $\sum_x f_{X|B}(x) = 1$ (hence it is a legitimate pmf). Indeed, the event B can be written as the countable union of pairwise disjoint events of the form $\{X = x\} \cap B$ (where x ranges through the countably many values taken by X) and so

$$\mathbb{P}(B) = \sum_x \mathbb{P}(X = x, B) = \sum_x f_{X|B}(x) \cdot \mathbb{P}(B) = \mathbb{P}(B)\sum_x f_{X|B}(x).$$

Let now X and Y be two discrete random variables associated with the same probability space. If we know that the value of Y is y (with $f_Y(y) > 0$), we can consider the conditional pmf of X given the event $\{Y = y\}$. Definition 1.7.1 adapts as follows: the **conditional pmf of X given $Y = y$** is the function

$$f_{X|Y}(x|y) \stackrel{\text{def}}{=} \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

The conditional pmf is particularly useful if we want to compute the joint pmf. Indeed, we have $f_{X,Y}(x, y) = f_{X|Y}(x|y) \cdot f_Y(y)$.

Example 1.7.2. Consider four independent rolls of a 6-sided die. Let X be the number of 1's and Y be the number of 2's obtained. What is the joint pmf of the discrete random variables X and Y?

Intuitively, X and Y are "related" and $f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y)$ should be easier to compute than $f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$. We then try to compute $f_Y(y)$ and $f_{X|Y}(x|y)$ and multiply them to get $f_{X,Y}(x, y)$. Notice first that X and Y are nothing but Binomial random variables with parameters $n = 4$ and $p = 1/6$. Indeed, nothing prevents you to think of the die as a biased coin in which the face 2 represents the outcome heads and all the other faces the outcome tails. Therefore,

$$f_Y(y) = \binom{4}{y}\left(\frac{1}{6}\right)^y\left(\frac{5}{6}\right)^{4-y},$$

for $y = 0, 1, 2, 3, 4$. Suppose now we have observed that $Y = y$. Then X is the number of 1's in the remaining $4 - y$ rolls, each of which can take one of the remaining values $\{1, 3, 4, 5, 6\}$ with probability $1/5$. This is again a Binomial random variable with parameters $n = 4 - y$ and $p = 1/5$ and so

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \binom{4-y}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{4-y-x},$$

for $x, y = 0, 1, 2, 3, 4$ with $0 \leq x + y \leq 4$.

The conditional pmf can also be used to compute one marginal pmf given the other. Indeed, Corollary 1.6.4 implies that

$$f_X(x) = \sum_y f_{X,Y}(x, y) = \sum_y f_{X|Y}(x|y) \cdot f_Y(y),$$

which is morally the same as the law of total probability.

1.8 Conditional expectation of discrete random variables

This is the same as ordinary expectation except that it refers to the conditional pmf.

Definition 1.8.1. Let X and Y be discrete random variables. The **conditional expectation of X given the event B** is

$$\mathbb{E}(X|B) = \sum_x x f_{X|B}(x),$$

provided that the series is absolutely convergent.

Adapting the above to events of the form $\{Y = y\}$, we obtain the **conditional expectation of X given $Y = y$**:

$$\mathbb{E}(X|Y = y) = \sum_x x f_{X|Y}(x|y).$$

Expectation and conditional expectation are related by the following important result. In words, it basically says that "the unconditional average can be obtained by averaging the conditional averages".

Theorem 1.8.2 (Total expectation theorem). *Let X and Y be discrete random variables. Then*

$$\mathbb{E}(X) = \sum_y \mathbb{E}(X|Y = y) \cdot f_Y(y),$$

provided that the expectations exist.

Proof. Recall that $f_X(x) = \sum_y f_{X|Y}(x|y) f_Y(y)$. Therefore,

$$\begin{aligned}
\mathbb{E}(X) &= \sum_x x f_X(x) \\
&= \sum_x x \sum_y f_{X|Y}(x|y) f_Y(y) \\
&= \sum_x \sum_y x f_{X|Y}(x|y) f_Y(y) \\
&= \sum_y \sum_x x f_{X|Y}(x|y) f_Y(y) \\
&= \sum_y f_Y(y) \sum_x x f_{X|Y}(x|y) \\
&= \sum_y f_Y(y) \cdot \mathbb{E}(X|Y=y),
\end{aligned}$$

where the exchange of summation is possible thanks to absolute convergence. □

Corollary 1.8.3. *Let $B_1, B_2, \ldots$ be a partition of Ω such that $\mathbb{P}(B_i) > 0$ for each i. Then*

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|B_i) \cdot \mathbb{P}(B_i).$$

Proof. Let Y be the discrete random variable that takes the value i if and only if B_i occurs. Clearly,

$$f_Y(i) = \mathbb{P}(Y=i) = \begin{cases} \mathbb{P}(B_i) & \text{for } i = 1, 2, \ldots; \\ 0 & \text{otherwise.} \end{cases}$$

By the Total expectation theorem,

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|Y=i) \cdot f_Y(i) = \sum_i \mathbb{E}(X|B_i) \cdot \mathbb{P}(B_i),$$

as $\omega \in B_i$ if and only if $\omega \in \{Y = i\}$. □

Theorem 1.8.2 and Corollary 1.8.3 are the "expectation versions" of the law of total probability.

Example 1.8.4. We have already computed the expectation of the geometric random variable in several different ways. To show the versatility of the Total expectation theorem, we provide yet another computation. Recall that the pmf of a geometric random variable X with parameter p is given by $f_X(x) = (1-p)^{x-1}p$. We use Corollary 1.8.3 by conditioning on the outcome of the first toss (as it is good practice when we have repeated independent trials). Therefore, consider the events $\{X = 1\}$ (i.e., the first toss gives heads) and its complement $\{X > 1\}$. Clearly, $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X > 1) = 1 - p$.

Given that the first toss is heads, the expected number of tosses before getting heads should be 1 i.e., $\mathbb{E}(X|X=1) = 1$. Indeed,

$$\mathbb{E}(X|X=1) = \sum_{x=1}^{\infty} x\mathbb{P}(X=x|X=1) = 1,$$

as only the first term of the series is nonzero.

Intuitively, given that the first toss is tails, the expected number of tosses before getting heads should be $\mathbb{E}(X|X>1) = \mathbb{E}(X) + 1$. Let's check it. We have that

$$\mathbb{E}(X|X>1) = \sum_{x=1}^{\infty} x\mathbb{P}(X=x|X>1) = \sum_{x=2}^{\infty} x\mathbb{P}(X=x|X>1).$$

But now observe that, for each $x \geq 2$,

$$\mathbb{P}(X = x|X > 1) = \frac{\mathbb{P}(X = x, X > 1)}{\mathbb{P}(X > 1)} = \frac{\mathbb{P}(X = x)}{\mathbb{P}(X > 1)} = (1-p)^{x-2}p = \mathbb{P}(X = x-1).$$

Therefore,

$$\begin{aligned}
\mathbb{E}(X|X > 1) &= \sum_{x=2}^{\infty} x\mathbb{P}(X = x-1) \\
&= \sum_{x=2}^{\infty} (x-1+1)\mathbb{P}(X = x-1) \\
&= \sum_{x=2}^{\infty} (x-1)\mathbb{P}(X = x-1) + \sum_{x=2}^{\infty} \mathbb{P}(X = x-1) \\
&= \mathbb{E}(X) + 1.
\end{aligned}$$

Corollary 1.8.3 then implies that

$$\mathbb{E}(X) = \mathbb{E}(X|X = 1)\mathbb{P}(X = 1) + \mathbb{E}(X|X > 1)\mathbb{P}(X > 1) = 1 \cdot p + (1 + \mathbb{E}(X))(1-p),$$

from which we obtain $\mathbb{E}(X) = 1/p$.

Exercise 1.8.5. *Show that the geometric random variable X has the **lack of memory** property. Namely, $\mathbb{P}(X > m+n|X > m) = \mathbb{P}(X > n)$, for each m and n in $\mathbb{N}$.*

1.9 Independence of discrete random variables

Definition 1.9.1. Two discrete random variables X and Y are **independent** if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for each pair $(x,y) \in \mathbb{R}^2$. More generally, a family of n discrete random variables $X_1, \ldots, X_n$ is independent if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

for each $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Finally, an arbitrary family of random variables is independent if each finite subfamily is.

Notice that X and Y are independent if and only if the events $\{X = x\}$ and $\{Y = y\}$ are independent for each $(x,y) \in \mathbb{R}^2$. Recall that $f_{X,Y}(x,y) = f_{X|Y}(x|y) \cdot f_Y(y)$. Therefore, X and Y are independent if and only if $f_{X|Y}(x|y) = f_X(x)$ for each y with $f_Y(y) > 0$ and for each x i.e., the experimental value of Y tells us nothing about the value of X.

Example 1.9.2. Consider again the random variables X_1, X_2 and $Y = 1 - X_1$ in Example 1.6.1. We have that X_1 and X_2 are independent but X_1 and Y are not.

Example 1.9.3. Let X_1 and X_2 be independent Poisson random variables with parameters λ_1 and λ_2, respectively. What is the pmf of $X_1 + X_2$? We need to determine $\mathbb{P}(X_1 + X_2 = n)$. We use the Law of

total probability and independence:

$$\begin{aligned}\mathbb{P}(X_1+X_2=n)=\sum_{k=0}^{n}\mathbb{P}(X_1=k,X_2=n-k)&=\sum_{k=0}^{n}\mathbb{P}(X_1=k)\mathbb{P}(X_2=n-k)\\&=\sum_{k=0}^{n}e^{-\lambda_1}\frac{\lambda_1^k}{k!}\cdot e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!}\\&=e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{\lambda_1^k\lambda_2^{n-k}}{k!(n-k)!}\\&=\frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k}\\&=e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!},\end{aligned}$$

where in the last equality we used the Binomial theorem. The computation shows that X_1+X_2 is a Poisson random variable with parameter $\lambda_1+\lambda_2$.

Exercise 1.9.4. *Let X_1 and X_2 be independent Poisson random variables with parameters λ_1 and λ_2, respectively.*

(i) Compute the conditional pmf of X_1 given that $X_1+X_2=n$.

(ii) Using (i), compute the conditional expectation of X_1 given that $X_1+X_2=n$.

(iii) Observe that in order to compute $\mathbb{E}(X_1|X_1+X_2=n)$ we can instead use symmetry and linearity of expectation: $\mathbb{E}(X_1|X_1+X_2=n)=\frac{1}{2}\mathbb{E}(X_1+X_2|X_1+X_2=n)$.

Example 1.9.5. Let $Y_1,Y_2,\dots$ be a family of independent discrete random variables. It is intuitively clear that the events $\{Y_{n+1}=y_{n+1}\}$ and $\{Y_1=y_1,Y_1+Y_2=y_2,\dots,Y_1+Y_2+\cdots+Y_n=y_n\}$ are independent. Let's formally show it. Call the last event A. We have that

$$A=\{Y_1=y_1,Y_2=y_2-y_1,\dots,Y_n=y_n-y_{n-1}-\cdots-y_1\}$$

and so

$$\begin{aligned}\mathbb{P}(Y_{n+1}=y_{n+1},A)&=\mathbb{P}(Y_{n+1}=y_{n+1},Y_1=y_1,Y_2=y_2-y_1,\dots,Y_n=y_n-y_{n-1}-\cdots-y_1)\\&=\mathbb{P}(Y_{n+1}=y_{n+1})\mathbb{P}(Y_1=y_1)\mathbb{P}(Y_2=y_2-y_1)\cdots\mathbb{P}(Y_n=y_n-y_{n-1}-\cdots-y_1)\\&=\mathbb{P}(Y_{n+1}=y_{n+1})\mathbb{P}(Y_1=y_1,Y_2=y_2-y_1,\dots,Y_n=y_n-y_{n-1}-\cdots-y_1)\\&=\mathbb{P}(Y_{n+1}=y_{n+1})\mathbb{P}(A),\end{aligned}$$

where we have simply used the definition of independence of random variables in the second and third equalities.

Theorem 1.9.6. *Let X and Y be independent discrete random variables. Then*

(a) For arbitrary Borel sets $A,B\subseteq\mathbb{R}$, we have that $\mathbb{P}(X\in A,Y\in B)=\mathbb{P}(X\in A)\mathbb{P}(Y\in B)$.

(b) For any functions $g,h\colon\mathbb{R}\to\mathbb{R}$, we have that $g(X)$ and $h(Y)$ are independent.

Notice that (a) extends to a family $X_1,\dots,X_n$ of n independent discrete random variables: for arbitrary Borel sets $S_1,\dots,S_n\subseteq\mathbb{R}$, we have

$$\mathbb{P}(X_1\in S_1,\dots,X_n\in S_n)=\prod_{i=1}^{n}\mathbb{P}(X_i\in S_i).$$

Proof. (a) We have that $\{X \in A, Y \in B\} = \bigcup_{x\in A,\, y\in B}\{X = x, Y = y\}$, where the union is countable since both X and Y are discrete random variables. But then countable additivity and independence imply that

$$\begin{aligned}\mathbb{P}(X \in A, Y \in B) &= \sum_{x\in A}\sum_{y\in B}\mathbb{P}(X = x, Y = y)\\ &= \sum_{x\in A}\sum_{y\in B}\mathbb{P}(X = x)\mathbb{P}(Y = y)\\ &= \Big(\sum_{x\in A}\mathbb{P}(X = x)\Big)\Big(\sum_{y\in B}\mathbb{P}(Y = y)\Big)\\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\end{aligned}$$

(b) It will be part of a homework assignment. □

Exercise 1.9.7. *Let $X_1, \ldots, X_n$ be a family of independent random variables. Show that the events $\{X_1 = x_1\}, \ldots, \{X_n = x_n\}$ are independent in the sense of Definition 1.3.1. In other words, you have to show that*

$$\mathbb{P}(\cap_{j\in J}\{X_j = x_j\}) = \prod_{j\in J}\mathbb{P}(\{X_j = x_j\}),$$

for each subset J of $\{1, \ldots, n\}$.

Exercise 1.9.8. *Let X and Y be independent geometric random variables with pmf's $f_X(x) = (1-\lambda)\lambda^{x-1}$ and $f_Y(y) = (1-\mu)\mu^{y-1}$, respectively. Find the pmf of $Z = \min\{X, Y\}$.*

It is in general not true that, given two discrete random variables X and Y, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ holds. Consider for example the random variable X taking values 1 and -1, each with probability $1/2$. Then $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^2) = 1$. Taking $Y = X$ we see that indeed $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ does not hold. However, the situation changes if X and Y are independent:

Theorem 1.9.9. *Let X and Y be independent discrete random variables such that $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ exist. Then $\mathbb{E}(XY)$ exists and $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Proof. We only show that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. We use Theorem 1.6.5 with the function $g\colon \mathbb{R}^2 \to \mathbb{R}$ given by $g(x, y) = xy$:

$$\mathbb{E}(XY) = \sum_x\sum_y xyf_{X,Y}(x, y) = \sum_x\sum_y xyf_X(x)f_Y(y) = \sum_x xf_X(x)\sum_y yf_Y(y) = \mathbb{E}(X)\mathbb{E}(Y),$$

where the second equality follows by independence. □

Remark 1.9.10. If X and Y are independent, we have seen that $g(X)$ and $h(Y)$ are independent as well and so, by Theorem 1.9.9, $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$.

Recall that the expectation is linear. In particular, $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ for any two random variables X and Y. Although not true in general[2], variance is linear for families of independent random variables.

Theorem 1.9.11. *If X and Y are independent discrete random variables, then* $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.

[2]For a counterexample consider again $X = Y$, where X takes values 1 and -1, each with probability $1/2$.

Proof. We have

$$\begin{aligned}\operatorname{var}(X+Y) &= \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2 \\ &= \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) - (\mathbb{E}(X))^2 - (\mathbb{E}(Y))^2 - 2\mathbb{E}(X)\mathbb{E}(Y) \\ &= \operatorname{var}(X) + \operatorname{var}(Y),\end{aligned}$$

where in the first equality we used Proposition 1.5.12(c), in the second we used LOTUS to expand the first term and linearity of expectation to expand the second term, and in the last equality we used Theorem 1.9.9. □

Example 1.9.12. Let us compute the variance of a Binomial random variable X with parameters n and p. Recall from Example 1.6.7 that X can be written as the sum of n Bernoulli random variables X_i with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$. By independence of coin tosses, $X_1, \ldots, X_n$ are independent and so $\operatorname{var}(X) = \operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n) = np(1-p)$.

We now introduce an indicator of "dependence" between two random variables:

Definition 1.9.13. The **covariance** of the discrete random variables X and Y is

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

and X and Y are **uncorrelated** if $\operatorname{cov}(X, Y) = 0$.

The covariance of two random variables is a measure of their tendency to be larger than their expected value together. A negative covariance means that when one of the variables is larger than its mean, the other is more likely to be less than its mean. By linearity of expectation, we have that

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

and so $\operatorname{cov}(X, Y) = 0$ if and only if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Even though independence implies uncorrelation, the converse is not true, as shown in the following:

Example 1.9.14. Let X be a discrete random variable such that $f_X(x) = f_X(-x)$ for each $x \in \operatorname{Im}(X)$. Suppose that $\mathbb{E}(X^3)$ exists and let $Y = X^2$. Clearly, X and Y are not independent. However, by LOTUS, we have

$$\mathbb{E}(XY) = \mathbb{E}(X^3) = \sum_{x>0} x^3(f_X(x) - f_X(-x)) = 0.$$

Similarly,

$$\mathbb{E}(X) = \sum_{x>0} x(f_X(x) - f_X(-x)) = 0$$

and so $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Exercise 1.9.15. *Show that a random variable X with the properties listed in Example 1.9.14 exists.*

1.10 Laws of large numbers and modes of convergence

Throughout this section, we assume that the random variables we are working with are discrete. Notice however that all stated results hold for any type of random variables.

Very often, in stochastics, we want to assert that some sequence of random variables tends to a limit in a suitable probabilistic sense. Limit theorems are useful for several reasons:

- They provide an interpretation of expectations in terms of a long sequence of identical independent experiments.
- They allow for an approximate analysis of the properties of random variables such as $\frac{X_1+\cdots+X_n}{n}$, where in contrast an exact analysis might reveal to be a complicated task.
- They describe the long term behavior of a stochastic process: We will see shortly that a stochastic process is nothing but a sequence $\{X_n\}$ of random variables indexed by time n.

We begin by considering the following classical situation. Let $X_1, X_2, \dots$ be a sequence of independent identically distributed (i.i.d. for short) random variables with expectation μ and variance σ^2. We look at the random variable

$$M_n = \frac{X_1 + \cdots + X_n}{n}.$$

By linearity of expectation,

$$\mathbb{E}(M_n) = \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) = \mu.$$

Since $X_1, \dots, X_n$ are independent, Proposition 1.5.12 and Theorem 1.9.11 imply that

$$\operatorname{var}(M_n) = \frac{1}{n^2}\operatorname{var}(X_1 + \cdots + X_n) = \frac{1}{n^2}(\operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n)) = \frac{\sigma^2}{n}.$$

In particular, the variance of M_n decreases to 0 as n increases. This phenomenon is the subject of the so-called Laws of large numbers, asserting that the random variables M_n converge to μ in a precise sense. We will see how this provides mathematical justification for the loose interpretation of the expectation of a random variable X as the average of a large number of independent samples drawn from the distribution of X.

In order to make the discussion above more precise, we need to introduce some probability inequalities. We remark that they hold for continuous random variables as well (with almost identical proofs, provided we define the expectation and variance of a continuous random variable) but we should content ourselves with discrete ones.

Theorem 1.10.1 (Markov's inequality). *Let X be a non-negative random variable. Then, for each $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. Fix $a > 0$ and consider the discrete random variable Y_a defined by

$$Y_a = \begin{cases} 0 & \text{if } X < a; \\ a & \text{if } X \geq a; \end{cases}$$

By construction, $X \geq Y_a$ (this should be understood as $X(\omega) \geq Y_a(\omega)$ for each $\omega \in \Omega$). But then, using monotonicity of expectation (show that it indeed holds!), we have

$$\mathbb{E}(X) \geq \mathbb{E}(Y_a) = a\mathbb{P}(Y_a = a) = a\mathbb{P}(X \geq a),$$

as claimed. □

In words, if a non-negative random variable has small expectation, then the probability that it takes a large value is small.

Theorem 1.10.2 (Chebyshev's inequality). *Let X be a random variable. Then, for each $c > 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq c) \leq \frac{\operatorname{var}(X)}{c^2}.$$

Proof. We apply Markov's inequality to the nonnegative random variable $(X - \mathbb{E}(X))^2$ and take $a = c^2$:

$$\mathbb{P}((X - \mathbb{E}(X))^2 \geq c^2) \leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{c^2} = \frac{\mathrm{var}(X)}{c^2}.$$

But the event $\{(X - \mathbb{E}(X))^2 \geq c^2\}$ is the same as the event $\{|X - \mathbb{E}(X)| \geq c\}$ and the conclusion follows. □

In words, if a random variable has small variance, then the probability that it takes a value far from its expectation is small.

Example 1.10.3. Let X be the random variable counting the number of students in a class of size n who receive their own homework after shuffling. Recall from Exercise 1.6.9 that $\mathbb{E}(X) = \mathrm{var}(X) = 1$. We now want to estimate $\mathbb{P}(X \geq 20)$. By monotonicity and Chebyshev's inequality,

$$\mathbb{P}(X \geq 20) \leq \mathbb{P}(|X - 1| \geq 19) \leq \frac{1}{19^2}.$$

Notice this is independent of the class size n.

Example 1.10.4. Let X be a discrete random variable with $\mathbb{E}(X) = \mathbb{E}(X^2) = 0$. Then $X = 0$ almost surely i.e., $\mathbb{P}(X = 0) = 1$. This should sound familiar, as we have already observed it. We use Chebyshev's inequality to deduce it again. Since $\mathrm{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 0$, Chebyshev's inequality implies that $\mathbb{P}(|X| \geq c) = 0$ for each $c > 0$. But $\{|X| > 0\} = \bigcup_{k \in \mathbb{N}} \{|X| > 1/k\}$ and so the union bound implies that

$$\mathbb{P}(|X| > 0) \leq \sum_{k=1}^{\infty} \mathbb{P}(|X| > 1/k) = 0,$$

from which we obtain that $\mathbb{P}(X = 0) = 1$.

Exercise 1.10.5. *Let X be a discrete random variable with mean $\mu = 10$ and $\sigma^2 = 5$. Estimate the probability $\mathbb{P}(3 < X < 15)$.*

Let's now go back to our sequence $X_1, X_2, \ldots$ of i.i.d. random variables with expectation μ and variance σ^2. We have seen that the random variable

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

has expectation $\mathbb{E}(M_n) = \mu$ and variance $\mathrm{var}(M_n) = \sigma^2/n$. Applying Chebyshev's inequality to M_n and taking $c = \varepsilon$, we have that for each $\varepsilon > 0$,

$$\mathbb{P}(|M_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

But for fixed $\varepsilon > 0$, the RHS goes to 0 as n tends to ∞. We have therefore proved the following:

Theorem 1.10.6 (Weak law of large numbers, WLLN). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ. Then, for each $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0$$

as $n \to \infty$.

Remark 1.10.7. Notice that, if we drop the independence assumption, the theorem will generally be false. Consider for example the case where $X_i = X$ for each i and the random variable X is not a constant random variable.

A consequence of the WLLN is the following interpretation of the expectation of a random variable: The arithmetic average of a sequence of independent observations of a random variable X converges with high probability to $\mathbb{E}(X)$. More precisely,

> we can estimate the expectation of a random variable with any amount of precision with arbitrary probability if we use a sufficiently large number of samples of its value.

Now, the WLLN is not completely satisfactory as it just states that the probability $\mathbb{P}(|M_n - \mu| \geq \varepsilon)$ of a significant deviation of M_n from μ goes to zero as $n \to \infty$. Still, for any $n \in \mathbb{N}$, this probability might be positive and it is conceivable that once in a while, even if infrequently, M_n deviates significantly from μ. The problem of the WLLN is that it deals with a somewhat weak notion of convergence, called convergence in probability. What would back our intuitive notion of expectation though, is another notion of convergence, called convergence almost surely, according to which M_n converges to μ with probability 1. We will see how this implies that, for any given $\varepsilon > 0$, the difference $|M_n - \mu|$ exceeds ε only finitely many times.

But what does it mean that "the sequence of random variables $\{X_n\}$ converges to the random variable X?" The X_n being random variables, they are in particular functions $\Omega \to \mathbb{R}$, so this involves convergence of functions. There are a number of possibilities.

Definition 1.10.8 (Modes of convergence). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, X_1, X_2, \ldots$ be a sequence of random variables.

- $\{X_n\}$ converges to X **pointwise** if $X_n(\omega) \to X(\omega)$ for each $\omega \in \Omega$ (this is the usual pointwise convergence of functions).
- $\{X_n\}$ converges to X **almost surely** (or, **almost everywhere**) if

  $$\mathbb{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1,$$

  denoted by $X_n \xrightarrow{\text{a.s.}} X$.
- $\{X_n\}$ converges to X **in probability** if, for each $\varepsilon > 0$,

  $$\mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}) \to 0 \text{ as } n \to \infty,$$

  denoted by $X_n \xrightarrow{\text{p}} X$.

Using the language just introduced, we then have that the WLLN asserts the convergence in probability of the sequence of random variables $\{M_n\}$ to the constant random variable μ.

Remark 1.10.9. Notice that, for each $\omega \in \Omega$, $\{X_n(\omega)\}$ is a sequence of real numbers and $X(\omega) \in \mathbb{R}$. The symbol $\to$ in the previous definition denotes the usual convergence of real sequences.

Remark 1.10.10. In the definition of convergence almost surely we are implicitly assuming that $\{\omega : X_n(\omega) \to X(\omega)\} \in \mathcal{F}$. The proof of this fact is extremely tedious and we happily omit it.

The notions of convergence in the definition above are listed in decreasing order of strength, namely

$$\text{pointwise convergence} \implies \text{convergence almost surely} \implies \text{convergence in probability.}$$

The last implication requires some work but we can already notice that pointwise convergence obviously implies convergence almost surely. Convergence almost surely is like pointwise convergence, except that there can be a set of probability zero on which the sequence fails to converge. Intuitively, what happens on a set of probability zero should not really matter and in fact convergence almost surely is the gold standard for convergence, the best you can hope for.

Example 1.10.11. Pointwise convergence is an extremely strong mode of convergence and almost never happens. Consider for example the experiment of repeatedly and independently tossing a fair coin. As usual, let M_n be the relative frequency of heads in the first n coin tosses. Does the sequence $\{M_n\}$ converge pointwise to $1/2$, the expectation of getting heads? Well, no. Just take ω to be the outcome $(H, H, H, \dots)$ consisting of all heads. We have that $M_n(\omega) = 1$ for each n and so $M_n(\omega) \to 1 \neq 1/2$. This shows that we cannot hope to replace convergence in probability in the statement of the WLLN with pointwise convergence. On the other hand, we will see that convergence almost surely will work!

Example 1.10.12. Suppose that $X_n \xrightarrow{\text{p}} X$ and that $X_n \xrightarrow{\text{p}} Y$ as well. Then we luckily have that $\mathbb{P}(X = Y) = 1$. In other words, we have uniqueness of the limit almost surely. Imagine how bad it would be if this failed to hold! Let's check it. Observe first that $\{X \neq Y\} = \bigcup_{k\in\mathbb{N}}\{|X - Y| > 1/k\}$. So if we show that $\mathbb{P}(|X - Y| > 1/k) = 0$ for each $k \in \mathbb{N}$, we can then invoke the union bound and conclude. But for each $\varepsilon > 0$, we have that

$$\mathbb{P}(|X - Y| > \varepsilon) = \mathbb{P}(|X - X_n + X_n - Y| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon/2) + \mathbb{P}(|X_n - Y| > \varepsilon/2), \quad (1.5)$$

where we used the triangle inequality $|X - X_n + X_n - Y| \leq |X_n - X| + |X_n - Y|$, which implies that

$$\{|X - X_n + X_n - Y| > \varepsilon\} \subseteq \{|X_n - X| > \varepsilon/2\} \cup \{|X_n - Y| > \varepsilon/2\}.$$

But the last two terms in Equation (1.5) go to 0 as $n \to \infty$, thanks to convergence in probability, and so indeed $\mathbb{P}(|X - Y| > \varepsilon) = 0$.

The following is a useful characterization of convergence almost surely. It essentially asserts that $X_n \xrightarrow{\text{a.s.}} X$ if and only if, almost surely, only finitely many X_n's deviate from X.

Lemma 1.10.13. *$X_n \xrightarrow{\text{a.s.}} X$ if and only if, for each $\varepsilon > 0$,*

$$\mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}\}) = 0.$$

Proof. Suppose first that $X_n \xrightarrow{\text{a.s.}} X$. Fix an outcome $\omega \in \{\omega : X_n(\omega) \to X(\omega)\}$. By definition of convergence of a sequence of real numbers, for each $\varepsilon > 0$, there exists $N_{\omega,\varepsilon}$ such that, if $n \geq N_{\omega,\varepsilon}$, then $|X_n(\omega) - X(\omega)| < \varepsilon$. Therefore, for such ω, $|X_n(\omega) - X(\omega)| \geq \varepsilon$ only finitely often and so

$$\{\omega : X_n(\omega) \to X(\omega)\} \subseteq \{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}\}^c.$$

Since the event on the LHS has probability one by assumption,

$$\mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}\}) = 0.$$

Conversely, suppose that, for each $\varepsilon > 0$, $\mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}\}) = 0$. Consider the event

$$A = \bigcup_{k\in\mathbb{N}} \{\omega : |X_n(\omega) - X(\omega)| \geq 1/k \text{ i.o.}\}.$$

By the union bound and our assumption, $\mathbb{P}(A) = 0$. We claim that $\{\omega : X_n(\omega) \to X(\omega)\}^c \subseteq A$. Indeed, let $\omega \in \{\omega : X_n(\omega) \to X(\omega)\}^c$. Then the real sequence $\{X_n(\omega)\}$ does not converge to $X(\omega)$ and so there exists $\varepsilon > 0$ such that $|X_n(\omega) - X(\omega)| \geq \varepsilon$ infinitely often. But then taking k such that $\varepsilon > 1/k$, we have that $\omega \in \{\omega : |X_n(\omega) - X(\omega)| \geq 1/k \text{ i.o.}\}$ and so $\omega \in A$. Since $\mathbb{P}(A) = 0$, we conclude that

$$\mathbb{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1,$$

as desired. □

Example 1.10.14. Convergence in probability does not imply convergence almost surely. Indeed, consider a sequence of independent Bernoulli random variables $\{X_n\}$ such that, for each $n \in \mathbb{N}$,

$$\mathbb{P}(X_n = 1) = \frac{1}{n} \text{ and } \mathbb{P}(X_n = 0) = 1 - \frac{1}{n}.$$

We claim that it converges in probability to the constant random variable 0. Indeed, for each $\varepsilon > 0$,

$$\mathbb{P}(\{\omega : |X_n(\omega) - 0| \geq \varepsilon\}) = \mathbb{P}(\{\omega : X_n(\omega) = 1\}) = \frac{1}{n} \to 0 \text{ as } n \to \infty.$$

On the other hand, we now verify that the sequence does not converge to 0 almost surely. We use Lemma 1.10.13. Given $\varepsilon > 0$, we have that

$$\mathbb{P}(\{\omega : |X_n(\omega)| \geq \varepsilon \text{ i.o.}\}) = \mathbb{P}(\{\omega : X_n(\omega) = 1 \text{ i.o.}\}).$$

But the second Borel-Cantelli lemma implies that this last probability is 1, as $\sum_{n=1}^{\infty} \mathbb{P}(X_n = 1) = \sum_{n=1}^{\infty} 1/n$ diverges. Therefore, by Lemma 1.10.13, we conclude that the sequence does not converge almost surely.

Example 1.10.15. Consider a random variable X such that $\mathbb{P}(|X| < 1) = 1$. We verify that the sequence $\{X/n\}$ converges to the constant random variable 0 almost surely. In view of Lemma 1.10.13, it is enough to show that, for each fixed $\varepsilon > 0$,

$$\mathbb{P}(\{\omega : |X(\omega)/n| \geq \varepsilon \text{ i.o.}\}) = 0.$$

This is screaming for Borel-Cantelli. Indeed, by the first Borel-Cantelli lemma, it is enough to show that $\sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n\varepsilon)$ converges. But there obviously exists $N \in \mathbb{N}$ such that $n\varepsilon \geq 1$ for each $n \geq N$. Since $\mathbb{P}(|X| \geq 1) = 0$, this implies that only finitely many terms of the series are non-zero and so the series converges.

Exercise 1.10.16. *Let X be the uniform random variable on $[a, b]$ (see Example 1.4.11), where $0 \leq a \leq b$. Show that the sequence $\{\frac{(-1)^n X}{n}\}$ converges to 0 almost surely.*

Exercise 1.10.17. *Let $\alpha > 0$ and let $\{X_n\}$ be a sequence of independent random variables such that, for each $n \in \mathbb{N}$,*

$$\mathbb{P}(X_n = n) = \frac{1}{n^\alpha} \text{ and } \mathbb{P}(X_n = 0) = 1 - \frac{1}{n^\alpha}.$$

Determine the values of α for which the sequence converges to 0 in probability and those for which it converges to 0 almost surely.

Exercise 1.10.18. *Show that, if $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $X_n + Y_n \xrightarrow{a.s.} X + Y$.*

Exercise 1.10.19. *Let X be a Bernoulli random variable with parameter $p = 1/2$. Let $\{X_n\}$ be a sequence of random variables such that, for each $n \in \mathbb{N}$, $X_{2n} = X$ and $X_{2n-1} = 1 - X$. Does $X_n \xrightarrow{p} X$?*

We will now show a technical result which will be used in the proof that convergence almost surely implies convergence in probability. It relates $\liminf$ and $\limsup$ of sequences of real numbers to $\liminf$ and $\limsup$ of sequences of sets.

Lemma 1.10.20. *Let $A_1, A_2, \ldots$ be events. Then*

$$\mathbb{P}(\liminf_n A_n) \leq \liminf_{n\to\infty} \mathbb{P}(A_n) \leq \limsup_{n\to\infty} \mathbb{P}(A_n) \leq \mathbb{P}(\limsup_n A_n)$$

Proof. The middle inequality follows from the properties of $\liminf_{n\to\infty}$ and $\liminf_{n\to\infty}$ of a sequence of real numbers. We show the last inequality and leave the first as an exercise. Let $B_m = \bigcup_{n\geq m} A_n$. By definition of $\limsup_n A_n$ and continuity of probability, we have

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}\Big(\bigcap_{m\geq 1} B_m\Big) = \lim_{m\to\infty} \mathbb{P}(B_m).$$

But $A_m \subseteq B_m$ and so $\mathbb{P}(B_m) \geq \mathbb{P}(A_m)$, from which we obtain the desired inequality. □

The following result, combined with Example 1.10.14, shows that convergence almost surely is stronger than convergence in probability.

Proposition 1.10.21. *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p} X$.*

Proof. Let $\varepsilon > 0$ and let $A_n = \{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}$. Since $X_n \xrightarrow{\text{a.s.}} X$, Lemma 1.10.13 implies that

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}\}) = 0,$$

and so, by Lemma 1.10.20,

$$\limsup_{n\to\infty} \mathbb{P}(A_n) \leq \mathbb{P}(\limsup_n A_n) = 0.$$

Therefore, the real sequence $\{\mathbb{P}(A_n)\}$ converges to 0 and this simply means that $X_n \xrightarrow{\text{p}} X$. □

We can finally prove the anticipated strengthening of the WLLN that guarantees convergence almost surely of $\{M_n\}$ to the mean μ. We will make use of the following classical inequality whose proof is omitted: It has an analytical counterpart that might be familiar to the reader and in fact the two results can be proved essentially in the same way.

Theorem 1.10.22 (Cauchy-Schwarz inequality). *Let X and Y be random variables such that $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ exist. Then*

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Theorem 1.10.23 (Strong law of large numbers, SLLN). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ. Then*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{a.s.} \mu.$$

Proof. For simplicity we show the result under the additional assumptions that $\mu = 0$ and $\mathbb{E}(X_i^4) \leq c$ for some $c \in \mathbb{R}$ and each i. In other words, the 4th moments are bounded.

Let $S_n = X_1 + \cdots + X_n$ and let $\varepsilon > 0$ be arbitrary. The random variable $|S_n|$ is nonnegative and the events $\{|S_n| \geq n\varepsilon\}$ and $\{S_n^4 \geq n^4\varepsilon^4\}$ coincide. Therefore, Markov's inequality implies that

$$\mathbb{P}(|S_n| \geq n\varepsilon) = \mathbb{P}(S_n^4 \geq n^4\varepsilon^4) \leq \frac{\mathbb{E}(S_n^4)}{n^4\varepsilon^4}. \tag{1.6}$$

We now expand $\mathbb{E}(S_n^4)$ using linearity:

$$\mathbb{E}(S_n^4) = \mathbb{E}((X_1 + \cdots + X_n)^4) = \mathbb{E}\Big(\sum_{i_1,\ldots,i_4=1}^{n} X_{i_1}X_{i_2}X_{i_3}X_{i_4}\Big) = \sum_{i_1,\ldots,i_4=1}^{n} \mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}).$$

Let us now look at what happens to the terms of the form $\mathbb{E}(X_i^3X_j)$, for $i \neq j$. Since X_i and X_j are independent, X_i^3 and X_j are independent (Theorem 1.9.6) and so, by Theorem 1.9.9, $\mathbb{E}(X_i^3X_j) = \mathbb{E}(X_i^3)\mathbb{E}(X_j) = 0$, as $\mathbb{E}(X_j) = \mu = 0$ by assumption. Similarly, $\mathbb{E}(X_i^2X_jX_k) = 0$ and $\mathbb{E}(X_iX_jX_kX_\ell) = 0$, where all considered indices are distinct (Exercise 1.10.24). But then only the terms of the form $\mathbb{E}(X_i^4)$ and $\mathbb{E}(X_i^2X_j^2)$ survive. We have n terms of the first type and $3n(n-1)$ terms of the second type (Exercise 1.10.25). We now bound $\mathbb{E}(S_n^4)$. By assumption $\mathbb{E}(X_i^4) \leq c$ and by the Cauchy-Schwarz inequality

$$\mathbb{E}(X_i^2X_j^2) \leq \sqrt{\mathbb{E}(X_i^4)\mathbb{E}(X_j^4)} \leq \sqrt{c^2} = c.$$

Therefore,

$$\mathbb{E}(S_n^4) \leq nc + 3n(n-1)c \leq 3n^2c.$$

We now use this bound in Equation (1.6) to obtain

$$\mathbb{P}\Big(\frac{|S_n|}{n} \geq \varepsilon\Big) \leq \frac{\mathbb{E}(S_n^4)}{n^4\varepsilon^4} \leq \frac{3n^2c}{n^4\varepsilon^4} = \frac{3c}{n^2\varepsilon^4}.$$

Since the series $\sum_{n=1}^{\infty} 1/n^2$ converges, we then have that the series

$$\sum_{n=1}^{\infty} \mathbb{P}\Big(\frac{|S_n|}{n} \geq \varepsilon\Big)$$

converges as well and so, by the first Borel-Cantelli, we have that

$$\mathbb{P}\Big(\frac{|S_n|}{n} \geq \varepsilon \text{ i.o.}\Big) = 0.$$

By Lemma 1.10.13, we then have that $\frac{X_1+\cdots+X_n}{n}$ converges to $\mu = 0$ almost surely. □

As a side remark, it is unclear who between Borel and Cantelli first proved the SLLN; what is certain is that it is a beautiful piece of mathematics. Kolmogorov made the autorship issue pointless by providing several generalizations some years after.

Exercise 1.10.24. *Show that if $X_1, X_2, \ldots$ is a (countable) family of independent random variables, then $\mathbb{E}(X_i^2X_jX_k) = 0$ and $\mathbb{E}(X_iX_jX_kX_\ell) = 0$, where all considered indices are distinct.*

Exercise 1.10.25. *Show that, in the expansion of $(X_1 + \cdots + X_n)^4$, there are n terms the form X_i^4 and $3n(n-1)$ terms of the form $X_i^2X_j^2$ $(i \neq j)$.*

According to the SLLN, with probability 1, $\{\frac{X_1+\cdots+X_n}{n}\}$ converges to the expectation μ of the X_i's. This means that, for any given $\varepsilon > 0$, the difference $|M_n - \mu|$ exceeds ε only for finitely many n (Lemma 1.10.13).

But the SLLN does in fact more than explaining the meaning of the expectation of a random variable. Recall that we intuitively identified the probability of an event with the frequency with which it occurs in an infinitely long sequence of independent trials. This was in fact our guiding intuition behind the abstract definition of the probability measure $\mathbb{P}$ (Definition 1.1.17). We now argue that the SLLN beautifully backs this intuition and establishes that the long-term frequency of occurrence of A is indeed equal

to $\mathbb{P}(A)$ almost surely i.e., with probability 1. We should be relieved: the very abstract theory built on the three probability axioms is consistent with our intuition!

Suppose indeed we run an experiment and we want to compute $\mathbb{P}(A)$, for some event A. Suppose the experiment is repeatable and each time results are independent of all other trials. Let X_i be the random variable with value 1 if A occurs and 0 otherwise (this is called the **indicator random variable of** A). Clearly, $\mathbb{E}(X_i) = 1 \cdot \mathbb{P}(X_i = 1) = \mathbb{P}(A)$. But then the SLLN implies that

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{\text{a.s.}} \mathbb{P}(A).$$

Since $\frac{X_1+\cdots+X_n}{n}$ is the fraction of time A occurred,

> we might indeed think of $\mathbb{P}(A)$ as the frequency of occurrence of A if we could repeat the experiment indefinitely and each time results were independent of all other trials.

Chapter 2

Markov Chains

We now change viewpoint and transition from classical probability to stochastic processes. Some of the same subjects (for example, sums of independent random variables and convergence) will return in different guises. But we will shift from a static to a dynamic viewpoint, from single random variables and limit theorems to processes which evolve in time.

We have seen that random variables can be interpreted as measurements of some random systems. Most systems evolve in time and one wants to be able to analyze such systems. This can be done, for example, by repeated measurements indexed by the time of the measurement. This is modelled by objects called stochastic processes. For the time being, we will content ourselves with considering the time discrete:

Definition 2.0.1. A sequence $\{X_n\}_{n\geq 0}$ of discrete random variables with values in a countable set E is a **discrete-time stochastic process with state space** E. The elements of E will be denoted by $i, j, k, \ldots$ and, if $X_n = i$, the process is said to be in state i at time n, or to visit state i at time n.

Sequences of independent random variables are stochastic processes but they are not very interesting as stochastic models, as they essentially behave in the same way. Similarly, sequences of partial averages are discrete-time stochastic processes. In order to introduce variability, we should allow some dependence on the past. In many real-life situations this happens only through the previous state. This limited amount of memory will still suffice to produce great diversity of behaviors.

Example 2.0.2 (Gambler's ruin again). Consider a gambling game in which on any turn you win $\$1$ with probability p or lose $\$1$ with probability $1-p$. Let X_n be the amount of money you have after n plays. The sequence $\{X_n\}_{n\geq 0}$ is a discrete-time stochastic process and, intuitively, given the current state X_n, any other information about the past is irrelevant for predicting the next state X_{n+1}:

$$\mathbb{P}(X_{n+1} = i+1 | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p = \mathbb{P}(X_{n+1} = i+1 | X_n = i),$$

for any $i, i_{n-1}, \ldots, i_0$. We will formally check this in Example 2.0.7.

The main reason for Markov to introduce the nowadays called Markov chains was to show that the requirement of independence in the SLLN could be relaxed. But he immediately noticed the great modelling power of Markov chains: they appeared to give an excellent description of the alternation of vowels and consonants and enabled him to calculate a very accurate estimate of the frequency at which consonants occur in Pushkin's poem *Eugene Onegin*[1].

Nowadays Markov chains are used in biology, social sciences, physics, computer science, operations research, etc. They are arguably the most successful class of stochastic processes as they are simple to describe but nevertheless can exhibit extremely varied and complex behaviors.

[1] https://www.americanscientist.org/article/first-links-in-the-markov-chain.

A discrete-time Markov chain is a stochastic process which is the simplest generalization of a sequence of independent random variables: the dependency of successive events goes back only one unit in time. In other words, the future probabilistic behavior of the process depends only on the present state of the process and is not influenced by its past history. The formal definition goes as follows:

Definition 2.0.3. A discrete-time stochastic process $\{X_n\}_{n\geq 0}$ with state space E is a **discrete-time Markov chain** if it satsfies the following **Markov property**: for each $n = 0, 1, \ldots$

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i),$$

for all $j, i, i_{n-1}, \ldots, i_0 \in E$.

The Markov chain is **temporally homogeneous** if there exist constants $p(i, j)$ such that

$$\mathbb{P}(X_{n+1} = j | X_n = i) = p(i, j)$$

holds for all n i.e., it is independent of the time parameter n.

The probabilities $p(i, j)$ are called **one-step transition probabilities** and the matrix $\mathbf{P}$ whose (i, j) entry is $p(i, j)$ is called the **transition matrix**.

Remark 2.0.4. You can remember the Markov property as "given the current state X_n, any other information about the past is irrelevant for predicting X_{n+1}".

The transition probabilities tell us how the process evolves. The laws of nature do not change with time, so if our Markov chain describes a physical process, and if the environment is not changing either, we would expect the process to evolve in the same way, regardless of what time the clock reads. In other words, we would expect the chain to be temporally homogeneous and that is indeed what we will assume from now on.

Remark 2.0.5. Notice that E might be infinite and so a transition matrix is not in general of the kind studied in linear algebra. However, the basic operations of addition and multiplication will be defined by the same formal rules. For instance, if $A = (a_{ij})_{i,j\in E}$ and $B = (b_{ij})_{i,j\in E}$, the product $C = AB$ is the matrix $(c_{ij})_{i,j\in E}$ with $c_{ij} = \sum_{k\in E} a_{ik} b_{kj}$.

Exercise 2.0.6. *Show that a sequence $\{X_n\}_{n\geq 0}$ of independent discrete random variables satisfies the Markov property.*

Notice that, for fixed i, $f(j) = p(i, j) = \mathbb{P}(X_{n+1} = j | X_n = i)$ is the conditional pmf of X_{n+1} given $X_n = i$ and so it inherits all the properties of a conditional pmf (see Section 1.7). In particular,

$$p(i, j) \geq 0 \quad \text{for all } i, j \in E \tag{2.1}$$

and

$$\sum_{j\in E} p(i, j) = 1 \quad \text{for each } i \in E. \tag{2.2}$$

(2.2) amounts to say that all row sums in the transition matrix $\mathbf{P}$ are equal to 1. A matrix with non-negative entries and such that all row sums are 1 is called **stochastic**. Hence the transition matrix of a Markov chain is a stochastic matrix.

A convenient way of representing a Markov chain is via a weighted directed graph, called the **transition graph**, whose nodes are the states in E, and for which there is a directed edge from $i \in E$ to $j \in E$ with weight $p(i, j)$ whenever this quantity is positive. Note that there may be "loops", corresponding to states i such that $p(i, i) > 0$.

Example 2.0.7 (Random walk on $\mathbb{Z}$). Suppose $Y_1, Y_2, \dots$ are i.i.d. integer-valued random variables. Let $X_0 = 0$ and, for each $n \geq 1$, let $X_n = \sum_{m=1}^{n} Y_m$. The stochastic process $\{X_n\}_{n\geq 0}$ is called random walk on $\mathbb{Z}$. You can think of a random walk as representing a quantity that changes over time (e.g., a stock price) such that its increments Y_i's are i.i.d. Let's verify it is a Markov chain. For all $j, i, i_{n-1}, \dots, i_0 \in E$, we have

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\
&= \mathbb{P}(X_n + Y_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\
&= \mathbb{P}(Y_{n+1} = j - i | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\
&= \mathbb{P}(Y_{n+1} = j - i | Y_1 + \cdots + Y_n = i, Y_1 + \cdots + Y_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\
&= \mathbb{P}(Y_{n+1} = j - i)
\end{aligned}$$

where in the last equality we used Example 1.9.5. The same computation gives $\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(Y_{n+1} = j - i)$ and so the Markov property holds. Moreover, since the Y_i's are identically distributed, we have that, for each $n \geq 0$, $\mathbb{P}(Y_{n+1} = j - i) = \mathbb{P}(Y_1 = j - i)$ and so the chain is temporally homogeneous with transition probabilities $p(i, j) = \mathbb{P}(Y_1 = j - i)$.

An interesting special case is when the increments Y_i's are Bernoulli random variables taking values 1 or -1 with $\mathbb{P}(Y_i = 1) = p$ and $P(Y_i = -1) = q$ (hence $p + q = 1$). In this case the chain $\{X_n\}_{n\geq 0}$ is called **simple random walk on** $\mathbb{Z}$. Its transition probabilities, for each i, are

$$p(i, i+1) = p, \qquad p(i, i-1) = q, \qquad p(i, j) = 0 \text{ for } j \notin \{i+1, i-1\}.$$

Notice that the gambler's ruin problem can be formulated as a simple random walk on $\mathbb{Z}$. The transition graph is as follows:

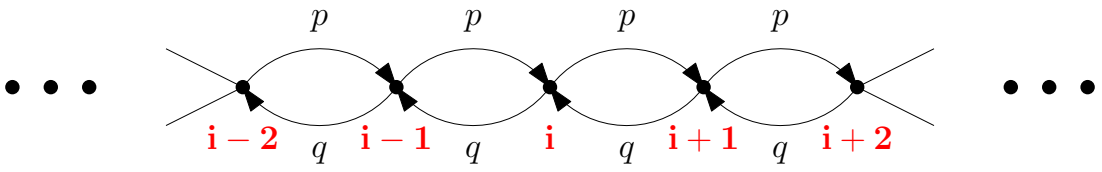


Figure 2.1: Transition graph of the simple random walk on $\mathbb{Z}$. Notice there are infinitely many states.

Example 2.0.8 (Ehrenfest chain). We have two urns A and B in which there are a total of N balls. We pick one of the N balls at random and move it to the other urn. Let X_n be the number of balls in A after the n-th draw. $\{X_n\}_{n\geq 0}$ has the Markov property. Indeed,

$$\mathbb{P}(X_{n+1} = i + 1 | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \frac{N - i}{N} = \mathbb{P}(X_{n+1} = i + 1 | X_n = i)$$

and

$$\mathbb{P}(X_{n+1} = i - 1 | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \frac{i}{N} = \mathbb{P}(X_{n+1} = i - 1 | X_n = i).$$

Moreover, if $|j - i| \geq 2$, then

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = 0 = \mathbb{P}(X_{n+1} = j | X_n = i).$$

The state space is clearly $\{1, \dots, N\}$ and the chain is temporally homogeneous.

Exercise 2.0.9. *State the transition matrix and draw the transition graph for the Ehrenfest chain with* $N = 4$.

Let $\{X_n\}_{n\geq 0}$ be a Markov chain with state space E. Every X_i, being a random variable, has a pmf, also called distribution (not to be confused with the notion of distribution function in Section 1.4). We can view this pmf as the row vector whose j-th component is $\mathbb{P}(X_i = j)$. Obviously, $\sum_{j\in E} \mathbb{P}(X_i = j) = 1$ and any vector whose entries are nonnegative numbers adding to is 1 is called a **distribution vector**.

In the previous example we started with a verbal description of the chain and figured out what the entries of its transition matrix are. However, we can describe a Markov chain by directly providing a legitimate transition matrix. Any matrix satisfying (2.1) and (2.2) gives rise to a Markov chain. This is the content of the next theorem whose proof we omit.

Theorem 2.0.10 (Existence of Markov chains). *Let* $\mathbf{P}$ *be a stochastic matrix and* α *a distribution vector. Then there exists on some probability space a sequence of random variables* $X_0, X_1, \ldots$ *which is a Markov chain with transition matrix* $\mathbb{P}$ *and initial distribution* α.

Example 2.0.11 (Social mobility). Let X_n be a family's social class in the n-th generation, which we assume is either 1 = lower, 2 = middle, or 3 = upper. In this simple version of sociology, changes of status are a Markov chain with transition matrix

$$\mathbf{P} = \begin{pmatrix} .7 & .2 & .1 \\ .3 & .5 & .2 \\ .2 & .4 & .4 \end{pmatrix}.$$

For example, we have the following one-step transition probabilities: $\mathbb{P}(X_1 = 1|X_0 = 2) = p(2,1) = .3$ and $\mathbb{P}(X_1 = 3|X_0 = 3) = p(3,3) = .4$.

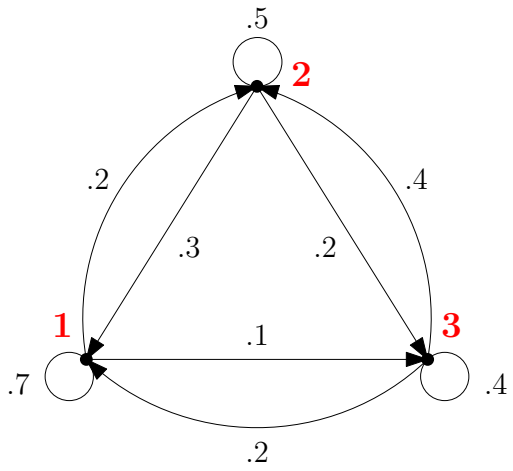


Figure 2.2: Transition graph of the social mobility Markov chain.

Given that your parents are middle class, what is the probability that you are upper class but your children are lower class? This probability is nothing but $\mathbb{P}(X_2 = 1, X_1 = 3|X_0 = 2)$. Intuitively, this is the probability that, starting in 2, we first jump to 3 and then from 3 we jump to 1. We would expect it

to be $p(2,3)p(3,1)$. Let's verify it:

$$\begin{aligned}\mathbb{P}(X_2=1,X_1=3|X_0=2) &= \frac{\mathbb{P}(X_2=1,X_1=3,X_0=2)}{\mathbb{P}(X_0=2)}\\ &= \frac{\mathbb{P}(X_2=1,X_1=3,X_0=2)}{\mathbb{P}(X_0=2,X_1=3)}\cdot\frac{\mathbb{P}(X_1=3,X_0=2)}{\mathbb{P}(X_0=2)}\\ &= \mathbb{P}(X_2=1|X_1=3,X_0=2)\cdot\mathbb{P}(X_1=3|X_0=2)\\ &= \mathbb{P}(X_2=1|X_1=3)\mathbb{P}(X_1=3|X_0=2)\\ &= p(3,1)p(2,3),\end{aligned}$$

where we have used the Markov property in the fourth equality. Loosely speaking, the probability of traversing the path $2,1,3$ in the transition graph can be computed by multiplying the probabilities on the edges.

What is the probability that your children are lower class given that your parents are middle class? This is just $\mathbb{P}(X_2=1|X_0=2)$. Now, if at time 0 we start in state 2 and at time 2 we end up in state 1, then we might have done this by being in either of the states $1,2,3$ at time 1. These are obviously disjoint events and so, since conditional probability is a probability measure, we can use finite additivity to obtain

$$\mathbb{P}(X_2=1|X_0=2)=\sum_{k=1}^{3}\mathbb{P}(X_2=1,X_1=k|X_0=2)=\sum_{k=1}^{3}p(2,k)p(k,1).$$

But the last sum should remind the reader matrix multiplication:

$$\begin{pmatrix}\cdot & \cdot & \cdot\\ p(2,1) & p(2,2) & p(2,3)\\ \cdot & \cdot & \cdot\end{pmatrix}\begin{pmatrix}p(1,1) & \cdot & \cdot\\ p(2,1) & \cdot & \cdot\\ p(3,1) & \cdot & \cdot\end{pmatrix}$$

So it seems that the probability that, starting in state i, we end up in state j after two steps is the (i,j) entry of $\mathbf{P}^2$. We will shortly prove and generalize this.

The following instructive exercise generalizes the first question in Example 2.0.11:

Exercise 2.0.12. *Show by induction on n that, for each $n\geq 1$,*

$$\mathbb{P}(X_n=i_n,\dots,X_1=i_1|X_0=i_0)=p(i_0,i_1)p(i_1,i_2)\cdots p(i_{n-1},i_n). \tag{2.3}$$

The message of eq. (2.3) is that we can think in terms of the transition graph:

> The probability that the Markov chain, starting in i_0, traverses the path $i_0,i_1,\dots,i_n$ is just the product $p(i_0,i_1)p(i_1,i_2)\cdots p(i_{n-1},i_n)$ of the transition probabilities. In other words, we can multiply probabilities along paths in the transition graphs.

Thanks to eq. (2.3), we can then compute the joint pmf of $X_0,\dots,X_n$ simply as

$$\mathbb{P}(X_0=i_0,\dots,X_n=i_n)=\mathbb{P}(X_0=i_0)\cdot p(i_0,i_1)p(i_1,i_2)\cdots p(i_{n-1},i_n).$$

Applying twice eq. (2.3), we obtain the following useful tower equality:

$$\begin{aligned}&\mathbb{P}(X_1=i_1,\dots,X_m=i_m,X_{m+1}=j_1,\dots,X_{m+n}=j_n|X_0=i_0)\\ &=[p(i_0,i_1)p(i_1,i_2)\cdots p(i_{m-1},i_m)]\cdot[p(i_m,j_1)\cdot p(j_1,j_2)\cdots p(j_{n-1},j_n)]\\ &=\mathbb{P}(X_1=i_1,\dots,X_m=i_m|X_0=i_0)\cdot\mathbb{P}(X_1=j_1,\dots,X_n=j_n|X_0=i_m).\end{aligned}$$

Loosely speaking, the tower equality says that traversing a certain path consisting of $m+n$ steps from i_0 is like doing the first m steps from i_0 to i_m and then doing the remaining n steps by starting afresh in i_m.

We can further generalize the tower equality as follows. Suppose I is a set of m-long sequences $(i_1, \ldots, i_{m-1}, j)$ of states, each with last state j, and J is a set of n-long sequences $(j_1, \ldots, j_n)$ of states. Since E is countable, I and J are countable as well. Therefore, we can write the event $\{(X_1, \ldots, X_m) \in I\}$ as a countable union of disjoint events of the form $\{X_1 = i_1, \ldots, X_{m-1} = i_{m-1}, X_m = j\}$. Similarly for the event $\{(X_{m+1}, \ldots, X_{m+n}) \in J\}$ and their intersection. But then, using countable additivity and the tower equality above, we obtain a **generalized tower equality** which will be useful later on:

$$\begin{aligned}
&\mathbb{P}((X_1, \ldots, X_m) \in I, (X_{m+1}, \ldots, X_{m+n}) \in J | X_0 = i) \\
&= \sum_{\substack{(i_1,\ldots,i_{m-1},j)\in I \\ (j_1,\ldots,j_n)\in J}} \mathbb{P}(X_1 = i_1, \ldots, X_{m-1} = i_{m-1}, X_m = j, X_{m+1} = j_1, \ldots, X_{m+n} = j_n | X_0 = i) \\
&= \sum_{\substack{(i_1,\ldots,i_{m-1},j)\in I \\ (j_1,\ldots,j_n)\in J}} \mathbb{P}(X_1 = i_1, \ldots, X_{m-1} = i_{m-1}, X_m = j | X_0 = i) \cdot \mathbb{P}(X_1 = j_1, \ldots, X_n = j_n | X_0 = j) \\
&= \mathbb{P}((X_1, \ldots, X_m) \in I | X_0 = i) \cdot \mathbb{P}((X_1, \ldots, X_n) \in J | X_0 = j).
\end{aligned}$$

The one-step transition probability $p(i,j) = \mathbb{P}(X_1 = j | X_0 = i)$ gives the probability of going from i to j in one step. But what is the probability of going from i to j in $n > 1$ steps, namely $\mathbb{P}(X_n = j | X_0 = i)$? We first observe that the assumption of temporal homogeneity for the one-step transition probabilities also implies temporal homogeneity for the n-step transition probabilities. In other words, the probability of going from i to j in n steps does not depend on the time at which we start our transition:

Lemma 2.0.13. *Let $\{X_n\}_{n\geq 0}$ be a Markov chain. Then $\mathbb{P}(X_{m+n} = j | X_m = i) = \mathbb{P}(X_n = j | X_0 = i)$.*

Proof. We proceed by induction on n, the base case $n = 1$ just being temporal homogeneity in Definition 2.0.3. Therefore, suppose the statement holds for n. We show it holds for $n+1$. We can write the event $\{X_{m+n+1} = j\}$ as the countable disjoint union $\bigcup_{k\in E}\{X_{m+n+1} = j, X_{m+n} = k\}$ and use countable additivity of the conditional probability:

$$\mathbb{P}(X_{m+n+1} = j | X_m = i) = \sum_{k\in E} \mathbb{P}(X_{m+n+1} = j, X_{m+n} = k | X_m = i).$$

We then simplify the terms of the series:

$$\begin{aligned}
\mathbb{P}(X_{m+n+1} = j, X_{m+n} = k | X_m = i) &= \frac{\mathbb{P}(X_{m+n+1} = j, X_{m+n} = k, X_m = i)}{\mathbb{P}(X_{m+n} = k, X_m = i)} \cdot \frac{\mathbb{P}(X_{m+n} = k, X_m = i)}{\mathbb{P}(X_m = i)} \\
&= \mathbb{P}(X_{m+n+1} = j | X_{m+n} = k, X_m = i) \cdot \mathbb{P}(X_{m+n} = k | X_m = i) \\
&= \mathbb{P}(X_{m+n+1} = j | X_{m+n} = k) \cdot \mathbb{P}(X_{m+n} = k | X_m = i) \\
&= \mathbb{P}(X_1 = j | X_0 = k) \cdot \mathbb{P}(X_n = k | X_0 = i),
\end{aligned}$$

where in the third equality we used the Markov property and in the fourth the induction hypothesis. Therefore,

$$\mathbb{P}(X_{m+n+1} = j | X_m = i) = \sum_{k\in E} \mathbb{P}(X_1 = j | X_0 = k) \cdot \mathbb{P}(X_n = k | X_0 = i).$$

But the RHS is independent of m and so the equality holds for each m, in particular for $m = 0$:

$$\mathbb{P}(X_{m+n+1} = j | X_m = i) = \sum_{k\in E} \mathbb{P}(X_1 = j | X_0 = k) \cdot \mathbb{P}(X_n = k | X_0 = i) = \mathbb{P}(X_{n+1} = j | X_0 = i),$$

as desired. □

Lemma 2.0.13 implies the following time-invariance principle:

> Conditional on the event $\{X_m = i\}$, the chain $\{X_{m+n}\}_{n\geq 0}$ has the same distribution as the Markov chain $\{X_n\}_{n\geq 0}$ with initial state $X_0 = i$. In other words, the Markov chain starts anew, or regenerates, at every determined time m.

We can finally find an expression for the n-step transition probabilities. It turns out that they are completely determined by the one-step transition probabilities, as shown by the following result. We denote by $p^n(i,j)$ the probability of going from i to j in n steps i.e., $\mathbb{P}(X_n = j|X_0 = i)$.

Theorem 2.0.14 (Chapman-Kolmogorov equation). *Let $\{X_n\}_{n\geq 0}$ be a Markov chain with state space E and transition matrix $\mathbf{P}$. Then, for any $m, n \geq 0$ and $i, j \in E$, we have*

$$p^{m+n}(i,j) = \sum_{k\in E} p^m(i,k)p^n(k,j).$$

In particular,

$$p^{m+1}(i,j) = \sum_{k\in E} p^m(i,k)p(k,j)$$

and $p^m(i,j)$ is the (i,j) entry of $\mathbf{P}^m$.

Proof. We begin by showing the first equation. We use the same idea as in Example 2.0.11. How can we go from state i to state j in $m+n$ steps? Well, we can first go in m steps to any of the possible states in E. We can write the event $\{X_{m+n} = j\}$ as the countable disjoint union $\bigcup_{k\in E}\{X_{m+n} = j, X_m = k\}$ and use countable additivity of the conditional probability:

$$\mathbb{P}(X_{m+n} = j|X_0 = i) = \sum_{k\in E} \mathbb{P}(X_{m+n} = j, X_m = k|X_0 = i).$$

But

$$\begin{aligned}
\mathbb{P}(X_{m+n} = j, X_m = k|X_0 = i) &= \frac{\mathbb{P}(X_{m+n} = j, X_m = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \frac{\mathbb{P}(X_{m+n} = j, X_m = k, X_0 = i)}{\mathbb{P}(X_m = k, X_0 = i)} \cdot \frac{\mathbb{P}(X_m = k, X_0 = i)}{\mathbb{P}(X_0 = i)} \\
&= \mathbb{P}(X_{m+n} = j|X_m = k, X_0 = i) \cdot \mathbb{P}(X_m = k|X_0 = i) \\
&= \mathbb{P}(X_{m+n} = j|X_m = k) \cdot \mathbb{P}(X_m = k|X_0 = i) \\
&= \mathbb{P}(X_n = j|X_0 = k) \cdot \mathbb{P}(X_m = k|X_0 = i) \\
&= p^n(k,j)p^m(i,k),
\end{aligned}$$

where we used the Markov property in the fourth equality and Lemma 2.0.13 in the fifth equality. To obtain the second equation, we simply let $n = 1$ in the first.

We finally show the last statement by induction. The case $m = 1$ holds by definition. Therefore, suppose that $p^m(i,j)$ is the (i,j) entry of $\mathbf{P}^m$. We show that $p^{m+1}(i,j)$ is the (i,j) entry of $\mathbf{P}^{m+1}$. We know that $p^{m+1}(i,j) = \sum_{k\in E} p^m(i,k)p(k,j)$ and, by the induction hypothesis, the terms of the form $p^m(i,k)$ in the sum are the (i,k) entries of $\mathbf{P}^m$ and so the sum is nothing but the product of the i-th row of $\mathbf{P}^m$ with the j-th column of $\mathbf{P}$ i.e., the (i,j) entry of $\mathbf{P}^{m+1}$. □

Remark 2.0.15. The Chapman-Kolmogorov equation simply asserts that the probability of going from i to j in $m+n$ steps is obtained by summing, over all possible states k, the probabilities of the mutually exclusive events of going first from i to k in m steps and then going from k to j in n steps.

Exercise 2.0.16. *Let $\{X_n\}_{n\geq 0}$ be a Markov chain with transition matrix $\mathbf{P}$ and let $Y_n = X_{kn}$, for some fixed k. Show that $\{Y_n\}_{n\geq 0}$ is a Markov chain and find its transition matrix.*

How is the pmf of X_n related to the pmf of the initial random variable X_0? By the Law of total probability, we have

$$\mathbb{P}(X_n = j) = \sum_{i\in E} \mathbb{P}(X_n = j|X_0 = i)\mathbb{P}(X_0 = i).$$

But we have just seen that $\mathbb{P}(X_n = j|X_0 = i)$ is the (i, j) entry of $\mathbf{P}^n$ and so the equation above tells us that we can obtain the distribution vector of X_n simply by multiplying the distribution vector of X_0 (i.e., the initial distribution), with the n-th power $\mathbf{P}^n$ of the transition matrix $\mathbf{P}$.

Example 2.0.17 (Two-state Markov chain). Arguably the easiest possible chain has two states and is defined via the following stochastic matrix $\mathbf{P}$, where $0 < p, q < 1$:

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

How do we compute $\mathbf{P}^n$? Recall that a square matrix A is **diagonalizable** if there exist a diagonal matrix D and an invertible matrix Q such that $A = QDQ^{-1}$. It is easy to compute the power of a diagonalizable matrix. Indeed, $A^n = (QDQ^{-1})(QDQ^{-1})\cdots(QDQ^{-1}) = QD^nQ^{-1}$ and D^n is simply the diagonal matrix whose (i, i) entry is the n-th power of the (i, i) entry of D. We recall the following characterization of diagonalizable matrices:

Theorem 2.0.18. *An $n \times n$ matrix A is diagonalizable if and only if A has n linearly independent eigenvectors. In this case, $A = QDQ^{-1}$, where the columns of Q are the right eigenvectors of A and the (i, i) entry of D is the eigenvalue corresponding to the eigenvector in the i-th column of Q.*

We now apply this result to our transition matrix $\mathbf{P}$. We first need to find its eigenvalues and eigenvectors. The eigenvalues are the zeros of its characteristic polynomial, namely the solutions of $\det(\mathbf{P} - \lambda I) = \lambda^2 + \lambda(p + q - 2) + 1 - q - p = 0$. The two eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 1 - p - q$. To find the corresponding eigenvectors we simply have to solve the systems $\mathbf{P}v = \lambda_i v$ for $i = 1, 2$. We get that the eigenvector corresponding to λ_1 is the column vector $(1, 1)^t$ and the eigenvector corresponding to λ_2 is the column vector $(-p, q)^t$. As they are clearly linearly independent, $\mathbf{P}$ is diagonalizable and

$$\mathbf{P}^n = \begin{pmatrix} 1 & -p \\ 1 & q \end{pmatrix} \begin{pmatrix} 1^n & 0 \\ 0 & (1-p-q)^n \end{pmatrix} \begin{pmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{-1}{p+q} & \frac{1}{p+q} \end{pmatrix} = \frac{1}{p+q} \begin{pmatrix} q + p(1-p-q)^n & p(1-(1-p-q)^n) \\ q(1-(1-p-q)^n) & p + q(1-p-q)^n \end{pmatrix}.$$

2.1 Classification of states

We now want to study the nature of the different states. How does the Markov chain evolve over time? Depending on its transition probabilities, a Markov chain may visit some states infinitely often and visit other states only a finite number of times over the infinite time horizon. Also, if a state is visited infinitely often, the mean time between visits may be infinite or finite. We will obtain a classification of states according to these properties.

For any event A, we denote by $P_i(A)$ the conditional probability $\mathbb{P}(A|X_0 = i)$. For example, using this notation, the Chapman-Kolmogorov equation says that $P_i(X_n = j) = p^n(i, j)$ is the (i, j) entry of $\mathbf{P}^n$.

Definition 2.1.1. A state $i \in E$ is **recurrent** if $P_i(X_n = i \text{ for some } n \geq 1) = 1$ and **transient** otherwise.

In words, a state i is recurrent if the chain will return to i after finitely many steps almost surely i.e., with probability 1.

Example 2.1.2. Consider the following transition matrix of a Markov chain with state space $E = \{0, 1, 2, 3, 4\}$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ .6 & 0 & .4 & 0 & 0 \\ 0 & .6 & 0 & .4 & 0 \\ 0 & 0 & .6 & 0 & .4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

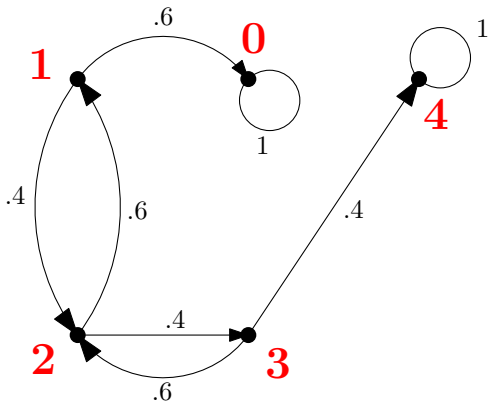


Figure 2.3: Transition graph.

Let's try to classify the states into transient and recurrent. State 0 is obviously recurrent as $P_0(X_n = 0 \text{ for some } n \geq 1) \geq P_0(X_1 = 0) = p(0, 0) = 1$. Similarly, 4 is recurrent. We claim that state 1 is transient. Indeed, the event of not returning to 1 contains the event of going from 1 to 0 and then staying in 0. This event happens with probability $p(1, 0) \cdot p(0, 0) = 0.6 > 0$. Therefore, $P_1(X_n = 1 \text{ for some } n \geq 1) < 1$. Similarly, the event of not returning to 2 contains the event of going from 2 to 1, then from 1 to 0 and then staying in 0. This event happens with probability $p(2, 1) \cdot p(1, 0) \cdot p(0, 0) = 0.6^2 > 0$. Therefore, 2 is transient. The same reasoning applies to 3.

Exercise 2.1.3. *Consider the Markov chain with state space $E = \{1, 2, 3\}$ and transition matrix*

$$\mathbf{P} = \begin{pmatrix} .1 & .2 & .7 \\ .3 & .4 & .3 \\ .5 & .4 & .1 \end{pmatrix}.$$

The initial distribution is $\mathbb{P}(X_0 = 1) = .6$, $\mathbb{P}(X_0 = 2) = .3, \mathbb{P}(X_0 = 3) = .1$.

(i) Find the distribution of X_1.

(ii) Compute $\mathbb{P}(X_1 = 1, X_2 = 2, X_3 = 3 | X_0 = 1)$.

(iii) Compute $\mathbb{P}(X_1 = 1, X_2 = 2)$.

Exercise 2.1.4. *Classify the states of the Markov chain with state space $E = \{1, 2, 3, 4, 5\}$ and transition matrix*

$$\mathbf{P} = \begin{pmatrix} .4 & .3 & .3 & 0 & 0 \\ 0 & .5 & 0 & .5 & 0 \\ .5 & 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & .5 & 0 \\ 0 & .3 & 0 & .3 & .4 \end{pmatrix}.$$

For $n \geq 1$, let $f_{i,j}^{(n)}$ be the probability that the first visit to state j from state i occurs at time n i.e.,

$$f_{i,j}^{(n)} = P_i(X_1 \neq j, \ldots, X_{n-1} \neq j, X_n = j).$$

Such an n is called **first-passage time**. Moreover, let $f_{i,j}$ be the probability that the chain ever visits state j starting in i. By countable additivity, we have

$$f_{i,j} = \sum_{n=1}^{\infty} f_{i,j}^{(n)}.$$

Notice that, by definition, $f_{i,i} = P_i(X_n = i \text{ for some } n \geq 1)$ and so a state i is recurrent if $f_{i,i} = 1$, or transient if $f_{i,i} < 1$.

Remark 2.1.5. $f_{i,j}$ should not be mistaken with the transition probability $p(i,j)$. We have that $p(i,j) \leq f_{i,j}$ and in general the inequality is strict.

Example 2.1.6. Consider the Markov chain whose state space is the set of positive integers $\{1, 2, \ldots\}$ and whose transition matrix is the infinite matrix

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & & & & \\ 1/3 & & 2/3 & & & \\ 1/4 & & & 3/4 & & \\ 1/5 & & & & 4/5 & \\ \vdots & & & & & \ddots \end{pmatrix}.$$

In other words, $p(k,1) = \frac{1}{k+1}$ and $p(k,k+1) = \frac{k}{k+1}$, for each $k \geq 1$, and all the other transition probabilities are 0. The transition graph looks as follows:

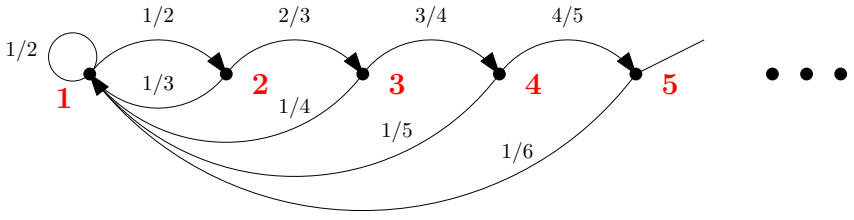


Figure 2.4: Transition graph.

Let's compute $f_{1,1}^{(n)}$, the probability that the first return to 1 occurs at time n. Clearly, $f_{1,1}^{(1)} = 1/2$. In general, we observe that the desired probability is nothing but the probability of traversing the path $1, 2, 3, \ldots, n, 1$ on the transition graph. But we know such a probability is

$$p(1,2)p(2,3)p(3,4)\cdots p(n-1,n)p(n,1) = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4}\cdots\frac{n-1}{n}\cdot\frac{1}{n+1} = \frac{1}{n}\cdot\frac{1}{n+1}.$$

Moreover,

$$f_{1,1} = \sum_{n=1}^{\infty} \frac{1}{n} \cdot \frac{1}{n+1} = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1} \right) = 1,$$

and so 1 is recurrent. An arguably easier way of reaching the same conclusion is to consider the probability of the event $\{X_n = 1 \text{ for no } n \geq 1\}$. By continuity of probability, this is just $\lim_{n\to\infty} 1/n = 0$.

Example 2.1.7. Consider the Markov chain with state space $E = \{1, 2, 3, 4, 5\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 1/3 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

(i) Compute $f_{1,1}, f_{4,4}, f_{5,5}, f_{4,3}$.

(ii) Classify the states.

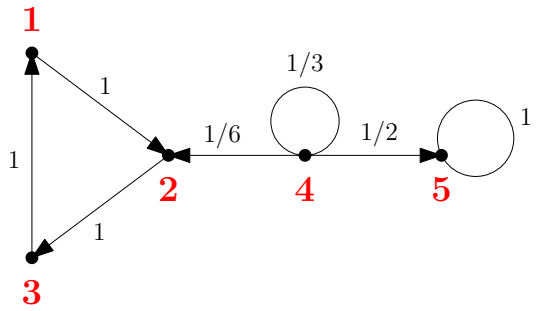


Figure 2.5: Transition graph.

Clearly, $f_{5,5} = 1$ as the chain, starting in 5, will be back after one step with probability 1. Therefore, 5 is recurrent. Consider now 1. The event that the chain ever visits 1 again contains the event of traversing the path $1, 2, 3, 1$ and the latter has probability $p(1,2) \cdot p(2,3) \cdot p(3,1) = 1$. Therefore, $f_{1,1} = 1$ i.e., 1 is recurrent. Similarly, $f_{2,2} = f_{3,3} = 1$ and 2 and 3 are recurrent as well. Consider now 4. The chain, starting in 4, visits 4 again if and only if it visits it after one step and so $f_{4,4} = p(4,4) = 1/3$ i.e., 4 is transient. Let's now compute $f_{4,3} = \sum_{n=1}^{\infty} f_{4,3}^{(n)}$. For $n \geq 2$, the probability $f_{4,3}^{(n)}$ that the first visit to 3 from 4 occurs at time n is the probability that the chain stays in 4 until time $n-2$, then moves to 2 and then to 3. The probability of this event is $(1/3)^{n-2} \cdot 1/6 \cdot 1$. Therefore,

$$f_{4,3} = \sum_{n=2}^{\infty} f_{4,3}^{(n)} = \left(\frac{1}{3} \right)^{n-2} \cdot \frac{1}{6} = \frac{1}{4}.$$

The goal now is to show that the recurrent or transient nature of a certain state depends on the number of visits the Markov chain makes to that state: A state j turns out to be recurrent if and only if the chain, starting in j, returns infinitely often almost surely i.e., with probability 1. Our characterization of recurrence and transience will involve the series $\sum_{n=1}^{\infty} p^n(j,j)$ of the n-step transition probabilities as well and we will see the significance of this.

We begin by computing the probability that, starting in i, the chain makes at least k visits to j:

Proposition 2.1.8. *Let $i, j \in E$ be states. Then*

$$P_i(\text{at least } k \text{ visits to } j) = f_{i,j} \cdot (f_{j,j})^{k-1}.$$

Proof. We show the result for $k = 2$. The proof of the case $k > 2$ follows the same line. We begin by computing $P_i(\text{first two visits at times } n_1, n_2)$, namely the probability that, starting in i, the first two visits to state j occur at at times n_1 and n_2 (where $n_1 < n_2$):

$$\begin{aligned}
&P_i(\text{first two visits at times } n_1, n_2) \\
&= P_i(X_1 \neq j, \dots, X_{n_1-1} \neq j, X_{n_1} = j, X_{n_1+1} \neq j, \dots, X_{n_2-1} \neq j, X_{n_2} = j) \\
&= P_i(X_1 \neq j, \dots, X_{n_1-1} \neq j, X_{n_1} = j) \cdot P_j(X_1 \neq j, \dots, X_{n_2-n_1-1} \neq j, X_{n_2-n_1} = j) \\
&= f_{i,j}^{(n_1)} \cdot f_{j,j}^{(n_2-n_1)},
\end{aligned}$$

where we used the generalized tower equality in the second equality. But then

$$\begin{aligned}
P_i(\text{at least two visits to } j) &= \sum_{n_1<n_2} P_i(\text{first two visits at times } n_1, n_2) \\
&= \sum_{n_1=1}^{\infty} \sum_{n_2=n_1}^{\infty} f_{i,j}^{(n_1)} \cdot f_{j,j}^{(n_2-n_1)} \\
&= f_{i,j} \cdot f_{j,j},
\end{aligned}$$

as desired. □

The previous result already tells us something meaningful:

> What happens when $k \to \infty$? Suppose first $i = j$. Then the probability that the chain visits i at least k times tends to 0 if i is transient (as $f_{i,i} < 1$), and to 1 if i is recurrent (as $f_{i,i} = 1$).
> Suppose now $i \neq j$. If j is transient, the probability that the chain visits j at least k times tends to 0, independently on the nature of i. On the other hand, if j is recurrent, then the probability is $f_{i,j}$, namely the probability that the chain ever visits j.

We now find a different expression for the n-step transition probabilities by a so-called first-passage decomposition: every path going from i to j in n steps visits j for the first time in m steps ($1 \leq m \leq n$) and then comes back to j in the remaining $n - m$ steps.

Lemma 2.1.9 (First-passage decomposition). *For any states i and j, we have*

$$p^n(i, j) = \sum_{m=1}^{n} f_{i,j}^{(m)} \cdot p^{n-m}(j, j).$$

Proof. We follow the idea mentioned above:

$$\begin{aligned}
P_i(X_n = j) &= \sum_{m=1}^{n} P_i(X_1 \neq j, \dots, X_{m-1} \neq j, X_m = j, X_n = j) \\
&= \sum_{m=1}^{n} P_i(X_1 \neq j, \dots, X_{m-1} \neq j, X_m = j) \cdot P_j(X_{n-m} = j) \\
&= \sum_{m=1}^{n} f_{i,j}^{(m)} \cdot p^{n-m}(j, j),
\end{aligned}$$

where in the first equality we used finite additivity and in the second equality we used the generalized tower equality. □

The following technical result will be used to discriminate the behavior of $\sum_{n=1}^{\infty} p^n(i,i)$ according to whether i is recurrent or transient.

Lemma 2.1.10. *For any states i and j, we have*

$$\sum_{n=1}^{N} p^n(i,j) \leq f_{i,j} \sum_{n=0}^{N} p^n(j,j).$$

Proof. By the first-passage decomposition,

$$p^n(i,j) = \sum_{m=1}^{n} f_{i,j}^{(m)} \cdot p^{n-m}(j,j).$$

But then the partial sums of the n-step transition probabilities can be rewritten as

$$\begin{aligned}
\sum_{n=1}^{N} p^n(i,j) = \sum_{n=1}^{N}\sum_{m=1}^{n} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) = \sum_{m=1}^{N}\sum_{n=m}^{N} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) &= \sum_{m=1}^{N} f_{i,j}^{(m)} \sum_{n=m}^{N} p^{n-m}(j,j) \\
&\leq \sum_{m=1}^{N} f_{i,j}^{(m)} \sum_{n=0}^{N} p^n(j,j) \\
&= \Big(\sum_{n=0}^{N} p^n(j,j)\Big)\Big(\sum_{m=1}^{N} f_{i,j}^{(m)}\Big) \\
&\leq f_{i,j} \sum_{n=0}^{N} p^n(j,j),
\end{aligned}$$

where in the third equality we exchanged the order of summation and the last inequality follows from the definition of $f_{i,j}$. □

We can finally characterize recurrent states:

Theorem 2.1.11. *For any state $i \in E$, the following are equivalent:*

1. *i is recurrent;*
2. *$\sum_{n=1}^{\infty} p^n(i,i) = \infty$;*
3. *$P_i(X_n = i \text{ i.o.}) = 1$.*

Proof. **3** $\Longrightarrow$ **2** : If it were $\sum_{n=1}^{\infty} P_i(X_n = i) < \infty$, then the first Borel-Cantelli lemma would imply that $P_i(X_n = i \text{ i.o.}) = 0$.
3 $\Longleftrightarrow$ **1** : Letting $i = j$ in Proposition 2.1.8, we obtain

$$P_i(\text{at least } k \text{ visits to } i) = f_{i,i} \cdot (f_{i,i})^{k-1} = (f_{i,i})^k.$$

Continuity of probability then implies that

$$P_i(X_n = i \text{ i.o.}) = P_i(\bigcap_k \{\text{at least } k \text{ visits to } i\}) = \lim_{k\to\infty} P_i(\text{at least } k \text{ visits to } i) = \begin{cases} 1 & \text{if } f_{i,i} = 1; \\ 0 & \text{if } f_{i,i} < 1. \end{cases}$$

2 $\Longrightarrow$ **3** : Thanks to the previous paragraph, it is enough to show that, if $\sum_{n=1}^{\infty} p^n(i,i) = \infty$, then $f_{i,i} = 1$. Therefore, suppose the series diverges. By Lemma 2.1.10,

$$\sum_{n=1}^{N} p^n(i,i) \leq f_{i,i} \sum_{n=0}^{N} p^n(i,i).$$

But $p^0(i,i) = 1$ and so we obtain

$$(1 - f_{i,i}) \sum_{n=1}^{N} p^n(i,i) \leq f_{i,i},$$

which implies that $f_{i,i} = 1$, or else $\sum_{n=1}^{\infty} p^n(i,i)$ would converge by the monotone convergence theorem, thus contradicting our assumption. Indeed, the sequence of partial sums $\sum_{n=1}^{N} p^n(i,i)$ is increasing and would be upper bounded by $\frac{f_{i,i}}{1-f_{i,i}}$. □

Using Theorem 2.1.11, we can immediately characterize transient states:

Theorem 2.1.12. *For any state $i \in E$, the following are equivalent:*

1. *i is transient;*
2. *$\sum_{n=1}^{\infty} p^n(i,i) < \infty$;*
3. *$P_i(X_n = i \text{ i.o.}) = 0$.*

Proof. **1** $\Longrightarrow$ **2** : If $\sum_{n=1}^{\infty} p^n(i,i)$ were divergent, then i would be recurrent by Theorem 2.1.11.
2 $\Longrightarrow$ **3** : This is just the first Borel-Cantelli lemma applied to the probability measure $P_i(\cdot)$.
3 $\Longrightarrow$ **1** : If i were recurrent, then Theorem 2.1.11 implies that $P_i(X_n = i \text{ i.o.}) = 1$. □

Remark 2.1.13. Theorem 2.1.11 and Theorem 2.1.12 give another example of a **zero-one law**: The probability of a certain event (in our case $\{X_n = i \text{ i.o.}\}$) must be either 0 or 1 and cannot take any intermediate value. Notice that, as opposed to the independence assumption is the second Borel-Cantelli lemma, we do not require the events $\{X_n = i\}$ to be independent!

Corollary 2.1.14. *If j is transient, then $\lim_{n\to\infty} p^n(i,j) = 0$, for each i.*

Proof. Fix an arbitrary i. Since j is transient, we know that $\sum_{n=1}^{\infty} p^n(j,j)$ converges. On the other hand, by Lemma 2.1.10,

$$\sum_{n=1}^{N} p^n(i,j) \leq f_{i,j} \sum_{n=0}^{N} p^n(j,j),$$

and so $\sum_{n=1}^{\infty} p^n(i,j)$ converges as well by the comparison test. But then the divergence test implies that $\lim_{n\to\infty} p^n(i,j) = 0$. □

We have just seen that the recurrent or transient nature of a state j depends on the number of visits the Markov chain makes to j. We could denote this number by N_j. This random number might be infinite and so N_j would not be a random variable of the type we have defined in Section 1.4. Although the following steps could be formally justified, it will be enough for us to just use them as an intuitive guidance. We write N_j as a sum of indicator random variables $N_j = \sum_{n=0}^{\infty} I_{\{X_n=j\}}$. We would then expect the following equality for the conditional mean value of N_j:

$$\mathbb{E}(N_j | X_0 = i) = \sum_{n=0}^{\infty} p^n(i,j).$$

In other words, we can interpret $\sum_{n=0}^{\infty} p^n(i,j)$ as the conditional mean number of visits that the chain makes to j. Consider for example a transient state j. We know that $f_{j,j} < 1$ and so, as seen in the proof of Theorem 2.1.11, the sequence $\sum_{n=1}^{N} p^n(j,j)$ of the partial sums is upper bounded by $\frac{f_{j,j}}{1-f_{j,j}}$ and hence $\sum_{n=0}^{N} p^n(j,j)$ is upper bounded by $\frac{1}{1-f_{j,j}}$. Therefore, if j is transient, it seems N_j behaves like a geometric random variable with parameter $1 - f_{j,j}$, which indeed could be formally proved.

Exercise 2.1.15. *The rooted binary tree is an infinite graph T with one distinguished vertex r from which comes a single edge; at every other vertex there are exactly three edges as depicted in Figure 2.6. For any vertex v of T, let $d(r,v)$ denote the distance between r and v i.e., the length of a shortest path on T between v and r. A flea, starting in r, jumps on T from a vertex along each available edge with equal probability. Let X_n be the random vertex it occupies at time n and $Y_n = d(X_n, r)$.*

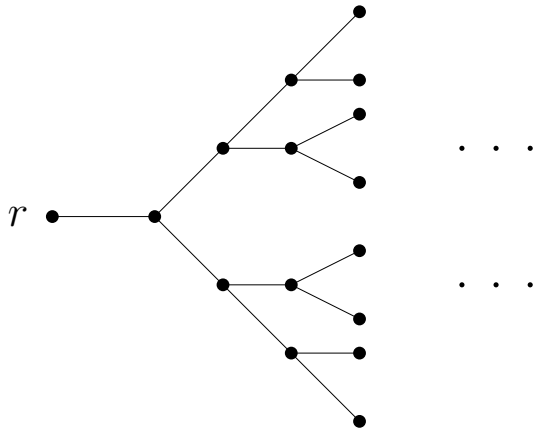


Figure 2.6: Rooted binary tree.

1. *Show that $\{Y_n\}_{n\geq 0}$ is a Markov chain with state space $\{0,1,2,\dots\}$.*
2. *Is the state 0 transient or recurrent?*
3. *What does the previous point imply for the flea?*

Example 2.1.16. Consider the simple random walk on $\mathbb{Z}$ of Example 2.0.7. We classify the states using the characterizations just obtained. For each state y, we look at $\sum_{n=1}^{\infty} p^n(y,y)$. If the series is divergent, then y is recurrent, otherwise it is transient. We first need to compute the n-step transition probabilities $p^n(y,y)$. We can represent the walk as in Figure 2.7. Clearly, we can be back at y only after an even number of steps, where each step is represented by an "up" or "down" movement in Figure 2.7: Indeed, we need the same number of steps "up" and steps "down". This means that each path of length $2n$ starting in y and returning to y occurs with probability $p^n q^n$ (thanks to Exercise 2.0.12). But there are $\binom{2n}{n}$ such paths (the number of choices of the times at which the n steps "up" occur). Therefore,

$$\sum_{n=1}^{\infty} p^n(y,y) = \sum_{n=1}^{\infty} \binom{2n}{n} p^n q^n.$$

To evaluate the series, we use the following estimation of the factorial:

Theorem 2.1.17 (Stirling formula).

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad \text{as } n \to \infty,$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$. In words, a_n and b_n behave the same way for large n.

Plugging in and simplifying, we get that the series

$$\sum_{n=1}^{\infty} \binom{2n}{n} p^n q^n$$

behaves like the series

$$\frac{\sqrt{\pi}}{\pi} \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} (4pq)^n.$$

If $p = q = 1/2$, the latter becomes

$$\frac{\sqrt{\pi}}{\pi} \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}},$$

which is divergent (it is a p-series with $p = 1/2 \leq 1$). On the other hand, if $p \neq q$, then $4pq < 1$. Since

$$\frac{1}{\sqrt{n}} (4pq)^n \leq (4pq)^n$$

and since the geometric series $\sum_{n=1}^{\infty} (4pq)^n$ converges, the comparison test implies that the series

$$\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} (4pq)^n$$

converges as well.

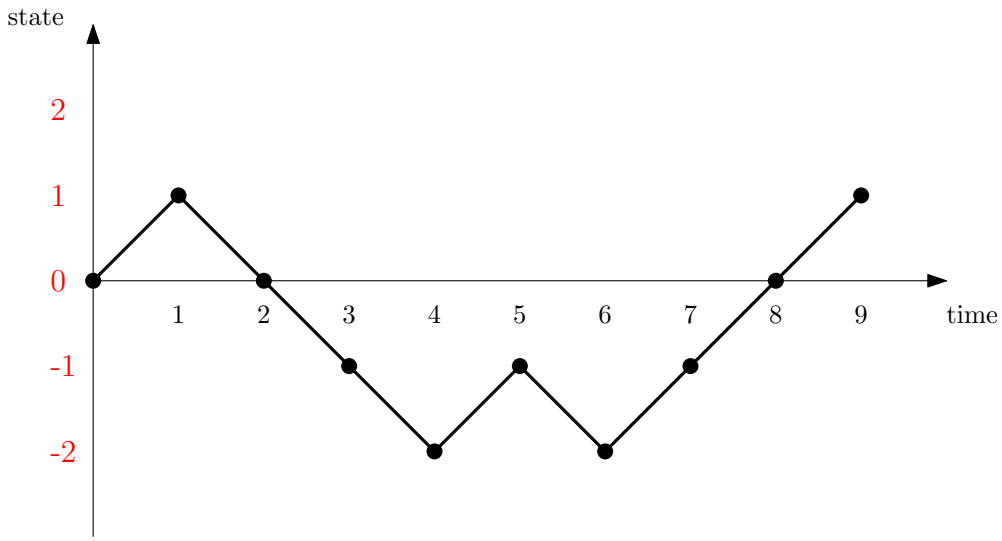


Figure 2.7: Example of evolution of the walk. At time 0 we are in state 0.

The computations above clearly hold for any state y and so we have obtained the following:

> In a simple random walk on $\mathbb{Z}$, all states are recurrent if $p = q$, whereas they are all transient if $p \neq q$. Therefore, in the case $p = q$, each integer is visited infinitely often almost surely, whereas in the case $p \neq q$, with positive probability the walk will never return to its starting point.

In the previous example we noticed that all states behaved the same way. As we will see, the reason is that they all communicate. This notion is defined as follows:

Definition 2.1.18. State j is **accessible from** state i, denoted by $i \to j$, if $p^n(i,j) > 0$ for some $n \geq 0$. States i and j **communicate**, denoted by $i \leftrightarrow j$, if $i \to j$ and $j \to i$.

Accessibility can be easily spotted in the transition graph: $i \to j$ if and only if there exists a directed path from i to j.

Lemma 2.1.19. *The communication relation $\leftrightarrow$ is an equivalence relation on E.*

Proof. Reflexivity ($i \leftrightarrow i$) and symmetry ($i \leftrightarrow j$ if and only if $j \leftrightarrow i$) are obvious. We show transitivity, namely that if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$. Since $i \to j$ and $j \to k$, there exist $n_1 \geq 0$ and $n_2 \geq 0$ such that $p^{n_1}(i,j) > 0$ and $p^{n_2}(j,k) > 0$. But then the Chapman-Kolmogorov equation implies that

$$p^{n_1+n_2}(i,k) = \sum_{\ell \in E} p^{n_1}(i,\ell) p^{n_2}(\ell,k) \geq p^{n_1}(i,j) p^{n_2}(j,k) > 0.$$

This shows that $i \to k$ and similarly we can show that $k \to i$. □

Since the communication relation is an equivalence relation, we can partition the state space E into equivalence classes under this relation, called **communication classes**. By definition of equivalence class, it might be possible to move from one class to another, but if that happens, it is then impossible to return. There are however certain sets of states from which it is impossible to get out; think about $\{1,2,3\}$ and $\{5\}$ in Example 2.1.7. These type of sets are called closed, as defined in the following:

Definition 2.1.20. A set of states $C \subseteq E$ is **closed** if no state outside C is accessible from any state in C. A state i is **absorbing** if $\{i\}$ is a closed set.

Trivially, the state space E is closed.

Remark 2.1.21. A set C is closed if $i \in C$ and $j \notin C$ implies that $p^n(i,j) = 0$ for all $n \geq 0$: it is impossible to get out of C. Indeed, for any $i \in C$, we have

$$P_i(\text{leaving } C) = P_i(\bigcup_{j \notin C} \bigcup_{n=0}^{\infty} \{X_n = j\}) \leq \sum_{j \notin C} P_i(\bigcup_{n=0}^{\infty} \{X_n = j\}) \leq \sum_{j \notin C} \sum_{n=0}^{\infty} P_i(X_n = j) = 0.$$

Notice that a state i is absorbing if and only if $p(i,i) = 1$. Absorbing states are then a special type of recurrent states: once the chain enters one of them, it will stay in that state almost surely.

A closed set may contain several communication classes. It is in fact a union of communication classes:

Exercise 2.1.22. *Show that every closed set is a union of communication classes.*

As mentioned, chains in which any two states communicate i.e., there is only one communication class, are particularly well-behaved.

Definition 2.1.23. Let C be a set of states. C is **irreducible** if $i \leftrightarrow j$ for any $i, j \in C$. C is **recurrent** or **transient** if all of its states are recurrent or transient, respectively. A Markov chain is irreducible (recurrent, transient) if its state space is irreducible (recurrent, transient).

Example 2.1.24. Each communication class is obviously irreducible. In fact, the communication class containing i is the maximal (under inclusion) irreducible set containing i.

Example 2.1.25. The Ehrenfest chain, the social mobility chain and the simple random walk on $\mathbb{Z}$ are all irreducible.

Example 2.1.26. Consider the Markov chain whose transition graph is depicted in Figure 2.8. Recall that in the transition graph we draw only the directed edges corresponding to non-zero one-step transition probabilities.

It is very easy to determine communication classes. To find the communication class a vertex i belongs to, simply look for the vertices j reachable from i by a directed path. j then belongs to the communication class of i iff there is a directed path from j to i. Repeating this procedure, we find all communication classes. In graph-theoretic language, these are the strongly connected components of the transition graph.

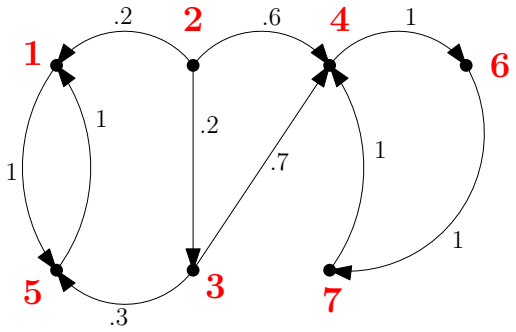


Figure 2.8: Transition graph.

In our example, the communication classes are $\{1,5\},\{2\},\{3\},\{4,6,7\}$. What are some examples of closed sets? Well, no state outside $\{1,5\}$ is accessible from it, so $\{1,5\}$ is closed. Similarly, $\{4,6,7\}$ is closed. Is there a larger closed set containing $\{1,5\}$? We know it must be a union of communication classes and $\{1,5,3,4,6,7\}$ would work, but for example $\{1,5,3\}$ would not. The chain is not irreducible, as it contains more than one communication class.

Can we identify recurrent and transient states? Well, 2 must be transient, as the event of not visiting 2 again contains the event of going to 1 in one step and the latter has non-zero probability. Hence, the probability of visiting 2 again is smaller than one. Similarly for 3. Consider now 1. It is easily seen that $f_{1,1} = 1$ and so 1 is recurrent. This shouldn't be surprising as 1 belongs to the closed communication class $\{1,5\}$, from which we cannot get out. We will soon prove this sort of behavior: a closed and finite communication class is recurrent.

Exercise 2.1.27. *For each of the chains in Examples 2.1.2, 2.1.6 and 2.1.7, determine the communication classes, the communication classes which are closed and the absorbing states. Which chains are irreducible?*

We now establish the major result that all states in an irreducible set are of the same type: either all recurrent or all transient. As every communication class is irreducible, we will then say that recurrence and transience are **class properties**. The theorem explains our observations in the previous examples.

Theorem 2.1.28. *Let C be an irreducible set. Then exactly one of the following holds:*

(i) C is recurrent, $\sum_n p^n(i,j) = \infty$ for all $i,j \in C$, and $P_j(X_n = i \text{ i.o. for all } i \in C) = 1$ for all $j \in C$.

(ii) C is transient, $\sum_n p^n(i,j) < \infty$ for all $i,j \in C$, and $P_j(X_n = i \text{ i.o. for some } i \in C) = 0$ for all $j \in C$.

Proof. The plan of the proof is as follows. We first suppose that C contains a recurrent state and show how this implies that C is recurrent and that the other two properties in (i) are satisfied. If however no state in C is recurrent, then C must be transient and we show that the other two properties in (ii) hold.

Therefore, suppose C contains a recurrent state i. By Theorem 2.1.11, $\sum_n p^n(i,i) = \infty$. Let now j be an arbitrary state in C distinct from i. We show that j is recurrent as well. Since C is irreducible,

$i \leftrightarrow j$, and so there exist n_1 and n_2 such that $p^{n_1}(j,i) > 0$ and $p^{n_2}(i,j) > 0$. For any $n \geq n_1 + n_2$, a double application of the Chapman-Kolmogorov equation gives

$$\begin{aligned} p^n(j,j) = \sum_{k\in E} p^{n_1}(j,k)p^{n-n_1}(k,j) &= \sum_{k\in E} p^{n_1}(j,k) \sum_{\ell\in E} p^{n-n_1-n_2}(k,\ell)p^{n_2}(\ell,j) \\ &= \sum_{k,\ell\in E} p^{n_1}(j,k)p^{n-n_1-n_2}(k,\ell)p^{n_2}(\ell,j) \\ &\geq p^{n_1}(j,i)p^{n-n_1-n_2}(i,i)p^{n_2}(i,j), \end{aligned}$$

where the inequality follows from the fact that the term in the RHS appears in the sum and each term in the sum is nonnegative. But

$$\sum_{n=n_1+n_2}^{\infty} p^{n_1}(j,i)p^{n_2}(i,j)p^{n-n_1-n_2}(i,i) = p^{n_1}(j,i)p^{n_2}(i,j) \sum_{n=n_1+n_2}^{\infty} p^{n-n_1-n_2}(i,i)$$

diverges and so the comparison test implies that $\sum_n p^n(j,j) = \infty$. Hence j is recurrent, as claimed. Since j was arbitrary, C is recurrent.

We now show that $\sum_{n=1}^{\infty} p^n(i,j) = \infty$ for all $i, j \in C$. Therefore, let $i, j \in C$ be arbitrary and choose n_1 such that $p^{n_1}(i,j) > 0$. For all $n \geq 1$, the Chapman-Kolmogorov equation implies that

$$p^{n_1+n}(i,j) = \sum_{k\in E} p^{n_1}(i,k)p^n(k,j) \geq p^{n_1}(i,j)p^n(j,j).$$

Since $\sum_{n=1}^{\infty} p^n(j,j)$ diverges, the comparison test implies that $\sum_{n=1}^{\infty} p^n(i,j)$ diverges as well.

We finally show the last assertion in (i):

$$P_j(X_n = i \text{ i.o. for all } i \in C) = P_j\left(\bigcap_{i\in C} \{X_n = i \text{ i.o.}\}\right) = 1.$$

For each n, we can upper bound the n-step transition probabilities as follows:

$$\begin{aligned} P_i(X_n = j) &= P_i(\{X_n = j\} \cap \{X_m = i \text{ i.o.}\}) \\ &\leq \sum_{m>n} P_i(X_n = j, X_{n+1} \neq i, \ldots, X_{m-1} \neq i, X_m = i) \\ &= \sum_{m>n} P_i(X_n = j) \cdot P_j(X_1 \neq i, \ldots, X_{m-n-1} \neq i, X_{m-n} = i) \\ &= p^n(i,j) \cdot f_{j,i}, \end{aligned}$$

where the first equality follows from Example 1.1.32 and the fact that $P_i(X_m = i \text{ i.o.}) = 1$ (Theorem 2.1.11), the inequality follows from the union bound, the second equality follows from the generalized tower equality and the last one from the definition of $f_{j,i}$. Take now n such that $p^n(i,j) > 0$. We just showed that $p^n(i,j) \leq p^n(i,j) \cdot f_{j,i}$ and so it must be $f_{j,i} = 1$. By Proposition 2.1.8 and continuity of probability, we have

$$P_j(X_n = i \text{ i.o.}) = P_j\left(\bigcap_k \{\text{at least } k \text{ visits to } i\}\right) = \lim_{k\to\infty} f_{j,i} \cdot (f_{i,i})^{k-1} = 1$$

and so, intersecting over all $i \in C$, we get $P_j(X_n = i \text{ i.o. for all } i \in C) = 1$, thanks to Exercise 1.1.36.

It remains to consider the case of C not containing any recurrent state. Then C is transient by definition. Let $i, j \in C$ be arbitrary states. By Lemma 2.1.10,

$$\sum_{n=1}^{N} p^n(i,j) \leq f_{i,j} \sum_{n=0}^{N} p^n(j,j).$$

Since when $N \to \infty$, the RHS converges (as j is transient), the comparison test implies that $\sum_{n=1}^{\infty} p^n(i,j)$ converges as well. Using this together with the first Borel-Cantelli lemma, we then obtain that $P_i(X_n = j \text{ i.o.}) = 0$ and unioning over all $j \in C$, we obtain $P_i(X_n = j \text{ i.o. for some } j \in C) = 0$, thanks to Exercise 1.1.36. This concludes the proof. □

Example 2.1.29. Consider the Markov chain whose state space is the set of nonnegative integers $\{0, 1, 2, \dots\}$ and whose transition matrix is the infinite matrix

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2^2 & 1/2^3 & 1/2^4 & \cdots \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & 1 & 0 & \\ & & & 1 & \ddots \\ & & & & \ddots \end{pmatrix}.$$

In other words, $p(0,i) = 1/2^{i+1}$, for each $i \in \{0, 1, \dots\}$, and $p(i, i-1) = 1$, for each $i \in \{1, 2, \dots\}$. All the other transition probabilities are 0. It is easy to see from the transition graph that the chain is irreducible. Notice also that, in order to check accessibility, the transition probabilities are irrelevant and we can omit them from the transition graph.

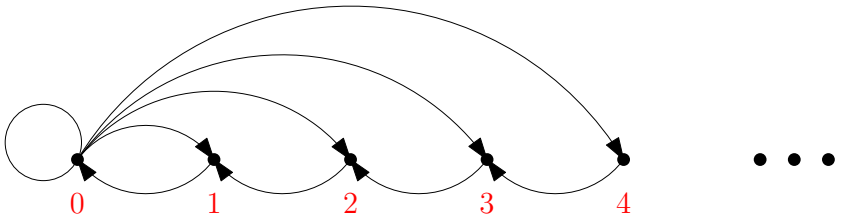


Figure 2.9: Transition graph.

By the previous theorem the chain is then either transient or recurrent. To establish which of the two holds, we simply look at the state 0 and compute $f_{0,0}$, the probability that the chain ever visits 0 again. This is easily obtained by first computing $f_{0,0}^{(n)}$ for each n (check it!). We provide another way of computing $f_{0,0}$. It should be intuitively clear that $f_{0,0} = 1$: starting in 0, no matter where we are at time 1, we will be back in 0 after finitely many steps. Formally, by the law of total probability,

$$f_{0,0} = \sum_{i=0}^{\infty} P_0(X_n = 0 \text{ for some } n \geq 1 | X_1 = i) P_0(X_1 = i) = \sum_{i=0}^{\infty} 1 \cdot P_0(X_1 = i) = 1.$$

Therefore, the chain is recurrent.

It should be intuitively clear that a communication class need not be closed (check for example the transient class in Example 2.1.2). However, the class is closed when it is recurrent.

Lemma 2.1.30. *A recurrent communication class is closed.*

Proof. We show that if a communication class is not closed then it is transient. This would conclude the proof. Therefore, suppose that C is a communication class which is not closed. Since C is not closed, there exist $j \notin C$ and $i \in C$ such that j is accessible from i i.e., $P_i(X_m = j) > 0$ for some $m \geq 0$. But since C is a communication class containing i and $j \notin C$, i is not accessible from j. This implies that a return to i is not possible if j is entered and so $1 - f_{i,i} \geq P_i(X_m = j) > 0$. Therefore, i is transient. □

Exercise 2.1.31. *Show that if $i \to j$ and j is transient, then i is transient.* *Hint: Use Lemma 2.1.30*

Exercise 2.1.32. *Consider a Markov chain with finite state space. Show that j is transient iff there exists k accessible from j but such that j is not accessible from k. Give a counterexample in the case the Markov chain has infinite state space.*

Exercise 2.1.33. *Show that every Markov chain with finite state space has at least one closed communication class.*

If the state space is finite, then we should expect not all the states to be transient, for otherwise after a finite number of steps the chain would leave every state never to return but have nowhere to go. Our intuition is confirmed by combining the following with Exercise 2.1.33.

Lemma 2.1.34. *If C is a closed and finite communication class, then C is recurrent.*

Proof. We use Theorem 2.1.28 and suppose, to the contrary, that C is transient. Then, for any $i, j \in C$ we have $\lim_{n\to\infty} p^n(i,j) = 0$ (Corollary 2.1.14). But since C is closed, we have $1 = P_i(X_n \in C)$ and so, since C is finite we can take the limit:

$$1 = P_i(X_n \in C) = \sum_{j \in C} p^n(i,j) \to 0 \quad \text{as} \quad n \to \infty,$$

a contradiction. □

Example 2.1.35. Lemma 2.1.34 immediately implies that the social mobility chain is recurrent.

Combining Lemma 2.1.30 and Lemma 2.1.34, we have that a finite communication class is recurrent if and only if it is closed. This gives the following easy procedure to classify the states of a Markov chain with finite state space. We simply find the communication classes: the closed classes are recurrent, the others are transient (check Example 2.1.26 again!).

Remark 2.1.36. Notice that the finiteness requirement is essential: we have seen that the simple random walk on $\mathbb{Z}$ with $p \neq q$ is irreducible and closed but transient.

Exercise 2.1.37. *Classify the states of every Markov chain with finite state space introduced so far.*

We now have all the ingredients to obtain a canonical decomposition of a Markov chain with state space E. We can first partition E into communication classes; the classes are irreducible, but not necessarily closed. Let now $R_1, R_2, \ldots$ denote the finite or infinite sequence of communication classes that are recurrent, and hence closed by Lemma 2.1.30. Then we set $T = E \setminus \bigcup_k R_k$. By Theorem 2.1.28, T is transient, as it consists of communication classes that are not recurrent. Notice that T is not necessarily closed and that it may be empty or equal to E. We have therefore proved the following:

Theorem 2.1.38 (Decomposition theorem for Markov chains). *The state space E of a Markov chain has the unique representation $E = T \cup (\bigcup_k R_k)$, where T is the set of transient states and each R_i is a closed irreducible recurrent set.*

Here are some consequences of the Decomposition theorem:

- If the chain starts in a recurrent set R_k, then it moves within that set forever.
- If the chain starts in the transient set T, then it moves within T and either enters one of the recurrent sets and remains in that set thereafter, or it remains in T forever, provided T is infinite (a finite T cannot be closed by Lemma 2.1.34).

Exercise 2.1.39. *Find the decomposition for the Markov chain with transition matrix*

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

2.2 Limiting and stationary distributions

In Markov chain models, we are often interested in the long-term state occupancy behavior i.e., in the n-step transition probabilities $p^n(i,j)$ for large n. We might oversimplify and suppose, for example, that a certain stock price is a Markov chain. Since transactions happen every second, we will be certainly interested in the long-term distribution. We begin by recalling what happens in the case of the simplest Markov chain.

Example 2.2.1. Consider again the two-state Markov chain defined via the following transition matrix $\mathbf{P}$, with $0 < p, q < 1$:

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

Recall that

$$\mathbf{P}^n = \begin{pmatrix} q + p(1-p-q)^n & p(1-(1-p-q)^n) \\ q(1-(1-p-q)^n) & p + q(1-p-q)^n \end{pmatrix}.$$

Since $|1-p-q| < 1$,

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{q}{p+q} & \frac{q}{p+q} \end{pmatrix}.$$

So the limit matrix exists and its rows are all equal. Our goal will be to determine to what extent this behavior is typical.

But what is the significance of the existence of such a limit? Suppose that indeed the limit matrix exists and its rows are all equal to a certain distribution vector π. We call such a vector the **limiting distribution**. In more precise terms, we have that $\lim_{n\to\infty} p^n(i,j) = \pi_j$ for all states i and j. We have seen that if X_0 has distribution vector $v^{(0)}$, then the distribution vector $v^{(n)}$ of X_n is $v^{(0)}\mathbf{P}^n$. Suppose for

simplicity[2] our Markov chain $\{X_n\}_{n\geq 0}$ has finite state space $\{0,\dots,k\}$. Then, taking the limit entrywise, we obtain

$$\lim_{n\to\infty} v^{(n)} = \lim_{n\to\infty} v^{(0)}\mathbf{P}^n = v^{(0)} \lim_{n\to\infty} \mathbf{P}^n = (v_0,\dots,v_k)\begin{pmatrix} \pi_0 & \cdots & \pi_k \\ \vdots & \vdots & \vdots \\ \pi_0 & \cdots & \pi_k \end{pmatrix} = \left(\pi_0 \sum_{i=0}^{k} v_i, \dots, \pi_k \sum_{i=0}^{k} v_i\right) = \pi.$$

This means that, no matter what the initial distribution is, the chain converges to an equilibrium distribution, namely all the random variables X_n have distribution π for sufficiently large n. In other words, at any large time n, the probability of being in state i is going to be π_i, regardless in which state the system was at time 0.

Example 2.2.2. Considering again the social mobility chain, we see that, for large n,

$$\mathbf{P}^n \approx \begin{pmatrix} .47 & .34 & .19 \\ .47 & .34 & .19 \\ .47 & .34 & .19 \end{pmatrix}.$$

The probability of being middle class is roughly $.34$, regardless of where we were at the beginning.

Observe now that if a limiting distribution π as above exists, then

$$\pi = \lim_{n\to\infty} v^{(0)}\mathbf{P}^{n+1} = (\lim_{n\to\infty} v^{(0)}\mathbf{P}^n)\mathbf{P} = \pi\mathbf{P}.$$

Such a property is so important that it deserves a definition:

Definition 2.2.3. A distribution vector π is a **stationary distribution** for $\mathbf{P}$ if $\pi\mathbf{P} = \pi$. A Markov chain admits a stationary distribution if its transition matrix admits one.

If a chain is started with a stationary distribution π as the initial distribution, it keeps the same distribution forever. Indeed, it is immediate to see by induction that $\pi\mathbf{P}^n = \pi$ for each n. We will see that many Markov chains automatically find their own way to a stationary distribution as the chain wanders through time, as the two examples above illustrates. This happens for many Markov chains, but not all. We will see the conditions required for the chain to find its way to a stationary distribution. We observed above that every limiting distribution is a stationary distribution. However, the converse is not always true (see Example 2.2.4) and so the notion of stationary distribution is more general. But the remarkable thing we will see is that in many interesting situations the two notions in fact coincide. If that is the case, computing limiting distributions becomes very easy.

Rephrasing Definition 2.2.3, we can equivalently say that a stationary distribution for $\mathbf{P}$ is a left eigenvector for $\mathbf{P}$ with eigenvalue 1. How do we check whether a stationary distribution exists and, if it does, how do we find it? Well, it's linear algebra. The condition $\pi\mathbf{P} = \pi$ gives a system of $|E|$ equations in $|E|$ unknowns (in the finite case) but we also have the condition that the sum of the entries of π is 1, as π is a distribution vector. This means that one of the equations in the system coming from $\pi\mathbf{P} = \pi$ is redundant. Indeed, summing these $|E|$ equations, it is easy to see that we get the one expressing the fact that π is a distribution.

[2] If the state space is countable the formula for the limit still holds, but one needs to be very cautious with the limit exchange.

Example 2.2.4. Let us look for a stationary distribution $\pi = (\pi_1, \pi_2)$ for the two-state Markov chain. In view of the previous comments, we already know that the limiting distribution $\pi = (\frac{q}{p+q}, \frac{p}{p+q})$ is a stationary distribution. But is there any other? We simply need to solve the system

$$\begin{cases} \pi_1(1-p) + \pi_2 q = \pi_1 \\ \pi_1 p + \pi_2(1-q) = \pi_2 \\ \pi_1 + \pi_2 = 1 \end{cases}$$

As remarked above, summing the first and second equation, we get the last. This means that we can discard one equation coming from $\pi\mathbf{P} = \pi$ and we easily find the unique solution $\pi = (\frac{q}{p+q}, \frac{p}{p+q})$.

Notice that if we allow p and q to take values in $[0, 1]$ and set $p = q = 1$, we obtain the transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

whose powers oscillates and have no limit. However, $\pi = (\frac{1}{2}, \frac{1}{2})$ is a stationary distribution for $\mathbf{P}$. So a stationary distribution might exist even if a limiting distribution does not.

The following three questions naturally arise from our previous observations:

1. Does every Markov chain admit a stationary distribution?
2. Is the stationary distribution unique?
3. When is that, for $n \to \infty$, $\mathbf{P}^n$ converges to the matrix whose rows are the stationary distribution vector, as seen in Example 2.2.1?

We start by answering the first question in the negative:

Example 2.2.5. Consider the symmetric random walk on $\mathbb{Z}$ i.e., the random walk in Example 2.0.7 with $p = q = 1/2$. Let us check whether it admits a stationary distribution. From the transition graph, we see that the j-th column of the infinite transition matrix $\mathbf{P}$ has only two non-zero entries, namely $p(j-1, j) = 1/2$ and $p(j+1, j) = 1/2$. We show that there is no stationary distribution π satisfying the system $\pi\mathbf{P} = \pi$ of infinitely many equations. For each $j \in \mathbb{Z}$, we must have that

$$\pi_j = \frac{1}{2}\pi_{j-1} + \frac{1}{2}\pi_{j+1}.$$

But this is just the recurrence relation we have seen in Example 1.3.6. We observed that π_j can be expressed in terms of π_1 and π_0 only: $\pi_j = j(\pi_1 - \pi_0) + \pi_0$. We also have that each π_j is a probability and so $\pi_j \in [0, 1]$. If $|\pi_1 - \pi_0| > 0$ then, for large j, we would have $|\pi_j - \pi_0| > 1$, a contradiction. Therefore, $\pi_1 - \pi_0 = 0$ and so $\pi_j = \pi_0$ for each $j \in \mathbb{Z}$. But

$$\sum_{j\in\mathbb{Z}} \pi_j = \sum_{j\in\mathbb{Z}} \pi_0$$

is clearly divergent (in particular, it can't be 1) and so this Markov chain has no stationary distribution.

The second question has a negative answer as well:

Example 2.2.6. Consider the Markov chain whose transition matrix is the identity I. Clearly, any distribution vector π satisfies $\pi I = \pi$ and so this Markov chain has infinitely many stationary distributions. Notice that this Markov chain is kind of trivial: every state is absorbing.

Here's another example of the same kind:

Example 2.2.7. Consider the Markov chain with state space $\{0,1,2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & .4 & .6 \\ 0 & 1 & 0 \end{pmatrix}.$$

Drawing the transition graph, it is easy to see that the chain is not irreducible, as the state 0 is absorbing. Suppose π is a stationary distribution. Let us write down the system of equations coming from $\pi\mathbf{P} = \pi$:

$$\begin{cases} \pi_0 = \pi_0 \\ 0.4\pi_1 + \pi_2 = \pi_1 \\ 0.6\pi_1 = \pi_2 \end{cases}$$

We also have that $\pi_0 + \pi_1 + \pi_2 = 1$. The first and second equation from the system are redundant. Plugging in the third into $\pi_0 + \pi_1 + \pi_2 = 1$, we get that $\pi_0 + 1.6\pi_1 = 1$. Clearly, we have infinitely many stationary distributions: any vector of the form

$$\left(\pi_0, \frac{1-\pi_0}{1.6}, 0.6 \cdot \frac{1-\pi_0}{1.6}\right)$$

with $\pi_0 \in [0,1]$.

The problem with the two examples above is that the chains are not irreducible: it is not indifferent where we start!

Before answering the third question, let's pause for a moment and see more examples of chains admitting a unique stationary distribution.

Example 2.2.8 (Ehrenfest chain again). Recalling the definition in Example 2.0.8, we have that, for $1 \leq i \leq N-1$, the i-th column of the transition matrix $\mathbf{P}$ has only two non-zero entries, namely $p(i-1,i) = \frac{N-(i-1)}{N}$ and $p(i+1,i) = \frac{i+1}{N}$. Let us look for a solution $\pi = (\pi_0, \ldots, \pi_N)$ to $\pi\mathbf{P} = \pi$. The entries of π need to satisfy

$$\pi_i = \pi_{i-1}\frac{N-(i-1)}{N} + \pi_{i+1}\frac{i+1}{N},$$

for $1 \leq i \leq N-1$, and $\pi_0 = \pi_1/N$ and $\pi_N = \pi_{N-1}/N$. Expressing the first values of π_i in terms of π_0, we make the educated guess that $\pi_i = \binom{N}{i}\pi_0$, for each $0 \leq i \leq N$, and this can be proved by induction. In order to determine π_0 we then use the fact that π is a distribution vector and so

$$1 = \sum_{i=0}^{N} \pi_i = \sum_{i=0}^{N} \binom{N}{i}\pi_0 = \pi_0 \sum_{i=0}^{N} \binom{N}{i} = \pi_0 \cdot 2^N,$$

where we used the binomial theorem in the last equality. Therefore, $\pi_0 = 1/2^N$ and the unique stationary distribution π is such that $\pi_i = \binom{N}{i} \cdot \frac{1}{2^N}$ for each $0 \leq i \leq N$. Notice that this is the probability of having i balls in urn A if balls are placed randomly and independently in either urn with probability $1/2$.

Example 2.2.9 (Random walk on a graph). A **finite simple graph** G is a collection of vertices and edges, where an edge connects two different vertices and any two vertices are connected by at most one edge. A graph can be concisely represented by its adjacency matrix A whose rows and columns are labelled with the vertices and whose (u,v) entry is 1, if there is an edge between the vertices u and v, and 0 otherwise. The degree of a vertex u is the quantity $d(u) = \sum_v A(u,v)$ i.e., the number of vertices

adjacent to u. The random walk on G is the Markov chain having as state space the set of vertices of G and whose transition matrix $\mathbf{P}$ is given by

$$p(u,v) = \frac{A(u,v)}{d(u)}.$$

It is easy to see that indeed the entries are nonnegative and the row sums are all 1. What this chain really does is that it starts at some vertex v of the graph, chooses a neighbor w of v (i.e., a vertex adjacent to v) uniformly at random, moves to w, and repeats. A simple check shows that the vector π whose u-th entry is $\pi_u = \frac{d(u)}{\sum_u d(u)}$ satisfies $\pi\mathbf{P} = \pi$ i.e., it is a stationary distribution.

2.3 Obstacles to convergence

Let us finally come back to our third question: When is that, for $n \to \infty$, $\mathbf{P}^n$ converges to the matrix whose rows are the stationary distribution vector? In order to answer this, we need to know when $\mathbf{P}^n$ converges in the first place. The following examples show what could go wrong.

Example 2.3.1. Consider again the Ehrenfest chain with $N = 3$. $\mathbf{P}$ and $\mathbf{P}^2$ are as follows:

$$\mathbf{P} = \begin{pmatrix} & 3/3 & & \\ 1/3 & & 2/3 & \\ & 2/3 & & 1/3 \\ & & 3/3 & \end{pmatrix} \qquad \mathbf{P}^2 = \begin{pmatrix} 1/3 & & 2/3 & \\ & 7/9 & & 2/9 \\ 2/9 & & 7/9 & \\ & 2/3 & & 1/3 \end{pmatrix}.$$

What happens to the diagonal entries of $\mathbf{P}^3$? Well, $p^n(i,i)$ is the probability of having i balls in urn A after n draws, given that we have i balls in A initially. Since each draw consists in removing a ball from one urn and putting it in the other, we have that the parity of the number of balls in A changes after each draw and so $p^n(i,i) = 0$ if n is odd. On the other hand, $p^n(i,i)$ is never 0 if n is even.

We have observed the same behavior for chains with infinite state space (see the simple random walk on $\mathbb{Z}$ in Example 2.0.7).

Definition 2.3.2. The **period** of a state i is the quantity $\gcd\{n \geq 1 : p^n(i,i) > 0\}$. If $p^n(i,i) = 0$ for all $n \geq 1$, we say i has period ∞. A chain is **aperiodic** if all its states have period 1.

The period is defined so that the time taken to get from state i back to state i again is always a multiple of the period. In the Ehrenfest chain all states have period 2, as $\{n \geq 1 : p^n(i,i) > 0\} = \{2, 4, 6, \dots\}$. The same is true for the simple random walk on $\mathbb{Z}$. The two-state Markov chain with $0 < p, q < 1$ is clearly aperiodic, as $p(i,i) > 0$ for each state i. More generally, if the diagonal entries of the transition matrix are all positive, then the chain is aperiodic.

The fact that all states in the Markov chains we just mentioned have the same period is no coincidence, as these chains are irreducible and periodicity is a class property. This is the content of the following result.

Lemma 2.3.3. *If $i \leftrightarrow j$, then i and j have the same period. In particular, all states in an irreducible set have the same period.*

Proof. Let t_i be the period of i and t_j that of j. Since i and j communicate, there exist n_1 and n_2 such that $p^{n_1}(i,j)$ and $p^{n_2}(j,i)$ are both positive. The Chapman-Kolmogorov equation implies that

$$p^{n_1+n_2}(i,i) \geq p^{n_1}(i,j)p^{n_2}(j,i) > 0$$

and so t_i divides $n_1 + n_2$. Let now $m \in \{n \geq 1 : p^n(j,j) > 0\}$. A double application of the Chapman-Kolmogorov equation gives

$$p^{n_1+m+n_2}(i,i) \geq p^{n_1}(i,j)p^m(j,j)p^{n_2}(j,i) > 0.$$

Therefore, t_i divides $n_1 + m + n_2$. But we have seen that t_i divides $n_1 + n_2$ and so it must divide $n_1 + m + n_2 - (n_1 + n_2) = m$. Since t_i divides all elements of $\{n \geq 1 : p^n(j,j) > 0\}$ and t_j is the largest number with such property, we then have $t_j \geq t_i$. Exchanging i and j in the previous argument, we obtain that $t_i \geq t_j$ and so $t_i = t_j$. □

Exercise 2.3.4. *Determine the period of each state for the Markov chain in Exercise 2.1.39.*

Exercise 2.3.5. *Suppose an irreducible Markov chain has a transition matrix* $\mathbf{P}$ *such that* $\mathbf{P}^2 = \mathbf{P}$. *Show that the chain is aperiodic.*

We can finally state the following remarkable result. If we exclude the obstructions we have identified so far (reducibility and periodicity), the existence of a stationary distribution implies convergence.

Theorem 2.3.6 (Convergence theorem). *If an irreducible and aperiodic Markov chain admits a stationary distribution* π, *then the chain is recurrent,* π *is unique and it is given by*

$$\pi_j = \lim_{n\to\infty} p^n(i,j),$$

where all π_j*'s are positive. In other words,* $\mathbf{P}^n$ *converges to the matrix whose rows are the stationary distribution vector.*

Proof. We show only the easy part, namely that if an irreducible Markov chain admits a stationary distribution π, then the chain is recurrent. Suppose this is not the case. Then, by Theorem 2.1.28, the chain is transient and $\lim_{n\to\infty} p^n(i,j) = 0$ for all states i and j. But since π is a stationary distribution, $\pi\mathbf{P}^n = \pi$, for each $n \geq 1$. Therefore,

$$\pi_i = \sum_j \pi_j \cdot p^n(j,i) \to 0 \quad \text{as} \quad n \to \infty,$$

contradicting the fact that $\sum_i \pi_i = 1$. As already mentioned, the fact that we can exchange the sum and the limit in the countable case requires justification, but we will not go into details. We simply know we can do that if the state space is finite. □

The Convergence theorem tells us that, for an irreducible and aperiodic Markov chain (with finite or infinite state space), the existence of a stationary distribution π ensures that the Markov chain will converge to π as $n \to \infty$. However, recall that if the state space is infinite, it is not guaranteed that a stationary distribution exists. We have seen this behavior for a periodic Markov chain (Example 2.2.5) and in fact it may occur even if the chain is aperiodic, as the following example shows.

Example 2.3.7. Consider again the Markov chain in Example 2.1.6. Checking the transition graph, the chain is clearly irreducible and aperiodic. However, it can be easily seen that it admits no stationary distribution.

If a stationary distribution does not exist, then the following occurs:

Theorem 2.3.8. *If a Markov chain is irreducible and aperiodic but admits no stationary distribution, then* $\lim_{n\to\infty} p^n(i,j) = 0$ *for all* i *and* j.

We have seen that if an irreducible Markov chain is transient, then $\lim_{n\to\infty} p^n(i,j) = 0$ for all i and j. In view of Theorem 2.3.8, it would then be tempting to guess that the absence of a stationary distribution is caused by transient states. It turns out this is not entirely correct: There are certain recurrent states exhibiting a similarly bad behavior. We know that every recurrent state is visited infinitely often almost surely, but the expected return time might be finite or not. This leads to the following distinction:

Definition 2.3.9. Let j be a recurrent state and let

$$\mu_j = \sum_{n=1}^{\infty} n f_{j,j}^{(n)}.$$

j is **positive recurrent** if μ_j converges and **null recurrent** otherwise.

For a recurrent state j, we can think of μ_j as the expected number of steps to first return to j given that $X_0 = j$. Since in general there is no upper bound on the number of steps to first return, this is not the expectation of a random variable as defined in Section 1.4. However, it is worth to know that we could generalize our definition of random variables so that μ_j becomes indeed the expectation of a random variable. The following result gives a characterization of positive and null recurrence.

Lemma 2.3.10. *Let j be a recurrent state and suppose that $\lim_{n\to\infty} p^n(j,j) = u$. Then $u > 0$ if and only if j is positive recurrent, in which case $u = 1/\mu_j$.*

Consider an irreducible and aperiodic Markov chain. If the chain is transient, then the Convergence theorem tells us it cannot admit a stationary distribution. On the other hand, if the chain is recurrent, then two possibilities may arise. If it admits no stationary distribution, then Theorem 2.3.8 tells us that $\lim_{n\to\infty} p^n(i,j) = 0$ for all i, j and Lemma 2.3.10 tells us that the chain is null recurrent. But if it admits a stationary distribution π then, by the Convergence theorem, $\lim_{n\to\infty} p^n(j,j) = \pi_j > 0$ for each j and so the chain is positive recurrent by Lemma 2.3.10. We can summarize this as follows:

For an irreducible and aperiodic Markov chain, exactly one of the following occurs:

1. The chain is transient, it admits no stationary distribution, $\lim_{n\to\infty} p^n(i,j) = 0$ for all i, j and in fact $\sum_n p^n(i,j) < \infty$.
2. The chain is recurrent, it admits no stationary distribution, $\lim_{n\to\infty} p^n(i,j) = 0$ for all i, j but $\sum_n p^n(i,j) = \infty$ and $\mu_j = \infty$.
3. The chain is recurrent, it admits a unique stationary distribution π, $\lim_{n\to\infty} p^n(i,j) = \pi_j > 0$ for all i, j and $\mu_j = 1/\pi_j < \infty$.

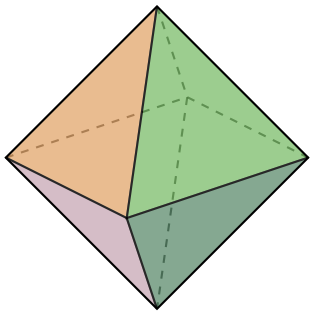
These three situations correspond to the irreducible chain being transient, null recurrent or positive recurrent, respectively.
Moreover, in the positive recurrent case, we can compute the expected number of steps to first return to j, which we interpreted to be μ_j, simply by computing the stationary distribution.

Example 2.3.11. Consider again the Markov chain in Example 2.1.6. What is the expected number of steps to first return to 1? Well, we figured out that $f_{1,1}^{(n)} = 1/n(n+1)$. Therefore,

$$\mu_1 = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty$$

and the expected number of steps to first return to 1 is not finite. We could have reached the same conclusion by recalling that the chain is recurrent and aperiodic but has no stationary distribution.

Example 2.3.12. A particle moves on the vertices of an octahedron in the following way: at each step the particle is equally likely to move to each of the adjacent vertices, independently of its past motion. Let i be the initial vertex occupied by the particle. What is the expected number of steps until the particle returns to i?



We immediately recognize this as a random walk on a graph (Example 2.2.9). The octahedron has 6 vertices, each of degree 4, and so the sum of the degrees of all vertices is $4 \cdot 6$. The random walk on the octahedron is clearly irreducible. It is also aperiodic. Indeed, for each vertex i, both $p^3(i,i)$ and $p^4(i,i)$ are positive. Since the stationary distribution vector has all entries equal to $1/6$, we have that the expected number of steps until return is 6.

Exercise 2.3.13. *Show that the random walk on a graph G is aperiodic if and only if the graph G contains a cycle of odd length.*

Exercise 2.3.14. *Consider the Markov chain in Example 2.1.29. Compute the expected number of steps to first return to state 3 in two ways: first by using Definition 2.3.9 and then by using the previous discussion.*

Exercise 2.3.15. *Consider the Markov chains with state space $\{1,2,3\}$ and transition matrices*

$$\begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

In both cases compute the expected number of steps to first return 1.

Under the convention $1/\infty = 0$, the relation $\lim_{n\to\infty} p^n(i,j) = 1/\mu_j$ in a recurrent chain has another interesting consequence. Let $V_n(j)$ be the number of visits to j up to time $n-1$. We mentioned that we can think of μ_j as the average number of steps to first return to j given that $X_0 = j$. But then, if the time from one visit to the next is about μ_j, we expect that $V_n(j)/n$ should be about $1/\mu_j$. Let's verify this. What is the expectation of $V_n(j)/n$? Well, we can write $V_n(j)$ as a sum of indicator random variables

$$V_n(j) = \sum_{k=0}^{n-1} I_{\{X_k = j\}}$$

and so

$$\mathbb{E}\Big(\frac{V_n(j)}{n} | X_0 = i\Big) = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}(I_{\{X_k=j\}} | X_0 = i) = \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{P}(X_k = j | X_0 = i) = \frac{1}{n}\sum_{k=0}^{n-1} p^k(i,j).$$

We can then use the following result from Analysis:

Lemma 2.3.16 (Cesàro's lemma). *If a real sequence $\{a_k\}$ converges to a, then the sequence of the partial averages $\{n^{-1}\sum_{k=0}^{n-1} a_k\}$ converges to a as well.*

Since $\lim_{n\to\infty} p^n(i,j) = 1/\mu_j$, Cesàro's lemma implies that

$$\lim_{n\to\infty} \mathbb{E}\Big(\frac{V_n(j)}{n}|X_0 = i\Big) = \frac{1}{\mu_j}.$$

The expected proportion of time that the chain is in state j converges to the inverse of the expected number of steps to first return to j.

Exercise 2.3.17. *A king confined to a 5×5 chessboard instantaneously makes standard king's moves each second in such a way that it is equally likely to move to any of the squares one move away from it. What long-run fraction of the time does it occupy the center square?*

We have seen in Example 2.3.7 that an irreducible and aperiodic Markov chain might not admit a stationary distribution. The example we gave had an infinite state space. The goal now is to show that if the state space is finite, then a (unique) stationary distribution always exists and this stationary distribution is the limiting distribution. The following result can then be viewed as the finite case of the Convergence theorem.

Theorem 2.3.18. *An irreducible and aperiodic Markov chain with finite state space has a unique stationary distribution π given by*

$$\pi_j = \lim_{n\to\infty} p^n(i,j).$$

Proof. We assume the existence of $\lim_{n\to\infty} p^n(j,j)$ for each j (which is the hard part of the proof) and show the remaining assertions. Namely,

(a) For any $i \neq j$,
$$\lim_{n\to\infty} p^n(i,j) = \lim_{n\to\infty} p^n(j,j).$$

(b) Letting $\pi_j = \lim_{n\to\infty} p^n(i,j)$ gives a stationary distribution.

(c) The stationary distribution is unique.

Proof of (a). Recall the first-passage decomposition

$$p^n(i,j) = \sum_{m=1}^{n} f_{i,j}^{(m)} \cdot p^{n-m}(j,j)$$

and that $f_{i,j} = \sum_{m=1}^{\infty} f_{i,j}^{(m)}$ is the probability that the chain ever visits state j starting in i. Since the chain is irreducible and finite, all states are recurrent (Lemma 2.1.34) and so

$$1 = f_{i,j} = \sum_{m=1}^{\infty} f_{i,j}^{(m)} = \lim_{n\to\infty} \sum_{m=1}^{n} f_{i,j}^{(m)}.$$

But then, for any $\varepsilon > 0$, there exists $m_1 \in \mathbb{N}$ such that $\sum_{m=1}^{m_1} f_{i,j}^{(m)} \geq 1 - \varepsilon$. Therefore,

$$\begin{aligned}\lim_{n\to\infty} p^n(i,j) = \lim_{n\to\infty} \sum_{m=1}^{n} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) \geq \lim_{n\to\infty} \sum_{m=1}^{m_1} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) &= \sum_{m=1}^{m_1} f_{i,j}^{(m)} \lim_{n\to\infty} p^{n-m}(j,j) \\ &= \sum_{m=1}^{m_1} f_{i,j}^{(m)} \lim_{n\to\infty} p^{n}(j,j) \\ &\geq (1-\varepsilon) \lim_{n\to\infty} p^n(j,j).\end{aligned}$$

Notice that we used finiteness to exchange sums and limits. We now use again the convergence of $\sum_{m=1}^{\infty} f_{i,j}^{(m)}$, this time as follows. Given our ε as above, there exists $m_2 \in \mathbb{N}$ such that $\sum_{m=m_2}^{\infty} f_{i,j}^{(m)} < \varepsilon$ (tails of a convergent series vanish). Then

$$\begin{aligned}p^n(i,j) = \sum_{m=1}^{n} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) &\leq \sum_{m=1}^{\infty} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) \\ &\leq \sum_{m=1}^{m_2} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) + \sum_{m=m_2}^{\infty} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) \\ &\leq \sum_{m=1}^{m_2} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) + \sum_{m=m_2}^{\infty} f_{i,j}^{(m)} \\ &< \sum_{m=1}^{m_2} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) + \varepsilon.\end{aligned}$$

But then

$$\lim_{n\to\infty} p^n(i,j) \leq \varepsilon + \lim_{n\to\infty} \sum_{m=1}^{m_2} f_{i,j}^{(m)} \cdot p^{n-m}(j,j) \leq \varepsilon + \sum_{m=1}^{m_2} f_{i,j}^{(m)} \cdot \lim_{n\to\infty} p^{n-m}(j,j) \leq \varepsilon + \lim_{n\to\infty} p^n(j,j),$$

where the last inequality follows from the fact that $\sum_{m=1}^{m_2} f_{i,j}^{(m)} \leq 1$. Combining

$$\lim_{n\to\infty} p^n(i,j) \geq (1-\varepsilon) \lim_{n\to\infty} p^n(j,j) \quad \text{with} \quad \lim_{n\to\infty} p^n(i,j) \leq \varepsilon + \lim_{n\to\infty} p^n(j,j),$$

we obtain that $\lim_{n\to\infty} p^n(i,j) = \lim_{n\to\infty} p^n(j,j)$ for each i and j.

Proof of (b). Let m be the number of states. We first show that $\pi_j = \lim_{n\to\infty} p^n(i,j)$ gives a distribution vector. Clearly, $\pi_j \geq 0$ for each j. Moreover, since $\sum_{j=1}^{m} p^n(i,j) = 1$, we have

$$1 = \lim_{n\to\infty} \sum_{j=1}^{m} p^n(i,j) = \sum_{j=1}^{m} \lim_{n\to\infty} p^n(i,j) = \sum_{j=1}^{m} \pi_j.$$

We now show that $\pi\mathbf{P} = \pi$. Since $p^{n+1}(i,j) = \sum_{k=1}^{m} p^n(i,k)p(k,j)$, letting $n \to \infty$, we obtain

$$\pi_j = \sum_{k=1}^{m} \pi_k \cdot p(k,j).$$

Proof of (c). Suppose ϕ is another stationary distribution. Then $\phi\mathbf{P}^n = \phi$, from which

$$\phi_j = \sum_{k=1}^{m} \phi_k \cdot p^n(k,j),$$

and letting $n \to \infty$, we obtain

$$\phi_j = \sum_{k=1}^{m} \phi_k \cdot \pi_j = \pi_j \sum_{k=1}^{m} \phi_k = \pi_j.$$

This concludes the proof. □

Remark 2.3.19. Notice that the interesting part in Theorem 2.3.18 is not the existence of a stationary distribution but rather the fact it is the limiting distribution. Indeed, any Markov chain with finite state space admits a stationary distribution. The reader familiar with Topology might want to check a nice proof of this result using Brouwer's fixed-point theorem![3]

Exercise 2.3.20. *Consider the Markov chain whose transition matrix is*

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

How many stationary distributions does it admit? Does this contradict Theorem 2.3.18?

We conclude this section with some remarks about positive and null recurrence. We know that any irreducible set C is either transient or recurrent. It turns out that if it is recurrent, then either all states in C are positive recurrent or all are null recurrent. In other words, positive and null recurrence are class properties.

Proposition 2.3.21. *Let C be an irreducible set. If C is recurrent, then either all states are positive recurrent or all are null recurrent.*

We have seen that every closed and finite communication class is recurrent. In fact something stronger is true: the class must be positive recurrent. This should not come as a surprise as it is natural to think that if the class is finite, the expected number of steps to first return to a given state should be finite.

Lemma 2.3.22. *If a closed communication class is finite, then it is positive recurrent. In particular, every irreducible Markov chain with finite state space is positive recurrent.*

The following is a particularly useful criterion for positive recurrence. It justifies our previous remark that the three distinct scenarios occurring for an irreducible and aperiodic Markov chain correspond to the cases of the chain being transient, positive recurrent or null recurrent.

Theorem 2.3.23. *An irreducible Markov chain is positive recurrent if and only if it admits a stationary distribution.*

Theorem 2.3.23 is a powerful tool for determining the nature of the states of an irreducible Markov chain, as the following examples show.

Example 2.3.24. Consider the symmetric random walk on $\mathbb{Z}$ (i.e., $p = q$). The chain is recurrent (Example 2.1.16) and admits no stationary distribution (Example 2.2.5). But then, being irreducible, Theorem 2.3.23 and Proposition 2.3.21 imply that it is null recurrent.

Example 2.3.25. Consider the two-state Markov chain with $0 < p, q < 1$. It is irreducible and admits a stationary distribution. Theorem 2.3.23 implies it is positive recurrent. The same holds for the Ehrenfest chain. Alternatively, we could have simply used Lemma 2.3.22.

[3] http://galton.uchicago.edu/~lalley/Courses/383/MarkovChains.pdf.

Example 2.3.26. We now use Theorem 2.3.23 to classify the states of the following Markov chain with state space $\{0, 1, 2, \dots\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} p_0 & p_1 & p_2 & p_3 & \cdots \\ 1 & & & & \\ 1 & & & & \\ 1 & & & & \\ \vdots & & & & \end{pmatrix},$$

where $\sum_i p_i = 1$ and $0 < p_i < 1$ for each i. Checking the transition graph, it is easy to see that the chain is irreducible. We now show it admits a stationary distribution π. If π exists, it has to satisfy the following system

$$\begin{cases} \pi_0 = \pi_0 p_0 + \sum_{i \geq 1} \pi_i & \\ \pi_k = \pi_0 p_k & \text{for each } k \geq 1 \\ \sum_i \pi_i = 1 & \end{cases}$$

Combining the first equation with the last, we get $\pi_0 = \pi_0 p_0 + (1 - \pi_0)$, from which $\pi_0 = \frac{1}{2-p_0}$. But then the values of π_k with $k \geq 1$ are given by $\pi_k = \frac{p_k}{2-p_0}$. In conclusion, a stationary distribution exists and so the chain is positive recurrent.

2.4 Absorbing chains

We now restrict ourselves to Markov chains with finite state space. Recall that a state i is absorbing if $\{i\}$ is a closed set. Therefore, $p(i, i) = 1$ and i is in particular recurrent.

Definition 2.4.1. A Markov chain is **absorbing** if it has at least one absorbing state and all non-absorbing states are transient.

Suppose our absorbing Markov chain has k states: t transient and $k - t$ absorbing. How does the transition matrix look like? We can label rows and columns of the transition matrix so that the transient states appear first. The transition matrix then looks like

$$\mathbf{P} = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right),$$

where Q is a $t \times t$ matrix indexed by the transient states, R is a $t \times (k - t)$ matrix, 0 is the $(k - t) \times t$ matrix whose entries are all zero and I is the $(k - t) \times (k - t)$ identity matrix. We have written $\mathbf{P}$ as a **block matrix**, where the submatrices $Q, 0, R, I$ are its blocks. These blocks can be treated as matrix entries while doing the usual matrix operations (check this!). We then have that

$$\mathbf{P}^2 = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right)\left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right) = \left(\begin{array}{c|c} Q^2 + R0 & QR + RI \\ \hline 0Q + I0 & 0R + I^2 \end{array}\right) = \left(\begin{array}{c|c} Q^2 & (Q + I)R \\ \hline 0 & I \end{array}\right).$$

In general, it easy to see by induction that

$$\mathbf{P}^n = \left(\begin{array}{c|c} Q^n & (I + Q + \cdots + Q^{n-1})R \\ \hline 0 & I \end{array}\right).$$

What is $\lim_{n\to\infty} \mathbf{P}^n$? Well, if it exists it should be

$$\lim_{n\to\infty} \mathbf{P}^n = \left(\begin{array}{c|c} \lim_{n\to\infty} Q^n & \lim_{n\to\infty}(I+Q+\cdots+Q^{n-1})R \\ \hline 0 & I \end{array}\right).$$

But the submatrix Q is indexed by the transient states and we know that, for any transient states i and j, we have $\lim_{n\to\infty} p^n(i,j) = 0$ (Corollary 2.1.14). This implies that $\lim_{n\to\infty} Q^n = 0$. In order to study $\lim_{n\to\infty} \mathbf{P}^n$, it is then enough to study $\lim_{n\to\infty}(I+Q+\cdots+Q^{n-1})R$. The following result is the matrix analogue of the sum of a geometric series.

Lemma 2.4.2. *Let A be an $n\times n$ matrix such that $A^n \to 0$ as $n\to\infty$. Then*

$$\lim_{n\to\infty}(I+A+\cdots+A^{n-1}) = (I-A)^{-1}.$$

Proof. For fixed n, we have that

$$(I-A)(I+A+A^2+\cdots A^n) = I+A+A^2+\cdots+A^n-(A+A^2+\cdots+A^n+A^{n+1}) = I-A^{n+1}.$$

Assuming for a moment that $I-A$ is invertible, we have that

$$I+A+A^2+\cdots+A^n = (I-A)^{-1}(I-A^{n+1})$$

and letting $n\to\infty$, we obtain the desired equality. It remains to show that $I-A$ is indeed invertible. Consider the linear system $(I-A)x = 0$. The invertibility of $I-A$ is equivalent to the fact that the only solution to this system is the zero vector $x=0$. We have that $0 = (I-A)x = x - Ax$ and so $x = Ax$. Iterating, we obtain

$$x = Ax = A(Ax) = A^2x = \cdots = A^nx,$$

for each $n\geq 1$. But then, passing to the limit, $x = \lim_{n\to\infty} A^nx = 0$. □

We can apply Lemma 2.4.2 to our setting as follows: Since we know that $Q^n\to 0$ as $n\to\infty$, we obtain that

$$\lim_{n\to\infty} \mathbf{P}^n = \left(\begin{array}{c|c} 0 & (I-Q)^{-1}R \\ \hline 0 & I \end{array}\right).$$

The moral is the following:

> The limiting submatrix $(I-Q)^{-1}R$ is indexed by transient rows and absorbing columns. Its (i,j)-entry is the long-term probability that the chain started in transient state i is absorbed in state j.

Definition 2.4.3. For an absorbing Markov chain, $(I-Q)^{-1}$ is called the **fundamental matrix**.

Example 2.4.4 (Gambler's ruin again). We can introduce a Markov chain in Example 1.3.6 as follows. We let X_n to be the gambler's fortune after the n-th toss. $\{X_n\}_{n\geq 0}$ gives a Markov chain with state space $\{0,\ldots,N\}$ and transition probabilities $p(i,i+1) = 1/2 = p(i,i-1)$, for each $1\leq i\leq N-1$, and $p(0,0) = 1 = p(N,N)$ (these are obtained by translating the rules of the gamble).

We then have that 0 and N are absorbing states and all the others are transient as $\{1,2,\ldots,N-1\}$ is not closed. The transition matrix $\mathbf{P}$ can then be written in block form as above. In this way, we can compute the probability that the gambler is ultimately bankrupted given that he starts with k units

Figure 2.10: Transition graph of gambler's ruin. States 0 and N are absorbing.

(recall we computed this in Example 1.3.6 by conditioning on the first gamble). We work out the case $N = 5$. The transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 0 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 0 & 5 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

We then have that

$$Q = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 \end{pmatrix} \end{array}$$

and

$$R = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 5 \end{array} \\ \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix} \end{array}$$

The probability that the gambler is ultimately bankrupted given that he starts with 3 units is then the $(3, 1)$ entry of $(I - Q)^{-1}R$. Similarly, the probability that the gambler ultimately reaches the amount $N = 5$ given that he starts with 3 units is the $(3, 2)$-entry of $(I - Q)^{-1}R$.

The fundamental matrix contains other important information related to an absorbing Markov chain:

Theorem 2.4.5. *Consider an absorbing Markov chain with k transient states and let $F = (I - Q)^{-1}$ be the $k \times k$ fundamental matrix. Then the (i, j)-entry F_{ij} of F is the expected number of visits to j given that the chain starts in i.*

The theorem above has the following important consequence. What is the expected number of steps from a certain transient state i until absorption? Well, for an absorbing Markov chain started in transient state i, the expected number of steps to reach an absorbing state is the sum of the number of transitions from i to each of the transient states, namely $\sum_{k \text{ transient}} F_{ik}$. To summarize what we have seen so far:

Absorption probability: The probability that from transient state i the chain is absorbed in j is the (i,j)-entry of $FR = (I-Q)^{-1}R$.

Absorption time: The expected number of steps from a transient state i until absorption is the sum of the entries on the i-th row of $F = (I-Q)^{-1}$.

Example 2.4.6 (Two year college). At a local two year college, 60% of freshmen become sophomores, 25% remain freshmen, and 15% drop out. 70% of sophomores graduate and transfer to a four year college, 20% remain sophomores and 10% drop out. What is the probability that a freshman will graduate?

We introduce a Markov chain with four states F, S, G, D, where the notation is self-explanatory. The description above tells us how the transition matrix looks like:

$$\mathbf{P} = \begin{array}{c} \\ F \\ S \\ G \\ D \end{array} \begin{array}{c} \begin{array}{cccc} F & S & G & D \end{array} \\ \begin{pmatrix} 0.25 & 0.6 & 0 & 0.15 \\ 0 & 0.2 & 0.7 & 0.1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

States G and D are obviously absorbing, whereas all the others are transient. The desired probability is nothing but the probability that from transient state F the chain is absorbed in G. We have seen that this can be computed as the entry corresponding to (F,G) in $(I-Q)^{-1}R$. But

$$Q = \begin{pmatrix} 0.25 & 0.6 \\ 0 & 0.2 \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 0 & 0.15 \\ 0.7 & 0.1 \end{pmatrix}.$$

Therefore, the desired probability is the $(1,1)$-entry of

$$(I-Q)^{-1}R = \begin{pmatrix} 0.75 & -0.6 \\ 0 & 0.8 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0.15 \\ 0.7 & 0.1 \end{pmatrix} = \frac{1}{0.75 \cdot 0.8} \begin{pmatrix} 0.8 & 0.6 \\ 0 & 0.75 \end{pmatrix} \begin{pmatrix} 0 & 0.15 \\ 0.7 & 0.1 \end{pmatrix},$$

which is

$$\frac{0.6 \cdot 0.7}{0.75 \cdot 0.8} = 0.7.$$

Consider now the following question. What is the expected time for a freshman to graduate or drop out? Well, this is the expected time from state F until absorption. We have seen that this value can be computed as the sum of the entries on the row corresponding to F in $(I-Q)^{-1}$ (don't confuse the state F with the fundamental matrix carrying the same name!). In other words, the sum of the entries on the first row of

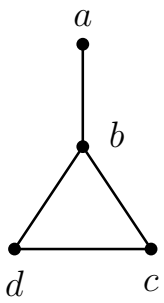
$$(I-Q)^{-1} = \begin{pmatrix} 0.75 & -0.6 \\ 0 & 0.8 \end{pmatrix}^{-1} = \frac{1}{0.75 \cdot 0.8} \begin{pmatrix} 0.8 & 0.6 \\ 0 & 0.75 \end{pmatrix},$$

which is

$$\frac{0.8}{0.75 \cdot 0.8} + \frac{0.6}{0.75 \cdot 0.8};$$

roughly 2.3 years.

The ideas above can be used to compute **expected hitting times**. Say that we have an irreducible Markov chain and we want to compute the expected time until state i is first hit. We modify the transition matrix $\mathbf{P}$ so that i becomes an absorbing state. By irreducibility and finiteness, all other states are now transient. But then we are in the setting introduced at the beginning of this section and the desired expected time is just the expected time until absorption in the modified chain. The following example is a summary of these type of problems.



Example 2.4.7. Consider the random walk on the graph depicted in figure and starting in a.

1. Find the expected number of steps to first return to a.
2. Find the expected number of steps to first hit d.
3. Find the probability that the walk hits c before d.

Consider 1. We can proceed in several ways. The transition matrix of the random walk is

$$\mathbf{P} = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix} \end{array}$$

Since the walk is irreducible, aperiodic (why?) and admits a stationary distribution π, the Convergence theorem and Lemma 2.3.10 tell us that the expected number of steps to first return to a is $1/\pi_a$, where π_a is the a-entry of the stationary distribution vector. But we know that

$$\pi_a = \frac{d(a)}{\sum_u d(u)} = \frac{1}{8}$$

and so the desired expectation is 8.

Alternatively, observe that after one step the walk is in b with probability 1 and so we can compute the expected number of steps to first return to a as the expected number of steps to first hit a starting in b plus 1. To compute the expected number of steps to first hit a starting in b, we first turn a into an absorbing state i.e., we consider the following modified Markov chain:

$$\mathbf{P}' = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix} \end{array}$$

We then compute the fundamental matrix $(I - Q)^{-1}$ of the new absorbing chain, where

$$Q = \begin{array}{c} \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{ccc} b & c & d \end{array} \\ \begin{pmatrix} 0 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} \end{array}$$

The desired expectation is the first row sum of $(I-Q)^{-1}$, which is 7. But then the expected number of steps to first return to a is 8 (as shown previously). To answer 2., we simply turn d in $\mathbf{P}$ into an absorbing state and proceed as in 1. The answer is $13/3$.

Consider now 3. We first make c and d absorbing states:

$$\mathbf{P}'' = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

In the modified chain with transition matrix $\mathbf{P}''$, a is transient and c and d are absorbing (notice however that in the original walk all states were positive recurrent). We then compute the probability that from transient state a the chain is absorbed in c. We know this is the (a,c)-entry of $(I-Q)^{-1}R$, where

$$Q = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1/3 & 0 \end{pmatrix} \end{array}$$

and

$$R = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} c & d \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1/3 & 1/3 \end{pmatrix} \end{array}$$

An easy computation gives the value $1/2$. But the modified walk, starting in a, is absorbed in c if and only if the original walk hits c before d. Therefore, the desired probability is exactly $1/2$. Notice that, in this specific example, we could have invoked symmetry. Indeed, either c is hit before d or d is hit before c, and these two events have equal probability given the symmetry of the graph.

An interesting special case of expected hitting times is the **expected time of sequence patterns in repeated experiments**. Suppose the elements of a set S are repeatedly sampled. A **pattern** is a sequence $s_n = p_1, \ldots, p_n$, where each p_i belongs to S. For example, we might repeatedly toss a coin. In this case, $S = \{H, T\}$ and the sequence H, T, H is a pattern. What is the expected time until a certain pattern s_n first appears? To answer this question, we introduce a Markov chain with state space $\{\varnothing, s_1, \ldots, s_n\}$, where each s_i is the subsequence of s_n consisting of the first i elements. We make s_n an absorbing state and we let X_n to be the largest subsequence of s_n appearing after the most recent samples. The desired expected time can then be computed using the technique above.

Example 2.4.8. Suppose a fair coin is tossed. What is the expected time until H, H first appears? We introduce a Markov chain with state space $\{\varnothing, H, HH\}$ and transition matrix

$$\mathbf{P} = \begin{array}{c} \\ \varnothing \\ H \\ HH \end{array} \begin{array}{c} \begin{array}{ccc} \varnothing & H & HH \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

The desired expected time is the sum of the entries on the first row of

$$(I-Q)^{-1} = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1 \end{pmatrix}^{-1} = 4 \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

The expected time is then 6.

Exercise 2.4.9. *A fair coin is tossed repeatedly until the sequence H, T, H appears. What is the expected number of tosses needed?*

Exercise 2.4.10. *A person has 3 umbrellas, some at office, some at home. Every day, he walks to the office in the morning and returns home in the evening. In each trip, he takes an umbrella with him only if it is raining. Suppose that, in every trip, the probability of rain is 0.2 (hence we know the gentleman is unlikely to live in Belfast).*

1. *What percentage of time does he get wet?*
2. *What is the expected number of trips until all umbrellas are at the same location?*

Chapter 3

Continuous Random Variables

In this section we are essentially going to revisit the notions introduced in Section 1.4, this time in the context of continuous random variables. Recall that a random variable X is continuous if its distribution function F_X can be expressed as

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u,$$

for some integrable function $f_X \colon \mathbb{R} \to [0, \infty)$ called the pdf of X. Moreover, if the distribution function is differentiable at $x \in \mathbb{R}$, then

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}.$$

The following properties hold for a continuous random variable X and its pdf f_X:

1. $\mathbb{P}(x < X \leq y) = \int_x^y f_X(u)\,\mathrm{d}u$.
2. $\mathbb{P}(X = x) = 0$.
3. $\int_{-\infty}^{\infty} f_X(u)\,\mathrm{d}u = 1$.
4. $\mathbb{P}(X \in B) = \int_B f_X(u)\,\mathrm{d}u$, for every $B \subseteq \mathbb{R}$ for which the (Riemann) integral exists.

Definition 3.0.1. Let X be a continuous random variable with pdf $f_X(x)$. The **expectation** of X is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x,$$

provided that the following improper integral converges:

$$\int_{-\infty}^{\infty} |x f_X(x)|\,\mathrm{d}x < \infty.$$

Remark 3.0.2. The convergence of the improper integral guarantees that the expectation is well-defined and finite.

We have seen that if X is an arbitrary random variable and $g \colon \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X)$ is a random variable. In fact, if X is discrete, $g(X)$ is discrete for any function g. On the other hand, if X is continuous, $g(X)$ can be either continuous or discrete. The former occurs for example if g is the identity, the latter by taking g as follows (why?):

$$g(x) = \begin{cases} 1 & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Given the pdf of X, how do we compute the pdf of $g(X)$? The standard technique is illustrated in the following example.

Example 3.0.3. Given a continuous random variable X, we compute the pdf of $Y = X^2$. The idea is to first obtain the distribution function of Y and then differentiate. For each $y > 0$, we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}),$$

where in the last equality we used Lemma 1.4.29. Therefore,

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{f_X(\sqrt{y})}{2\sqrt{y}} + \frac{f_X(-\sqrt{y})}{2\sqrt{y}},$$

where the last equality follows from the Chain rule.

If we are just interested in the expectation of $g(X)$, we can however skip the computation of the pdf of $g(X)$, thanks to the following result. It is the continuous analogue of the Law of the unconscious statistician.

Theorem 3.0.4 (LOTUS). *If X and $g(X)$ are continuous random variables, then*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, \mathrm{d}x.$$

Example 3.0.5. Let X be a continuous random variable with pdf

$$f_X(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Compute the expectation of $Y = X^2$. We can proceed in two ways. Either we compute the pdf of Y and use the definition of expectation or just apply LOTUS. As for the first way, we have seen in Example 3.0.3 that

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} = \begin{cases} \dfrac{3y}{2\sqrt{y}} & \text{if } 0 < y < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y)\, \mathrm{d}y = \int_0^1 \frac{3}{2} y^{3/2}\, \mathrm{d}y = \frac{3}{5}.$$

Using LOTUS, we immediately get

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} x^2 f_X(x)\, \mathrm{d}x = \int_0^1 3x^4\, \mathrm{d}x = \frac{3}{5}.$$

Definition 3.0.6. A continuous random variable X with pdf given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

is called **exponential** with parameter λ.

The function $f_X(x)$ above is a legitimate pdf. Indeed,

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \int_0^{\infty} \lambda e^{-\lambda x}\,\mathrm{d}x = 1.$$

Moreover, for any $a \geq 0$,

$$\mathbb{P}(X \geq a) = \int_a^{\infty} \lambda e^{-\lambda x}\,\mathrm{d}x = e^{-\lambda a} \tag{3.1}$$

and so the probability that X exceeds a falls exponentially. The exponential random variable is a good model for the amount of time until a certain event occurs. For example, the amount of time until a piece of equipment breaks down, until a car accident occurs or until the next earthquake. Using integration by parts, it is easy to see that the expectation of an exponential random variable X with parameter λ is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_0^{\infty} x\lambda e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}.$$

Example 3.0.7. Suppose that the duration of a phone call in minutes is an exponential random variable X with parameter $\lambda = 1/10$. What is the probability that the phone call lasts more than 10 minutes? This is just $\mathbb{P}(X > 10) = e^{-1}$. Suppose now we know that the phone call has already lasted 10 minutes. What is the probability that it will last at least 10 more minutes? The probability we are interested in is

$$\mathbb{P}(X > 20 | X > 10) = \frac{\mathbb{P}(X > 20, X > 10)}{\mathbb{P}(X > 10)} = \frac{\mathbb{P}(X > 20)}{\mathbb{P}(X > 10)} = \frac{e^{-2}}{e^{-1}} = e^{-1}.$$

The same argument used in the previous example shows that if Y is exponential, then

$$\mathbb{P}(Y > t | Y > s) = \mathbb{P}(Y > t - s),$$

for each $t > s$. But this is the lack of memory property and we have seen that, in the discrete world, the geometric random variable has this property. It turns out that the exponential random variable can be viewed as the continuous analogue of the geometric random variable in the following sense. Suppose X is a geometric random variable with parameter p, for p small and recall that $\mathbb{P}(X > n) = (1-p)^n$. Let's now consider the rescaled random variable $X/\mathbb{E}(X) = pX$. We have that

$$\mathbb{P}(pX > t) = \mathbb{P}(X > t/p) \approx (1-p)^{\frac{t}{p}} \approx e^{-t},$$

where we used the fact that $e = \lim_{n\to\infty}(1 + 1/n)^n$. In other words, the rescaled geometric pX behaves like the exponential with parameter $\lambda = 1$ when p is small.

Definition 3.0.8. The **variance** of a continuous random variable X is defined exactly as in the discrete case: $\mathrm{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$.

By LOTUS, the variance can be computed as follows:

$$\begin{aligned}
\mathrm{var}(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x - 2\mathbb{E}(X)\int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x + \mathbb{E}(X)^2 \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x \\
&= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2 \\
&= \mathbb{E}(X^2) - \mathbb{E}(X)^2,
\end{aligned}$$

exactly as in the discrete case.

Example 3.0.9. The variance of the exponential random variable X with parameter λ is $1/\lambda^2$. Indeed, we know that $\mathbb{E}(X) = 1/\lambda$. Moreover, by LOTUS and integration by parts, we have that

$$\mathbb{E}(X^2) = \int_0^{\infty} x^2 f_X(x)\,\mathrm{d}x = \frac{2}{\lambda^2}.$$

3.1 Multiple continuous random variables

Definition 3.1.1. Two continuous random variables X and Y admit **joint pdf** if there exists a nonnegative integrable function $f_{X,Y}\colon \mathbb{R}^2 \to \mathbb{R}$ such that

$$\mathbb{P}((X,Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y,$$

for every $B \subseteq \mathbb{R}^2$ for which the double integral exists.

Suppose X and Y admit joint pdf $f_{X,Y}$. If B is the rectangle $[a,b]\times[c,d] \subseteq \mathbb{R}^2$, then

$$\mathbb{P}(a \le X \le b, c \le Y \le d) = \int_c^d \int_a^b f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

On the other, letting $B = \mathbb{R}^2$, we obtain

$$1 = \mathbb{P}((X,Y) \in \mathbb{R}^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

As in the discrete case, we can recover the densities of X and Y from the joint pdf:

Lemma 3.1.2. *For continuous random variables X and Y with joint pdf $f_{X,Y}$, we have*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \quad \textit{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x.$$

Remark 3.1.3. Contrary to the discrete case, a joint pdf might not exist. Here is the intuition. Take X as the uniform random variable on $[0,1]$ and $Y = X$. Suppose that X and Y admit a joint pdf $f_{X,Y}$ and let $B = \{(x,y) \in [0,1]\times[0,1] : x = y\}$ be the main diagonal of the unit square. We have that $\mathbb{P}((X,Y)\in B) = 1$ and so

$$1 = \iint_{(x,y)\in B} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

But the double integral gives the volume under the surface $z = f_{X,Y}(x,y)$ and above B, which has area 0, and so it cannot be 1.

Example 3.1.4 (Two-dimensional uniform pdf). Suppose $S \subseteq \mathbb{R}^2$ has finite area. We say that the pair (X,Y) of continuous random variables is **uniformly distributed** over S if the joint pdf of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\mathrm{area}(S)} & \text{if } (x,y)\in S; \\ 0 & \text{otherwise.} \end{cases}$$

We then have that, for $B \subseteq \mathbb{R}^2$,

$$\mathbb{P}((X,Y)\in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = \frac{1}{\mathrm{area}(S)}\iint_{(x,y)\in B\cap S}\mathrm{d}x\mathrm{d}y = \frac{\mathrm{area}(B\cap S)}{\mathrm{area}(S)}.$$

Suppose that a point is chosen at random from an open unit disk. What is the probability that the sum of its coordinates is larger than 1?

Let $B_1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ be the open unit disk centered at the origin and let (X, Y) be the coordinates of our random point. It is reasonable to assume that (X, Y) is uniformly distributed over B_1:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi} & \text{if } (x, y) \in B_1; \\ 0 & \text{otherwise.} \end{cases}$$

We need to compute $\mathbb{P}(X + Y > 1)$. It is then enough to compute the area of the intersection between B_1 and the half-plane $\{(x, y) \in \mathbb{R}^2 : x + y > 1\}$. Drawing a picture, it is easy to see that this area is $\frac{\pi}{4} - \frac{1}{2}$ and so $\mathbb{P}(X + Y > 1) = \frac{1}{4} - \frac{1}{2\pi}$.

Example 3.1.5 (Buffon's needle). We throw a needle of length ℓ at random on a surface marked with horizontal lines at distance d (see Figure 3.1). Assume $\ell < d$, so that the needle can intersect at most one horizontal line. What is the probability that the needle will intersect one of these lines?

Figure 3.1

Consider the midpoint of the needle and the vertical segment between the midpoint and the closest horizontal line (the dotted lines in Figure 3.1). Let X be the length of this segment and let Θ be the acute angle between the needle and the segment. The pair of random variables (X, Θ) uniquely determines the position of the needle and we may assume it is uniformly distributed over $R = [0, \frac{d}{2}] \times [0, \frac{\pi}{2}]$. We then have that

$$f_{X,\Theta}(x, \theta) = \begin{cases} \frac{4}{\pi d} & \text{if } (x, \theta) \in R; \\ 0 & \text{otherwise.} \end{cases}$$

The needle will intersect one of the lines if and only if

$$\frac{X}{\cos \Theta} < \frac{\ell}{2}.$$

Therefore, the desired probability is

$$\iint_{(x,\theta)\in A} f_{X,\Theta}(x, \theta)\, \mathrm{d}x\mathrm{d}\theta,$$

where $A = \{(x, \theta) \in \mathbb{R}^2 : 0 \le x \le d/2, 0 \le \theta \le \pi/2, x < \ell \cos\theta/2\}$. We then have that

$$\iint_{(x,\theta)\in A} f_{X,\Theta}(x, \theta)\, \mathrm{d}x\mathrm{d}\theta = \int_0^{\frac{\pi}{2}} \int_0^{\frac{\ell\cos\theta}{2}} \frac{4}{\pi d}\, \mathrm{d}x\mathrm{d}\theta = \frac{2\ell}{\pi d}.$$

This formula suggests a way to calculate π: Throw the needle a large number of times, count the number of intersections in the first n tosses and divide by n. This will give an estimate of the true probability $2\ell/\pi d$ and so

$$\pi \sim \frac{2n\ell}{\#\{\text{intersections in first } n \text{ tosses}\} \cdot d}.$$

A generalized version of the LOTUS still holds:

Lemma 3.1.6. *Let X and Y be continuous random variables with joint pdf $f_{X,Y}(x,y)$ and let $g\colon \mathbb{R}^2 \to \mathbb{R}$. Then*

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

In particular, $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.

3.2 Conditioning continuous random variables

Definition 3.2.1. Let X be a continuous random variable and let A be an event with $\mathbb{P}(A) > 0$. The **conditional pdf** of X is the nonnegative integrable function $f_{X|A}$ satisfying

$$\mathbb{P}(X \in B|A) = \int_B f_{X|A}(x)\,\mathrm{d}x,$$

for every $B \subseteq \mathbb{R}$ for which the integral exists.

It is the same as an ordinary pdf except that it refers to a universe in which A has occurred. Suppose now we condition on an event of the form $X \in A$. We have that

$$\int_B f_{X|X\in A}(x)\,\mathrm{d}x = \mathbb{P}(X \in B|X \in A) = \frac{\mathbb{P}(X \in B, X \in A)}{\mathbb{P}(X \in A)} = \frac{\int_{A\cap B} f_X(x)\,\mathrm{d}x}{\mathbb{P}(X \in A)}.$$

Since this is true for every $B \subseteq \mathbb{R}$ for which the integrals exist, it must be that the integrands coincide on A i.e.,

$$f_{X|X\in A}(x) = \begin{cases} \dfrac{f_X(x)}{\mathbb{P}(X \in A)} & \text{if } x \in A; \\ 0 & \text{otherwise.} \end{cases}$$

This means that, within the conditioning set A, the conditional pdf has the same shape as the ordinary pdf: it is just rescaled by $1/\mathbb{P}(X \in A)$ so that $\int_A f_{X|X\in A}(x)\,\mathrm{d}x = 1$.

Definition 3.2.2. The **conditional expectation** of a continuous random variable X is defined as

$$\mathbb{E}(X|A) = \int_{-\infty}^{\infty} x f_{X|A}(x)\,\mathrm{d}x.$$

We then have the following continuous analogue of Corollary 1.8.3.

Theorem 3.2.3. *Let $A_1, \ldots, A_n$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i and let X be a continuous random variable. Then*

(a) $f_X(x) = \sum_{i=1}^n \mathbb{P}(A_i) f_{X|A_i}(x)$.

(b) $\mathbb{E}(X) = \sum_{i=1}^n \mathbb{P}(A_i)\mathbb{E}(X|A_i)$.

Example 3.2.4. The metro train arrives every 15 minutes starting at 6am. You walk into the station every morning between 7:10am and 7:30am with the time of arrival in this interval being a uniform random variable. What is the pdf of the time you have to wait for the first train?

Let X be the time of arrival. We know it is a uniform random variable on the interval between 7:10 and 7:30. Let Y be the waiting time. We want $f_Y(y)$. As the waiting time depends on whether you manage to take the 7:15 train or not, we consider the following partition:

$$A_1 = \{7:10 \leq X \leq 7:15\} \quad A_2 = \{7:15 < X \leq 7:30\}.$$

Conditioned on A_1, the arrival time is uniform on the interval 7:10-7:15 and so the waiting time is uniform on $[0, 5]$. In other words,

$$f_{Y|A_1}(y) = \begin{cases} \frac{1}{5} & \text{if } 0 \leq y \leq 5; \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, conditioned on A_2, the arrival time is uniform on the interval 7:15-7:30 and so the waiting time is uniform on $[0, 15]$. In other words,

$$f_{Y|A_2}(y) = \begin{cases} \frac{1}{15} & \text{if } 0 \leq y \leq 15; \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 3.2.3, we have that

$$f_Y(y) = \mathbb{P}(A_1) f_{Y|A_1}(y) + \mathbb{P}(A_2) f_{Y|A_2}(y).$$

Since X is uniform on the interval 7:10-7:30 of length 20, we know that $\mathbb{P}(A_1) = 5/20$ and $\mathbb{P}(A_2) = 15/20$. Combining, we obtain

$$f_Y(y) = \begin{cases} 1/10 & \text{if } 0 \leq y \leq 5; \\ 1/20 & \text{if } 5 < y \leq 15. \end{cases}$$

Continuing the analogy with discrete random variables, we would now like to condition on events of the form $Y = y$. But we know that if Y is continuous, $\mathbb{P}(Y = y) = 0$. How do we interpret then probabilities of the form $\mathbb{P}(X \in A | Y = y)$? We will make use of the following notion.

Definition 3.2.5. Let X and Y be continuous random variables with joint pdf $f_{X,Y}$. For any fixed y with $f_Y(y) > 0$, the **conditional pdf of X given that $Y = y$** is defined by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Notice that it agrees with the definition of conditional pmf in the discrete case. Viewing $f_{X|Y}(x|y)$ as a function of x, it has the same shape as $f_{X,Y}$. The normalization by $f_Y(y)$ implies that $f_{X|Y}(x|y)$ is a legitimate pdf:

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x, y)}{f_Y(y)}\, \mathrm{d}x = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}x = \frac{f_Y(y)}{f_Y(y)} = 1.$$

But how do we interpret $f_{X|Y}(x|y)$? Fix small δ_1 and δ_2 and consider the following conditional probability:

$$\begin{aligned}\mathbb{P}(x \le X \le x+\delta_1 | y \le Y \le y+\delta_2) &= \frac{\mathbb{P}(x \le X \le x+\delta_1, y \le Y \le y+\delta_2)}{\mathbb{P}(y \le Y \le y+\delta_2)} \\ &= \frac{\int_x^{x+\delta_1}\int_y^{y+\delta_2} f_{X,Y}(x,y)\,\mathrm{d}y\mathrm{d}x}{\int_y^{y+\delta_2} f_Y(y)\,\mathrm{d}y} \\ &\approx \frac{f_{X,Y}(x,y)\delta_1\delta_2}{f_Y(y)\delta_2} \\ &= f_{X|Y}(x|y)\delta_1.\end{aligned}$$

Letting $\delta_2 \to 0$, we have that $\mathbb{P}(x \le X \le x+\delta_1 | Y = y)$ should approximately be $f_{X|Y}(x|y)\delta_1$ for small δ_1. We then make the following definition in the continuous case:

$$\mathbb{P}(X \in A | Y = y) \stackrel{\text{def}}{=} \int_A f_{X|Y}(x|y)\,\mathrm{d}x.$$

Example 3.2.6. We throw a dart at a circular target of radius r. We assume that we always hit the target and that all points of impact (X, Y) are equally likely. In other words, we assume that the joint pdf of X and Y is

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\pi r^2} & \text{if } x^2+y^2 \le r^2; \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional pdf $f_{X|Y}(x|y)$? We first compute the marginal $f_Y(y)$. By Lemma 3.1.2,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \begin{cases} 0 & \text{if } |y| > r; \\ \displaystyle\int_{x:x^2+y^2\le r^2} \frac{1}{\pi r^2}\,\mathrm{d}x = \int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} \frac{1}{\pi r^2}\,\mathrm{d}x = \frac{2}{\pi r^2}\sqrt{r^2-y^2} & \text{if } |y| \le r. \end{cases}$$

Therefore, $f_{X|Y}(x|y) = \frac{1}{2\sqrt{r^2-y^2}}$.

Definition 3.2.7. The conditional expectation $\mathbb{E}(X|Y=y)$ is defined as $\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x$.

We have that $\mathbb{E}(g(X)|Y=y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\,\mathrm{d}x$. Moreover, the following version of the total expectation theorem holds:

Theorem 3.2.8. *Let X and Y be continuous random variables. Then*

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \mathbb{E}(X|Y=y) f_Y(y)\,\mathrm{d}y.$$

Independence is defined exactly as in the discrete case.

Definition 3.2.9. Two continuous random variables X and Y are **independent** if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for each x, y.

Exactly as in the discrete case, if X and Y are independent, then the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. Indeed,

$$\begin{aligned}\mathbb{P}(X \in A, Y \in B) &= \int_{X\in A}\int_{Y\in B} f_{X,Y}(x,y)\,\mathrm{d}y\mathrm{d}x \\ &= \int_{X\in A}\int_{Y\in B} f_X(x)f_Y(y)\,\mathrm{d}y\mathrm{d}x \\ &= \int_{X\in A} f_X(x)\int_{Y\in B} f_Y(y)\,\mathrm{d}y\mathrm{d}x \\ &= \Big(\int_{Y\in B} f_Y(y)\,\mathrm{d}y\Big)\Big(\int_{X\in A} f_X(x)\,\mathrm{d}x\Big) \\ &= \mathbb{P}(Y \in B)\mathbb{P}(X \in A).\end{aligned}$$

The following properties, which were shown for discrete random variables, remain true in the continuous case. Proofs are similar and hence omitted.

Lemma 3.2.10. *Let X and Y be independent continuous random variables and let g and h be two functions such that $g(X)$ and $h(Y)$ are continuous. The following hold:*

- *$g(X)$ and $h(Y)$ are independent;*
- *$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$;*
- *$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$;*
- *$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.*

3.3 Normal random variables

Definition 3.3.1. A continuous random variable X is **normal** if it has pdf given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

for some parameters μ, σ with $\sigma > 0$. If $\mu = 0$ and $\sigma = 1$, X is called **standard normal**.

It can be shown that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = 1$$

and so $f_X(x)$ is a legitimate pdf. The parameter μ is the "center" of the density. Indeed, $f_X(x)$ is symmetric around μ i.e., $f_X(\mu + x) = f_X(\mu - x)$. The parameter σ is the "spread" of the density. The graph of $f_X(x)$ has a characteristic bell shape symmetric around the line $x = \mu$.

The importance of the normal random variable is mainly due to the Central limit theorem. Loosely speaking, it asserts that the distribution of the sum of a large number of i.i.d. random variables is approximated by the normal distribution.

Lemma 3.3.2. *If X is normal, then $\mathbb{E}(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$.*

Theorem 3.3.3. *Normality is preserved by linear transformations. Namely, if X is normal with mean μ and variance σ^2, then $Y = aX + b$ is normal with $\mathbb{E}(Y) = a\mu + b$ and $\mathrm{var}(Y) = a^2\sigma^2$.*

Proof. We look for the pdf of Y and obtain it by differentiating the distribution function. Suppose that $a > 0$ (the other case is similar).

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \mathbb{P}\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

By the Chain rule and using the fact that X is normal with mean μ and variance σ^2,

$$f_Y(y) = \frac{\mathrm{d}F_Y}{\mathrm{d}y}(y) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right) = \frac{1}{a} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{((y-b)/a-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}},$$

as claimed. □

If X is normal, we have that

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x.$$

Unfortunately, the function e^{-x^2} has no elementary antiderivative i.e., its antiderivative cannot be expressed as a sum, product, composition of finitely many polynomials, rational functions, trigonometric and exponential functions, and their inverse functions. On the other hand, in order to compute probabilities involving the normal random variable, we need to somehow compute the integral above. The lack of an elementary antiderivative is bypassed by computing approximations of the integral above, in the case $\mu = 0$ and $\sigma = 1$, via numerical integration. These approximated values are then stored in tables (see Figure 3.2) and allow to determine an approximate value of $\mathbb{P}(X \leq x)$, for each x. Notice that the distribution function $F_X(x) = \mathbb{P}(X \leq x)$ of a standard normal is usually denoted by $\Phi(x)$.

But in order to use these tables, how do we pass from a normal random variable X with parameters μ and σ to a standard normal? The answer is already in Theorem 3.3.3: The random variable $Y = \frac{X-\mu}{\sigma}$ is normal with mean $\mu = 0$ and variance $\sigma^2 = 1$. We can then use this linear transformation and its inverse to jump from generic normal to standard normal and vice versa.

Example 3.3.4. The annual snowfall at a particular location is modelled as a normal random variable with mean $\mu = 60$ (in inches) and $\sigma = 20$. What is the probability that this year's snowfall will be at least 80 inches?

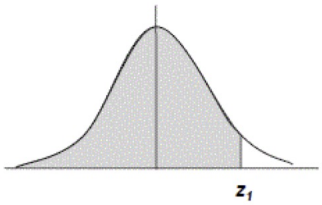
Let X be the snow accumulation. We need to compute $\mathbb{P}(X \geq 80) = 1 - \mathbb{P}(X \leq 80)$. To compute the latter, since we need to resort to the tables, we first pass to the standard normal random variable $Y = \frac{X-60}{20}$. We have that $\mathbb{P}(X \leq 80) = \mathbb{P}(20Y + 60 \leq 80) = \mathbb{P}(Y \leq 1)$. We then check the approximate value of $\mathbb{P}(Y \leq 1)$ in the tables: it is 0.8413 (see Figure 3.2).

Example 3.3.5. A binary message is transmitted as a signal which is either -1 or 1. The communication channel corrupts the transmission with additive normal noise with mean $\mu = 0$ and variance σ^2. The receiver concludes that the signal -1 (or 1) was transmitted if the value received is smaller than 0 (or at least 0). What is the probability of an error?

Let N be the noise and S be the signal. We have an error if -1 is transmitted and $N \geq 1$ (as this gives $N + S \geq 0$) or if 1 is transmitted and $N < -1$ (as this gives $N + S < 0$). We want to compute $\mathbb{P}(N \geq 1)$ and $\mathbb{P}(N < -1)$. As N is normal with $\mu = 0$, we know that these two values are the same. We then pass to the standard normal $N' = \frac{N-\mu}{\sigma} = \frac{N}{\sigma}$ and compute

$$\mathbb{P}(N \geq 1) = 1 - \mathbb{P}(N < 1) = 1 - \mathbb{P}(\sigma N' < 1) = 1 - \mathbb{P}(N' < 1/\sigma).$$

Standard Normal Distribution



$$p(z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{-\frac{1}{2}z^2}\, dz$$

z_1	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 3.2: Table storing values of $\Phi(z)$.

3.4 Moment generating functions

We now introduce moment generating functions. These functions are useful in several different ways. Among others, they provide an easy way of calculating the moments of a random variable, they provide tools for dealing with sums of independent random variables and for proving limit theorems, such as the

Central limit theorem.

Definition 3.4.1. The **moment generating function** of a random variable X is the function $M_X(s) = \mathbb{E}(e^{sX})$. Notice that M_X is defined only for those values of s for which $\mathbb{E}(e^{sX})$ exists i.e., $\mathbb{E}(e^{sX}) \in \mathbb{R}$.

Remark 3.4.2. By LOTUS, M_X can be written as follows: If X is discrete with pmf $f_X(x)$,

$$M_X(s) = \sum_x e^{sx} f_X(x)$$

and if X is continuous with pdf $f_X(x)$,

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, \mathrm{d}x.$$

Remark 3.4.3. Notice that M_X is always defined at 0: $M_X(0) = \mathbb{E}(1) = 1$. In fact, M_X is always defined in a neighborhood of 0. Indeed, suppose that $M_X(s_0)$ exists for some $s_0 > 0$ and let $s \in [0, s_0]$. On $\{X < 0\}$, we have that $e^{sX} \leq 1$, and on $\{X \geq 0\}$, we have that $e^{sX} \leq e^{s_0 X}$. Therefore, $0 \leq e^{sX} \leq 1 + e^{s_0 X}$ and so $\mathbb{E}(e^{sX}) \leq 1 + \mathbb{E}(e^{s_0 X}) < \infty$. Similarly, if $M_X(s_0)$ exists for some $s_0 < 0$, then $M_X(s)$ exists for all $s \in [s_0, 0]$.

Example 3.4.4. Compute the mgf of a Poisson random variable X with parameter λ. Recall that X has pmf $f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, for $x = 0, 1, 2, \ldots$. Therefore,

$$M_X(s) = \sum_{x=0}^{\infty} e^{sx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^s \lambda)^x}{x!} = e^{-\lambda} e^{e^s \lambda} = e^{\lambda(e^s - 1)}.$$

Example 3.4.5. Compute the mgf of a normal random variable X with mean μ and variance σ^2. Let us first consider the standard normal $Y = \frac{X - \mu}{\sigma}$ and suppose for a moment we know the mgf of Y. We can compute the mgf of X by recalling properties of expectation:

$$M_X(s) = \mathbb{E}(e^{s(\sigma Y + \mu)}) = \mathbb{E}(e^{s\sigma Y} \cdot e^{s\mu}) = e^{s\mu} \mathbb{E}(e^{s\sigma Y}) = e^{s\mu} M_Y(s\sigma).$$

We then compute the mgf of the standard normal Y:

$$\begin{aligned}
M_Y(s) &= \int_{-\infty}^{\infty} e^{sy} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, \mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^2 - 2sy)}\, \mathrm{d}y \\
&= \frac{e^{\frac{s^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^2 - 2sy + s^2)}\, \mathrm{d}y \\
&= \frac{e^{\frac{s^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y - s)^2}\, \mathrm{d}y \\
&= e^{\frac{s^2}{2}},
\end{aligned}$$

where the last equality follows from the fact that we are integrating over the real line the pdf $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-s)^2}$ of a normal random variable with parameters $\mu = s$ and $\sigma = 1$. This integral has then to be 1. Combining the two results above, we have that the mgf of X is

$$M_X(s) = e^{\frac{s^2 \sigma^2}{2} + s\mu}.$$

If we know the mgf of X, we can compute its moments. Hence the suggestive name.

Theorem 3.4.6. *The n-th derivative with respect to s of M_X evaluated at 0 gives the n-th moment $\mathbb{E}(X^n)$.*

The proof of this result is non-trivial and we just provide intuition for the case $n = 1$ and X continuous. We first differentiate both sides of the equality

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, \mathrm{d}x.$$

We obtain

$$\begin{aligned}\frac{\mathrm{d}M_X}{\mathrm{d}s}(s) &= \frac{\mathrm{d}}{\mathrm{d}s}\int_{-\infty}^{\infty} e^{sx} f_X(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}s} e^{sx} f_X(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} x e^{sx} f_X(x)\, \mathrm{d}x.\end{aligned}$$

Evaluating at $s = 0$, we obtain

$$\mathbb{E}(X) = \frac{\mathrm{d}M_X}{\mathrm{d}s}(0).$$

Notice that the interchange of integration and differentiation is not allowed in general and requires justification. The reason we can do this under our assumptions comes from a powerful theorem in Analysis, called the Dominated convergence theorem (hand-waving here).

Exercise 3.4.7. *Compute mean and variance of a normal random variable using its mgf.*

The following important theorem asserts that random variables are completely determined by their moment generating functions.

Theorem 3.4.8 (Inversion theorem). *Suppose that $M_X(s)$ exists for all $s \in [-a, a]$, where $a > 0$. Then M_X determines uniquely the distribution function of the random variable X. In particular, if $M_X(s) = M_Y(s)$ for all $s \in [-a, a]$, then X and Y have the same distribution function.*

Mgf are particularly useful when dealing with sums of independent random variables. Let X and Y be two independent random variables and let $W = X + Y$. We have that

$$M_W(s) = \mathbb{E}(e^{sW}) = \mathbb{E}(e^{s(X+Y)}) = \mathbb{E}(e^{sX} \cdot e^{sY}).$$

But for a fixed value of s, e^{sX} and e^{sY} are independent random variables and so

$$M_W(s) = \mathbb{E}(e^{sX} \cdot e^{sY}) = \mathbb{E}(e^{sX})\mathbb{E}(e^{sY}) = M_X(s)M_Y(s).$$

Similarly, if $X_1, \ldots, X_n$ are independent random variables and $W = X_1 + \cdots + X_n$, then

$$M_W(s) = M_{X_1}(s) \cdots M_{X_n}(s).$$

Example 3.4.9. Let X and Y be independent normal random variables with means μ_x, μ_y and variances σ_x^2, σ_y^2, respectively. Let $W = X + Y$. By Example 3.4.5 and the previous paragraph, we have that

$$M_W(s) = e^{\frac{s^2\sigma_x^2}{2} + s\mu_x} \cdot e^{\frac{s^2\sigma_y^2}{2} + s\mu_y} = e^{\frac{s^2(\sigma_x^2+\sigma_y^2)}{2} + s(\mu_x+\mu_y)}.$$

This implies that W has the same mgf as a normal with mean $\mu_x+\mu_y$ and variance $\sigma_x^2+\sigma_y^2$. Theorem 3.4.8 then implies that W *has to be* a normal with mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$.

Exercise 3.4.10. *Using moment generating functions, show that the sum of two independent Poisson random variables with parameters λ_1 and λ_2 is a Poisson with parameter $\lambda_1 + \lambda_2$. We already showed this in Example 1.9.3.*

3.5 Central limit theorem

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2 and let $S_n = X_1 + \cdots + X_n$. The Weak law of large numbers tells us that the distribution of S_n/n concentrates around its mean μ as n becomes large. The Central limit theorem goes further and quantifies the behavior of the "fluctuations" of S_n around its mean $n\mu$. It is in fact another statement about convergence, where the mode of convergence involved is the following:

Definition 3.5.1. Let $X, X_1, X_2, \ldots$ be a sequence of random variables (not necessarily defined on the same probability space) with distribution functions $F, F_1, F_2, \ldots$, respectively. $\{X_n\}$ converges to X **in distribution**, denoted $X_n \xrightarrow{\mathrm{d}} X$, if for each x such that F is continuous at x, $F_n(x) \to F(x)$ as $n \to \infty$.

Remark 3.5.2. Convergence in distribution is really about convergence of distributions rather than convergence of random variables.

We have already seen several examples of convergence in distribution in disguise:

Example 3.5.3. Let $\{X_n\}$ be a sequence of binomial random variables where X_n has parameters $(n, \lambda/n)$. Then $\{X_n\}$ converges in distribution to the Poisson random variable with parameter λ (see Example 1.4.16).

Let $\{X_n\}$ be a sequence of geometric random variables where X_n has parameter p_n. If $p_n \to 0$ as $n \to \infty$, then $\{p_n X_n\}$ converges in distribution to the exponential random variable with parameter 1 (see the discussion after Example 3.0.7).

Convergence in distribution is the weakest form of convergence we have introduced, as the following results show.

Lemma 3.5.4. *If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.*

Example 3.5.5. Let X and Y be random variables having the same distribution function and such that $\mathbb{P}(X = Y) < 1$. Letting $X_n = X$ for each $n \geq 1$, the sequence $\{X_n\}$ clearly converges to Y in distribution but not in probability.

Although convergence in distribution is weaker than convergence in probability, there is a partial converse in the case the limit is deterministic:

Exercise 3.5.6. *Show that if $\{X_n\}$ converges in distribution to the constant random variable c, then $\{X_n\}$ converges in probability to c.*

Recall that we are interested in studying the behavior of the deviations of S_n from its mean $n\mu$. We rescale $S_n - \mathbb{E}(S_n)$ as follows:

$$Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{var}(S_n)}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Theorem 3.5.7 (Central limit theorem). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2 and let*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then $\{Z_n\}$ converges in distribution to the standard normal. In other words, $\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z)$ for any z, where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, \mathrm{d}x.$$

Proof. In order to show convergence in distribution, we will make use of the following sufficient condition in terms of the moment generating functions:

Theorem 3.5.8. *Suppose Y and $X_1, X_2, \ldots$ are random variables such that $M_Y(s)$ and $M_{X_1}(s), M_{X_2}(s), \ldots$ exist for all $s \in [-a, a]$, for some $a > 0$. If $M_{X_n}(s) \to M_Y(s)$ as $n \to \infty$ for all $s \in [-a, a]$, then $X_n \xrightarrow{d} Y$.*

Back to the proof, let $Y_i = X_i - \mu$. Notice that $\mathbb{E}(Y_i) = 0$ and $\mathrm{var}(Y_i) = \mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2 = \mathbb{E}(Y_i^2)$. Let Y be a random variable with the same distribution as the Y_i's. We first rewrite Z_n as a sum:

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 - \mu}{\sigma\sqrt{n}} + \cdots + \frac{X_n - \mu}{\sigma\sqrt{n}} = \frac{Y_1}{\sigma\sqrt{n}} + \cdots + \frac{Y_n}{\sigma\sqrt{n}}$$

But

$$M_{\frac{Y_i}{\sigma\sqrt{n}}}(t) = \mathbb{E}\Big(e^{t\frac{Y_i}{\sigma\sqrt{n}}}\Big) = M_Y\Big(\frac{t}{\sigma\sqrt{n}}\Big)$$

and so, since $\frac{Y_1}{\sigma\sqrt{n}}, \ldots, \frac{Y_n}{\sigma\sqrt{n}}$ are independent,

$$M_{Z_n}(t) = M_{\frac{Y_1}{\sigma\sqrt{n}}}(t) \cdots M_{\frac{Y_n}{\sigma\sqrt{n}}}(t) = \Big(M_Y\Big(\frac{t}{\sigma\sqrt{n}}\Big)\Big)^n.$$

Since the function M_Y is differentiable twice at 0 (as the Y_i's admit mean and variance), Taylor's theorem tells us that, for $h \to 0$,

$$\begin{aligned} M_Y(h) &= M_Y(0) + hM_Y'(0) + \frac{h^2}{2}M_Y''(0) + o(h^2) \\ &= 1 + h\mathbb{E}(Y) + \frac{h^2}{2}\mathrm{var}(Y) + o(h^2) \\ &= 1 + \frac{h^2\sigma^2}{2} + o(h^2), \end{aligned}$$

where $o(h^2)$ denotes a function which goes to 0 faster than h^2 as $h \to 0$. Using the Taylor approximation above for $n \to \infty$, we obtain

$$M_Y\Big(\frac{t}{\sigma\sqrt{n}}\Big) = 1 + \frac{t^2}{2n} + o\Big(\frac{t^2}{\sigma^2 n}\Big)$$

and so

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\Big(1 + \frac{t^2}{2n}\Big)^n = e^{t^2/2},$$

where the last equality follows from the fact that $(1 + \frac{a_n}{n})^n \to e^a$ for any sequence $a_n \to a$.

Since $M_Z(t) = e^{t^2/2}$ is the mgf of the standard normal (Example 3.4.5), Theorem 3.5.8 then implies that $\{Z_n\}$ converges in distribution to the standard normal. □

The Central limit theorem has several important consequences:

> It first tells us that the "fluctuations" of S_n around its mean $n\mu$ are of order $\sqrt{n}$. Moreover, the behavior of these fluctuations is universal: no matter what the distribution of the X_i's is, the asymptotic distribution of the "fluctuations" is standard normal.

It also answers the question: How does S_n behave for large n?

For n large, probabilities of the form $\mathbb{P}(S_n \leq c)$ can be approximated as follows:

1. Compute mean $n\mu$ and variance $n\sigma^2$ of S_n.
2. Compute $z = \frac{c-n\mu}{\sigma\sqrt{n}}$.
3. Use the approximation $\mathbb{P}(S_n \leq c) \approx \Phi(z)$, where the value of $\Phi(z)$ can be found from tables.

Let us justify these steps. For large n, the Central limit theorem implies that

$$\Phi(z) \approx \mathbb{P}(Z_n \leq z) = \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \mathbb{P}(S_n \leq z\sigma\sqrt{n} + n\mu).$$

Therefore, letting z as in 2., we obtain the approximation in 3.

Example 3.5.9. The number of students X who are going to fail in the exam is a Poisson random variable with mean 100. What is the probability that at least 120 students will fail?

Since X is Poisson with mean 100 (and so $\lambda = 100$), we know that the desired probability is

$$\mathbb{P}(X \geq 120) = 1 - \mathbb{P}(X \leq 119) = 1 - \sum_{k=0}^{119} \frac{e^{-100}100^k}{k!}.$$

If we are just interested in an approximate value of this complicated sum, we can use the procedure described above. We can express X as a sum of 100 independent Poisson random variables $X_1, \ldots, X_{100}$, each with mean 1 and variance 1 (for example, by Exercise 3.4.10). Checking Figure 3.2, we then have that

$$\mathbb{P}(X \leq 119) \approx \Phi\left(\frac{119 - 100 \cdot 1}{1 \cdot \sqrt{100}}\right) = \Phi(1.9) = 0.9713.$$

Example 3.5.10. We load on a plane 100 packages whose weights are independent random variables uniformly distributed between 5kg and 50kg. What is the probability that the total weight will exceed 3000kg?

Let S_{100} be the sum of weights of the 100 packages. We compute an approximate value for the desired probability $\mathbb{P}(S_{100} > 3000)$ by following the procedure above. We first need mean and variance of a uniform random variable on $[5, 50]$. In general, mean and variance of a uniform random variable X on $[a, b]$ are $\mathbb{E}(X) = \frac{a+b}{2}$ and $\mathrm{var}(X) = \frac{(a-b)^2}{12}$. You can compute these in two ways: using the definitions of mean and variance or using moment generating functions. In our case, $\mu = 27.5$ and $\sigma^2 = 168.75$. Letting $z = \frac{3000-100\cdot 27.5}{\sqrt{168.75\cdot 100}} = 1.92$, we get $\mathbb{P}(S_{100} \leq 3000) \approx \Phi(1.92) = 0.9726$.

Exercise 3.5.11. *Compute mean and variance of a uniform random variable X on $[a, b]$ using moment generating functions.*

Exercise 3.5.12. *A machine processes parts one at a time. The processing times of different parts are independent random variables uniformly distributed on $[1, 5]$. Find an approximate value for the probability that the number of parts processed within 320 time units is at least 100.*

Chapter 4

Poisson Processes

A Poisson process is a stochastic process used to model the occurrence, or arrival, of events over a continuous interval of time. There are several ways to define a Poisson process. One can focus, for example,

1. on the number of events that occur in fixed intervals;
2. on the times at which events occur and the times between occurrences.

We begin with 1.

Definition 4.0.1. A sequence $\{N(s) : s \geq 0\}$ of random variables indexed by the continuous parameter s is a **Poisson process** if it satisfies the following:

(i) $N(0) = 0$;

(ii) $N(t+s) - N(s)$ is Poisson with parameter λt, where λ is called the **rate** of the Poisson process;

(iii) $N(t)$ has **independent increments** i.e., if $t_0 < t_1 < \cdots < t_n$, then $N(t_1) - N(t_0), \ldots, N(t_n) - N(t_{n-1})$ are independent random variables.

Remark 4.0.2. Notice that a Poisson process consists of uncountably many random variables and is an example of a **continuous-time stochastic process**, as opposed to the discrete-time Markov chains we have seen in Chapter 2.

The typical use of a Poisson process is as a model for the number of arrivals $N(s)$ in time $[0, s]$ to a certain facility, say an ATM. We can then think of (ii) as saying that the average rate at which customers arrive is constant and of (iii) as saying that the number of customers arriving during a certain time interval does not affect the number of customers arriving during a different time interval. Why are these assumptions reasonable? Consider the following situation. Suppose that each of n students flips a coin, with probability λ/n of heads, to decide if he will go to a fixed ATM between 12:17 and 12:18. The probability that the number of students X going to the ATM during this interval is exactly k is

$$\mathbb{P}(X = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

But we have seen that if $n \to \infty$, then

$$\binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}.$$

Example 4.0.3. Starting at 6 a.m., customers arrive at a bakery according to a Poisson process at the rate of 30 customers per hour. What is the probability that more than 65 customers arrive between 9 and 11 a.m.? We let $s = 0$ represent 6 a.m. The desired probability is then $\mathbb{P}(N(5) - N(3) > 65)$. Since $N(5) - N(3)$ is a Poisson with parameter $30 \cdot 2 = 60$, we have

$$\mathbb{P}(N(5) - N(3) > 65) = 1 - \mathbb{P}(N(3) - N(2) \leq 65) = 1 - \sum_{k=0}^{65} e^{-60} \frac{60^k}{k!}.$$

We can now construct Poisson processes in another way, following 2, with emphasis on the times of arrivals. The fact that these two definitions are equivalent is a nontrivial result. In the new definition, the random variables $N(s)$ are built as follows:

Definition 4.0.4. Let $\tau_1, \tau_2, \ldots$ be independent exponential random variables with parameter λ. Let $T_n = \tau_1 + \cdots + \tau_n$ for $n \geq 1$, $T_0 = 0$ and let $N(s) = \max\{n : T_n \leq s\}$.

The τ_n's can be thought of as the times between arrivals of customers (**interarrival times**), say to our ATM, and so T_n is the arrival time of the n-th customer and $N(s)$ is the number of arrivals up to time s.

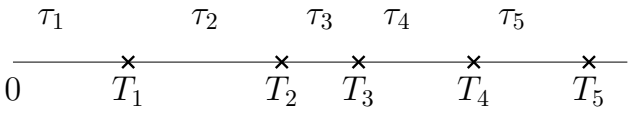


Figure 4.1: With the interpretation above, if $T_4 \leq s < T_5$, then $N(s) = 4$.

We now show that $N(s)$, as defined in Definition 4.0.4, satisfies (i), (ii), (iii) in Definition 4.0.1 and so gives indeed rise to a Poisson process. (i) is trivial. To show (ii), we will in fact show the following:

Proposition 4.0.5. *$N(s)$ is a Poisson random variable with parameter λs.*

The proof of Proposition 4.0.5 requires several auxiliary results. The key is establishing the nature of the random variables T_n. They will turn out to be Gamma random variables:

Definition 4.0.6. A continuous random variable X is **Gamma** with parameters (α, β), for some $\alpha, \beta > 0$, if its pdf is given by

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

where $\Gamma(\alpha)$ is the **Gamma function** defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, \mathrm{d}t.$$

Remark 4.0.7. The Gamma function extends the factorial function to positive real numbers. Namely, if α is a positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$. Notice also that if $\alpha = 1$, then the Gamma random variable with parameters $(\alpha, \beta) = (1, \beta)$ is an exponential with parameter $1/\beta$ and so the Gamma random variable generalizes the exponential random variable.

Comparing the mgf of T_n with that of the Gamma random variable, we obtain the following:

Lemma 4.0.8. *T_n is a Gamma random variable with parameters $(n, 1/\lambda)$.*

Proof. Let us first compute the moment generating function of the Gamma random variable X with parameters (α, β). Suppose that $t < 1/\beta$.

$$M_X(t) = \int_0^\infty e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \, \mathrm{d}x = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x(\frac{1-\beta t}{\beta})} \, \mathrm{d}x.$$

We now make the substitution $x(\frac{1-\beta t}{\beta}) = u$ and $\frac{1-\beta t}{\beta} \, \mathrm{d}x = \mathrm{d}u$. We obtain

$$M_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty u^{\alpha-1} \left(\frac{\beta}{1-\beta t}\right)^{\alpha-1} e^{-u} \frac{\beta}{1-\beta t} \, \mathrm{d}u = \left(\frac{1}{1-\beta t}\right)^\alpha.$$

Let us now compute the moment generating function of the exponential random variable Y with parameter λ.

$$M_Y(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} \, \mathrm{d}x = \lambda \int_0^\infty e^{-x(\lambda - t)} \, \mathrm{d}x = \frac{\lambda}{\lambda - t} = \frac{1}{1 - t/\lambda}.$$

Since T_n is a sum of n independent exponential random variables with parameter λ, we have that

$$M_{T_n}(t) = \left(\frac{1}{1-t/\lambda}\right)^n.$$

Using the Inversion theorem, we then conclude that T_n is a Gamma random variable with parameters $(n, 1/\lambda)$. □

We are finally ready to show Proposition 4.0.5. Recall that we are using the definition of $N(s)$ as in Definition 4.0.4.

Proof of Proposition 4.0.5. Observe first that $N(s) = n$ if and only if $T_n \leq s < T_{n+1}$. Moreover,

$$\mathbb{P}(N(s) = n) = \mathbb{P}(N(s) \geq n) - \mathbb{P}(N(s) \geq n + 1).$$

But the event $\{N(s) \geq n\}$ ("up to time s we have at least n arrivals") coincides with the event $\{T_n \leq s\}$ ("the n-th arrival happens before time s") and so $\mathbb{P}(N(s) \geq n) = \mathbb{P}(T_n \leq s)$, where $T_n = \tau_1 + \cdots + \tau_n$ is a sum of n independent exponential random variables with parameter λ.

We know use the following equality which holds for all nonnegative real numbers x and all positive integers n, and which can be proved by showing that the derivatives of the two sides are equal:

$$\int_0^x \frac{\lambda^n}{(n-1)!} u^{n-1} e^{-\lambda u} \, \mathrm{d}u = \sum_{i=n}^\infty e^{-\lambda x} \frac{(\lambda x)^i}{i!}.$$

Using the equality above and Lemma 4.0.8, we obtain

$$\mathbb{P}(T_n \leq s) = \int_0^s \frac{\lambda^n}{(n-1)!} u^{n-1} e^{-\lambda u} \, \mathrm{d}u = \sum_{i=n}^\infty e^{-\lambda s} \frac{(\lambda s)^i}{i!}, \tag{4.1}$$

for each $n \geq 1$. Moreover, if $n = 0$, then $\mathbb{P}(T_0 \leq s) = \mathbb{P}(N(s) \geq 0) = 1$ and

$$\sum_{i=0}^\infty e^{-\lambda s} \frac{(\lambda s)^i}{i!} = 1$$

(pmf of Poisson) and so (4.1) holds for each $n \geq 0$. But then

$$\mathbb{P}(N(s) = n) = \mathbb{P}(T_n \leq s) - \mathbb{P}(T_{n+1} \leq s) = \sum_{i=n}^\infty e^{-\lambda s} \frac{(\lambda s)^i}{i!} - \sum_{i=n+1}^\infty e^{-\lambda s} \frac{(\lambda s)^i}{i!} = e^{-\lambda s} \frac{(\lambda s)^n}{n!}.$$

This implies that $N(s)$ is indeed a Poisson with parameter λs. □

Using Proposition 4.0.5, it can then be shown that that (ii) and (iii) in Definition 4.0.1 hold as well: If $0 = t_0 < t_1 < t_2 < \cdots < t_d$, then $N(t_i) - N(t_{i-1})$ is Poisson with parameter $\lambda(t_i - t_{i-1})$ and $N(t_1) - N(t_0), N(t_2) - N(t_1), \ldots, N(t_d) - N(t_{d-1})$ are independent. We will not show this, but just notice that $N(t_1) - N(t_0) = N(t_1)$ is indeed Poisson with parameter $\lambda(t_1 - t_0) = \lambda t_1$.

It can also be shown that Definition 4.0.1 implies Definition 4.0.4 and so the two are equivalent. This means that any property obtained from one definition holds with the other. Let us just observe (using Definition 4.0.1) the special case that the first interarrival time τ_1 is indeed exponential with parameter λ. We have that $\tau_1 > t$ if and only if $N(t) = 0$ and so $\mathbb{P}(\tau_1 > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t}$.

Example 4.0.9. Let $\{N(s) : s \geq 0\}$ be a Poisson process where the interarrival times τ_i are exponential with parameter 5. Compute $\mathbb{P}(N(3) = 12)$ and $\mathbb{P}(N(2) = 3, N(5) = 4)$.

We know that $N(3) = N(3) - N(0)$ is Poisson with parameter $3 \cdot 5 = 15$. Therefore,

$$\mathbb{P}(N(3) = 12) = e^{-15}\frac{15^{12}}{12!}.$$

As for the second question, $N(2)$ and $N(5)$ might not be independent, but we use the fact that the increments $N(2) - N(0) = N(2)$ and $N(5) - N(2)$ are. Clearly, the event $\{N(2) = 3, N(5) = 4\}$ is the same as the event $\{N(2) = 3, N(5) - N(2) = 1\}$ and so

$$\begin{aligned}\mathbb{P}(N(2) = 3, N(5) = 4) &= \mathbb{P}(N(2) = 3, N(5) - N(2) = 1)\\ &= \mathbb{P}(N(2) = 3)\mathbb{P}(N(5) - N(2) = 1)\\ &= e^{-10}\frac{10^3}{3!} \cdot e^{-15}\frac{15^1}{1!}.\end{aligned}$$

Example 4.0.10. Consider a Poisson process with rate λ. Compute

(a) $\mathbb{E}$(time of 10th arrival).

(b) $\mathbb{P}$(10th arrival occurs two or more time units after 9th).

(c) $\mathbb{P}$(10th arrival occurs later than time 20).

(d) $\mathbb{P}$(there are two arrivals in [1, 4] and three arrivals in [3, 5]).

(a) We want $\mathbb{E}(T_{10})$. We know that $T_{10} = \tau_1 + \cdots + \tau_{10}$, where each τ_i is exponential with parameter λ. We know that $\mathbb{E}(\tau_i) = 1/\lambda$ and so linearity of expectation implies that $\mathbb{E}(T_{10}) = 10/\lambda$. Let us obtain again the mean of the exponential, this time using its moment generating function $M_Y(t) = \frac{1}{1-t/\lambda}$ computed in the proof of Lemma 4.0.8. Since

$$M_Y'(t) = \frac{1}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-2},$$

we obtain again $\mathbb{E}(Y) = M_Y'(0) = 1/\lambda$. Another way of computing $\mathbb{E}(T_{10})$ is by recalling that T_{10} is a Gamma random variable with parameters $(10, \frac{1}{\lambda})$ and using the moment generating function $M_{T_{10}}(t)$ we computed in the proof of Lemma 4.0.8.

(b) We want $\mathbb{P}(T_{10} \geq T_9 + 2)$. We have

$$\mathbb{P}(T_{10} \geq T_9 + 2) = \mathbb{P}(T_{10} - T_9 \geq 2) = \mathbb{P}(\tau_{10} \geq 2) = e^{-2\lambda},$$

where in the last equality we recalled (3.1).

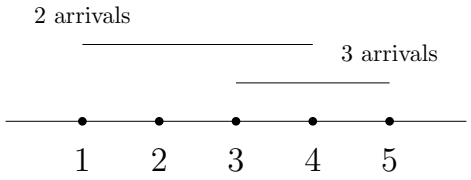
(c) We want $\mathbb{P}(T_{10} > 20)$. By Lemma 4.0.8, T_{10} is Gamma with parameter $(10, 1/\lambda)$. Recalling the pdf of the Gamma random variable, we obtain

$$\mathbb{P}(T_{10} > 20) = \int_{20}^{\infty} \frac{\lambda^{10}}{9!} u^9 e^{-\lambda u}\, \mathrm{d}u.$$

Alternatively, observe that the event $\{T_{10} > 20\}$ is the same as the event $\{N(20) < 10\}$. Since $N(20)$ is Poisson with parameter 20λ,

$$\mathbb{P}(T_{10} > 20) = \mathbb{P}(N(20) < 10) = \sum_{j=0}^{9} e^{-20\lambda} \frac{(20\lambda)^j}{j!}.$$

(d) We want $\mathbb{P}(N(4) - N(1) = 2, N(5) - N(3) = 3)$. Since the intervals $[1, 4]$ and $[3, 5]$ overlap, we cannot use the independent increments property directly, so we have first to reduce to non-overlapping intervals. We do this by conditioning on the number of arrivals in the intersection $[3, 4]$, which can be 0, 1 or 2.



By the law of total probability,

$$\begin{aligned}
\mathbb{P}(2 \text{ in } [1,4], 3 \text{ in } [3,5]) &= \sum_{k=0}^{2} \mathbb{P}(2 \text{ in } [1,4], 3 \text{ in } [3,5] \mid k \text{ in } [3,4]) \mathbb{P}(k \text{ in } [3,4]) \\
&= \sum_{k=0}^{2} \mathbb{P}(2-k \text{ in } [1,3], 3-k \text{ in } [4,5]) \mathbb{P}(k \text{ in } [3,4]) \\
&= \sum_{k=0}^{2} \mathbb{P}(N(3) - N(1) = 2-k, N(5) - N(4) = 3-k) \mathbb{P}(N(4) - N(3) = k) \\
&= \sum_{k=0}^{2} \mathbb{P}(N(3) - N(1) = 2-k) \mathbb{P}(N(5) - N(4) = 3-k) \mathbb{P}(N(4) - N(3) = k) \\
&= \sum_{k=0}^{2} e^{-2\lambda} \frac{(2\lambda)^{2-k}}{(2-k)!} \cdot e^{-\lambda} \frac{\lambda^{3-k}}{(3-k)!} \cdot e^{-\lambda} \frac{\lambda^k}{k!}.
\end{aligned}$$

Example 4.0.11. Consider a Poisson process with rate λ. Suppose we know that a single event occured in $[0, s]$. What is the probability that it occured before time t?

We need to compute $\mathbb{P}(T_1 \leq t | T_1 \leq s)$, where $0 < t < s$. We have

$$\begin{aligned}
\mathbb{P}(T_1 \leq t | T_1 \leq s) &= \mathbb{P}(N(t) = 1 | N(s) = 1) \\
&= \frac{\mathbb{P}(N(t) = 1, N(s) = 1)}{\mathbb{P}(N(s) = 1)} \\
&= \frac{\mathbb{P}(N(t) - N(0) = 1, N(s) - N(t) = 0)}{\mathbb{P}(N(s) = 1)} \\
&= \frac{\mathbb{P}(N(t) = 1) \mathbb{P}(N(s) - N(t) = 0)}{\mathbb{P}(N(s) = 1)} \\
&= \frac{e^{-\lambda t} \lambda t \cdot e^{-\lambda(s-t)}}{e^{-\lambda s} \lambda s} \\
&= \frac{t}{s},
\end{aligned}$$

where the third equality follows from the independence of the increments $N(t) - N(0) = N(t)$ and $N(t) - N(s)$. We observe that the conditional distribution is uniform (see Example 1.4.11). For example, suppose we heard that a football team won a match 1-0. Is it more likely that the goal was scored in the first half or second half? Letting T_1 to be the time the goal was scored (and under the assumption T_1 is as in Definition 4.0.4), the goal is equally likely to have been scored in the first half or second half.

Exercise 4.0.12. *Let $\{N(t) : t \geq 0\}$ be a Poisson process with rate $\lambda = 3$.*

1. *What is the probability that $\tau_3 > 6$ given that $N(5) < 2$?*
2. *What is the expected time of the 5-th arrival?*
3. *Compute $\mathbb{E}(N(4) - N(2)|N(1) = 3)$*

Exercise 4.0.13. *Let $\{N(t) : t \geq 0\}$ be a Poisson process with rate λ. Show that $\mathbb{E}(N(t)N(t+s)) = \lambda t + (\lambda t)^2 + \lambda s \lambda t$.*

4.1 Compound Poisson processes

We now associate i.i.d. random variables Y_i with each arrival. Independent means that the Y_i's and the T_i's are all independent. For example,

(a) At a drive-thru cars arrive between noon and 1:00pm according to a Poisson process. We let Y_i to be the number of people in the i-th car.

(b) Messages arrive at a computer to be transmitted across the Internet. Arrival times can be modeled by a Poisson process. We let Y_i to be the size in bytes of the i-th message.

We can consider the sum of the Y_i's up to time t:

$$S(t) = Y_1 + Y_2 + \cdots + Y_{N(t)},$$

where we set $S(t) = 0$ if $N(t) = 0$. In (a), $S(t)$ is the number of customers up to time t, whereas in (b), $S(t)$ is the total amount of bytes received up to time t. What are mean and variance of $S(t)$? The following more general results allow to compute mean and variance of the sum of a random number of i.i.d. random variables.

Theorem 4.1.1 (Wald's equation). *Let $Y_1, Y_2, \ldots$ be i.i.d. discrete random variables, let N be an independent (of the Y_i's) nonnegative integer-valued random variable and let $S = Y_1 + \cdots + Y_N$, where $S = 0$ if $N = 0$.*

(i) If $\mathbb{E}(|Y_i|), \mathbb{E}(N) < \infty$, then $\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(Y_i)$.

(ii) If $\mathbb{E}(Y_i^2), \mathbb{E}(N^2) < \infty$, then $\mathrm{var}(S) = \mathbb{E}(N)\mathrm{var}(Y_i) + \mathrm{var}(N)\mathbb{E}(Y_i)^2$.

(iii) If N is Poisson with parameter λ, then $\mathbb{E}(S) = \lambda\mathbb{E}(Y_i)$ and $\mathrm{var}(S) = \lambda\mathbb{E}(Y_i^2)$.

Proof. (i) We use the Total expectation theorem (Theorem 1.8.2):

$$\begin{aligned}\mathbb{E}(S) = \sum_{n=0}^{\infty} \mathbb{E}(S|N=n)\mathbb{P}(N=n) &= \sum_{n=0}^{\infty} \mathbb{E}(Y_1 + \cdots + Y_n)\mathbb{P}(N=n) \\ &= \sum_{n=0}^{\infty} n\mathbb{E}(Y_i)\mathbb{P}(N=n) = \mathbb{E}(Y_i)\sum_{n=0}^{\infty} n\mathbb{P}(N=n) = \mathbb{E}(Y_i)\mathbb{E}(N).\end{aligned}$$

(ii) Recall that $\mathrm{var}(S) = \mathbb{E}(S^2) - \mathbb{E}(S)^2$. Moreover, again by the Total expectation theorem,

$$\mathbb{E}(S^2) = \sum_{n=0}^{\infty} \mathbb{E}(S^2|N=n)\mathbb{P}(N=n) = \sum_{n=0}^{\infty} \mathbb{E}((Y_1+\cdots+Y_n)^2)\mathbb{P}(N=n).$$

But since $Y_1, Y_2, \ldots$ are independent,

$$\begin{aligned} n\mathrm{var}(Y_i) = \mathrm{var}(Y_1+\cdots+Y_n) &= \mathbb{E}((Y_1+\cdots+Y_n)^2) - \mathbb{E}(Y_1+\cdots+Y_n)^2 \\ &= \mathbb{E}((Y_1+\cdots+Y_n)^2) - n^2\mathbb{E}(Y_i)^2. \end{aligned}$$

Using the latter equality in the formula above, we obtain

$$\begin{aligned} \mathbb{E}(S^2) = \sum_{n=0}^{\infty}(n\mathrm{var}(Y_i) + n^2\mathbb{E}(Y_i)^2)\mathbb{P}(N=n) &= \mathrm{var}(Y_i)\sum_{n=0}^{\infty} n\mathbb{P}(N=n) + \mathbb{E}(Y_i)^2\sum_{n=0}^{\infty} n^2\mathbb{P}(N=n) \\ &= \mathrm{var}(Y_i)\mathbb{E}(N) + \mathbb{E}(Y_i)^2\mathbb{E}(N^2), \end{aligned}$$

where in the last equality we used LOTUS. Therefore, using (i),

$$\mathrm{var}(S) = \mathrm{var}(Y_i)\mathbb{E}(N) + \mathbb{E}(Y_i)^2\mathbb{E}(N^2) - \mathbb{E}(Y_i)^2\mathbb{E}(N)^2 = \mathrm{var}(Y_i)\mathbb{E}(N) + \mathbb{E}(Y_i)^2\mathrm{var}(N).$$

(iii) Since N is Poisson, we know that $\mathbb{E}(N) = \mathrm{var}(N) = \lambda$ and so, by (i), $\mathbb{E}(S) = \lambda\mathbb{E}(Y_i)$, and by (ii),

$$\mathrm{var}(S) = \lambda\mathrm{var}(Y_i) + \lambda\mathbb{E}(Y_i)^2 = \lambda\mathbb{E}(Y_i^2).$$ □

Example 4.1.2. The number of customers at a shop in a day is Poisson with mean 81 and each customer spends an average of $\$8$ with a standard deviation of $\$6$. What is the mean revenue? What is the variance of the revenue?

Let Y_i be the amount spent by the i-th customer and let N be the number of customers. We want to compute $\mathbb{E}(S)$ and $\mathrm{var}(S)$, where $S = Y_1+\cdots+Y_N$. We know that $\mathbb{E}(Y_i) = 8$, $\mathrm{var}(Y_i) = 6^2$ and $\mathbb{E}(N) = 81$. Therefore, $\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(Y_i) = 81 \cdot 8$ and $\mathrm{var}(S) = \mathbb{E}(N)\mathrm{var}(Y_i) + \mathrm{var}(N)\mathbb{E}(Y_i)^2 = 81 \cdot 36 + 81 \cdot 64$.

4.2 Thinning

Thinning is the operation of splitting a Poisson process into several ones using the associated Y_i's. Let $N_j(t)$ be the number of $i \leq N(t)$ such that $Y_i = j$. For example, in (a), $N_j(t)$ is the number of cars with exactly j people arrived at the drive-thru up to time t. Notice that

$$N(t) = \sum_j N_j(t),$$

where the sum runs through all the values j taken by the i.i.d. random variables $Y_1, Y_2, \ldots$. For each value j of Y_i we can then consider the sequence of random variables $\{N_j(t) : t \geq 0\}$.

Consider for simplicity the case in which Y_i takes exactly two values. Each arrival can then be classified, independent of the other arrivals, of type 1 with probability p or of type 2 with probability $1-p$. We then have that $N(t) = N_1(t) + N_2(t)$. The nice feature of thinning is that we end up with independent Poisson processes:

Theorem 4.2.1. *$\{N_j(t) : t \geq 0\}$ are independent Poisson processes with rate $\lambda\mathbb{P}(Y_i = j)$. Independent means that, for any $t_0 < \cdots < t_k$, the families $\{N_1(t_i) : 0 \leq i \leq k\}, \ldots, \{N_n(t_\ell) : 0 \leq \ell \leq k\}$ are independent.*

Informal proof. For simplicity, we consider again the case where each Y_i takes two values and hence we have arrivals of type 1, with probability p, and of type 2, with probability $1-p$. Observe that, given $N(t) = n+m$, $N_1(t)$ is Binomial with parameters $(n+m, p)$. Therefore,

$$\begin{aligned}
\mathbb{P}(N_1(t)=n, N_2(t)=m) &= \mathbb{P}(N_1(t)=n, N_2(t)=m | N(t)=n+m)\mathbb{P}(N(t)=n+m) \\
&= \binom{n+m}{n} p^n (1-p)^m \cdot e^{-\lambda t}\frac{(\lambda t)^{n+m}}{(n+m)!} \\
&= \frac{(n+m)!}{n!m!} p^n (1-p)^m \cdot e^{-\lambda t}\frac{(\lambda t)^{n+m}}{(n+m)!} \\
&= e^{-\lambda tp} e^{\lambda tp} \cdot e^{-\lambda t}\frac{(\lambda tp)^n (\lambda t(1-p))^m}{n!m!} \\
&= e^{-\lambda tp}\frac{(\lambda tp)^n}{n!} \cdot e^{-\lambda t(1-p)}\frac{(\lambda t(1-p))^m}{m!}.
\end{aligned}$$

The above hints at the fact that $N_1(t)$ is a Poisson with parameter λtp, that $N_2(t)$ is a Poisson with parameter $\lambda t(1-p)$ and that $N_1(t)$ and $N_2(t)$ are independent. □

Example 4.2.2. A fisherman catches fish at times of a Poisson process with rate 2 per hour. 40% of fish is salmon and 60% is trout. What is the probability that he will catch exactly one salmon and exactly two trouts if he fishes for 2.5 hours?

As we want to distinguish between salmons and trouts, we apply thinning. Let $N_s(t)$ be the number of salmons fished up to time t and $N_t(t)$ be the number of trouts fished up to time t. Theorem 4.2.1 tells us that $\{N_s(t) : t \geq 0\}$ and $\{N_t(t) : t \geq 0\}$ are independent Poisson processes with rates $2 \cdot \frac{40}{100}$ and $2 \cdot \frac{60}{100}$, respectively. This implies that $N_s(t)$ is a Poisson random variable with parameter $0.8t$ and $N_t(t)$ is a Poisson random variable with parameter $1.2t$ As the two are independent, the probability we are interested in is

$$\mathbb{P}(N_s(2.5)=1, N_t(2.5)=2) = \mathbb{P}(N_s(2.5)=1)\mathbb{P}(N_t(2.5)=2) = e^{-2}\frac{2^1}{1!} \cdot e^{-3}\frac{3^2}{2!} = 9e^{-5}.$$

Example 4.2.3. According to the United Nations Population Division, the worldwide sex ratio at birth is 108 boys to 100 i.e., the probability that any birth is a boy is $p = 108/(108+100) = 0.519$. Suppose that births occur on a maternity ward according to a Poisson process with rate 2 births per hour.

(i) On an 8-hour shift, what is the expectation of the number of female births?

(ii) Find the probability that only girls were born between 2 and 5 p.m.

(iii) Assume that five babies were born on the ward yesterday. Find the probability that two are boys.

(i) Let $N_m(t)$ and $N_f(t)$ be the number of male and female births, respectively, up to time t. Theorem 4.2.1 tells us that $\{N_m(t) : t \geq 0\}$ and $\{N_f(t) : t \geq 0\}$ are independent Poisson processes with rates $2 \cdot 0.519$ and $2 \cdot 0.481$, respectively. Therefore, $N_f(8)$ is Poisson with parameter $8 \cdot 2 \cdot 0.481$, which is also its expectation.
(ii) The desired probability is $\mathbb{P}(N_m(3)=0, N_f(3)>0)$. Since the two processes are independent,

$$\mathbb{P}(N_m(3)=0, N_f(3)>0) = \mathbb{P}(N_m(3)=0)\mathbb{P}(N_f(3)>0) = \mathbb{P}(N_m(3)=0)(1-\mathbb{P}(N_f(3)=0)).$$

(iii) Conditional on there being five births in a given interval, the number of boys in that interval is Binomial with parameters $n=5$ and $p=0.519$.

Exercise 4.2.4. *Accidents occur at a busy intersection according to a Poisson process at the rate of two accidents per week. Three out of four accidents involve the use of alcohol.*

(i) What is the probability that three accidents involving alcohol will occur next week?

(ii) What is the probability that at least one accident occurs tomorrow?

(iii) If six accidents occur in February (four weeks), what is the probability that less than half of them involve alcohol?

4.3 Superposition

Suppose women and men arrive to a shop according to Poisson processes with rates λ_1 and λ_2, respectively. We can combine the two processes into one by considering all arrivals, regardless of gender. If the two initial processes are independent, it turns out that the combined process is also a Poisson process, and its rate is $\lambda_1 + \lambda_2$. We obtained a new process by superposition, which can be viewed as the inverse of thinning.

Proposition 4.3.1. *Suppose $\{N_1(t) : t \geq 0\}, \ldots, \{N_k(t) : t \geq 0\}$ are independent Poisson processes with rates $\lambda_1, \ldots, \lambda_k$, respectively. Then, letting $N(t) = N_1(t) + \cdots + N_k(t)$, we have that $\{N(t) : t \geq 0\}$ is a Poisson process with rate $\lambda_1 + \cdots + \lambda_k$.*

Proof. We verify that the three properties in Definition 4.0.1 hold:
(i) This follows from the fact that $N_i(0) = 0$ for each i.
(ii) We have that $N(t+s) - N(s) = \sum_{i=1}^{k} N_i(t+s) - N_i(s)$ is a sum of k independent Poisson random variables with parameters $\lambda_1 t, \ldots, \lambda_k t$, which is a Poisson random variable with parameter $(\lambda_1 + \cdots + \lambda_k)t$ (for example, by Exercise 3.4.10).
(iii) The superposition process has independent increments because all the initial processes do and they are in addition independent. □

The following result addresses the order of events in independent Poisson processes:

Lemma 4.3.2. *Let $\{N_1(t) : t \geq 0\}$ and $\{N_2(t) : t \geq 0\}$ be two independent Poisson processes with rates λ_1, λ_2, respectively. The probability that n arrivals occur in the first process before m arrivals occur in the second is*

$$\sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n+m-1-k}.$$

Proof. Let us consider a Poisson process with rate $\lambda_1 + \lambda_2$. We independently decide for each arrival that it belongs to the first process with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$, or to the second process with probability $\frac{\lambda_2}{\lambda_1+\lambda_2}$. By thinning (Theorem 4.2.1), the obtained processes are independent and have rates λ_1 and λ_2. The probability we are interested in is the probability that among the first $m+n-1$ arrivals in the combined process, n or more belong to the first process, and this is exactly the probability in the statement. □

Example 4.3.3. A radioactive material emits α-, β-, γ- and δ-particles according to four independent Poisson processes with rates $\lambda_\alpha, \lambda_\beta, \lambda_\gamma$ and λ_δ, respectively. A particle counter counts all emitted particles. Let $N(t)$ be the number of emissions (registrations) during $(0, t]$, for $t \geq 0$.

(a) What is the expected duration until a particle is registered?

(B) What is the expected duration until a β-particle is registered?

(a) By Proposition 4.3.1, $\{N(t) : t \geq 0\}$ is a Poisson process with rate $\lambda = \lambda_\alpha + \lambda_\beta + \lambda_\gamma + \lambda_\delta$. The duration τ until a particle is registered is then an exponential random variable with parameter λ and so $\mathbb{E}(\tau) = 1/\lambda$.
(b) Since β-particles are emitted according to a Poisson process with rate β independently of the other Poisson processes, it follows that the desired expectation is $1/\lambda_\beta$.

We have seen in Example 4.0.11 that, for $0 < t < s$, $\mathbb{P}(N(t) = 1|N(s) = 1) = t/s$. It turns out that the following generalization holds: The conditional pmf of $N(t)$ given that $N(s) = n$ is Binomial with parameters n and $p = t/s$, as the following result shows.

Lemma 4.3.4. *If $t < s$ and $0 \leq m \leq n$, then*

$$\mathbb{P}(N(t) = m|N(s) = n) = \binom{n}{m}\left(\frac{t}{s}\right)^m\left(1 - \frac{t}{s}\right)^{n-m}.$$

Proof. By independent increments,

$$\begin{aligned}\mathbb{P}(N(t) = m|N(s) = n) &= \frac{\mathbb{P}(N(t) = m)\mathbb{P}(N(s) - N(t) = n - m)}{\mathbb{P}(N(s) = n)} \\ &= e^{-\lambda t}\frac{(\lambda t)^m}{m!} \cdot e^{-\lambda(s-t)}\frac{(\lambda(s-t))^{n-m}}{(n-m)!} \cdot \frac{1}{e^{-\lambda s}}\frac{n!}{(\lambda s)^n} \\ &= \binom{n}{m}\left(\frac{t}{s}\right)^m\left(1 - \frac{t}{s}\right)^{n-m}.\end{aligned}$$

□

Example 4.3.5. Trucks and cars on a highway are independent Poisson processes with rate 40 and 100 per hour, respectively. $1/8$ of trucks and $1/10$ of cars get off on exit A.

(a) What is the probability that exactly six trucks get off on exit A between noon and 1pm?

(b) Given that six trucks got off on A between noon and 1pm, what is the probability that exactly two got off between 12:20 and 12:40?

(c) If we start watching at noon, what is the probability that four cars exit before two trucks do?

(d) Suppose all trucks have one passenger, whereas 30% of cars have one passenger, 50% have two and 20% have four. Find the mean number of people getting off on A in one hour.

(a) As we are interested in the trucks getting off on A, we apply thinning. Let $N(t)$ be the number of trucks getting off on A up to time t. By Theorem 4.2.1, $\{N(t) : t \geq 0\}$ is a Poisson process with rate $40 \cdot \frac{1}{8} = 5$. Therefore,

$$\mathbb{P}(N(1) = 6) = e^{-5}\frac{5^6}{6!}.$$

(b) The desired probability can be computed using Lemma 4.3.4:

$$\mathbb{P}(N(1/3) = 2|N(1) = 6) = \binom{6}{2}\left(\frac{1}{3}\right)^2\left(1 - \frac{1}{3}\right)^{6-2}.$$

(c) By thinning, cars getting off on A is a Poisson processes with rate $100 \cdot \frac{1}{10}$. As this is independent of the Poisson process of trucks getting off on A, we can apply Lemma 4.3.2: the probability that four cars exit before two trucks is

$$\sum_{k=4}^{5}\binom{5}{k}\left(\frac{10}{15}\right)^k\left(\frac{5}{15}\right)^{5-k}.$$

(d) We associate with each arrival of a car at exit A a random variable Y_i taking values $1, 2, 4$ with probabilities $0.3, 0.5, 0.2$, respectively. We then have that $\mathbb{E}(Y_i) = (0.3)1 + (0.5)2 + (0.2)4 = 2.1$. By linearity and Theorem 4.1.1, the desired mean is $5 \cdot 1 + 10 \cdot 2.1 = 26$.

Exercise 4.3.6. *Customers arrive at a store at a rate of* 10 *per hour. Each is either male or female with probability* $1/2$*. Suppose that exactly* 10 *women entered within some hour (say, 10 to 11am). Compute the probability that exactly* 10 *men also entered and the probability that at least* 20 *customers have entered.*

Chapter 5

Martingales

Suppose you are at a "fair" casino, placing bets on various games and watching your total wealth rise and fall randomly. The casino is "fair" if, whenever you play a game there, the expected change in your total wealth is always 0, no matter what the history of the process has been. A martingale is a stochastic process that models the time evolution of your total wealth according to these assumptions. In order to formally define martingales, we first need to review the notion of conditional expectation, as seen in Sections 1.8 and 3.2 in the case of discrete and continuous random variables, respectively.

For any fixed number y, $\mathbb{E}(X|Y=y)$ is also a number. As y varies, so does $\mathbb{E}(X|Y=y)$, and we can therefore view $\mathbb{E}(X|Y=y)$ as a function of y. Since y is the experimental value of the random variable Y, $\mathbb{E}(X|Y=y)$ can then be viewed as a function of a random variable, hence a new random variable. Guided by this, we make the following definition:

Definition 5.0.1. $\mathbb{E}(X|Y)$ is the random variable whose value is $\mathbb{E}(X|Y=y)$ when the value taken by Y is y. In other words, $\mathbb{E}(X|Y)$ is the function $\Omega \to \mathbb{R}$ mapping the outcome ω to $\mathbb{E}(X|Y(\omega))$.

Example 5.0.2. We roll a fair die until we get a 6. Let Y be the total number of rolls and let X be the number of 1's we get. We compute $\mathbb{E}(X|Y=y)$. If $Y=y$, then the first $y-1$ rolls were not a 6 and the y-th roll was a 6. Given this event, X is a binomial random variable with parameters $n=y-1$ and $p=1/5$. So

$$\mathbb{E}(X|Y=y) = np = \frac{1}{5}(y-1).$$

This means that $\mathbb{E}(X|Y)$ is the random variable $\frac{1}{5}(Y-1)$.

Example 5.0.3. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, where each X_i is Bernoulli with parameter p i.e., $\mathbb{P}(X_i=1)=p$ and $\mathbb{P}(X_i=0)=1-p$. Let $Y = X_1 + X_2 + \cdots + X_n$. We know that Y is a Binomial random variable with parameters n and p. But what kind of random variable is $\mathbb{E}(X_1|Y)$? We need to compute $\mathbb{E}(X_1|Y=r)$ for each $r \geq 0$. By definition of conditional expectation,

$$\mathbb{E}(X_1|Y=r) = 0 \cdot \mathbb{P}(X_1=0|Y=r) + 1 \cdot \mathbb{P}(X_1=1|Y=r).$$

Moreover,

$$\begin{aligned}
\mathbb{P}(X_1=1|Y=r) &= \frac{\mathbb{P}(X_1=1, Y=r)}{\mathbb{P}(Y=r)} = \frac{\mathbb{P}(X_1=1, X_2+\cdots+X_n=r-1)}{\mathbb{P}(Y=r)} \\
&= \frac{\mathbb{P}(X_1=1)\mathbb{P}(X_2+\cdots+X_n=r-1)}{\mathbb{P}(Y=r)} \\
&= \frac{p \cdot \binom{n-1}{r-1} p^{r-1}(1-p)^{n-r}}{\binom{n}{r} p^r (1-p)^{n-r}} = \frac{r}{n}.
\end{aligned}$$

This implies that $\mathbb{E}(X_1|Y) = \frac{Y}{n}$.

Exercise 5.0.4. *Let X and Y be continuous random variables with joint pdf $f_{X,Y}(x,y) = \frac{1}{x}$, for $0 < y \leq x \leq 1$. Find $\mathbb{E}(Y|X)$.*

We will be interested in conditioning on more than one random variable Y. To this end, we extend the notions of conditional pmf and pdf and of conditional expectation as follows:

Definition 5.0.5. Let X and $Y_1, \ldots, Y_n$ be random variables (either all discrete or all continuous) associated with the same experiment. The **conditional pmf or pdf of X given** $Y_1 = y_1, \ldots, Y_n = y_n$ is the function

$$f_{X|Y_1,\ldots,Y_n}(x|y_1,\ldots,y_n) = \frac{f_{X,Y_1,\ldots,Y_n}(x,y_1,\ldots,y_n)}{f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n)},$$

where the functions on the RHS are joint pmf, if all the random variables are discrete, and joint pdf, if all the random variables are continuous. If X and $Y_1, \ldots, Y_n$ are all discrete, we let

$$\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n) = \sum_x x f_{X|Y_1,\ldots,Y_n}(x|y_1,\ldots,y_n).$$

If X and $Y_1, \ldots, Y_n$ are all continuous, we let

$$\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n) = \int_{-\infty}^{\infty} x f_{X|Y_1,\ldots,Y_n}(x|y_1,\ldots,y_n)\, \mathrm{d}x.$$

Notice that these definitions all agree with the previous definitions in the case $n = 1$. We can finally extend Definition 5.0.1 as follows.

Definition 5.0.6. $\mathbb{E}(X|Y_1, \ldots, Y_n)$ is the random variable whose value is $\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n)$ when the value taken by Y_i is y_i for each $1 \leq i \leq n$.

Notice that we are implicitly assuming that the expectations $\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n)$ exist i.e., they are finite real numbers. We now list some useful properties of this newly defined random variable $\mathbb{E}(X|Y_1, \ldots, Y_n)$.

Theorem 5.0.7. *Let $X, Z, Y_1, \ldots, Y_n$ be random variables (either all discrete or all continuous), let $a, b \in \mathbb{R}$ and let $g\colon \mathbb{R}^n \to \mathbb{R}$. The following hold:*

(i) $\mathbb{E}(a|Y_1, \ldots, Y_n) = a$.

(ii) $\mathbb{E}(aX + bZ|Y_1, \ldots, Y_n) = a\mathbb{E}(X|Y_1, \ldots, Y_n) + b\mathbb{E}(Z|Y_1, \ldots, Y_n)$.

(iii) $\mathbb{E}(X|Y_1, \ldots, Y_n) \geq 0$ *if* $X \geq 0$.

(iv) $\mathbb{E}(X|Y_1, \ldots, Y_n) = \mathbb{E}(X)$ *if $X, Y_1, \ldots, Y_n$ are independent.*

(v) $\mathbb{E}(\mathbb{E}(X|Y_1, \ldots, Y_n)) = \mathbb{E}(X)$.

(vi) $\mathbb{E}(Xg(Y_1, \ldots, Y_n)|Y_1, \ldots, Y_n) = g(Y_1, \ldots, Y_n)\mathbb{E}(X|Y_1, \ldots, Y_n)$.
In particular, by (i), $\mathbb{E}(g(Y_1, \ldots, Y_n)|Y_1, \ldots, Y_n) = g(Y_1, \ldots, Y_n)$.

Proof. We suppose that all random variables are discrete. The continuous case is analogous.

(i) We look at the values taken by the random variable $\mathbb{E}(a|Y_1, \ldots, Y_n)$ i.e., compute $\mathbb{E}(a|Y_1 = y_1, \ldots, Y_n = y_n)$. By linearity of expectation, $\mathbb{E}(a|Y_1 = y_1, \ldots, Y_n = y_n) = a$ and so $\mathbb{E}(a|Y_1, \ldots, Y_n)$ is just the constant random variable a.

(ii) As in (i), we first compute the values $\mathbb{E}(aX + bZ|Y_1 = y_1, \ldots, Y_n = y_n)$ taken by $\mathbb{E}(aX + bZ|Y_1, \ldots, Y_n)$. By linearity of expectation,

$$\mathbb{E}(aX + bZ|Y_1 = y_1, \ldots, Y_n = y_n) = a\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n) + b\mathbb{E}(Z|Y_1 = y_1, \ldots, Y_n = y_n)$$

and so $\mathbb{E}(aX + bZ|Y_1, \ldots, Y_n) = a\mathbb{E}(X|Y_1, \ldots, Y_n) + b\mathbb{E}(Z|Y_1, \ldots, Y_n)$.

(iii) The values taken by $\mathbb{E}(X|Y_1, \ldots, Y_n)$ are $\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n)$. But we know that

$$\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n) = \sum_x x f_{X|Y_1,\ldots,Y_n}(x|y_1, \ldots, y_n),$$

where x ranges over the values taken by X. Therefore, if X is nonnegative, we have that $\mathbb{E}(X|Y_1, \ldots, Y_n)$ is nonnegative as well.

(iv) If $X, Y_1, \ldots, Y_n$ are independent, then their conditional pmf is given by

$$\begin{aligned} f_{X|Y_1,\ldots,Y_n}(x|y_1, \ldots, y_n) &= \frac{f_{X,Y_1,\ldots,Y_n}(x, y_1, \ldots, y_n)}{f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)} \\ &= \frac{f_X(x) f_{Y_1}(y_1) \cdots f_{Y_n}(y_n)}{f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)} \\ &= \frac{f_X(x) f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)}{f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)} \\ &= f_X(x). \end{aligned}$$

Therefore,

$$\mathbb{E}(X|Y_1 = y_1, \cdots, Y_n = y_n) = \sum_x x f_{X|Y_1,\ldots,Y_n}(x|y_1, \ldots, y_n) = \sum_x x f_X(x) = \mathbb{E}(X).$$

(v) We need to compute the expectation of the random variable $\mathbb{E}(X|Y_1, \ldots, Y_n)$. This random variable is a function of $Y_1, \ldots, Y_n$ taking the value $\mathbb{E}(X|Y_1 = y_1, \cdots, Y_n = y_n)$ when $Y_1 = y_1, \ldots, Y_n = y_n$. We have that,

$$\mathbb{E}(\mathbb{E}(X|Y_1, \ldots, Y_n)) = \sum_{(y_1,\ldots,y_n)} \mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n)\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n) = \mathbb{E}(X),$$

where the first equality follows from LOTUS and the second equality is the total expectation theorem (Theorem 1.8.2).

(vi) We first compute $\mathbb{E}(Xg(Y_1, \ldots, Y_n)|Y_1 = y_1, \ldots, Y_n = y_n)$. Given that $Y_1 = y_1, \ldots, Y_n = y_n$, the possible values of $Xg(Y_1, \ldots, Y_n)$ are $xg(y_1, \ldots, y_n)$, where x varies over the range of X. The probability that $Xg(Y_1, \ldots, Y_n)$ takes value $xg(y_1, \ldots, y_n)$ given that $Y_1 = y_1, \ldots, Y_n = y_n$ is just $\mathbb{P}(X = x|Y_1 = y_1, \ldots, Y_n = y_n)$. Therefore, by LOTUS,

$$\begin{aligned} \mathbb{E}(Xg(Y_1, \ldots, Y_n)|Y_1 = y_1, \ldots, Y_n = y_n) &= \sum_x xg(y_1, \ldots, y_n)\mathbb{P}(X = x|Y_1 = y_1, \ldots, Y_n = y_n) \\ &= g(y_1, \ldots, y_n) \sum_x x\mathbb{P}(X = x|Y_1 = y_1, \ldots, Y_n = y_n) \\ &= g(y_1, \ldots, y_n)\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n). \end{aligned}$$

This shows that $\mathbb{E}(Xg(Y_1, \ldots, Y_n)|Y_1, \ldots, Y_n) = g(Y_1, \ldots, Y_n)\mathbb{E}(X|Y_1, \ldots, Y_n)$. □

We are now ready to define martingales. Given a discrete-time stochastic process $\{W_k\}_{k\geq 0}$, we let $W_{m,n}$ to be the portion $W_m, W_{m+1}, \ldots, W_n$ of the process from time m up to time n.

Definition 5.0.8. A discrete-time stochastic process $\{M_n\}_{n\geq 0}$ is a **martingale** if $\mathbb{E}(|M_n|) < \infty$ and

$$\mathbb{E}(M_{n+1}|M_{0,n}) = M_n,$$

for each $n \geq 0$.

We can think of M_n as the fortune we have at time n. The condition $\mathbb{E}(|M_n|) < \infty$ is the technical condition guaranteeing finiteness of the conditional expectations. The condition $\mathbb{E}(M_{n+1}|M_{0,n}) = M_n$ is the crucial requirement: it represents "game fairness". If we are playing a fair game, we expect neither to win nor to lose money on average. Given the history of our fortunes up to time n, our expected fortune M_{n+1} at time $n+1$ should just be the fortune M_n that we have at time n.

The following generalization of the previous definition is sometimes useful.

Definition 5.0.9. A discrete-time stochastic process $\{M_n\}_{n\geq 0}$ is a **martingale** with respect to another process $\{W_n\}_{n\geq 0}$ if $\mathbb{E}(|M_n|) < \infty$ and

$$\mathbb{E}(M_{n+1}|W_{0,n}) = M_n,$$

for each $n \geq 0$.

Example 5.0.10 (Random walk). Let $X_1, X_2, \ldots$ be i.i.d. random variables and let $S_n = \sum_{k=1}^{n} X_k$ with $S_0 = 0$. $\{S_n\}_{n\geq 0}$ is a stochastic process called random walk. Notice that, in the special case each X_i takes values in $\{1, -1\}$ with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = -1) = 1 - p$, we obtain the random walk on $\mathbb{Z}$ of Example 2.0.7. If the random variables X_t have mean 0, then $\{S_n\}_{n\geq 0}$ is a martingale. Indeed,

$$\mathbb{E}(S_{n+1}|S_{0,n}) = \mathbb{E}(S_n + X_{n+1}|S_{0,n}) = \mathbb{E}(S_n|S_{0,n}) + \mathbb{E}(X_{n+1}|S_{0,n}) = S_n + \mathbb{E}(X_{n+1}|S_{0,n}) = S_n + \mathbb{E}(X_{n+1}) = S_n.$$

In the first equality we used the definition of S_n. In the second, Theorem 5.0.7(ii). In the third, Theorem 5.0.7(vi). In the fourth, we used the fact that the events $\{X_{n+1} = x_{n+1}\}$ and $B = \{S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n\}$ are independent (Example 1.9.5) and the observation that this implies

$$\mathbb{E}(X_{n+1}|S_0 = 0, S_1 = s_1, \ldots, S_n = s_n) = \sum_x x f_{X_{n+1}|B}(x) = \sum_x x f_{X_{n+1}}(x) = \mathbb{E}(X_{n+1}). \tag{5.1}$$

Exercise 5.0.11. *Consider a random walk as above where each X_i is Bernoulli with parameter p. For $m < n$, determine $\mathbb{E}(S_m|S_n)$.*

Example 5.0.12 (Product of independent random variables). Let $X_0, X_1, \ldots$ be i.i.d. random variables and let $Z_n = \prod_{k=0}^{n} X_k$. If the random variables X_t have mean 1, then $\{Z_n\}_{n\geq 0}$ is a martingale. Indeed,

$$\mathbb{E}(Z_{n+1}|Z_{0,n}) = \mathbb{E}(Z_n X_{n+1}|Z_{0,n}) = Z_n \mathbb{E}(X_{n+1}|Z_{0,n}) = Z_n \mathbb{E}(X_{n+1}) = Z_n.$$

In the first equality we used the definition of Z_n. In the second, Theorem 5.0.7(vi). In the third, a reasoning similar to Equation (5.1).

Example 5.0.13 (Pólya's urn). Suppose we start at time 2 with one black ball and one white ball in an urn. At each discrete time, we randomly take out a ball from the urn and we return it to the urn together with a new ball of the same color. Let X_n denote the number of white balls at time n. Given that $X_n = k$, with probability k/n we draw a white ball so that $X_{n+1} = k + 1$, and with probability $1 - k/n$ we draw a black ball so that $X_{n+1} = k$. Letting $M_n = X_n/n$ be the fraction of white balls at time n, we have that

$$\mathbb{E}(M_{n+1}|X_{2,n}) = \mathbb{E}\left(\frac{X_{n+1}}{n+1}|X_{2,n}\right) = \mathbb{E}\left(\frac{X_{n+1}}{n+1}|X_n\right) = \frac{1}{n+1}\left((X_n+1)\frac{X_n}{n} + X_n\left(1 - \frac{X_n}{n}\right)\right) = \frac{X_n}{n} = M_n.$$

In the second equality we used the fact that X_{n+1} depends only on X_n. In the third, Theorem 5.0.7(ii) and the definition of $\mathbb{E}(X_{n+1}|X_n)$ (details are left as an exercise). Therefore, $\{M_n\}_{n\geq 2}$ is a martingale with respect to $\{X_n\}_{n\geq 2}$.

Exercise 5.0.14. *Let X_n be defined as in Example 5.0.13. Show that*

$$\mathbb{E}\Big(\frac{X_{n+1}}{n+1}|X_n\Big) = \frac{1}{n+1}\Big((X_n+1)\frac{X_n}{n} + X_n\Big(1-\frac{X_n}{n}\Big)\Big).$$

Example 5.0.15 (Doob's martingale). Let $X_0, X_1, \dots$ be an arbitrary sequence of random variables and let Y be another random variable. Let $M_n = \mathbb{E}(Y|X_{0,n})$, for each $n \geq 0$. We claim that $\{M_n\}_{n\geq 0}$ is a martingale with respect to the process $\{X_n\}_{n\geq 0}$. Indeed,

$$\mathbb{E}(M_{n+1}|X_{0,n}) = \mathbb{E}(\mathbb{E}(Y|X_{0,n+1}) \mid X_{0,n}) = \mathbb{E}(\mathbb{E}(Y|X_{0,n}, X_{n+1}) \mid X_{0,n}) = \mathbb{E}(Y|X_{0,n}) = M_n.$$

In the first equality we used the definition of M_n. In the second, the definition of $X_{0,n}$. The third equality is an instructive exercise (Exercise 5.0.16).

Doob's martingales have the following interpretation. Imagine that you are to receive some future reward Y. You observe the random variables $X_0, X_1, \dots$ sequentially (at time n, you observe the value of X_n). You do not know the value of the random variable Y, but assume that you know from the beginning the joint distribution of the random variables, so that you can compute expectations and so on. At time n, if you had to guess the value of Y, then your best guess would be to consider the conditional expectation of Y given all the information at your disposal so far, that is, $\mathbb{E}(Y|X_{0,n})$. We showed that the sequence of guesses forms a martingale. This makes sense. You do not expect tomorrow's guess to be systematically higher than today's: if you did expect this, that would mean that you think today's guess is too low and so it would not be your best guess!

Exercise 5.0.16. *Let X, Y and W be discrete random variables. Show that $\mathbb{E}(\mathbb{E}(X|Y,W)|Y) = \mathbb{E}(X|Y)$.*

We now introduce two new processes, called submartingales and supermartingales, which are "better than fair" and "worse than fair", respectively.

Definition 5.0.17. A discrete-time stochastic process $\{X_n\}_{n\geq 0}$ is a **submartingale** with respect to another process $\{W_n\}_{n\geq 0}$ if

$$\mathbb{E}(X_{n+1}|W_{0,n}) \geq X_n,$$

for each $n \geq 0$, and a **supermartingale** with respect to $\{W_n\}_{n\geq 0}$ if

$$\mathbb{E}(X_{n+1}|W_{0,n}) \leq X_n,$$

for each $n \geq 0$.

These names go somehow against our intuition. Looking at the definition, if we would like to make money, we would bet on a submartingale, not on a supermartingale. So why having these names then? Well, there are two ways to look at the inequalities in the definition: in a submartingale we have $X_n \leq \mathbb{E}(X_{n+1}|W_{0,n})$, whereas in a supermartingale we have $X_n \geq \mathbb{E}(X_{n+1}|W_{0,n})$. At each time, a submartingale is below its future expected value, whereas a supermartingale is above its future expected value.

Example 5.0.18. Consider again the random walk $\{S_n\}_{n\geq 0}$ in Example 5.0.10. If the random variables X_t have mean 0, we have seen it is a martingale. If they have positive mean, it is a submartingale, and if they have negative mean, a supermartingale.

Exercise 5.0.19. *Let $X_1, X_2, \dots$ be independent random variables with $\mathbb{E}(X_i) = 0$ and $\mathbb{E}(X_i)^2 < \infty$, for each i. For each $n \geq 0$, let $M_n = M_0 + X_1 + \cdots + X_n$ and $T_n = (M_n)^2$. Show that $\{T_n\}_{n\geq 0}$ is a submartingale with respect to $\{M_n\}_{n\geq 0}$.*

5.1 Optional sampling

In this section, we will show an important property of martingales: the "conservation of fairness" property or "you can't beat the system" property, technically known as optional sampling. Let us begin with the following simple observation which justifies the informal definition of martingale stated at the beginning of the chapter.

Lemma 5.1.1. *Let $\{M_n\}_{n\geq 0}$ be a martingale with respect to $\{W_n\}_{n\geq 0}$. Then $\mathbb{E}(M_n) = \mathbb{E}(M_0)$, for all times $n \geq 0$.*

Proof. By definition of martingale, we have that $\mathbb{E}(M_{n+1}|W_{0,n}) = M_n$, for each $n \geq 0$. Taking expectations of both sides and using Theorem 5.0.7(v), we obtain

$$\mathbb{E}(M_n) = \mathbb{E}(\mathbb{E}(M_{n+1}|W_{0,n})) = \mathbb{E}(M_{n+1}),$$

for each $n \geq 0$. □

> The moral is that I can say "stop" at any *predetermined* time t, say $t = 8$, and my winnings will be "fair": $\mathbb{E}(M_8) = \mathbb{E}(M_0)$.

But what if I say "stop" at a time that is not *predetermined* but *random* i.e., that depends on the observed sample path of the game? Is it still true that $\mathbb{E}(M_T) = \mathbb{E}(M_0)$ if T is a random time? There are two obvious obstructions to this "conservation of fairness":

1. I wait an indefinitely long time to say "stop". Well, if I am able just to keep waiting until I see something I like, that seems clearly unfair.

   Consider, for example, the simple symmetric random walk $S_n = X_1 + \cdots + X_n$ on $\mathbb{Z}$. If I say "stop" at time $T = \inf\{n : S_n = 1\}$, then clearly

   $$\mathbb{E}(S_T) = 1 > 0 = \mathbb{E}(S_0).$$

2. I retract moves i.e., I change my mind about something I did in the past. I could use the information I collected up to time t to go back at some time $s < t$ and claim "I meant to say stop then!". This obviously violates fairness: I am supposed to say "stop" using only the information available up to that time.

   For example, consider again the simple symmetric random walk $\{S_n\}_{n\geq 0}$ on $\mathbb{Z}$. Let $T \in [0, 3]$ be the random time at which the random walk takes its maximum value $\max\{S_n : 0 \leq n \leq 3\}$. We have that $S_T > 0$ with positive probability. Indeed, since $S_1 = 1$ with probability $1/2$, $\mathbb{P}(S_T \geq 1) \geq 1/2$. Therefore,

   $$\mathbb{E}(S_T) > 0 = \mathbb{E}(S_0).$$

Notice that these are two distinct obstructions: in 2., the time T to say stop is bounded by 3.

If we want $\mathbb{E}(M_T) = \mathbb{E}(M_0)$ to hold for a random time T, we then have to rule out these two obstructions. Ruling out arbitrarily long times is done by assuming that the random time T is **bounded**: there exists $b \in \mathbb{R}$ such that $T \leq b$ holds with probability 1. Ruling out peeks into the future is done by forcing T to be a stopping time:

Definition 5.1.2. Given a stochastic process $\{X_n\}_{n\geq 0}$, a **stopping time** T is a discrete random variable defined on the same probability space as X_n, taking values in $\{0, 1, 2, \dots\} \cup \{\infty\}$, and such that the event $\{T = n\}$ is completely determined by the familiy $X_{0,n}$ i.e., $\{T = n\}$ can be written in terms of events of the form $\{X_0 \in A_0, \dots, X_n \in A_n\}$.

Example 5.1.3. Consider a Markov chain $\{X_n\}_{n\geq 0}$. The first-passage time T_j to state j is a stopping time. Indeed, $\{T_j = n\} = \{X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j\}$.

Once we get rid of the obvious obstructions, we indeed have $\mathbb{E}(M_T) = \mathbb{E}(M_0)$:

Theorem 5.1.4 (Optional sampling theorem). *Let $\{M_n\}_{n\geq 0}$ be a martingale with respect to $\{W_n\}_{n\geq 0}$ and let T be a bounded stopping time. Then $\mathbb{E}(M_T) = \mathbb{E}(M_0)$.*

In a market setting, the moral is that you can't make money (in expectation) by buying and selling an asset whose price is a martingale. More precisely, if you buy the asset at some time and adopt any strategy at all for deciding when to sell it, then the expected price at the time you sell is the price you originally paid. In other words, if the market price is a martingale, you cannot make money in expectation by "timing the market."

Proof of Theorem 5.1.4. Suppose the stopping time T is bounded by n i.e., $T(\omega) \leq n$, for all $\omega \in \Omega$. Consider the indicator random variable $I_{\{T=j\}}$ and recall that

$$I_{\{T=j\}} = \begin{cases} 1 & \text{if } \{T = j\} \text{ occurs;} \\ 0 & \text{otherwise.} \end{cases}$$

Since M_T is the random variable which equals M_j if $T = j$, we can write

$$M_T = \sum_{j=0}^{n} M_j I_{\{T=j\}}.$$

Therefore,

$$\mathbb{E}(M_T|W_{0,n-1}) = \mathbb{E}\Big(\sum_{j=0}^{n} M_j I_{\{T=j\}}|W_{0,n-1}\Big) = \mathbb{E}(M_n I_{\{T=n\}}|W_{0,n-1}) + \sum_{j=0}^{n-1} \mathbb{E}(M_j I_{\{T=j\}}|W_{0,n-1}), \quad (5.2)$$

where in the last equality we used Theorem 5.0.7(ii). But for each $j \in \{0, \ldots, n-1\}$, $M_j I_{\{T=j\}}$ can be written as a function of $W_{0,n-1}$ (as T is a stopping time and $\{M_n\}_{n\geq 0}$ is a martingale) and so Theorem 5.0.7(vi) implies that

$$\mathbb{E}(M_j I_{\{T=j\}}|W_{0,n-1}) = M_j I_{\{T=j\}}. \quad (5.3)$$

Therefore (5.2) becomes

$$\mathbb{E}(M_T|W_{0,n-1}) = \mathbb{E}(M_n I_{\{T=n\}}|W_{0,n-1}) + \sum_{j=0}^{n-1} M_j I_{\{T=j\}}. \quad (5.4)$$

Let us now look at $\mathbb{E}(M_n I_{\{T=n\}}|W_{0,n-1})$. Since $T \leq n$, the event $\{T = n\}$ is the same as the event $\{T > n-1\}$. But T is a stopping time and so $I_{\{T>n-1\}}$ can be written as a function of $W_{0,n-1}$. Theorem 5.0.7(vi) then implies that

$$\mathbb{E}(M_n I_{\{T=n\}}|W_{0,n-1}) = \mathbb{E}(M_n I_{\{T>n-1\}}|W_{0,n-1}) = I_{\{T>n-1\}}\mathbb{E}(M_n|W_{0,n-1}). \quad (5.5)$$

Since $\{M_n\}_{n\geq 0}$ is a martingale with respect to $\{W_n\}_{n\geq 0}$, we have that $\mathbb{E}(M_n|W_{0,n-1}) = M_{n-1}$. Therefore, (5.4) becomes

$$\begin{aligned}\mathbb{E}(M_T|W_{0,n-1}) &= \mathbb{E}(M_n I_{\{T=n\}}|W_{0,n-1}) + \sum_{j=0}^{n-1} M_j I_{\{T=j\}} \\ &= I_{\{T>n-1\}} M_{n-1} + \sum_{j=0}^{n-1} M_j I_{\{T=j\}} \\ &= M_{n-1}\big(I_{\{T>n-1\}} + I_{\{T=n-1\}}\big) + \sum_{j=0}^{n-2} M_j I_{\{T=j\}} \\ &= I_{\{T>n-2\}} M_{n-1} + \sum_{j=0}^{n-2} M_j I_{\{T=j\}}.\end{aligned}$$

We can now condition the random variable $\mathbb{E}(M_T|W_{0,n-1})$ on the set of random variables $W_{0,n-2}$. We obtain

$$\begin{aligned}\mathbb{E}(\mathbb{E}(M_T|W_{0,n-1})|W_{0,n-2}) &= \mathbb{E}\Big(I_{\{T>n-2\}} M_{n-1} + \sum_{j=0}^{n-2} M_j I_{\{T=j\}}|W_{0,n-2}\Big) \\ &= \mathbb{E}(I_{\{T>n-2\}} M_{n-1}|W_{0,n-2}) + \sum_{j=0}^{n-2} \mathbb{E}(M_j I_{\{T=j\}}|W_{0,n-2}) \\ &= I_{\{T>n-2\}} M_{n-2} + \sum_{j=0}^{n-2} M_j I_{\{T=j\}} \\ &= I_{\{T>n-3\}} M_{n-2} + \sum_{j=0}^{n-3} M_j I_{\{T=j\}},\end{aligned}$$

where the third equality follows from (5.3) and (5.5) applied to $n-2$ instead of $n-1$. On the other hand, recalling Exercise 5.0.16, we have

$$\mathbb{E}(\mathbb{E}(M_T|W_{0,n-1})|W_{0,n-2}) = \mathbb{E}(M_T|W_{0,n-2})$$

and so

$$\mathbb{E}(M_T|W_{0,n-2}) = I_{\{T>n-3\}} M_{n-2} + \sum_{j=0}^{n-3} M_j I_{\{T=j\}}.$$

Repeating this argument again (until $n-2$ reaches 0), we end up with

$$\mathbb{E}(M_T|W_{0,0}) = M_0$$

and so, taking expectations of both sides, $\mathbb{E}(\mathbb{E}(M_T|W_{0,0})) = \mathbb{E}(M_0)$. But Theorem 5.0.7(v) implies that $\mathbb{E}(\mathbb{E}(M_T|W_{0,0})) = \mathbb{E}(M_T)$ and so $\mathbb{E}(M_T) = \mathbb{E}(M_0)$, as desired. □

We now give some examples showing how the Optional sampling theorem can be applied and a standard trick that can be used to deal with unbounded stopping times.

Example 5.1.5. Consider the symmetric simple random walk $\{S_n\}_{n\geq 0}$ on $\mathbb{Z}$ with $S_0 = 0$. We have seen in Example 5.0.10 that it is a martingale, as the mean of each X_t is 0. Let a and b be integers with $a < 0 < b$. We will compute the probabilities of reaching b before a and a before b.

Let $T_a = \inf\{n : S_n = a\}$ and $T_b = \inf\{n : S_n = b\}$. T_a and T_b are both stopping times. Indeed, $\{T_b = n\}$ coincides with the event $\{S_0 < b, S_1 < b, \ldots, S_{n-1} < b, S_n = b\}$. Similarly for T_a (work out the details). Moreover, the following holds:

Exercise 5.1.6. *If T_1 and T_2 are stopping times, then $T = \min\{T_1, T_2\}$ is a stopping time as well.*

In particular, $T = \min\{T_a, T_b\}$ is a stopping time. Notice that T is the first time the random walk hits either a or b. However, it is not necessarily bounded! In order to apply Theorem 5.1.4, we define $T_m = \min\{T, m\}$, for each m. Exercise 5.1.6 implies that T_m is a stopping time and T_m is clearly bounded (by m). Therefore, by Theorem 5.1.4,

$$\mathbb{E}(S_{T_m}) = \mathbb{E}(S_0) = 0,$$

for each m.

We now show that $\mathbb{E}(S_T) = 0$. Since the symmetric simple random walk is recurrent (Example 2.1.16), we have that $\mathbb{P}(T < \infty) = 1$. Moreover, for each ω such that $T(\omega) < \infty$, we have $T_m(\omega) = T(\omega)$ for sufficiently large m (we can simply take m such that $m \geq T(\omega)$). Therefore, for each ω such that $T(\omega) < \infty$, we have $S_{T_m}(\omega) = S_T(\omega)$ for sufficiently large m and so $S_{T_m} \xrightarrow{\text{a.s.}} S_T$. Moreover, $|S_{T_m}| \leq \max\{|a|, |b|\}$ for each m. We now make use of the following important result which we already mentioned. One of its many formulations reads as follows:

Theorem 5.1.7 (Dominated convergence theorem). *If $X_n \xrightarrow{\text{a.s.}} X$ and there exists a random variable Y such that $|X_n| \leq Y$ for each n and $\mathbb{E}(Y) < \infty$, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$ as $n \to \infty$.*

Using the Dominated convergence theorem, we conclude that $\mathbb{E}(S_{T_m}) \to \mathbb{E}(S_T)$. But since $\mathbb{E}(S_{T_m}) = 0$ for each m, we must have $\mathbb{E}(S_T) = 0$, as claimed.

Notice now that the random variable S_T takes two values: a with probability $\mathbb{P}(S_T = a)$ and b with probability $\mathbb{P}(S_T = b)$. Therefore,

$$0 = \mathbb{E}(S_T) = a\mathbb{P}(S_T = a) + b\mathbb{P}(S_T = b) = a(1 - \mathbb{P}(S_T = b)) + b\mathbb{P}(S_T = b),$$

from which

$$\mathbb{P}(S_T = b) = \frac{-a}{b-a} \quad \text{and} \quad \mathbb{P}(S_T = a) = \frac{b}{b-a}.$$

These are the probabilities of reaching b before a and a before b, respectively.

Example 5.1.8. Let us consider again the random walk $\{S_n\}_{n\geq 0}$ and the stopping time T from the previous example. We will compute $\mathbb{E}(T)$, the expected time until the walk reaches a or b. To do this, we consider another martingale associated with the random walk, namely $M_n = S_n^2 - n$. We first show it is indeed a martingale. By Theorem 5.0.7(ii) and (i), we have that

$$\mathbb{E}(S_{n+1}^2 - (n+1)|S_{0,n}) = \mathbb{E}(S_{n+1}^2|S_{0,n}) - \mathbb{E}(n+1|S_{0,n}) = \mathbb{E}(S_{n+1}^2|S_{0,n}) - (n+1).$$

Moreover, since $S_{n+1} = S_n + X_{n+1}$, we have

$$\begin{aligned}
\mathbb{E}(S_{n+1}^2|S_{0,n}) &= \mathbb{E}(S_n^2 + X_{n+1}^2 + 2S_nX_{n+1}|S_{0,n}) \\
&= \mathbb{E}(S_n^2|S_{0,n}) + \mathbb{E}(X_{n+1}^2|S_{0,n}) + \mathbb{E}(2S_nX_{n+1}|S_{0,n}) \\
&= S_n^2 + \mathbb{E}(X_{n+1}^2|S_{0,n}) + 2S_n\mathbb{E}(X_{n+1}|S_{0,n}) \\
&= S_n^2 + \mathbb{E}(X_{n+1}^2) + 2S_n\mathbb{E}(X_{n+1}) \\
&= S_n^2 + 1 + 0,
\end{aligned}$$

where in the second equality we used Theorem 5.0.7(ii), in the third, Theorem 5.0.7(vi), in the fourth, Equation (5.1) and in the fifth the fact that $\mathbb{E}(X_{n+1}) = 0$ and $\mathbb{E}(X_{n+1}^2) = 1 \cdot \frac{1}{2} + (-1)^2 \cdot \frac{1}{2} = 1$. Therefore,

$$\mathbb{E}(S_{n+1}^2 - (n+1)|S_{0,n}) = S_{n+1}^2 + 1 - (n+1) = S_n^2 - n,$$

and so $\{M_n\}_{n\geq 0}$ is indeed a martingale.

As in the previous example, consider the bounded stopping time $T_m = \min\{T, m\}$. By Theorem 5.1.4 applied to $\{M_n\}_{n\geq 0}$ and T_m, we have that

$$\mathbb{E}(S_{T_m}^2 - T_m) = \mathbb{E}(M_{T_m}) = \mathbb{E}(M_0) = \mathbb{E}(S_0^2 - 0) = 0$$

and so

$$\mathbb{E}(S_{T_m}^2) = \mathbb{E}(T_m), \tag{5.6}$$

for all m. But we have seen that, for each ω such that $T(\omega) < \infty$, $T_m(\omega) = T(\omega)$ for sufficiently large m and so $S_{T_m}^2 \xrightarrow{\text{a.s.}} S_T^2$. Applying the Dominated convergence theorem to the sequences $\{T_m\}$ and $\{S_{T_m}^2\}$, we then obtain that $\mathbb{E}(T_m) \to \mathbb{E}(T)$ and $\mathbb{E}(S_{T_m}^2) \to \mathbb{E}(S_T^2)$. By (5.6) and uniqueness of limits, we conclude that $\mathbb{E}(T) = \mathbb{E}(S_T^2)$. Recall now from the previous example that

$$\mathbb{P}(S_T = b) = \frac{-a}{b-a} \quad \text{and} \quad \mathbb{P}(S_T = a) = \frac{b}{b-a}$$

and so

$$\mathbb{E}(T) = \mathbb{E}(S_T^2) = a^2 \cdot \frac{b}{b-a} + b^2 \cdot \frac{-a}{b-a} = -ab = |a|b.$$

This tells us, for example, that the expected time until a random walk wanders 100 units in either direction away from its starting position is 100^2.

We can apply the results in Examples 5.1.5 and 5.1.8 to the gambler's ruin problem:

Example 5.1.9 (Gambler's ruin one more time). Recall that a man wants to buy a car at a cost of N units of money. He starts with k units, for some $0 < k < N$ and tries to win the remainder by tossing a fair coin repeatedly: If heads comes up, then he wins one unit, if tails comes up, he loses one unit. He plays the game repeatedly until one of two events occurs: either he runs out of money and is bankrupted or he wins enough to buy the car.

We can interpret the gamble as a simple random walk on $\mathbb{Z}$ starting at 0, where the probability of moving in either direction is $1/2$. The probability of being bankrupted is exactly the probability that the random walk reaches $-k$ before reaching $N - k$. The computation in Example 5.1.5 shows that this probability is

$$\frac{N-k}{N} = 1 - \frac{k}{N},$$

in accordance with our old computation (see Example 1.3.6). We can also compute the expected duration of the gamble. By Example 5.1.8, it is $k(N-k)$. The moral is that:

> In a fair gamble, you expect to play for the product of the amount you are willing to lose times the amount you want to win.

5.2 Option pricing

A derivative security is a financial contract whose value is derived from the value of another underlying security, such as a stock or a bond. For example, a derivative security based on an underlying stock would pay off various amounts at various times depending on the behavior of the price of the stock. In this section, we give an idea of the theory predicting the prices of such derivative securities and its connection to martingales. We consider discrete-time models.

Let S_t be the stock price at time t. The stock price process is the discrete-time stochastic process $\{S_t\}_{t\geq 0}$. We assume that at an agreed-upon future time n, the derivative security pays an amount X that is some function $X = g(S_{0,n})$ of the stock price history up to that time. Let us give an example of a derivative security: a call option.

Example 5.2.1. A **call option** on a given underlying stock is the right to buy a share of the stock at a certain fixed price c (the **strike price**) at a certain fixed time n in the future (the **maturity date**). If I buy a call option from you, I am paying you some money in return for the right to force you to sell me a share of the stock, if I want it, at the strike price on the maturity date. If $S_n > c$, then the buyer of the option will exercise his right at time n, buying the stock for c and selling it for S_n, gaining a net of $S_n - c$. If $S_n \leq c$, then it is not profitable to buy the stock at price c, so the option is not exercised, and the gain at time n is 0. Here we are obviously assuming rational behavior. In summary,

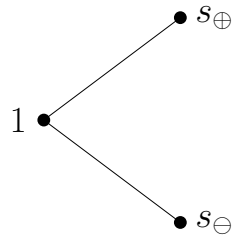
$$X = g(S_{0,n}) = \begin{cases} S_n - c & \text{if } S_n > c; \\ 0 & \text{otherwise.} \end{cases}$$

The natural problem of pricing derivative securities arises: How do we assure a "fair" price? This can be done assuming the no-arbitrage principle, one of the cornerstones of modern finance. An **arbitrage** is a transaction that makes money without risk, that is, with no chance of losing money. Such transactions should not exist. Here is the rough motivation. Suppose that the price of A is so low that some clever set of transactions involving buying A and perhaps selling something else is guaranteed to make a riskless profit. Many eager arbitrage seekers would try to perform these transactions many times. The resulting increased demand for A would cause its price to increase, thereby destroying the arbitrage opportunity. The no-arbitrage principle can be summarized as "There is no such thing as a free lunch", an adage popularized by Milton Friedman. A consequence of the no-arbitrage principle is the following: If there are two portfolios that give the same sets of payoffs at the same times (for example, two different combinations of securities that produce the same rewards under all circumstances), then those portfolios must have the same price.

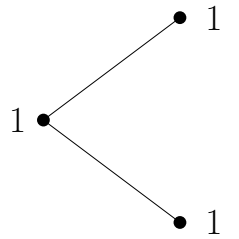
Given the no-arbitrage assumption, the "fair" price of a derivative security can be computed based on the following result: The derivative security is actually *redundant*, in the sense that there is a portfolio involving just the stock and a bond that produces exactly the same payoffs as the derivative security itself. Assuming this result, and since we are given the prices of the stock and the bond, we can then compute the price of the reproducing portfolio which, by the no-arbitrage principle, must be the same as the price of the derivative security. This is exemplified in the following.

Example 5.2.2. Consider the Mickey Mouse model of the market with only one period and two states corresponding to the stock price rising or falling. At time 0 the stock price is 1 and at time 1 the stock price is either $s_\oplus$ or $s_\ominus$, where $s_\ominus < 1 < s_\oplus$. Suppose we also have a bond in our portfolio, for simplicity with interest rate zero: investing $\$1$ at time 0 returns exactly $\$1$ at time 1. We can think of it this way. Buying b shares of the bond corresponds to lending out $\$b$ for one period: we lose $\$b$ at time 0 but then gain back $\$b$ at time 1. If $b < 0$ this corresponds to borrowing $\$b$ for one period. In other words, assuming an interest rate of zero means that we can lend or borrow money with no cost.

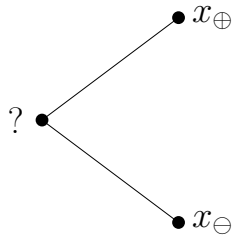
The redundant derivative security has payoffs that are a function of $s_\oplus$ and $s_\ominus$. Suppose that it pays $x_\oplus$ if the stock price goes up to $s_\oplus$ and $x_\ominus$ if the stock price goes down to $s_\ominus$. We want to determine the no-arbitrage price of the derivative security.



stock



bond



derivative security

We assume the stock and bond can be traded in continuous amounts, so that one can buy 2.718 shares of stock or sell 3.14 bonds, for example. Let a and b denote the number of stock and bond shares in the portfolio, respectively. Since the payoffs at time 1 from such a portfolio are $as_\oplus + b$ if the stock goes up and $as_\ominus + b$ if the stock goes down, the requirement that the portfolio reproduces the payoffs of the redundant derivative security consists of the equations

$$as_\oplus + b = x_\oplus \quad \text{and} \quad as_\ominus + b = x_\ominus.$$

Solving the system, we obtain

$$a = \frac{x_\oplus - x_\ominus}{s_\oplus - s_\ominus} \quad \text{and} \quad b = \frac{x_\ominus s_\oplus - x_\oplus s_\ominus}{s_\oplus - s_\ominus}.$$

Therefore, the price π that we pay for this portfolio at time 0 is

$$\pi = a + b = x_\oplus\left(\frac{1 - s_\ominus}{s_\oplus - s_\ominus}\right) + x_\ominus\left(\frac{s_\oplus - 1}{s_\oplus - s_\ominus}\right).$$

If the price of the redundant derivative security were anything other than π, we would have two investments (the redundant security and the reproducing portfolio) that have different prices but exactly the same payoffs, thus violating the no-arbitrage principle. Therefore, the price for the redundant derivative security implied by the no-arbitrage assumption is π.

There is a nice interpretation of π that makes it easy to remember and points toward the connection with martingales. Letting

$$p = \frac{1 - s_\ominus}{s_\oplus - s_\ominus},$$

we found that

$$\pi = x_\oplus p + x_\ominus(1 - p). \tag{5.7}$$

If, for some reason, p is the probability that the stock price rises, then the price π would simply be the expected value of the payoff of the redundant derivative security. But the no-arbitrage argument we used to determine π is independent of the probability of the stock price rising! Hence the magic of option pricing: the "probability" p has nothing to do with the probability of the stock price rising. It does, however, have an interesting and useful interpretation: the value $p = \frac{1-s_\ominus}{s_\oplus - s_\ominus}$ is the probability that makes the stock price a martingale. Indeed, if p' is the probability of the stock price rising, the expected value of the stock price at time 1 is $s_\oplus p' + s_\ominus(1 - p')$. For the stock price to be a martingale, this last

expression must be the same as the expected value of the stock price at 0, which is 1. This happens precisely when

$$p' = \frac{1 - s_\ominus}{s_\oplus - s_\ominus}.$$

The equality (5.7) says that the price of the redundant derivative security is its expected payoff, where the expectation is taken under the probability measure that makes the stock price a martingale.

We now allow more than one period in our model of the market. Imagine that the stock price process can be described as a bifurcating tree, where for each possible history of the stock price up to time t, there are just two possible values for the price at time $t+1$ (see Figure 5.1). We describe a path, or history, in the stock price tree up to time t by a sequence of binary random variables $W_1, W_2, \ldots, W_t$, where $W_k = \oplus$ means that the price took the larger of the two possible values at time k, and $W_k = \ominus$ means that the price took the smaller of the two possible values at time k. We assume that the initial stock price s_0 is known. The stock price S_t at time t is then determined by the history $W_{1,t}$ and so we can view it as a function of $W_{1,t}$. To emphasize this, we will write $S_t = S_t(W_{1,t})$.

Example 5.2.3. Suppose that, in each period, the stock price either double or half. For example, if the stock price starts out at $s_0 = 8$ and goes up in the first 3 periods, then at time 3 the price is $S_3(\oplus, \oplus, \oplus) = 64$. The price after 3 "down" periods would be $S_3(\ominus, \ominus, \ominus) = 1$. The possible paths of the stock price for the first 3 periods are depicted in Figure 5.1.

This model might appear very artificial, as we would not expect a stock to either double or half each period. But it becomes somewhat realistic when there are many very short periods and the stock price can gain or lose some very small percentage each period.

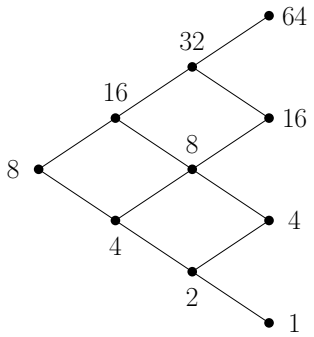


Figure 5.1: Stock price tree.

We have not yet specified the probabilities of the various stock price paths, namely the probabilities of the 8 paths $(\ominus, \ominus, \ominus), (\ominus, \ominus, \oplus), \ldots, (\oplus, \oplus, \oplus)$. For example, we might assume that all 8 paths are equally likely, which is equivalent to assuming a probability measure $\mathbb{P}$ under which the random variables W_1, W_2, W_3 are independent with $\mathbb{P}(W_i = \oplus) = \mathbb{P}(W_i = \ominus) = 1/2$, as depicted in Figure 5.2. In fact, similarly to the one-period case, it will turn out that these probabilities do not matter.

Given the stock price $S_t = S_t(W_{1,t})$ at time t, the stock price S_{t+1} has the two possible values $S_{t+1}(W_{1,t}, \oplus)$ or $S_{t+1}(W_{1,t}, \ominus)$. Let us assume that the current stock price is always strictly between the two possible prices at the next period i.e.,

$$S_{t+1}(W_{1,t}, \ominus) < S_t(W_{1,t}) < S_{t+1}(W_{1,t}, \oplus). \tag{5.8}$$

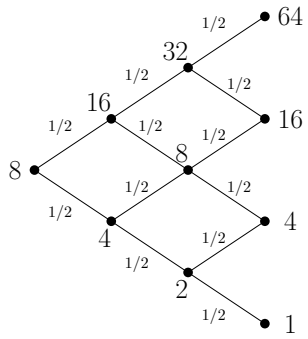


Figure 5.2

We now want to price a derivative security X whose value at time n is a function of $W_{1,n}$. For example, for the stock price process in Example 5.2.3, a call option with strike price 10 at time 3 would have the payoffs depicted in Figure 5.3.

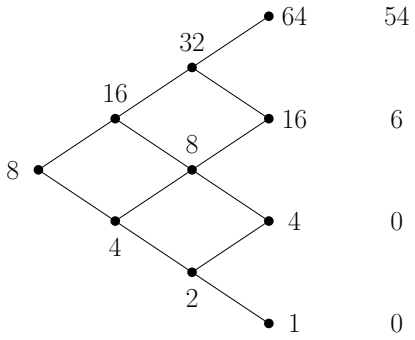


Figure 5.3: Payoffs of call option with strike price 10.

As in the one-period case, the key for pricing the derivative security is to show that it is redundant: its payoffs can be duplicated using the stock and a bond. In order to do that, we need the following notion. A **trading strategy** consists of a specification of the number of stock shares that we hold at each time period. Let H_t denote the number of shares held at time t. Think of it this way: at time t we buy H_t shares to hold over the interval from t to $t+1$. This choice is based on the history of the stock price process up to time t and cannot depend on future information. The payoff of this strategy in the first period is $H_0(S_1 - S_0)$, the number of shares held at time 0 times the amount the stock price rises. Accumulating such gains over the first n periods, the gain at time n from the strategy $H_0, H_1, \ldots$ is

$$(H \bullet S)_n = H_0(S_1 - S_0) + H_1(S_2 - S_1) + \cdots + H_{n-1}(S_n - S_{n-1}). \tag{5.9}$$

We can think of it the following way. At time 0, we buy H_0 shares, which costs $H_0 S_0$. In fact, let us imagine that we borrow $H_0 S_0$ at time 0 to finance the stock purchase. So in fact we neither gain nor lose money at time 0, since we gain $H_0 S_0$ by borrowing it and then spend that same amount of money to buy the stock. At time 1, we sell our H_0 shares (remember that H_0 is the number of shares we hold just for the first period), gaining $H_0 S_1$, and we pay off the money we borrowed, namely $H_0 S_0$. So the strategy $H_0, H_1, \ldots$ produces payoffs $H_0(S_1 - S_0)$ at time 1 and costs nothing to perform. The reasoning can then be repeated.

Remark 5.2.4. As a side remark, the process $H \bullet S$ is an example of a **discrete-time stochastic integral**. Loosely speaking, an integral is a sum of products, or a limit of sums of products. For example, the so-called Riemann-Stieltjes integral $\int_a^b f \, \mathrm{d}g$ is defined to be a limit of sums of the form $\sum_{i=0}^{k-1} f(x_i)(g(x_{i+1}) -$

$g(x_i))$ as $k \to \infty$, where $a = x_0 < x_1 < \cdots < x_k = b$ and $\max_{i<k}(x_{i+1} - x_i) \to 0$ as $k \to \infty$. Notice that the Riemann integral $\int_a^b f(x)\,\mathrm{d}x$ you are familiar with is the special case of the Riemann-Stieltjes integral where g is the identity function $g(x) = x$. $\int f\,\mathrm{d}g$ is then a sum of products of values of f with changes in values of g. Analogously, (5.9) is a sensible way to define an integral $\int H\,\mathrm{d}S$ for discrete-time processes.

It turns out that in the type of bifurcating tree models of stock prices we are considering, the payoffs given by any derivative security X with maturity date n may be reproduced precisely by the sum of a constant and some trading strategy:

$$X = x_0 + (H \bullet S)_n, \tag{5.10}$$

for some $x_0 \in \mathbb{R}$ and trading strategy H. Loosely speaking, we can reproduce the payoffs of the derivative security X by trading the stock. We can obtain (5.10) from the following martingale representation result:

Theorem 5.2.5 (Martingale representation). *Let $\{S_t\}_{t\geq 0}$ be a martingale with respect to the probability measure $\mathbb{Q}$. For any other martingale $\{M_t\}_{t\geq 0}$ with respect to $\mathbb{Q}$, there is a trading strategy H such that*

$$M_t = M_0 + (H \bullet S)_t.$$

Proof of (5.10) *using Theorem 5.2.5.* It is enough to define a martingale $\{M_t\}_{t\geq 0}$ such that $M_n = X$. Consider the Doob's martingale $M_t = \mathbb{E}_{\mathbb{Q}}(X|W_{1,t})$ (with the notation $\mathbb{E}_{\mathbb{Q}}(\cdot)$ we stress the fact that expectations are taken with respect to the probability measure $\mathbb{Q}$). Notice that $M_n = X$, as we have assumed that X is a function of $W_{1,n}$. Moreover, $M_0 = \mathbb{E}_{\mathbb{Q}}(X)$. Thus, applying the martingale representation theorem to $\{M_t\}_{t\geq 0}$, there is a trading strategy H such that

$$X = \mathbb{E}_{\mathbb{Q}}(X) + (H \bullet S)_n,$$

as desired. □

A consequence of (5.10) is the following:

Lemma 5.2.6. *Suppose that $S_{t+1}(W_{1,t}, \ominus) < S_t(W_{1,t}) < S_{t+1}(W_{1,t}, \oplus)$ holds for each t and that each path in the tree has positive probability. If the derivative security X satisfies* (5.10)*, then the no-arbitrage price of X is x_0.*

The lemma reduces the problem of pricing X to that of finding x_0 in the representation (5.10). A stochastic process, such as the price process for the stock, takes various possible paths with various probabilities. Different probability measures will allocate probability among paths differently. For discrete processes of the type we have been discussing, some paths will have positive probability and some will have zero probability.

Definition 5.2.7. Two probability measures $\mathbb{P}$ and $\mathbb{Q}$ for a process are **equivalent** if they agree on which sets of paths have zero probability and which sets of paths have positive probability.

In the case of Example 5.2.3 with the probability measure $\mathbb{P}$ depicted in Figure 5.2, any equivalent probability measure $\mathbb{Q}$ will simply reassign probabilities among the set of paths already taken by $\mathbb{P}$.

Definition 5.2.8. Given a process and a probability measure $\mathbb{P}$, a probability measure $\mathbb{Q}$ is an **equivalent martingale measure** if $\mathbb{Q}$ is equivalent to $\mathbb{P}$ and the process is a martingale under the measure $\mathbb{Q}$.

A martingale measure $\mathbb{Q}$ makes the identification of the desired price x_0 in (5.10) easy. Indeed, since the martingale property gives $\mathbb{E}_{\mathbb{Q}}(S_{t+1} - S_t|W_{1,t}) = 0$, we have

$$\mathbb{E}_{\mathbb{Q}}\big(H_t(S_{t+1} - S_t)\big) = \mathbb{E}_{\mathbb{Q}}\big(\mathbb{E}_{\mathbb{Q}}(H_t(S_{t+1} - S_t)|W_{1,t})\big) = \mathbb{E}_{\mathbb{Q}}\big(H_t\mathbb{E}_{\mathbb{Q}}(S_{t+1} - S_t|W_{1,t})\big) = 0,$$

where the first equality follows from Theorem 5.0.7(v) and the second from Theorem 5.0.7(vi). Therefore,

$$\mathbb{E}_{\mathbb{Q}}((H \bullet S)_n) = \mathbb{E}_{\mathbb{Q}}(H_0(S_1 - S_0)) + \cdots + \mathbb{E}_{\mathbb{Q}}(H_{n-1}(S_n - S_{n-1})) = 0.$$

and taking the expectation under $\mathbb{Q}$ of both sides of (5.10), we obtain

$$x_0 = \mathbb{E}_{\mathbb{Q}}(X),$$

in accordance with the one-period case. To summarize:

> If the stock price is governed by a probability measure $\mathbb{P}$ on the paths in a bifurcating tree, letting $\mathbb{Q}$ be a probability measure equivalent to $\mathbb{P}$ under which the stock price is a martingale, the no-arbitrage price of the derivative security X is $\mathbb{E}_{\mathbb{Q}}(X)$.

Example 5.2.9. The measure $\mathbb{P}$ in Example 5.2.3 is not a martingale measure. Indeed, if the current price is 4, the price under $\mathbb{P}$ at the next period will be either 2 or 8 with equal probability, giving an expected value of $5 \neq 4$. On the other hand, the equivalent measure $\mathbb{Q}$ such that $\mathbb{Q}(W_k = \oplus) = 1/3$ and $\mathbb{Q}(W_k = \ominus) = 2/3$ is a martingale measure (check this!).

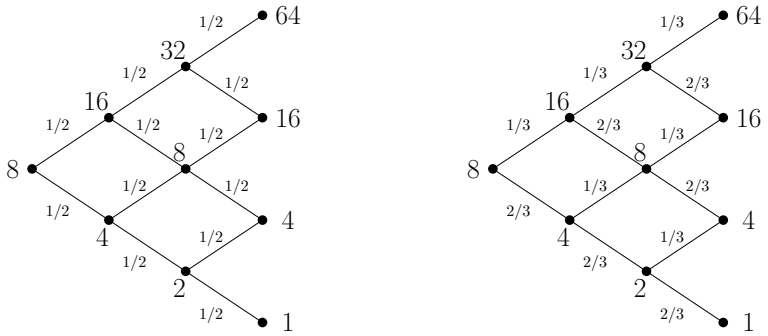


Figure 5.4: The measure $\mathbb{P}$ on the left and the measure $\mathbb{Q}$ on the right.

The no-arbitrage price of the call option with strike price 10 is then

$$54\left(\left(\frac{1}{3}\right)^3\right) + 6\left(3\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)\right) = \frac{10}{3}.$$

5.3 Ballot theorem

In this section we consider a famous result obtained by Bertrand in 1887 and called Ballot theorem. Suppose that two candidates A and B run for an election. Candidate A obtains a votes and candidate B obtains $b < a$ votes. The votes are counted in a random order chosen uniformly at random from all permutations on the $a + b$ votes. What is the probability that A is always ahead in the count? We show that this probability is

$$\frac{a-b}{a+b}.$$

We provide two proofs, one using martingales and the Optional sampling theorem and one using an elementary technique called the *reflection principle*. This technique will also allow us to obtain several interesting results for random walks.

We first prove the Ballot theorem using martingales. Let $n = a + b$ be the total number of votes and let S_k be the random number of votes by which A is leading after k votes are counted (note that S_k might be negative). Clearly, $S_n = a - b$. We build a finite martingale $\{X_k\}_{0 \leq k \leq n-1}$ as follows:

$$X_k = \frac{S_{n-k}}{n-k}.$$

Let us verify it is indeed a martingale. Consider $\mathbb{E}(X_k | X_{0,k-1})$ and observe that conditioning on $X_{0,k-1}$ is equivalent to conditioning on $S_n, \ldots, S_{n-k+1}$, which in turn is equivalent to conditioning on the values of the count when counting the last $k-1$ votes. Let a_k be the number of votes for A after the first k votes are counted and, similarly, let b_k be the number of votes for B after the first k votes are counted. We have that

$$a_{n-k+1} = \frac{a_{n-k+1} + b_{n-k+1} + (a_{n-k+1} - b_{n-k+1})}{2} = \frac{n-k+1+S_{n-k+1}}{2}$$

and

$$b_{n-k+1} = \frac{a_{n-k+1} + b_{n-k+1} - (a_{n-k+1} - b_{n-k+1})}{2} = \frac{n-k+1-S_{n-k+1}}{2}.$$

Now, the $(n-k+1)$th vote is a random vote among the first $n-k+1$ votes. Moreover, S_{n-k} is equal to $S_{n-k+1}+1$ if the $(n-k+1)$th vote was for B, and equal to $S_{n-k+1}-1$ if the $(n-k+1)$th vote was for A. Therefore, for $k \geq 1$, we obtain

$$\begin{aligned}\mathbb{E}(S_{n-k} | S_{n-k+1}) &= (S_{n-k+1}+1)\frac{n-k+1-S_{n-k+1}}{2(n-k+1)} + (S_{n-k+1}-1)\frac{n-k+1+S_{n-k+1}}{2(n-k+1)} \\ &= S_{n-k+1} \cdot \frac{n-k}{n-k+1}.\end{aligned}$$

But then

$$\mathbb{E}(X_k | X_{0,k-1}) = \mathbb{E}\left(\frac{S_{n-k}}{n-k} | S_n, \ldots, S_{n-k+1}\right) = \frac{1}{n-k}\mathbb{E}(S_{n-k} | S_{n-k+1}) = \frac{S_{n-k+1}}{n-k+1} = X_{k-1},$$

showing that $X_0, \ldots, X_{n-1}$ is indeed a martingale.

Let now T be the minimum k such that $X_k = 0$, if such k exists, and $T = n-1$ otherwise. In this way we obtain a bounded stopping time and so, by the Optional sampling theorem,

$$\mathbb{E}(X_T) = \mathbb{E}(X_0) = \frac{\mathbb{E}(S_n)}{n} = \frac{a-b}{a+b}.$$

We now consider two cases:

1. A leads throughout the count. In this case, $S_{n-k} > 0$ for each $k \in \{0, \ldots, n-1\}$ (and so all X_k are positive), $T = n-1$ and $X_T = X_{n-1} = S_1 = 1$.

2. A does not lead throughout the count. In this case we claim that $X_k = 0$ for some $k < n-1$. Indeed, since A has more votes at the end, if B ever leads then there must be some intermediate point k such that $S_{n-k} = 0$. Taking the minimum such k, we have that $X_k = 0$, from which $T = k < n-1$ and $X_T = 0$.

But then

$$\frac{a-b}{a+b} = \mathbb{E}(X_T) = 1 \cdot \mathbb{P}(\text{case 1 occurs}) + 0 \cdot \mathbb{P}(\text{case 2 occurs}),$$

from which we obtain that the probability that case 1 occurs i.e., that A leads throughout the count, is indeed $\frac{a-b}{a+b}$.

We now provide another proof of the Ballot theorem, this time with a combinatorial flavor. It is based on the reflection principle, an elementary but useful result in the context of random walks. Consider the following random walk on $\mathbb{Z}$ with starting point a: we have a sequence $X_1, X_2, \ldots$ of i.i.d. random variables taking value 1 with probability p and -1 with probability $q = 1-p$ and we let $S_n = a + \sum_{k=1}^{n} X_k$. We keep a record of the random walk through its path $\{(n, S_n) : n \geq 0\}$ as depicted in Figure 5.5.

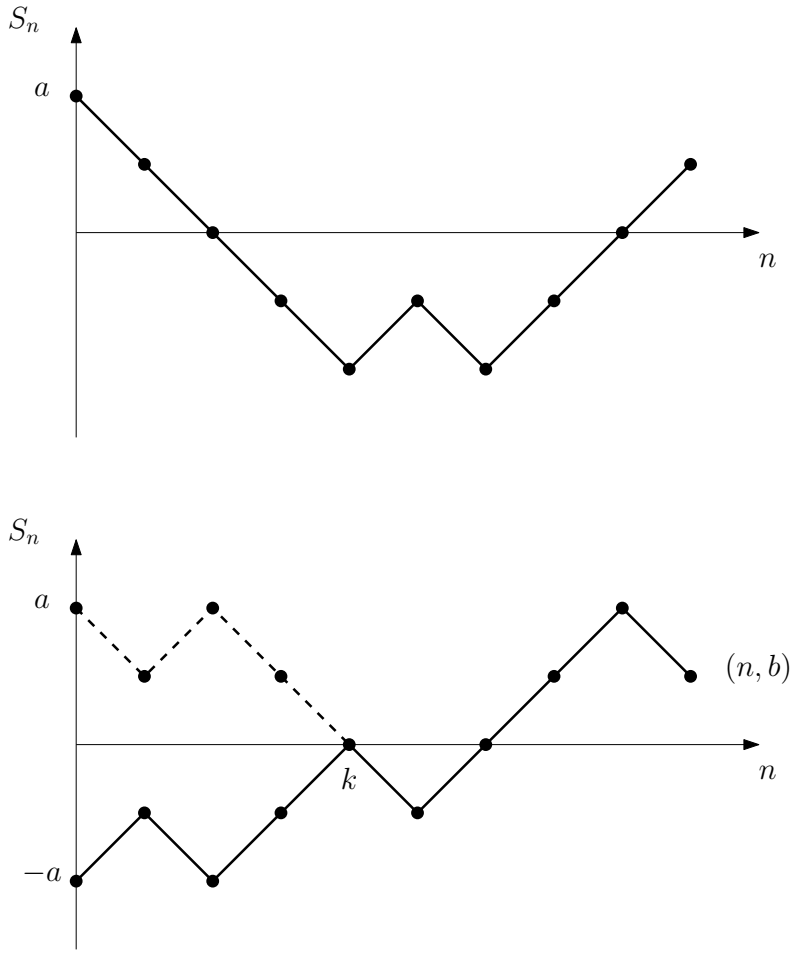


Figure 5.5: The path describing the random walk (above) and the bijection between the paths from $(0, a)$ to (n, b) containing some point $(k, 0)$ and the paths from $(0, -a)$ to (n, b) (below).

Suppose we know that $S_n = b$. The random walk may or may not have visited the origin between times 0 and n. Let $N_n(a, b)$ be the number of possible paths from $(0, a)$ to (n, b) and let $N_n^0(a, b)$ be the number of such paths which contain some point $(k, 0)$ on the x-axis.

Theorem 5.3.1 (Reflection principle). *If $a, b > 0$, then $N_n^0(a, b) = N_n(-a, b)$*

Proof. We provide a bijection between the family of paths from $(0, -a)$ to (n, b) and the family of paths from $(0, a)$ to (n, b) containing some point $(k, 0)$ on the x-axis. Consider a path from $(0, -a)$ to (n, b). It intersects the x-axis for the first time at some point $(k, 0)$. We reflect the subpath with x-coordinates

between 0 and k in the x-axis to obtain a path from $(0,a)$ to (n,b) intersecting the x-axis. This operation provides the desired bijection. □

We can now count the number of paths from $(0,a)$ to (n,b). We use the same idea as in Example 2.1.16.

Lemma 5.3.2.

$$N_n(a,b) = \binom{n}{\frac{n+b-a}{2}}$$

Proof. Consider a path from $(0,a)$ to (n,b) and let α and β be the number of steps up and down, respectively, in this path. Clearly, $\alpha+\beta = n$ and $\alpha-\beta = b-a$, from which $\alpha = \frac{n+b-a}{2}$. The number of such paths is the number of ways of picking α steps up from the n steps made i.e.,

$$N_n(a,b) = \binom{n}{\alpha} = \binom{n}{\frac{n+b-a}{2}}.$$ □

Corollary 5.3.3. *If $b>0$, then the number of paths from $(0,0)$ to (n,b) which do not revisit the x-axis is $\frac{b}{n}\cdot N_n(0,b)$.*

Proof. Since $b>0$, the first step of all such paths is $(1,1)$ and so the number of such paths is

$$N_{n-1}(1,b) - N^0_{n-1}(1,b) = N_{n-1}(1,b) - N_{n-1}(-1,b)$$

by the Reflection principle. Using Lemma 5.3.2, we then obtain

$$N_{n-1}(1,b) - N_{n-1}(-1,b) = \binom{n-1}{\frac{n+b}{2}-1} - \binom{n-1}{\frac{n+b}{2}} = \frac{b}{n}\binom{n}{\frac{n+b}{2}} = \frac{b}{n}\cdot N_n(0,b).$$ □

Back to our ballot problem. The probability that A is always ahead in the count is simply the proportion of paths from $(0,0)$ to $(a+b,a-b)$ which do not revisit the x-axis. By Corollary 5.3.3, this proportion is $\frac{a-b}{a+b}$.

We will now see some other interesting consequences of the reflection principle. Suppose that $S_0=0$. What is the probability that the walk does not revisit its starting point 0 in the first n steps?

Proposition 5.3.4. *If $S_0=0$, then*

$$\mathbb{P}(S_1\neq 0,\dots,S_n\neq 0) = \frac{\mathbb{E}(|S_n|)}{n}.$$

Proof. By countable additivity,

$$\mathbb{P}(S_1\neq 0,\dots,S_n\neq 0) = \sum_{b\in\mathbb{Z}} \mathbb{P}(S_1\neq 0,\dots,S_n\neq 0, S_n=b).$$

Let us compute $\mathbb{P}(S_1\neq 0,\dots,S_n\neq 0,S_n=b)$ when $b>0$. The event in question occurs if and only if the path of the random walk does not visit the x-axis in the time interval $[1,n]$. By Corollary 5.3.3, the number of such paths is $\frac{b}{n}\cdot N_n(0,b)$ and each such path has $(n+b)/2$ steps up and $(n-b)/2$ steps down. Therefore,

$$\mathbb{P}(S_1\neq 0,\dots,S_n\neq 0,S_n=b) = \frac{b}{n}\cdot N_n(0,b)\cdot p^{(n+b)/2}q^{(n-b)/2} = \frac{b}{n}\cdot\mathbb{P}(S_n=b).$$

When $b < 0$, a reflection in the x-axis reduces the problem to the previous case and we obtain

$$\mathbb{P}(S_1 \neq 0, \ldots, S_n \neq 0, S_n = b) = \frac{-b}{n} \cdot N_n(0, -b) \cdot p^{(n+b)/2} q^{(n-b)/2} = \frac{|b|}{n} \cdot \mathbb{P}(S_n = b).$$

Therefore,

$$\mathbb{P}(S_1 \neq 0, \ldots, S_n \neq 0) = \sum_{b \in \mathbb{Z}} \frac{|b|}{n} \cdot \mathbb{P}(S_n = b) = \frac{\mathbb{E}(|S_n|)}{n},$$

where the last equality follows from LOTUS. □

Another feature of interest is the maximum value attained by the random walk. Let $M_n = \max\{S_i : 0 \leq i \leq n\}$ be the maximum value up to time n. Clearly, $\mathbb{P}(M_n \geq r, S_n = b) = \mathbb{P}(S_n = b)$ if $b \geq r$. Therefore, the nontrivial case is when $b < r$:

Proposition 5.3.5. *Suppose that $S_0 = 0$. Then, for any $r \geq 1$ and $b < r$,*

$$\mathbb{P}(M_n \geq r, S_n = b) = \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r - b).$$

Proof. Let $N_n^r(0, b)$ be the number of paths from $(0, 0)$ to (n, b) which contain some point at height r i.e., a point of the form (i, r) for some $0 < i < n$. For any such path π, let (i_π, r) be the earliest such point. We reflect the subpath traversed in the time interval $[i_\pi, n]$ in the line $y = r$. We thus obtain a path π' from $(0, 0)$ to $(n, 2r - b)$ (see Figure 5.6). Any such path π' is obtained from a unique path π and so $N_n^r(0, b) = N_n(0, 2r - b)$. Therefore,

$$\begin{aligned}
\mathbb{P}(M_n \geq r, S_n = b) &= N_n^r(0, b) p^{(n+b)/2} q^{(n-b)/2} \\
&= \left(\frac{q}{p}\right)^{r-b} N_n(0, 2r - b) p^{(n+2r-b)/2} q^{(n-2r+b)/2} \\
&= \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r - b).
\end{aligned}$$

□

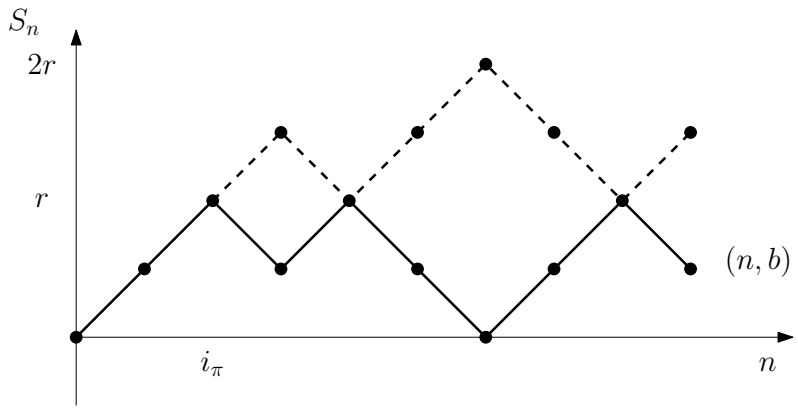


Figure 5.6: Bijection between paths from $(0, 0)$ to (n, b) containing some point at height r and paths from $(0, 0)$ to $(n, 2r - b)$.

Proposition 5.3.5 allows us to obtain an expression for $\mathbb{P}(M_n \geq r)$ when $r \geq 1$. Indeed, by countable

additivity,

$$\begin{aligned}\mathbb{P}(M_n \geq r) = \sum_{b\in\mathbb{Z}} \mathbb{P}(M_n \geq r, S_n = b) &= \sum_{b=r}^{\infty} \mathbb{P}(M_n \geq r, S_n = b) + \sum_{b=-\infty}^{r-1} \mathbb{P}(M_n \geq r, S_n = b) \\ &= \sum_{b=r}^{\infty} \mathbb{P}(S_n = b) + \sum_{b=-\infty}^{r-1} \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r - b) \\ &= \mathbb{P}(S_n = r) + \sum_{c=r+1}^{\infty} \mathbb{P}(S_n = c) + \sum_{c=r+1}^{\infty} \left(\frac{q}{p}\right)^{c-r} \mathbb{P}(S_n = c) \\ &= \mathbb{P}(S_n = r) + \sum_{c=r+1}^{\infty} \left(1 + \left(\frac{q}{p}\right)^{c-r}\right) \mathbb{P}(S_n = c),\end{aligned}$$

which in the symmetric case reduces to

$$\mathbb{P}(M_n \geq r) = \mathbb{P}(S_n = r) + 2 \sum_{c=r+1}^{\infty} \mathbb{P}(S_n = c) = \mathbb{P}(S_n = r) + 2\mathbb{P}(S_n \geq r + 1).$$

With the aid of Proposition 5.3.5 we can also compute first-passage probabilities for a random walk starting in 0:

Corollary 5.3.6. *For each $n \geq 1$,*

$$f_{0,b}^{(n)} = \frac{|b|}{n} \cdot \mathbb{P}(S_n = b).$$

Proof. We suppose that $b > 0$, the case $b < 0$ being similar. Observe that

$$\begin{aligned} f_{0,b}^{(n)} &= \mathbb{P}(M_{n-1} = S_{n-1} = b - 1, S_n = b) \\ &= p \cdot \mathbb{P}(M_{n-1} = S_{n-1} = b - 1) \\ &= p \cdot (\mathbb{P}(M_{n-1} \geq b - 1, S_{n-1} = b - 1) - \mathbb{P}(M_{n-1} \geq b, S_{n-1} = b - 1)) \\ &= p \cdot (\mathbb{P}(S_{n-1} = b - 1) - \mathbb{P}(M_{n-1} \geq b, S_{n-1} = b - 1)) \\ &= p \cdot \left(\mathbb{P}(S_{n-1} = b - 1) - \frac{q}{p} \cdot \mathbb{P}(S_{n-1} = b + 1)\right) \\ &= \frac{b}{n} \cdot \mathbb{P}(S_n = b), \end{aligned}$$

where the first equality follows from the definition of $f_{0,b}^{(n)}$, the second follows from the independence of the X_i's, the third follows from finite additivity, the fifth follows from Proposition 5.3.5 and the last is left as an exercise. □

5.4 Martingale convergence theorem

In this section we establish the almost sure convergence of martingales under some mild conditions. This goes under the name of Martingale convergence theorem. Recall the elementary analysis fact that every convergent sequence of real numbers is bounded, but a bounded sequence is not necessarily convergent. It turns out that martingales are much better behaved:

Theorem 5.4.1 (Martingale convergence theorem). *If $\{S_n\}_{n\geq 1}$ is a martingale such that $\mathbb{E}(S_n^2) < M$ for some $M \in \mathbb{R}$ and all $n \geq 1$, then there exists a random variable S such that $S_n \xrightarrow{a.s.} S$.*

Before turning to the proof, we need some auxiliary results. By definition, if $\{S_n\}_{n\geq 1}$ is a martingale with respect to $\{X_n\}_{n\geq 1}$, then $\mathbb{E}(S_{m+1}|X_{1,m}) = S_m$. The following generalized property holds:

Lemma 5.4.2. *Let $\{S_n\}_{n\geq 1}$ be a martingale with respect to $\{X_n\}_{n\geq 1}$. Then $\mathbb{E}(S_{m+n}|X_{1,m}) = S_m$ for each $n, m \geq 1$.*

Proof. We have seen in Exercise 5.0.16 that $\mathbb{E}(\mathbb{E}(X|Y,W)|Y) = \mathbb{E}(X|Y)$. It is not difficult to see that the same holds if we replace Y and W with families of random variables. We then obtain,

$$\mathbb{E}(S_{m+n}|X_{1,m}) = \mathbb{E}(\mathbb{E}(S_{m+n}|X_{1,m}, X_{m+1,m+n-1})|X_{1,m}) = \mathbb{E}(S_{m+n-1}|X_{1,m})$$

and iterating gives the desired equality. □

We now establish a technical but useful result, which is an analogue of Markov's and Chebyshev's inequalities. Recall that Markov's inequality asserts that, for any random variable X and $\varepsilon > 0$,

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}(|X|)}{\varepsilon}.$$

Martingales satisfy a similar, but much more powerful inequality, which bounds the *maximum* of the process.

Theorem 5.4.3 (Doob-Kolmogorov inequality). *Let $\{S_n\}_{n\geq 1}$ be a martingale with respect to $\{X_n\}_{n\geq 1}$. Then, for each $\varepsilon > 0$,*

$$\mathbb{P}(\max_{1\leq i\leq n} |S_i| \geq \varepsilon) \leq \frac{\mathbb{E}(S_n^2)}{\varepsilon^2}.$$

Proof. Let $A_n = \{\omega : |S_i(\omega)| < \varepsilon \text{ for each } i \leq n\}$ and let $B_n = A_{n-1} \cap \{\omega : |S_n(\omega)| \geq \varepsilon\}$. In other words, A_n is the event that none of the first n random variables S_i's deviates from 0 by at least ε, whereas B_n is the event that the first such deviation occurs at the n-th random variable S_n. Clearly, we can write Ω as the following union of pariwise disjoint events:

$$\Omega = A_n \cup \Big(\bigcup_{i=1}^{n} B_i\Big).$$

But then we have a partition of Ω and so, by Corollary 1.8.3,

$$\mathbb{E}(S_n^2) = \mathbb{E}(S_n^2|A_n)\cdot\mathbb{P}(A_n) + \sum_{i=1}^{n}\mathbb{E}(S_n^2|B_i)\cdot\mathbb{P}(B_i) = \mathbb{E}(S_n^2 I_{A_n}) + \sum_{i=1}^{n}\mathbb{E}(S_n^2 I_{B_i}) \geq \sum_{i=1}^{n}\mathbb{E}(S_n^2 I_{B_i}).$$

For each term in the sum, we have

$$\mathbb{E}(S_n^2 I_{B_i}) = \mathbb{E}((S_n + S_i - S_i)^2 I_{B_i}) = \mathbb{E}((S_n - S_i)^2 I_{B_i}) + 2\mathbb{E}((S_n - S_i)S_i I_{B_i}) + \mathbb{E}(S_i^2 I_{B_i}).$$

Clearly, $\mathbb{E}((S_n + S_i - S_i)^2 I_{B_i}) \geq 0$. Consider now $\mathbb{E}(S_i^2 I_{B_i})$. If B_i occurs, then $|S_i| \geq \varepsilon$ and so

$$\mathbb{E}(S_i^2 I_{B_i}) = \mathbb{E}(S_i^2|B_i)\cdot\mathbb{P}(B_i) \geq \varepsilon^2\cdot\mathbb{P}(B_i).$$

It remains to consider the middle term:

$$\mathbb{E}((S_n - S_i)S_i I_{B_i}) = \mathbb{E}(\mathbb{E}((S_n - S_i)S_i I_{B_i})|X_{1,i}) = \mathbb{E}(S_i I_{B_i}\mathbb{E}(S_n - S_i|X_{1,i})) = \mathbb{E}(S_i I_{B_i}(S_i - S_i)) = 0,$$

where the first equality follows from Theorem 5.0.7(v), the second follows from Theorem 5.0.7(vi) and the third follows from Lemma 5.4.2. Combining the lower bounds obtained so far, $\mathbb{E}(S_n^2 I_{B_i}) \geq \varepsilon^2 \cdot \mathbb{P}(B_i)$, from which

$$\mathbb{E}(S_n^2) \geq \sum_{i=1}^{n} \varepsilon^2 \cdot \mathbb{P}(B_i) \geq \varepsilon^2 \cdot \mathbb{P}(\max_{1 \leq i \leq n} |S_i| \geq \varepsilon),$$

where the last inequality follows from monotonicity and the fact that $\{\omega : \max_{1 \leq i \leq n} |S_i(\omega)| \geq \varepsilon\}$ is contained in $\bigcup_{i=1}^{n} B_i$. Indeed, if the maximum deviates by at least ε, then at least one of the S_i's must deviate by at least ε. □

We can finally prove the Martingale convergence theorem. The proof is split into four parts:

1. $\{\mathbb{E}(S_n^2)\}$ is a non-decreasing real sequence.
2. Find an expression for the event of non-convergence (the non-convergence set) and evaluate its probability.
3. For fixed m, let $Y_n = S_{m+n} - S_m$. Then $\{Y_n\}_{n \geq 1}$ is a martingale with respect to itself.
4. The probability of the non-convergence set is zero.

Some comments are in order. Recall that if we want to show that a sequence of real numbers converges but we do not have a plausible candidate for the limit, it is convenient to show that the sequence is Cauchy convergent. Indeed, a sequence is convergent if and only if it is Cauchy convergent. Let us recall the definition. A sequence of real numbers $\{x_n\}$ is **Cauchy convergent** if, for each $\varepsilon > 0$, there exists N such that $|x_m - x_n| < \varepsilon$, for each $m, n \geq N$. In the case of sequences of random variables, this notion translates as follows:

Definition 5.4.4 (Cauchy convergence for sequences of random variables). A sequence of random variables $\{X_n\}$ is Cauchy convergent if, for each $\varepsilon > 0$, there exists N such that

$$\mathbb{P}(\{\omega : |X_m(\omega) - X_n(\omega)| < \varepsilon, \text{ for each } m, n \geq N\}) = 1.$$

Recall now that $X_n \xrightarrow{\text{a.s.}} X$ for some X means that there exists a random variable X such that

$$\mathbb{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1.$$

Therefore, if we manage to show that $\{X_n\}$ is Cauchy convergent in the sense of Definition 5.4.4, we would then obtain that $X_n \xrightarrow{\text{a.s.}} X$ for some X. This is indeed what will be done for our martingale $\{S_n\}$. We define the **convergence set** C (which is the complement of the non-convergence set mentioned in 2.) as $C = \{\omega : \{S_n(\omega)\} \text{ is Cauchy convergent}\}$ and we will show that $\mathbb{P}(C^c) = 0$, thus implying $\mathbb{P}(C) = 1$.

Proof of Theorem 5.4.1. We follow the four steps highlighted above:
1.

$$\mathbb{E}(S_{m+n}^2) = \mathbb{E}((S_{m+n} + S_m - S_m)^2) = \mathbb{E}(S_m^2) + 2\mathbb{E}(S_m(S_{m+n} - S_m)) + \mathbb{E}((S_{m+n} - S_m)^2).$$

Let us look at the middle term:

$$\mathbb{E}(S_m(S_{m+n} - S_m)) = \mathbb{E}(\mathbb{E}(S_m(S_{m+n} - S_m)|S_{1,m})) = \mathbb{E}(S_m \mathbb{E}(S_{m+n} - S_m|S_{1,m})) = \mathbb{E}(S_m(S_m - S_m)) = 0,$$

where the first equality follows from Theorem 5.0.7(v), the second follows from Theorem 5.0.7(vi) and the third follows from linearity and Lemma 5.4.2. But then

$$\mathbb{E}(S_{m+n}^2) = \mathbb{E}(S_m^2) + \mathbb{E}((S_{m+n} - S_m)^2) \geq \mathbb{E}(S_m^2) \tag{5.11}$$

and so $\{\mathbb{E}(S_n^2)\}$ is a nondecreasing real sequence.

2.

$$\begin{aligned} C &= \{\omega : \{S_n(\omega)\} \text{ is Cauchy convergent}\} \\ &= \{\omega : \text{ for each } \varepsilon > 0, \text{ there exists an } m \text{ such that } |S_{m+i}(\omega) - S_{m+j}(\omega)| < \varepsilon, \text{ for each } i, j \geq 1\}. \end{aligned}$$

On the other hand, by the triangle inequality,

$$|S_{m+i} - S_{m+j}| = |(S_{m+i} - S_m) + (S_m - S_{m+j})| \leq |S_{m+i} - S_m| + |S_{m+j} - S_m|$$

and so the last even above can be rewritten as

$$\begin{aligned} C &= \{\omega : \text{ for each } \varepsilon > 0, \text{ there exists an } m \text{ such that } |S_{m+i}(\omega) - S_m(\omega)| < \varepsilon \text{ for each } i \geq 1\} \\ &= \bigcap_{\varepsilon > 0} \bigcup_m \{\omega : |S_{m+i}(\omega) - S_m(\omega)| < \varepsilon \text{ for each } i \geq 1\}. \end{aligned}$$

Notice that the intersection above is over all positive $\varepsilon \in \mathbb{Q}$, hence countable. Here we are using the fact that between any two real numbers there is a rational number. Alternatively, if the reader feels more comfortable, we could simply take ε of the form $1/n$. We now pass to the complement

$$C^c = \bigcup_{\varepsilon > 0} \bigcap_m \{\omega : |S_{m+i}(\omega) - S_m(\omega)| \geq \varepsilon \text{ for some } i \geq 1\} = \bigcup_{\varepsilon > 0} \bigcap_m A_m(\varepsilon),$$

where $A_m(\varepsilon) = \{\omega : |S_{m+i}(\omega) - S_m(\omega)| \geq \varepsilon \text{ for some } i \geq 1\}$. Clearly, if $\varepsilon \geq \varepsilon'$, then $A_m(\varepsilon) \subseteq A_m(\varepsilon')$ and so, by continuity of probability,

$$\mathbb{P}(C^c) = \mathbb{P}\Big(\bigcup_{\varepsilon > 0} \bigcap_m A_m(\varepsilon)\Big) = \lim_{\varepsilon \to 0} \mathbb{P}\Big(\bigcap_m A_m(\varepsilon)\Big) \leq \lim_{\varepsilon \to 0} \lim_{m \to \infty} \mathbb{P}(A_m(\varepsilon)). \tag{5.12}$$

3. For fixed m, let $Y_n = S_{m+n} - S_m$. We show that $\{Y_n\}_{n \geq 1}$ is a martingale with respect to itself:

$$\begin{aligned} \mathbb{E}(Y_{n+1}|Y_{1,n}) &= \mathbb{E}(S_{m+n+1} - S_m | S_{m,m+n}) \\ &= \mathbb{E}(\mathbb{E}(S_{m+n+1} - S_m | S_{1,m-1}, S_{m,n+m}) | S_{m,m+n}) \\ &= \mathbb{E}(\mathbb{E}(S_{m+n+1} | S_{1,n+m}) - \mathbb{E}(S_m | S_{1,n+m}) | S_{m,m+n}) \\ &= \mathbb{E}(S_{m+n} - S_m | S_{m,m+n}) \\ &= S_{m+n} - S_m = Y_n, \end{aligned}$$

where the first equality follows from the definition of $\{Y_n\}$, the second follows from the fact that $\mathbb{E}(\mathbb{E}(X|Y,W)|Y) = \mathbb{E}(X|Y)$ (see Example 5.0.15 and the comment in the proof of Lemma 5.4.2), the third follows from linearity, the forth follows from the fact that $\{S_n\}$ is a martingale and from Theorem 5.0.7(vi) and the fifth follows again from Theorem 5.0.7(vi).

4. We apply the Doob-Kolmogorov inequality to $\{Y_n\}$. For each $\varepsilon > 0$,

$$\mathbb{P}(\max_{1 \leq i \leq n} |S_{m+i} - S_m| \geq \varepsilon) \leq \frac{\mathbb{E}((S_{m+n} - S_m)^2)}{\varepsilon^2}.$$

Therefore,

$$\begin{aligned} \mathbb{P}(|S_{m+i} - S_m| \geq \varepsilon \text{ for some } 1 \leq i \leq n) &\leq \mathbb{P}(\max_{1 \leq i \leq n} |S_{m+i} - S_m| \geq \varepsilon) \\ &\leq \frac{\mathbb{E}((S_{m+n} - S_m)^2)}{\varepsilon^2} \\ &= \frac{\mathbb{E}(S_{m+n}^2) - \mathbb{E}(S_m^2)}{\varepsilon^2} \\ &\leq \frac{\sup_m \{\mathbb{E}(S_m^2)\} - \mathbb{E}(S_m^2)}{\varepsilon^2}, \end{aligned}$$

where the first inequality follows from monotonicity and the equality follows from Equation (5.11). Letting $n \to \infty$ in the above, we obtain

$$\mathbb{P}(A_m(\varepsilon)) \leq \frac{\sup_m\{\mathbb{E}(S_m^2)\} - \mathbb{E}(S_m^2)}{\varepsilon^2}.$$

But in 1. we have shown that $\{\mathbb{E}(S_m^2)\}$ is a nondecreasing real sequence which is upper bounded by assumption and hence convergent to its sup. This implies that, for fixed ε, as $m \to \infty$, we have that $\mathbb{P}(A_m(\varepsilon)) \to 0$. But then, by Equation (5.12), $\mathbb{P}(C^c) = 0$, as claimed. This concludes the proof. □

Example 5.4.5 (Pólya's urn again). Recall Example 5.0.13. We start at time 2 with one black ball and one white ball in an urn. At each discrete time, we randomly take out a ball from the urn and we return it to the urn together with a new ball of the same color. Letting X_n denote the number of white balls at time n and $M_n = X_n/n$ the fraction of white balls at time n, we verified that $\{M_n\}_{n\geq 2}$ is a martingale with respect to $\{X_n\}_{n\geq 2}$. Since $M_n = X_n/n \leq 1$, we have that $M_n^2 \leq 1$ and so $\mathbb{E}(M_n^2) \leq 1$. Therefore, the martingale convergence theorem implies that $\{M_n\}$ converges almost surely. This means that the proportion of white balls does not fluctuate between 0 and 1 infinitely often. It can be proved that $\{M_n\}$ converges almost surely to a uniform random variable on $[0, 1]$.

Example 5.4.6 (Random harmonic series). It is known from analysis that the harmonic series $\sum_{j=1}^{\infty} 1/j$ diverges while the alternating harmonic series $\sum_{j=1}^{\infty}(-1)^{j+1}/j$ converges. What about choosing pluses and minuses at random? Let $X_1, X_2, \ldots$ be i.i.d random variables with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Let $M_0 = 0$ and for $n > 0$,

$$M_n = \sum_{j=1}^{n} \frac{1}{j} X_j.$$

Since $\mathbb{E}(M_n) = 0$, the same computation as in Example 5.0.10 shows that $\{M_n\}$ is a martingale. Moreover,

$$\mathbb{E}(M_n^2) = \operatorname{var}(M_n) = \sum_{j=1}^{n} \operatorname{var}\left(\frac{1}{j} X_j\right) = \sum_{j=1}^{n} \frac{1}{j^2} \leq \sum_{j=1}^{\infty} \frac{1}{j^2}.$$

Since the latter series converges, the second moments are bounded and so $\{M_n\}$ converges almost surely.

5.5 Martingale concentration inequalities

We conclude our chapter on martingales with a brief overview of concentration inequalities. The interested reader should refer to [5] for a more in-depth coverage.

It is often useful to bound the probability that a random variable deviates from some other value, usually its mean. Recall that Chebyshev's inequality tells us that

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\operatorname{var}(X)}{t^2}.$$

Therefore, if $t \gg \operatorname{var}(X)$, the probability of deviating by more than t from $\mathbb{E}(X)$ is small. However, it is often desirable that the probability of large deviations is *very* small i.e., that X is *concentrated* around its mean. Concentration inequalities are an essential tool in the probabilistic analysis of algorithms and in the study of randomized algorithms. Our interest will be in concentration inequalities in which the deviation probabilities decay exponentially. The most basic such inequality is the Azuma-Hoeffding inequality for bounded martingales:

Theorem 5.5.1 (Azuma-Hoeffding inequality). *Let $\{Y_n\}_{n\geq 0}$ be a martingale with respect to $\{X_n\}_{n\geq 0}$ such that, for each $i \geq 1$,*

$$|Y_i - Y_{i-1}| \leq d_i,$$

for some real numbers d_i. Then, for all $n \geq 0$ and $t > 0$,

$$\mathbb{P}(|Y_n - Y_0| \geq t) \leq 2e^{-2t^2/(\sum_{i=1}^n d_i^2)}.$$

Example 5.5.2 (Divergence of random walk). Consider the simple symmetric random walk $\{S_n\}_{n\geq 0}$ on $\mathbb{Z}$ with starting point S_0. S_n denotes the position at time n. What is the likelihood of the random walk diverging far from its starting point? Notice that $|S_i - S_{i-1}| \leq 1$ for each $i \geq 1$ and so, by the Azuma-Hoeffding inequality,

$$\mathbb{P}(|S_n - S_0| \geq t) \leq 2e^{-2t^2/n}.$$

This implies that the random walk, in the first n steps, is likely to stay within an interval of radius $\sqrt{n}$ around the starting point. Indeed,

$$\mathbb{P}(|S_n - S_0| \geq \sqrt{n}) \leq 2e^{-2} < 0.28.$$

Example 5.5.3 (Pattern matching). In many scenarios, including examining DNA structure, a goal is to find "interesting" patterns in a sequence of characters, where "interesting" refers to strings that occur more often than one would expect if the characters were generated randomly. This notion of "interesting" is reasonable if the number of occurrences of a string is concentrated around its mean in the random model. We now obtain a concentration result for this setting.

Let $X = (X_1, \ldots, X_n)$ be a sequence of characters chosen independently and uniformly at random from an alphabet Σ, where $s = |\Sigma|$. Let $B = (b_1, \ldots, b_k)$ be a fixed string of k characters from Σ and let F be the number of occurrences of the fixed string B in the random string X. What is $\mathbb{E}(F)$? Let A_i be the event that B occurs in X starting at position i. By linearity of expectation,

$$\mathbb{E}(F) = \sum_{i=1}^{n-k+1} \mathbb{E}(I_{A_i}) = \sum_{i=1}^{n-k+1} 1 \cdot \mathbb{P}(A_i) = \sum_{i=1}^{n-k+1} \left(\frac{1}{s}\right)^k = (n+k-1)\left(\frac{1}{s}\right)^k.$$

We now use a Doob-type martingale and the Azuma-Hoeffding inequality to show that if k is relatively small with respect to n, then the number of occurrences of B in X is highly concentrated around its mean.

Let $Z_0 = \mathbb{E}(F)$ and, for $1 \leq i \leq n$, let $Z_i = \mathbb{E}(F|X_{1,i})$. A computation similar to that in Example 5.0.15 shows that $Z_0, \ldots, Z_n$ is a martingale with respect to $X_1, \ldots, X_n$. Z_i defines the expected number of occurrences of the pattern in the entire sequence, given only the first i characters. Clearly, $Z_n = F$. Notice that, when a new character X_{i+1} is exposed, it adds at most k new occurrences of B in expectation (from the leftmost one with $b_k = X_{i+1}$ to the rightmost one with $b_1 = X_{i+1}$). Hence $|Z_{i+1} - Z_i| \leq k$. Therefore, by the Azuma-Hoeffding inequality,

$$\mathbb{P}(|F - \mathbb{E}(F)| \geq t) \leq 2e^{-2t^2/(nk^2)}.$$

The most common way Azuma-Hoeffding is applied is by considering Lipschitz functions:

Definition 5.5.4 (Lipschitz property). A function $f(x_1, \ldots, x_n)$ of n variables satisfies the Lipschitz property (or the bounded differences condition) with constants d_i if

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq d_i,$$

whenever the n-dimensional vectors $\mathbf{x}$ and $\mathbf{x}'$ differ in just the i-th coordinate.

In words, the Lipschitz property states that changing the value of any single coordinate can change the value of f by at most a constant. The following result is obtained by applying the Azuma-Hoeffding inequality to Doob's martingales in a way similar to Example 5.5.3.

Corollary 5.5.5 (Method of bounded differences). *If f satisfies the Lipschitz property with constants d_i and $X_1, \ldots, X_n$ are independent random variables, then*

$$\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}(f(X_1, \ldots, X_n)| \geq t) \leq 2e^{-2t^2/(\sum_{i=1}^n d_i^2)}.$$

It is important to remark that the independence requirement in Corollary 5.5.5 is necessary.

Example 5.5.6 (Bin packing problem). We are given n items of sizes in the interval $[0, 1]$ and are required to pack them into the fewest number of unit-capacity bins as possible. This problem is computationally hard: it is a so-called NP-complete problem i.e., a problem which does not admit an "efficient" algorithm unless $\mathsf{P} = \mathsf{NP}$.

In the stochastic version of this problem, the item sizes are independent random variables $X_1, \ldots, X_n$ in the interval $[0, 1]$. Let $B_n = B_n(X_1, \ldots, X_n)$ denote the optimum value i.e., the minimum number of bins that suffice. Since the Lipschitz condition holds for B_n with constants 1 (why?), Corollary 5.5.5 gives the following concentration result:

$$\mathbb{P}(|B_n - \mathbb{E}(B_n)| \geq t) \leq 2e^{-2t^2/n}.$$

Notice that in this case it is not easy to compute $\mathbb{E}(B_n)$. Nevertheless, even if we can't compute the mean, we can still obtain a concentration bound!

Example 5.5.7 (Balls in bins). m balls are thrown independently at random into n bins ($m \geq n$) and we are interested in the number of empty bins. For each $i \in \{1, \ldots, n\}$, let A_i be the event that the i-th bin is empty. Then the random variable $Z = \sum_{i=1}^n I_{A_i}$ counts the number of empty bins. Clearly,

$$\mathbb{E}(Z) = n \cdot \mathbb{P}(A_i) = n\left(1 - \frac{1}{n}\right)^m.$$

For each $k \in \{1, \ldots, m\}$, let X_k be the random variable taking values in $\{1, \ldots, n\}$ and indicating the bin in which ball k lands. We can consider Z as a function $Z(X_1, \ldots, X_m)$. We claim that this function satisfies the Lipschitz property with constants 1. Indeed, if the i-th ball is moved from one bin to another, keeping all other balls where they are, then the number of empty bins can at most either go up by 1 or down by 1. Therefore,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2e^{-2t^2/n}.$$

Bibliography

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2nd edition, 2008.

[2] P. Billingsley. *Probability and Measure*. Wiley, anniversary edition, 2012.

[3] J. T. Chang. Stochastic Processes. http://www.stat.yale.edu/~pollard/Courses/251.spring2013/Handouts/Chang-notes.pdf, 2007.

[4] R. P. Dobrow. *Introduction to Stochastic Processes with R*. Wiley, 2016.

[5] D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[6] R. Durrett. *Essentials of Stochastic Processes*. Springer, 3rd edition, 2016.

[7] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 3rd edition, 2001.

[8] G. F. Lawler. *Introduction to Stochastic Processes*. Chapman & Hall/CRC, 2nd edition, 2006.

[9] M. Mitzenmacher and E. Upfal. *Probability and Computing - Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[10] J. R. Norris. *Markov Chains*. Cambridge University Press, 2009.

[11] R. Serfozo. *Basics of Applied Stochastic Processes*. Springer, 2009.

[12] D. Stirzaker. *Elementary Probability*. Cambridge University Press, 2nd edition, 2003.

[13] S. S. Venkatesh. *The Theory of Probability*. Cambridge University Press, 2013.

[14] J. B. Walsh. *Knowing the Odds - An Introduction to Probability*. AMS, 2012.