Ramon van Handel

Probability and Random Processes

ORF 309/MAT 380 Lecture Notes
Princeton University

This version: February 22, 2016

Preface

These lecture notes are intended for a one-semester undergraduate course in applied probability. Such a course has been taught at Princeton for many years by Erhan Çinlar. The choice of material in these notes was greatly inspired by Çinlar's course, though my own biases regarding the material and presentation are inevitably reflected in the present incarnation.

As always, some choices had to be made regarding what to present:

- It should be emphasized that this course is *not* intended for a pure mathematics audience, for whom an entirely different approach would be indicated. The course is taken by a diverse range of undergraduates in the sciences, engineering, and applied mathematics. For this reason, the focus is on probabilistic intuition rather than rigorous proofs, and the choice of material emphasizes exact computations rather than inequalities or asymptotics. The main aim is to introduce an applied audience to a range of basic probabilistic notions and to quantitative probabilistic reasoning.
- A principle I have tried to follow as much as possible is not to introduce any concept out of the blue, but rather to have a natural progression of topics. For example, every new distribution that is encountered is derived naturally from a probabilistic model, rather than being defined abstractly. My hope is that this helps students develop a feeling for the big picture and for the connections between the different topics.
- The range of topics is quite large for a first course on probability, and the pace is rapid. The main missing topic is an introduction to martingales; I hope to add a chapter on this at the end at some point in the future.

It is a fact of life that lecture notes are a perpetual construction zone. Surely errors remain to be fixed and presentation remains to be improved. Many thanks are due to all students who provided me with corrections in the past, and I will be grateful to continue to receive such feedback in the future.

Princeton,
January 2016

Contents

0

Introduction

0.1 What is probability?

Most simply stated, probability is the study of randomness. Randomness is of course everywhere around us—this statement surely needs no justification! One of the remarkable aspects of this subject is that it touches almost every area of the natural sciences, engineering, social sciences, and even pure mathematics. The following random examples are only a drop in the bucket.

- Physics: quantities such as temperature and pressure arise as a direct consequence of the random motion of atoms and molecules. Quantum mechanics tells us that the world is random at an even more basic level.
- Biology and medicine: random mutations are the key driving force behind evolution, which has led to the amazing diversity of life that we see today. Random models are essential in understanding the spread of disease, both in a population (epidemics) or in the human body (cancer).
- Chemistry: chemical reactions happen when molecules randomly meet. Random models of chemical kinetics are particularly important in systems with very low concentrations, such as biochemical reactions in a single cell.
- Electrical engineering: noise is the universal bane of accurate transmission of information. The effect of random noise must be well understood in order to design reliable communication protocols that you use on a daily basis in your cell phones. The modelling of data, such as English text, using random models is a key ingredient in many data compression schemes.
- Computer science: randomness is an important resource in the design of algorithms. In many situations, randomized algorithms provide the best known methods to solve hard problems.
- Civil engineering: the design of buildings and structures that can reliably withstand unpredictable effects, such as vibrations, variable rainfall and wind, etc., requires one to take randomness into account.

- Finance and economics: stock and bond prices are inherently unpredictable; as such, random models form the basis for almost all work in the financial industry. The modelling of randomly occurring rare events forms the basis for all insurance policies, and for risk management in banks.

- Sociology: random models provide basic understanding of the formation of social networks and of the nature of voting schemes, and form the basis for principled methodology for surveys and other data collection methods.

- Statistics and machine learning: random models form the foundation for almost all of data science. The random nature of data must be well understood in order to draw reliable conclusions from large data sets.

- Pure mathematics: probability theory is a mathematical field in its own right, but is also widely used in many problems throughout pure mathematics in areas such as combinatorics, analysis, and number theory.

- ... (insert your favorite subject here)

As a probabilist[1], I find it fascinating that the same basic principles lie at the heart of such a diverse list of interesting phenomena: probability theory is the foundation that ties all these and innumerable other areas together. This should already be enough motivation in its own right to convince you (in case you were not already convinced) that we are on to an exciting topic.

Before we can have a meaningful discussion, we should at least have a basic idea of what randomness means. Let us first consider the opposite notion. Suppose I throw a ball many times at exactly the same angle and speed and under exactly the same conditions. Every time we run this experiment, the ball will land in exactly the same place: we can predict exactly what is going to happen. This is an example of a *deterministic* system. *Randomness* is the opposite of determinism: a random phenomenon is one that can yield different outcomes in repeated experiments, even if we use exactly the same conditions in each experiment. For example, if we flip a coin, we know in advance that it will either come up heads or tails, but we cannot predict before any given experiment which of these outcomes will occur. Our challenge is to develop a framework to reason precisely about random phenomena.

0.2 Why do we need a mathematical theory?

It is not at all obvious at first sight that it is possible to develop a rigorous theory of probability: how can one make precise predictions about a phenomenon whose behavior is inherently unpredictable? This philosophical hurdle hampered the development of probability theory for many centuries.

[1] Official definition from the Oxford English Dictionary: "probabilist, *n.* An expert or specialist in the mathematical theory of probability."

To illustrate the pitfalls of an intuitive approach to probability, let us consider a seemingly plausible definition. You probably think of the probability that an event E happens as the fraction of outcomes in which E occurs (this is not entirely unreasonable). We could posit this as a tentative definition

$$\text{Probability of } E = \frac{\text{Number of outcomes where } E \text{ occurs}}{\text{Number of all possible outcomes}}.$$

This sort of intuitive definition may look at first sight like it matches your experience. However, it is totally meaningless: we can easily use it to come to entirely different conclusions.

Example 0.2.1. Suppose that we flip two coins. What is the probability that we obtain one heads (H) and one tails (T)?

- Solution 1: The possible outcomes are HH, HT, TH, TT. The outcomes where we have one heads and one tails are HT, TH. Hence,

$$\text{Probability of one heads and one tails} = \frac{2}{4} = \frac{1}{2}.$$

- Solution 2: The possible outcomes are *two heads, one heads and one tails, two tails*. Only one of these outcomes has one heads and one tails. Hence,

$$\text{Probability of one heads and one tails} = \frac{1}{3}.$$

Now, you may come up with various objections to one or the other of these solutions. But the fact of the matter is that both of these solutions are perfectly reasonable interpretations of the "intuitive" attempt at a definition of probability given above. (While our modern understanding of probability corresponds to Solution 1, the eminent mathematician and physicist d'Alembert forcefully argued for Solution 2 in the 1750s in his famous encyclopedia). We therefore immediately see that an intuitive approach to probability is not adequate. In order to reason reliably about random phenomena, it is essential to develop a rigorous mathematical foundation that leaves no room for ambiguous interpretation. This is the goal of probability theory:

Probability theory is the mathematical study of random phenomena.

It took many centuries to develop such a theory. The first steps in this direction have their origin in a popular pastime of the 17th century: gambling (I suppose it is still popular). A French writer, Chevalier de Méré, wanted to know how to bet in the following game. A pair of dice is thrown 24 times; should one bet on the occurence of at least one double six? An intuitive computation led him to believe that betting on this outcome is favorable, but repeated "experiments" led him to the opposite conclusion. De Méré decided to consult his friend, the famous mathematician Blaise Pascal, who started corresponding about this problem with another famous mathematician, Pierre de Fermat.

This correspondence marked the first serious attempt at understanding probabilities mathematically, and led to important works by Christiaan Huygens, Jacob Bernoulli, Abraham de Moivre, and Pierre-Simon de Laplace in the next two centuries. It was only in 1933, however, that a truly satisfactory mathematical foundation to probability theory was developed by the eminent Russian mathematician Andrey Kolmogorov. With this solid foundation in place, the door was finally open to the systematic development of probability theory and its applications. It is Kolmogorov's theory that is used universally today, and this will also be the starting point for our course.

0.3 This course

In the following chapter, we are going to develop the basic mathematical principles of probability. This solid mathematical foundation will allow us to systematically build ever more complex random models, and to analyze the behavior of such models, without running any risk of the type of ambiguous conclusions that we saw in the example above. With precision comes necessarily a bit of abstraction, but this is nothing to worry about: the basic principles of probability are little more than "common sense" properly formulated in mathematical language. In the end, the success of Kolmogorov's theory is due to the fact that it genuinely captures our real-world observations about randomness.

Once we are comfortable with the basic framework of probability theory, we will start developing increasingly sophisticated models of random phenomena. We will pay particular attention to models of *random processes* where the randomness develops over time. The notion of time is intimately related with randomness: one can argue that the future is random, but the past is not. Indeed, we already know what happened in the past, and thus it is perfectly predictable; on the other hand, we typically cannot predict what will happen in the future, and thus the future is random. While this idea might seem somewhat philosophical now, it will lead us to notions such as random walks, branching processes, Poisson processes, Brownian motion, and Markov chains, which form the basis for many complex models that are used in numerous applications. At the end of the course, you might want to look back at the humble point at which we started. I hope you will find yourself convinced that a mathematical theory of probability is worth the effort.

. . .

This course is aimed at a broad audience and is not a theorem-proof style course.[2] That does not mean, however, that this course does not require rigorous thinking. The goal of this course is to teach you how to reason precisely about randomness and, most importantly of all, how to *think probabilistically*.

[2] Students seeking a mathematician's approach to probability should take ORF 526.

1

Basic Principles of Probability

The goal of this chapter is to introduce the basic ingredients of a mathematical theory of probability that will form the basis for all further developments. As was emphasized in the introduction, these ingredients are little more than "common sense" expressed in mathematical form. You will quickly become comfortable with this basic machinery as we start using it in the sequel.

1.1 Sample space

A *random experiment* is an experiment whose outcome cannot be predicted before the experiment is performed. We do, however, know in advance what outcomes are possible in the experiment. For example, if you flip a coin, you know it will come up either heads or tails; you just do not know which of these outcomes will actually occur in a given experiment.

The first ingredient of any probability model is the specification of all possible outcomes of a random experiment.

Definition 1.1.1. *The* sample space Ω *is the set of all possible outcomes of a random experiment.*

Example 1.1.2 (Two dice). Consider the random experiment of throwing one red die and one blue die. We denote by (i, j) the outcome that the red die comes up i and the blue die comes up j. Hence, we define the sample space

$$\Omega = \{(i, j) \colon 1 \leq i, j \leq 6\}.$$

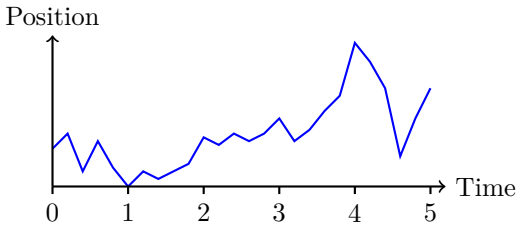
In this experiment, there are only $6^2 = 36$ possible outcomes.

Example 1.1.3 (Waiting for the bus). Consider the random experiment of waiting for a bus that will arrive at a random time in the future. In this case, the outcome of the experiment can be any real number $t \geq 0$ ($t = 0$ means the bus comes immediately, $t = 1.5$ means the bus comes after 1.5 hours, etc.) We can therefore define the sample space

$$\Omega = [0, +\infty[.$$

In this experiment, there are infinitely many possible outcomes.

Example 1.1.4 (Flight of the bumblebee). A bee is buzzing around, and we track its flight trajectory for 5 seconds. What possible outcomes are there in such a random experiment? A flight path of the bee might look something like this:



(Of course, the true position of the bee in three dimensions is a point in $\mathbb{R}^3$; we have plotted one coordinate for illustration). As bees have not yet discovered the secret of teleportation, their flight path cannot have any jumps (it must be continuous), but otherwise they could in principle follow any continuous path. So, the sample space for this experiment can be chosen as

$$\Omega = \{\text{all continuous paths } \omega \colon [0, 5] \to \mathbb{R}^3\}.$$

This is a huge sample space. But this is not a problem: Ω faithfully describes all possible outcomes of this random experiment.

1.2 Events

Once we have defined all possible outcomes of a random experiment, we should discuss what types of questions we can ask about such outcomes. This leads us to the notion of *events*. Informally, *an event is a statement for which we can determine whether it is true or false after the experiment has been performed.* Before we give a formal definition, let us consider some simple examples.

Example 1.2.1 (Two dice). In Example 1.1.2, consider the following event:

"The sum of the numbers on the dice is 7."

Note that this event occurs in a given experiment if and only if the outcome of the experiment happens to lie in the following subset of all possible outcomes:

$$\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \subset \Omega.$$

We cannot predict in advance whether this event will occur, but we can determine whether it has occured once the outcome of the experiment is known.

Example 1.2.2 (Bus). In Example 1.1.3, consider the following event:

"The bus comes within the first hour."

Note that this event occurs in a given experiment if and only if the outcome of the experiment happens to lie in the following subset of all possible outcomes:

$$[0,1] \subset \Omega.$$

Example 1.2.3 (Bumblebee). In Example 1.1.4, suppose there is an object (say, a wall or a chair) that takes up some volume $A \subset \mathbb{R}^3$ of space. We want to know whether or not the bee will hit this object in the first second of its flight. For example, we can consider the following event:

"The bumblebee stays outside the set A in the first second."

Note that this event occurs in a given experiment if and only if the outcome of the experiment happens to lie in the following subset of all possible outcomes:

$$\{\text{continuous paths } \omega : [0,5] \to \mathbb{R}^3 : \omega(t) \notin A \text{ for all } t \in [0,1]\} \subset \Omega.$$

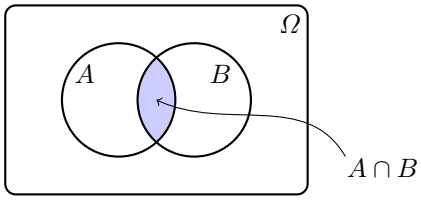
As the above examples show, every event can be naturally identified with the subset of all possible outcomes of the random experiment for which the event is true. Indeed, take a moment to convince yourself that the verbal description of events (such as "the bus comes within an hour") is completely equivalent to the mathematical description as a subset of the sample space (such as $[0,1]$). This observation allows us to give a formal definition.

Definition 1.2.4. *An* event *is a subset A of the sample space Ω.*

The formal definition allows us to translate our common sense reasoning about events into mathematical language.

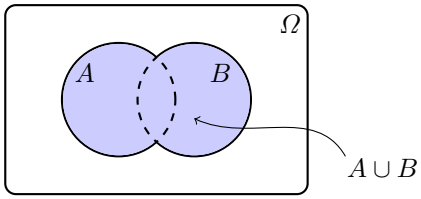
Example 1.2.5 (Combining events). Consider two events A, B.

- The intersection $A \cap B$ is the event that A *and* B occur simultaneously. It might be helpful to draw a picture:



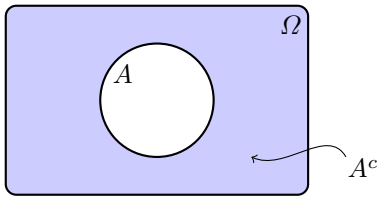
The set $A \subset \Omega$ consists of all outcomes for which the event A occurs, while $B \subset \Omega$ consists of all outcomes for which the event B occurs. Thus $A \cap B$ is the set of all outcomes for which both A *and* B occur.

- The union $A \cup B$ is the event that A *or* B occurs:



(When we say A *or* B, we mean that either A or B or both occur.)

- The complement $A^c := \Omega \backslash A$ is the event that A does *not* occur:



Along the same lines, any common sense combination of events can be translated into mathematical language. For example, the event "A occurs *or* at most one of B, C and D occur" can be written as (why?)

$$A \cup ((B \cap C)^c \cap (B \cap D)^c \cap (C \cap D)^c).$$

After a bit of practice, you will get used to expressing common sense statements in terms of sets. Conversely, when you see such a statement in terms of sets, you should always keep the common sense meaning of the statement in the back of your mind: for example, when you see $A \cap B$, you should automatically read that as "the events A and B occur," rather than the much less helpful "the intersection of sets A and B."

1.3 Probability measure

We have now specified the sample space Ω of all possible outcomes, and the events $A \subseteq \Omega$ about which we can reason. Given these ingredients, how does a random experiment work? Each time we run a random experiment, the goddess of chance Tyche (Τύχη) picks one outcome $\omega \in \Omega$ from the set of all possible outcomes. Once this outcome is revealed to us, we can check for any event $A \subseteq \Omega$ whether or not that event occurred in this realization of the experiment by checking whether or not $\omega \in A$.

Unfortunately, we have no way of predicting which outcome Tyche will pick before conducting the experiment. We therefore also do not know in advance whether or not some event A will occur. To model a random experiment, we will specify for each event A our "degree of confidence" about whether this event will occur. This degree of confidence is specified by assigning a number $0 \leq \mathbf{P}(A) \leq 1$, called a *probability*, to every event A. If $\mathbf{P}(A) = 1$, then we are certain that the event A will occur: in this case A will happen *every* time we perform the experiment. If $\mathbf{P}(A) = 0$, we are certain the event A will not occur: in this case A *never* happens in any experiment. If $\mathbf{P}(A) = 0.7$, say, then the event will occur in some realizations of the experiment and not in others: before we run the experiment, we are 70% confident that the event will happen. What this means in practice is discussed further below.

In order for probabilities to make sense, we cannot assign arbitrary numbers between zero and one to every event: these numbers must obey some rules that encode our common sense about how random experiments work. These rules form the basis on which all of probability theory is built.

Definition 1.3.1. *A* probability measure *is an assignment of a number* $\mathbf{P}(A)$ *to every event* A *such that the following rules are satisfied.*

a. $0 \leq \mathbf{P}(A) \leq 1$ *(probability is a "degree of confidence").*

b. $\mathbf{P}(\Omega) = 1$ *(we are certain that* something *will happen).*

c. If A, B *are events with* $A \cap B = \varnothing$*, then*

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$$

(the probabilities of mutually exclusive events add up).
More generally, if events $E_1, E_2, \ldots$ *satisfy* $E_i \cap E_j = \varnothing$ *for all* $i \neq j$*,*

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(E_i).$$

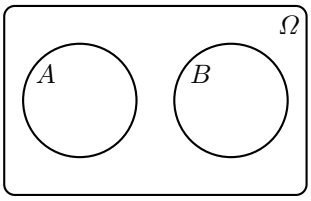
Remark 1.3.2 (Probabilities, frequencies, and common sense). You probably have an intuitive idea about what probability means. If we flip a coin many times, then the coin will come up heads roughly half the time. Thus we say that the probability that the coin will come up heads is one half. More generally, our common sense intuition about probabilities is in terms of *frequency*: if we repeated a random experiment many times, the probability of an event is the fraction of these experiments in which the event occurs.

The problem with this idea is that it is not clear how to use it to define a precise mathematical theory: we saw in the introduction that a heuristic definition in terms of fractions can lead to ambiguous conclusions. This is why we do not *define* probabilities as frequencies. Instead, we make an unambiguous mathematical definition of probability as a number $\mathbf{P}(A)$ assigned to every event A. We encode common sense into mathematics by insisting that these numbers must satisfy some rules that are precisely the properties that frequencies should have. What are these rules?

a. The fraction of experiments in which an event occurs must obviously, by definition, be a number between 0 and 1.

b. As Ω is the set of all possible outcomes, the fraction of experiments where the outcome lies in Ω is obviously 1 by definition.

c. Let A and B be two events such that $A \cap B = \varnothing$:



This means that the events A and B can never occur in the same experiment: these events are *mutually exclusive*. Now suppose we repeated the experiment many times. As A and B cannot occur simultaneously in the same experiment, the number of experiments in which A *or* B occurs is precisely the sum of the number of experiments where A occurs and where B occurs. Thus the fraction of experiments in which $A \cup B$ occurs is the sum of the fraction of experiments in which A occurs and in which B occurs. A similar conclusion holds for mutually exclusive events $E_1, E_2, \ldots$

These three properties of frequencies are precisely the rules that we require probability measures to satisfy in Definition 1.3.1. Once again, we see that the basic principles of probability theory are little more than common sense expressed in mathematical language.

Some of you might be concerned at this point that we have traded mathematics for reality: by making a precise mathematical definition, we had to give up our intuitive interpretation of probabilities as frequencies. It turns out that

this is not a problem even at the philosophical level. Even though we have not *defined* probabilities in terms of frequencies, we will later be able to *prove* using our theory that when we repeat an experiment many times, an event of probability p will occur in a fraction p of the experiments! This important result, called the *law of large numbers*, is an extremely convincing sign that have made the "right" definition of probabilities: our unambiguous mathematical theory manages to reproduce our common sense notion of probability, while avoiding the pitfalls of a heuristic definition (as in the introduction). We will prove the law of large numbers later in this course. In the meantime, you can rest assured that our mathematical definition of probability faithfully reproduces our everyday experience with randomness.

Our definition of a probability measure requires that probabilities satisfy some common sense properties. However, there are many other common sense properties that are not listed in Definition 1.3.1. It turns out that the three rules of Definition 1.3.1 are sufficient: we can derive many other natural properties as a consequence. Here are two simple examples.

Example 1.3.3. Let A be an event. Clearly A and its complement A^c are mutually exclusive, that is, $A \cap A^c = \varnothing$ (an event cannot occur and not occur at the same time!) On the other hand, we have $A \cup A^c = \Omega$ by definition (in any experiment, either A occurs or A does not occur; there are no other options!) Hence, by properties b and c in the definition of probability measure, we have

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(A \cup A^c) = \mathbf{P}(A) + \mathbf{P}(A^c),$$

which implies the common sense rule

$$\boxed{\mathbf{P}(A^c) = 1 - \mathbf{P}(A).}$$

You can verify, for example, that this rule corresponds to your intuitive interpretation of probabilities as frequencies. As a special case, suppose that $\mathbf{P}(A) = 1$, that is, we are certain that the event A will happen. Then the above rule shows that $\mathbf{P}(A^c) = 0$, that is, we are certain that A will not *not* happen. That had better be true if our theory is to make any sense!

Example 1.3.4. Let A, B be events such that $A \subseteq B$. The common sense interpretation of this assumption is that "A *implies* B". Indeed, if $A \subseteq B$, then for every outcome for which A occurs necessarily also B occurs; thus occurrence of the event A implies the occurrence of the event B.

If A implies B, then you naturally expect that B is at least as likely than A, so $\mathbf{P}(B)$ should not be smaller than $\mathbf{P}(A)$. We can derive this from Definition 1.3.1. To do that, note that we can write $B = A \cup (B \backslash A)$, where A and $B \backslash A$ are mutually exclusive. Hence, by properties a and c of Definition 1.3.1

$$\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \backslash A) \geq \mathbf{P}(A),$$

where we have used that probabilities are always nonnegative numbers.

There are many more examples of this kind. If you are ever in doubt about the correctness of a certain statement about probabilities, you should go back to Definition 1.3.1 and try to derive your statement using only the basic rules that probability measures must satisfy.

1.4 Probabilistic modelling

We have now described the three basic ingredients of any probability model: the *sample space*; the *events*; and the *probability measure*. In most problems, it is straightforward to define a suitable sample space and to describe the events of interest (see the examples in sections 1.1 and 1.2). There is no general recipe, however, for defining the probability measure: we have specified the basic rules that the probability measure must satisfy, but the precise probabilities of particular events are specific to every model. The basic problem of *probabilistic modelling* is to assign probabilities to events in a manner that captures the random phenomenon or system that we are trying to model. Sometimes, imposing natural modelling assumptions is enough to fix the probability measure. In most real-world situations, defining a good probability model requires the combination of suitable modelling assumptions and experimental measurements to determine the right parameters of the model.

As an example, let us consider some simple probability models. We will see numerous other probability models throughout this course.

Example 1.4.1 (Throwing a die). Consider the random experiment of throwing a die. As the die has six sides, the natural sample space for this model is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

To assign probabilities to events, we will make the modelling assumption that each outcome of the die is equally likely to occur. What does this mean? Before the experiment is performed, we are equally certain that the die will come up as 1, as 2, etc. The event that the die comes up as i is the set $\{i\} \subset \Omega$. Our modelling assumption can therefore be written as follows:

$$\mathbf{P}\{1\} = \mathbf{P}\{2\} = \cdots = \mathbf{P}\{6\}.$$

(In the sequel, we will frequently write $\mathbf{P}\{i\}$ for simplicity rather than $\mathbf{P}(\{i\})$ when dealing with probabilities of explicitly defined events.)

It turns out that this is enough to define the probability measure for this model. To see this, let us use the rules that a probability measure must satisfy. First, note that the events $\{1\}, \ldots, \{6\}$ are mutually exclusive (the die cannot come up both 1 and 2 simultaneously!) We therefore obtain

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(\{1\} \cup \cdots \cup \{6\}) = \mathbf{P}\{1\} + \cdots + \mathbf{P}\{6\} = 6\,\mathbf{P}\{1\},$$

where we used the modelling assumption in the last step. This implies

$$\mathbf{P}\{1\} = \mathbf{P}\{2\} = \cdots = \mathbf{P}\{6\} = \frac{1}{6}.$$

We have still only defined the probabilities of very special events $\{i\}$ of the form "the die comes up i". In the definition of a probability measure, we must define the probability of any event (such as "the die comes up an even number" $\{2, 4, 6\}$). However, we can again use the basic rules of probability measures to define these probabilities. Indeed, note that for any event $A \subseteq \Omega$,

$$\mathbf{P}(A) = \mathbf{P}\left(\bigcup_{i \in A} \{i\}\right) = \sum_{i \in A} \mathbf{P}\{i\} = \frac{|A|}{6},$$

where $|A|$ denotes the number of points in the set A (that is, the number of outcomes for which A is true). So, for example, we can compute

$$\mathbf{P}\{\text{the die comes up an even number}\} = \frac{|\{2, 4, 6\}|}{6} = \frac{3}{6} = \frac{1}{2}.$$

In this example, we encountered a useful general fact.

Example 1.4.2 (Discrete probability models). If the sample space Ω is discrete (it contains a finite or countable number of points), then in order to specify the probability $\mathbf{P}(A)$ of any event it is enough to only specify for each outcome $\omega \in \Omega$ the probability of the event $\{\omega\}$ that this outcome occurs: by Definition 1.3.1, we can then express the probability of any event $A \subseteq \Omega$ as

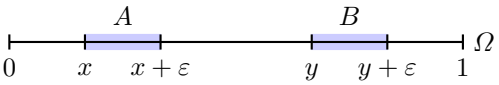
$$\mathbf{P}(A) = \mathbf{P}\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} \mathbf{P}\{\omega\}.$$

We can in principle choose $\mathbf{P}\{\omega\}$ to be any numbers, as long as they are non-negative and sum to 1 (as is required by rules a and b of Definition 1.3.1). In the special case that each outcome is equally likely, as in the previous example, the corresponding probability measure is called the *uniform distribution*.

This example may make you wonder why we bothered to define the probability of every *event* $A \subseteq \Omega$. Would it not be enough to specify the probability of every *outcome* $\omega \in \Omega$? In the next example, we will see that this is a bad idea: it is in general necessary to assign probabilities to events and not just to individual outcomes. This is why Definition 1.3.1 is the way it is.

Example 1.4.3 (Waiting for the bus). We are waiting for a bus at the bus stop. Based on previous experience, we make the following modelling assumption: the bus always comes within at most one hour; but it is equally likely to come at any time within this hour.

Of course, the natural sample space for this problem is $\Omega = [0, 1]$. The more interesting question is how to define the probability measure. To this end, consider for example the following subsets A and B of Ω:



Because our modelling assumption is that the bus is equally likely to come at any time in $[0, 1]$, the probability that the bus arrives at a time in the set A should be equal to the probability that the bus arrives at a time in B: these intervals contain *the same length of time*. In particular, this suggests that we should define the probability of any event $A \subseteq \Omega$ as

$$\mathbf{P}(A) = \text{length}(A).$$

We can check that this choice satisfies the properties of Definition 1.3.1: clearly $\mathbf{P}(\Omega) = \text{length}([0, 1]) = 1$; $\mathbf{P}(A) = \text{length}(A) \in [0, 1]$ for every $A \subseteq [0, 1]$; and the length of the union of two disjoint sets is the sum of their lengths.

In this example, the probability that the bus comes in an interval $A = [a, a + \varepsilon]$ is $\mathbf{P}(A) = \varepsilon$. For example, we can compute the probability

$$\mathbf{P}(\text{The bus comes within the first 15 minutes}) = \mathbf{P}([0, 0.25]) = 0.25.$$

This has an interesting consequence, however: the probability that the bus comes *exactly* at a given time $x \in [0, 1]$ is

$$\mathbf{P}(\text{The bus comes at } \textit{exactly} \text{ time } x) = \mathbf{P}\{x\} = 0.$$

Thus if you guess in advance of the experiment that the bus will come at exactly time x, then you will always be wrong no matter what number x you chose. This may seem paradoxical, but that is the way it must be! There are infinitely many times in the interval $[0, 1]$, and each is equally likely; so each *exact* time must occur with probability zero.

A good way to think about this phenomenon is as follows. Suppose you have a watch on which you are timing the arrival of the bus. No watch has infinite precision. An ordinary wristwatch can maybe determine the time within the precision of one second. A good stopwatch might be able to determine time with the precision of 10 milliseconds. If you want to measure even more precisely, you have to get your physicist buddy to rig up a timing device based on femtosecond lasers. For every positive amount of precision, the probability of measuring a given time to that precision is nonzero. For example,

$$\mathbf{P}(\text{The bus comes at 15 minutes } \pm 1 \text{ second}) = 2 \times 60^{-2} \approx 0.0006,$$

while

$$\mathbf{P}(\text{The bus comes at 15 minutes } \pm 10 \text{ milliseconds}) \approx 0.000006.$$

The more precisely you are asked to guess the time the bus will arrive, the less likely it is be that you will be correct. The logical conclusion must therefore be that if you are asked to guess the time the bus will arrive *exactly* (with infinite precision), you will never be correct. There is no contradiction: this is a fact of life, which is automatically built into our theory of probability.

One word of caution is in order. As the events $\{x\}$ and $\{y\}$ are mutually exclusive for $x \neq y$, you may be tempted to reason (incorrectly!) that

$$1 = \mathbf{P}(\Omega) = \mathbf{P}\left(\bigcup_{x \in [0,1]} \{x\} \right) \stackrel{?}{=} \sum_{x \in [0,1]} \mathbf{P}\{x\} = 0$$

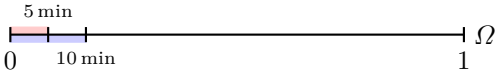
as in Example 1.4.2. Of course, this cannot be true. Where did we go wrong? In Definition 1.3.1, we only allowed the sum rule $\mathbf{P}(\bigcup E_i) = \sum_i \mathbf{P}(E_i)$ to hold for a *sequence* of mutually exclusive events $E_1, E_2, E_3, \ldots$ As it is impossible to arrange the numbers in $[0, 1]$ in a sequence (the set $[0, 1]$ is *uncountable*), this resolves our paradox: unlike in discrete models (Example 1.4.2), we can no longer deduce the probabilities of arbitrary events from the probabilities of the individual outcomes $\{\omega\}$ in continuous models. The rules that a probability measure must satisfy have been carefully designed in order to allow us to define probabilities of continuous events without getting nonsensical answers: as in the tale of Goldilocks and the Three Bears, our theory of probability turns out to be "just right" for everything to conclude in a happy ending.

Remark 1.4.4 (FYI—you may read and then immediately forget this remark!). As this is not a pure mathematics course, we will sometimes sweep some purely technical issues under the rug. One of these issues arises in continuous probability models. In the example above, we defined the probability of a subset of $[0, 1]$ as the length of that subset. It is easy to compute the length of a set if it is an interval $[a, b]$, or a finite union of intervals, or even for much more complicated sets. However, mathematicians have the uncanny ability to define really bizarre sets for which it is not clear how to measure their length. It turns out that it is impossible to define the probability of *every* subset of $[0, 1]$ in such a way that the rules of Definition 1.3.1 are satisfied: there are simply too many subsets of $[0, 1]$ in order to be able to satisfy all the constraints. In a pure mathematics course, one would therefore restrict attention only to those events that are sufficiently non-pathological that we can define their probabilities. These events are called *measurable*. For the purposes of this course, we are going to completely ignore this issue, and you should never worry about it. Any event you will encounter in any application will always be measurable: it is essentially impossible to describe a nonmeasurable event

in the English language, and so you can never run into trouble unless you are doing abstract mathematics. Careful attention to this issue is needed if you want to prove theorems, but that is beyond the level of our course.

1.5 Conditional probability

This morning, I caught the bus whose arrival time is modelled as in Example 1.4.3. I ask you to guess whether bus came within the first 5 minutes. In the absence of any further information, you would say this is quite unlikely: the probability that this happens is $\mathbf{P}([0, \frac{1}{12}]) = \frac{1}{12}$. Now suppose, however, that I give you a hint: I tell you that the bus actually came in the first 10 minutes. This still does not allow us to determine with certainty whether the bus came in the first 5 minutes. Nonetheless, given this additional information, it seems much more likely that the bus came in the first 5 minutes than without this information. Intuitively, you might expect that the probability that the bus came within the first 5 minutes, given the knowledge that it came in the first 10 minutes, is $\frac{1}{2}$, as is suggested by the following picture:



This example illustrates that probabilities should change if we gain information about what events occur. The above computation is intuitive, however, and it is not entirely obvious why it is correct. To reason rigorously in problems of this kind, we introduce the formal notion of *conditional probability*.

Definition 1.5.1. *The* conditional probability $\mathbf{P}(A|B)$ *of an event A given that the event B occurs is defined as*

$$\mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

provided $\mathbf{P}(B) > 0$ *(if* $\mathbf{P}(B) = 0$*, then the event B never occurs and so we do not need to define what it means to condition on this event ocurring).*

This definition is quite natural. When we condition on B, only outcomes where B occurs remain possible. This restricts the set of all possible outcomes to B, and the set of favorable outcomes to $A \cap B$. Thus the conditional probability of A given B is $\mathbf{P}(A \cap B)/\mathbf{P}(B)$: we must normalize by $\mathbf{P}(B)$ to ensure that the probability of all possible outcomes is one (that is, $\mathbf{P}(B|B) = 1$).

Remark 1.5.2 (Frequency interpretation). In Remark 1.3.2, we argued that the rules that define probability measures have a common sense interpretation

in terms of frequencies: if we repeat a random experiment many times, the probability $\mathbf{P}(A)$ is the fraction of the experiments in which A occurs.

We can give a similar common sense interpretation to conditional probabilities. Suppose we want to measure the probability that A occurs, given that B occurs. To do this, we first repeat the experiment many times. As we are only interested in what happens when the event B occurs, we discard all those experiments in which B does not occur. Then $\mathbf{P}(A|B)$ is the fraction of the remaining experiments (that is, of those experiments in which B occurs) in which the event A occurs. We can now see that this intuitive interpretation of conditional probability naturally leads to the above formula. Indeed, the fraction of the experiments in which B occurs where A also occurs is the ratio of the number of experiments in which B *and* A occur, and the total number of experiments in which B occurs. Thus $\mathbf{P}(A|B)$ must be the ratio of $\mathbf{P}(A \cap B)$ and $\mathbf{P}(B)$, which is exactly what we use as the defining property of conditional probabilities in our mathematical theory. Once again, we see that Definition 1.5.1 is just common sense expressed in mathematical form.

Even though we have not defined probabilities as frequencies in our mathematical theory, we already noted in Remark 1.3.2 that we will be able to *prove* the connection between frequencies and probabilities using our theory later in this course. When we do this, it will be easy to prove the frequency interpretation of conditional probabilities as well. In the meantime, you should not hesitate to let this interpretation guide your intuition.

In order for our definition of conditional probability to make sense, one thing we should check is that it does in fact satisfy the defining properties of probabilities! It is a simple exercise to verify that this is the case:

- It is trival that $\mathbf{P}(B|B) = \mathbf{P}(B \cap B)/\mathbf{P}(B) = 1$ (this is common sense: given the knowledge that B happens, we are certain that B happens!)
- As probabilities are nonnegative and as $A \cap B \subseteq B$, it follows easily that $0 \leq \mathbf{P}(A|B) \leq 1$ (conditional probabilities are in fact probabilities!)
- Let A, C be mutually exclusive events. Then $(A \cup C) \cap B = (A \cap B) \cup (C \cap B)$ (why? draw a picture, or express this statement in English). We therefore have $\mathbf{P}(A \cup C|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B) + \mathbf{P}(C \cap B)/\mathbf{P}(B) = \mathbf{P}(A|B) + \mathbf{P}(C|B)$ (conditional probabilities behave like probability measures). The same conclusion follows for a sequence of mutually exclusive events $E_1, E_2, \ldots$

That is, conditional probabilities behave exactly like probabilities.

Let us now derive a more interesting property of conditional probabilities. By the definition of conditional probability, we have

$$\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\mathbf{P}(B),$$
$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A|B^c)\mathbf{P}(B^c).$$

As $A = (A \cap B) \cup (A \cap B^c)$ (why?), this implies

$$\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c) = \mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(A|B^c)\mathbf{P}(B^c).$$

This allows us to derive the following very useful **Bayes formula**:

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(B \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A|B)\mathbf{P}(B)}{\mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(A|B^c)\mathbf{P}(B^c)}.$$

The beauty of the formula is that it allows us to turn around the role of the conditioning and conditioned event: that is, it expresses $\mathbf{P}(B|A)$ in terms of $\mathbf{P}(A|B)$ and $\mathbf{P}(A|B^c)$. In many problems, one of these conditional probabilities is much easier to compute than the other. In fact, our modelling assumptions will often use conditional probabilities to *define* probability models.

Example 1.5.3 (Medical diagnostics). You are not feeling well, and go to the doctor. The doctor fears your symptoms may be consistent with a serious disease (say, cancer). To be sure, he sends you to undergo a medical test. Unfortunately, no medical test is always accurate: even the best test has a small probability of giving a false positive or false negative. In the present case, medical statistics show that patients who have this disease test positive 95% of the time, while patients who do not have the disease test positive 2% of the time (this is a pretty accurate test!) In the general population, one in a thousand people in your age group have this disease.

A week after the test, your doctor calls you back with the result. The test came back positive! Given this information, what is the probability that you actually have the feared disease?

This problem is the source of a common fallacy among doctors which can have serious consequences. Many doctors might conclude that given that the test came back positive, the probability that you have the disease is 95% (extremely likely). Nothing is further from the truth: we will see that this test provides almost no information. On the basis of such misconceptions, doctors may prescribe invasive treatment programs that can be very detrimental to patient's health. Please do not commit this crime against probability.

Let us do a careful computation to see how things really are. Denote by A the event that the test comes back positive, and by B the event that you have the disease. Our modelling assumptions are:

- $\mathbf{P}(A|B) = 0.95$ (95% of patients with the disease test positive).
- $\mathbf{P}(A|B^c) = 0.02$ (2% of patients without the disease test positive).
- $\mathbf{P}(B) = 0.001$ (one in a thousand people have the disease).

We want to compute the probability $\mathbf{P}(B|A)$ the you have the disease given that the test comes back positive. By the Bayes formula,

$$\mathbf{P}(B|A) = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.02 \cdot 0.999} \approx 0.045.$$

Thus there is only 4.5% chance that you have the disease, despite that your test came back positive! This very different conclusion than the incorrect conclusion suggested above will lead you to be much more careful in making any treatment decisions on the basis of such a test.

What is the intuitive explanation for having such a small probability that you have the disease despite that the test is very reliable? Even though this test will rarely give a false positive, the disease that is being tested is even more rare. This means that if we collect a random sample from the general population, we are much more likely to encounter a false positive than someone who actually has the disease. This makes it very difficult to draw any conclusions from the outcome of the test. In order for a test to be useful, the fraction of the cases where it gives a false positive must be much smaller than the fraction of the population that has the disease.

You can see in the above example that the Bayes formula is very useful. Nonetheless, I recommend that you do not memorize this formula, but rather learn how to derive it from the basic properties of conditional probabilities. By forcing yourself to get very comfortable with conditional probabilities, you will minimize the potential to make conceptual mistakes.

1.6 Independent events

Consider a random experiment where two people each flip a coin. The probability of each coin coming up heads in one half. If the one person flips her coin first, and tells us the outcome, should that change our degree of confidence in the outcome of the other coin? Unless there is a conspiracy between the two people, the answer, of course, if no: as they are flipping their coins independently, knowing one of the outcomes cannot affect the other.

This notion of independence of events has a completely natural interpretation in terms of conditional probabilities: if we condition on the outcome of one of the coins, that does not change the probability of the other coin.

Definition 1.6.1. *Events* A, B *are* independent *if* $\mathbf{P}(A|B) = \mathbf{P}(A)$.

One concern you might have about this definition is that it seems somewhat arbitrary: would it not be equally logical to require $\mathbf{P}(B|A) = \mathbf{P}(B)$? It turns out that these two definitions are equivalent. In fact, it is useful to give a more symmetric definition of independence. Note that if A, B are independent, then the definition of conditional probability implies

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

We can use this as a completely equivalent definition of independence.

Definition 1.6.2. *Events* A, B *are* independent *if* $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

There is no difference between Definitions 1.6.1 and 1.6.2: they are two different ways of writing the same thing. In performing computations, it is often Definition 1.6.2 that is particularly useful. However, Definition 1.6.1 is much more intuitive, as it expresses directly our common sense notion of independence. You can use either one as you please.

Once we have defined a probabilistic model, we can use the definition of independence to check whether two given events are independent. However, we will often use independence in the opposite direction as an extremely useful *modelling assumption*. That is, in order to model a random system, we will use our intuition and understanding of the problem to determine which events must be independent, and then design our model accordingly.

Example 1.6.3 (Flipping two coins). We flip two coins independently; E_i is the event that the ith coin comes up heads ($i = 1, 2$). Let $\mathbf{P}(E_i) = p$ (the coins may be biased, that is, heads and tails are not necessarily equally likely).

Our modelling assumption is that E_1 and E_2 are independent. This implies

$$\mathbf{P}(\text{both coins come up heads}) = \mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1)\mathbf{P}(E_2) = p^2.$$

How about the probability the the first coin comes of heads and the second comes up tails, that is, $\mathbf{P}(E_1 \cap E_2^c)$? Our intuition tells us that if E_1 and E_2 are independent, then E_1 and E_2^c should also be independent (if your degree of confidence in the outcome of the first coin is unchanged if you condition on the second coin coming up heads, it should also be unchanged if you condition on the second coin *not* coming up heads). This is not obvious from the definition, however, so we must prove it. Note that we can write $E_1 = (E_1 \cap E_2) \cup (E_1 \cap E_2^c)$ as the union of two mutually exclusive events (why?), so

$$\mathbf{P}(E_1) = \mathbf{P}(E_1 \cap E_2) + \mathbf{P}(E_1 \cap E_2^c).$$

This implies using independence of E_1 and E_2

$$\begin{aligned} \mathbf{P}(E_1 \cap E_2^c) &= \mathbf{P}(E_1) - \mathbf{P}(E_1 \cap E_2) \\ &= \mathbf{P}(E_1) - \mathbf{P}(E_1)\mathbf{P}(E_2) \\ &= \mathbf{P}(E_1)(1 - \mathbf{P}(E_2)) \\ &= \mathbf{P}(E_1)\mathbf{P}(E_2^c). \end{aligned}$$

We have therefore verified the intuitively obvious fact that the independence of E_1, E_2 implies the independence of E_1, E_2^c (as well as independence of E_1^c, E_2^c, etc.; despite that this is common sense, we should check at least once that our theory is still doing the right thing!) In the present example, this gives

$$\mathbf{P}(\text{coin 1 comes up heads and coin 2 comes up tails}) = p(1-p).$$

We can similarly compute $\mathbf{P}(E_1^c \cap E_2) = p(1-p)$ and $\mathbf{P}(E_1^c \cap E_2^c) = (1-p)^2$.

Remark 1.6.4. We have emphasized the three basic ingredients of probabilistic models: the sample space (which contains all possible outcomes), the events (which are subsets of the sample space), and the probability measure (which assigns a probability to every event). In the above example, however, we never defined explicitly the sample space Ω and the probability $\mathbf{P}(A)$ of every possible event A. We started from our basic modelling assumptions, and worked from there directly to arrive at the answers to our questions.

There is nothing wrong or fishy going on here. The sample space and probability measure really are there: we simply did not bother to write them down. The probabilistic model is being defined and used implicitly as the basis for our computations. If we wanted to, we could make every part of the model explicit. For sake of illustration, let us go through this exercise. As we are flipping two coins, a good sample space would be

$$\Omega = \{HH, HT, TH, TT\},$$

and the events E_1 and E_2 can then be expressed as subsets

$$E_1 = \{HH, HT\}, \qquad E_2 = \{HH, TH\}.$$

By the computations that we have performed above, our modelling assumptions and the rules of probability imply that we must choose

$$\mathbf{P}\{HH\} = p^2, \qquad \mathbf{P}\{HT\} = \mathbf{P}\{TH\} = p(1-p), \qquad \mathbf{P}\{TT\} = (1-p)^2,$$

and we can subsequently define the probability of any event as

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

We have now worked out exactly every ingredient of the probabilistic model that corresponds to our assumptions. While this reassures us that our model makes sense, this did not really help us to do any computations: by using our modelling assumptions and applying the rules of probability, we could get the answers we wanted without having to write down all these details.

As our models become increasingly complicated, we will usually omit to write out every single detail of the underlying sample space and probabilities, but instead work directly from our modelling assumptions. While this will save a lot of trees, we can always rest secure that somewhere under the hood is a bona fide sample space and probability measure that are working for us to produce common sense using mathematics. You typically do not need to worry about this: however, if you are ever in doubt, you can always go back to expressing your model carefully in terms of the basic ingredients of probability theory in order to make sure that everything is unambiguous.

So far, we have only discussed the independence of two events. We must extend our notion of independence to describe the situation where ten people each flip a coin independently. In this setting, our degree of confidence about one of the coins cannot be affected by knowing that any subset of the other coins came up heads. This idea is directly expressed by the following definition.

Definition 1.6.5. *Events* $E_1, \ldots, E_n$ *are* independent *if*

$$\mathbf{P}(E_i | E_{j_1} \cap \ldots \cap E_{j_k}) = \mathbf{P}(E_i)$$

for all $i \neq j_1 \neq \ldots \neq j_k$, $1 \leq k < n$.

As in the case of two events, we can express this notion of independence in a more symmetric fashion. Note that

$$\begin{aligned}
\mathbf{P}(E_{j_1} \cap \ldots \cap E_{j_k}) &= \mathbf{P}(E_{j_1} | E_{j_2} \cap \ldots \cap E_{j_k}) \mathbf{P}(E_{j_2} \cap \ldots \cap E_{j_k}) \\
&= \mathbf{P}(E_{j_1}) \mathbf{P}(E_{j_2} \cap \ldots \cap E_{j_k}) \\
&= \mathbf{P}(E_{j_1}) \mathbf{P}(E_{j_2} | E_{j_3} \cap \ldots \cap E_{j_k}) \mathbf{P}(E_{j_3} \cap \ldots \cap E_{j_k}) \\
&= \ldots \\
&= \mathbf{P}(E_{j_1}) \mathbf{P}(E_{j_2}) \ldots \mathbf{P}(E_{j_k}).
\end{aligned}$$

That is, we obtain the following definition that is equivalent to Definition 1.6.5.

Definition 1.6.6. *Events* $E_1, \ldots, E_n$ *are* independent *if*

$$\mathbf{P}(E_{j_1} \cap \ldots \cap E_{j_k}) = \mathbf{P}(E_{j_1}) \mathbf{P}(E_{j_2}) \cdots \mathbf{P}(E_{j_k})$$

for all $j_1 \neq \ldots \neq j_k$, $1 \leq k \leq n$.

Example 1.6.7 (Flipping five coins). We flip five coins independently; E_i is the event that the ith coin comes up heads ($1 \leq i \leq 5$). Let $\mathbf{P}(E_i) = p$. What is the probability that four of the coins come up heads and one of the coins come up tails? This event can be written as the union of five mutually exclusive event: that the first coin comes up tails and the rest are heads; that the second coin comes up tails and the rest are heads, etc. Therefore

$$\begin{aligned}
&\mathbf{P}(\text{four heads, one tails}) \\
&= \mathbf{P}(E_1^c \cap E_2 \cap E_3 \cap E_4 \cap E_5) + \mathbf{P}(E_1 \cap E_2^c \cap E_3 \cap E_4 \cap E_5) \\
&\quad + \mathbf{P}(E_1 \cap E_2 \cap E_3^c \cap E_4 \cap E_5) + \mathbf{P}(E_1 \cap E_2 \cap E_3 \cap E_4^c \cap E_5) \\
&\quad + \mathbf{P}(E_1 \cap E_2 \cap E_3 \cap E_4 \cap E_5^c) \\
&= 5(1-p)p^4.
\end{aligned}$$

1.7 Random variables

Up to this point, the only kind of random objects we have encountered are events: statements that are either true or false depending on the outcome of the experiment. In most situations, we will be interested in random quantities that are not necessarily represented by a yes/no question. For example, if we flip five coins, the number of heads is a random number between 0 and 5. Such random quantities are called *random variables*.

In the following definition, Ω is the sample space and D is a set of values.

Definition 1.7.1. *A* random variable *taking values in D is a function X that assigns a value $X(\omega) \in D$ to every possible outcome $\omega \in \Omega$.*

The idea behind the definition of a random variable is quite similar to that of an event. While we do not know prior to performing the experiment what value the random variable X will take, we can specify for every possible outcome $\omega \in \Omega$ of the experiment what the value $X(\omega)$ of the random variable would be if Tyche happened to choose that outcome ω. We can now reason about the outcomes that may occur in a given experiment by looking at the probabilities of appropriate events. For example, the set of outcomes

$$\{\omega \in \Omega : X(\omega) = i\}$$

is the event that the random variable X takes the value i (which may be true or false in any given experiment). As writing out such long expressions will get tiresome, we will usually write this event simply as $\{X = i\}$. The probability that X takes the value i can consequently be expressed as $\mathbf{P}\{X = i\}$.

Example 1.7.2. We flip three coins. A good sample space for this problem is

$$\Omega = \{HHH, THH, HTH, HHT, TTH, THT, HTT, TTT\}.$$

Let X be the total number of heads that are flipped in this experiment. Then X is a random variable that can be defined explicitly as follows:

$$\begin{aligned}
&X(HHH) = 3, \quad X(THH) = X(HTH) = X(HHT) = 2, \\
&X(TTH) = X(THT) = X(HTT) = 1, \quad X(TTT) = 0.
\end{aligned}$$

Of course, there is no need in practice to write out explicitly the value of $X(\omega)$ for every possible outcome $\omega \in \Omega$: we have just done that here for sake of illustration. The verbal description "X is the total number of heads" is ample to define this random variable unambiguously.

Example 1.7.3 (Repeated coin flips). Let us repeatedly flip a coin with probability p of coming up heads. We flip independently each time, and we repeat indefinitely. A natural sample space for such an infinite sequence of flips is

$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \ldots) : \omega_i \in \{H, T\} \text{ for every } i\}.$$

Denote by E_i the event that the ith coin comes up heads: that is,

$$E_i = \{\omega \in \Omega : \omega_i = H\}.$$

Our modelling assumption is that $\mathbf{P}(E_i) = p$ and that all E_i are independent.

We would like to know how long we have to wait until we flip a heads for the first time. To study this question, let X be the first time that a coin comes up heads. Then the random variable X is a function

$$X : \Omega \to \{1, 2, 3, \ldots\} \cup \{+\infty\}.$$

For example, for an outcome that starts as $\omega = (T, T, T, H, T, H, T, \ldots)$, we have $X(\omega) = 4$ (as the fourth coin is the first to come up heads). The random variable X takes the value $+\infty$ if no heads is every flipped.

Let us now discuss some examples of probabilities involving X.

a. The event that we get heads for the first time exactly on the nth flip is

$$\{\omega \in \Omega : X(\omega) = n\} = \{\omega = (\underbrace{T, T, \ldots, T}_{n-1 \text{ times}}, H, \underbrace{\ldots}_{\text{anything}})\}.$$

A better way to express this event is

$$\{X = n\} = E_1^c \cap E_2^c \cap \cdots \cap E_{n-1}^c \cap E_n.$$

We can therefore compute by independence

$$\mathbf{P}\{X = n\} = \mathbf{P}(E_1^c)\mathbf{P}(E_2^c) \cdots \mathbf{P}(E_{n-1}^c)\mathbf{P}(E_n) = p(1-p)^{n-1}.$$

b. The event that we get heads for the first time at time 6 or later is

$$\{\omega \in \Omega : X(\omega) \geq 6\} = \{\omega = (T, T, T, T, T, \underbrace{\ldots}_{\text{anything}})\}$$

(reason carefully why this is true!), or equivalently

$$\{X \geq 6\} = E_1^c \cap E_2^c \cap E_3^c \cap E_4^c \cap E_5^c.$$

We can therefore compute

$$\mathbf{P}\{X \geq 6\} = \mathbf{P}(E_1^c)\mathbf{P}(E_2^c) \cdots \mathbf{P}(E_5^c) = (1-p)^5.$$

c. The event that we get heads for the first time at the latest by time 5 is much more easily studied by considering its negation:

$$\{X \leq 5\} = \{X \geq 6\}^c.$$

We can therefore compute immediately

$$\mathbf{P}\{X \leq 5\} = 1 - \mathbf{P}\{X \geq 6\} = 1 - (1-p)^5.$$

There is a good lesson here: sometimes it is easier to compute the probability of the negation of an event than of the event itself!
We could have done this computation differently. Note that

$$\{X \leq 5\} = \{X = 1\} \cup \{X = 2\} \cup \{X = 3\} \cup \{X = 4\} \cup \{X = 5\}.$$

As the right-hand side is the union of mutually exclusive events, we obtain

$$\mathbf{P}\{X \leq 5\} = \sum_{n=1}^{5} \mathbf{P}\{X = n\} = \sum_{n=1}^{5} p(1-p)^{n-1}.$$

You are encouraged as an exercise to show that the geometric sum that we get in this manner does indeed equal $1 - (1-p)^5$ (no matter how you compute, you should end up with the same answer!)

In principle, random variables can take values in an arbitrary set D. This set need not be numerical: for example, if X is the suit of a randomly drawn card from a deck of cards, it is perfectly reasonable to define this as a D-valued random variable with $D = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. This does not present any difficulties in our theory. It will however prove to be convenient, in first instance, to consider random variables with values in a finite or countable set D.

Definition 1.7.4. *A random variable $X \colon \Omega \to D$ is called* discrete *if it takes values in a finite or countable set D.*

For example, the suit of a randomly drawn card is discrete (it can only take four possible values), and the number of customers that arrive at a store in a given day is also discrete (it takes values in $\{0, 1, 2, \ldots\}$ which is a countable set). However, the height of a random person can take any value in $\mathbb{R}_+$ which is not a countable set, and this is therefore not a discrete random variable. The latter type of random variable is called *continuous*.

Of course, a useful theory of probability must be able to deal with both discrete and continuous random variables, and both these notions will appear numerous times throughout this course. Since it is mathematically easier to deal with the discrete case, however, we will initially focus our attention on discrete random variables. Later in this course, after we have some experience working with discrete random variables, we will come back to the basic principles of probability and extend our theory to continuous random variables.

Remark 1.7.5. It is customary to denote random variables by uppercase letters $X, Y, \ldots$ and nonrandom values by lowercase letters $x, y, i, j, \ldots$ For example, $\{X = x\}$ is the event that the random variable X takes the given value x. It is good to get into the habit of following this notational convention, as it helps clarify what parts of a problem are random and what parts are not.

1.8 Expectation and distributions

Recall that the probability $\mathbf{P}\{X = i\}$ has a natural interpretation as the fraction of repeated experiments in which the random variable X takes the value i. This does not necessarily give us an immediate indication of how large X typically is. In many cases, we would like to know how large X is "on average," that is, what is the average value of X over many repeated experiments? Of course, such questions only make sense if the random variable X is numerical, that is, of the set of values D is a set of numbers (for example, it does not make much sense to compute the average of $\diamondsuit$ and $\heartsuit$). In general, we can compute an average of $\varphi(X)$ if we associate a numerical value $\varphi(i) \in \mathbb{R}$ to every possible value $i \in D$ taken by the random variable.

Definition 1.8.1. *Let $X : \Omega \to D$ be a discrete random variable, and let $\varphi : D \to \mathbb{R}$ be a function. The* expectation *of $\varphi(X)$ is defined as*

$$\mathbf{E}(\varphi(X)) = \sum_{i \in D} \varphi(i)\, \mathbf{P}\{X = i\}.$$

If X is numerical ($D \subset \mathbb{R}$), then $\mathbf{E}(X)$ is defined by choosing $\varphi(i) = i$.

Remark 1.8.2 (Frequency interpretation). We motivated the notion of expectation as the average of the value of a random variable X over repeated experiments. How is this reflected in the definition of expectation? Suppose we repeat the experiment n times and observe the values $x_1, x_2, \ldots, x_n \in D$ in each experiment. The average of the values $\varphi(x_i)$ can be expressed as (why?)

$$\frac{1}{n} \sum_{k=1}^{n} \varphi(x_k) = \sum_{i \in D} \varphi(i)\, \frac{\#\{k \le n : x_k = i\}}{n}.$$

The quantity $\frac{1}{n}\#\{k \le n : x_k = i\}$ is the fraction of experiments in which the random variable takes the value i, which corresponds intuitively to the probability $\mathbf{P}\{X = i\}$. We can therefore interpret the formal definition of expectation given in Definition 1.8.1 as an entirely common sense relation between averages and frequencies. Once we prove the law of large numbers (which shows that probabilities are indeed frequencies), we will also be able to make precise the notion that expectations are really averages.

Remark 1.8.3. In general, random variables X do not need to be numerical. However, numerical random variables are extremely common, and many concepts (such as $\mathbf{E}(X)$) are defined only for numerical variables. We will therefore often refer to numerical random variables simply as random variables when the fact that they must be numerical is obvious from the context.

Remark 1.8.4. Note that Definition 1.8.1 does not make sense as written for continuous random variables. For example, let X be the length of time we must wait for the bus in Example 1.4.3. We have seen that $\mathbf{P}\{X = x\} = 0$ for any $x \in [0, 1]$, yet clearly the average amount of time we must wait for the bus is not zero in this example. This technical difficulty means we must be a bit more careful when we define the expectation of a continuous random variable. In order not to get distracted by such issues, we only consider discrete random variables for the time being, and extend the notion of expectation to continuous random variables later in this course.

Example 1.8.5 (A simple card game). We draw a card uniformly at random from a deck of cards. We agree that if its suit is clubs, you lose \$1; if it is diamonds, you win or lose nothing; of it is hearts, you win \$1; and if it is spades, you win \$7. What are your expected winnings?

To model this game, let $X \in D = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$ be the suit of the randomly drawn card. As each suit appears equally often in the deck, it is not difficult to see that $\mathbf{P}\{X = i\} = \frac{1}{4}$ for every suit $i \in D$ (work out the sample space and probability measure if this is not obvious!) The function $\varphi : D \to \mathbb{R}$ defines how much money we win for every suit; that is, $\varphi(\clubsuit) = -1$, $\varphi(\diamondsuit) = 0$, $\varphi(\heartsuit) = 1$, and $\varphi(\spadesuit) = 7$. We can therefore compute our expected winnings

$$\mathbf{E}(\varphi(X)) = \frac{1}{4}(-1 + 0 + 1 + 7) = 1.75.$$

Note that it does not make any sense in this example to talk about the "expected suit" $\mathbf{E}(X)$, as we cannot average suits. However, it makes perfect sense to talk about the expected winnings $\mathbf{E}(\varphi(X))$.

Example 1.8.6 (Indicator functions). The following general idea is often useful. For any event A, define the *indicator function* $\mathbf{1}_A : \Omega \to \{0, 1\}$ as

$$\mathbf{1}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

That is, $\mathbf{1}_A$ is the discrete random variable that takes the value 1 if A occurs and takes the value 0 if A does not occur. The expectation of $\mathbf{1}_A$ is

$$\mathbf{E}(\mathbf{1}_A) = 1 \cdot \mathbf{P}\{\mathbf{1}_A = 1\} + 0 \cdot \mathbf{P}\{\mathbf{1}_A = 0\} = \mathbf{P}(A).$$

We therefore see that the probability of an event A is just the expectation of the random variable $\mathbf{1}_A$. This observation is useful in many computations.

From Definition 1.8.1, it is clear that in order to compute the expectation $\mathbf{E}(\varphi(X))$ for any function φ, we only need to know the probability $\mathbf{P}\{X = i\}$ that X can take any value $i \in D$: we do not need to know precisely how the random variable $X : \Omega \to D$ is defined as a function on the sample space. It is often natural to model a random variable by specifying the probabilities $\mathbf{P}\{X = i\}$ of each of its possible values $i \in D$. We then say that we have specified the *distribution* of the random variable.

Definition 1.8.7. *Let $X : \Omega \to D$ be a discrete random variable. The collection $(\mathbf{P}\{X = i\})_{i \in D}$ is called the* distribution *of X.*

It is perfectly possible, and very common, for different random variables to have the same distribution. Suppose, for example, that we throw two dice. Let X be the outcome of the first die, and Y be the outcome of the second die. Then X and Y are different random variables (when you throw two dice, you typically get two different numbers!) But X and Y have the same distribution, as each die is equally likely to yield every outcome $\mathbf{P}\{X = i\} = \mathbf{P}\{Y = i\} = \frac{1}{6}$ for $1 \leq i \leq 6$ (this is called the *uniform distribution*).

Let us now consider a pair $X : \Omega \to D$ and $Y : \Omega \to D'$ of discrete random variables. How can we compute the expectation of a function $\varphi(X, Y)$ of two random variables? You could just think of the pair $Z = (X, Y)$ as a new discrete random variable, which takes values in the set of pairs of outcomes $\{(i, j) : i \in D, j \in D'\}$. Once you realize this, the expectation of $\varphi(X, Y) = \varphi(Z)$ is just the expectation of another discrete random variable:

$$\mathbf{E}(\varphi(X, Y)) = \sum_{i \in D, j \in D'} \varphi(i, j)\, \mathbf{P}\{X = i, Y = j\}.$$

To compute $\mathbf{E}(\varphi(X, Y))$ for any function φ, it is not enough to know just the probabilities $\mathbf{P}\{X = i\}$ and $\mathbf{P}\{Y = j\}$: we need to know the probabilities $\mathbf{P}\{X = i, Y = j\}$. The latter specify the *joint distribution* of X and Y.

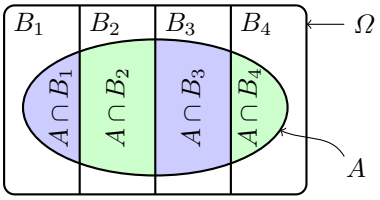
Definition 1.8.8. *Let $X : \Omega \to D$ and $Y : \Omega \to D'$ be discrete random variables. Then $(\mathbf{P}\{X = i, Y = j\})_{i \in D, j \in D'}$ is their* joint distribution.

Given the *joint* distribution $\mathbf{P}\{X = i, Y = j\}$ of two random variables X and Y, you can always recover the *marginal* distributions $\mathbf{P}\{X = i\}$ and $\mathbf{P}\{Y = j\}$. To see how, let us first do a simple probability exercise.

Example 1.8.9. Let A and $B_1, \ldots, B_n$ be events such that $B_i \cap B_j = \varnothing$ for all $i \neq j$ (the events $B_1, \ldots, B_n$ are mutually exclusive) and $B_1 \cup \cdots \cup B_n = \Omega$ (in any given experiments, one of the events $B_1, \ldots, B_n$ must occur). Then

$$A = A \cap (B_1 \cup \cdots \cup B_n) = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n).$$

What have we done here? Say A is the event that a coin flip comes up heads, B_1 is the event that it rains in Brazil, and B_2 is the event that it does not rain in Brazil. Then the event "the coin comes up heads" can be written as "the coin comes up heads *and* it rains in Brazil, *or* the coin comes up heads *and* it does not rain in Brazil". It does not matter that the rain in Brazil has nothing to do with the coin flip: the equality between these events is obviously true. If you prefer to think in terms of sets, this is illustrated in the following figure:



Because the events $A \cap B_i$ are disjoint, we can compute

$$\mathbf{P}(A) = \mathbf{P}(A \cap B_1) + \mathbf{P}(A \cap B_2) + \cdots + \mathbf{P}(A \cap B_n).$$

This formula often very useful in cases where $\mathbf{P}(A)$ may be difficult to compute directly, but $\mathbf{P}(A \cap B_i)$ can be easy to compute.

Let us now return to the problem of computing the marginal distribution $\mathbf{P}\{X = i\}$ from the joint distribution $\mathbf{P}\{X = i, Y = j\}$. We will apply the previous example with $A = \{X = i\}$ and $B_j = \{Y = j\}$:

$$\mathbf{P}\{X = i\} = \sum_{j \in D'} \mathbf{P}\{X = i, Y = j\}.$$

Therefore, to compute the distribution of X from the joint distribution of X and Y, we must simply sum over all possible outcomes of Y.

To see how this can be useful, let us investigate an important property of the expectation of random variables. Let $X : \Omega \to D$ and $Y : \Omega \to D'$ be discrete random variables. Then $X + Y$, the sum of the two variables, is also a discrete random variable. If we think about expectation as the average over many repeated experiments, it would seem to be evident that we must have

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y).$$

After all, the average of a sum of two quantities $\frac{1}{n}\sum_{k=1}^{n}(x_k + y_k)$ is just the sum of the averages $\frac{1}{n}\sum_{k=1}^{n} x_k + \frac{1}{n}\sum_{k=1}^{n} y_k$. This common sense property is not entirely obvious, however, from the formal definition of expectation. We can easily prove it, however, if we view the sum as a function $\varphi(x, y) = x + y$:

$$\begin{aligned}
\mathbf{E}(X+Y) &= \sum_{x\in D}\sum_{y\in D'} (x+y)\,\mathbf{P}\{X=x, Y=y\} \\
&= \sum_{x\in D}\sum_{y\in D'} x\,\mathbf{P}\{X=x, Y=y\} + \sum_{x\in D}\sum_{y\in D'} y\,\mathbf{P}\{X=x, Y=y\} \\
&= \sum_{x\in D} x\,\mathbf{P}\{X=x\} + \sum_{y\in D'} y\,\mathbf{P}\{Y=y\} \\
&= \mathbf{E}(X) + \mathbf{E}(Y).
\end{aligned}$$

Thus the common sense property $\mathbf{E}(X+Y) = \mathbf{E}(X)+\mathbf{E}(Y)$ does indeed hold in our mathematical theory of probability, as it should!

Example 1.8.10 (Flipping coins). We flip n coins. Each coin has probability p of coming up heads. What is the expected number of times that we flip heads?

Let E_i be the event that the ith coin comes up heads. Our modelling assumption is that $\mathbf{P}(E_i) = p$ for all $i \le n$. Let X be the total number of heads that we flip. We can write the random variable X as (why?)

$$X = \mathbf{1}_{E_1} + \mathbf{1}_{E_2} + \cdots + \mathbf{1}_{E_n}.$$

We can therefore compute

$$\begin{aligned}
\mathbf{E}(X) &= \mathbf{E}(\mathbf{1}_{E_1}) + \cdots + \mathbf{E}(\mathbf{1}_{E_n}) \\
&= \mathbf{P}(E_1) + \cdots + \mathbf{P}(E_n) = np.
\end{aligned}$$

Note that we did not need to assume the coins are flipped independently for this to be true! All we are using here is linearity of the expectation.

1.9 Independence and conditioning

So far, we have only defined what it means for events to be independent. It makes perfect sense, however, to say that two random variables are independent. The definition extends in the obvious manner: X is independent of Y if conditioning on the outcome of Y does not affect the distribution of X.

Definition 1.9.1. *Let $X : \Omega \to D_X$, $Y : \Omega \to D_Y$ be discrete random variables. Then X, Y are* independent *(sometimes denoted as $X \perp\!\!\!\perp Y$) if*

$$\mathbf{P}\{X = x | Y = y\} = \mathbf{P}\{X = x\}$$

for every $x \in D_X$ and $y \in D_Y$.

Exactly as in the case of independent events, you can easily check that X and Y are independent if and only if for every $x \in D_X$ and $y \in D_Y$

$$\mathbf{P}\{X = x, Y = y\} = \mathbf{P}\{X = x\}\mathbf{P}\{Y = y\}.$$

This form of the independence property is often useful in practice.

Given random variables X and Y, we can use the above definition to verify that they are independent. We will often use independence in the opposite direction, however: we impose independence as a modelling assumption in order to define our model. For example, consider the experiment of throwing two dice. Let X be the outcome of the first die, and Y be the outcome of the second die, both taking values in $\{1, \ldots, 6\}$. It is natural to assume that every outcome of each die is equally likely, and that the outcomes of the two dice are independent. Then X and Y are *independent and identically distributed*, a notion so ubiquitous that it has its own abbreviation *i.i.d.* In the present case, our assumptions are sufficient to specify everything we need to know about this model: because of independence, the joint distribution of X and Y is given by $\mathbf{P}\{X = i, Y = j\} = \mathbf{P}\{X = i\}\mathbf{P}\{Y = j\} = \frac{1}{36}$ for every i, j.

Example 1.9.2 (Raising the stakes). We play the same card game as in Example 1.8.5, but now we add a twist. In addition to drawing a card, we independently throw a die. The amount of winnings (or loss) in the card game is multiplied by the number on the die. What are your expected winnings?

Let the suit X of the randomly drawn card and the winnings function φ in the card game be defined as in Example 1.8.5. In addition, let Y be the number on the die. The total winnings in this game are given by the random variable $\varphi(X)Y$. As X and Y are independent, we can compute

$$\begin{aligned}\mathbf{E}(\varphi(X)Y) &= \sum_{i\in D}\sum_{k=1}^{6}\varphi(i)k\,\mathbf{P}\{X=i,Y=k\}\\ &= \sum_{i\in D}\sum_{k=1}^{6}\varphi(i)k\,\mathbf{P}\{X=i\}\mathbf{P}\{Y=k\}\\ &= \sum_{i\in D}\varphi(i)\,\mathbf{P}\{X=i\}\sum_{k=1}^{6}k\,\mathbf{P}\{Y=k\}\\ &= \mathbf{E}(\varphi(X))\,\mathbf{E}(Y) = 1.75\cdot 3.5 = 6.125.\end{aligned}$$

More generally, the same reasoning shows that if X, Y are independent (discrete) random variables, then $\mathbf{E}(f(X)g(Y)) = \mathbf{E}(f(X))\,\mathbf{E}(g(Y))$ for any f, g.

Just as in the case of events, we can define what it means for more than two random variables to be independent: the distribution of each random variable is not affected by conditioning on the outcomes of the other random variables.

Definition 1.9.3. *Discrete variables $X_1, \ldots, X_n$ are* independent *if*

$$\mathbf{P}\{X_i = x_i | X_{j_1} = x_{j_1}, \ldots, X_{j_k} = x_{j_k}\} = \mathbf{P}\{X_i = x_i\}$$

for all $x_i, x_{j_1}, \ldots, x_{j_k}$, $i \neq j_1 \neq \cdots \neq j_k$, $1 \leq k < n$.

So far, we considered only the conditional probability $\mathbf{P}\{X = x|Y = y\}$ that a random variable X takes the value x, given that Y takes the value y. When dealing with random variables, we often would like to know the expected value of X, given that Y takes the value y. We can define the *conditional expectation* in exactly the same manner as we define the expectation, where we simply replace the probabilities by the corresponding conditional probabilities.

Definition 1.9.4. *Let $X : \Omega \to D$ and $Y : \Omega \to D'$ be discrete random variables. The* conditional expectation *of $\varphi(X)$ given $Y = y$ is defined as*

$$\mathbf{E}(\varphi(X)|Y = y) = \sum_{x\in D}\varphi(x)\,\mathbf{P}\{X = x|Y = y\}.$$

Intuitively, the conditional expectation $\mathbf{E}(X|Y = y)$ is the average outcome of X over many repeated experiments, if we discard all experiments except those in which Y actually took the value y. As before, this intuition will be justified when we develop the law of large numbers.

Example 1.9.5. Suppose that we repeatedly throw a die until the first time that we get a 6. What is the expected number of 1s that we throw, given that we see a 6 for the first time in the kth throw?

To answer this question, let us define the following random variables:

$$\begin{aligned} Y &= \text{first time we throw a 6,} \\ X &= \text{number of 1s thrown,} \\ Z_i &= \text{outcome of the } i\text{th throw.} \end{aligned}$$

We are asked to compute $\mathbf{E}(X|Y=k)$. If the experiment terminates in the kth throw, then X is the number of ones in the first $k-1$ throws. Therefore

$$\begin{aligned} \mathbf{E}(X|Y=k) &= \mathbf{E}\left(\sum_{i=1}^{k-1} \mathbf{1}_{Z_i=1} \middle| Y=k \right) \\ &= \sum_{i=1}^{k-1} \mathbf{E}(\mathbf{1}_{Z_i=1}|Y=k) \\ &= \sum_{i=1}^{k-1} \mathbf{P}\{Z_i=1|Y=k\}. \end{aligned}$$

Note that $\{Y=k\} = \{Z_1 \neq 6, Z_2 \neq 6, \ldots, Z_{k-1} \neq 6, Z_k = 6\}$. Therefore

$$\begin{aligned} &\mathbf{P}\{Z_1=1|Y=k\} \\ &= \frac{\mathbf{P}\{Z_1=1, Y=k\}}{\mathbf{P}\{Y=k\}} \\ &= \frac{\mathbf{P}\{Z_1=1, Z_1\neq 6, \ldots, Z_{k-1}\neq 6, Z_k=6\}}{\mathbf{P}\{Z_1\neq 6, \ldots, Z_{k-1}\neq 6, Z_k=6\}} \\ &= \frac{\mathbf{P}\{Z_1=1, Z_1\neq 6\}\,\mathbf{P}\{Z_2\neq 6\}\cdots\mathbf{P}\{Z_{k-1}\neq 6\}\,\mathbf{P}\{Z_k=6\}}{\mathbf{P}\{Z_1\neq 6\}\cdots\mathbf{P}\{Z_{k-1}\neq 6\}\,\mathbf{P}\{Z_k=6\}} \\ &= \frac{\mathbf{P}\{Z_1=1\}}{\mathbf{P}\{Z_1\neq 6\}} = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5}, \end{aligned}$$

where we used that $Z_1, \ldots, Z_k$ are independent. Exactly the same computation shows that $\mathbf{P}\{Z_i=1|Y=k\} = \frac{1}{5}$ for every $i \leq k-1$. Therefore

$$\mathbf{E}(X|Y=k) = \sum_{i=1}^{k-1} \mathbf{P}\{Z_i=1|Y=k\} = \frac{k-1}{5}.$$

That is, if we repeat the experiment many times, and consider only those experiments where it took exactly k throws to get the first six, then the average number of ones we throw in those experiments is $(k-1)/5$.

In order to compute the conditional expectation of $\varphi(X)$ given the outcome of the random variable Y, we only need to know the conditional probabilities $\mathbf{P}\{X=i|Y=j\}$. This motivates the following definition.

Definition 1.9.6. *Let $X : \Omega \to D$, $Y : \Omega \to D'$ be discrete variables.* $(\mathbf{P}\{X = i|Y = j\})_{i\in D, j\in D'}$ *is the* conditional distribution *of X given Y.*

Given two random variables X and Y, we can readily compute the conditional distribution of X given Y. Just as for the notion of independence, however, conditional distributions are often used as a modelling assumption. In such cases, one is interested in going in the opposite direction: we must compute (unconditional) expectations of random variables on the basis of their conditional distributions. This is easily done, however: note that

$$\mathbf{P}\{X = x\} = \sum_y \mathbf{P}\{X = x, Y = y\} = \sum_y \mathbf{P}\{X = x|Y = y\}\,\mathbf{P}\{Y = y\},$$

and hence,

$$\mathbf{E}(X) = \sum_{x,y} x\,\mathbf{P}\{X = x|Y = y\}\,\mathbf{P}\{Y = y\} = \sum_y \mathbf{E}(X|Y = y)\,\mathbf{P}\{Y = y\}.$$

This is a very useful fact: it is often easier to compute first $\mathbf{E}(X|Y)$ (the conditional expectation of X given the outcome of the random variable Y), and then to compute the expectation of this random variable to obtain $\mathbf{E}(X)$, than it is to compute the expectation $\mathbf{E}(X)$ directly.

Example 1.9.7. There are two security lines at the airport. You are assigned randomly to one of these lines with equal probability. The number of people waiting in the first line is a random variable with mean μ, and the number of people waiting in the second line is a random variable with mean ν. Each person takes 5 minutes to go through the security check. What is the expected amount of time you have to wait until it is your turn to go through security?

To formalize the problem, let us define two random variables:

$$N = \text{length of your line},$$
$$Y = \text{the line you are in (1 or 2)}.$$

Our modelling assumptions are that $\mathbf{P}\{L = 1\} = \mathbf{P}\{L = 2\} = \frac{1}{2}$, and that $\mathbf{E}(N|L = 1) = \mu$ and $\mathbf{E}(N|L = 2) = \nu$. We can therefore compute the expected length of your line as

$$\mathbf{E}(N) = \mathbf{E}(N|L = 1)\mathbf{P}\{L = 1\} + \mathbf{E}(N|L = 2)\mathbf{P}\{L = 2\} = \frac{\mu + \nu}{2}.$$

As every person takes five minutes to get through security, the expected amount of time you have to wait in line is $\mathbf{E}(5N) = \frac{5}{2}(\mu + \nu)$.

. . .

We have by no means developed all the basic principles of probability in this chapter. However, we now have enough of a foundation in order to do many interesting things. Let us therefore take a break from theory building and start looking at some interesting problems. In the rest of this course, we will introduce new probabilistic concepts as they are needed.

2

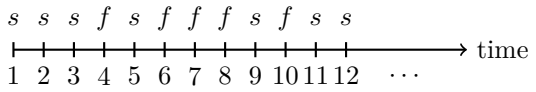
Bernoulli Processes

In this chapter, we will study the simplest kind of random process: a sequence of independent trials where success or failure occur with given probabilities. This will lead us investigate several interesting questions, and to develop notions such as the Binomial, geometric, and Poisson distributions. It will also finally allow us to give precise meaning to the frequency interpretation of probabilities in the form of the law of large numbers.

2.1 Counting successes and binomial distribution

A *Bernoulli variable* is a random variable that takes values in the set $\{0, 1\}$. A *Bernoulli process* is a sequence $X_1, X_2, X_3, \ldots$ of independent Bernoulli variables with $\mathbf{P}\{X_i = 1\} = p$, $\mathbf{P}\{X_i = 0\} = q = 1 - p$. These notions are named after Jacob Bernoulli (1655–1705), who first studied such processes in his famous treatise on probability *Ars Conjectandi*.[1]

You should think of a Bernoulli process as a sequence of independent trials. The ith trial is a success if $X_i = 1$ and a failure if $X_i = 0$. For example, think of a sequence of independent coin flips (success is heads, say), or a sequence of questions you are attempting to answer on an exam (success means you get the correct answer). One possible sequence of trials might look like this:



Even in this very simple model, there are many natural questions we could try to answer. For example:

[1] Jacob Bernoulli is one of a large family of famous mathematicians and physicists. Among them is Johann Bernoulli, his brother, who was an early researcher in the (then new) field of calculus, and the teacher of Leonhard Euler; his nephew Daniel Bernoulli, who developed the Bernoulli principle in fluid mechanics; etc.

- How many successes do we have by time n?
- How long must we wait for the first success? Between successes?
- Will we get success infinitely often? Or can we fail forever?

We will answer all these questions, and more, in the rest of this chapter.

In this section, we will investigate how many successes we have obtained by time n. Note that we can write

$$\begin{aligned} S_n &= \#\text{ successes by time } n \\ &= X_1 + X_2 + \ldots + X_n. \end{aligned}$$

As the outcome X_k of each trial is random, the number of successes S_n is also a random variable. What can we say about it?

Example 2.1.1. Let us compute $\mathbf{P}\{S_8 = 2\}$, the probability that we have exactly 2 successes in the first 8 trials. How should we go about doing this? First, let us write out all possible ways that we can get two successes in eight trials:

$$\{S_8 = 2\} = \{ssffffff\cdots, sfsffffff\cdots, fsfsffff\cdots, \ldots\}$$

(we write $\cdots$ to indicate that the outcomes of the remaining trials after time 8 are arbitrary). We can easily compute the probability of each of these distinct possibilities: for example, using independence of the trials, we have

$$\begin{aligned} \mathbf{P}\{ssffffff\cdots\} &= \mathbf{P}\{X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0, \ldots, X_8 = 0\} \\ &= \mathbf{P}\{X_1 = 1\}\mathbf{P}\{X_2 = 1\}\mathbf{P}\{X_3 = 0\}\cdots\mathbf{P}\{X_8 = 0\} = p^2q^6. \end{aligned}$$

We therefore obtain

$$\mathbf{P}\{S_8 = 2\} = p^2q^6 \times \#\{\text{sequences with 2 successes out of 8 trials}\}.$$

Now we have to count how many such sequences there are. This is the same question as: in how many ways can we put 2 balls in 8 bins? (Each success is a ball, and each trial is a bin.) We have 8 bins in which to put the first ball, and then 7 bins remaining in which to put the second ball. We therefore have $8 \cdot 7$ ways to put two balls into eight bins if we take the order (*first* ball, *second* ball) into account. However, the order of the balls does not matter to us: the same outcome $ssffffff\cdots$ is generated either by putting the first ball in the first bin and the second ball in the second bin, or by putting the first ball in the second bin and the second ball in the first bin. That is, each outcome appears twice in the above count. Thus there are $7 \cdot 8/2 = 28$ ways to put two balls in eight bins, and we have shown that $\mathbf{P}\{S_8 = 2\} = 28p^2q^6$.

There is nothing special about this example: it generalizes readily to any number of trials n and successes k. In exact analogy with the above computation, the number of possible ways that we can put k balls in n bins is

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots 1} \quad \begin{matrix} \leftarrow \#\text{ ordered ways to put } k \text{ balls in } n \text{ bins} \\ \leftarrow \#\text{ permutations of } k \text{ balls} \end{matrix}$$
$$= \frac{n!}{k!(n-k)!} =: \binom{n}{k},$$

and thus the probability of exactly k successes in n trials can be written as

$$\mathbf{P}\{S_n = k\} = \binom{n}{k} p^k q^{n-k}$$

in terms of the binomial coefficients $\binom{n}{k}$ (pronounced "n choose k").

Definition 2.1.2. *Let $p \in [0,1]$, $n \in \mathbb{N}$. A random variable Z with*

$$\mathbf{P}\{Z = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \qquad 0 \le k \le n$$

is called a binomial random variable *with parameters n and p.*

Remark 2.1.3. If a random variable Z has the binomial distribution with parameters n and p, we will often write this as $Z \sim \text{Binom}(n,p)$. The symbol $\sim$ means "has the distribution," and $\text{Binom}(n,p)$ indicates the binomial distribution with parameters n and p as defined in Definition 2.1.2.

Let us begin with a sanity check. The number of successes in n trials must obviously be an integer between 0 and n. Thus the probability that S_n is an integer between 0 and n must be one. Let us check this using Definition 2.1.2:

$$\sum_{k=0}^{n} \mathbf{P}\{S_n = k\} = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1,$$

where we have used the binomial theorem that you know from calculus. Thus everything works as expected. However, we did not need to use (or even know) the binomial theorem: we *know* that the probabilities must sum to one simply by the construction of our probability model. Thus it pays to have faith in probability theory; when we are confident that our model makes sense, we can often avoid difficult computations by using probabilistic common sense. (In the present case, this could even serve as a proof of the binomial theorem!)

Let us now consider a more interesting question: what is the average number of successes by time n? Evidently, we are asking to compute the expectation $\mathbf{E}(S_n)$. We can try to do this directly from the definition of expectation:

$$\begin{aligned}\mathbf{E}(S_n) &= \sum_{k=0}^{n} k\,\mathbf{P}\{S_n = k\} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k} \\ &= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} q^{n-k} \\ &= np \sum_{k=1}^{n} \mathbf{P}\{S_{n-1} = k-1\} = np.\end{aligned}$$

Thus the expected number of successes in n trials is $\mathbf{E}(S_n) = np$. We could have obtained this result much more easily, however, if we remember how we defined the number of successes S_n in the first place:

$$\mathbf{E}(S_n) = \mathbf{E}(X_1 + X_2 + \cdots + X_n) = \mathbf{E}(X_1) + \mathbf{E}(X_2) + \cdots + \mathbf{E}(X_n) = np.$$

Now the answer makes perfect intuitive sense: we have n trials, and each trial has probability p of success (that is, success should occur in a fraction p of the trials), so the expected number of successes is np. Both the above approaches are correct; but it is much easier and more intuitive to arrive at the answer by exploiting the fact that we understand what random model gives rise to the binomial random variable S_n, than to do a brute-force computation.

So far, we considered the number of successes S_n at a fixed time n. However, the number of successes changes over time as we perform more trials. We should therefore be able to say something about how the number of successes at different times are related. Let us illustrate this with two examples.

Example 2.1.4. The random variable $S_{18} - S_8$ can be written as

$$S_{18} - S_8 = \sum_{k=1}^{18} X_k - \sum_{k=1}^{8} X_k = \sum_{k=9}^{18} X_k.$$

We can therefore interpret $S_{18} - S_8$ as the number of successes that occur from time 9 through time 18. What is the distribution of this random variable? Evidently $S_{18} - S_8$ is the sum of 10 independent Bernoulli random variables with probability of success p. Similarly, S_{10} (the number of successes by time 10) is also the sum of 10 independent Bernoulli random variables with probability of success p. Thus $S_{18} - S_8$ must have the same distribution as S_{10}, even though these are clearly different random variables—in most experiments, there will be a different number of successes between times 1 and 10 than there are between times 9 and 18, but each number of successes will occur with the same probability in both these intervals. Thus we can compute, for example,

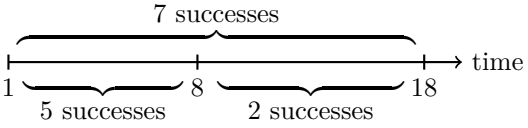
$$\mathbf{P}\{S_{18} - S_8 = 2\} = \mathbf{P}\{S_{10} = 2\} = \binom{10}{2} p^2 q^8.$$

Remark 2.1.5. The fact that S_{10} and $S_{18} - S_8$ have the same distribution has a natural interpretation. The Bernoulli process can be thought of as a robot that repeatedly performs the same task at every time, in which it succeeds with the same probability at every time. If we start observing the robot at time 1 and count successes in the first 10 trials, we observe S_{10}. If we only start observing the robot at time 9 and count successes in the first 10 trials, we observe $S_{18} - S_8$. Because the robot always does the same thing, however, it should not make any difference whether we start the clock at time 1 or at time 9: the random behavior of the robot is always the same. This explains why S_{10} and $S_{18} - S_9$ must have the same distribution.

Example 2.1.6. What is the probability that we see 5 successes by time 8, and 7 successes by time 18? That is, we would like to compute $\mathbf{P}\{S_8 = 5, S_{18} = 7\}$.

You might be tempted to try to use independence to split the probability. However, you cannot do that directly: the events $\{S_8 = 5\}$ and $\{S_{18} = 7\}$ are *not* independent! One way to see this is to note that $S_{18} \geq S_8$ (we must have at least as many successes by time 18 as we had by time 8). If we increase the number of successes by time 8, then we also increase the number of successes by time 18: thus the random variables S_{18} and S_8 cannot be independent.

How, then, do we do this computation? We must try to rewrite our problem in a way that allows us to use the independence of the different trials. To do this, let us argue as follows. Saying that we have 5 successes by time 8 and 7 successes by time 18 is completely equivalent to saying that we have 5 successes by time 8 and 2 successes between times 9 and 18:



In terms of events, this means that

$$\{S_8 = 5, S_{18} = 7\} = \{S_8 = 5, S_{18} - S_8 = 2\} = \{S_8 = 5\} \cap \{S_{18} - S_8 = 2\}.$$

What have we accomplished? Note that S_8 only depends on the trials up to time 8, while $S_{18} - S_8$ only depends on the trials at times 9 through 18. As the different trials are independent, the events $\{S_8 = 5\}$ and $\{S_{18} - S_8 = 2\}$ *are* independent. We can therefore compute

$$\mathbf{P}\{S_8 = 5, S_{18} = 7\} = \mathbf{P}\{S_8 = 5\}\,\mathbf{P}\{S_{18} - S_8 = 2\} = \binom{8}{5} p^5 q^3 \binom{10}{2} p^2 q^8.$$

Note that even though our original question was not stated in terms of independent events, the key to the solution was to reformulate the problem in terms of independent events: that is, we had to *find* the independence.

2.2 Arrival times and geometric distribution

In this section, we will be interested in how long we have to wait until a success occurs (how many questions must you attempt on the exam before you get one right?) More generally, let us define the random variables

$$T_k = k\text{th time of success.}$$

For example, for the sequence of outcomes $\omega = sfssfffsfss\cdots$, we have $T_1(\omega) = 1$, $T_2(\omega) = 3$, $T_3(\omega) = 4$, $T_4(\omega) = 8$, etc.

What can we say about the distribution of T_k? Let us take a moment to think about what it means that the kth success occurs at time n. If this is the case, then we certainly must have success at time n; moreover, there must be exactly $k-1$ successes before time n. Conversely, if there are $k-1$ successes before time n and success at time n, then the kth success occurs at time n. In mathematical terms, we have just argued that

$$\{T_k = n\} = \{X_n = 1, S_{n-1} = k-1\}.$$

Now note that the number of successes S_{n-1} before time n depends only on the trials before time n, and is therefore independent of the trial X_n at time n. We can therefore immediately compute the distribution of T_k:

$$\mathbf{P}\{T_k = n\} = \mathbf{P}\{X_n = 1\}\,\mathbf{P}\{S_{n-1} = k-1\} = \binom{n-1}{k-1} p^k q^{n-k}, \quad n \geq k$$

(note that $T_k \geq k$: the kth success cannot occur before the kth trial!) For some obscure reason, a random variable with this distribution is called a *negative binomial* random variable with parameters k and p.

Remark 2.2.1. In general, there is no recipe you can follow to compute the distribution of a random variable Z. You simply must sit down and think about how to approach the particular random variable you are trying to study. In many cases, a clever way of writing the event $\{Z = z\}$ will allow you to simplify the problem, as we have done here for T_k. In other cases, computing the distribution involves a direct computation and/or some combinatorics, as we have seen in the case of the binomial random variable S_n.

How long must we wait until the first success occurs? This is just the special case T_1 of the random variables considered above. Setting $k = 1$ in the above formula, we immediately find $\mathbf{P}\{T_1 = n\} = q^{n-1}p$. We could have easily obtained this result by reasoning directly. If the first success occurs at time n, this is equivalent to saying that we first fail $n-1$ times, and then succeed. The probability that this happens is

$$\begin{aligned}\mathbf{P}\{T_1 = n\} &= \mathbf{P}\{X_1 = 0, X_2 = 0, \ldots, X_{n-1} = 0, X_n = 1\} \\ &= \mathbf{P}\{X_1 = 0\}\mathbf{P}\{X_2 = 0\}\cdots\mathbf{P}\{X_{n-1} = 0\}\mathbf{P}\{X_n = 1\} = q^{n-1}p.\end{aligned}$$

Thus the expression we have obtained for the distribution of T_1 makes perfect sense. This distribution appears in many problems, so we give it a name.

Definition 2.2.2. *Let $p \in [0,1]$. A random variable Z with distribution*

$$\mathbf{P}\{Z = n\} = (1-p)^{n-1}p, \qquad 1 \leq n < \infty$$

is called a geometric random variable *with parameter p ($Z \sim \text{Geom}(p)$).*

Let us investigate some basic properties of geometric random variables. The first question we must answer is whether it is possible that $T_1 = \infty$? This would mean that success never occurs or, in other words, that we fail forever. To compute the probability that this happens, note that

$$\begin{aligned}\mathbf{P}\{T_1 = \infty\} &= \mathbf{P}\{X_1 = 0, X_2 = 0, X_3 = 0, \ldots\} \\ &= \mathbf{P}\{X_1 = 0\}\mathbf{P}\{X_2 = 0\}\mathbf{P}\{X_3 = 0\}\cdots = q^\infty.\end{aligned}$$

There are two possibilities:

- If $q = 1$, then $\mathbf{P}\{T_1 = \infty\} = 1$. This makes perfect sense: if the probability of failure is 1, then we must fail forever.
- If $q < 1$, then $\mathbf{P}\{T_1 = \infty\} = 0$. That is, if there is any chance of success, then we will eventually succeed.

Of course, the same must be true for failures: if there is any chance of failure, then we will eventually fail. This statement is known as Murphy's law (anything that can go wrong will go wrong!)

Remark 2.2.3. You might find this statement absurd. Every time you walk to class, there is some (hopefully small) probability that you will be attacked by an angry bird and die. Therefore, if you keep coming to class, you will eventually get attacked by a bird and die. To the best of my knowledge, this has never happened in the history of Princton. However, Murphy's law only says that if we repeat the trials *infinitely* often, then eventually success will occur. It does not state how long it will take (we will compute this below). When the probability p is very small, as in this example, the amount of time it takes until success happens is astronomically long. As we only live for a finite amount of time, there is no contradiction between Murphy's law and reality: the probability that success will occur in a lifetime is exceedingly small.

Let us consider a different formulation of Murphy's law. If $q < 1$, we have seen that $T_1 < \infty$. Thus $T_1 = 1$, or $T_1 = 2$, or $T_1 = 3$, or ..., so we must have

$$1 = \sum_{n=1}^{\infty} \mathbf{P}\{T_1 = n\} = \sum_{n=1}^{\infty} q^{n-1} p.$$

If we divide by $p = 1 - q$, this expression can be written as

$$\frac{1}{1-q} = \sum_{k=0}^{\infty} q^k.$$

This expression for a geometric series is one that you know from calculus. However, we used no calculus in this proof, only probability: you could view this calculus identity as nothing other than Murphy's law in disguise.

Now that we know that success will eventually happen, the next most natural question is how long we must wait on average until the first success occurs. That is, we would like to compute the expectation $\mathbf{E}(T_1)$. By the definition of the expectation, we can write this as

$$\mathbf{E}(T_1) = \sum_{n=1}^{\infty} n \cdot \mathbf{P}\{T_1 = n\} + \infty \cdot \mathbf{P}\{T_1 = \infty\}.$$

Assuming that $q < 1$, the second term is zero, and we can use a calculus trick to perform the computation: as $nq^{n-1} = dq^n/dq$, we have

$$\mathbf{E}(T_1) = \sum_{n=1}^{\infty} nq^{n-1}p = p\,\frac{d}{dq}\sum_{n=1}^{\infty} q^n = p\,\frac{d}{dq}\frac{q}{1-q} = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

That is, the expected time we must wait until success is one divided by the probability of success. In particular, if the probability of success is small, we must wait a long time for success to happen.

Remark 2.2.4. There is an intuitive way to remember this result. If you think of p as the "fraction of experiments in which success occurs", then we have (on average) p successes/unit time. That means that the amount of time/success should be $1/p$. However, this physicist's argument should be taken with a grain of salt: it is easy to remember, but a careful justification requires us to compute the expectation of a geometric random variable as we did above.

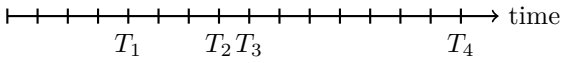
So far we have talked mostly about how long we must wait for the first success. We may be interested in more general questions; for example, how long do we have to wait between successes, or how often do we get successes? To study such problems, let us first consider a simple example.

Example 2.2.5. Let us try to compute the probability that the first success comes in the 5th trial, the following success comes after 3 more trials, the

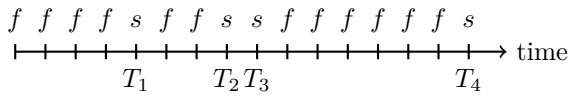
following success comes after 1 more trial, and the following success comes after 7 more trials. That is, we want to compute the following probability:

$$\mathbf{P}\{T_1 = 5,\ T_2 - T_1 = 3,\ T_3 - T_2 = 1,\ T_4 - T_3 = 7\} = ?$$

(note that $T_k - T_{k-1}$ is the number ot trials between the $(k-1)$st and kth success). To compute this probability, let us draw a picture:



We have success at every time T_k (this is the time of the kth success), and between the times of success we must have failure. So the only way in which the above event can occur is if we have the following sequence of outcomes:



The probability that this happens is therefore

$$\begin{aligned}
&\mathbf{P}\{T_1 = 5,\ T_2 - T_1 = 3,\ T_3 - T_2 = 1,\ T_4 - T_3 = 7\} \\
&= \mathbf{P}\{fffffsffssfffffffs\cdots\} = q^4 p\, q^2 p\, q^0 p\, q^6 p \\
&= \mathbf{P}\{T_1 = 5\}\,\mathbf{P}\{T_1 = 3\}\,\mathbf{P}\{T_1 = 1\}\,\mathbf{P}\{T_1 = 7\}.
\end{aligned}$$

The last line is just the observation that the answer can be written as a product of geometric probabilities. We will shortly argue that this is no coincidence!

There is nothing special about this example: we have simply worked out one particular case for sake of illustration. By exactly the same argument, you can easily convince yourself that for every $k \geq 1$ and $n_i \geq 1$

$$\begin{aligned}
&\mathbf{P}\{T_1 = n_1,\ T_2 - T_1 = n_2,\ \ \ldots,\ T_k - T_{k-1} = n_k\} \\
&\qquad = \mathbf{P}\{T_1 = n_1\}\,\mathbf{P}\{T_1 = n_2\}\cdots\mathbf{P}\{T_1 = n_k\}.
\end{aligned}$$

What does this formula mean? First of all, note that (why?)

$$\begin{aligned}
\mathbf{P}\{T_k - T_{k-1} = n_k\} &= \sum_{n_1,\ldots,n_{k-1}} \mathbf{P}\{T_1 = n_1,\ \ldots,\ T_k - T_{k-1} = n_k\} \\
&= \mathbf{P}\{T_1 = n_k\},
\end{aligned}$$

so that $T_k - T_{k-1}$ has the same distribution as T_1. Moreover the probability that $\{T_1 = n_1\}$ *and* $\{T_2 - T_1 = n_2\}$ *and* ... occur is the product of the probabilities. We have therefore shown the following:

- The times $T_1,\ T_2-T_1,\ T_3-T_2,\ \ldots$ between successes are *independent*.
- Each time $T_k - T_{k-1}$ is *geometric* with parameter p.

These fundamental properties of the times between consecutive successes are extremely useful: they allow us to do many computations.

Remark 2.2.6. These conclusions are not at all surprising. As in Remark 2.1.5, let us think of the Bernoulli process as a robot that repeatedly performs the same task at every time with the same probability of success. If we start our stopwatch at time 1, then the time we measure until the first success is T_1. We could also start our stopwatch after the first time of success, however, in which case the time we measure until the next success is T_2-T_1. The point is that the robot does not care when we start the clock: it is doing the same thing at every time regardless of when we start looking at it. Thus $T_2 - T_1$ and T_1 should have the same distribution: these random variables correspond to watching the robot for the same length of time doing the same thing, just started at two different times. Moreover, as the robot is performing independent trials, whatever happens *after* the first success should be independent of how long we had to wait to achieve the first success. What we have done above is to show carefully that this intuition is in fact captured by our probability model.

Let us give some simple illustrations of how we can use these properties. First, consider the following question. Murphy's law states that if there is any chance of success, then success will occur eventually. However, could it be that we eventually get success, and then succeed never again thereafter? The answer is no: if $q < 1$, then we must succeed *infinitely often*. Indeed, note that

$$\begin{aligned}
&\mathbf{P}\{\text{success occurs infinitely often}\} \\
&= \mathbf{P}\{T_1 < \infty,\ T_2 - T_1 < \infty,\ T_3 - T_2 < \infty,\ \ldots\} \\
&= \mathbf{P}\{T_1 < \infty\}\,\mathbf{P}\{T_1 < \infty\}\,\mathbf{P}\{T_1 < \infty\}\cdots \\
&= 1 \cdot 1 \cdot 1 \cdots = 1,
\end{aligned}$$

where we have used the two properties given above of the times between consecutive successes. Exactly the same conclusion follows for failure: if anything can go wrong, then it will go wrong over and over again. This depressing conclusion is a stronger version of Murphy's law.

Here is another application. We have not yet computed the expected time $\mathbf{E}(T_k)$ at which the kth success occurs. We did compute the distribution of T_k, so we could try to compute the expectation by explicitly summing up the probabilities in the definition $\mathbf{E}(T_k) = \sum_{n=k}^{\infty} n\,\mathbf{P}\{T_k = n\}$. This will give the right answer, but is tedious! Here is a much simpler solution. Note that

$$T_k = T_1 + (T_2 - T_1) + (T_3 - T_2) + \cdots + (T_k - T_{k-1}) :$$

that is, the kth success occurs after we first wait until the first success, then we wait the additional amount of time until the second success, etc. As each of the terms in this sum is geometric with parameter p, we immediately obtain

$$\mathbf{E}(T_k) = \mathbf{E}(T_1) + \mathbf{E}(T_2 - T_1) + \cdots + \mathbf{E}(T_k - T_{k-1}) = \frac{k}{p}.$$

Once again, we see that the right probabilistic idea greatly simplifies our math.

2.3 The law of large numbers

Consider a given event A with probability $\mathbf{P}(A)$. In Chapter 1, we repeatedly invoked the follwing *intuitive* notion of probability:

> If we repeat the random experiment many times, the probability $\mathbf{P}(A)$ is the fraction of these experiments in which A occurs.

This statement is too vague to have a precise meaning. To give it an unambiguous interpretation, we need a more precise mathematical formulation. We are now finally ready to give a proper formulation of this idea.

In order to reason rigorously about repeated experiments, imagine a new "super-experiment" that consists of repeating the original random experiment infinitely many times. In this super-experiment we have events $A_1, A_2, A_3, \ldots$, where A_i denotes the event that A occurs in the ith time that we repeat our original experiment. Because we are repeating the *same* experiment, every event A_i must have the same probability as the original event A, that is, $\mathbf{P}(A_i) = \mathbf{P}(A)$. Moreover, if you think about our intuitive notion of "repeating" an experiment, you will readily agree that what we really mean by this is that $A_1, A_2, A_3, \ldots$ are *independent*: that is, we perform each repetition of the experiment independently from the rest.

How often does A occur in these repeated experiments? In the first n experiments, the number of times that A occurs is (recall Example 1.8.10)

$$\text{Number of times } A \text{ occurs in first } n \text{ experiments} = S_n = \sum_{k=1}^{n} \mathbf{1}_{A_k}.$$

This random variable is already familiar to us! Notice that the indicator functions $\mathbf{1}_{A_k}$ are independent Bernoulli variables: they only take the values 0 and 1. Thus $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \mathbf{1}_{A_3}, \ldots$ is none other than a Bernoulli process with $\mathbf{P}\{\mathbf{1}_{A_k} = 1\} = \mathbf{P}(A_k) = \mathbf{P}(A)$. In particular, this implies that the number of times that A occurs in the the first n experiments has the distribution

$$S_n \sim \text{Binom}(n, \mathbf{P}(A)).$$

This allows us to compute the probability $\mathbf{P}\{S_n = k\}$ that the event A will occur precisely k times if we repeat the experiment n times.

Example 2.3.1 (Throwing a die). You throw a die 7 times; what is the probability that you will throw 6 or 3 exactly twice? You can think of this problem as a repeated experiment: in each experiment, we throw a die, and we look at the event A that the outcome was either 6 or 3. The probability of this event is $\mathbf{P}(A) = \frac{2}{6} = \frac{1}{3}$. We now repeat this experiment 7 times. Then the probability that A occurs exactly twice in these seven experiments is

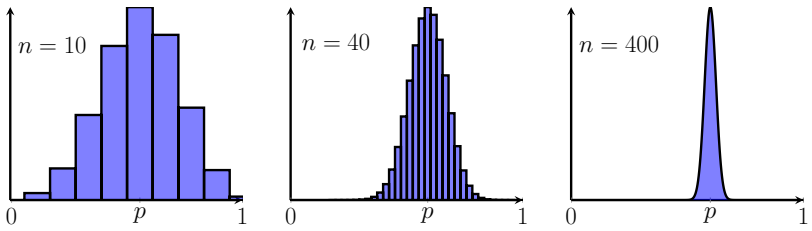
$$\mathbf{P}\{\text{Binom}(7, \tfrac{1}{3}) = 2\} = \binom{7}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^5.$$

There are many different ways to arrive at this answer. However, here we got there very quickly by thinking of this problem as a repeated experiment.

In our intuitive interpretation of probability, we are interested in the fraction of the repeated experiments in which A occurs. Of course, the fraction of experiments in which A occurs is just the number of experiments in which A occurs divided by the total number of experiments, that is,

$$\text{Fraction of times } A \text{ occurs in first } n \text{ experiments} = \frac{S_n}{n} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{A_k}.$$

What does this random variable look like? As we know the distribution of S_n, we also know the distribution $\mathbf{P}\{S_n/n = x\} = \mathbf{P}\{S_n = nx\}$ $(0 \le x \le 1)$. In the following figure, we plot a histogram of this distribution for a few different values of the total number of experiments n (here $p := \mathbf{P}(A) = 0.6$):



You can see in these pictures that something interesting is going on. For every total number n of experiments, the fraction S_n/n of experiments in which A occurs is $\mathbf{P}(A)$ *on average*. We can immediately see this without pictures:

$$\mathbf{E}\left(\frac{S_n}{n}\right) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}(A_k) = \mathbf{P}(A).$$

However, as we increase the total number of experiments n, the distribution of S_n/n changes significantly. When n is small, S_n/n is quite random: even though its expectation is $\mathbf{P}(A)$, in any given sequence of n repeated experiments the fraction of times in which A occurs can be quite far from $\mathbf{P}(A)$. However, as we increase the total number of experiments n, the distribution

becomes less and less random: when n is very large, $S_n/n \approx \mathbf{P}(A)$ almost all the time. We therefore expect that if we perform *infinitely many* experiments, the fraction of experiments in which A occurs is *exactly* $\mathbf{P}(A)$. This is precisely the correct mathematical interpretation of our intuition about probabilities: the statement "if the experiment is repeated many times, then the fraction of experiments in which A occurs is $\mathbf{P}(A)$" should be interpreted as follows.

Theorem 2.3.2 (Law of large numbers). *With probability one*

$$\lim_{n\to\infty} \frac{S_n}{n} = \mathbf{P}(A).$$

Remark 2.3.3. In the previous chapter, the idea that probabilities are frequencies was only a guiding intuition: we did not define probability in this manner, but only used this intuition as inspiration to create a mathematically unambiguous definition. We now see that, by the law of large numbers, the frequency interpretation of probability actually follows as a consequence of our theory! This successful recovery of our real-world experience is a very convincing sign that we have created the "right" theory of probability.

We are going to prove (a form of) the law of large numbers. Before we do so, however, let us consider two other important notions for which we introduced intuitive interpretations in the previous chapter: expectations and conditional probabilities. In each case, we can derive the correct mathematical statement of corresponding intuitive interpretation from Theorem 2.3.2.

Example 2.3.4 (The law of large numbers for random variables). Recall the intuitive notion of the expectation of a random variable:

> If we repeat the random experiment many times, the expectation $\mathbf{E}(X)$ is the average of the outcomes of X over all these experiments.

The precise formulation of this notion is as follows. Let $X_1, X_2, X_3, \ldots$ be independent random variables such that X_i has the same distribution as X for every i. You should think of X_i as the outcome of the random variable X in the ith experiment. The following form of the law of large numbers shows that the average over all these random variables is precisely $\mathbf{E}(X)$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mathbf{E}(X).$$

The simplest way to derive this result is to use the following very useful property of discrete random variables $X : \Omega \to D$:

$$X = \sum_{u \in D} u \, \mathbf{1}_{\{X=u\}}.$$

Why is this true? The indicator function $\mathbf{1}_{\{X=u\}}$ is equal to one if $X = u$, and is zero otherwise. Therefore, if Tyche chooses an an outcome $\omega \in \Omega$ of our random experiment such that $X(\omega) = r$, then the term $u = r$ in the above sum is equal to r and all the remaining terms are zero. As this holds for every possible outcome ω, the above identity is established.

Let us now use this identity to compute the average of the random variables X_k. As we can always rearrange the order of taking sums, we have

$$\frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\sum_{k=1}^{n}\sum_{u \in D} u \, \mathbf{1}_{\{X_k=u\}} = \sum_{u \in D} u \, \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_k=u\}}.$$

The average inside the outer sum is nothing other than the fraction of experiments in which X takes the value u. By the law of large numbers

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_k=u\}} = \mathbf{P}\{X = u\}.$$

Therefore

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \lim_{n\to\infty} \sum_{u \in D} u \, \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_k=u\}} = \sum_{u \in D} u \, \mathbf{P}\{X = u\} = \mathbf{E}(X),$$

and thus the fact that expectations are averages follows from Theorem 2.3.2.

Example 2.3.5 (The law of large numbers for conditional probability). Recall the intuitive notion of the conditional probability of A given B:

> If we repeat the random experiment many times, and discard all experiments except those in which B occurs, the fraction of the remaining experiments in which A occurs is $\mathbf{P}(A|B)$.

Let us make this precise. Consider $A_1, A_2, A_3, \ldots$ and $B_1, B_2, B_3, \ldots$, where A_i (or B_i) is the event that A (or B) occurs in the ith time that we repeat the experiment. Suppose we consider the first n experiments and retain only those in which B occurs. Then the fraction of these experiments in which A occurs is simply the number of experiments in which B *and* A occur, divided by the total number of experiments in which B occurs. Therefore

$$\frac{\#\text{ of first } n \text{ experiments where } B \text{ and } A \text{ occur}}{\#\text{ of first } n \text{ experiments where } B \text{ occurs}} = \frac{\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{A_k \cap B_k}}{\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{B_k}}.$$

If we let $n \to \infty$, we can apply the law of large numbers separately to the numerator and to the denominator:

$$\lim_{n\to\infty} \frac{\#\text{ of first } n \text{ experiments where } B \text{ and } A \text{ occur}}{\#\text{ of first } n \text{ experiments where } B \text{ occurs}} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \mathbf{P}(A|B).$$

This gives precise meaning to our intuitive notion of conditional probability. It is left as an exercise to make our intuition precise for conditional expectation.

In order to prove the law of large numbers, we have to find a way to make precise the idea that the random variable S_n/n becomes less and less random as we increase n (so far, we have only seen some pictures that suggest that this is the case). To this end, we need to introduce a measure of "how random" a random variable is. Before we can do this, let us begin with a simpler question: when is a random variable X *not* random? Clearly, this is the case when X takes only one possible value x with probability $\mathbf{P}\{X = x\} = 1$ (that is, we see the same outcome x in every experiment). In this case, we can compute

$$\mathbf{E}(X) = x\,\mathbf{P}\{X = x\} = x.$$

Hence X is *not* random when it is always equal to its expectation $X = \mathbf{E}(X)$. With this idea in mind, it makes sense to think of X as being "almost nonrandom" when $X \approx \mathbf{E}(X)$, while X is "very random" if X is often far from $\mathbf{E}(X)$. A natural measure of "how random" X is should therefore quantify the difference between X and its expectation $\mathbf{E}(X)$. One of the most useful measures of this kind is the *variance* of a random variable.

Definition 2.3.6. *The* variance *of a random variable X is defined as*

$$\mathrm{Var}(X) := \mathbf{E}((X - \mathbf{E}(X))^2).$$

That is, $\mathrm{Var}(X)$ *is the mean square difference between X and* $\mathbf{E}(X)$.

The larger the variance $\mathrm{Var}(X)$, the "more random" is the random variable X. The utility of the notion of variance stems from the fact that it has many nice properties. The following properties are particularly important:

a. $\mathrm{Var}(X) \geq 0$.

b. $\mathrm{Var}(X) = 0$ if and only if X is nonrandom.

c. $\mathrm{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2$.

d. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if $X \perp\!\!\!\perp Y$.

e. $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$ for $a \in \mathbb{R}$.

Let us discuss why each of these properties is true.

a. It is easy to see that the variance is nonnegative: $(X-\mathbf{E}(X))^2$ is a nonnegative random variable (it is a square!), so its expectation is also nonnegative. Thus the minimal possible value of the variance is $\mathrm{Var}(X) = 0$.

b. When X is nonrandom, that is, $X = \mathbf{E}(X)$, the definition of variance immediately implies that X has zero variance. We claim that the converse is also true: a random variable with zero variance must be nonrandom. To see why, note that $(X - \mathbf{E}(X))^2 \geq 0$ and $\mathbf{E}((X - \mathbf{E}(X))^2) = 0$ imply that $(X - \mathbf{E}(X))^2 = 0$ with probability one by the following useful remark.

Remark 2.3.7. The following fact is often useful: if Z is a random variable such that $Z \geq 0$ and $\mathbf{E}(Z) = 0$, then $Z = 0$ (more precisely, $\mathbf{P}\{Z = 0\} = 1$). This statement is common sense: the average of nonnegative numbers can be zero only if all these numbers are zero (draw a picture!) Be careful, however: at some point late at night after you have been drinking, you will be tempted to solve a problem using that $\mathbf{E}(Z) = 0$ implies $Z = 0$ in a case where Z is not nonnegative. This is *not* correct! (Think of the random variable Z that takes the values $+1, -1$ with probability $\frac{1}{2}$.)

c. To establish the third property, we expand the square in Definition 2.3.6:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbf{E}(X^2 - 2X\mathbf{E}(X) + \mathbf{E}(X)^2) \\ &= \mathbf{E}(X^2) - (2\mathbf{E}(X))\mathbf{E}(X) + \mathbf{E}(X)^2 \\ &= \mathbf{E}(X^2) - \mathbf{E}(X)^2.\end{aligned}$$

Here we have used that $\mathbf{E}(X)^2$ is nonrandom, so $\mathbf{E}(\mathbf{E}(X)^2) = \mathbf{E}(X)^2$; and that $2\mathbf{E}(X)$ is nonrandom, so $\mathbf{E}(2X\mathbf{E}(X)) = (2\mathbf{E}(X))\mathbf{E}(X)$. While its interpretation is less intuitive than that of Definition 2.3.6, this formula for the variance is often easy to use in computations.

Example 2.3.8. Let X be a Bernoulli variable with $\mathbf{P}\{X = 1\} = p$. Then the expectation of X is $\mathbf{E}(X) = p$. What is its variance? Note that $0^2 = 0$ and $1^2 = 1$, so $X^2 = X$. Therefore, in this special case,

$$\mathrm{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2 = \mathbf{E}(X) - \mathbf{E}(X)^2 = p(1-p).$$

The variance is zero when $p = 0$ or $p = 1$. This makes perfect sense: if the outcome of the Bernoulli variable is always failure, then it is nonrandom; and similiarly if the outcome is always success, then it is nonrandom. On the other hand, the variance is maximized for $p = \frac{1}{2}$: the "most random" Bernoulli variable is the one where success and failure occur with equal probability. This also makes intuitive sense, as this is the case in which it is the hardest to predict in advance whether we will succeed or fail.

d. Let X and Y be *independent* random variables. Then

$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathbf{E}((X+Y-\mathbf{E}(X)-\mathbf{E}(Y))^2)\\ &= \mathbf{E}((X-\mathbf{E}(X))^2)+\mathbf{E}((Y-\mathbf{E}(Y))^2)\\ &\qquad +\mathbf{E}(2(X-\mathbf{E}(X))(Y-\mathbf{E}(Y)))\\ &= \operatorname{Var}(X)+\operatorname{Var}(Y)+2\,\mathbf{E}(X-\mathbf{E}(X))\,\mathbf{E}(Y-\mathbf{E}(Y))\\ &= \operatorname{Var}(X)+\operatorname{Var}(Y).\end{aligned}$$

Here we have used $\mathbf{E}((X-\mathbf{E}(X))(Y-\mathbf{E}(Y))) = \mathbf{E}(X-\mathbf{E}(X))\,\mathbf{E}(Y-\mathbf{E}(Y))$ as X and Y are independent, and $\mathbf{E}(X-\mathbf{E}(X)) = \mathbf{E}(X)-\mathbf{E}(X) = 0$.

Remark 2.3.9. Recall that $\mathbf{E}(X+Y) = \mathbf{E}(X)+\mathbf{E}(Y)$ for *any* random variables X, Y. However, the variance does *not* necessarily satisfy $\operatorname{Var}(X+Y) = \operatorname{Var}(X)+\operatorname{Var}(Y)$ when X and Y are not independent!

e. Finally, let X be a random variable and $a \in \mathbb{R}$. Then

$$\operatorname{Var}(aX) = \mathbf{E}((aX-\mathbf{E}(aX))^2) = a^2\,\mathbf{E}((X-\mathbf{E}(X))^2) = a^2\operatorname{Var}(X).$$

Do not confuse this with the property of expectations $\mathbf{E}(aX) = a\,\mathbf{E}(X)$!

Remark 2.3.10. You might be wondering what is so special about the variance as a "measure of randomness" of a random variable. Would a quantity such as $\mathbf{E}(|X-\mathbf{E}(X)|)$, for example, not serve equally well as a measure of randomness? This is perfectly reasonable: there is indeed nothing special about the variance, it is just one of many possible notions of "how random" a random variable is. What makes the variance useful, however, is that it satisfies the nice properties that we have developed. For example, the perfectly reasonable measure of randomness $\mathbf{E}(|X-\mathbf{E}(X)|)$ does not satisfy any analogue of the property $\operatorname{Var}(X+Y) = \operatorname{Var}(X)+\operatorname{Var}(Y)$ when X, Y are independent. It is because the variance has so many nice properties that it is particularly convenient to use it as a measure of randomness.

We are now ready to return to the law of large numbers. We recall that the *expected* fraction of the first n experiments in which A occurs is given by

$$\mathbf{E}\left(\frac{S_n}{n}\right) = \mathbf{E}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{A_k}\right) = \frac{1}{n}\sum_{k=1}^{n}\mathbf{P}(A_k) = \mathbf{P}(A).$$

We would like to show that the fraction of experiments is *exactly* $\mathbf{P}(A)$ when we perform an infinite number of experiments. That is, we must show that S_n/n becomes nonrandom as the number of experiments $n \to \infty$. We can now achieve this using the nice properties of the variance:

$$\operatorname{Var}\left(\frac{S_n}{n}\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{A_k}\right) = \frac{1}{n^2}\sum_{k=1}^{n}\operatorname{Var}(\mathbf{1}_{A_k}) = \frac{\mathbf{P}(A)(1-\mathbf{P}(A))}{n}.$$

It follows immediately that

$$\lim_{n\to\infty} \mathbf{E}\left(\left(\frac{S_n}{n} - \mathbf{P}(A)\right)^2\right) = \lim_{n\to\infty} \mathrm{Var}\left(\frac{S_n}{n}\right) = 0.$$

This shows that S_n/n gets closer and closer to $\mathbf{P}(A)$, and becomes less and less random, as we increase the number of experiments n. This is, in essence, all there is to the law of large numbers!

Some more advanced ideas

The material in the rest of this section is more advanced than this course, and is not needed in the sequel. It is only included for those of you who are curious: you can safely skip to the next section.

The law of large numbers states that

$$\lim_{n\to\infty} \frac{S_n}{n} = \mathbf{P}(A).$$

If we want to be precise, we have to be careful by what we mean by "lim". If you look carefully at what we have actually proved, you will see that we have shown that the expectation of the square difference between S_n/n and $\mathbf{P}(A)$ goes to zero. This type of convergence is known as *mean square convergence*, and the corresponding form of the law of large numbers is called the *weak law of large numbers*. This result is quite satisfying: its interpretation is that when the number of experiments is very large, then the fraction of experiments in which A occurs is, on average, very close to $\mathbf{P}(A)$.

However, we might expect a somewhat stronger result to hold. If we genuinely were to perform an infinite number of experiments, we expect that the limit of S_n/n as $n \to \infty$ is exactly $\mathbf{P}(A)$, *every* time we run this experiment. We therefore expect that the limit in the law of large numbers in fact holds *with probability one*, not just in mean square. It is not entirely obvious that this does not follow from what we have shown. You might expect that

$$\mathbf{E}\left(\lim_{n\to\infty}\left(\frac{S_n}{n} - \mathbf{P}(A)\right)^2\right) = \lim_{n\to\infty} \mathbf{E}\left(\left(\frac{S_n}{n} - \mathbf{P}(A)\right)^2\right) = 0,$$

from which it would follow that

$$\lim_{n\to\infty}\left(\frac{S_n}{n} - \mathbf{P}(A)\right)^2 = 0 \quad \text{with probability 1.}$$

It turns out that you can often, but not always, exchange a limit and an expectation: there are some cases in which mean square convergence does not imply convergence with probability one (in a rigorous mathematics course on probability theory, much time is devoted to understanding conditions under which one form of convergence implies another). It is therefore necessary to work a little bit harder to show that the limit in the law of large numbers

does indeed hold with probability one. This result, called the *strong law of large numbers*, is the one that we actually stated in Theorem 2.3.2.

The easiest way to prove the strong law of large numbers is by exploiting a useful trick that is illustrated in the following example.

Example 2.3.11. Let $Z_1, Z_2, Z_3, \ldots$ be *nonnegative* random variables such that

$$\sum_{k=1}^{\infty} \mathbf{E}(Z_k) < \infty.$$

By the linearity of the expectation, this implies

$$\mathbf{E}\left(\sum_{k=1}^{\infty} Z_k \right) = \sum_{k=1}^{\infty} \mathbf{E}(Z_k) < \infty,$$

and so we must have

$$\sum_{k=1}^{\infty} Z_k < \infty \quad \text{with probability 1 :}$$

after all, if a nonnegative random variable takes the value infinity with positive probability, then its expectation would have to be infinity.

But if the sum of a sequence of nonnegative numbers is finite, then this sequence must converge to zero (why?). Therefore

$$\sum_{k=1}^{\infty} \mathbf{E}(Z_k) < \infty \quad \text{implies} \quad \lim_{k \to \infty} Z_k = 0 \quad \text{with probability 1.}$$

The sneaky trick that we have pulled off here is that, even though we cannot always exchange an expectation and a limit, we can always exchange an expectation and a sum (the expectation is linear).

We would really like to apply this example to the random variables

$$Z_n = \left(\frac{S_n}{n} - \mathbf{P}(A) \right)^2.$$

We already computed $\mathbf{E}(Z_n)$, but unfortunately

$$\sum_{n=1}^{\infty} \mathbf{E}(Z_n) = \mathbf{P}(A)(1 - \mathbf{P}(A)) \sum_{n=1}^{\infty} \frac{1}{n} = \infty \, !$$

So this was a nice try, but it does not work. Nonetheless, this is a good idea. As is often the case with good ideas, even if they do not work on the first attempt, we can fix what is wrong and make them work after all.

In the present case, the problem we had was that $\mathbf{E}(Z_n) \sim n^{-1}$ which is not a summable sequence. However, it seems like a pretty good guess that if

$\mathbf{E}(Z_n) \sim n^{-1}$, then $\mathbf{E}(Z_n^2) \sim n^{-2}$. If that is indeed true, then we would get a summable sequence and life is good after all! Let us try it. Define

$$Y_n = Z_n^2 = \left(\frac{S_n}{n} - \mathbf{P}(A) \right)^4.$$

To compute the expectation, we expand the power as follows:

$$\begin{aligned} \mathbf{E}(Y_n) &= \mathbf{E}\left(\left(\frac{1}{n} \sum_{k=1}^{n} X_k \right)^4 \right) \\ &= \frac{1}{n^4} \sum_{k_1,k_2,k_3,k_4=1}^{n} \mathbf{E}(X_{k_1} X_{k_2} X_{k_3} X_{k_4}), \end{aligned}$$

where we denote $X_k = \mathbf{1}_{A_k} - \mathbf{P}(A)$ for simplicity. Computing this sum is annoying because there are lots of different terms, but it is not really difficult. The main thing you need to notice is that when $k_1 \neq k_2, k_3, k_4$, then the random variable X_{k_1} is independent of $X_{k_2}, X_{k_3}, X_{k_4}$, so we have

$$\mathbf{E}(X_{k_1} X_{k_2} X_{k_3} X_{k_4}) = \mathbf{E}(X_{k_1}) \mathbf{E}(X_{k_2} X_{k_3} X_{k_4}) = 0$$

(because $\mathbf{E}(X_k) = \mathbf{E}(\mathbf{1}_{A_k}) - \mathbf{P}(A) = 0$). Similarly, if any of the indices k_i is distinct from all the others, then this term is zero. Therefore, there are only four types of terms that are nonzero: either $k_1 = k_2 = k_3 = k_4$, or $k_1 = k_2 \neq k_3 = k_4$ or $k_1 = k_3 \neq k_2 = k_4$, or $k_1 = k_4 \neq k_2 = k_3$. If you compute the sum of each of these terms, you get

$$\mathbf{E}(Y_n) = \frac{1}{n^3} \mathbf{E}((\mathbf{1}_A - \mathbf{P}(A))^4) + \frac{3(n-1)}{n^3} \mathbf{E}((\mathbf{1}_A - \mathbf{P}(A))^2)^2.$$

If you are so inclined, you can check this computation carefully.

On the other hand, we have now achieved exactly what we want, because

$$\sum_{n=1}^{\infty} \mathbf{E}(Y_n) = \mathbf{E}((\mathbf{1}_A - \mathbf{P}(A))^4) \sum_{n=1}^{\infty} \frac{1}{n^3} + 3\,\mathbf{E}((\mathbf{1}_A - \mathbf{P}(A))^2)^2 \sum_{n=1}^{\infty} \frac{n-1}{n^3} < \infty.$$

It follows that

$$\lim_{n\to\infty} \left(\frac{S_n}{n} - \mathbf{P}(A) \right)^4 = \lim_{n\to\infty} Y_n = 0,$$

and we have proved the strong law of large numbers.

2.4 From discrete to continuous arrivals

Let us consider the following example. You work at the Fruity Yogurt store in Princeton. In order to optimize your napping time in the back of the store,

you are trying to model the arrival of customers at the store. You know from experience that, on average, μ customers enter the store in one hour. What can we say about the distribution of the number of customers that arrive in one hour? How long will it take for the first, second, or third customer to arrive? In this section, we will try to answer these questions.

Of course, just saying that μ customers arrive on average in one hour is not enough to specify the model. We must make a more precise assumption. A very natural idea, which is used in many problems involving arrival times, is to introduce a modelling assumption that reads roughly as follows: *at every time, the arrival of a customer happens with equal probability and independently.* This modelling assumption makes intuitive sense! You should think of the mechanism that this represents as follows: each person who lives in Princeton decides independently at each time whether or not they will go get some Fruity Yogurt; as the craving for Fruity Yogurt can occur at any time, it will be equally likely at every time that a customer will enter the store.

The problem with this description is that it is not entirely clear how to turn it into mathematics: what do we mean by "at every time independently with equal probability?" To understand this, we first consider a simpler problem.

Suppose we divide each hour in n intervals of the same length (for example, if $n = 60$, then each interval is one minute long). Inspired by the above modelling assumption, it is natural to assume that in every interval, a customer enters a store with probability p, and that the arrivals of customers in different intervals are independent. Denote by X_k the random variable whose value is one if a customer arrives in the kth interval, and zero otherwise. Then $X_1, X_2, X_3, \ldots$ is none other than a Bernoulli process! In particular, the number of customers that arrive in one hour (that is, after n intervals) is

$$S_n = \sum_{k=1}^{n} X_k \sim \text{Binom}(n, p).$$

This implies that the expected number of customers that arrive in one hour is $\mathbf{E}(S_n) = np$. But our modelling assumption stated that μ customers arrive on average in one hour. Therefore, the probability that a customer arrives in a single interval must be chosen as follows:

$$p = \frac{\mu}{n}.$$

Remark 2.4.1. Of course, a probability must be between zero and one, so we need to assume that we made the time intervals small enough that $n \geq \mu$.

In some sense, this model captures a form of the modelling assumption that we have made. However, it is not quite satisfactory. The problem with this model is that the size of the time intervals is somehow arbitrary: why do people in Princeton only decide exactly at the beginning of every minute whether they will enter the Fruity Yogurt store? The craving for Fruity Yogurt can arrive

at any time: at the beginning of some second, or some femtosecond, or... In order to model this, we should therefore shrink the size of the time intervals to zero, so that people can decide *at every time* whether to go to the store—as we stated informally in our modelling assumption above! Mathematically, this modelling assumption corresponds to taking the limit $n \to \infty$ as the size of the time intervals goes to zero. This idea is called the *continuous time limit* of our discrete time model. In this process, we must remember to scale the probability of arrival in each interval $p = \mu/n$ with n, so that the modelling assumption that an average of μ customers arrive in an hour is preserved.

We will now investigate what happens in this continuous time limit. Let us begin by computing the distribution of the number of customers N that arrive in one hour. To this end, we must take the continuous time limit of the binomial random variable S_n, which describes the number of customers that arrive in an hour in the discrete time model with n time intervals:

$$\mathbf{P}\{N = k\} = \lim_{\substack{n\to\infty \\ p=\mu/n}} \mathbf{P}\{S_n = k\} = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k}.$$

This limit looks scarier than it is. It is helpful to rearrange it as follows:

$$\binom{n}{k} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} = \frac{\mu^k}{k!} \frac{n!}{(n-k)!n^k} \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)^{-k}.$$

Your knowledge of calculus allows you to compute

$$\lim_{n\to\infty} \frac{n!}{(n-k)!n^k} = \lim_{n\to\infty} \frac{n}{n} \frac{n-1}{n} \frac{n-2}{n} \cdots \frac{n-k+1}{n} = 1,$$

$$\lim_{n\to\infty} \left(1 - \frac{\mu}{n}\right)^n = e^{-\mu},$$

$$\lim_{n\to\infty} \left(1 - \frac{\mu}{n}\right)^{-k} = 1.$$

We have therefore shown that in our continuous time model, the probability that k customers arrive in the first hour is given by

$$\mathbf{P}\{N = k\} = \frac{e^{-\mu}\mu^k}{k!}.$$

This distribution is so important that it is named after its inventor, the renowned French mathematician Siméon Poisson (1781–1840).

Definition 2.4.2. *A random variable Z with distribution*

$$\mathbf{P}\{Z = k\} = \frac{e^{-\mu}\mu^k}{k!}, \qquad 0 \le k < \infty$$

is called a Poisson random variable *with mean μ ($Z \sim Pois(\mu)$).*

Example 2.4.3 (Sanity checks). Because you might be a bit suspicious still about the continuous time limit, it is instructive to perform some sanity checks on the formula we have obtained for the Poisson distribution to make sure everything we did made sense. First, let us compute the probability that the number of customers N that arrive in one hour takes some value $0 \leq k < \infty$:

$$\sum_{k=0}^{\infty} \mathbf{P}\{N = k\} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} e^{\mu} = 1,$$

where we have noticed that the sum that appears here is just the Taylor expansion of the exponential function (review this material in your calculus textbook if you do not remember it). Of course, if the answer had been anything else, we would have been in trouble! Similarly, we can compute the expected number of customers that arrive in one hour:

$$\mathbf{E}(N) = \sum_{k=0}^{\infty} k\, \mathbf{P}\{N = k\} = e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!} = e^{-\mu} \mu \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu.$$

We knew in advance that this would be the answer, however: the fact that μ customers arrive on average per hour was one of our modelling assumptions, which we strictly enforced in the continuous time limit.

We now turn to the following question: what can we say about the time

$$T_k = \text{time at which } k\text{th customer arrives ?}$$

To reason about these arrival times, it is useful to think of the number of customers that have entered the store as a function of time:

$$N_t = \#\text{ customers that arrive by time } t.$$

Thus $N_1 = N$ is the number of customers that arrive in the first hour, $N_{0.5}$ is the number of customers that arrive in the first half hour, etc. As μ customers arrive per hour on average, and as customers arrive independently with the same probability at every time, it must be that μt customers arrive on average by time t: you can think of μ as the *rate* at which customers arrive. Thus

$$N_t \sim \text{Pois}(\mu t), \qquad t \geq 0.$$

This is our first encounter with the random process $\{N_t\}_{t \geq 0}$ that is known as a *Poisson process*. We will develop the theory of Poisson processes more carefully and in much more detail later in this course. For the moment, let us simply use it as a tool to study the arrival time T_k of the kth customer.

Example 2.4.4. What is the probability that the 5th customer arrives at Fruity Yogurt by time t? Here is an easy way to think about this problem. If the fifth customer arrived by time t (that is, $T_5 \leq t$), then the number of customers that arrived by time t must be at least five. Conversely, if at least five customers arrived by time t, then the fifth customer must certainly have arrived by time t. We have therefore shown that the following events are equal:

$$\{T_5 \leq t\} = \{N_t \geq 5\}.$$

We know how to compute the probability of the latter event, because we know that N_t is a Poisson random variable. It is slightly easier to compute the probability of the negation of this event, as follows:

$$\begin{aligned}\mathbf{P}\{T_5 \leq t\} &= \mathbf{P}\{N_t \geq 5\} = 1 - \mathbf{P}\{N_t < 5\} \\ &= 1 - \sum_{n=0}^{4} \mathbf{P}\{N_t = n\} = 1 - \sum_{n=0}^{4} \frac{e^{-\mu t}(\mu t)^n}{n!}.\end{aligned}$$

We cannot simplify this expression further: the answer is what it is!

There is nothing special about this example; by exactly the same reasoning, we can compute the probability that the kth customer arrives by time t:

$$\mathbf{P}\{T_k \leq t\} = 1 - \sum_{n=0}^{k-1} \frac{e^{-\mu t}(\mu t)^n}{n!}.$$

Such random variables have a name.

Definition 2.4.5. *A random variable T that satisfies*

$$\mathbf{P}\{T \leq t\} = 1 - \sum_{n=0}^{k-1} \frac{e^{-\mu t}(\mu t)^n}{n!}, \qquad t \geq 0$$

is a Gamma variable *with shape k and scale μ ($T \sim Gamma(k, \mu)$).*

Of particular interest to the Fruity Yogurt salesman is the arrival time T_1 of the first customer (which determines how long he can nap). From the above formula, we immediately read off the following special case:

$$\mathbf{P}\{T_1 \leq t\} = 1 - e^{-\mu t}.$$

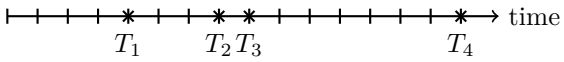
Such random variables appear in many different problems.

Definition 2.4.6. *A random variable T that satisfies*

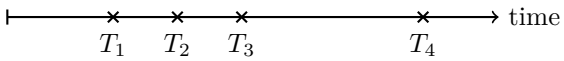
$$\mathbf{P}\{T \le t\} = 1 - e^{-\mu t}, \qquad t \ge 0$$

is called an exponential random variable *with rate μ ($T \sim Expon(\mu)$).*

Notice that while N_t is a discrete random variable (the number of customers that arrive by time t can take the values $0, 1, 2, \ldots$), the random variables T_k are *continuous*: the arrival time of the kth customer can be any nonnegative real number $t \ge 0$. That is, unlike the arrival times of the kth success of a Bernoulli process, which could look like



the craving for Fruity Yogurt can occur at any time, perhaps like this:



The basic rules of probability apply equally to continuous random variables.

Example 2.4.7. What is the probability that the first customer arrives between times a and b? Note that the customer arrives by time b if and only if either he arrives by time a, or he arrives between times a and b. That is,

$$\{T_1 \le b\} = \{T_1 \le a\} \cup \{a < T_1 \le b\},$$

and the two events on the right are mutually exclusive. Therefore

$$\mathbf{P}\{T_1 \le b\} = \mathbf{P}\{T_1 \le a\} + \mathbf{P}\{a < T_1 \le b\},$$

from which we can compute

$$\mathbf{P}\{a < T_1 \le b\} = \mathbf{P}\{T_1 \le b\} - \mathbf{P}\{T_1 \le a\} = e^{-\mu a} - e^{-\mu b}.$$

From this example, we can compute in particular the probability that the first customer will arrive at *exactly* time t:

$$\mathbf{P}\{T_1 = t\} = \lim_{\varepsilon \to 0} \mathbf{P}\{t - \varepsilon < T_1 \le t\} = 0.$$

That is, the probability that the first customer will arrive at exactly time t is zero, for every time t! As was explained in Example 1.4.3, this is precisely how it should be: if we are asked to guess the arrival time of a customer with infinite precision, we will never guess correctly (so the probability is zero). There is no contradiction here between probability theory and common sense.

Nonetheless, this property of continuous variables introduces complications when we want to define notions such as the *expectation* of a continuous random variable. Recall that we have so far only defined the expectation $\mathbf{E}(X) = \sum_{k \in D} k\, \mathbf{P}\{X = k\}$ for discrete random variables: clearly this definition makes no sense for continuous variables (it would suggest that the expectation of the arrival time of the first customer is zero, which is absurd!) In order to work with continuous quantities, we must extend our definitions of notions such as expectation and conditional probability in a manner that makes sense for continuous variables. This is our next order of business.

3

Continuous Random Variables

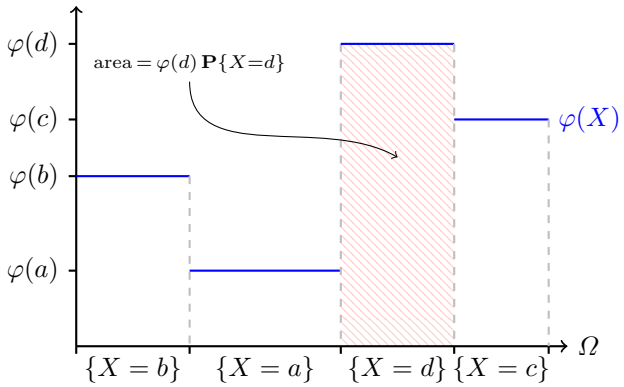
In Chapter 1, we defined the notions of expectation, conditioning, and independence for discrete random variables. The goal of this chapter is to extend these basic principles of probability theory to continuous random variables. In the process, we introduce cumulative distributions functions and probability densities, which are the basic ingredients of continuous models.

3.1 Expectation and integrals

Recall that if $X : \Omega \to D$ is a discrete random variable (where D is a finite or countable subset of $\mathbb{R}$), then the expectation of $\varphi(X)$ is defined as

$$\mathbf{E}(\varphi(X)) = \sum_{i \in D} \varphi(i)\, \mathbf{P}\{X = i\}.$$

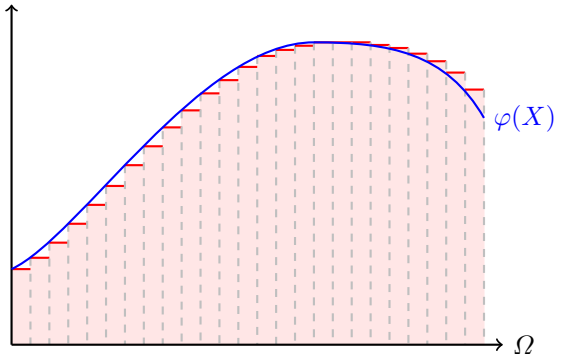
To get some more insight into this definition, let us illustrate it in a picture:



In this picture, we have represented the subsets $\{X = i\}$ on the horizontal axis as intervals whose length is proportional to their probability $\mathbf{P}\{X = i\}$.

When the function is plotted in this manner, you directly see that the area under the graph of the function in the set $\{X = i\}$ is precisely $\varphi(i)\,\mathbf{P}\{X = i\}$. The expectation $\mathbf{E}(X)$ is therefore nothing other than the total area under the graph of the function $\varphi(X)$, or, equivalently, the *integral* of this function! The interpretation of expectation as an integral is the key idea that allows us to extend this notion to continuous random variables.

Let us now consider a continuous random variable $X : \Omega \to \mathbb{R}$. How should we define its expectation? We can use the same idea as you used in calculus to define integrals of continuous functions: draw increasingly fine discrete functions that approximate the graph (whose integrals we can compute):



If we take the limit of the integrals of such discrete functions as the discretization gets finer and finer, we will end up with the area under the graph. This is precisely how you defined the Riemann integral in your calculus course. We do exactly the same thing for the expectation: the only thing we must remember is that the "width" of each of the intervals is actually its probability.

Definition 3.1.1. *The expectation of a random variable $X : \Omega \to \mathbb{R}$ is*

$$\mathbf{E}(\varphi(X)) = \int \varphi(x)\,\mathbf{P}\{X \in dx\} := \lim_{|x_{i+1}-x_i|\to 0} \sum_i \varphi(x_i)\,\mathbf{P}\{x_i < X \le x_{i+1}\},$$

where the limit is taken over increasingly fine partitions $x_1 < x_2 < x_3 < \cdots$.

In calculus, dx is often thought of as an "infinitely small number" (an *infinitesimal*). In probability theory, it is more natural to think of dx as an "infinitely small interval" at the point x, that is, something like $]x, x + dx]$. We can now see that our definition of expectation in the continuous case is a direct extension of the discrete case: the right way to resolve our difficulties is to integrate over $\mathbf{P}\{x \le X \le x + dx\}$ (which is "infinitely small"), rather than summing over $\mathbf{P}\{X = x\}$ (which is typically zero).

At this point, you might be inclined to be unhappy: the definition of expectation looks incredibly ugly. You would not want to have to compute a bunch of limits every time you compute an expectation! Don't panic: the same also

happened in calculus. When you defined the Riemann integral, you did that in terms of a limit just like the one above. But rather than computing integrals from first principles, you quickly learned some rules that allowed you to compute integrals without ever taking any limits. The same thing will happen here. In fact, we will shortly show how to rewrite the "probability integral" as an ordinary Riemann integral that we all know and love from calculus.

Example 3.1.2 (Uniform distribution). Let X be a random variable that is equally likely to take any value in $[0,1]$. As was discussed in Example 1.4.3, in this case we must have $\mathbf{P}\{X \in A\} = \text{length}(A)$ for $A \subseteq [0,1]$. We then say that X is *uniformly distributed* in the interval $[0,1]$.

What is the expectation of $\varphi(X)$? In this special case

$$\mathbf{E}(\varphi(X)) = \lim_{|x_{i+1}-x_i|\to 0} \sum_i \varphi(x_i)\,(x_{i+1}-x_i) = \int_0^1 \varphi(x)\,dx$$

is just the ordinary Riemann integral from calculus, where we have used that $\mathbf{P}\{x_i < X \le x_{i+1}\} = x_{i+1} - x_i$. For example,

$$\mathbf{E}(X) = \int_0^1 x\,dx = \frac{1}{2}.$$

This makes perfect sense: the average of all numbers in $[0,1]$, when each number is equally likely, should be one half.

In the special case of the uniform distribution, the expectation really is an ordinary calculus integral. In other cases, it is not entirely obvious whether this is still true. We will now show that one can compute expectations quite generally using ordinary calculus. To see why, note first that (why?)

$$\mathbf{P}\{x_i < X \le x_{i+1}\} = \mathbf{P}\{X \le x_{i+1}\} - \mathbf{P}\{X \le x_i\}.$$

If we define the function $F(x) = \mathbf{P}\{X \le x\}$, then

$$\lim_{|x_{i+1}-x_i|\to 0} \frac{\mathbf{P}\{x_i < X \le x_{i+1}\}}{x_{i+1}-x_i} = \lim_{|x_{i+1}-x_i|\to 0} \frac{F(x_{i+1})-F(x_i)}{x_{i+1}-x_i} = \frac{dF}{dx}(x_i).$$

We can therefore compute

$$\begin{aligned}
\mathbf{E}(\varphi(X)) &= \lim_{|x_{i+1}-x_i|\to 0} \sum_i \varphi(x_i)\,\mathbf{P}\{x_i < X \le x_{i+1}\} \\
&= \lim_{|x_{i+1}-x_i|\to 0} \sum_i \varphi(x_i)\,\frac{\mathbf{P}\{x_i < X \le x_{i+1}\}}{x_{i+1}-x_i}\,(x_{i+1}-x_i) \\
&= \lim_{|x_{i+1}-x_i|\to 0} \sum_i \varphi(x_i)\,\frac{dF}{dx}(x_i)\,(x_{i+1}-x_i) \\
&= \int \varphi(x)\,\frac{dF(x)}{dx}\,dx.
\end{aligned}$$

Thus as long as we can compute the derivative of the function $F(x)$, we can compute the expectation $\mathbf{E}(\varphi(X))$ using ordinary calculus. This is almost always the case for continuous random variables.

Remark 3.1.3. It is easy to remember this formula without taking limits just from the integral notation. If you wish, you can formally write

$$\mathbf{P}\{X \in dx\} = \frac{\mathbf{P}\{X \in dx\}}{dx}\, dx = \left(\frac{d}{dx}\mathbf{P}\{X \le x\}\right) dx,$$

so that

$$\mathbf{E}(\varphi(X)) = \int \varphi(x)\, \mathbf{P}\{X \in dx\} = \int \varphi(x)\, \frac{d}{dx}\mathbf{P}\{X \le x\}\, dx.$$

Clearly the function $F(x)$ and its derivative are extremely important for continuous random variables. We therefore give them names.

Definition 3.1.4. *The function* $F(x) = \mathbf{P}\{X \le x\}$ *is called the* cumulative distribution function (CDF) *of the random variable* X*. Its derivative* $f(x) = \frac{d}{dx}F(x)$ *is called the* density *of the random variable* X*.*

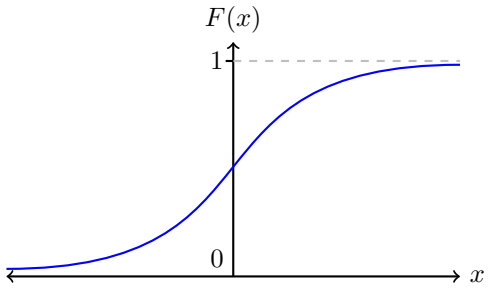
The *distribution* of a random variable is the information needed to compute expectations of the form $\mathbf{E}(\varphi(X))$ for any function φ. Evidently, specifying either the CDF or the density of a random variable determines its distribution.

What do the functions $F(x)$ and $f(x)$ look like?

- Note that $\{X \le x\} \subseteq \{X \le x'\}$ and thus $\mathbf{P}\{X \le x\} \le \mathbf{P}\{X \le x'\}$ for $x \le x'$. This implies that $F(x)$ is an *increasing* function. Moreover,

$$F(+\infty) = \mathbf{P}\{X < +\infty\} = 1,$$
$$F(-\infty) = \mathbf{P}\{X = -\infty\} = 0.$$

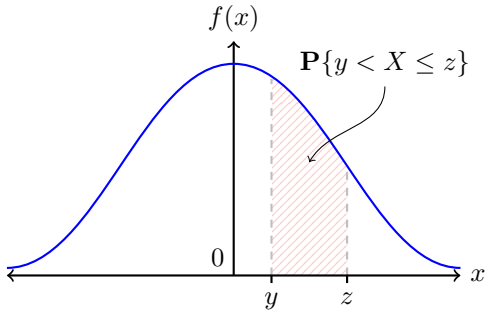
Thus the CDF of a random variable could look something like this:



- As the CDF $F(x)$ is increasing, its derivative $f(x) = \frac{d}{dx}F(x)$ is non-negative. Furthermore, it must be the case that

$$\int f(x)\,dx = \mathbf{E}(1) = 1.$$

Thus the density of a random variable could look something like this:



Note that we can compute

$$\mathbf{P}\{y < X \leq z\} = \mathbf{E}(\mathbf{1}_{]y,z]}(X)) = \int_y^z f(x)\,dx.$$

Therefore, the area under the density in an interval corresponds to the probability that the random variable will be in that interval.

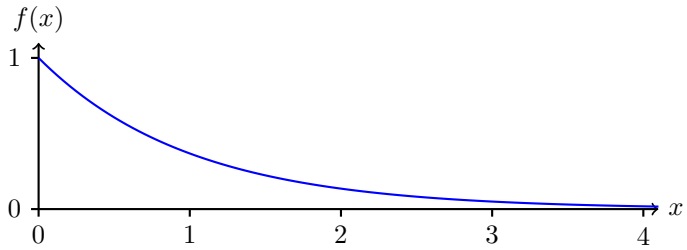
Remark 3.1.5. It is easy to get an idea of how a random variable behaves by looking at its density: the random variable takes values with more probability in regions where there is more area under the curve. It is tempting to think of $f(x)$ as being the "probability at the point x" so that the density is large for points with large probability, but this is not quite correct: for continuous

random variables $\mathbf{P}\{X = x\} = 0$ (in fact, unlike probabilities, the density may well be more than one!) The right way to interpret $f(x)$ is to note that the probability that X is in a very small interval $]x, x + \Delta x]$ is approximately $f(x)\Delta x$; so the density at a point x gives a probability of being in a small neighborhood of x when you multiply it by the size of the neighborhood.

Example 3.1.6 (Exponential random variables). Let $X \sim \text{Expon}(\mu)$. We defined such exponential random variables in the previous chapter by specifying their CDF $\mathbf{P}\{X \leq x\} = 1 - e^{-\mu x}$. The density of X is therefore

$$\frac{d}{dx}\mathbf{P}\{X \leq x\} = \mu e^{-\mu x}.$$

The exponential density looks like this (for $\mu = 1$, say):



We can now compute

$$\mathbf{E}(X) = \int_0^\infty x\, \mu e^{-\mu x}\, dx = -\mu \frac{d}{d\mu} \int_0^\infty e^{-\mu x}\, dx = -\mu \frac{d}{d\mu} \frac{1}{\mu} = \frac{1}{\mu}.$$

This makes perfect sense! Recall that X is the time we must wait for the arrival of a customer, when μ customers arrive on average per hour. Intuitively, if there are μ customers per hour, then there are $1/\mu$ hours per customer, and thus the first customer is expected to arrive after $\sim 1/\mu$ hours. Similarly,

$$\mathbf{E}(X^2) = \int_0^\infty x^2\, \mu e^{-\mu x}\, dx = \mu \frac{d^2}{d\mu^2} \int_0^\infty e^{-\mu x}\, dx = \mu \frac{d^2}{d\mu^2} \frac{1}{\mu} = \frac{2}{\mu^2}.$$

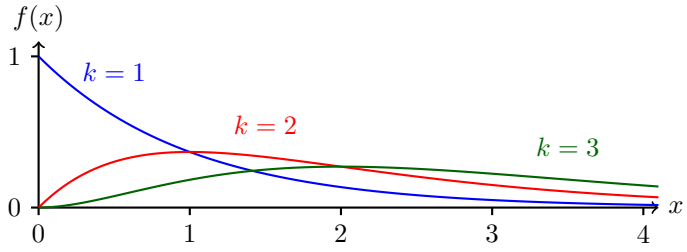
The variance of an exponential random variable is therefore

$$\text{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2 = \frac{1}{\mu^2}.$$

Example 3.1.7 (Gamma random variables). Let $X \sim \text{Gamma}(k, \mu)$. We defined Gamma random variables in the previous chapter by specifying their CDF $\mathbf{P}\{X \le x\} = 1 - \sum_{n=0}^{k-1} e^{-\mu x}(\mu x)^n/n!$. The density of X is therefore

$$\frac{d}{dx}\mathbf{P}\{X \le x\} = \frac{\mu e^{-\mu x}(\mu x)^{k-1}}{(k-1)!}.$$

The Gamma density looks like this (for $\mu = 1$, say):



We can now compute

$$\begin{aligned}\mathbf{E}(X) &= \int_0^\infty x\,\frac{\mu e^{-\mu x}(\mu x)^{k-1}}{(k-1)!}\,dx \\ &= \frac{k}{\mu}\int_0^\infty \frac{\mu e^{-\mu x}(\mu x)^{k}}{k!}\,dx \\ &= \frac{k}{\mu}\,\mathbf{P}\{0 \le Y < \infty\} = \frac{k}{\mu},\end{aligned}$$

where Y is a random variable with distribution $\text{Gamma}(k+1, \mu)$. This makes perfect sense! Recall that X is the time of arrival of the kth customer, when μ customers arrive on average per hour. As each customer takes $1/\mu$ hours on average to arrive, the average arrival time of the kth customer should be k/μ.

Remark 3.1.8. Note how we used probability in the last example to avoid having to compute a nasty integral. You could have shown that

$$\int_0^\infty \frac{\mu e^{-\mu x}(\mu x)^{k}}{k!}\,dx = 1$$

by integrating by parts k times. However, we did not need to do that: we know the integrand is the density of a Gamma random variable, and thus its integral must be one (as is the case for all probability densities). This sort of reasoning can often be used to simplify computations.

3.2 Joint and conditional densities

Let X and Y be two random variables. How can we compute the expectation of $\varphi(X,Y)$? For discrete random variables, we could write

$$\mathbf{E}(\varphi(X,Y)) = \sum_{x,y} \varphi(x,y)\,\mathbf{P}\{X=x, Y=y\}.$$

in terms of the joint distribution of X and Y. Exactly the same thing happens in the continuous case: if X and Y are continuous random variables, we have

$$\mathbf{E}(\varphi(X,Y)) = \int \varphi(x,y)\,\mathbf{P}\{X\in dx, Y\in dx\}.$$

You can define and justify this integral as a limit in exactly the same way as we did for the expectation of a single random variable $\mathbf{E}(\varphi(X))$.

In order to compute the integral, we use that

$$\begin{aligned}\mathbf{P}\{X\in dx, Y\in dy\} &= \frac{\mathbf{P}\{X\in dx, Y\in dy\}}{dx\,dy}\,dx\,dy \\ &= \frac{\partial^2}{\partial x\,\partial y}\mathbf{P}\{X\le x, Y\le y\}\,dx\,dy.\end{aligned}$$

The function

$$f(x,y) = \frac{\partial^2}{\partial x\,\partial y}\mathbf{P}\{X\le x, Y\le y\}$$

is called the *joint density* of X and Y. Once we have computed the joint density, we can readily compute using ordinary calculus

$$\mathbf{E}(\varphi(X,Y)) = \int \varphi(x,y)\,f(x,y)\,dx\,dy.$$

Example 3.2.1 (Joint and marginal densities). Suppose we model continuous random variables X, Y by specifying their joint density $\mathbf{P}\{X\in dx, Y\in dy\} = f(x,y)\,dx\,dy$. How can we compute the *marginal* density of X?

Note that $\varphi(x)$ can be viewed as a special case of a function of two variables $\psi(x,y) = \varphi(x)$ that happens not to depend on y. Therefore

$$\mathbf{E}(\varphi(X)) = \mathbf{E}(\psi(X,Y)) = \int \varphi(x)\,f(x,y)\,dx\,dy = \int \varphi(x)\Big(\int f(x,y)\,dy\Big)\,dx.$$

Therefore, the marginal density $f_X(x)$ of X is obtained from the joint density $f(x,y)$ of X and Y by integrating over all possible values of Y:

$$f_X(x) = \int f(x,y)\,dy.$$

This is completely analogous to the discrete case: the marginal distribution of a discrete random variable X can be obtained from the joint distribution of X and Y by summing over all possible outcomes of Y.

Example 3.2.2 (Additivity of expectation). For discrete random variables, we proved the common sense property $\mathbf{E}(X+Y) = \mathbf{E}(X)+\mathbf{E}(Y)$ of expectations. Let us verify that this important property also holds for continuous variables. To this end, we compute the expectation of the function $\varphi(X,Y) = X+Y$:

$$\begin{aligned}\mathbf{E}(X+Y) &= \int (x+y)\,f(x,y)\,dx\,dy \\ &= \int x\,f(x,y)\,dx\,dy + \int y\,f(x,y)\,dx\,dy \\ &= \mathbf{E}(X) + \mathbf{E}(Y).\end{aligned}$$

In the last equality, we identified the two integrals as the expectation of the functions $\varphi(X,Y) = X$ and $\varphi(X,Y) = Y$, respectively.

We now turn to the problem of defining *conditional* distributions and expectations for continuous random variables. For discrete random variables, we defined the conditional distribution of X given that Y takes the value y as

$$\mathbf{P}\{X = x | Y = y\} = \frac{\mathbf{P}\{X = x, Y = y\}}{\mathbf{P}\{Y = y\}}.$$

This definion does not make sense for continuous random variables, however, as the event $\{Y = y\}$ has zero probability: therefore

$$\mathbf{P}\{X \in dx | Y = y\} \stackrel{?}{=} \frac{\mathbf{P}\{X \in dx, Y = y\}}{\mathbf{P}\{Y = y\}} = \frac{0}{0}$$

does not make any sense! We solve this problem in the same manner as when we defined the expectation of continuous random variables. Even though the event $\{Y = y\}$ has zero probability, the event $\{y \le Y \le y + \Delta y\}$ has positive probability for every $\Delta y > 0$. We can therefore define $\mathbf{P}\{X \in dx | Y = y\}$ as the limit of $\mathbf{P}\{X \in dx | y \le Y \le y + \Delta y\}$ as $\Delta y \to 0$. This yields the following.

Definition 3.2.3. *The* conditional distribution *of X given $Y = y$ is*

$$\mathbf{P}\{X \in dx | Y = y\} = \frac{\mathbf{P}\{X \in dx, Y \in dy\}}{\mathbf{P}\{Y \in dy\}}.$$

Remark 3.2.4. On the left-hand side of this definition, we condition on $Y = y$. On the right-hand side, however, we have $Y \in dy$. The point is that the dy's on the right-hand side cancel, as they appear both in the numerator and the denominator. Thus the bookkeeping works out: the same number of dx's and dy's appear on either side of the equality, as must be the case.

If the joint density of X and Y is $\mathbf{P}\{X \in dx, Y \in dy\} = f(x,y)\,dx\,dy$,

$$\mathbf{P}\{X \in dx | Y = y\} = \frac{f(x,y)\,dx\,dy}{(\int f(x',y)\,dx')\,dy} = \frac{f(x,y)}{(\int f(x',y)\,dx')}\,dx =: f(x|y)\,dx,$$

where we have used the expression for the marginal density of Y in the denominator. Note again how the dy's cancel on the right-hand side. The function $f(x|y)$ is called the *conditional density* of x given y.

Given the conditional distribution, we can now compute the conditional expectation of any function $\varphi(X)$ given $Y = y$ as follows:

$$\mathbf{E}(\varphi(X)|Y = y) = \int \varphi(x)\,\mathbf{P}\{X \in dx | Y = y\} = \int \varphi(x)\,f(x|y)\,dx.$$

As in the case of discrete random variables, we can also compute the (unconditional) expectation $\mathbf{E}(\varphi(X))$ in terms of the conditional expectation:

$$\begin{aligned}\mathbf{E}(\varphi(X)) &= \int \varphi(x)\,\mathbf{P}\{X \in dx, Y \in dy\} \\ &= \int \varphi(x)\,\mathbf{P}\{X \in dx | Y = y\}\,\mathbf{P}\{Y \in dy\} \\ &= \int \mathbf{E}(\varphi(X)|Y = y)\,\mathbf{P}\{Y \in dy\}.\end{aligned}$$

This is particularly useful in situations where conditional distributions are specified as a modelling assumption: in such cases it is often easier to compute first $\mathbf{E}(X|Y)$ (the conditional expectation of X given the outcome of the random variable Y, and then to compute the expectation of this random variable to obtain $\mathbf{E}(X)$, than it is to compute $\mathbf{E}(X)$ directly.

Example 3.2.5 (Shopping robot). A robot is sent by aliens from outer space to go shopping on Earth. The robot does not want to draw attention to himself

by attempting to shop at an alarming rate (to Earthlings). In order to blend in, he calibrates his shopping schedule as follows. First, he stakes out a store and measures the time it takes for the first customer to arrive. If the first customer arrives after x hours, the robot calibrates himself to shop at a rate of entering the store $1/x$ times per hour. How long do we need to wait on average for the robot to enter the store (after he completed his calibration)?

What are our modelling assumptions? If we assume μ customers enter the store per hour on average, then the arrival time of the first customer is $X \sim \text{Expon}(\mu)$. The arrival time Y of the robot to the store (after his calibration is complete) is modelled by specifying its conditional distribution $\text{Expon}(1/x)$ given $X = x$: that is, $\mathbf{P}\{Y \in dy | X = x\} = \frac{1}{x} e^{-y/x}\, dy$. Therefore,

$$\mathbf{E}(Y) = \int \mathbf{E}(Y|X = x)\, \mathbf{P}\{X \in dx\} = \int_0^\infty x\, \mu e^{-\mu x}\, dx = \mathbf{E}(X) = \frac{1}{\mu}.$$

That is, the average time it takes the robot to enter the store is the same as the average time it takes for the customer to enter the store. Unlike the customer, however, the marginal distribution for the robot is not exponential. To see this, note that the joint distribution is given by

$$\mathbf{P}\{X \in dx, Y \in dy\} = \mathbf{P}\{X \in dx\}\mathbf{P}\{Y \in dy | X = x\} = \mu e^{-\mu x}\, \frac{1}{x} e^{-y/x}\, dx\, dy,$$

so the marginal density $f_Y(y)$ of Y is given by

$$f_Y(y) = \int_0^\infty \frac{\mu}{x}\, e^{-\mu x - y/x}\, dx.$$

I do not want to compute this integral (the answer is nasty in terms of so-called Bessel functions), but the point is that it is certainly not exponential. Nonetheless, we have seen that we could easily compute $\mathbf{E}(Y)$.

Example 3.2.6 (A guessing game). Three people play the following game. The first person generates a random number $X \sim \text{Expon}(\mu)$, and gives its outcome $X = x$ to the second person. The second person draws a random number $Y \sim \text{Expon}(x)$, and gives its outcome to the third person. The third person is asked to guess the original number x that was drawn by the first person.

What are our modelling assumptions? We are given the distribution of X and the conditional distribution of Y given $X = x$:

$$\mathbf{P}\{X \in dx\} = \mu e^{-\mu x}\, dx, \qquad \mathbf{P}\{Y \in dy | X = x\} = x e^{-xy}\, dy.$$

The joint distribution of X and Y is therefore given by

$$\mathbf{P}\{X \in dx, Y \in dy\} = \mathbf{P}\{X \in dx\}\mathbf{P}\{Y \in dy | X = x\} = \mu x e^{-(\mu+y)x}\, dx\, dy,$$

and the marginal distribution of Y is given by

$$\mathbf{P}\{Y \in dy\} = \left(\int_0^\infty \mu x e^{-(\mu+y)x}\, dx \right) dy = \frac{\mu}{(\mu+y)^2}\, dy$$

(you can compute the integral easily by noting that $\int_0^\infty x(\mu+y)e^{-(\mu+y)x}dx$ is the expectation of a random variable with distribution Expon$(\mu+y)$).

The goal of the third player is to estimate the outcome of X, given that $Y = y$. He should therefore compute the conditional distribution

$$\mathbf{P}\{X \in dx | Y = y\} = \frac{\mathbf{P}\{X \in dx, Y \in dy\}}{\mathbf{P}\{Y \in dy\}} = (\mu+y)x\,(\mu+y)e^{-(\mu+y)x}\, dx.$$

That is, the conditional distribution of X given $Y = y$ is Gamma$(2, \mu+y)$, so

$$\mathbf{E}(X|Y=y) = \frac{2}{\mu+y}.$$

Thus the expected outcome of X, given the outcome of Y (which is the only information available to the third player), is $2/(\mu+Y)$.

The last two examples illustrate the notion of *randomization of parameters*. Many distributions, such as Expon(μ) or Bin(n,p), depend on some numerical parameters (μ or n, p) that must be chosen either experimentally or on the basis of some modelling assumption when the model is constructed. In some cases, however, the parameters can themselves be determined by some other random variable in the model: for example, in the previous example the random variable Y was modelled by an exponential distribution whose parameter was itself a random variable drawn from an exponential distribution. As we have seen, this idea is readily implemented using conditional distributions.

3.3 Independence

Recall that discrete random variables are said to be independent if conditioning on one of the variables does not affect the distribution of the other variable. This idea is entirely in line with our real-world intuition about what "independence" means. For continuous random variables, the notion of independence is defined in exactly the same manner.

Definition 3.3.1. *Random variables X, Y are said to be* independent *if*

$$\mathbf{P}\{X \in dx | Y = y\} = \mathbf{P}\{X \in dx\}.$$

If we multiply both sides of this definition by $\mathbf{P}\{Y \in dy\}$, we obtain the following equivalent definition: X and Y are independent if

$$\mathbf{P}\{X \in dx, Y \in dy\} = \mathbf{P}\{X \in dx\}\,\mathbf{P}\{Y \in dy\}.$$

An entirely analogous definition was given for discrete random variables.

Example 3.3.2 (How many people get on the bus?). On average, μ people arrive at a bus stop per hour, and ν buses arrive per hour. The arrivals of the people and of the buses are independent. How many people get on the first bus?

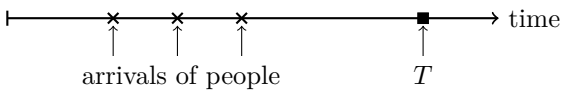
Let us model each of the ingredients of this problem. Define

$$T = \text{time of arrival of first bus},$$
$$N_t = \#\text{ people waiting at time } t.$$

As we argued in section 2.4, the distributions of these random variables are

$$T \sim \text{Expon}(\nu), \qquad N_t \sim \text{Pois}(\mu t),$$

and T, N_t are independent random variables. For example, the arrivals of people and buses in a given experiment could look something like this:



We would like to compute the expectation and distribution of N_T, the number of people at the bus stop at the time of arrival of the bus. How should you interpret this? In any experiment, Tyche chooses one outcome $\omega \in \Omega$; in this experiment, $N_t(\omega)$ is the number of people waiting for the bus at time t, and $T(\omega)$ is the time that the first bus arrives. Thus in this experiment, the number of people at the bus stop when the bus arrives is $N_{T(\omega)}(\omega)$. As the outcome depends on ω, this defines a random variable which we denote N_T.

How should we compute the expectation $\mathbf{E}(N_T)$? We do not know, a priori, the distribution of N_T, but we do know the distribution of N_t for a fixed t. In particular, we know the conditional distribution

$$\mathbf{E}(N_T|T=t) = \mathbf{E}(N_t|T=t) = \mathbf{E}(N_t) = \mu t.$$

Here we have used first that *conditionally* on $T = t$, we must have $N_T = N_t$; and second, that N_t and T are independent. We can now use the formula from the previous section to conclude that

$$\mathbf{E}(N_T) = \int \mathbf{E}(N_T|T=t)\,\mathbf{P}\{T \in dt\} = \int \mu t\,\mathbf{P}\{T \in dt\} = \mu\,\mathbf{E}(T) = \frac{\mu}{\nu}.$$

This makes perfect sense: on average μ people arrive per hour and the first bus arrives after $1/\nu$ hours, so on average μ/ν people get on the first bus.

With a little more work, we can compute the entire distribution of the number of people who get on the first bus. Note that

$$\mathbf{P}\{N_T = k | T = t\} = \mathbf{P}\{N_t = k | T = t\} = \mathbf{P}\{N_t = k\} = \frac{e^{-\mu t}(\mu t)^k}{k!},$$

where we have used again conditioning and independence. Therefore

$$\begin{aligned}
\mathbf{P}\{N_T = k\} &= \int \mathbf{P}\{N_T = k | T = t\}\, \mathbf{P}\{T \in dt\} \\
&= \int \frac{e^{-\mu t}(\mu t)^k}{k!}\, \nu e^{-\nu t}\, dt \\
&= \frac{\nu}{\mu + \nu} \left(\frac{\mu}{\mu + \nu}\right)^k \int \frac{e^{-(\mu+\nu)t}(\mu + \nu)^{k+1} t^k}{k!}\, dt.
\end{aligned}$$

But you can recognize the integrand in the last integral as the density of a random variable with distribution Gamma$(k+1, \mu+\nu)$, and thus this integral must be one (as is the integral of every probability density). Therefore

$$\mathbf{P}\{N_T = k\} = \left(\frac{\mu}{\mu + \nu}\right)^k \frac{\nu}{\mu + \nu}, \qquad k = 0, 1, 2, \ldots$$

That is, $N_T + 1$ is a geometric random variable with parameter $\nu/(\mu + \nu)$.

4

Lifetimes and Reliability

In this chapter, we will create models of the lifetime of (potentially complicated) systems. This is a problem that arises in applications ranging from insurance to computer networks. More importantly, it is a good exercise in modelling and using continuous random variables.

4.1 Lifetimes

Let L be the lifetime of a person, of a device, of a complex operation, ... In most cases of interest, we can (and will) clearly assume that

- L is a random variable; and
- $0 < L < \infty$ (death occurs eventually but not instantly).

Like any continuous random variable, one can define its distribution by specifying the CDF $F(t) := \mathbf{P}\{L \leq t\}$.

In the case of lifetimes, however, it proves to be convenient to define the distribution by specifying the probability $\mathbf{P}\{L > t\} = 1 - F(t)$ that the lifetime exceeds t in terms of functions $H(t)$ or $h(t)$ as follows.

Definition 4.1.1. *A random variable L is a* lifetime *if $0 < L < \infty$. The* (cumulative) hazard function $H(t)$ *and* hazard rate $h(t)$ *are defined by*

$$\mathbf{P}\{L > t\} = e^{-H(t)} = e^{-\int_0^t h(s)ds}.$$

Let us discuss the significance of these functions.

- Note that as $0 < L < \infty$, we must have

$$e^{-H(0)} = \mathbf{P}\{L > 0\} = 1,$$
$$e^{-H(\infty)} = \mathbf{P}\{L = \infty\} = 0.$$

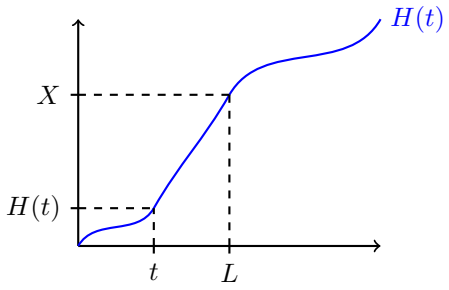
Moreover, as $\{L > t\} \supset \{L > s\}$ if $s \geq t$, the function $\mathbf{P}\{L > t\}$ must be decreasing. Therefore, the cumulative hazard function must satisfy

$$H(0) = 0, \qquad H(\infty) = \infty, \qquad H(t) \text{ is increasing.}$$

Let us note an interesting property of the lifetime L. Let $X \sim \text{Expon}(1)$ be an exponential random variable, that is, $\mathbf{P}\{X > t\} = e^{-t}$. Then

$$\mathbf{P}\{H^{-1}(X) > t\} = \mathbf{P}\{X > H(t)\} = e^{-H(t)} = \mathbf{P}\{L > t\}.$$

Therefore, the lifetime L has the same distribution as the random variable $H^{-1}(X)$, as is illustrated in the following figure:



Here is one possible intuitive (though biologically questionable) interpretation of this picture. When a person is born, Tyche decides how many heartbeats that person will have in their lifetime. This number X is different for each person, and thus it is a random variable. There is nothing you can do to change it. However, you have a lot of control on how you lead your life. You can spend a lot of time jogging, in which case your heart will beat very quickly; or you can sleep a lot, in which case your heart will beat much more slowly. Thus the function $H(t)$, the number of heartbeats you have spent after t years, is under your control. Once the number of heartbeats $H(t)$ reaches your quota X, you die. This happens precisely at the time $L = H^{-1}(X)$, which is your lifetime.

Of course, this story is somewhat made up. But the idea is clear: the cumulative hazard function $H(t)$ measures the "degree of deterioration" of the object in question by time t. The modelling of lifetimes can therefore be done quite naturally by specifying this deterioration function.

- Because $H(0) = 0$, we can write $H(t) = \int_0^t h(s)ds$ with $h(s) = \frac{d}{ds}H(s)$ by the fundamental theorem of calculus. As $H(t)$ is increasing, its derivative $h(t)$ must be nonnegative. As $H(\infty) = \infty$, the hazard rate must satisfy

$$h(t) \geq 0, \qquad \int_0^\infty h(s)\,ds = \infty.$$

In order to understand the significance of the hazard rate, let us first compute the conditional probability that you will live for another u years, given that you made it to be t years old:

$$\begin{aligned}\mathbf{P}\{L > t+u | L > t\} &= \frac{\mathbf{P}\{L > t+u, L > t\}}{\mathbf{P}\{L > t\}} = \frac{\mathbf{P}\{L > t+u\}}{\mathbf{P}\{L > t\}} \\ &= e^{-(H(t+u)-H(t))} = e^{-\int_t^{t+u} h(s)ds}.\end{aligned}$$

Therefore, the probability that you will die in the next u years, given that you made it to be t years old, is given by

$$\mathbf{P}\{L \leq t+u | L > t\} = 1 - \mathbf{P}\{L > t+u | L > t\} = 1 - e^{-\int_t^{t+u} h(s)ds}.$$

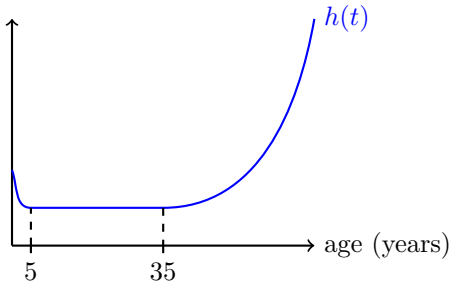
If we define the function $F_t(u) = 1 - e^{-\int_t^{t+u} h(s)ds}$, then

$$\lim_{u\to 0} \frac{1}{u}\mathbf{P}\{L \leq t+u | L > t\} = \lim_{u\to 0} \frac{F_t(u) - F_t(0)}{u} = \frac{\partial}{\partial u} F_t(u)\bigg|_{u=0} = h(t).$$

This means that when Δt is very small,

$$\mathbf{P}\{L \leq t + \Delta t | L > t\} \approx h(t)\Delta t.$$

You can therefore think of the hazard rate $h(t)$ as being approximately the probability that you will die in the next year, given that you have made it to be t years old. You can imagine that an insurance company would be very interested in this number. For the human lifetime, for example, the hazard rate typically looks something like this:



Infants are fragile and have a somewhat elevated death rate; once they get a few years old, they are sufficiently mature that their death rate is at its

lowest point. It stays low until age 35–40, after which the death rate grows increasingly rapidly as you get older.

It should be clear that the hazard rate provides a very natural method to model the distribution of a lifetime: one is essentially specifying the probability of death as a function of age. If the hazard rate $h(t)$ is increasing, we say that the model exhibits *ageing*; if it is decreasing, it exhibits *maturation*. For humans, both maturation and ageing effects appear.

Let us give two classical examples of lifetime models.

Example 4.1.2 (Exponential distribution). Let $L \sim \text{Expon}(\mu)$. Then

$$\mathbf{P}\{L > t\} = e^{-\mu t}, \qquad H(t) = \mu t, \qquad h(t) = \mu.$$

In this case, the hazard rate is constant: the probability of death does not change over time. In particular, there is no ageing and no maturation.

A useful way to think about this property is as follows: note that

$$\mathbf{P}\{L > t + u | L > t\} = e^{-(H(t+u)-H(t))} = e^{-\mu u} = \mathbf{P}\{L > u\}.$$

That is, if you made it to age t, the probability that you will live another u years is the same as the probability that you make it to age u: there is no memory of the fact that you have already been alive for t years. This is called the *memorylessness* property of the exponential distribution.

This is of course a highly unrealistic model for the lifetime of a person or of an electronic or mechanical device, which exhibit significant ageing. But, as we have already seen, it is a good model for the time until it takes a customer to enter the store, in which case no ageing effects arise.

Example 4.1.3 (Weibull distribution). The Weibull distribution is defined by

$$\mathbf{P}\{L > t\} = e^{-\mu t^\alpha}, \qquad H(t) = \mu t^\alpha, \qquad h(t) = \mu \alpha t^{\alpha - 1}$$

for $\mu, \alpha > 0$. It is named for a Swedish engineer, Waloddi Weibull (1887–1979).

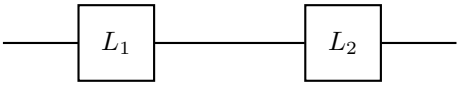
The Weibull distribution can model either ageing of maturation. When $\alpha < 1$, the hazard rate is decreasing and thus the model exhibits maturation. When $\alpha > 1$, the hazard rate is increasing and thus the model exchibits ageing. When $\alpha = 1$, the hazard rate is constant: the special case of the Weibull distribution with $\alpha = 1$ is none other than the exponential distribution.

4.2 Minima and maxima

Now that we have a general approach for modelling lifetimes, we can investigate what happens when we put multiple objects with different lifetimes

together to form a larger system. This does not necessarily make sense for people, but is very useful for modelling electronic devices. Each electric component has its own lifetime, and they are combined together in an electronic circuit. How long does it take until the circuit breaks? This type of question arises naturally in many problems concerning *reliability* of complex systems. In this section, we will investigate the two simplest cases; in the next section, we will develop a more systematic method to approach such problems.

Let us begin with the simplest configuration: two components in *series*.



Think of two lights in a string of lights on a Christmas tree. As you likely know from experience, if any of the lights breaks, then the circuit is broken and all the lights go off. So, when two components are put in series, then the system breaks as soon as one of the components breaks. Therefore, if the lifetimes of the two components are L_1 and L_2, respectively, then the lifetime L of the full system is the minimum of the lifetimes of the components: $L = L_1 \wedge L_2$.

Remark 4.2.1. It is customary in probability theory to use so-called lattice-theoretic notation for minima and maxima: for two numbers x and y, we write $x \wedge y := \min(x, y)$ and $x \vee y := \max(x, y)$. The notation is easy to remember if you recall the analogous notation for sets: $A \cap B$ is *smaller* than the sets A, B, while $A \cup B$ is *larger* than the sets A, B.

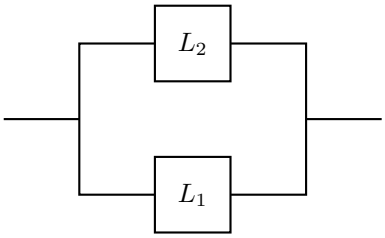
Now suppose that L_1 has hazard function H_1 and L_2 has hazard function H_2. What is the hazard function of the total lifetime L? It is natural to assume that the lifetimes of each of the components (e.g., lights on a christmas tree) is independent, that is, $L_1 \perp\!\!\!\perp L_2$. Then

$$\begin{aligned}\mathbf{P}\{L > t\} &= \mathbf{P}\{L_1 \wedge L_2 > t\} \\ &= \mathbf{P}\{L_1 > t, L_2 > t\} \\ &= \mathbf{P}\{L_1 > t\}\,\mathbf{P}\{L_2 > t\} \\ &= e^{-(H_1(t)+H_2(t))}.\end{aligned}$$

That is, if L_1 has hazard function H_1, and L_2 has hazard function H_2, and if $L_1 \perp\!\!\!\perp L_2$, then the minmum $L_1 \wedge L_2$ has hazard function $H = H_1 + H_2$.

Example 4.2.2 (Minimum of independent exponentials). The following very useful property follows immediately: if $X \sim \text{Expon}(\mu)$ and $Y \sim \text{Expon}(\nu)$, and if $X \perp\!\!\!\perp Y$, then $X \wedge Y \sim \text{Expon}(\mu + \nu)$.

Now suppose we would like to avoid that the entire system breaks when one of the components breaks. To combat this annoying behavior, we might place the components in *parallel* rather than in series:



In this case, if one of the lights fails, the other keeps functioning. Therefore, the light goes off only when *all* the lights break. Thus the total lifetime of this system is the maximum of the lifetimes of the components: $L = L_1 \vee L_2$.

Computing the hazard function of the maximum of two independent lifetimes $L_1 \perp\!\!\!\perp L_2$ is more annoying than computing the minimum. For the maximum, it is more convenient to compute $\mathbf{P}\{L \le t\}$ rather than $\mathbf{P}\{L > t\}$:

$$\begin{aligned}
\mathbf{P}\{L > t\} &= 1 - \mathbf{P}\{L \le t\} \\
&= 1 - \mathbf{P}\{L_1 \vee L_2 \le t\} \\
&= 1 - \mathbf{P}\{L_1 \le t, L_2 \le t\} \\
&= 1 - \mathbf{P}\{L_1 \le t\}\,\mathbf{P}\{L_2 \le t\} \\
&= 1 - (1 - e^{-H_1(t)})(1 - e^{-H_2(t)}).
\end{aligned}$$

Thus if $L_1 \perp\!\!\!\perp L_2$, then the hazard function of $L_1 \vee L_2$ is given by

$$H(t) = -\log(1 - (1 - e^{-H_1(t)})(1 - e^{-H_2(t)})).$$

This is ugly! Nonetheless, we can perform useful computations concerning the maximum of independent lifetimes. Perhaps the easiest trick in the book is the following example; we will develop a more systematic approach shortly.

Example 4.2.3. Notice that for any two numbers x and y, we can write

$$x \vee y = x + y - x \wedge y.$$

Thus to compute the *expected* lifetime of two components in parallel, we only need the expected lifetimes of each component and of the system in series:

$$\mathbf{E}(L_1 \vee L_2) = \mathbf{E}(L_1) + \mathbf{E}(L_2) - \mathbf{E}(L_1 \wedge L_2).$$

For example, if $L_1 \sim \text{Expon}(\mu)$, $L_2 \sim \text{Expon}(\nu)$, and $L_1 \perp\!\!\!\perp L_2$, then

$$\mathbf{E}(L_1 \vee L_2) = \frac{1}{\mu} + \frac{1}{\nu} - \frac{1}{\mu + \nu}.$$

The approach of Example 4.2.3 is a bit ad-hoc, and would be difficult to apply to more complicated lifetimes. Fortunately, there is a more systematic way to perform computations with lifetimes that is often convenient. In particular, we will presently develop a useful trick to compute the expected lifetime $\mathbf{E}(L)$.

In principle, we should compute $\mathbf{E}(L)$ using its definition:

$$\mathbf{E}(L) = \int_0^\infty t\,\mathbf{P}\{L \in dt\} = \int_0^\infty t\,\frac{d}{dt}\mathbf{P}\{L \le t\}\,dt = \int_0^\infty t\,\frac{d}{dt}(1-\mathbf{P}\{L > t\})\,dt.$$

As we have an expression for $\mathbf{P}\{L > t\}$, we can plug it in to compute $\mathbf{E}(L)$. This is a bit annoying, however, as it involves taking a derivative. We can avoid this using a simple trick. Note that for any number $x \ge 0$,

$$x = \int_0^x dt = \int_0^x 1\,dt + \int_x^\infty 0\,dt = \int_0^\infty \mathbf{1}_{x>t}\,dt.$$

We can therefore write

$$L = \int_0^\infty \mathbf{1}_{L>t}\,dt.$$

Taking the expectation on both sides gives

$$\boxed{\mathbf{E}(L) = \int_0^\infty \mathbf{E}(\mathbf{1}_{L>t})\,dt = \int_0^\infty \mathbf{P}\{L > t\}\,dt.}$$

This formula is very convenient: it allows us to compute $\mathbf{E}(L)$ directly by integrating $\mathbf{P}\{L > t\}$, without having to take derivatives.

Remark 4.2.4. The formula that we have obtained for $\mathbf{E}(L)$ can be applied to any nonnegative random variable, not just to lifetimes. However, it can *not* be applied to random variables that may be negative.

Remark 4.2.5. If you are a lover of calculus, you could have also obtained this formula from the definition of $\mathbf{E}(L)$ by integrating by parts.

Example 4.2.6 (Expected lifetime of two components in parallel). Consider a system consisting of two independent components in parallel. Suppose that the component lifetimes are exponentially distributed with $L_1 \sim \text{Expon}(\mu)$ and $L_2 \sim \text{Expon}(\nu)$. The system lifetime L satisfies

$$\begin{aligned}\mathbf{P}\{L > t\} &= 1 - (1 - \mathbf{P}\{L_1 > t\})(1 - \mathbf{P}\{L_2 > t\}) \\ &= 1 - (1 - e^{-\mu t})(1 - e^{-\nu t}) \\ &= e^{-\mu t} + e^{-\nu t} - e^{-(\mu+\nu)t}.\end{aligned}$$

We can therefore compute

$$\mathbf{E}(L) = \int_0^\infty (e^{-\mu t} + e^{-\nu t} - e^{-(\mu+\nu)t})\, dt = \frac{1}{\mu} + \frac{1}{\nu} - \frac{1}{\mu+\nu}.$$

We obtained precisely the same answer as we did in Example 4.2.3. The present approach is much more systematic, however.

Example 4.2.7 (Expected remaining lifetime of two components in parallel). Consider again the model from the previous example. Now we are interested in the following question: supposing the system made it to age t, how long do we expect it to remain functioning after that? That is, we are interested in the conditional expectation of the remaining life $L-t$ after time t, given that the system made it to age t (that is, given that $L > t$). By exactly the same reasoning as above, we can compute

$$\mathbf{E}(L-t|L>t) = \int_0^\infty \mathbf{P}\{L-t>u|L>t\}\, du,$$

where by definition

$$\begin{aligned}\mathbf{P}\{L-t>u|L>t\} &= \frac{\mathbf{P}\{L>t+u, L>t\}}{\mathbf{P}\{L>t\}} \\ &= \frac{\mathbf{P}\{L>t+u\}}{\mathbf{P}\{L>t\}} \\ &= \frac{e^{-\mu(t+u)} + e^{-\nu(t+u)} - e^{-(\mu+\nu)(t+u)}}{e^{-\mu t} + e^{-\nu t} - e^{-(\mu+\nu)t}}.\end{aligned}$$

Integrating with respect to u gives

$$\mathbf{E}(L-t|L>t) = \frac{\frac{1}{\mu}e^{-\mu t} + \frac{1}{\nu}e^{-\nu t} - \frac{1}{\mu+\nu}e^{-(\mu+\nu)t}}{e^{-\mu t} + e^{-\nu t} - e^{-(\mu+\nu)t}}.$$

It is a bit hard to see in this expression what is going on. To simplify matters a bit, let us consider the special case $\mu = \nu$ where each component has the same average lifetime. In this case, the expression simplifies to

$$\mathbf{E}(L-t|L>t) = \frac{\frac{2}{\mu}e^{-\mu t} - \frac{1}{2\mu}e^{-2\mu t}}{2e^{-\mu t} - e^{-2\mu t}} = \frac{1}{\mu}\left(1 + \frac{1}{4e^{\mu t} - 2}\right).$$

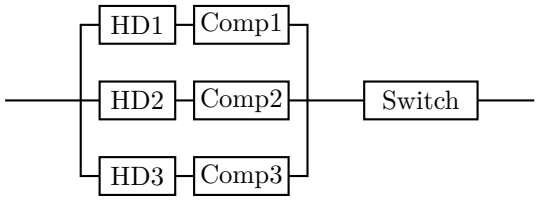
Note that as t increases, the expected remaining lifetime at age t decreases. Even though the components have exponential lifetimes, so that each component is memoryless, the total system is not memoryless: it exhibits ageing! The explanation is simple. When t is small, there will most likely be two functioning components, so the expected lifetime is roughly the maximum lifetime of two components. When t is large, even if we know the system is still functioning, most likely one of the two components has died; and therefore the remaining lifetime of the system is roughly the liftime of only one component.

4.3 * Reliability

We will skip the material in this section in 2015. It is included in case you are interested, but will not appear on the homework or exams.

In the previous section, we investigated the lifetimes of two simple systems of components: two components in series, and two components in parallel. Realistic systems (such as jet engines or nuclear reactors) are composed of thousands or even millions of components. To reason about the reliability of complex systems, we must develop a more systematic way to think about their structure. Even in simple examples such as the following one, performing computations directly can quickly become tedious.

Example 4.3.1 (Web server). A web server consists of three computers, each with their own hard drive. They are configured in parallel so that failure of one component does not bring down the server. All three computers talk to the internet through one network switch. Thus the system looks like this:



Note that each computer and its hard drive are in series: a computer cannot function once its hard drive breaks, and vice versa.

If we denote the lifetimes of the three computers as L_1, L_3, L_5, the lifetimes of the three hard drives as L_2, L_4, L_6, and the lifetime of the switch as L_7, then the total lifetime L of this system can be expressed by combining the series and parallel lifetimes that we derived in the previous section:

$$L = ((L_1 \wedge L_2) \vee (L_3 \wedge L_4) \vee (L_5 \wedge L_6)) \wedge L_7.$$

Doing computations directly with this expression is quite tedious, however.

To investigate the lifetime or reliability of such complicated systems, we need a systematic way to describe the structure of the system in terms of its components. To that end, suppose that we have a system with n components that function independently. For each i, we define

$$X_i = \begin{cases} 1 & \text{if component } i \text{ is functioning,} \\ 0 & \text{if component } i \text{ is dead.} \end{cases}$$

X_i is called the *state variable* of component i. Note that $X_1, \ldots, X_n$ are independent Bernoulli variables. For convenince, let

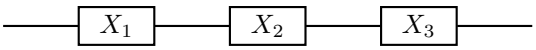
$$p_i = \mathbf{P}\{X_i = 1\} = \mathbf{E}(X_i)$$

be the probability that component i is functioning. Whether or not the entire system is functioning depends on whether each component is functioning according to the structure of the system. We encode the behavior of the system in terms of a *structure function* $\varphi : \{0,1\}^n \to \{0,1\}$ such that

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if the system is functioning,} \\ 0 & \text{if the system is dead.} \end{cases}$$

We must learn how to define the structure function for a given system. This is most easily seen by working out a few examples.

Example 4.3.2 (Systems in series). Consider three components in series:



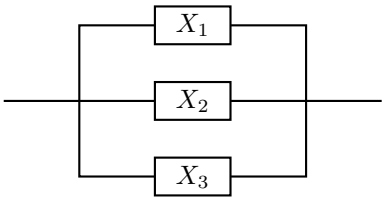
The system is functioning only if all three components are functioning. Therefore, the structure function for this system is given by

$$\varphi(x_1, x_2, x_3) = x_1 x_2 x_3.$$

To check that this structure function is correct, note that if one of x_1, x_2, x_3 is zero, then $\varphi(x_1, x_2, x_3) = 0$, but if all x_1, x_2, x_3 are one, then $\varphi(x_1, x_2, x_3) = 1$.

More generally, suppose you are given two systems with n_1 and n_2 components, respectively, whose structure functions are φ_1 and φ_2. If you place these two systems in series, you get a new system whose structure function is given by $\varphi(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) = \varphi_1(x_1, \ldots, x_{n_1})\, \varphi_2(y_1, \ldots, y_{n_2})$.

Example 4.3.3 (Systems in parallel). Consider three components in parallel:



This is, in a sense, exactly the opposite of components in series: this system is *not* functioning only if all three components are *not* functioning. So, we can obtain the parallel case from the series case by putting "not" in front of every "functioning". What does this mean mathematically? Component i is *not* functioning if $1 - X_i = 1$, and the entire system is *not* functioning if $1 - \varphi(X_1, X_2, X_3) = 1$. Therefore, the structure function is given by

$$\varphi(x_1, x_2, x_3) = 1 - (1 - x_1)(1 - x_2)(1 - x_3).$$

To check that this structure function is correct, note that if one of x_1, x_2, x_3 is one, then $\varphi(x_1, x_2, x_3) = 1$, but if all x_1, x_2, x_3 are zero, then $\varphi(x_1, x_2, x_3) = 0$.

More generally, suppose you are given two systems with n_1 and n_2 components, respectively, whose structure functions are φ_1 and φ_2. If you place these two systems in parallel, you get a new system whose structure function is given by $\varphi(x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) = 1 - (1 - \varphi_1(x_1, \ldots, x_{n_1}))(1 - \varphi_2(y_1, \ldots, y_{n_2}))$.

Example 4.3.4 (Web server). Consider the web server system of Example 4.3.1. We can decompose this into subsystems: first, the computers are placed in series with their hard drives. Then, the three computer/hard drive combos are placed in parallel. Finally, this server farm is placed in series with the network switch. Putting these together gives the structure function

$$\varphi(x_1, \ldots, x_7) = (1 - (1 - x_1x_2)(1 - x_3x_4)(1 - x_5x_6))x_7.$$

(Here x_i corresponds to the component with lifetime L_i in Example 4.3.1.)

Now that we can describe the structure of a system, we can use this to compute the probability that the system is functioning in terms of the probabilities that each of the components is functioning. To this end, note that

$$\mathbf{P}\{\text{system is functioning}\} = \mathbf{P}\{\varphi(X_1, \ldots, X_n) = 1\} = \mathbf{E}(\varphi(X_1, \ldots, X_n)).$$

The latter expectation is usually easy to compute, as we know that $X_1, \ldots, X_n$ are independent Bernoulli variables with probabilities $p_1, \ldots, p_n$.

Example 4.3.5 (Components in series and in parallel). In the setting of three components in series as in Example 4.3.2, we readily compute

$$\mathbf{P}\{\text{system is functioning}\} = \mathbf{E}(X_1X_2X_3) = \mathbf{E}(X_1)\,\mathbf{E}(X_2)\,\mathbf{E}(X_3) = p_1p_2p_3,$$

where we have used that the expectation of the product of independent random variables is the product of the expectations. In exactly the same manner, we compute for three components in parallel as in Example 4.3.3

$$\begin{aligned}\mathbf{P}\{\text{system is functioning}\} &= \mathbf{E}(1 - (1 - X_1)(1 - X_2)(1 - X_3)) \\ &= 1 - \mathbf{E}(1 - X_1)\,\mathbf{E}(1 - X_2)\,\mathbf{E}(1 - X_3) \\ &= 1 - (1 - p_1)(1 - p_2)(1 - p_3),\end{aligned}$$

where we have used linearity of the expectation and independence.

Example 4.3.6 (Web server). In the web server of Example 4.3.1, we have

$$\begin{aligned}\mathbf{P}\{\text{system is functioning}\} &= \mathbf{E}((1 - (1 - X_1X_2)(1 - X_3X_4)(1 - X_5X_6))X_7) \\ &= (1 - (1 - p_1p_2)(1 - p_3p_4)(1 - p_5p_6))p_7.\end{aligned}$$

Remark 4.3.7. In all three examples above, it happens that

$$\mathbf{P}\{\text{system is functioning}\} = \mathbf{E}(\varphi(X_1, \ldots, X_n)) = \varphi(p_1, \ldots, p_n).$$

That is very convenient! It turns out that this is always the case for systems of independent components that are configured in series and parallel. This is because the structure function of a system with series and parallel components can always be written as a sum of products of distinct state variables, as we have seen above; and as the expectation of a product of independent random variables is the product of the expectations. This observation makes it very easy to compute the probabilty that such a system is functioning.

However, *be very careful:* it is *not* true in general that

$$\mathbf{E}(\varphi(X_1, \ldots, X_n)) \stackrel{?}{=} \varphi(\mathbf{E}(X_1), \ldots, \mathbf{E}(X_n))$$

for any function φ and random variables $X_1, \ldots, X_n$. The fact that this is true for systems of independent components in series and parallel is somewhat of a (convenient) coincidence. In more general cases, you can no longer reason in this manner. For example, given three components, consider the "two-out-of-three" system that is functioning if at least two of the three components are functioning. This cannot be described as a system of series and parallel components; the structure function for this system is given by (why?)

$$\varphi(x_1, x_2, x_3) = 1 - (1 - x_1x_2)(1 - x_2x_3)(1 - x_1x_3).$$

Note that the random variables $1 - X_1X_2$, $1 - X_2X_3$, and $1 - X_1X_3$ are not independent: each pair shares one of the variables X_i! It is therefore *not* true that the probability that the system is functioning is $\varphi(p_1, p_2, p_3)$; to compute the correct probability, you must multiply out the products above, which gives

$$\mathbf{E}(\varphi(X_1, X_2, X_3)) = p_1p_2 + p_2p_3 + p_1p_3 - 2p_1p_2p_3 \neq \varphi(p_1, p_2, p_3).$$

The lesson is that as long as you only have components in series and parallel, you can use the above convenient observation to compute the probability that the system is functioning very quickly; but you must be very careful not to abuse this idea where it does not apply.

So far, we have only considered the probability that a component or the system is functioning right now. What can we say about the lifetime of the system? To bring time back into the picture, let us suppose that the ith component has lifetime L_i, and that the lifetimes are independent. At any time t, we can check whether or not each component is functioning. Note that component i is functioning at time t if its time of death is in the future, that is, if $L_i > t$. Therefore, the state variable of component i at time t is

$$\mathbf{1}_{\{L_i > t\}} = \begin{cases} 1 & \text{if component } i \text{ is functioning at time } t, \\ 0 & \text{if component } i \text{ is dead by time } t. \end{cases}$$

Similarly, if L is the lifetime of the entire system, then

$$\mathbf{1}_{\{L>t\}} = \begin{cases} 1 & \text{if the system is functioning at time } t, \\ 0 & \text{if the system is dead by time } t. \end{cases}$$

We therefore immediately obtain

$$\mathbf{1}_{L>t} = \varphi(\mathbf{1}_{\{L_1>t\}}, \ldots, \mathbf{1}_{\{L_n>t\}}),$$

which allows us to compute the distribution of the system lifetime as

$$\mathbf{P}\{L > t\} = \mathbf{E}(\varphi(\mathbf{1}_{\{L_1>t\}}, \ldots, \mathbf{1}_{\{L_n>t\}})).$$

In the special case of two components in series or two components in parallel, this gives exactly the same answer as we obtained in the previous section.

4.4 * A random process perspective

We will skip the material in this section in 2015. It is included in case you are interested, but will not appear on the homework or exams.

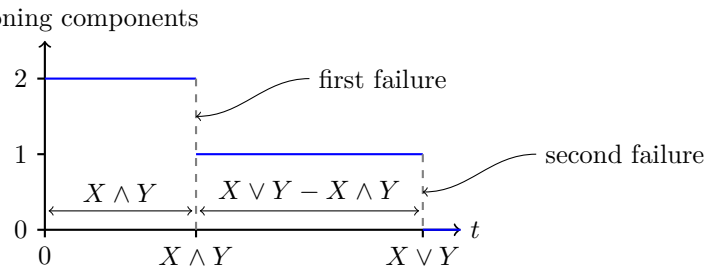
In the previous sections, we saw how to express the probability $\mathbf{P}\{L > t\}$ that a system is functioning at a fixed time t in terms of the probabilities $\mathbf{P}\{L_i > t\}$ that its components are functioning. We could subsequently use that to compute expected lifetimes, conditional distributions, etc. Nonetheless, this is essentially a static picture: when we fix a time t, the lifetimes only enter the picture to the extent that they determine the state variables of each of the components at that time. That components are malfunctioning one after another as time passes does not enter our analysis.

In this section, we will take a more dynamic view: we are going to think of the number of functioning components as a random process, that is, a quantity that changes over time. This point of view is particularly useful for exponential lifetimes, which we will consider throughout this section. The random process perspective will lead us to some new insights, and will help is perform certain computations much more easily. In the rest of this course, random processes of different kinds will play a central role, and we will often see that they help us solve problems that would otherwise be much more difficult.

Let us begin by considering the simplest possible case of two independent components with exponential lifetimes

$$X \sim \text{Expon}(\mu), \qquad Y \sim \text{Expon}(\nu), \qquad X \perp\!\!\!\perp Y.$$

What happens over time? Initially, both components are functioning. Then, at time $X \wedge Y$, one of the components fails. Subsequently, at time $X \vee Y$, the other component fails as well. This is illustrated in the following plot:



Note that we can write

$$X \vee Y = X \wedge Y + (X \vee Y - X \wedge Y).$$

That, is the time $X \vee Y$ at which the last component fails is the time $X \wedge Y$ at which the first component fails, plus the amount of time $X \vee Y - X \wedge Y$ between the first failure the last failure.

We have seen that $X \wedge Y$, the minimum of two exponential random variables with parameters μ and ν, is itself an exponential random variable with parameter $\mu + \nu$. The distribution of $X \vee Y$ was much less attractive. What can we say about the distribution of the time $X \vee Y - X \wedge Y$ between failures? Here is the intuition. Suppose that component X is the first to fail. Then, at the time when it fails, only component Y remains functional. However, by the memoryless property of the exponential, if we know component Y was functional at the time when the other component died, then its remaining lifetime should still be Expon(ν) (the same as its original distribution). On the other hand, if it was Y that died first, then the remaining lifetime of X has distribution Expon(μ). To make this intuition precise, we have to condition on the events $Y > X$ that X dies first or $X > Y$ that Y dies first:

$$\begin{aligned}\mathbf{P}\{X \vee Y - X \wedge Y > t\} = {} & \mathbf{P}\{X \vee Y - X \wedge Y > t | X > Y\}\,\mathbf{P}\{X > Y\} + \\ & \mathbf{P}\{X \vee Y - X \wedge Y > t | X < Y\}\,\mathbf{P}\{X < Y\}.\end{aligned}$$

(Note that we ignored the possibility that $X = Y$; however, the probability that this happens is zero, so we can eliminate it from our computations.) Using that $X \vee Y = X$ and $X \wedge Y = Y$ if $X > Y$, and vice versa, this becomes

$$\begin{aligned}\mathbf{P}\{X \vee Y - X \wedge Y > t\} = {} & \mathbf{P}\{X - Y > t | X > Y\}\,\mathbf{P}\{X > Y\} + \\ & \mathbf{P}\{Y - X > t | Y > X\}\,\mathbf{P}\{Y > X\}.\end{aligned}$$

We would like to argue, by the memoryless property of the exponential, that $\mathbf{P}\{X - Y > t | X > Y\} = \mathbf{P}\{X > t\}$. That would certainly be true if Y were replaced by a nonrandom number $y \in \mathbb{R}$. But does the memoryless property still hold when the time from which we are computing the remaining lifetime is random (but independent of the exponential variable)? Fortunately, the answer is yes! To see that this is true, we condition on Y:

$$\begin{aligned}
\mathbf{P}\{X > Y + t | X > Y\} &= \frac{\mathbf{P}\{X > Y + t\}}{\mathbf{P}\{X > Y\}} \\
&= \frac{\int_0^\infty \mathbf{P}\{X > Y + t | Y = s\}\, \mathbf{P}\{Y \in ds\}}{\int_0^\infty \mathbf{P}\{X > Y | Y = s\}\, \mathbf{P}\{Y \in ds\}} \\
&= \frac{\int_0^\infty \mathbf{P}\{X > s + t | Y = s\}\, \mathbf{P}\{Y \in ds\}}{\int_0^\infty \mathbf{P}\{X > s | Y = s\}\, \mathbf{P}\{Y \in ds\}} \\
&= \frac{\int_0^\infty \mathbf{P}\{X > s + t\}\, \mathbf{P}\{Y \in ds\}}{\int_0^\infty \mathbf{P}\{X > s\}\, \mathbf{P}\{Y \in ds\}} \\
&= \frac{\int_0^\infty e^{-\mu(s+t)}\, \nu e^{-\nu s}\, ds}{\int_0^\infty e^{-\mu s}\, \nu e^{-\nu s}\, ds} \\
&= e^{-\mu t} \frac{\int_0^\infty e^{-\mu s}\, \nu e^{-\nu s}\, ds}{\int_0^\infty e^{-\mu s}\, \nu e^{-\nu s}\, ds} \\
&= e^{-\mu t} = \mathbf{P}\{X > t\}.
\end{aligned}$$

On the other hand, what is the probability $\mathbf{P}\{X > Y\}$ that Y dies before X? This is easy to compute by conditioning as well:

$$\begin{aligned}
\mathbf{P}\{X > Y\} &= \int_0^\infty \mathbf{P}\{X > Y | Y = y\}\, \mathbf{P}\{Y \in dy\} \\
&= \int_0^\infty \mathbf{P}\{X > y | Y = y\}\, \mathbf{P}\{Y \in dy\} \\
&= \int_0^\infty \mathbf{P}\{X > y\}\, \mathbf{P}\{Y \in dy\} \\
&= \int_0^\infty e^{-\mu y}\, \nu e^{-\nu y}\, dy \\
&= \frac{\nu}{\mu + \nu}.
\end{aligned}$$

While these computations look a bit lengthy, they are not hard. We are doing something that we often do when we have to deal with two independent random variables: we condition on the outcome of one of the variables, so that we can separate the probabilities involving the two variables. The hard work has paid off, because we can now put things together to obtain our final expression for the distribution of the time between the first and second failures:

$$\begin{aligned}
\mathbf{P}\{X \vee Y - X \wedge Y > t\} &= \mathbf{P}\{X > t\}\, \mathbf{P}\{X > Y\} + \mathbf{P}\{Y > t\}\, \mathbf{P}\{Y > X\} \\
&= \frac{\nu}{\mu + \nu}\, e^{-\mu t} + \frac{\mu}{\mu + \nu}\, e^{-\nu t}.
\end{aligned}$$

This formula should be seen as the precise mathematical statement of the intuition that we started out with: if X dies before Y, then the time between failures has the same distribution as Y; and if Y dies before X, then the time between failures has the same distribution as X.

Example 4.4.1. Let us use the above formula to compute the expectation of the maximum of two exponential variables. Note that

$$\begin{aligned}\mathbf{E}(X \vee Y - X \wedge Y) &= \int_0^\infty \mathbf{P}\{X \vee Y - X \wedge Y > t\}\, dt \\ &= \frac{1}{\mu+\nu}\left(\frac{\nu}{\mu} + \frac{\mu}{\nu}\right).\end{aligned}$$

Therefore, as $X \wedge Y \sim \text{Expon}(\mu+\nu)$, we have

$$\mathbf{E}(X \vee Y) = \mathbf{E}(X \wedge Y) + \mathbf{E}(X \vee Y - X \wedge Y) = \frac{1}{\mu+\nu}\left(1 + \frac{\nu}{\mu} + \frac{\mu}{\nu}\right).$$

At first sight, this looks nothing like the answer we got in Example 4.2.3. Nonetheless, the are the same! Let us check that this is true. Note that

$$1 + \frac{\nu}{\mu} + \frac{\mu}{\nu} = \frac{\mu\nu + \nu^2 + \mu^2}{\mu\nu} = \frac{(\mu+\nu)^2 - \mu\nu}{\mu\nu}.$$

Thus

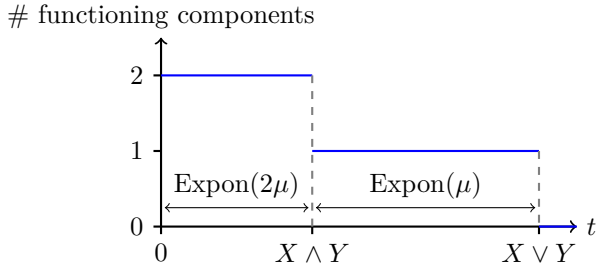
$$\frac{1}{\mu+\nu}\left(1 + \frac{\nu}{\mu} + \frac{\mu}{\nu}\right) = \frac{1}{\mu} + \frac{1}{\nu} - \frac{1}{\mu+\nu},$$

which is precisely the expression we obtained in Example 4.2.3.

The most interesting case is when the two components have the same average lifetimes, i.e., $\mu = \nu$. In this case, our computation reduces to

$$\mathbf{P}\{X \vee Y - X \wedge Y > t\} = e^{-\mu t},$$

so the time between failures $X \vee Y - X \wedge Y$ is an exponential $\text{Expon}(\mu)$ random variable! As in this case $X \wedge Y \sim \text{Expon}(2\mu)$, we have the following picture:



Again, this is completely intuitive. Initially, we have two functioning independent components with lifetime distribution $\text{Expon}(\mu)$. Thus the time until the first component fails is the minimum of the two lifetimes, which has distribution $\text{Expon}(2\mu)$. Once one of the components fails, only one functioning component remains. By the memoryless property of the exponential distribution, the remaining lifetime of this component is just the lifetime of a single component, which has distribution $\text{Expon}(\mu)$.

In this case where the two components have the same mean, we claim that something even more interesting happens: the time of first failure and the time between the two failures are *independent* random variables, that is,

$$X \wedge Y \perp\!\!\!\perp (X \vee Y - X \wedge Y).$$

To see this, we must show that

$$\mathbf{P}\{X \vee Y - X \wedge Y > t, X \wedge Y > s\} = \mathbf{P}\{X \vee Y - X \wedge Y > t\}\,\mathbf{P}\{X \wedge Y > s\}.$$

The ingredients of this computation are by now familiar to us:

$$\begin{aligned}
&\mathbf{P}\{X \vee Y - X \wedge Y > t, X \wedge Y > s\} \\
&= \mathbf{P}\{X - Y > t, Y > s, X > Y\} + \mathbf{P}\{Y - X > t, X > s, Y > X\} \\
&= \mathbf{P}\{X - Y > t, Y > s\} + \mathbf{P}\{Y - X > t, X > s\} \\
&= 2\,\mathbf{P}\{X - Y > t, Y > s\} \\
&= 2\int_0^\infty \mathbf{P}\{X - Y > t, Y > s | Y = u\}\,\mathbf{P}\{Y \in du\} \\
&= 2\int_0^\infty \mathbf{P}\{X - u > t, u > s\}\,\mathbf{P}\{Y \in du\} \\
&= 2\int_s^\infty \mathbf{P}\{X - u > t\}\,\mathbf{P}\{Y \in du\} \\
&= 2\int_s^\infty e^{-\mu(t+u)}\,\mu e^{-\mu u}\,du \\
&= e^{-\mu t} e^{-2\mu s} \\
&= \mathbf{P}\{X \vee Y - X \wedge Y > t\}\,\mathbf{P}\{X \wedge Y > s\}.
\end{aligned}$$

What have we done here? First, we split the probability according to whether X or Y fails first. Then, we notice that if $X - Y > t$ (X fails time t after Y), this already implies that $X > Y$ (Y fails first), so we can remove the latter from our probabilities. Next, we notice that in the present case X and Y are identical components, so the two probabilities must be the same. Finally, we used the conditioning formula as usual. Thus independence is established.

We have now obtained an extremely useful representation for the maximum of two i.i.d. exponential random variables with rate μ as the sum of an exponential random variable with rate 2μ (the time until the first failure), and an independent exponential random variable with rate μ (the time between failures). This makes certain computations extremely simple.

Example 4.4.2. We can immediately compute

$$\mathbf{E}(X \vee Y) = \mathbf{E}(X \wedge Y) + \mathbf{E}(X \vee Y - X \wedge Y) = \frac{1}{2\mu} + \frac{1}{\mu}.$$

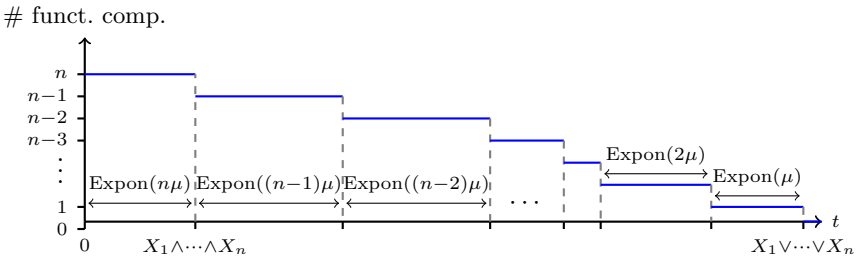
Similarly, using independence, we can write

$$\mathrm{Var}(X \vee Y) = \mathrm{Var}(X \wedge Y) + \mathrm{Var}(X \vee Y - X \wedge Y) = \frac{1}{(2\mu)^2} + \frac{1}{\mu^2}.$$

This would have been more work to compute directly!

The beauty of the ideas that we developed in this section is that they can be extended, in exactly the same manner, to maxima of more than two exponential random variables. Let $X_1, \ldots, X_n$ be independent $\mathrm{Expon}(\mu)$ random variables. The first failure occurs at the minimum of these times: this time has distribution $\mathrm{Expon}(n\mu)$ (as the minimum of exponential random variables is an exponential random variable whose rate is the sum of the rates of the components). At this point, one component has failed and $n-1$ components remain functioning. Therefore, the time until the next failure is an $\mathrm{Expon}((n-1)\mu)$ random variable, independent of the time until the first failure. Now there are $n-2$ functioning components remaining, etc.

Proceeding in this fashion, we find that the maximum $X_1 \vee X_2 \vee \cdots \vee X_n$ of independent exponential random variables with rate μ has the same distribution of the sum of independent exponential random variables with rates $n\mu$, $(n-1)\mu$, …, 3μ, 2μ, μ, as is illustrated in the following picture:



Once you have the right intuition about what is going on, this property of maxima of exponentials becomes very easy to remember. Using this representation, we can immediately compute the mean and variance of the maximum:

$$\mathbf{E}(X_1 \vee \cdots \vee X_n) = \sum_{k=1}^{n} \frac{1}{k\mu},$$

$$\mathrm{Var}(X_1 \vee \cdots \vee X_n) = \sum_{k=1}^{n} \frac{1}{(k\mu)^2}.$$

This would have been much more difficult to compute directly!

Remark 4.4.3. In this section, we did several computations that are a bit tricky. In this course, I am not expecting you to be able to come up with

such ideas on your own, at least not without significant help/hints. On the other hand, the intuition behind what is going on here is very clear, as is the final result—the computations were just needed to convince us that these results are correct! You should be able to use the representation of the maxima of independent exponentials that we learned here to do interesting computations.

5

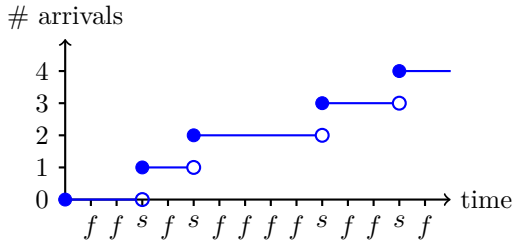
Poisson Processes

In section 2.4, we first introduced the basic model of arrivals in continuous time by taking the continuous time limit of a Bernoulli process. In this chapter, we are going to study such arrival processes and their properties systematically, and show how they can be used to answer various interesting questions.

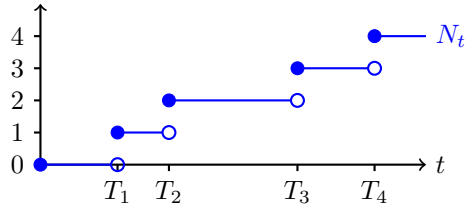
5.1 Counting processes and Poisson processes

The aim of this chapter is to model the process of arrivals over time. Such models arise in many different contexts: for example, customers arriving at a store; buses arriving at a bus stop; airplanes landing at the airport; individual photons emitted from a weak light source (such as a distant star); clicks on a Geiger counter; etc. We denote by N_t the number of arrivals by time t. Then the evolution of the number of arrivals over time is a *random process* $\{N_t\}_{t\geq 0}$, that is, a collection of random variables N_t indexed by time t.

The basic modelling assumption that we will make it that at every time, an arrival occurs independently and with equal probability. We saw in section 2.4 how this idea can be made precise by taking the continuous time limit of a Bernoulli process. Let us recall how this was done. First, we discretize time into n intervals per unit time. At the beginning of each interval, a customer arrives with probability p. Therefore, we can view the arrivals of customers at the beginning of each interval as a Bernoulli process. The number of arrivals over time in this discrete model is illustrated in the following figure:



In real life, however, arrivals do not only happen once a minute, or once a second: an arrival can occur at any time. To capture this, we have to take the limit of the discrete arrival process in the limit where we have increasingly many possible times of arrival in every unit of time. That is, we must let the number of intervals per unit time become infinite $n \to \infty$. We must be careful, however, to do this in a way that the average number of arrivals per unit time λ remains fixed: that is, the probability of arrival in each interval must be scaled as $p = \lambda/n$. When we take this limit correctly, the number of arrivals N_t as a function of time t will look something like this:



The (random) time T_i of the ith arrival can now be any positive number: the process $\{N_t\}_{t\geq 0}$ models arrivals in continuous time.

As you can see in the above figure, N_t should be really thought of as counting the arrivals as they occur: each time a new arrival occurs, we increase the value of N_t by one. A process of this kind is called a *counting process*.

Definition 5.1.1. *A random process $\{N_t\}$ is called a* counting process *if*

1. $N_0 = 0$;
2. N_t is increasing in t;
3. N_t increases by one every time that it changes.

Clearly, our arrivals process $\{N_t\}_{t\geq 0}$ must be a counting process. But there are many other counting processes as well! For example, consider the (nonrandom) process that increases by one precisely at the beginning of every minute. This process is called a *clock*. Much as that would be good for business, customers arriving at a store cannot be expected to arrive like clockwork; this is a different counting process. We should therefore figure out what properties characterize the special arrivals process that we obtained from the continuous time limit of a Bernoulli process as discussed above and in section 2.4.

Remark 5.1.2. Note that we have excluded the possibility that two or more arrivals occur simultaneously. If the arrivals occur independently in continuous time, then two simultaneous arrivals have zero probability of occuring: two people who choose their arrival time independently according to a continuous distribution will never arrive at precisely the same instant. Of course, there may be situations in which you would like to model the simultaneous arrivals

of two or more customers. In this case, you would use a more general type of counting process. The above definition is sufficient for our purposes, however.

Let us collect some useful properties of the arrival process $\{N_t\}_{t\geq 0}$ that are direct consequences of the continuous time limit.

a. We alreay computed the distribution of N_t in section 2.4:

$$N_t \sim \text{Pois}(\lambda t), \qquad t \geq 0.$$

To obtain this, we noted that in the discrete arrivals process with n arrivals per unit time, the number of arrivals by time t has the Binomial distribution $\text{Binom}(nt, \lambda/n)$. Taking the limit as $n \to \infty$ of this Binomial distribution was precisely how we derived the Poisson distribution in the first place.

b. For times $t \geq s$, the quantity

$$N_t - N_s = \#\text{ arrivals between times } s \text{ and } t$$

is called an *increment* of the process $\{N_t\}_{t\geq 0}$. In the discrete model, the increments have a very natural property: the number of arrivals between times s and t depends only on the Bernoulli trials between times s and t, and is therefore independent of the arrivals before time s (Example 2.1.6). When we take the continuous time limit, this property remains true:

$$N_t - N_s \perp\!\!\!\perp \{N_r\}_{r\leq s}, \qquad t \geq s.$$

This is called the *independent increments* property. It makes precise the idea that our arrivals are occuring independently at every time.

c. In a Bernoulli process, the number of successes in trial numbers $n+1, \ldots, m$ and the number of successes in trial numbers $1, \ldots, m-n$ are both sums of $m-n$ independent Bernoulli variables. Therefore, these two numbers of successes *have the same distribution*. When we take the continuous time limit, this property remains true:

$$N_t - N_s \text{ has the same distribution as } N_{t-s}, \qquad t \geq s.$$

This is called the *stationary increments* property. It makes precise the idea that there is equal probability of an arrival occuring at every time.

Remark 5.1.3. The stationary increments property states that the distribution of the number of arrivals in any time interval depends only on the length of that interval. Thus, for example, the same number of arrivals occur on average between 10 and 11AM as between 2 and 3AM. This might be realistic for counting radioactive emissions on a Geiger counter (atoms do not sleep), but is less realistic if we model customers arriving at a store: most stores close at night. Even when the store is open, there are some times of the day (for

example, around 5PM when people get off work) when there will be more customers on average than during other times. This sort of behavior is not included in the simplest model of arrivals that we are investigating here. Once we have understood how to model arrivals, however, we will be able to create more complicated models that include these sorts of time-dependent effects.

The three properties that we described above characterize precisely those counting processes that are based on the modelling assumption that arrivals occur independently and with equal probability at every time. We call counting processes of this kind *Poisson processes.*

Definition 5.1.4. *A counting process* $\{N_t\}_{t\geq 0}$ *is called a* Poisson process *with rate* $\lambda > 0$ *if it satisfies the following properties:*

1. $N_t - N_s \perp\!\!\!\perp \{N_r\}_{r\leq s}$ *for* $t \geq s$ *(independent increments).*
2. $N_t - N_s \sim \text{Pois}(\lambda(t-s))$ *for* $t \geq s$ *(stationary increments).*

Note that the rate λ of a Poisson process can be interpreted precisely as the average number of arrivals per unit time.

Remark 5.1.5. It turns out that it is not essential to assume that $N_t - N_s$ has a Poisson distribution: one can prove that every counting process with stationary independent increments is automatically a Poisson process (for *some* rate λ). However, as we automatically obtained the Poisson distribution from our construction of the continuous arrivals process, there is no harm in including it as part of our definition of the Poisson process.

Example 5.1.6. Consider a store where customers arrive according to a Poisson process $\{N_t\}_{t\geq 0}$ with rate λ. Let $t > s$ and $j \geq i$. The event that there are i customers arrived by time s and j customers by time t is $\{N_s = i, N_t = j\}$. Let us compute the probability that this occurs:

$$\begin{aligned}\mathbf{P}\{N_s = i, N_t = j\} &= \mathbf{P}\{N_s = i, N_t - N_s = j - i\} \\ &= \mathbf{P}\{N_s = i\}\,\mathbf{P}\{N_t - N_s = j - i\} \\ &= \frac{e^{-\lambda s}(\lambda s)^i}{i!} \cdot \frac{e^{-\lambda(t-s)}(\lambda(t-s))^{j-i}}{(j-i)!}.\end{aligned}$$

Note that N_s and N_t are *not* independent; we can therefore not split the probability as a product of the probabilities of $N_s = i$ and $N_t = j$! To solve the problem, we must reformulate the question in terms of the increments, so that we can use the independent increments property of the Poisson process. The second equality follows by the independent increments property, and the third equality is simply using the definition of the Poisson distribution.

Let us now investigate the time T_k at which the kth arrival occurs. Consider the first arrival T_1. Saying that $T_1 > t$ (the first arrival will occur in the future) is the same as saying that $N_t = 0$ (no arrivals have occured yet). Therefore,

$$\mathbf{P}\{T_1 > t\} = \mathbf{P}\{N_t = 0\} = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

We have therefore shown that $T_1 \sim \text{Expon}(\lambda)$. In precisely the same manner, we showed in section 2.4 that $T_k \sim \text{Gamma}(k, \lambda)$ for every $k \geq 1$. It will be very useful, however, to develop an alternative way to think about the consecutive arrival times of a Poisson process.

What can we say about the relationship between the random variables $T_1, T_2, \ldots$? Clearly, they are not independent: for example, if we know that $T_1 > t$, then we must certainly have $T_2 > t$ also; so having information about T_1 most certainly affects what we know about T_2. Recall, however, the following useful property of Bernoulli processes (see section 2.2). Even though the times of the first, second, third, ... success in a Bernoulli process are not independent, the times *between* consecutive successes are independent and have the same distribution. When we take the continuous time limit, this property remains true: the times between consecutive arrivals

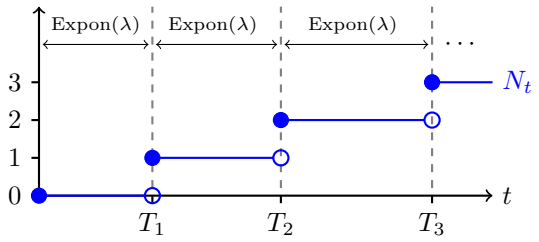
$$T_1,\ T_2 - T_1,\ T_3 - T_2,\ \ldots$$

are *independent and identically distributed.*

This implies, in particular, that

$$T_k = T_1 + (T_2 - T_1) + (T_3 - T_2) + \cdots + (T_k - T_{k-1})$$

can be written as a sum of k independent random variables with distribution $\text{Expon}(\lambda)$, as is illustrated in the following figure:



As we already showed that $T_k \sim \text{Gamma}(k, \lambda)$, this gives us a new and very useful interpretation of the Gamma distribution! For example, this implies

$$\mathbf{E}(T_k) = \mathbf{E}(T_1) + \mathbf{E}(T_2 - T_1) + \cdots + \mathbf{E}(T_k - T_{k-1}) = \frac{k}{\lambda},$$

and (using that the variance of the sum of independent random variables is the sum of the variances of those random variables)

$$\mathrm{Var}(T_k) = \mathrm{Var}(T_1) + \mathrm{Var}(T_2 - T_1) + \cdots + \mathrm{Var}(T_k - T_{k-1}) = \frac{k}{\lambda^2}.$$

Of course, we could have also computed the expectation and variance of a Gamma variable by writing this out as an integral over the Gamma density and integrating by parts numerous times. Here we arrived at the answer much more quickly by using just a bit of probabilistic insight!

5.2 Superposition and thinning

The goal of this section is to develop two important operations involving Poisson processes: superposition and thinning. Together, these two principles prove to be powerful tools to answer various different questions.

Before we can develop the superposition principle, we consider a simple property of Poisson random variables. This also provides a nice example of how Poisson processes can help us give very simple answers to certain questions.

Example 5.2.1 (Sum of independent Poisson variables). Let $X \sim \mathrm{Pois}(\mu)$ and $Y \sim \mathrm{Pois}(\nu)$ be independent Poisson random variables. What is the distribution of $X + Y$? We could write out the probabilities $\mathbf{P}\{X + Y = k\}$ in terms of the joint distribution of X and Y, and then compute an infinite sum. But Poisson processes provide us with a much quicker route to the answer.

Let $\{N_t\}_{t\geq 0}$ be a Poisson process with rate 1. Then $N_\mu \sim \mathrm{Pois}(\mu)$ has the same distribution as X, while $N_{\nu+\mu} - N_\mu \sim \mathrm{Pois}(\nu)$ has the same distribution as Y. Moreover, by the independent increments property, $N_{\nu+\mu} - N_\mu$ and N_μ are independent (as are X and Y). Therefore, $X+Y$ has the same distribution as $N_{\nu+\mu} - N_\mu + N_\mu = N_{\nu+\mu} \sim \mathrm{Pois}(\mu + \nu)$. That is, *the sum of independent Pois*(μ) *and Pois*(ν) *random variables has distribution Pois*$(\mu + \nu)$.

We now develop the superposition principle. To illustrate this principle, let us consider a simple example.

Example 5.2.2 (Men and women customers). Both men and women shop at a certain store (for example, a lingerie store), albeit at different rates. On average, μ men and ν women arrive per hour. The arrivals of men and women are modelled by independent Poisson processes. What can we say about the process that counts the total arrivals of all customers at the store?

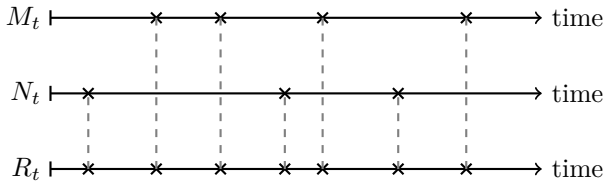
Let us define

$$M_t = \#\text{ men that arrived by time } t,$$
$$N_t = \#\text{ women that arrived by time } t,$$

so that $\{M_t\}_{t\geq 0}$ is a Poisson process with rate μ and $\{N_t\}_{t\geq 0}$ is a Poisson process with rate ν. Denote by

$$R_t = \#\text{ customers that arrived by time } t = M_t + N_t$$

the total number of customers. It is useful to think of these three counting processes in terms of their arrival times, as is illustrated in the following picture (the xs denote the times at which each process increases by one):



What can we say about the properties of the process $\{R_t\}_{t\geq 0}$?

a. Note that

$$R_t - R_s = (M_t - M_s) + (N_t - N_s).$$

As $\{M_t\}_{t\geq 0}$ is a Poisson process, the independent increments property implies that $M_t - M_s \perp\!\!\!\perp \{M_u\}_{u\leq s}$, while $M_t - M_s \perp\!\!\!\perp \{N_u\}_{u\leq s}$ as $\{M\}_{t\geq 0}$ and $\{N\}_{t\geq 0}$ are assumed to be independent. Exactly the same reasoning applies to $N_t - N_s$. Therefore, we have shown that

$$R_t - R_s \perp\!\!\!\perp \{R_u\}_{u\leq s}.$$

b. Similarly, note that $M_t - M_s$ and $N_t - N_s$ are independent Poisson random variables with means $\mu(t-s)$ and $\nu(t-s)$. Therefore,

$$R_t - R_s \sim \text{Pois}((\mu+\nu)(t-s)).$$

Thus we have shown that $\{R_t\}_{t\geq 0}$ is a Poisson process with rate $\mu + \nu$! This is intuitive: if men arrive at an average rate of μ per hour, and women arrive at a rate of ν per hour, then on average $\mu + \nu$ customers arrive per hour.

More generally, we have the following **principle of superposition**:

> If $N_t^1, \ldots, N_t^k$ are independent Poisson processes with rates $\mu_1, \ldots, \mu_k$, then $N_t^1 + \cdots + N_t^k$ is a Poisson process with rate $\mu_1 + \cdots + \mu_k$.

The principle of thinning is the precise opposite of superposition. To illustrate it, let us reconsider the above example from a different perspective.

Example 5.2.3 (Men and women customers revisited). On average, μ customers arrive at a store per hour. Each customer is (independently) a man with probability p and a woman with probability $1-p$. What can we say about the processes that count the arrivals of men and of women at the store?

Let us make some careful definitions. Let

$$R_t = \#\text{ customers that arrived by time } t$$

be a Poisson process with rate μ. Let also

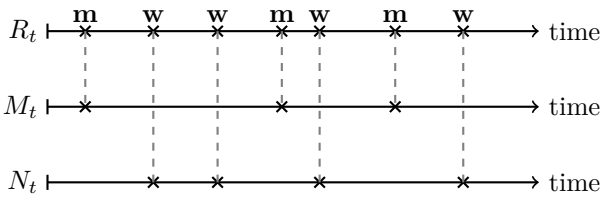
$$X_k = \text{gender of customer } k$$

be i.i.d. variables with $\mathbf{P}\{X_k = \mathbf{m}\} = p$ and $\mathbf{P}\{X_k = \mathbf{w}\} = 1-p$. Then

$$M_t = \#\text{ men that arrived by time } t = \sum_{k=1}^{R_t} \mathbf{1}_{\{X_k=\mathbf{m}\}},$$

$$N_t = \#\text{ women that arrived by time } t = \sum_{k=1}^{R_t} \mathbf{1}_{\{X_k=\mathbf{w}\}}.$$

This is illustrated in the following figure:



For sake of illustration, let us compute the distribution of M_t:

$$\begin{aligned}
\mathbf{P}\{M_t = k\} &= \sum_{r=0}^{\infty} \mathbf{P}\{M_t = k | R_t = r\}\, \mathbf{P}\{R_t = r\} \\
&= \sum_{r=0}^{\infty} \mathbf{P}\left\{ \sum_{i=1}^{R_t} \mathbf{1}_{\{X_i=\mathbf{m}\}} = k \,\middle|\, R_t = r \right\} \mathbf{P}\{R_t = r\} \\
&= \sum_{r=0}^{\infty} \mathbf{P}\left\{ \sum_{i=1}^{r} \mathbf{1}_{\{X_i=\mathbf{m}\}} = k \,\middle|\, R_t = r \right\} \mathbf{P}\{R_t = r\} \\
&= \sum_{r=0}^{\infty} \mathbf{P}\left\{ \sum_{i=1}^{r} \mathbf{1}_{\{X_i=\mathbf{m}\}} = k \right\} \mathbf{P}\{R_t = r\},
\end{aligned}$$

where we have used that R_t and X_i are independent. As $\mathbf{1}_{\{X_i=\mathbf{m}\}}$ are Bernoulli variables with success probability p, the sum of r such variables is a Binomial random variable with parameters r and p. Moreover, we know that R_t is Poisson with mean μt. Therefore,

$$\begin{aligned}
\mathbf{P}\{M_t = k\} &= \sum_{r=k}^{\infty} \binom{r}{k} p^k (1-p)^{r-k} \, \frac{e^{-\mu t}(\mu t)^r}{r!} \\
&= \sum_{r=k}^{\infty} \frac{r!}{k!(r-k)!} p^k (1-p)^{r-k} \, \frac{e^{-\mu t}(\mu t)^r}{r!} \\
&= \frac{e^{-\mu t}(p\mu t)^k}{k!} \sum_{r=k}^{\infty} \frac{((1-p)\mu t)^{r-k}}{(r-k)!} \\
&= \frac{e^{-\mu t}(p\mu t)^k}{k!} \, e^{(1-p)\mu t} = \frac{e^{-p\mu t}(p\mu t)^k}{k!}.
\end{aligned}$$

That is, $M_t \sim \text{Pois}(p\mu t)$. In fact, you can easily check that $\{M_t\}_{t\geq 0}$ has stationary independent increments, so that the counting process that counts only the men $\{M_t\}_{t\geq 0}$ is itself a Poisson process with rate $p\mu$! This is very intuitive: if customers arrive at an average rate of μ per hour, and each customer is a man with probability p, then there should be on average $p\mu$ men arriving per hour. In precisely the same way, we find that the number of women $\{N_t\}_{t\geq 0}$ is a Poisson process with rate $(1-p)\mu$.

In the present case, we can say even more: the processes $\{M_t\}_{t\geq 0}$ and $\{N_t\}_{t\geq 0}$ that count men and women are in fact *independent*! Let us verify, for example, that M_t and N_t are independent. As $M_t + N_t = R_t$, we have

$$\begin{aligned}
\mathbf{P}\{M_t = m, N_t = n\} &= \mathbf{P}\{M_t = m, R_t = m+n\} \\
&= \mathbf{P}\left\{ \sum_{i=1}^{R_t} \mathbf{1}_{\{X_i = \mathbf{m}\}} = m, \ R_t = m+n \right\} \\
&= \mathbf{P}\left\{ \sum_{i=1}^{m+n} \mathbf{1}_{\{X_i = \mathbf{m}\}} = m, \ R_t = m+n \right\} \\
&= \mathbf{P}\left\{ \sum_{i=1}^{m+n} \mathbf{1}_{\{X_i = \mathbf{m}\}} = m \right\} \mathbf{P}\{R_t = m+n\} \\
&= \binom{m+n}{m} p^m (1-p)^n \, \frac{e^{-\mu t}(\mu t)^{m+n}}{(m+n)!} \\
&= \frac{(m+n)!}{m!n!} \, p^m (1-p)^n \, \frac{e^{-\mu t}(\mu t)^{m+n}}{(m+n)!} \\
&= \frac{e^{-p\mu t}(p\mu t)^m}{m!} \, \frac{e^{-(1-p)\mu t}((1-p)\mu t)^n}{n!} \\
&= \mathbf{P}\{M_t = m\} \, \mathbf{P}\{N_t = n\}.
\end{aligned}$$

What can we conclude? If we have a Poisson process with rate μ, and we put each arrival randomly in two bins with probabilities p and $1-p$, then the arrivals in each bin follow an independent Poisson process with rates $p\mu$ and $(1-p)\mu$. This is called *thinning* (because you are "thinning" the arrivals by rejecting each one with some probability).

More generally, we have the following **principle of thinning**:

> If N_t is a Poisson process with rate μ, and if each arrival is independently assigned one of k lables with probabilities $p_1, \ldots, p_k$, then the counting processes $N_t^1, \ldots, N_t^k$ that count the arrivals of each type are independent Poisson processes with rates $p_1\mu, \ldots, p_k\mu$.

Note that the principle of thinning is exactly the converse of the principle of superposition: we can split a single Poisson process up into independent Poisson processes by thinning, and we can put them together again to form the original Poisson process by superposition.

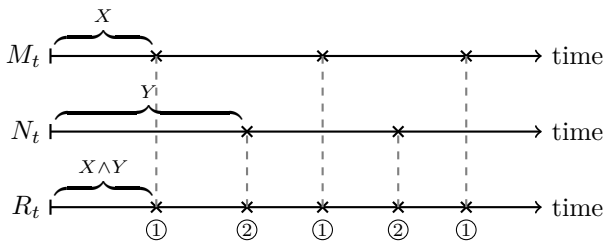
The combination of superposition and thinning is a surprisingly powerful tool. Let us use it to solve a number of interesting problems.

Example 5.2.4 (Exponential random variables). Let $X \sim \text{Expon}(\mu)$ and $Y \sim \text{Expon}(\nu)$ be independent random variables. We have previously already proved, by explicit computation, the following facts:

$$X \wedge Y \sim \text{Expon}(\mu + \nu), \qquad \mathbf{P}\{X \leq Y\} = \frac{\mu}{\mu + \nu}.$$

In this example, we will reprove these facts using Poisson processes. This requires less explicit computation, but more thinking: we must apply the superposition and thinning principles in a clever way.

Let $\{M_t\}_{t\geq 0}$ and $\{N_t\}_{t\geq 0}$ be independent Poisson processes with rates μ and ν, respectively. Then we can define X to be the first arrival time of $\{M_t\}$ and Y to be the first arrival time of $\{N_t\}$: these variables are independent and have precisely the desired distributions! Let us denote the total arrivals process by $R_t = M_t + N_t$. This is illustrated in the following figure:



Note that $X \wedge Y$ is precisely the first arrival time of $\{R_t\}$. By the superposition principle, the latter is a Poisson process with rate $\mu + \nu$. It therefore follows immediately that $X \wedge Y \sim \text{Expon}(\mu + \nu)$.

In the above figure, we have labelled each arrival of $\{R_t\}$ by which of the Poisson processes $\{M_t\}$ or $\{N_t\}$ it came from (① for $\{M_t\}$, ② for $\{N_t\}$). Let us now invert the picture. Suppose we start with the Poisson process $\{R_t\}$ with rate $\mu + \nu$, and assign each arrival an independent label X_k such that

$$\mathbf{P}\{X_k = ①\} = \frac{\mu}{\mu+\nu}, \qquad \mathbf{P}\{X_k = ②\} = \frac{\nu}{\mu+\nu}.$$

Let M_t be the number of ①-arrivals and let N_t be the number of ②-arrivals by time t. By the thinning principle, $\{M_t\}$ and $\{N_t\}$ are independent Poisson processes with rates μ and ν, respectively, and $R_t = M_t + N_t$. Therefore

$$\mathbf{P}\{X \le Y\} = \mathbf{P}\{X \wedge Y = X\} = \mathbf{P}\{\text{first label is } ①\} = \frac{\mu}{\mu+\nu},$$

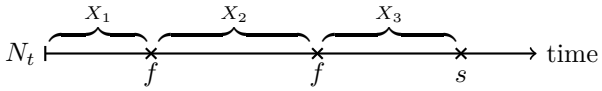
which is exactly what we set out to show.

Example 5.2.5 (Geometric sum of exponentials). Let $X_1, X_2, \ldots$ be independent exponential random variables with parameter λ, and let $K \sim \text{Geom}(p)$ be independent of $\{X_i\}$. What is the distribution of

$$S = \sum_{i=1}^{K} X_i \quad ?$$

You could answer this question by conditioning on K and performing an explicit computation (an infinite sum). However, using some probabilistic intuition, we can arrive at the answer much more quickly.

Let $\{N_t\}$ be a Poisson process with rate λ, and denote by T_k its kth arrival time. Then we can define $X_1 = T_1$ and $X_k = T_k - T_{k-1}$ for $k > 1$: the times between arrivals of a Poisson process are independent $\text{Expon}(\lambda)$ random variables. On the other hand, a geometric random variable $K \sim \text{Geom}(p)$ is the time of first success when we perform a sequence of independent trials, each of which is successful with probability p. Let us therefore attach to every arrival of $\{N_t\}$ an independent variable Y_k such that $\mathbf{P}\{Y_k = s\} = p$ and $\mathbf{P}\{Y_k = f\} = 1 - p$, and let K be the first arrival k such that $Y_k = s$. Then the random variable S is none other than the first time of an arrival of type s, as is illustrated in the following figure:



By the thinning principle, the arrivals of type s form a Poisson process with rate $p\lambda$, so the first arrival time S of this process must satisfy $S \sim \text{Expon}(p\lambda)$.

As we see in the previous example, it is helpful not just to remember the definitions of different distributions: you should also remember their meaning! If you know that exponential random variables are the times between arrivals of a Poisson process, and that a geometric random variable is the time of first success in independent trials, then you can use probabilistic reasoning to solve the above problem much more quickly than if you had tried to compute the distribution $\mathbf{P}\{S \le x\}$ by direct computation (which would, of course, give the same answer). The following is another example of this kind.

Example 5.2.6 (Waiting for the bus). Passengers arrive at a bus stop according to a Poisson process with rate λ. The bus arrives at time $T \sim \text{Expon}(\mu)$, independently from the arrivals of the passengers. What is the distribution of the number of passengers that get on the bus?

Let $\{N_t\}$ and $\{M_t\}$ be independent Poisson processes with rates λ and μ, respectively. N_t is the number of passengers that arrived at the bus stop by time t, and M_t is the number of buses that arrived by time t. The problem statement only refers to one bus, so we simply let T be the first arrival time of $\{M_t\}$ (which is indeed exponential with parameter μ).

Let $R_t = N_t + M_t$ count the total arrivals of people and buses at the stop. By the superposition principle, this is a Poisson process with rate $\lambda + \mu$. We can generate N_t and M_t from R_t by the thinning principle: if every arrival of R_t is independently chosen to be a person with probability $\frac{\lambda}{\lambda+\mu}$ and a bus with probability $\frac{\mu}{\lambda+\mu}$, then we can let N_t be the number of people and M_t be the number of buses that arrived by time t. But then it is easy to compute

$$\begin{aligned}
&\mathbf{P}\{k \text{ passengers get on bus}\} \\
&= \mathbf{P}\{\text{first } k \text{ arrivals are people, } (k+1)\text{st arrival is a bus}\} \\
&= \left(\frac{\lambda}{\lambda+\mu}\right)^k \frac{\mu}{\lambda+\mu}.
\end{aligned}$$

That is, we have shown that the number of passengers that get on the bus plus one is a geometric random variable with parameter $\frac{\mu}{\lambda+\mu}$.

To conclude this section, let us consider a more tricky example where the properties of Poisson processes make our life much simpler.

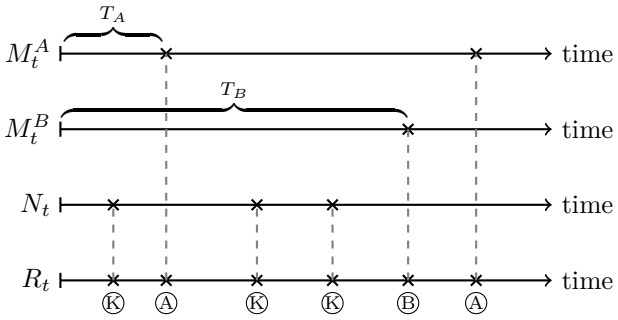
Example 5.2.7 (Kidney transplants). Two patients A and B are in need of a kidney transplant. Kidneys from organ donors are a scarce commodity: they arrive only occasionally, according to a Poisson process with rate λ. If the patients A and B do not get a new kidney, they will die from kidney failure at exponentially distributed times with parameters μ_A and μ_B, respectively (of course, each patient and the arrivals of the kidneys are independent).

Suppose that A is first on the waiting list for kidney transplants, and B is second. When a new kidney arrives, it is given to the patient who is highest on the list. When a patient gets a kidney or dies, he is removed from the list. What is the probability that B gets a new kidney before his old kidney fails?

Let us set up this problem using Poisson processes. As in the previous examples, we can assume that the lifetimes T_A and T_B of patients A and B are the first arrival times of Poisson processes $\{M_t^A\}$ and $\{M_t^B\}$ with rates μ_A and μ_B, respectively. Let $\{N_t\}$ be a Poisson process with rate λ that counts the arrivals of kidneys, and assume that the three Poisson processes that we have defined are independent. By the superposition principle, the total arrivals process $R_t = N_t + M_t^A + M_t^B$ is a Poisson process with rate $\lambda + \mu_A + \mu_B$. Moreover, by the thinning principle, we can define each of the processes $\{N_t\}, \{M_t^A\}, \{M_t^B\}$ by randomly labelling the arrivals of $\{R_t\}$ as

- Ⓚ if the arrival is a kidney (with probability $\frac{\lambda}{\lambda+\mu_A+\mu_B}$),
- Ⓐ if the arrival is a death of A (with probability $\frac{\mu_A}{\lambda+\mu_A+\mu_B}$),
- Ⓑ if the arrival is a death of B (with probability $\frac{\mu_B}{\lambda+\mu_A+\mu_B}$).

This is illustrated in the following figure.



We now want to compute the probability of the event

$$E = \{B \text{ gets a kidney before he dies}\}.$$

How should we think about this event? In order for B to get a kidney, first A must be removed from the waiting list (as B is lower on the list than A). Therefore, the first arrival of the process $\{R_t\}$ must be either Ⓐ or Ⓚ. In either case, A is removed from the waiting list, and what remains is for B to get a kidney before he dies. We can therefore write $E = E_1 \cap E_2$, where

- $E_1 = \{$the first arrival of R_t is of type Ⓐ or Ⓚ$\}$.
- $E_2 = \{$in the subsequent arrivals, Ⓚ arrives before Ⓑ$\}$.

We now make an important observation: *the events E_1 and E_2 are independent*. Indeed, by the thinning principle, the label of every arrival of $\{R_t\}$ is independent. Therefore, as E_1 only depends on the label of the first arrival and E_2 only depends on the labels of the second, third, fourth, ... arrivals, these two events must be independent. We can therefore write

$$\mathbf{P}(E) = \mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1)\mathbf{P}(E_2).$$

We can immediately compute

$$\mathbf{P}(E_1) = \frac{\lambda}{\lambda+\mu_A+\mu_B} + \frac{\mu_A}{\lambda+\mu_A+\mu_B},$$

and it remains to compute the probability of E_2.

To this end, we first make another useful observation: we claim that

$$\mathbf{P}(E_2) = \mathbf{P}\{\text{Ⓚ arrives before Ⓑ}\}.$$

Indeed, as the labels of the arrivals of R_t are independent and identically distributed, the probability that Ⓚ arrives before Ⓑ in arrivals $2, 3, 4, \ldots$ is the same as the probability that Ⓚ arrives before Ⓑ in arrivals $1, 2, 3, \ldots$

Now we can easily finish the proof by applying again the thinning principle. As we no longer care about the arrivals of type Ⓐ, we now only consider the arrival process $S_t = N_t + M_t^B$ of arrivals of type Ⓑ an Ⓚ. We therefore immediately compute using the thinning principle

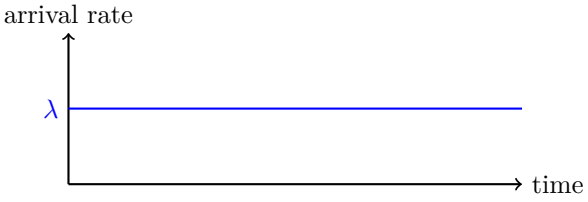
$$\mathbf{P}(E_2) = \mathbf{P}\{\text{the first arrival of } S_t \text{ is of type Ⓚ}\} = \frac{\lambda}{\lambda + \mu_B}.$$

Putting everything together, we have shown that

$$\mathbf{P}\{B \text{ gets a kidney before he dies}\} = \frac{\lambda + \mu_A}{\lambda + \mu_A + \mu_B} \frac{\lambda}{\lambda + \mu_B}.$$

5.3 Nonhomogeneous Poisson processes

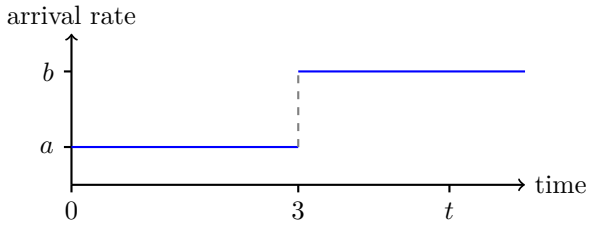
Let $\{N_t\}$ be a Poisson process. Then the distribution of the number of arrivals $N_t - N_s$ between times s and t depends only on the length of the time interval $t - s$. That is, Poisson processes, by definition, have *stationary increments*. In particular, the average number of arrivals per unit time is constant:



This could be a natural modelling assumption for some random phenomena, such as the emission of photons from a laser. In other cases, however, the stationary increments property may be less realistic. For example, if we model the number of arrivals of customers at a store by a Poisson process, then the average number of arrivals between 5PM and 6PM would be the same as the average number of arrivals between 3AM and 4AM. This seems unlikely, as the store is probably closed at night (in which case no customers arrive at all); and even if the store is open 24/7, there are likely many fewer customers who shop in the middle of the night than during the 5-6PM rush hour.

To create more realistic models, we must often allow the arrival rate to change over time. Let us begin by investigating the simplest possible example, which we will subsequently expand into a general definition.

Example 5.3.1 (Two arrival rates). Let us make a simple model where the arrival rate can take one of two values a and b. Up to time 3, customers arrive according to a Poisson process with rate a. After time 3, customers arrive according to a Poisson process with rate b. That is, we have:



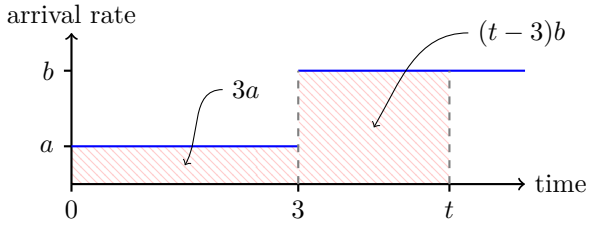
In this model, what is the distribution of the number N_t of customers that arrived by time $t > 3$? Note that we can write

$$N_t = N_3 + (N_t - N_3).$$

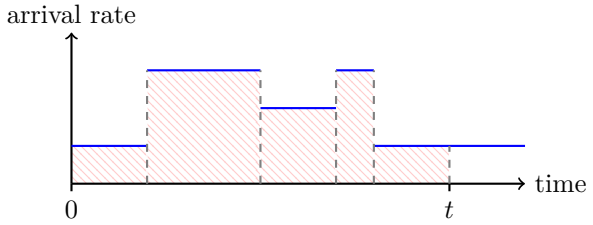
Before time 3, customers arrive with rate a, so the number of customers that arrived by time 3 is given by $N_3 \sim \text{Pois}(3a)$. After time 3, customers arrive with rate b, so the number of customers that arrive between times 3 and t is given by $N_t - N_3 \sim \text{Pois}((t-3)b)$. As customers are still arriving independently at each time, and so N_3 and $N_t - N_3$ must be independent. It follows that

$$N_t \sim \text{Pois}(3a + (t-3)b).$$

Notice a particularly useful fact: the average number $3a + (t-3)b$ of customers that arrived by time t is precisely the area under the arrival rate function!

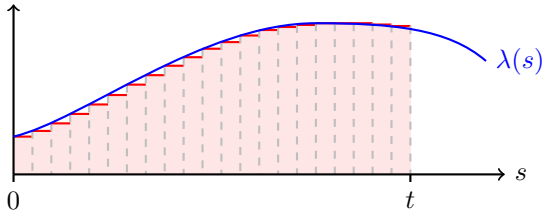


The above example can be easily generalized to the case where the arrival rate takes a finite number of values, as in the following figure:



It is an easy exercise to verify that the number of customers N_t that arrived by time t has the Poisson distribution whose mean is given by the area under the arrival rate function: all you need to do is to add up the number of customers that arrive in each time interval where the arrival rate is constant.

We can now generalize this idea. In principle, we can choose any nonnegative function $\lambda(t)$ to define the arrival rate at every time t. Then the number of customers that arrived by time t is Poisson distributed whose mean is given by the area under the arrival rate function $\int_0^t \lambda(s)ds$. This can be justified by approximating the function λ by a piecewise constant function, exactly as we did when we defined the expectation of a continuous random variable:



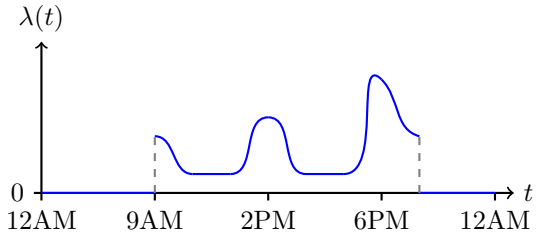
This idea provides us with a very general model of independent arrivals whose arrival rate may be time-dependent. We formalize this idea as follows.

Definition 5.3.2. *A counting process* $\{N_t\}_{t\geq 0}$ *is called a* nonhomogeneous Poisson process *with* rate function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ *if it satisfies*

1. $N_t - N_s$ *and* $\{N_r\}_{r\leq s}$ *are independent for* $t > s$;
2. $N_t - N_s \sim \text{Pois}(\int_s^t \lambda(u)du)$ *for* $t > s$.

Notice that a nonhomogeneous Poisson process still has independent increments (the first property), just like an ordinary Poisson process: this encodes the fact that the arrivals occur independently at each time. On the other hand, the average number of arrivals per unit time is no longer constant, but may change over time according to the arrival rate function λ. Thus a nonhomogeneous Poisson process differs from the ordinary Poisson process in that it is no longer required to have stationary increments. Of course, the ordinary Poisson process is recovered as a special case of a nonhomogeneous Poisson process if we take the arrival rate function to be constant $\lambda(t) \equiv \lambda$.

Example 5.3.3 (Customers at a store). The arrival rate function of customers arriving at a store might look something like this:



The store is closed from 9PM-9AM, so the arrival rate during this period is zero. During opening hours, the arrival rate depends on the time of day: there is a small rush around opening time and lunchtime, and a larger rush from 5PM until closing time. Of course, this plot is entirely fabricated. In practice, if you were modelling the arrival rate of customers at a particular store, you would establish the arrival rate function experimentally: you first collect customer arrival data over the course of a few weeks, and then fit the rate function to the data. (For example, you could divide the day into 15 minute intervals, and compute the average number of custumers arriving in each interval over all the days for which you have data.)

Many of the properties of ordinary Poisson processes can be generalized to nonhomogeneous Poisson processes. For sake of illustration, let us conclude by deriving one simple by useful property.

Example 5.3.4 (First arrival time). Let $\{N_t\}$ be a nonhomogeneous Poisson process with rate function $\lambda(t)$. Let T be the time of first arrival. What is the distribution of T? We can use exactly the same proof as we did for the ordinary Poisson process: we note that the event $\{T > t\}$ that the first arrival time is to the future of time t coincides with the event $\{N_t = 0\}$ that no customers have yet arrived by time t. Therefore,

$$\mathbf{P}\{T > t\} = \mathbf{P}\{N_t = 0\} = e^{-\int_0^t \lambda(u)du}.$$

In other words, we see that T is a lifetime with hazard rate function $\lambda(t)$!

We can therefore view any lifetime as the first arrival time of a nonhomogeneous Poisson processes. (As a sanity check, we note that if the arrival rate function is constant $\lambda(t) \equiv \lambda$, then we recover the property of the ordinary Poisson process that the first arrival time is exponentially distributed.)

6

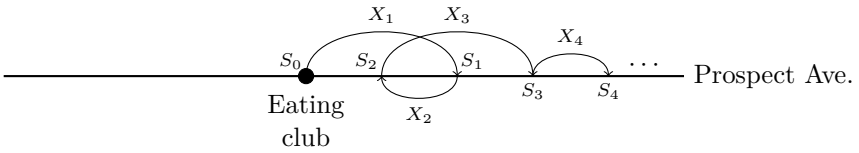
Random Walks

In this chapter, we investigate the simplest possible model of random motion that arises in numerous applications from physics to gambling: the *random walk*. While random walks are very simple processes, we will see that we can answer some interesting questions using new tools. Random walks also form a basic starting point for the investigation of both Brownian motion and Markov chains, which we will undertake in the following chapters.

6.1 What is a random walk?

Before giving a precise definition of random walk, let us first consider two examples. We will revisit these examples throughout this chapter.

Example 6.1.1 (Drunk on Prospect Ave.). After a productive evening at your eating club, you depart in a less than sober state. As a result, you wander around drunkenly on Prospect Ave. First, you might randomly make a step to the right. Then, forgetting where you are, you might make a step to the left. Totally disoriented, you make another random step... For example, for a few steps you might end up walking something like this:



Let us model this as a random process. The location of the eating club is S_0 (your initial position after zero steps). In the kth step, you move by an amount X_k, which can be positive or negative (depending on whether you move left or right). Your position S_n after n steps is therefore given by

$$S_n = S_0 + X_1 + X_2 + \cdots + X_n.$$

As you are drunk, you forget completely where you are after every step. We therefore model the steps $X_1, X_2, \ldots$ as independent and identically distributed (i.i.d.) random variables that are independent of the initial position S_0. This gives us our first example of a random walk.

Example 6.1.2 (Gambling). Suppose that you play the following game with your friend. In every round, each player bets \$1 and a fair coin is flipped independently. If the coin comes up heads, then you win, that is, you get back your \$1 and you also get the \$1 of your friend. If the coin comes up tails, then your friend wins the money. You repeatedly play rounds of this game, until either of you decides that it is time to cut your losses.

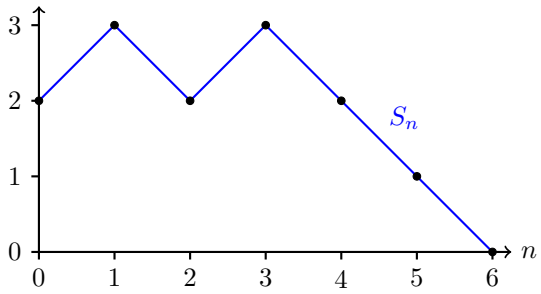
To model your wealth during the game as a random process, let us suppose that you start the game with initial wealth $\$S_0$. Note that your *net gain* in each round is either $+1$ or -1. Let X_i be your net gain in the ith round. Since the coin is fair, you lose or gain \$1 with equal probability

$$\mathbf{P}\{X_i = +1\} = \mathbf{P}\{X_i = -1\} = \frac{1}{2},$$

and we assume that $S_0, X_1, X_2, \ldots$ are independent. Your wealth S_n after the nth round is therefore given by

$$S_n = S_0 + X_1 + X_2 + \cdots + X_n.$$

One possible outcome of a few rounds of this game is illustrated here:



This gives us another natural example of a random walk.

In both these examples, we make a sequence of i.i.d. steps X_k starting at an initial point S_0. Such processes are called *random walks*. The case where the random walk makes only steps of size one, as in the second example, is particularly important; this is called a *simple* random walk.

Definition 6.1.3. *A* random walk $\{S_n\}_{n\geq 0}$ *is the random process*

$$S_n = S_0 + X_1 + \ldots + X_n,$$

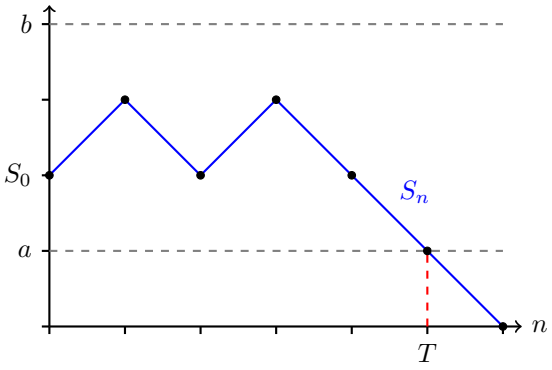
where $X_1, X_2, \ldots$ are i.i.d. random variables independent of S_0.

If S_0 is integer and the steps X_i only take the values ± 1, the random walk is called simple *(in this case, S_n is always integer as well). A simple random walk is called* symmetric *if* $\mathbf{P}\{X_i = +1\} = \mathbf{P}\{X_i = -1\} = \frac{1}{2}$.

The drunkard's walk on Prospect Ave. is an example of a random walk. The wealth process of a gambler is an example of a symmetric simple random walk.

6.2 Hitting times

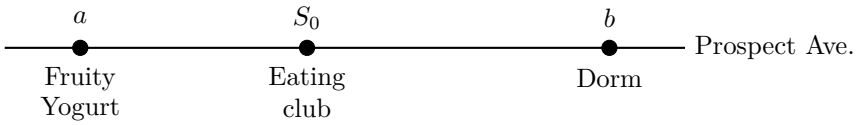
For the rest of this section, suppose that $\{S_n\}$ is a symmetric simple random walk. Let us fix two integers a, b such that $a \leq S_0 \leq b$. The *hitting time* T of the barriers $\{a, b\}$ by the random walk $\{S_n\}$ is the first time that the random walk hits a or b, as is illustrated in the following figure:



Formally, we define the hitting time as follows:

$$T = \min\{n \geq 0 \colon S_n = a \text{ or } S_n = b\}.$$

Example 6.2.1 (Drunk on Prospect Ave.). Suppose that your dorm is on the right end of the Propsect Street (at point b), while the Fruity Yogurt store is at the left end (at point a), as illustrated in the following figure:



When you depart drunkenly from your eating club (at point S_0), you perform a random walk along Prospect Ave. You may or may not eventually end up either in your dorm (where you can sleep it off), or the Fruity Yogurt store

(when you can sober up with some yummy non-fat yogurt). Let T be the time at which you first arrive at one of these two locations. The following natural questions arise immediately:

- Will you ever reach the Fruity Yogurt store or the dorm (i.e., is $T < \infty$)? Or is it possible that you will wander drunkenly back and forth on Prospect Ave. forever, without ever reaching one of the two endpoints?
- If you do reach one of the endpoints, how long does this take on average? And how likely are you to end up at the dorm vs. the Fruity Yogurt store?

We will shortly obtain answers to all these questions.

Example 6.2.2 (Gambling). Suppose that you start the gambling game with $\$0 < \$S_0 < \$100$. The game continues until either you are bankrupt $S_n = 0$, in which case you have no more money to wager, or when your wealth reaches $S_n = 100$, at which point you are declared the winner and your friend loses interest in throwing away more of his money. You might want to know:

- Will the game ever end, or can it go on forever?
- If it ends, what is the probability that you will win the game?
- How long will it take on average for the game to end?

We will now proceed to answer these questions.

Will we ever hit a boundary?

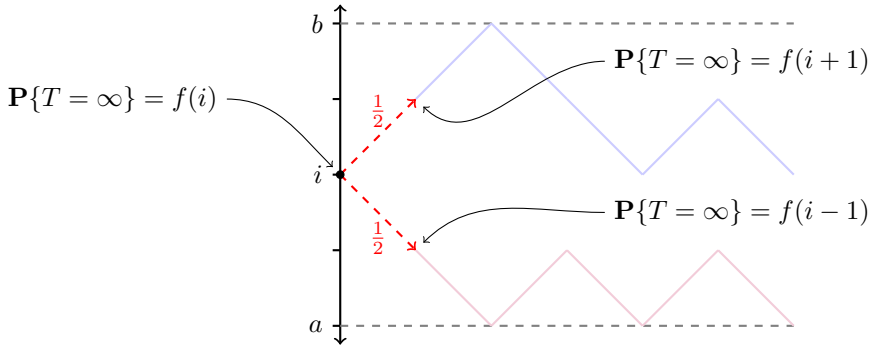
The first question we must investigate is whether the random walk will in fact ever hit the boundary $\{a, b\}$, that is, whether $T < \infty$. Of course, the probability of ever hitting the boundary might in principle depend on the starting point S_0. Let us therefore introduce the function

$$f(i) := \mathbf{P}\{T = \infty | S_0 = i\}, \qquad a \le i \le b,$$

that is, $f(i)$ is the probability that we never hit either of the boundaries a or b (in other words, the random walk stays forever strictly between these two boundaries), given that we started at the point $S_0 = i$. We would like to compute these probabilities. To this end, we will introduce a very useful tool for studying the behavior of random walks: the *first step analysis*.

Let us first explain the idea behind first step analysis. Suppose that we start at a point $a < i < b$ strictly between the boundaries. Then the earliest we could hit the boundary is after one step of the random walk. Moreover, as this is a simple symmetric random walk, there are only two possible locations where we can be after one step, each occurring with equal probability:

Because the steps of the random walk are i.i.d., the random walk simply restarts after every step at its current position. So, for example, if in the first step we happen to go from $i \to i+1$, then from that point onwards the random walk simply behaves as a random walk that is started from the point $i+1$. In particular, the first step is $i \to i+1$, then the probability of never hitteng the boundary afterwards is $f(i+1)$, and similarly for the other possible step:



By putting these facts together, we can write an equation for $f(i)$ in terms of $f(i+1)$ and $f(i-1)$. Once we get this equation, we can solve it to find $f(i)$!

Let us now work out this idea mathematically. Fix $a < i < b$. If we start at $S_0 = i$, then either $S_1 = i+1$ or $S_1 = i-1$ at time one. By conditioning,

$$\begin{aligned} f(i) = {} & \mathbf{P}\{T = \infty | S_1 = i+1, S_0 = i\}\, \mathbf{P}\{S_1 = i+1 | S_0 = i\} \\ & + \mathbf{P}\{T = \infty | S_1 = i-1, S_0 = i\}\, \mathbf{P}\{S_1 = i-1 | S_0 = i\}. \end{aligned}$$

Let us compute each of these four probabilities. Note that

$$\mathbf{P}\{S_1 = i+1 | S_0 = i\} = \mathbf{P}\{X_1 = +1 | S_0 = i\} = \mathbf{P}\{X_1 = +1\} = \frac{1}{2},$$

since X_1 and S_0 are independent. Similarly,

$$\mathbf{P}\{S_1 = i-1 | S_0 = i\} = \frac{1}{2}.$$

On the other hand, we have already argued that

$$\mathbf{P}\{T = \infty | S_1 = i+1, S_0 = i\} = \mathbf{P}\{T = \infty | S_0 = i+1\} = f(i+1),$$
$$\mathbf{P}\{T = \infty | S_1 = i-1, S_0 = i\} = \mathbf{P}\{T = \infty | S_0 = i-1\} = f(i-1).$$

Putting these pieces together, we get the equation

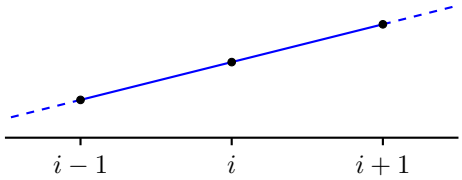
$$f(i) = \frac{1}{2} f(i+1) + \frac{1}{2} f(i-1), \quad a < i < b.$$

On the other hand, suppose that we start at $S_0 = a$ or $S_0 = b$. Then we are already starting at the boundary at the outset, so the probability that we never hit the boundary is zero. In particular, this implies that $f(a) = f(b) = 0$. Our first step analysis therefore leaves us with the following equations:

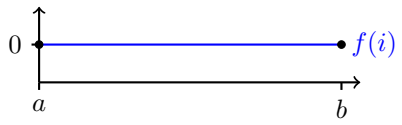
$$\begin{cases} f(i) = \frac{1}{2} f(i+1) + \frac{1}{2} f(i-1), & a < i < b, \\ f(a) = f(b) = 0. \end{cases}$$

This is a system of linear equations that we should be able to solve!

There is a particularly neat trick that allows us to solve these equations easily. Note that, according to our equation, the value $f(i)$ must lie precisely half-way between the values $f(i+1)$ and $f(i-1)$. This implies that these values must lie on a straight line:



Therefore, the function $i \mapsto f(i)$ must be linear. However, the only straight line that is 0 at both a and b is the function $f(i) = 0$:



We have therefore shown that $f(i) = 0$ for all $a \le i \le b$, or, in other words,

$$\mathbf{P}\{T = \infty | S_0 = i\} = 0 \quad \text{for all } a \le i \le b.$$

In particular, we have shown that *the random walk always hits a or b eventually, regardless of its starting point $a \le S_0 \le b$.*

Which boundary will we hit?

Now that we know that we will always eventualy hit one of the boundaries a or b, we might wonder which one we hit first? For example, in the gambling problem, the fact that $T < \infty$ is saying that you will eventually either win \$100 or go bankrupt. In practice, you are probably quite interested in which of these outcomes is more likely to occur...

To gain insight into this problem, let us phrase it in mathematical terms. At the first time T at which we reach either a or b, the value S_T must obviously be either a or b. We are interested in computing the probability that $S_T = b$, say (in which case the random walk reaches b before it reaches a). As this probability could again depends on where we start the random walk, let us define the function

$$r(i) := \mathbf{P}\{S_T = b | S_0 = i\}.$$

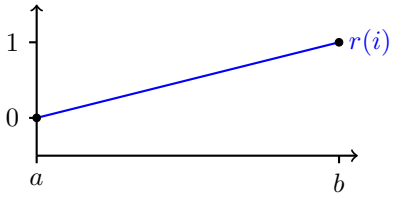
To compute $r(i)$, we can again use the first step analysis. If $a < i < b$, then we can hit one of the boundaries at the earliest after one step. After one step, we are either at $i+1$ or $i-1$ with equal probability, and the probability that we will hit b before a is then given by $r(i+1)$ and $r(i-1)$, respectively. Arguing exactly as above, we therefore find that

$$r(i) = \frac{1}{2} r(i+1) + \frac{1}{2} r(i-1), \quad a < i < b.$$

What is different here are the boundary conditions. In the present case, if we start at $S_0 = b$, then we obviously reach b before a (as we start already at b!), so $r(b) = 1$. On the other hand, if $S_0 = a$, then we obviously reach a before b, so $r(b) = 0$. We therefore obtain the following linear equations:

$$\begin{cases} r(i) = \frac{1}{2} r(i+1) + \frac{1}{2} r(i-1), & a < i < b, \\ r(a) = 0, \quad r(b) = 1. \end{cases}$$

In particular, this implies that $i \mapsto r(i)$ is a straight line with $r(a) = 0$ and $r(b) = 1$, and there can only be one such line:



The formula for this solution is given by

$$\mathbf{P}\{S_T = b | S_0 = i\} = r(i) = \frac{i-a}{b-a}, \qquad a \le i \le b.$$

Note in particular that the closer we start to b, the more likely we are to hit b before a, while the closer we start to a, the more likely we are to hit a before b. This makes perfect sense intuitively!

Example 6.2.3 (Gambling). We run the gambling game as in Example 6.2.2, that is, we continue playing until you either go bankrupt or have reached a wealth of \$100. Let T be the time when the game ends. We have shown that $T < \infty$ (the game ends eventually), and that the probability that you reach \$100 is

$$\mathbf{P}\{S_T = 100 | S_0 = i\} = \frac{i}{100}, \quad 0 \leq i \leq 100.$$

Therefore, the richer you are, the more likely you are to get even richer; while the poorer you are, the more likely you are to go bankrupt. This is a time-tested principle on which casinos are based.

Depending on your political leanings, it may seem unfair to you that the rich get richer, and the poor get poorer. However, in a sense, the form of gambling we are modelling here is really a *fair game*. To see why, let us compute the expected amount of money we have when the game ends:

$$\mathbf{E}(S_T | S_0 = i) = 0 \times \mathbf{P}\{S_T = 0 | S_0 = i\} + 100 \times \mathbf{P}\{S_T = 100 | S_0 = i\} = i.$$

This means that, *on average*, you have just as much money when the game ends as you started with! In this sense, the game is eminently fair: on average, neither you or your opponent has any advantage. This is in no way in contradiction to the "rich get richer" property that we saw above: gambling is inherently a risky business. If you are risk-averse, you should not gamble.

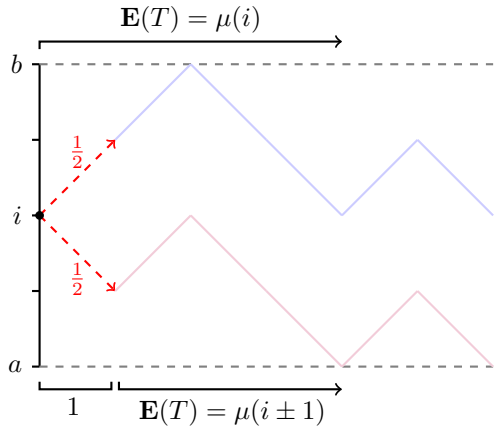
How long will it take?

Beside knowing which boundary we will hit, we would also like to know how long it takes, on average, to hit the boundary. How many rounds must you gamble, on average, before you either hit the jackpot or go bankrupt? How long to you have to wander around drunkenly, on average, before you reach your dorm or the Fruity Yogurt store? These questions are at least as interesting as the ones we have already answered.

To answer such questions, let us define the expected hitting time

$$\mu(i) := \mathbf{E}(T | S_0 = i).$$

We would like to compute $\mu(i)$ for every $a \leq i \leq b$. To this end, we will again perform first step analysis. If $a < i < b$, then we can hit one of the boundaries at the earliest after one step. After one step, we are either at $i+1$ or $i-1$ with equal probability, and the expected *remaining* time until we hit the boundary is given by $\mu(i+1)$ and $\mu(i-1)$, respectively. Thus the *total* time until we hit the boundary after moving to $i+1$ or $i-1$ in the first step is $1+\mu(i+1)$ or $1+\mu(i-1)$, respectively. This is illustrated in the following figure:



By the same argument as in the our previous applications of first step analysis, we obtain the following equation for $a < i < b$:

$$\begin{aligned}\mu(i) &= \frac{1}{2}(\mu(i+1)+1) + \frac{1}{2}(\mu(i-1)+1) \\ &= 1 + \frac{1}{2}\mu(i+1) + \frac{1}{2}\mu(i-1).\end{aligned}$$

On the other hand, if $S_0 = b$ or $S_0 = b$, then we already start at the boundary and thus $T = 0$ This evidently implies that on the boundary $\mu(a) = \mu(b) = 0$. We therefore obtain the following linear equations:

$$\begin{cases} \mu(i) = 1 + \frac{1}{2}\mu(i+1) + \frac{1}{2}\mu(i-1), & a < i < b, \\ \mu(a) = \mu(b) = 0. \end{cases}$$

Because of the extra $1+$ in the equation for $\mu(i)$, the solution of this equation is no longer a straight line as in our earlier computations. We must therefore find a different way of solving this equation.

It is convenient to rearrange our equation as follows. If we write $\mu(i) = \frac{1}{2}\mu(i) + \frac{1}{2}\mu(i)$, and then move one term from the left side of the equation to the right, and one term from the right side to the left, we obtain

$$\mu(i+1) - \mu(i) = \mu(i) - \mu(i-1) - 2$$

(check that this is correct!) We know that $\mu(a) = 0$. We do not know the value of $\mu(a+1) - \mu(a)$, so let us call this unknown quantity u for the time being (we will solve for u later). At this point, however, the above equation allows us to compute $\mu(a+2) - \mu(a-1)$, and then $\mu(a+3) - \mu(a+2)$, etc. Let is make a list of all of the quantities we can compute in this way:

$$\begin{aligned}
\mu(a) &= 0 \\
\mu(a+1) - \mu(a) &= u \\
\mu(a+2) - \mu(a+1) &= u - 2 \\
\mu(a+3) - \mu(a+2) &= u - 4 \\
&\dots \\
\mu(a+i) - \mu(a+i-1) &= u - 2(i-1)
\end{aligned}$$

If we add up all the quantities on the left- and right-hand sides, we obtain

$$\begin{aligned}
\mu(a+i) &= iu - 2(0 + 1 + \ldots + (i-1)) \\
&= iu - (i-1)i \\
&= i(u - i + 1).
\end{aligned}$$

We still do not know what u is. However, we have not yet used our other boundary condition $\mu(b) = 0$; this allows us to solve for u:

$$\mu(b) = \mu(a + (b-a)) = (b-a)(u - b + a + 1) = 0$$

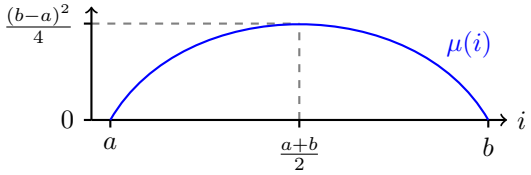
implies that $u = b - a - 1$. We have therefore shown that

$$\mu(a+i) = i(b-a-i),$$

or, equivalently,

$$\mathbf{E}(T|S_0 = i) = \mu(i) = (i-a)(b-i) \quad \text{for } a \le i \le b.$$

Thus the expected time to hit the boundary looks something like this.



Note that the closer one starts to one of the boundaries, the shorter it takes on average to hit the boundary. Conversely, the average time to hit the boundary is maximized if we start exactly half-way between the two boundaries. Just like our previous results, this makes perfect intuitive sense!

6.3 Gambler's ruin

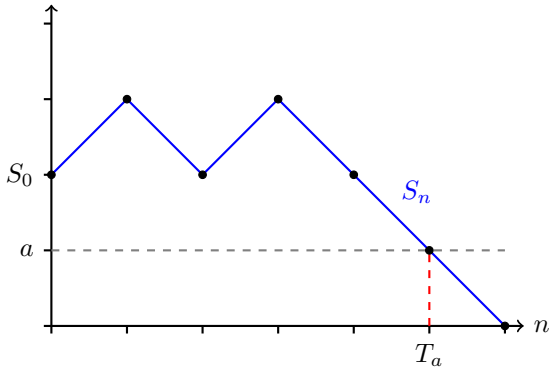
Suppose we play a gambling game where in every round you win or lose \$1 with equal probability, as in the previous section. You start initially with $\$S_0$.

If at any time your wealth S_n reaches zero, you go bankrupt and must stop playing. What is the probability that you will ever go bankrupt?

At first sight, this looks very much like the hitting time problems that we solved in the previous section. Let us define for every integer a

$$T_a := \min\{n : S_n = a\}.$$

Thus T_0 is the time at which we go bankrupt. This is illustrated here:



We would like to know whether $T_0 < \infty$. Unlike the hitting time T in the previous section, however, here we only have a one-sided boundary: we stop when we hit zero, but our wealth is allowed to get arbitrarily large. It is therefore not at all obvious whether it is still the case that $T_0 < \infty$. Perhaps it is possible that our wealth keeps growing larger and larger forever, so that we never hit the bankruptcy point! That is the dream of every gambler.

The easiest way to understand what is going on in this setting is to use the results of the previous section. Let $0 \le S_0 \le b$, and define

$$T = \min\{n : S_n = 0 \text{ or } S_n = b\},$$

that is, T is the time at which we either go bankrupt, or reach a wealth of b dollars. We know from the previous section that $T < \infty$. We now use the following trick. If it so happens that $S_T = 0$, that is, we go bankrupt before reaching a wealth of b dollars, then $T_0 = T < \infty$. This implies that (why?)

$$\{S_T = 0\} \subseteq \{T_0 < \infty\}.$$

Consequently,

$$\mathbf{P}\{T_0 < \infty | S_0 = i\} \ge \mathbf{P}\{S_T = 0 | S_0 = i\} = 1 - \frac{i}{b}.$$

However, the time T_0 at which we go bankrupt does not depend at all on the level b—we only introduced the latter in order to define the time T! Therefore, we can take the limit as $b \to \infty$ in this inequality, which gives

$$1 = \lim_{b\to\infty} \left(1 - \frac{i}{b}\right) \leq \mathbf{P}\{T_0 < \infty | S_0 = i\} \leq 1.$$

We have therefore shown that

$$\mathbf{P}\{T_0 < \infty\} = 1.$$

That is, *a gambler who keeps playing indefinitely will eventually go bankrupt.* This sad fact of probability is known as Gambler's ruin.

Remark 6.3.1. What did we really do here? If you start with $\$S_0$, and you are playing against an adversary who starts with $\$(b - S_0)$, we could view T as the time at which either you or your adversary goes bankrupt. We have seen in the previous section that one of you must go bankrupt eventually. In the Gambler's ruin problem, however, we assume that your adversary cannot go bankrupt, and so we must give him unlimited funds: that is, we let $b \to \infty$.

Having absorbed the bad news that the gambler will inevitably go bankrupt, he would like to know how long that will take on average. To this end, we again use the results of the previous section. First, notice that T is the first time that we hit either 0 or b. Therefore,

$$T = T_0 \wedge T_b.$$

We claim that

$$\lim_{b\to\infty} T = T_0 \wedge \lim_{b\to\infty} T_b = T_0.$$

Why is this true? Notice that in order to reach wealth b, we must win at least $b - S_0$ dollars. As we can only win one dollar in each round, we must play at least $b - S_0$ rounds before we can reach wealth b. Therefore, $T_b \geq b - S_0$, and so $T_b \to \infty$ as $b \to \infty$, which implies the claim. We can now use the result of the previous section to compute

$$\mathbf{E}(T_0 | S_0 = i) = \lim_{b\to\infty} \mathbf{E}(T | S_0 = i) = \lim_{b\to\infty} i(b - i) = \begin{cases} 0 & \text{if } i = 0, \\ \infty & \text{if } i > 0. \end{cases}$$

That $\mathbf{E}(T_0) = 0$ when $S_0 = 0$ is obvious: if we start with no money, then we are bankrupt already before we can start gambling. What you might find surprising, however, is that if we start with any nonzero amount of money, then $\mathbf{E}(T_0) = \infty$. That is, *even though we always go bankrupt at a finite time, the* expected *time at which we go bankrupt is infinite!*

What does this mean? So far, we have always tacitly assumed that our random variables have finite expectation. But this does not have to be true, even if the random variable itself is finite: here we see a very natural example. What this means is that if we were to compute the average time at which we go bankrupt over many experiments, then we will see larger and larger numbers ("large outliers") on a regular basis, and so the average converges to

infinity. Therefore, even though we know the time is finite, we cannot make an accurate prediction of how large this time is before running the experiment: hence the expected time is infinity. This is the silver lining for our gambler: while a gambler who plays infinitely often goes bankrupt eventually, there is a good chance that will take so long that it may never happen in his lifetime.

Example 6.3.2 (Doubling strategies). We can turn around the conclusions of the Gambler's ruin problem to say something about a popular gambling system of the 19th century: the *doubling strategy*. This strategy works as follows. Say we start with $\$S_0$ dollars, and we start gambling. The strategy is that now matter how much money you lose, borrow more money from the bank to keep gambling until your wealth reaches twice the amount that you started with. That is, if $S_0 = i$, you stop gambling at the time T_{2i} and collect your winnings.

Probability theory seems to support this strategy as a good idea. Indeed,

$$\mathbf{P}\{T_{2i} < \infty | S_0 = i\} = 1,$$

so if you are patient and just keep gambling through thick or thin, you will eventually double your initial wealth. This seems like the sure bet that every gambler is waiting for! What is the catch? The expected number of rounds that you must gamble before your double your initial wealth is infinite:

$$\mathbf{E}(T_{2i} | S_0 = i) = \infty.$$

This means this might well never happen in your lifetime. Even worse, before you finally manage to double your wealth, you might have to go arbitrarily far into debt, which few banks will let you do. If your bank imposes a credit limit (as they would be wise to do if you are gambling with their money), then you would typically be much more likely to go bankrupt than to double your wealth. Ultimately, few gamblers have found the doubling strategy to be profitable in the long run. This sentiment is bluntly expressed by the 19th century English novelist W. M. Thackeray:

> "You have not played as yet? Do not do so; above all avoid a martingale, if you do. [...] I have calculated infallibly, and what has been the effect? Gousset empty, tiroirs empty, necessaire parted for Strasbourg!" (*The Newcomes*, 1854)

The word "martingale" is the old term for a doubling strategy. In modern probability theory, the word martingale has a different meaning; that topic is more advanced than the level of this course.

6.4 Biased random walks

In the definition of a *simple* random walk

$$S_n = S_0 + X_1 + \cdots + X_n,$$

the step sizes $\{X_k\}$ are i.i.d. with values in $\{-1, +1\}$. That is, in each step the random walk moves up or down by a single unit (say, one dollar in a gambling game). Up to this point, we have considered only *symmetric* simple random walks that are equally likely to go up or down in each step, that is, $\mathbf{P}\{X_k = +1\} = \mathbf{P}\{X_k = -1\} = \frac{1}{2}$. The gambling interpretation of a symmetric random walk is that it models a *fair* game: as $\mathbf{E}(X_k) = 0$, your expected wealth always equals the amount you started with, that is,

$$\mathbf{E}(S_n | S_0 = i) = i + \sum_{k=1}^{n} \mathbf{E}(X_k) = i.$$

In particular, neither you or your opponent have an advantage over each other.

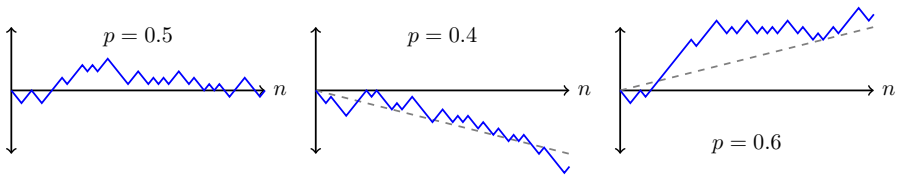
Of course, as everyone knows, real life is not fair. It is quite likely that your opponent will try to cheat by flipping a coin that is biased in his favor. In this case, in every round the random walk will be more likely to go down than up. If you are the cheater, then in every step the random walk will be more likely to go up than down. This is modelled by a *biased* simple random walk, where the step sizes $\{X_k\}$ are i.i.d. with

$$\mathbf{P}\{X_k = 1\} = p, \qquad \mathbf{P}\{X_k = -1\} = q = 1 - p.$$

The random walk is symmetric when $p = \frac{1}{2}$. On the other hand, if $p > q$, then the random walk is more likely to go up in every step, while if $p < q$ the random walk is more likely to go down. In this setting, your expected wealth is no longer constant: in fact, as $\mathbf{E}(X_k) = p - q$, we have

$$\mathbf{E}(S_n | S_0 = i) = i + \sum_{k=1}^{n} \mathbf{E}(X_k) = i + (p - q)n.$$

Thus if the game is biased in your favor, your expected wealth increases on average; and if the game is biased in your opponent's favor, your expected wealth will decrease on average. A typical sample path of the random walk and its expectation in each case is illustrated in the following figure.



As in the previous sections, let us consider for $a \le S_0 \le b$ the hitting time

$$T := \min\{n : S_n = a \text{ or } S_n = b\}.$$

For the symmetric random walk, we have seen that T is always finite, and we computed its expectation and the probability of hitting b before a. How do these properties change when the random walk is biased? Once again, first step analysis is the right tool for answering these questions.

Will we ever hit a boundary?

Let us define, as before, the probability that the (biased) random walk never hits either of the levels a or b:

$$f(i) := \mathbf{P}\{T = \infty | S_0 = i\}.$$

In the symmetric case, we have seen that this probability is zero: the random walk always hits one of the boundaries eventually. Intuitively, we might expect that this is still true for a biased random walk; if anything, this is even more obvious than in the symmetric case, as a biased random walk is actually moving towards one of the boundaries on average. Let us show that this intuition is indeed correct.

Just as in the symmetric case, we perform a first step analysis. In the present case, if we start at the point $a < i < b$, then the probabilities of being at $i+1$ or $i-1$ after one step are p and q, respectively, while the probability of never hitting the boundary becomes $f(i+1)$ or $f(i-1)$. We therefore have

$$f(i) = pf(i+1) + qf(i-1), \quad a < i < b.$$

On the other hand, if we start at one of the boundaries, then clearly the probability of never hitting the boundary is zero. This implies

$$f(a) = f(b) = 0.$$

Clearly $f(i) = 0$ for all i is a solution to these equations, so we have shown

$$\mathbf{P}\{T = \infty | S_0 = i\} = 0 \quad \text{for all } a \le i \le b$$

exactly as for symmetric random walks.

Which boundary will we hit?

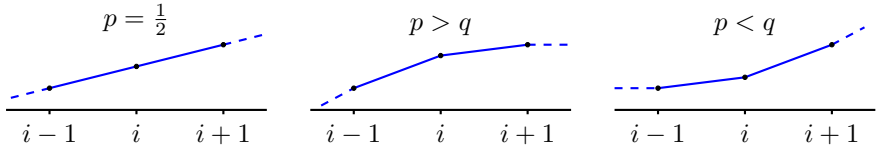
Next, let us compute the probability

$$r(i) := \mathbf{P}\{S_T = b | S_0 = i\}$$

that we hit b before a. First step analysis gives

$$\begin{cases} r(i) = pr(i+1) + qr(i-1), & a < i < b, \\ r(a) = 0, \quad r(b) = 1. \end{cases}$$

Unlike in the symmetric case, the function $i \mapsto r(i)$ is no longer a straight line when $p \neq \frac{1}{2}$. When $p > q$, the equation implies that the value of $r(i)$ is closer to $r(i+1)$ than to $r(i-1)$, so the function $i \mapsto r(i)$ must increase in a concave manner. On the other hand, when $p < q$, the value of $r(i)$ must be closer to $r(i-1)$ than to $r(i+1)$, so the function $i \mapsto r(i)$ must increase in a convex manner. This is illustrated in the following figure:



We can interpret these properties intuitively. If $p > q$, the random walk is moving toward b on average, so the probability of hitting b before a should increase. On the other hand, if $p < q$, then the random walk is moving away from b on average, so the probability of hitting b before a should decrease.

Now that we have some intuition about what the function $i \mapsto r(i)$ should look like, let us actually solve the equation precisely. The computation is similar to the computation of the expected time to hit the boundary for the symmetric random walk. If we substitute $r(i) = pr(i) + qr(i)$ (this is true because $p + q = 1$!) into the equation for $r(i)$ and rearrange, we obtain

$$r(i+1) - r(i) = \frac{q}{p}(r(i) - r(i-1)).$$

We know that $r(a) = 0$. The first increment $r(a+1) - r(a)$ is unknown, so let us call it u for the time being. From this point onward, we can use the equation to compute $r(a+2) - r(a+1)$, etc. This gives:

$$\begin{aligned} r(a) &= 0, \\ r(a+1) - r(a) &= u, \\ r(a+2) - r(a) &= (q/p)u, \\ r(a+3) - r(a+2) &= (q/p)^2 u, \\ &\cdots \\ r(a+i) - r(a+i-1) &= (q/p)^{i-1} u. \end{aligned}$$

If we add up all the quantities on the left- and right-hand sides, we obtain

$$\begin{aligned} r(a+i) &= u(1 + q/p + (q/p)^2 + \cdots + (q/p)^{i-1}) \\ &= \frac{(q/p)^i - 1}{q/p - 1}\, u. \end{aligned}$$

In particular, we have

$$r(i) = \frac{(q/p)^{i-a} - 1}{q/p - 1}\, u.$$

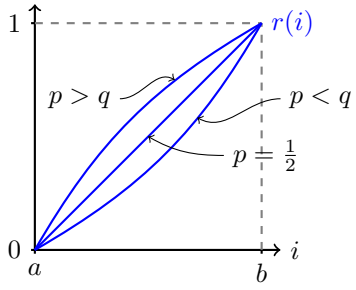
To compute the unknown u we use the remaining boundary condition $r(b) = 1$:

$$r(b) = \frac{(q/p)^{b-a} - 1}{q/p - 1}\, u = 1 \quad \Longrightarrow \quad u = \frac{q/p - 1}{(q/p)^{b-a} - 1}.$$

We have therefore shown that

$$\mathbf{P}\{S_T = b | S_0 = i\} = r(i) = \frac{(q/p)^{i-a} - 1}{(q/p)^{b-a} - 1} \qquad \text{for } a \le i \le b.$$

As expected, this function looks something like this:



How long will it take?

Next, we would like to compute the expect time to hit a boundary:

$$\mu(i) := \mathbf{E}(T | S_0 = i).$$

We are by now such experts on first step analysis that there is no need to explain. The first step analysis yields the following equation:

$$\begin{cases} \mu(i) = 1 + p\mu(i+1) + q\mu(i-1), & a < i < b, \\ \mu(a) = \mu(b) = 0. \end{cases}$$

One can solve this system of equations by writing $\mu(i) = p\mu(i+1) + q\mu(i-1)$ and then rearranging the terms in analogy with our earlier computations. Thus will work, but it is a bit tedious. Instead, let us take a shortcut.

The annoying part of the above equation is the extra $1 +$ that appears here. We are going to remove this additional term by performing a change of variables. To this end, let us define the transformation

$$\tilde{\mu}(i) := \mu(i) + \frac{i-a}{p-q}.$$

As $\mu(a) = 0$, we must have $\tilde{\mu}(a) = 0$ as well. On the other hand,

$$\begin{aligned}\tilde{\mu}(i) &= \frac{i-a}{p-q} + 1 + p\left(\tilde{\mu}(i+1) - \frac{i+1-a}{p-q}\right) + q\left(\tilde{\mu}(i-1) - \frac{i-1-a}{p-q}\right) \\ &= \frac{i-a}{p-q} + 1 + p\tilde{\mu}(i+1) - \frac{p}{p-q} - (p+q)\frac{i-a}{p-q} + q\tilde{\mu}(i-1) + \frac{q}{p-q} \\ &= p\tilde{\mu}(i+1) + q\tilde{\mu}(i-1),\end{aligned}$$

where we used the equation for μ and $p+q=1$. We have therefore transformed the equation for μ into the following equation for $\tilde{\mu}$:

$$\begin{cases} \tilde{\mu}(i) = p\tilde{\mu}(i+1) + q\tilde{\mu}(i-1), & a < i < b, \\ \tilde{\mu}(a) = 0, \quad \tilde{\mu}(b) = \dfrac{b-a}{p-q}. \end{cases}$$

This is the same equation as for $r(i)$, which we have already solved! We recall that the general solution is of the form

$$\tilde{\mu}(i) = \frac{(q/p)^{i-a} - 1}{q/p - 1}\, u,$$

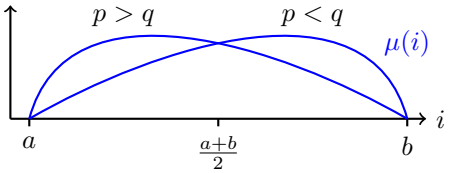
where $u = \tilde{\mu}(a+1) - \tilde{\mu}(a)$. To solve for the unknown u, we use the boundary condition for $\tilde{\mu}(b)$, which gives the following:

$$\tilde{\mu}(i) = \frac{(q/p)^{i-a} - 1}{(q/p)^{b-a} - 1}\,\frac{b-a}{p-q}, \quad a \le i \le b.$$

We have therefore shown that

$$\mathbf{E}(T|S_0 = i) = \mu(i) = \frac{(q/p)^{i-a} - 1}{(q/p)^{b-a} - 1}\,\frac{b-a}{p-q} - \frac{i-a}{p-q}, \qquad a \le i \le b.$$

This function looks something like this:



This picture makes intuitive sense! If $p > q$, for example, the random walk moves on average towards the point b. Therefore, if you start close to b, you are walking towards the nearest boundary and you will get there faster than for the symmetric random walk; but if you start close to a, you are walking away from the nearest boundary and thus it takes longer to reach the boundary.

Gambler's ruin?

For the symmetric random walk, we also considered the one-sided hitting time

$$T_a := \min\{n : S_n = a\}.$$

For example, if S_n is your wealth after n rounds of gambling, then T_0 is the time at which you go bankrupt. We have seen in the previous section that for a symmetric random walk (that is, in a fair game), $T_0 < \infty$ (you will eventually go bankrupt) but $\mathbf{E}(T) = \infty$ (it might take a very long time!)

What happens when the game is unfair? We can repeat the same computations as in the previous section, but we must be careful to distinguish between the three cases $p > q$ (the game is biased in your favor), $p = q$ (fair game), and $p < q$ (the game is biased against you):

$$\mathbf{P}\{T_a < \infty | S_0 = i\} = \lim_{b\to\infty} \mathbf{P}\{S_T = a | S_0 = i\} = \begin{cases} 1 & \text{for } p \le q, \\ (q/p)^{i-a} & \text{for } p > q. \end{cases},$$

$$\mathbf{E}(T_a | S_0 = i) = \lim_{b\to\infty} \mathbf{E}(T | S_0 = i) = \begin{cases} \infty & \text{for } p \ge q, \\ \frac{i-a}{q-p} & \text{for } p < q \end{cases}$$

for $i > a$. In particular, we have the following interesting table.

	$\mathbf{P}\{T_a < \infty\}$	$\mathbf{E}(T_a)$
$p < q$	1	$< \infty$
$p = \frac{1}{2}$	1	∞
$p > q$	< 1	∞

The three possible cases are all different. If the game is biased against you, then not only will you eventually go bankrupt, but the expected time at which this will occur is finite. If the game is fair, you will still eventually go bankrupt, but the expected time at which this happens is infinite. Finally, if the game is biased in your favor, then there is a nonzero probability that you can keep gambling forever without ever going bankrupt!

7

Brownian Motion

In this chapter we will introduce the notion of Brownian motion, which is in essence nothing more than the continuous time limit of a random walk. Nonetheless, Brownian motion shows up in numerous applications across different areas of science and engineering. Moreover, the study of Brownian motion will naturally lead us to an even more important topic: Gaussian variables.

7.1 The continuous time limit of a random walk

Around 1828, the Scottish botanist Robert Brown made a remarkable observation. Brown was observing grains of pollen suspended in water under his microscope. To his surprise, Brown noticed that the pollen particles were continually jittering around in an apparently random fashion. At first, Brown thought that the fact that the pollen particles were moving was a sign that they are alive. However, he rejected this hypothesis after observing that the same phenomenon occurs for indisputably inanimate particles, such as glass powder, various minerals, and even a fragment of the Sphinx! Brown was not able to explain this phenomenon, and the mystery remained in place until it was conclusively resolved by Albert Einstein in a famous 1905 paper.

Einstein's explanation is as follows. Water is not really a continuous fluid, as many physicists thought at the time; rather, water consists of zillions of microscopic particles called "molecules." At room temperature, the laws of physics suggest that these particles must be moving around in their container extremely rapidly with random directions and velocities, like a huge number of tiny billiards that are stuck bouncing around between the microscope's slides. The pollen particle that is suspended in water is therefore constantly being bombarded by these water molecules. Each time the pollen particle is hit by a water molecule, it gets pushed a little bit in a random direction (because the direction in which the water molecule was going is random). The aggregate effect of all these bombardments is that the pollen particle moves around randomly, precisely as was observed by Brown.

Remark 7.1.1. The historical importance of Einstein's explanation goes much beyond the fact that he was able to understand the phenomenon that Brown observed. While today children learn in elementary school that matter consists of molecules, in 1905 this was an extremely controversial idea that was disputed by many top physicists: one cannot see molecules under a microscope, so what evidence is there for their existence? Einstein's explanation of Brownian motion was the first tangible proof that matter consists of molecules: the jittering motion observed by Brown could not be explained convincingly otherwise. This greatly accelerated the acceptance of the molecular hypothesis.

Let us try to create the simplest model of Einstein's idea. For simplicity, we assume the pollen particle moves in one dimension. Let us denote by X_k the displacement of the pollen particle due to its bombardment by the kth water molecule. From Einstein's explanation, it is clear that $\{X_k\}$ should be chosen to be i.i.d. random variables with zero expectation (as the water is not flowing in any particular direction). Therefore, if the pollen particle starts at the position S_0, its position after being bombarded by n water molecules is

$$S_n = S_0 + X_1 + X_2 + \cdots + X_n.$$

In other words, Einstein's explanation implies that the pollen particle performs a random walk!

If you were to look at the pollen particle under a microscope, however, you might have a hard time actually seeing a random walk. Because the pollen particle is humongously large as compared to the tiny water molecules, each bombardment only results in a tiny displacement of the pollen particle that is much too small to be seen by the naked eye (even under a microscope). On the other hand, there is a enormous number of such bombardments that happen every second: these occur much too fast to be seen by the naked eye. Therefore, what we can observe under the microscope is not the individual steps of the random walk, but rather the aggregate effect of a huge number of tiny bombardments. To model this idea, we will create an idealized model of Brownian motion in which we have infinitely many bombardments per second, each of which give rise to an infinitesimal displacement of the pollen particle. That is, we will model Brownian motion as the *continuous time limit* of a random walk! This is precisely the same idea as we used to derive the Poisson process as the continuous time limit of Bernoulli processes.

To understand this limit, let us start for simplicity with a simple symmetric random walk $\mathbf{P}\{X_k = 1\} = \mathbf{P}\{X_k = -1\} = \frac{1}{2}$ that starts at the origin at time zero. Suppose there are N bombardments by water molecules per unit time, and that each bombardment causes a displacement of $\pm\varepsilon$. Then the position of the pollen particle at time t is given by

$$B_t^N = \varepsilon X_1 + \varepsilon X_2 + \cdots + \varepsilon X_{tN}.$$

Remark 7.1.2. The above formula makes sense if tN is an integer. Otherwise, we should really round tN down to the nearest integer $\lfloor tN \rfloor$. As it will be annoying to keep writing $\lfloor \cdot \rfloor$, we will make our life easy by sweeping this minor issue under the rug. If you feel so inclined, you can check that everything we will do makes sense for all times t if you round appropriately.

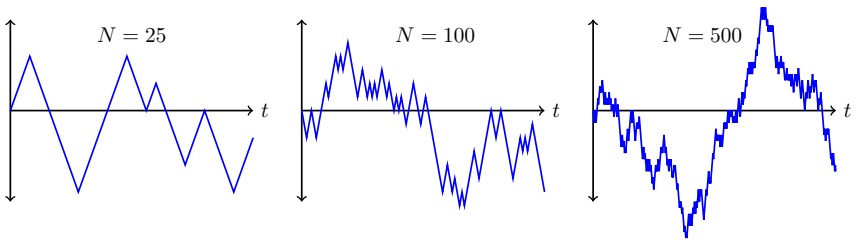
In the continuous time limit, we would like to let $N \to \infty$ (a huge number of bombardments per unit time) and $\varepsilon \to 0$ (each bombardment causes a tiny displacement). But how small should we actually take ε? In real life, a particle must move by a finite amount in finite time. Let us therefore fix the mean square distance travelled by the particle per unit time to be σ^2: that is, $\mathbf{E}((B_1^N)^2) = \sigma^2$. As the variables $\{X_k\}$ are i.i.d. and $\mathbf{E}(B_1^N) = 0$, this implies

$$\mathbf{E}((B_1^N)^2) = \mathrm{Var}(B_1^N) = \mathrm{Var}(\varepsilon X_1) + \cdots + \mathrm{Var}(\varepsilon X_N) = \varepsilon^2 N = \sigma^2.$$

We must therefore choose $\varepsilon = \sigma/\sqrt{N}$, and our microscopic model becomes

$$B_t^N = \frac{\sigma}{\sqrt{N}}(X_1 + X_2 + \cdots + X_{tN}).$$

The process $\{B_t^N\}$ is illustrated below for different values of N:



As we increase the number of bombardments N per unit time (under the correct scaling of the displacement in each bombardment), the paths of the random process $\{B_t^N\}$ look increasingly like what you might observe under a microscope. To model Brownian motion, we now take the limit as the number of bombardments per unit time becomes infinite:

$$B_t = \lim_{N \to \infty} B_t^N.$$

This defines *Brownian motion* $\{B_t\}$ as the continuous limit of a random walk. We will define a *standard* Brownian motion to be the case where $\sigma = 1$ (clearly the general Brownian motion is just σ times a standard Brownian motion).

Remark 7.1.3. Brownian motion turns out to be an incredibly useful process, even if you are not a botanist. Brownian behavior (and the closely related notion of *diffusion*) plays an important role in the understanding of how different molecules find each other in chemical reactions, whether these occur in a chemist's test tube, a large-scale chemical plant, a nuclear reactor, or inside

a living cell. On the other hand, Brownian motion forms the foundation for almost all models of stock price fluctuations used in economics and in finance. In fact, a mathematical description very similar in spirit to Einstein's was already given five years earlier, in the year 1900, by Louis Bachelier for the purpose of modeling the fluctuations of prices on the Parisian bond market. In this setting, the microscopic displacements are due to price movements caused by individual trades on the market, and the continuous time limit arises in liquid markets where there are huge numbers of small trades going on every day. There are numerous other applications of Brownian motion in different areas of physics, chemistry, biology, geoscience, engineering, and statistics.

Remark 7.1.4. Neither Einstein's nor Bachelier's work can be thought of as being mathematically rigorous. The first mathematician to study Brownian motion as a solid mathematical object (for an entirely different purpose!) was Norbert Wiener in a paper published in 1923. The random process $\{B_t\}$ is therefore often also called a *Wiener process* in the mathematical literature.

Remark 7.1.5. The logic behind the continuous time limit we have taken here is very similar to the logic behind the definition of the Poisson process as the continuous time limit of a Bernoulli process. In the latter case, we split time into N intervals per unit time; in each interval, a customer arrives with probability p. The continuous time limit takes $N \to \infty$ and $p \to 0$. But how small should p be? Even in continuous time, there should only be a finite number of customers in finite time; we therefore fixed the average number of customers per unit time to a value λ, and this immediately implies that we must choose $p = \lambda/n$. In the same way, we have insisted here that a Brownian particle can only move by a finite amount in finite time, and this allowed us to fix the size of the steps ε. It is instructive to convince yourself that any other choice of ε would give an uninteresting result: the limit would either blow up to infinity (particle is out of control) or go to zero (particle is not moving).

Now that we have created a basic model of Brownian motion, we must understand its properties. For example:

- What is the distribution of the random variable B_t?
- Can we characterize the properties of the random process $\{B_t\}$?

Recall that we faced the same problem when we defined the Poisson process; there we were able to show that the distribution of N_t is Poisson, and that $\{N_t\}$ is a process with stationary independent increments. We are presently going to derive the analogous properties of Brownian motion.

7.2 Brownian motion and Gaussian distribution

In the previous section, we have seen how Brownian motion can be naturally modelled as the continuous time limit $\{B_t\}$ of random walks $\{B_t^N\}$. However,

defining Brownian motion as a limit is not very useful in practice: it makes it difficult to do computations! In this section, we will compute explicitly the distribution of the limiting random process defined in the previous section. The computation we are about to do is somewhat tedious: we actually follow in the footsteps of the very first person to do this computation, the French mathematician Abraham de Moivre (ca. 1733). The good news is that we only need to do these computations once: once we obtained the answer, we never have to take a limit again, as we can directly define Brownian motion in terms of its important properties (just like we did for the Poisson process!)

Let us begin by computing the distribution of the random variable

$$B_t = \lim_{N\to\infty} B_t^N.$$

Before we can take the limit, we must compute the distribution of B_t^N. To this end, it is convenient to note that as X_k takes the values ± 1 with equal probability, the random variable $Z_k = (X_k + 1)/2$ takes the values $\{0, 1\}$ with equal probability: that is, Z_k is Bernoulli with success probability $\frac{1}{2}$! Thus

$$B_t^N = \frac{\sigma}{\sqrt{N}} \sum_{k=1}^{Nt} (2Z_k - 1) = \frac{2\sigma}{\sqrt{N}} \left(\sum_{k=1}^{Nt} Z_k - \frac{Nt}{2} \right).$$

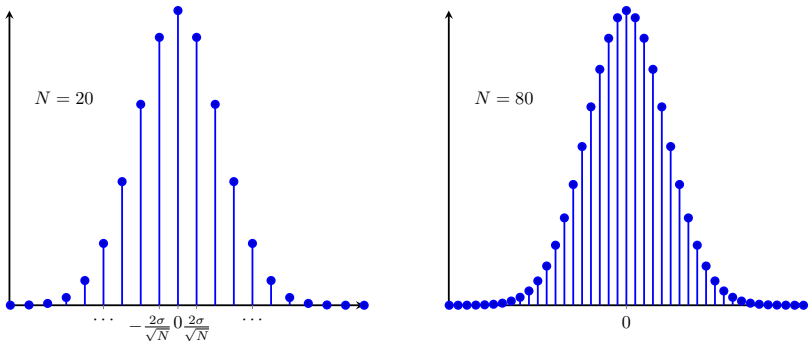
But we know the distribution of the sum of Bernoulli variables: this is just the number of successes in N trials, which has the distribution

$$\sum_{k=1}^{Nt} Z_k \sim \text{Binom}(Nt, \tfrac{1}{2}).$$

Thus B_t^N is a discrete random variable with probabilities

$$\mathbf{P}\left\{ B_t^N = \frac{2\sigma}{\sqrt{N}}\, k \right\} = \mathbf{P}\left\{ \text{Binom}(Nt, \tfrac{1}{2}) = k + \frac{Nt}{2} \right\} = 2^{-Nt} \binom{Nt}{\frac{Nt}{2} + k}$$

for $k = -\frac{Nt}{2}, -\frac{Nt}{2} + 1, \ldots, \frac{Nt}{2}$. The probabilities $\mathbf{P}\{B_t^N = x\}$ are illustrated in the following figure for two different values of N:



Evidently the random variables B_t^N are discrete; however, as N increases, the possible values that B_t^N can take get closer and closer together. Therefore, in the limit as $N \to \infty$, we expect to obtain a *continuous* random variable B_t! This continuous random variable cannot be described by the probabilities $\mathbf{P}\{B_t = x\}$, which must be zero. Instead, we will try to compute the density f of this random variable, that is, $\mathbf{P}\{B_t \in dx\} = f(x)dx$. Recall that

$$f(x) = \frac{d}{dx}\mathbf{P}\{B_t \le x\} = \lim_{\delta \to 0} \frac{\mathbf{P}\{B_t \in [x, x+\delta[\}}{\delta}.$$

It is convenient to choose $\delta = \frac{2\sigma}{\sqrt{N}} \to 0$ as $N \to \infty$, so that B_t^N can take one value in the interval $[x, x+\delta[$ (because the possible values this discrete random variable can take are spaced by precisely δ). Then we should have

$$f(x) = \lim_{N\to\infty} \frac{\mathbf{P}\{B_t^N \in [x, x+\delta[\}}{\delta} = \lim_{N\to\infty} 2^{-Nt} \binom{Nt}{\frac{Nt}{2} + \frac{x\sqrt{N}}{2\sigma}} \frac{\sqrt{N}}{2\sigma}$$

We have now turned the problem of computing the density of B_t into a purely mathematical problem: we must compute the limit of an expression that involves binomial coefficients. This is a tedious business, but it can be done! Let us grit our teeth and get through this ordeal.

To do the computation, notice that we can write

$$\begin{aligned}
\binom{2n}{n+i} &= \frac{(2n)!}{(n+i)!(n-i)!} \\
&= \frac{(2n)!}{(n!)^2} \cdot \frac{n(n-1)\cdots(n-i+1)}{(n+i)(n+i-1)\cdots(n+1)} \\
&= \binom{2n}{n} \cdot \frac{n(n-1)\cdots(n-i+1)}{(n+i)(n+i-1)\cdots(n+1)}
\end{aligned}$$

This implies that

$$\begin{aligned}
f(x) &= \lim_{N\to\infty} 2^{-Nt} \binom{Nt}{\frac{Nt}{2}} \frac{\sqrt{N}}{2\sigma} \frac{\frac{Nt}{2}(\frac{Nt}{2}-1)\cdots(\frac{Nt}{2}-\frac{x\sqrt{N}}{2\sigma}+1)}{(\frac{Nt}{2}+1)(\frac{Nt}{2}+2)\cdots(\frac{Nt}{2}+\frac{x\sqrt{N}}{2\sigma})} \\
&= \lim_{N\to\infty} 2^{-Nt} \binom{Nt}{\frac{Nt}{2}} \frac{\sqrt{N}}{2\sigma} \frac{(1-\frac{2}{Nt})(1-\frac{4}{Nt})\cdots(1-\frac{x\sqrt{N}}{2\sigma}\frac{2}{Nt}+\frac{2}{Nt})}{(1+\frac{2}{Nt})(1+\frac{4}{Nt})\cdots(1+\frac{x\sqrt{N}}{2\sigma}\frac{2}{Nt})},
\end{aligned}$$

where in the second equality we have divided each term in the numerator and denominator by $\frac{Nt}{2}$. Now notice that each of these terms is of the form $1+x$ for constants x that converge to zero as $N \to \infty$. As

$$\lim_{x\to 0} \frac{e^x - 1}{x} = \left.\frac{de^x}{dx}\right|_{x=0} = 1,$$

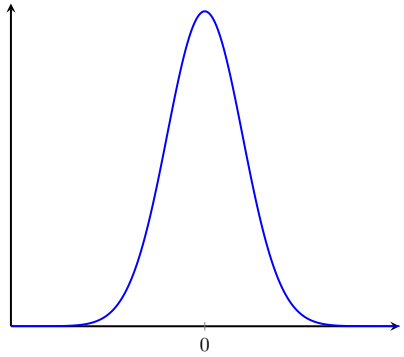
we have $1 + x \approx e^x$ when $x \approx 0$. This allows us to write

$$\begin{aligned}\frac{(1-\frac{2}{Nt})(1-\frac{4}{Nt})\cdots(1-\frac{x\sqrt{N}}{2\sigma}\frac{2}{Nt}+\frac{2}{Nt})}{(1+\frac{2}{Nt})(1+\frac{4}{Nt})\cdots(1+\frac{x\sqrt{N}}{2\sigma}\frac{2}{Nt})} &\approx \frac{e^{-\frac{2}{Nt}}e^{-\frac{4}{Nt}}\cdots e^{-\frac{x\sqrt{N}}{2\sigma}\frac{2}{Nt}}}{e^{+\frac{2}{Nt}}e^{+\frac{4}{Nt}}\cdots e^{+\frac{x\sqrt{N}}{2\sigma}\frac{2}{Nt}}}\\ &= e^{-\frac{4}{Nt}\sum_{k=1}^{\frac{x\sqrt{N}}{2\sigma}}k}\\ &= e^{-\frac{2}{Nt}\left(\left(\frac{x\sqrt{N}}{2\sigma}\right)^2+\frac{x\sqrt{N}}{2\sigma}\right)}\\ &= e^{-\frac{x^2}{2\sigma^2 t}-\frac{x}{\sigma\sqrt{Nt}}}\\ &\xrightarrow{N\to\infty} e^{-\frac{x^2}{2\sigma^2 t}}.\end{aligned}$$

We have therefore shown that the density $f(x)$ of B_t must be of the form

$$f(x) = Ce^{-\frac{x^2}{2\sigma^2 t}}, \qquad C = \lim_{N\to\infty} 2^{-Nt}\binom{Nt}{\frac{Nt}{2}}\frac{\sqrt{N}}{2\sigma}.$$

This function looks like this:



This formula clearly explains the beautiful bell shape that we saw appearing when we plotted the binomial distributions—and the only thing that is left for us to compute is the constant C! We defined the constant C as a certain limit, but this is annoying to compute. Instead, we will take a shortcut. Recall that any probability density must integrate to one, so

$$C\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2 t}}\,dx = 1.$$

To compute the constant C, it is therefore enough to compute the integral in this expression. We can do this using a clever calculus trick: we compute the *square* of this integral by transforming to polar coordinates!

$$\begin{aligned}\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2 t}}\,dx\right)^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2\sigma^2 t}}\,dx\,dy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2\sigma^2 t}}\,r\,dr\,d\varphi \\ &= \pi\int_0^{\infty} e^{-\frac{r^2}{2\sigma^2 t}}\,dr^2 \\ &= -2\pi\sigma^2 t\, e^{-\frac{s}{2\sigma^2 t}}\Big|_{s=0}^{\infty} \\ &= 2\pi\sigma^2 t.\end{aligned}$$

Plugging this into our equation for C, we have finally shown that

$$\mathbf{P}\{B_t \in dx\} = \frac{e^{-\frac{x^2}{2\sigma^2 t}}}{\sqrt{2\pi\sigma^2 t}}\,dx.$$

This is one of the most important distributions in probability theory.

Definition 7.2.1. *A random variable X with distribution*

$$\mathbf{P}\{X \in dx\} = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\,dx$$

is called Gaussian *with mean* 0 *and variance* σ^2 *($X \sim N(0,\sigma^2)$).*

In particular, we have shown that $B_t \sim N(0,\sigma^2 t)$ for every time t.

Remark 7.2.2. We followed the derivation of the Gaussian distribution due to Abraham de Moivre in 1733. His work was later made more precise by the Marquis de Laplace in 1774. The German mathematician Carl Friedrich Gauss, motivated by completely different problems of statistics, published a paper about the Gaussian distribution in 1809. It is an all-too-usual twist of fate that this distribution has come to be known as the "Gaussian" distribution. In statistics, the Gaussian distribution is often called the Normal distribution, a name popularized by the English statistician Karl Pearson (1857–1936). This is the origin of the notation $N(0,\sigma^2)$. I do not like this terminology (it implies that other distributions are somehow abnormal, which is of course nonsense), but the notation has become standard and thus we will use it as well.

Remark 7.2.3. By construction, $\{B_t^N\}$ satisfies $\mathbf{E}(B_1^N) = 0$ and $\mathrm{Var}(B_1^N) = \sigma^2$. Therefore, letting $N \to \infty$, we must also have $\mathbf{E}(B_1) = 0$ and $\mathrm{Var}(B_1) = \sigma^2$. Thus a random variable with distribution $N(0,\sigma^2)$ does indeed have mean 0 and variance σ^2, as suggested in the definition.

We have now computed the distribution of Brownian motion B_t at a given time t. Let us collect two further important properties of Brownian motion:

a. For times $r \leq s \leq t$, the random variables

$$B_t^N - B_s^N = \frac{\sigma}{\sqrt{N}}(X_{Ns+1} + \ldots + X_{Nt}),$$
$$B_r^N = \frac{\sigma}{\sqrt{N}}(X_1 + \ldots + X_{Nr})$$

are independent: each of the steps X_i is independent, and we can see that $B_t^N - B_s^N$ and B_r^N depend on disjoint subsets of these variables. Taking the limit as $N \to \infty$, it follows immediately that

$$B_t - B_s \perp\!\!\!\perp \{B_r\}_{r \leq s} \quad \text{for } s \leq t,$$

that is, Brownian motion has *independent increments*. Note that the increment $B_t - B_s$ measures the distance travelled by the pollen particle between times s and t. The independent increments property states that the distance travelled between times s and t is independent of what happened to the particle before time s. This is completely intuitive: the particle gets bombarded by an independent water molecule at every time.

b. Note that $B_t^N - B_s^N$ is the sum of $N(t-s)$ individual steps. As the steps in our model are i.i.d., this implies that $B_t^N - B_s^N$ has the same distribution as B_{t-s}^N. In particular, in the limit as $N \to \infty$, we find that

$$B_t - B_s \sim N(0, \sigma^2(t-s)) \quad \text{for } s \leq t.$$

That is, Brownian motion has *stationary increments*. This property captures the fact that all bombardments of the pollen particle have the same distribution (that is, all water molecules are alike!)

These two facts together capture all the important properties of Brownian motion. There are a few little choices that we made that are less important. First of all, we assumed that the random walks start at zero $B_0^N = 0$, so $B_0 = 0$ as well. Second, we assumed that the mean-square distance travelled by time one is σ^2. There is little harm in assuming $\sigma = 1$: obtaining any other value of σ is then accomplished by multiplying the Brownian motion by the constant σ. We therefore introduce the following standard definition.

Definition 7.2.4. *A continuous random process* $\{B_t\}_{t \geq 0}$ *is called* (standard) Brownian motion *if*

a. $B_0 = 0$.
b. $B_t - B_s \perp\!\!\!\perp \{B_r\}_{r \leq s}$ *for* $s \leq t$ *(independent increments).*
c. $B_t - B_s \sim N(0, t-s)$ *for* $s \leq t$ *(stationary Gaussian increments).*

The term *standard* means that $B_0 = 0$ and $\mathrm{Var}(B_1) = 1$. It is conventional to assume that a Brownian motion is standard, unless explicitly stated otherwise. If $\{B_t\}$ is a standard Brownian motion, you can create a Brownian motion $\{B'_t\}$ with arbitrary starting point a and variance σ by setting $B'_t = a + \sigma B_t$.

Remark 7.2.5. Notice how similar the definition of Brownian motion is to the definition of a Poisson process: both processes have stationary independent increments. On the other hand, these processes look nothing alike: the Poisson process is a counting process whose increments are Poisson, while Brownian motion is a continuous process whose increments are Gaussian.

Remark 7.2.6. Definition 7.2.4 *contains everything you need to perform computations with Brownian motion.* Even though we derived this definition from the continuous time limit of random walks, you will never have to take any limit of this kind again once you know that Brownian motion behaves as in Definition 7.2.4. This is great news, because the limits that we took in this section are the nastiest limits we will take in this course (you are not expected to be able to reproduce them!), and we would hate to have to compute such limits every time we wanted to use Brownian motion. Nonetheless, it was important to have seen once the continuous time limit, if only because it shows that Brownian motion and the Gaussian distribution follow naturally from a simple modelling problem: these are not mysterious concepts, but rather natural objects that you would discover by yourself if you ask the right questions.

7.3 The central limit theorem

In this section we are going to derive an important result in probability and statistics. Before we can do that, however, we must understand some very useful properties of Gaussian random variables.

Example 7.3.1 (Sums of independent Gaussian variables). Let $X \sim N(0, \sigma^2)$ and $Y \sim N(0, \tau^2)$ be independent Gaussian random variables. What is the distribution of of the random variable $X + Y$?

This question is most easily answered using Brownian motion. Notice that $B_{\sigma^2} \sim N(0, \sigma^2)$, $B_{\tau^2+\sigma^2} - B_{\sigma^2} \sim N(0, \tau^2)$, and $B_{\tau^2+\sigma^2} - B_{\sigma^2} \perp\!\!\!\perp B_{\sigma^2}$. We can therefore define $X = B_{\sigma^2}$ and $Y = B_{\tau^2+\sigma^2} - B_{\sigma^2}$, which implies

$$X + Y = B_{\tau^2+\sigma^2} \sim N(0, \sigma^2 + \tau^2).$$

Thus we have shown that if $X \sim N(0, \sigma^2)$ and $Y \sim N(0, \tau^2)$ are *independent* Gaussian variables, then $X + Y \sim N(0, \sigma^2 + \tau^2)$ is again Gaussian.

Example 7.3.2 (Linear functions of Gaussian variables). Let $X \sim N(0, \sigma^2)$ and $a, b \in \mathbb{R}$. What is the distribution of $aX + b$? To answer this question, let us compute the density of this random variable:

$$\begin{aligned}\mathbf{P}\{aX + b \in dx\} &= \mathbf{P}\{aX + b \in [x, x + dx[\} \\ &= \mathbf{P}\{X \in [\tfrac{x-b}{a}, \tfrac{x-b+dx}{a}[\} = \frac{e^{-\frac{(x-b)^2}{2\sigma^2 a^2}}}{\sqrt{2\pi\sigma^2 a^2}}\, dx.\end{aligned}$$

Definition 7.3.3. *A random variable X with distribution*

$$\mathbf{P}\{X \in dx\} = \frac{e^{-\frac{(x-b)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\, dx$$

is called Gaussian *with mean b and variance σ^2 ($X \sim N(b, \sigma^2)$).*

Thus if $X \sim N(0, \sigma^2)$, then $aX + b = N(b, \sigma^2 a^2)$. In particular, a linear function of a Gaussian variable is again Gaussian.

Example 7.3.4 (Linear combinations of standard Gaussians). If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(0, 1)$ random variables and $a_1, a_2, \ldots, a_n \in \mathbb{R}$, then

$$a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \sim N(0, a_1^2 + a_2^2 + \cdots + a_n^2)$$

This follows immediately by combining the previous two examples.

Example 7.3.5 (Moments of Gaussian variables). Let $X \sim N(0, 1)$. Can we compute $\mathbf{E}(X^p)$ for $p = 1, 2, 3, 4, 5$?

The most direct way to solve this problem is to compute the integral

$$\mathbf{E}(X^p) = \int_{-\infty}^{\infty} x^p\, \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx.$$

This road is quite tedious, however. Let us therefore develop a different method to do this computation that is useful for many other distributions (not just Gaussian) as well, using the notion of a *moment generating function*.

Definition 7.3.6. *The moment generating function of a random variable X is defined as $\varphi(\alpha) = \mathbf{E}(e^{\alpha X})$ for $\alpha \in \mathbb{R}$.*

What is the point of this bizarre definition? Note that

$$\frac{d^p}{d\alpha^p}\varphi(\alpha) = \mathbf{E}\left(\frac{d^p}{d\alpha^p}e^{\alpha X}\right) = \mathbf{E}(X^p e^{\alpha X}),$$

so that

$$\frac{d^p}{d\alpha^p}\varphi(\alpha)\bigg|_{\alpha=0} = \mathbf{E}(X^p).$$

We can therefore compute quantities such as $\mathbf{E}(X^p)$ by taking derivatives of the moment generating function! For some distributions, this is no easier than to compute $\mathbf{E}(X^p)$ directly. In other cases, however, computing the moment generating function is significantly simpler than a direct computation. This is the case for Gaussian random variables.

To see why, let us compute the moment generating function $\varphi(\alpha)$ of the Gaussian random variable $X \sim N(0, \sigma^2)$. Note that

$$\varphi(\alpha) = \mathbf{E}(e^{\alpha X}) = \int_{-\infty}^{\infty} e^{\alpha x} \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\, dx = \int_{-\infty}^{\infty} \frac{e^{(2\sigma^2\alpha x - x^2)/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\, dx.$$

The reason the moment generating function of the Gaussian is easy to compute is that we can complete the square:

$$2\sigma^2\alpha x - x^2 = -(x - \sigma^2\alpha)^2 + \sigma^4\alpha^2.$$

We can therefore compute

$$\varphi(\alpha) = \int_{-\infty}^{\infty} e^{\sigma^2\alpha^2/2} \frac{e^{-(x-\sigma^2\alpha)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\, dx = e^{\sigma^2\alpha^2/2}$$

(as $e^{\sigma^2\alpha^2/2}$ is a constant and $\frac{e^{-(x-\sigma^2\alpha)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$ is the density of a random variable with distribution $N(\sigma^2\alpha, \sigma^2)$, so must integrate to one). Thus we have shown:

> The moment generating function of $X \sim N(0, \sigma^2)$ is $\varphi(\alpha) = e^{\sigma^2\alpha^2/2}$.

We can now easily compute the first few expectations $\mathbf{E}(X^p)$:

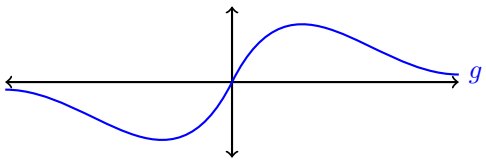
$$\begin{aligned}
\mathbf{E}(X) &= \frac{d\varphi(\alpha)}{d\alpha}\bigg|_{\alpha=0} = \alpha\sigma^2 e^{\alpha^2\sigma^2/2}\Big|_{\alpha=0} = 0,\\
\mathbf{E}(X^2) &= \frac{d^2\varphi(\alpha)}{d\alpha^2}\bigg|_{\alpha=0} = (\sigma^2 + (\alpha\sigma^2)^2)e^{\alpha^2\sigma^2/2}\Big|_{\alpha=0} = \sigma^2,\\
\mathbf{E}(X^3) &= \frac{d^3\varphi(\alpha)}{d\alpha^3}\bigg|_{\alpha=0} = (3\alpha\sigma^4 + (\alpha\sigma^2)^3)e^{\alpha^2\sigma^2/2}\Big|_{\alpha=0} = 0,\\
\mathbf{E}(X^4) &= \frac{d^4\varphi(\alpha)}{d\alpha^4}\bigg|_{\alpha=0} = (3\sigma^4 + 6\alpha^2\sigma^6 + (\alpha\sigma^2)^4)e^{\alpha^2\sigma^2/2}\Big|_{\alpha=0} = 3\sigma^4,\\
\mathbf{E}(X^5) &= \frac{d^5\varphi(\alpha)}{d\alpha^5}\bigg|_{\alpha=0} = (15\alpha\sigma^6 + 10\alpha^3\sigma^8 + (\alpha\sigma^2)^5)e^{\alpha^2\sigma^2/2}\Big|_{\alpha=0} = 0,
\end{aligned}$$

etcetera. As a sanity check, note that we already know that the first two expressions are correct: a Gaussian $X \sim N(0, \sigma^2)$ must certainly have mean $\mathbf{E}(X) = 0$ and variance $\mathbf{E}(X^2) - \mathbf{E}(X)^2 = \mathbf{E}(X^2) = \sigma^2$.

Remark 7.3.7. In the computation above, we always had $\mathbf{E}(X^p) = 0$ when p is an odd number. Can we explain this behavior? To understand what is going on, let us write the expectation as an integral in the usual way:

$$\mathbf{E}(X^p) = \int_{-\infty}^{\infty} x^p \, \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \, dx.$$

Notice that the function $g(x) = x^p e^{-x^2/2\sigma^2}$ is odd when p is odd: it satisfies $g(-x) = -g(x)$. The function looks something like this:



Note that the area under the graph of g for $x \leq 0$ is precisely minus the area under the graph of g for $x \geq 0$. This implies that the integral of this function must be zero, which explains why $\mathbf{E}(X^p) = 0$ whenever p is odd.

We have shown that the moment generating function of a Gaussian random variable $X \sim N(0, \sigma^2)$ is $\varphi(\alpha) = e^{\sigma^2\alpha^2/2}$. However, it turns out that the converse is also true: if a random variable X has moment generating function $\varphi(\alpha) = e^{\sigma^2\alpha^2/2}$, then it must be the case the $X \sim N(0, \sigma^2)$. More generally, it is (almost always) the case that the moment generating function encodes the distribution of a random variable, that is, the moment generating function of X contains enough information to allow us to compute $\mathbf{E}(f(X))$ for any function f! Why? If we Taylor expand $f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$, we can write $\mathbf{E}(f(X)) = a_0 + a_1\mathbf{E}(X) + a_2\mathbf{E}(X^2) + a_3\mathbf{E}(X^3) + \cdots$. As we have seen that the moment generating function allows us to compute $\mathbf{E}(X^p)$ for any p, we can (in principle!) compute the expectation of any function using only the moment generating function. Of course, you would never actually do the computation this way; but this idea allows us to reason about Gaussian (and other) distributions by looking at their moment generating functions.

Example 7.3.8. Let us once again show that if $X \sim N(0, \sigma^2)$ and $Y \sim N(0, \tau^2)$ are independent Gaussian random variables, then $X + Y \sim N(0, \sigma^2 + \tau^2)$. We have shown at this beginning of this section that this is true by using Brownian motion. Let us now give an alternative proof of this fact using moment

generating functions. Let $\varphi_X, \varphi_Y, \varphi$ be the moment generating functions of $X, Y, X+Y$, respectively. Note that as $X \perp\!\!\!\perp Y$, we can compute

$$\varphi(\alpha) = \mathbf{E}(e^{\alpha(X+Y)}) = \mathbf{E}(e^{\alpha X} e^{\alpha Y}) = \mathbf{E}(e^{\alpha X})\,\mathbf{E}(e^{\alpha Y}) = \varphi_X(\alpha)\varphi_Y(\alpha).$$

As $\varphi_X(\alpha) = e^{\sigma^2\alpha^2/2}$ and $\varphi_Y(\alpha) = e^{\tau^2\alpha^2/2}$, we obtain

$$\varphi(\alpha) = e^{(\sigma^2+\tau^2)\alpha^2/2}.$$

Thus the moment generating function of $X+Y$ is that of a $N(0, \sigma^2+\tau^2)$ random variable, which implies $X+Y \sim N(0, \sigma^2+\tau^2)$.

Remark 7.3.9. Of course, we have skimmed over some technicalities. For example, some random variables have $\mathbf{E}(X^p) = \infty$ or $\mathbf{E}(X^p)$ growing very quickly with p, in which case our Taylor expansion does not work. Also, not all functions are sufficiently smooth to have a convergent Taylor expansion. For the purposes of this course, we will sweep these technical issues under the rug and not worry about them. In the vast majority of cases, if the moment generating function of a random variable makes sense, then it completely determines its distribution, and in our course you can always assume this is the case.

Now that we have consumed the appetizers, we can finally turn to the main course of this section: the *central limit theorem*.

Recall how we obtained the Gaussian distribution as the limit of a symmetric simple random walk: we showed in the previous section that

$$\text{the distribution of } \frac{X_1 + X_2 + \cdots + X_N}{\sqrt{N}} \xrightarrow{N\to\infty} N(0,1)$$

if $\{X_k\}$ are i.i.d. random variables with $\mathbf{P}\{X_k = 1\} = \mathbf{P}\{X_k = -1\} = \frac{1}{2}$. We motivated this limit as a model of how a pollen particle behaves when it is bombarded by many water molecules. However, as a physical model, this is quite stupid: it claims that the particle moves left or right by exactly the same amount ($\pm\frac{1}{\sqrt{N}}$) in every bombardment. In practice, it is likely that the displacement X_k caused by a bombarding water molecule has a more complicated distribution. How would that change our model of Brownian motion?

As a first hint of what might happen, let us try to assume instead that the step sizes X_k are i.i.d. *Gaussian* random variables $X_k \sim N(0,1)$. What happens to the above limit? In this case, life is even easier:

$$\text{the distribution of } \frac{X_1 + X_2 + \cdots + X_N}{\sqrt{N}} = N(0,1)$$

for *any* N, even without taking the limit! After all, each step $X_k/\sqrt{N}$ is Gaussian with mean 0 and variance $1/N$, so the sum of N such independent steps is Gaussian with mean 0 and variance 1. Thus in this case the limit is precisely the same as before, even though we started with a very different distribution of the individual steps of the random walk.

It turns out that this is a much more general phenomenon: we can start with i.i.d. steps $\{X_k\}$ that have *any* distribution with the same mean and variance, and in the limit as $N \to \infty$ we always get the same answer. This means that it does not matter at all how precisely we model the displacement caused by each water molecule: no matter what we do, we always end up with the same Brownian motion in the continuous time limit! This is a very remarkable fact: "macroscopic" behavior (the displacement of our particle after many bombardments) does not depend on the details of our "microscopic" model (what happens in each bombardment). Physicists call this type of phenomenon *universality*. In the present case, the precise statement is as follows.

Theorem 7.3.10 (Central limit theorem). *If $\{X_k\}$ are i.i.d. random variables with $\mathbf{E}(X_k) = 0$ and $\mathrm{Var}(X_k) = \sigma^2$, then*

$$\text{the distribution of } \frac{X_1 + X_2 + \cdots + X_N}{\sqrt{N}} \xrightarrow{N\to\infty} N(0, \sigma^2).$$

The central limit theorem is a very important result (this is why it is called the "central" limit theorem—the name is due to George Pólya). Not only does it explain universality in physics, but it allows us to compute confidence intervals in statistics, and it justifies the use of the Gaussian distribution in many models ranging from biology to electrical engineering.

To understand why the central limit theorem is true, we will use moment generating functions. Suppose that $\{X_k\}$ are i.i.d. random variables with $\mathbf{E}(X_k) = 0$ and $\mathrm{Var}(X_k) = \sigma^2$. Denote by $\varphi(\alpha)$ and $\varphi_N(\alpha)$ the moment generating functions of X_k and of $(X_1 + \cdots + X_N)/\sqrt{N}$, respectively. Then

$$\begin{aligned}
\varphi_N(\alpha) &= \mathbf{E}(e^{\alpha(X_1+X_2+\cdots+X_N)/\sqrt{N}}) \\
&= \mathbf{E}(e^{\alpha X_1/\sqrt{N}} e^{\alpha X_2/\sqrt{N}} \cdots e^{\alpha X_N/\sqrt{N}}) \\
&= \mathbf{E}(e^{\alpha X_1/\sqrt{N}})\, \mathbf{E}(e^{\alpha X_2/\sqrt{N}}) \cdots \mathbf{E}(e^{\alpha X_N/\sqrt{N}}) \\
&= \varphi(\alpha/\sqrt{N})^N = e^{N \log \varphi(\alpha/\sqrt{N})}.
\end{aligned}$$

To see what happens to the distribution in the limit $N \to \infty$, we would like to compute the limit of the moment generating function φ_N. Note that

$$\begin{aligned}
\log \varphi(\alpha)\Big|_{\alpha=0} &= 0, \\
\frac{d}{d\alpha} \log \varphi(\alpha)\Big|_{\alpha=0} &= \frac{\varphi'(0)}{\varphi(0)} = \mathbf{E}(X_k) = 0, \\
\frac{d^2}{d\alpha^2} \log \varphi(\alpha)\Big|_{\alpha=0} &= \frac{\varphi''(0)}{\varphi(0)} - \frac{\varphi'(0)^2}{\varphi(0)^2} = \mathbf{E}(X_k^2) - \mathbf{E}(X_k)^2 = \sigma^2.
\end{aligned}$$

Therefore, we can Taylor expand the function $\log \varphi$ to obtain

$$\log \varphi(\alpha) = \frac{\sigma^2}{2}\alpha^2 + \frac{c_3}{3!}\alpha^3 + \frac{c_4}{4!}\alpha^4 + \cdots$$

This means that

$$N \log \varphi(\alpha/\sqrt{N}) = \frac{\sigma^2}{2}\alpha^2 + \frac{1}{\sqrt{N}}\frac{c_3}{3!}\alpha^3 + \frac{1}{N}\frac{c_4}{4!}\alpha^4 + \cdots$$

Therefore, in the limit as $N \to \infty$, all terms except the first disappear:

$$\lim_{N\to\infty} N \log \varphi(\alpha/\sqrt{N}) = \frac{\sigma^2\alpha^2}{2},$$

no matter what moment generating function φ we started with! In particular, the moment generating function $\varphi_N(\alpha)$ of $(X_1 + \cdots + X_N)/\sqrt{N}$ converges to

$$\lim_{N\to\infty} \varphi_N(\alpha) = e^{\sigma^2\alpha^2/2},$$

which is the moment generating function of a Gaussian $N(0, \sigma^2)$ random variable. We have therefore established the central limit theorem!

Example 7.3.11 (Moving truck). A moving truck has a capacity of 600 cu ft. A customer delivers 100 boxes to be moved, each of which has a different size. The sizes of the boxes are i.i.d., the average size is 5 cu ft, and the variance of the size is 4 (cu ft)^2. What is the probability that more than one truck will be needed to move all the boxes (which would greatly increase the expense)?

This question is not very concrete. Let us try to make it more precise. If X_k is the size of the kth box, then our assumptions state that $X_1, X_2, \ldots, X_{100}$ are i.i.d. random variables with $\mathbf{E}(X_k) = 5$ and $\mathrm{Var}(X_k) = 4$. We are asked to compute the probability that the total size of all the boxes exceeds the capacity of the truck, that is, we would like to compute

$$\mathbf{P}\left\{\sum_{k=1}^{100} X_k > 600\right\}.$$

Unfortunately, it is impossible to compute this probability on the basis of the data given in the problem, as *we have not been told the distribution of X_k!* Indeed, the customer was too lazy to make a precise measurement of the distribution of the box sizes, so he only told us the mean and variance. In the absence of further information, there is no way we can compute the above probability exactly. However, the central limit theorem allows us to *approximate* this probability without knowing the distribution of the random variables X_k: as the question concerns the sum of a large number of i.i.d. random variables, the central limit theorem tells us that this sum must approximately have the Gaussian distribution, regardless of the original distribution of the

summands. This is a typical example where the universality principle helps us do otherwise intractable computations.

Let us turn all this talk into action. Because the random variables X_k do not have mean zero, we cannot directly apply the central limit theorem (which works only for mean zero random variables). Let us therefore define $Z_k = X_k - 5$, which is imply the size of the kth box minus the average box size. Note that $\mathbf{E}(Z_k) = 0$ and $\mathrm{Var}(Z_k) = 4$. Moreover,

$$\sum_{k=1}^{100} X_k = \sum_{k=1}^{100} (Z_k + 5) = \sum_{k=1}^{100} Z_k + 500.$$

The central limit theorem tells us that

$$\frac{1}{\sqrt{100}} \sum_{k=1}^{100} Z_k \approx N(0, 4).$$

Therefore,

$$\mathbf{P}\left\{ \sum_{k=1}^{100} X_k > 600 \right\} = \mathbf{P}\left\{ \frac{1}{\sqrt{100}} \sum_{k=1}^{100} Z_k > 10 \right\} \approx \mathbf{P}\{N(0,4) > 10\}.$$

The latter is an explicit probability that we can hope to compute. In fact, as $X \sim N(0, \sigma^2)$ implies $aX \sim N(0, a^2\sigma^2)$, an $N(0,4)$ random variable is just two times an $N(0,1)$ random variable. Therefore,

$$\mathbf{P}\{N(0,4) > 10\} = \mathbf{P}\{N(0,1) > 5\} = \int_5^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx.$$

The function $z \mapsto \int_z^\infty \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$ cannot be simplified, so you should not bother to try to compute it explicitly; this is an unfortunate fact of life in dealing with Gaussian variables. Fortunately, you can compute it very accurately on any computer using something called the *erf* function (consult the manual of your favorite compute program). The computer finds that the probability that we need more than one truck to move the boxes is approximately 2.87×10^{-7}, a reassuringly miniscule number for the cash-strapped customer.

7.4 Jointly Gaussian variables

We have seen that if X, Y are *independent* Gaussian random variables, then the linear combination $aX + bY$ is also Gaussian for every $a, b \in \mathbb{R}$. The independence assumption is crucial here! Here is an extreme example.

Example 7.4.1 (Dependent Gaussians may be evil). Let $X \sim N(0,1)$, and let $S \perp\!\!\!\perp X$ be another random variable with $\mathbf{P}\{S = +1\} = \mathbf{P}\{S = -1\} = \frac{1}{2}$.

Define the random variable $Y = SX$. We claim that X and Y are both $N(0,1)$ random variables (that are *not* independent), but $X + Y$ is not Gaussian.

Why is $Y \sim N(0,1)$? Recall that $aX \sim N(0, a^2)$ for every a. In particular, if we let $a = -1$, we obtain $-X \sim N(0,1)$: the distribution of a Gaussian does not change if we multiply the Gaussian by -1. We can use this to compute

$$\begin{aligned}\mathbf{P}\{Y \le t\} &= \mathbf{P}\{Y \le t | S = +1\}\mathbf{P}\{S = +1\} + \mathbf{P}\{Y \le t | S = -1\}\mathbf{P}\{S = -1\} \\ &= \tfrac{1}{2}\mathbf{P}\{X \le t\} + \tfrac{1}{2}\mathbf{P}\{-X \le t\} \\ &= \mathbf{P}\{X \le t\}.\end{aligned}$$

Thus X and Y have the same CDF, and consequently the same distribution. This shows that X and Y are both $N(0,1)$ random variables.

However, unlike in the case of indepenedent Gaussian random variables, in this case $X+Y$ cannot be Gaussian. To see this, note that $X+Y = (S+1)X$, and $S+1$ takes the values 0 or 2 with equal probability. Thus

$$\mathbf{P}\{X + Y = 0\} = \tfrac{1}{2}.$$

Thus $X + Y$ is very much not Gaussian: a Gaussian $N(0, \sigma^2)$ is a continuous random variable for any $\sigma > 0$, and must therefore have $\mathbf{P}\{N(0,\sigma^2) = 0\} = 0$.

There is nothing particularly important about this example, and there is no need to remember it. It is included to drive home the point that you must be careful when you do computations with Gaussian random variables that are not independent. One safe solution would be to always work only with independent Gaussian random variables. In practice, however, there is frequently the need to model dependent random variables. For example, suppose we were to model the stock prices of IBM and Microsoft using Gaussian random variables. If the tech industry is doing well, the probably both prices will be high, while if the tech industry is doing badly, it is likely that both prices are low. Thus the two prices must be modelled as dependent random variables: we would like to capture the fact that they are more likely to be high together or be low together, than that one is low and the other is high.

In order to build useful dependent models, we would like to have a way of constructing dependent Gaussian random variables that still allows us to do computations as easily as in the independent case. To this end, we introduce the notion of *jointly Gaussian* random variables.

Definition 7.4.2. *Random variable X, Y are called* jointly Gaussian *if $aX + bY$ is Gaussian for every $a, b \in \mathbb{R}$. More generally, $X_1, X_2, \ldots, X_n$ are jointly Gaussian if $a_1X_1 + \cdots + a_nX_n$ is Gaussian for every $a_1, \ldots, a_n \in \mathbb{R}$.*

Remark 7.4.3. For notational simplicity, we will concentrate on the case of two jointly Gaussian random variables in this section. Everything we will say extends directly to more than two Gaussian random variables.

We have seen in Example 7.4.1 that it is perfectly possible to have two Gaussian random variables X, Y that are not jointly Gaussian. On the other hand, we have seen that independent Gaussian random variables are always jointly Gaussian. The good news is that one can easily generate jointly Gaussian random variables that are dependent, as we will shortly see. This makes jointly Gaussian variables a particularly useful tool for modelling dependence.

It is helpful to think about jointly Gaussian variables as vectors: if

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}, \qquad r = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Then the linear combination $aX + bY$ can be written as

$$aX + bY = r \cdot Z.$$

The assumption that X, Y are jointly Gaussian is now simply the statement that $r \cdot Z$ is Gaussian for every $r \in \mathbb{R}^2$. Let us compute its mean and variance:

$$\mathbf{E}(r \cdot Z) = a\mathbf{E}(X) + b\mathbf{E}(Y) = r \cdot \mu$$

and

$$\mathrm{Var}(r \cdot Z) = a^2\mathrm{Var}(X) + 2ab\mathrm{Cov}(X, Y) + b^2\mathrm{Var}(Y) = r \cdot \Sigma r,$$

where we defined the *mean vector* μ ans *covariance matrix* Σ as

$$\mu = \begin{pmatrix} \mathbf{E}(X) \\ \mathbf{E}(Y) \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \mathrm{Var}(Y) \end{pmatrix},$$

and the *covariance* between X and Y is defined as

$$\mathrm{Cov}(X, Y) := \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))).$$

We have therefore shown that if X, Y are jointly Gaussian, then

$$r \cdot Z \sim N(r \cdot \mu, r \cdot \Sigma r) \quad \text{for all } r.$$

When this is the case, we will write

$$Z \sim N(\mu, \Sigma).$$

Remark 7.4.4. Implicit in our notation $Z \sim N(\mu, \Sigma)$ is that the distribution of a jointly Gaussian vector *is completely determined by its mean vector and covariance matrix*. Why is this true? As we know the moment generating function of a Gaussian, we can easily compute

$$\varphi(r) := \mathbf{E}(e^{r \cdot Z}) = e^{r \cdot \mu + \frac{1}{2} r \cdot \Sigma r}.$$

Note that $\frac{\partial^p}{\partial x^p}\frac{\partial^q}{\partial y^q}\varphi(x,y)|_{x,y=0} = \mathbf{E}(X^pY^q)$ for every $p,q \geq 1$. Thus all moments $\mathbf{E}(X^pY^q)$ are determined by μ and Σ, and so by Taylor expansion we can compute $\mathbf{E}(f(X,Y))$ is determined by μ and Σ as well. This shows that the distribution of jointly Gaussian variables is completely determined by μ and Σ. It is even possible to obtain an explicit expression for the joint density of jointly Gaussian random variables X, Y:

$$\mathbf{P}\{X \in dx, Y \in dy\} = \frac{e^{-(r-\mu)\cdot\Sigma^{-1}(r-\mu)/2}}{\sqrt{\det(2\pi\Sigma)}}\,dx\,dy.$$

You can verify, for example, that this expression yields the moment generating function $\mathbf{E}(e^{r\cdot Z}) = e^{r\cdot\mu+\frac{1}{2}r\cdot\Sigma r}$. The explicit expression is somewhat tedious to use, however. In most cases, it is much more helpful for computations to use directly the definition of jointly Gaussian variables or the moment generating function than to use the explict form of the density.

Jointly Gaussian random variables provide a particularly convenient way of modelling dependent random variables, as their dependence is completely determined by their covariance matrix. Indeed, looking at the definition of the covariance, you can readily see that $\mathrm{Cov}(X,Y) > 0$ means that the variables are *positively correlated*: they are more likely to both exceed their mean or both fall short of their mean, than that one exceeds and the other falls short (that is, these two variables are usually both large or both small). This could be a good model of stock prices of companies in the same industry: IBM and Microsoft both do well when the technology sector is strong, and both do poorly when the technology sector is weak, so their ups and downs are likely in the same direction. Conversely, $\mathrm{Cov}(X,Y) < 0$ means that the variables are *negatively correlated*: if one is large then the other is more likely to be small, and vice versa. This could be a good model of stock prices of two competing industries, such as McDonalds and Subway: when health is popular Subway is up and McDonalds is down, while outside a health kick McDonalds might do better at the expense of Subway. Jointly Gaussian variables easily capture these interactions by choosing the covariance matrix appropriately.

When $\mathrm{Cov}(X,Y) = 0$, the variables X and Y are *uncorrelated*: they do not affect each other either positively or negatively. It turns out that if X, Y are uncorrelated jointly Gaussian random variables, then they are independent! To see this, we only have to note that its moment generating function is

$$\mathbf{E}(e^{aX+bY}) = e^{a\mathbf{E}(X)+b\mathbf{E}(Y)+\frac{1}{2}a^2\mathrm{Var}(X)+\frac{1}{2}b^2\mathrm{Var}(Y)} = \mathbf{E}(e^{aX})\mathbf{E}(e^{bY}),$$

which is precisely the moment generating function of independent Gaussian random variables X, Y. It seems very intuitive that uncorrelated random variables are independent. Beware, however: this is only true in general for *jointly* Gaussian random variables! As you can see in Example 7.4.1, uncorrelated Gaussian random variables are not necessarily independent when they are not jointly Gaussian. (X, Y are called independent if $\mathbf{E}(f(X)g(Y)) =$

$\mathbf{E}(f(X))\mathbf{E}(g(Y))$ for all functions f, g, while X, Y are uncorrelated if $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$: in general, this is a much weaker condition!)

In practice, the easiest way to generate jointly Gaussian random variables is often to start from independent Gaussian random variables. For example, if $Z_1, Z_2 \sim N(0,1)$ are independent random variables and $a_i, b_i \in \mathbb{R}$, then

$$X = a_0 + a_1 Z_1 + a_2 Z_2, \qquad Y = b_0 + b_1 Z_1 + b_2 Z_2$$

are jointly Gaussian (why?) You can choose the constants a_i, b_i to generate any covariance that you like.

Example 7.4.5. Let $X, Y, Z \sim N(0,1)$ be independent Gaussian random variables. We claim that $X + Y + Z$ and $X + Y - 2Z$ are independent random variables. To see why, note that $(X+Y+Z, X+Y-2Z)$ is a jointly Gaussian vector. Hence, it is enough to observe that

$$\mathrm{Cov}(X + Y + Z, X + Y - 2Z) = \mathbf{E}((X + Y + Z)(X + Y - 2Z)) = 0.$$

It would have been much harder to come to this conclusion by computing the joint density of $(X + Y + Z, X + Y - 2Z)$, say!

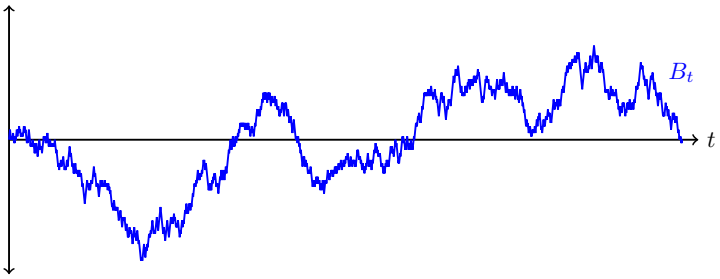
Example 7.4.6. Let $\{B_t\}$ be standard Brownian motion and $s < t$. We claim that B_s, B_t are jointly Gaussian. Why is this true? From the properties of Brownian motion, we know that B_s and $B_t - B_s$ are independent Gaussian random variables. Therefore, B_s and $B_t = (B_t - B_s) + B_s$ are jointly Gaussian. We can compute their covariance as

$$\begin{aligned}
\mathrm{Cov}(B_s, B_t) &= \mathbf{E}(B_s B_t) \\
&= \mathbf{E}(B_s(B_t - B_s)) + \mathbf{E}(B_s^2) \\
&= \mathbf{E}(B_s)\mathbf{E}(B_t - B_s) + s \\
&= s.
\end{aligned}$$

More generally, for any $t_1, \ldots, t_n \geq 0$, the random variables $B_{t_1}, \ldots, B_{t_n}$ are jointly Gaussian. A random process with this property is called a *Gaussian process* (so Brownian motion is a special kind of Gaussian process).

7.5 Sample paths of Brownian motion

Up to this point, we have mostly concentrated on understanding the distribution of Brownian motion B_t at a fixed time t. This led us to discover the Gaussian distribution, the central limit theorem, and several other interesting ideas. In this section, we return to studying the properties of Brownian motion as a random process: that is, how does the path of a Brownian particle vary over time? A random sample path is illustrated in the following figure:



The paths of Brownian motion have some remarkable properties, only some of which we will investigate (there are entire books written about this topic!)

Unlike the Poisson process, which is a counting process and therefore has jumps, the paths of Brownian motion are continuous (as you can see in the above picture): this is a good thing, as a pollen particle cannot teleport itself! The natural question we ask next is with what "speed" the particle is moving, that is, what is the derivative of the path? That is, let us try to compute

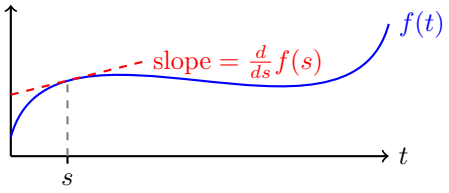
$$\frac{d}{dt}B_t = \lim_{\varepsilon \downarrow 0} \frac{B_{t+\varepsilon} - B_t}{\varepsilon}.$$

Recall from the definition of Brownian motion that $B_{t+\varepsilon} - B_t \sim N(0, \varepsilon)$, so

$$\frac{B_{t+\varepsilon} - B_t}{\varepsilon} \sim N(0, \tfrac{1}{\varepsilon}).$$

As $\varepsilon \downarrow 0$, the right-hand side blows up! Did we make a mistake? We did not: the derivative of Brownian motion does not exist. That is, the path $t \mapsto B_t$ of Brownian motion is *continuous but nowhere differentiable*. If you have never taken a course in real analysis, you may perhaps be surprised that this is even possible: every continuous function you encounter in a garden-variety calculus course tends to be also differentiable, at least at most points. Brownian motion provides a completely natural example where this is not the case.

While this property of Brownian motion might sound scary, it actually makes a lot of sense. Let us try to explain informally why it is true. Suppose that the function $f(t)$ is differentiable. Then $\frac{d}{dt}f(t)$ is its *slope* at time t:



If the the slope at time t is positive $\frac{d}{dt}f(t) > 0$, then we know that the function will increase at least in the immediate future; while if the slope is negative $\frac{d}{dt}f(t) < 0$, then we know that the function will decrease in the immediate

future. Therefore, a differentiable function has hidden in its definition a form of predictability: even though we have seen the function only until the present, we can predict whether it will go up or down in the immediate future by looking at its derivative. This sort of clairvoyance is impossible, however, for Brownian motion. The pollen particle is bombarded at each time by an independent water molecule; if it is bombarded from the left then the position will increase, and if it is bombarded from the right its position will decrease. Because of the independence of the water molecules, it is impossible to predict which of these will occur before we have seen the outcome of the bombardment. Therefore, it would be impossible for a path that is defined in this way to be differentiable. This is the logical conclusion of the intuitive fact that we are surprised at every time about what direction the particle will move.

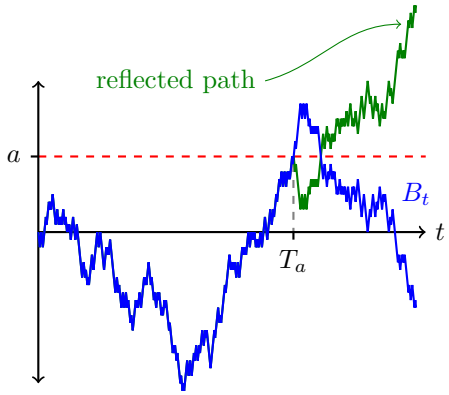
Remark 7.5.1. Our explanation of the non-differentiability of Brownian motion is very important in financial mathematics. Suppose stock prices were differentiable; then we could always make profit without running any risk whatsoever by buying stock when the derivative of the stock price is positive and selling stock when the derivative of the stock price is negative. That is, we can always make money "for free" because knowing the slope of the stock price allows us to look into the future. As you might have noticed, however, it does not appear to be the case that we are all incredibly rich. Therefore, there must be something wrong with this idea. The reason it fails is that stock prices behave like Brownian motion: they do not have a derivative, and so there is no way to predict whether they will go up and down until it is too late.

Let us now turn to another computation involving the properties of Brownian paths. When we discussed the Poisson process, we tried to understand the distribution of the time of the first arrival, the second arrival, etc. The time of kth arrival of a Poisson process $\{N_t\}$ is the first time t such that $N_t = k$. The analogous quantity for Brownian motion $\{B_t\}$ is the hitting time

$$T_a := \min\{t : B_t = a\}$$

for $a > 0$. We are going to compute the distribution of this random variable. To this end, we are first going to develop a seemingly very different property of Brownian paths that is known as the *reflection principle*.

Consider the Brownian path from time T_a onward. At each time, the path goes up or down with equal probability. Therefore, if we were to exchange the up and down directions after time T_a, the new *reflected* path must have the same likelihood of occuring as the original path:



Because the original path and the reflected path have the same probability, the Brownian motion is equally likely to lie above $B_t > a$ or below $B_t < a$ at any time $t \geq T_a$: that is, we have established the interesting identity

$$\mathbf{P}\{B_t > a | T_a \leq t\} = \frac{1}{2}.$$

This is called the "reflection principle" of Brownian paths.

Remark 7.5.2. Of course, because Brownian motion has a continuous distribution, the probability of any individual path is zero. Therefore, the statement that "the original path and the reflected path have the same probability" is a bit imprecise. Rather, you should notice that we can run the argument without such difficulties for a simple symmetric random walk, and then take the continuous time limit to obtain the result for Brownian motion.

The reflection principle turns out to make it very easy to compute the distribution of T_a. To see why, let us compute its cumulative distribution function. Notice that, by the reflection principle,

$$\begin{aligned} \tfrac{1}{2}\,\mathbf{P}\{T_a \leq t\} &= \mathbf{P}\{B_t > a | T_a \leq t\}\,\mathbf{P}\{T_a \leq t\} \\ &= \mathbf{P}\{B_t > a, T_a \leq t\} \\ &= \mathbf{P}\{B_t > a\}, \end{aligned}$$

where we have used that $B_t > a$ already implies $T_a \leq t$ (if the Brownian motion is more than a at time t and zero at time 0, then it must have hit a at some time in between). Therefore, the CDF of T_a is given by

$$\mathbf{P}\{T_a \leq t\} = 2\mathbf{P}\{B_t > a\} = 2\mathbf{P}\left\{N(0,1) > \frac{a}{\sqrt{t}}\right\} = 2\int_{a/\sqrt{t}}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx,$$

and its density is given by

$$\frac{d}{dt}\mathbf{P}\{T_a \leq t\} = \frac{ae^{-a^2/2t}}{\sqrt{2\pi t^3}}, \qquad t \geq 0.$$

This might look superficially like a Gaussian distribution, but it is completely different: notice that the variable t of the density appears in the denominator of the exponential, not in the numerator. This distribution is quite new to us. To my knowledge, it does not have a name.

Now that we have derived the distribution of T_a, we can compute various basic properties. For example, note that

$$\mathbf{P}\{T_a < \infty\} = 2\int_0^\infty \frac{e^{-x^2}}{\sqrt{2\pi}}\,dx = \int_{-\infty}^\infty \frac{e^{-x^2}}{\sqrt{2\pi}}\,dx = 1,$$

where we have used that e^{-x^2} is a symmetric function and that the integral of the $N(0,1)$ density is one (as it is for any probability density). Therefore, Brownian motion will always eventually hit the level a. However,

$$\mathbf{E}(T_a) = \int_0^\infty t\,\frac{ae^{-a^2/2t}}{\sqrt{2\pi t^3}}\,dt = \int_0^\infty \frac{1}{\sqrt{t}}\frac{ae^{-a^2/2t}}{\sqrt{2\pi}}\,dt = \infty,$$

as the integrand is of order $\sim 1/\sqrt{t}$ when t is large (the integral of the function $1/\sqrt{t}$ is infinity). Therefore, the expected time until Brownian motion hits the level a is infinite. These properties should not surprise you: we derived exactly the same properties for the simple symmetric random walk.

Using similar arguments, we can derive various other interesting properties of Brownian paths. Let us consider two amusing examples.

Example 7.5.3 (Gas price betting pool). The price of gas is modelled as a standard Brownian motion $\{B_t\}$ (this is not terribly realistic, and Brownian motion starts at zero and might go negative—an unlikely scenario for gas prices in my experience; still, it is a reasonable starting point for a model). You and your friend have made a bet. You believe that the maximal gas price over the next week will exceed some level a; your friend thinks you're crazy. What is the probability you will win your bet?

The goal of this problem is evidently to compute

$$\mathbf{P}\left\{\max_{0\le s\le 1} B_s \ge a\right\}.$$

How can we do this? Notice that if the maximal gas price in the next week exceeds a, then the gas price must hit a at some point during the week (and vice versa). As we know the distribution of T_a, we immediately compute

$$\mathbf{P}\left\{\max_{0\le s\le 1} B_s \ge a\right\} = \mathbf{P}\{T_a \le 1\} = 2\int_a^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx.$$

Note that there is nothing special about considering the one-week period. The maximal gas price after t weeks satisfies

$$\mathbf{P}\left\{\max_{0\le s\le t} B_s \ge a\right\} = \mathbf{P}\{T_a \le t\} = 2\,\mathbf{P}\{B_t \ge a\} = \mathbf{P}\{|B_t| \ge a\},$$

where we used that

$$\begin{aligned}\mathbf{P}\{|B_t| \geq a\} &= \mathbf{P}\{B_t \geq a \text{ or } -B_t \geq a\} \\ &= \mathbf{P}\{-B_t \geq a\} + \mathbf{P}\{B_t \geq a\} \\ &= 2\mathbf{P}\{B_t \geq a\}.\end{aligned}$$

We have therefore established the remarkable fact that the random variables

$$\max_{0\leq s\leq t} B_t \quad \text{and} \quad |B_t| \quad \text{have the same distribution.}$$

This is surprising, as the random processes $t \mapsto \max_{0\leq s\leq t} B_t$ and $t \mapsto |B_t|$ look nothing like one another! (The first is always increasing, while the second looks like a Brownian motion that is reflected by the horizontal axis.) There is no good explanation for this: it is one of the miracles of Brownian motion.

Example 7.5.4 (Another miracle). Let us derive another fun fact about the distribution of the hitting time T_a. Notice that

$$\mathbf{P}\{T_a \leq t\} = 2\,\mathbf{P}\{B_t \geq a\} = \mathbf{P}\{|B_t| \geq a\} = \mathbf{P}\{B_t^2 \geq a^2\}.$$

As $B_t \sim N(0,t)$, this random variable has the same distribution as $\sqrt{t}Z$ where $Z \sim N(0,1)$. We can therefore write

$$\mathbf{P}\{T_a \leq t\} = \mathbf{P}\{tZ^2 \geq a^2\} = \mathbf{P}\left\{\frac{a^2}{Z^2} \leq t\right\}.$$

Therefore,

$$T_a \quad \text{and} \quad \frac{a^2}{Z^2} \quad \text{have the same distribution.}$$

This completely useless fact is nonetheless extremely amusing.

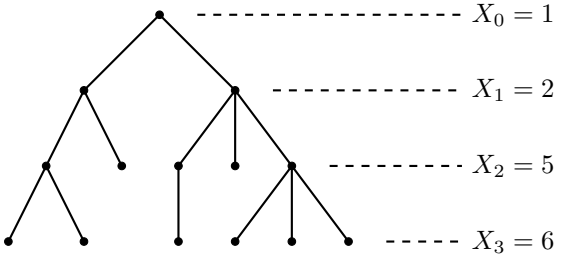
8

Branching Processes

In this short chapter, we introduce a basic model of a random tree that is used in evolutionary biology, genetics, physics, and in many other areas: the *branching process*. It provides a nice illustration of the methods that we have developed so far, and also serves as another step towards the notion of Markov chains that will be introduced in the next chapter.

8.1 The Galton-Watson process

Consider an individual starting a family tree. This individual forms the 0th generation of the tree and has a random number of children, which form the first generation. Each of these children will independently have a random number of their own children that form the second generation; and so on. We will denote by X_n the number of descendants in the nth generation (note that, by assumption, $X_0 = 1$). One realization of such a family tree could be:



We will assume that each person has children independent of the others in his generation and of all individuals in the past genearations. Let us denote by p_k the probability that an individual has k children, where $k \geq 0$ (so p_0 is the probability that an individual has no children, p_1 is the probability that an individual has exactly one child, etc.) Of course, as every individual has *some* number of children, we must have $p_0 + p_1 + p_2 + \cdots = 1$.

The random family tree that we have just described is known as a *Galton-Watson tree* or, more generally, as a (type of) *branching process*. Such models are used in various different areas of science, engineering, and even in pure mathematics (to study the properties of random graphs).

Example 8.1.1 (British aristocracy). In England, aristocratic family names are passed down from father to son. In the disasterous event of a marriage that leads exclusively to daughters (or to no children at all), the family name becomes extinct. The anthropologist Sir Francis Galton (1822–1911) was interested in computing the probability that this would happen, and posted this as an open problem in the *Educational Times* in 1873. The problem was solved by the Reverend Henry Watson (1827–1903), who had studied mathmatics at Cambridge. The Galton-Watson tree is named after them. (Incidentally, Galton was not the nicest character by today's standards; for example, he is the inventor and strong proponent of eugenics).

In this model, each "individual" is a male, so that only male children are counted (that is, p_0 is the probability of having zero sons, p_1 is the probability of having one son, etc.) Then X_n is the number of male descendents in the nth generation. Galton and Watson were interested in understanding the probability that the family name goes extinct (i.e., that $X_n = 0$ after a finite number of generations). We will compute this probability in the next section.

It turns out that the Galton-Watson model is not a terribly good model of family name inheritance, as individuals tend to change their family names over time. However, the model is very useful in evolutionary biology and genetics.

Example 8.1.2 (Nuclear chain reactions). In the 1930s, the Hungarian physicist Leó Szilárd (1898–1964) conceived the notion of a nuclear chain reaction. To model it, he essentially reinvented the notion of a Galton-Watson process. His computations, which convinced him that nuclear chain reactions are possible, had major international consequences. He wrote the letter in 1939 (which was signed by Einstein) that convinced President Franklin D. Roosevelt to start the Manhattan Project that created the atomic bomb.

A nuclear chain reaction starts with a single neutron that is shot into a mass of uranium atoms. With some probability, the neutron will pass right through without hitting an atom, in which case nothing happens. However, if the neutron hits an atom, three neutrons are produced that now travel independently. If any of these three neutrons hits an atom, another three neutrons are produced, etc. An atomic bomb explodes if the number of neutrons grows exponentially (as would be the case if every neutron hits an atom with probability one). One is interested in computing the probability of explosion.

Before we turn to more involved analysis of the Galton-Watson tree, let us perform a simple computation of the expected size $\mathbf{E}(X_n)$ of the nth generation. For convenience, let us define the quantity

$$\mu = \sum_{k=0}^{\infty} k p_k :$$

μ is the expected number of children that a single individual has. Note that

$$\mathbf{E}(X_{n+1}|X_n = k) = k\mu,$$

as each of the k individuals in generation n has μ children on average. Thus

$$\begin{aligned}\mathbf{E}(X_{n+1}) &= \sum_{k=0}^{\infty} \mathbf{E}(X_{n+1}|X_n = k)\,\mathbf{P}\{X_n = k\} \\ &= \mu \sum_{k=0}^{\infty} k\,\mathbf{P}\{X_n = k\} = \mu\,\mathbf{E}(X_n).\end{aligned}$$

As $X_0 = 1$, we can iterate this identity to compue $\mathbf{E}(X_n)$: we have $\mathbf{E}(X_1) = \mu\mathbf{E}(X_0) = \mu$, $\mathbf{E}(X_2) = \mu\mathbf{E}(X_1) = \mu^2$, etc. That is, we have shown that

$$\mathbf{E}(X_n) = \mu^n.$$

Note that if $\mu < 1$ (each individual has less than one child on average), then $\mathbf{E}(X_n) \to 0$ as $n \to \infty$: in this case, the family tree becomes extinct. But if $\mu > 1$ (each individual has more than one child on average), then $\mathbf{E}(X_n) \to \infty$ exponentially fast: in this case, the expected size of the family tree explodes.

Example 8.1.3 (Nuclear chain reactions). In Example 8.1.2, p_0 is the probability that a neutron does not hit an atom: in this case the neutron has no descendents. On the other hand, p_3 is the probability that a neutron hits an atom and hence creates three new neutrons. As these are the only two possibilities (in our simple model), we have $p_0 + p_3 = 1$. Therefore,

$$\mu = 0 \times p_0 + 3 \times p_3 = 3p_3.$$

Hence, if the probability of hitting an atom $p_3 < \frac{1}{3}$, then the system fizzles out. If $p_3 > \frac{1}{3}$, then the nuclear chain reaction could result in an explosion!

8.2 Extinction probability

In the previous section, we compute the expected number of individuals $\mathbf{E}(X_n)$ in the nth generation of a Galton-Watson tree. In practice, a question that is often of interest is the probability that the Galton-Watson tree will go extinct (that is, that $X_n = 0$ after a finite number of generations). Even if the expected number of individuals $\mathbf{E}(X_n)$ blows up as $n \to \infty$, this does not mean that the tree cannot go extinct with positive probability.

Example 8.2.1 (Expected population vs. extinction probability). Consider a sequence of random variables X_n with $\mathbf{P}\{X_n = 0\} = \mathbf{P}\{X_n = 2^n\} = \frac{1}{2}$. Then $\mathbf{E}(X_n) \to \infty$ exponentially fast, but X_n goes extinct with probability $\frac{1}{2}$.

Of course, these random variables X_n are not actually generated by any branching process: they are just engineered to illustrate that the expected population size cannot tell us whether or not the process goes extinct. We will shortly see that the same phenomenon occurs in Galton-Watson trees.

Let us define the *extinction probability*

$$\eta := \mathbf{P}\Big\{ \lim_{n\to\infty} X_n = 0 \Big\}.$$

We also define

$$\eta_n := \mathbf{P}\{X_n = 0\},$$

the probability that the tree is extinct by the nth generation. Note that:

a. $\eta_0 = \mathbf{P}\{X_0 = 0\} = 0$ because we always have $X_0 = 1$.

b. If the population is extinct by generation n, then it is still extinct in generation $n+1$. In other words, $\{X_n = 0\} \subset \{X_{n+1} = 0\}$. This implies that $\eta_1 \le \eta_2 \le \eta_3 \le \ldots$, that is, $\{\eta_n\}$ is an increasing sequence.

c. The extinction probability $\eta = \lim_{n\to\infty} \eta_n$.

We will first try to compute the probabilities η_n, and then take the limit as $n \to \infty$ to obtain a formula of η.

To compute η_n, we use first step analysis, which we already know from the study of random walks! While a branching process is not a random walk, the idea is similar. Conditioning on the size of the first generation, we get

$$\eta_{n+1} = \mathbf{P}\{X_{n+1} = 0\} = \sum_{k=0}^{\infty} \mathbf{P}\{X_{n+1} = 0 | X_1 = k\}\, \mathbf{P}\{X_1 = k\}.$$

$\mathbf{P}\{X_{n+1} = 0 | X_1 = k\}$ is the probability of extinction by time $n+1$ given that there are k individuals in the first generation. For the family tree to extinct, the line of descendants from each of these k individuals must go extinct in n generations (from generation 1 to generation $n+1$). These k lines can be considered as k *independent* family trees up to generation n, so

$$\mathbf{P}\{X_{n+1} = 0 | X_1 = k\} = \prod_{i=1}^{k} \mathbf{P}\{X_n = 0\} = \prod_{i=1}^{k} \eta_n = \eta_n^k.$$

Substituting into our first step analysis, we obtain

$$\eta_{n+1} = \sum_{k=0}^{\infty} \eta_n^k p_k.$$

We can rewrite this relation in a more compact way by defining the *generating function* $g(z)$ of the Galton-Watson tree by

$$g(x) := \sum_{k=0}^{\infty} x^k p_k.$$

We have shown that

$$\eta_{n+1} = g(\eta_n), \quad n \geq 0,$$
$$\eta_0 = 0.$$

In principle, this allows us to compute the extinction probability η_n in every generation n: for example, $\eta_5 = g(g(g(g(g(0)))))$. This is a bit annoying. Fortunately, we are not so much interested in η_n, but in $\eta = \lim_{n\to\infty} \eta_n$. We can use our equation to obtain a very useful formula for computing η.

To compute η, we note that as g is a continuous function

$$\eta = \lim_{n\to\infty} \eta_n = \lim_{n\to\infty} g(\eta_{n-1}) = g\left(\lim_{n\to\infty} \eta_{n-1}\right) = g(\eta).$$

Therefore, η is a solution of the equation

$$x = g(x).$$

This is very useful—it means that we can find η by looking where the graphs of the functions $y = x$ and $y = g(x)$ intersect! The problem is that these graph might intersect at more than one place, in which case it is not immediately clear which point is actually the correct extinction probability.

To understand better what is going on, let us do a bit a high-school style plotting. First, as η is a *probability*, it must be between 0 and 1. We therefore only have to worry about what happens to our functions for $x \in [0, 1]$. Next,

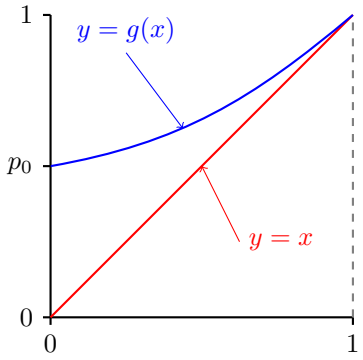
$$g(1) = \sum_{k=0}^{\infty} p_k = 1, \qquad g(0) = p_0,$$

and

$$g'(1) = \sum_{k=1}^{\infty} k p_k = \mu, \qquad g''(x) = \sum_{k=2}^{\infty} k(k-1)x^{k-2} p_k \geq 0$$

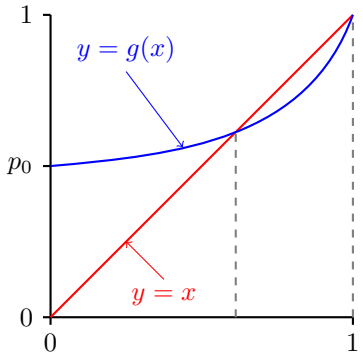
for $x \geq 0$. Therefore, the function $x \mapsto g(x)$ is convex, $g(0) = p_0$, $g(1) = 1$, $g'(1) = \mu$. There are two possible situations:

a. If $\mu \leq 1$, then $g'(1) \leq 1$ and the plot looks something like this:



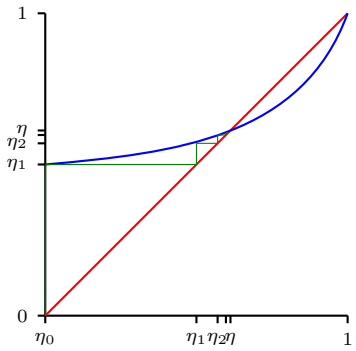
In this case, clearly the only solution of $x = g(x)$ with $x \in [0, 1]$ is $x = 1$. Therefore, if $\mu \leq 1$, then the extinction probability $\eta = 1$: that is, if each individual has at most one child on average, the human race would go extinct. (Motto: go forth and multiply!)

b. If $\mu > 1$, then $g'(1) > 1$ and the plot looks something like this:



Now there are two solutions of $x = g(x)$ for $X \in [0, 1]$! Which one is the correct extinction probability?

To understand this, recall that $\eta_{n+1} = g(\eta_n)$ with $\eta_0 = 0$. We can compute η_0, η_1, η_2, etc. by following a zig-zag path on the above plot:



You can see in this plot that the zig-zag path cannot cross the smallest point at which the graphs of $y = g(x)$ and $y = x$ intersect. Thus we have shown:

> *The extinction probability η is the* smallest *solution of the equation $x = g(x)$ in the interval $x \in [0, 1]$.*

Once we are given the probabilities $p_0, p_1, p_2, \ldots$, we can compute the generating function $g(x)$ and use this principle to compute the extinction probability.

Example 8.2.2 (Nuclear chain reactions). In Example 8.1.2, suppose that $p_0 = p_3 = \frac{1}{2}$, that is, each neutron has equal probability of hitting or not hitting an

atom. In this case, the expected offspring of each neutron is $\mu = 3p_3 = \frac{3}{2} > 1$ neutrons. Therefore, the expected number of neutrons in the nth generation

$$\mathbf{E}(X_n) = \left(\frac{3}{2}\right)^n$$

blows up exponentially fast as $n \to \infty$. This does not necessarily mean, however, that every atomic bomb explodes!

Let us compute the probability η that the atomic bomb fizzles out. In the present case, the generating function is given by

$$g(x) = p_0 + p_3 x^3 = \frac{1 + x^3}{2}.$$

We therefore want to solve the equation

$$g(x) = x \;\Leftrightarrow\; x^3 - 2x + 1 = 0.$$

Since $x = 1$ is always a solution, we can factor out $(x - 1)$:

$$x^3 - 2x + 1 = (x - 1)(x^2 + x - 1) = 0.$$

The solutions of the quadratic equation $x^2 + x - 1 = 0$ are $x = \frac{-1-\sqrt{5}}{2} < 0$ and $x = \frac{-1+\sqrt{5}}{2} \approx 0.62$. Thus $x = 1$ and $x = \frac{-1+\sqrt{5}}{2}$ are the only solutions in $[0, 1]$, and we conclude that the probability of extinction is the smallest solution

$$\eta = \frac{-1 + \sqrt{5}}{2} \approx 0.62.$$

Note that even though the expected number of neutrons grows exponentially, the probability that this atomic bomb fizzles out is even more than one half! As we have discussed above, there is no contradiction between these statements.

9

Markov Chains

Many random processes that we have encountered so far (Bernoulli and Poisson processes, random walks, Brownian motion) were based on the idea that what happens in the future is independent of what has happened to date. For example, the arrivals of customers in the future are independent of the arrivals to date; and the motion of a Brownian particle in the future is independent of its path to date. There are many situations where this is not the case, however. For example, your mood tomorrow is probably quite dependent on your mood today: if you are in a good mood today, you will be more likely to be in a good mood tomorrow also. To capture this idea, we introduce in this chapter an important class of random processes called *Markov Chains*.

For simplicity, we will concentrate on the case of Markov chains with *finite* state space and discrete time. A theory in general state spaces or continuous time exists, but is more advanced than the level of this course.

9.1 Markov chains and transition probabilities

Informally, *a Markov chain is a random process whose future behavior depends only on its present state.* As this definition is rather vague, we immediately begin by turning it into a sensible mathematical statement.

Definition 9.1.1. *A random process $\{X_n\}_{n\geq 0}$, where each X_n takes values in a finite set D, is called a* Markov chain *if*

$$\mathbf{P}\{X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0\} = \mathbf{P}\{X_{n+1} = x_{n+1} | X_n = x_n\}$$

for all $n \geq 0$ and $x_1, \ldots, x_{n+1} \in D$.

In words, a Markov chain is a random process whose state tomorrow (at time $n+1$), conditionally on the history of the process to date, depends only on the value of the process today (at time n).

The definition of a Markov chain leaves open the possibility that what happens tomorrow depends not just on the state today, but also on the day of the week (because $\mathbf{P}\{X_{n+1} = j | X_n = i\}$ might depend on n). We will almost always consider Markov chains for which this is not the case.

Definition 9.1.2. *A Markov chain $\{X_n\}$ is called* time-homogeneous *if*

$$\mathbf{P}\{X_{n+1} = j | X_n = i\} = P_{ij}$$

for all $n \geq 0$. We call P_{ij} the transition probability *from i to j, and the matrix $P = (P_{ij})_{i,j \in D}$ is called the* transition probability matrix.

Remark 9.1.3. Unless stated otherwise, we use the convention that any Markov chain is by default considered to be time-homogeneous.

Example 9.1.4 (Mood swings). Let us model whether you are in a good mood (g) or bad mood (b). The state space of this Markov chain is $D = \{g, b\}$. If you are in a good mood today, then the probability that you are in a bad mood tomorrow is p; if you are in a bad mood today, the probability that you are in a good mood tomorrow is q. Therefore, we must have

$$P_{gb} = p, \quad P_{gg} = 1 - p, \quad P_{bg} = q, \quad P_{bb} = 1 - q.$$

Example 9.1.5 (English text). For many purposes (for example, data compression or predictive texting), it is necessary to have a simple model of English text. A common way to do this is to model text as a Markov chain whose state space $D = \{a, b, c, \ldots, z\}$ is the alphabet. The transition probability matrix P determines for each given letter the fraction of occurences in English text in which it is followed by another given letter (this is something you can easily estimate from data, for example, from the works of Shakespeare).

There are incredibly many other potential examples. In fact, it turns out that Markov chains appear, in one form or another, in almost every area of science and engineering. While Markov chains were originally developed by the Russian mathematician A. A. Markov in 1906 from a purely mathematical motivation, he quickly realized the utility of his discovery of applications. In 1913, Markov himself applied his chains to study the alternation of vowels and consonants in Pushkin's verses. Only a few years later, in 1917, Markov chains were used by Erlang to study waiting times in the telephone network. Thousands of other application of Markov chains have been developed since.

The central object in the theory of Markov chains is the transition probability matrix P: it determines, given the state today, the probability of each state tomorrow. What properties must P satisfy?

- $0 \le P_{ij} \le 1$ (as P_{ij} is a probability!)
- $\sum_{j \in D} P_{ij} = \sum_{j \in D} \mathbf{P}\{X_{n+1} = j | X_n = i\} = 1$ (probabilities sum to one).

Thus P is a nonnegative matrix whose rows sum to one. Such matrices are called *stochastic matrices*. Every stochastic matrix defines the transition probability matrix of a Markov chain.

In addition to the transition probability matrix, we must also specify the starting point of the Markov chain. Let us define the convenient notation

$$\mathbf{P}_{x_0}\{\,\cdot\,\} := \mathbf{P}\{\,\cdot\,|X_0 = x_0\}.$$

We claim that once we specify the starting point x_0 and transition probability matrix P, the distribution of the Markov chain is completely determined. To see why this is true, note for example that

$$\begin{aligned}
&\mathbf{P}_{x_0}\{X_3 = x_3, X_2 = x_2, X_1 = x_1\} \\
&= \mathbf{P}_{x_0}\{X_3 = x_3 | X_2 = x_2, X_1 = x_1\}\, \mathbf{P}_{x_0}\{X_2 = x_2 | X_1 = x_1\}\, \mathbf{P}_{x_0}\{X_1 = x_1\} \\
&= \mathbf{P}\{X_3 = x_3 | X_2 = x_2\}\, \mathbf{P}\{X_2 = x_2 | X_1 = x_1\}\, \mathbf{P}\{X_1 = x_1 | X_0 = x_0\} \\
&= P_{x_0 x_1} P_{x_1 x_2} P_{x_2 x_3},
\end{aligned}$$

where we used the definition of conditional probability in the first equality and the Markov property in the second equality. More generally, we have

$$\mathbf{P}_{x_0}\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = P_{x_0 x_1} P_{x_1 x_2} \cdots P_{x_{n-1} x_n}.$$

Therefore, once x_0 and P have been specified, the distribution of the Markov chain $\{X_n\}$ is completely determined. This is very convenient: to model a Markov chain, all we need to do is design a stochastic matrix P!

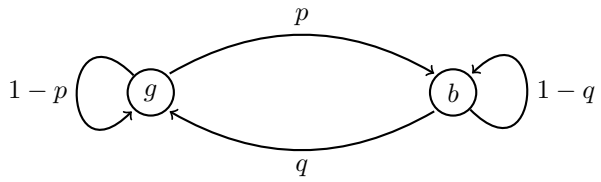
Example 9.1.6 (Mood swings). In our mood swings example, we had

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

We can therefore compute, for example,

$$\mathbf{P}_g\{X_1 = g, X_2 = g, X_3 = b, X_4 = g\} = P_{gg} P_{gg} P_{gb} P_{bg} = (1-p)^2 pq.$$

It is often helpful to represent Markov chains graphically in the form of a state diagram. Each vertex of this diagram is one state, and arrows indicate each possible transition and its probability (we do not include impossible transitions, that is, those that occur with probability zero). For example, our mood swings example corresponds to the following state diagram:



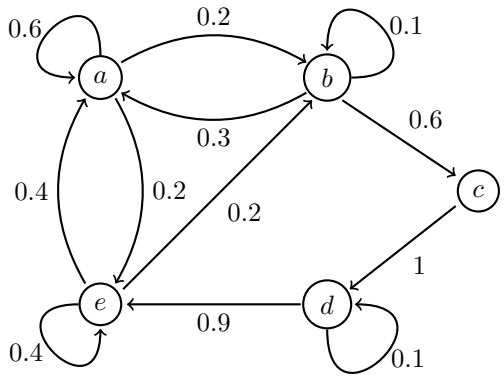
You should imagine the Markov chain as being generated by a frog jumping around on this graph. In each step, the frog randomly chooses one of the outgoing arrows with the indicated probabilities, and then jumps to the end of that arrow. The process repeats in the next step, etc.

Here is a more complicated example along the same lines.

Example 9.1.7 (State diagram). Consider the Markov chain on $D = \{a, b, c, d, e\}$ with the following transition probability matrix:

$$P = \begin{pmatrix} 0.6 & 0.2 & 0 & 0 & 0.2 \\ 0.3 & 0.1 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.1 & 0.9 \\ 0.4 & 0.2 & 0 & 0 & 0.4 \end{pmatrix}.$$

Note that this is indeed a stochastic matrix: the entries are nonnegative and each row sums to one (as is required of transition probabilities!) The state diagram for this Markov chain is easily drawn, as follows:



Once again, the state diagram helps us visualize what the path of the Markov chain looks like: imagine a frog jumping on this graph.

Let $\{X_n\}$ be a Markov chain. Given that we start at some state i, what is the probability that we land at a state j after k steps? For example, if you

are currently in a good mood, what is the probability that you will be in a bad mood one week from today? If $k = 1$, of course, the answer is

$$\mathbf{P}\{X_1 = j | X_0 = i\} = P_{ij}$$

by the definition of the Markov chain. Let us therefore try the case $k = 2$:

$$\begin{aligned}
&\mathbf{P}\{X_2 = j | X_0 = i\} \\
&= \sum_{a \in D} \mathbf{P}\{X_2 = j, X_1 = a | X_0 = i\} \\
&= \sum_{a \in D} \mathbf{P}\{X_2 = j | X_0 = i, X_1 = a\} \, \mathbf{P}\{X_1 = a | X_0 = i\} \\
&= \sum_{a \in D} p_{ia} p_{aj} \\
&= (P^2)_{ij},
\end{aligned}$$

where P^2 denotes the product PP (in the sense of matrix multiplication). Similarly, for $k = 3$, we can compute in the same way

$$\begin{aligned}
&\mathbf{P}\{X_3 = j | X_0 = i\} \\
&= \sum_{a,b \in D} \mathbf{P}\{X_3 = j, X_2 = b, X_1 = a | X_0 = i\} \\
&= \sum_{a,b \in D} p_{ia} p_{ab} p_{bj} \\
&= (P^3)_{ij}.
\end{aligned}$$

In general, we have the following result:

$$\mathbf{P}\{X_k = j | X_0 = i\} = (P^k)_{ij}.$$

It is not for nothing that we arranged our transition probabilities in a matrix P: matrix multiplication is a very useful tool in the study of Markov chains.

Example 9.1.8 (Accumulating rewards). Let $\{X_n\}$ be a Markov chain with state space D and transition probability matrix P. Suppose that every time we visit state j, we receive a reward of $f(j)$ dollars. Thus the reward received at time k is $f(X_k)$ dollars. Let $r(i)$ be the expected total reward that we accumulated up to time n given that the Markov chain starts at state i, that is,

$$r(i) = \mathbf{E}_i \left(\sum_{k=0}^{n} f(X_k) \right)$$

where $\mathbf{E}_i(Z) = \mathbf{E}(Z | X_0 = i)$. How can we compute $r(i)$? Note that

$$\begin{aligned} r(i) &= \sum_{k=0}^{n} \mathbf{E}_i(f(X_k)) \\ &= \sum_{k=0}^{n} \sum_{j \in D} f(j)\, \mathbf{P}\{X_k = j | X_0 = i\} \\ &= \sum_{k=0}^{n} \sum_{j \in D} (P^k)_{ij} f(j). \end{aligned}$$

It is convenient to express this expression in matrix-vector notation. If we write $D = \{i_1, i_2, \ldots, i_d\}$, we can represent any function on D as a d-dimensional column vector: for example, in the case of the functions f and r, we can write

$$f = \begin{pmatrix} f(i_1) \\ f(i_2) \\ \vdots \\ f(i_d) \end{pmatrix}, \qquad r = \begin{pmatrix} r(i_1) \\ r(i_2) \\ \vdots \\ r(i_d) \end{pmatrix}.$$

Moreover, the transition probability matrix P is itself a $d \times d$ matrix. Then the above expression can be written as follows:

$$r = \sum_{k=0}^{n} P^k f = (I + P + P^2 + \cdots + P^n) f.$$

In particular, we have reduced our probabilistic problem into a problem of matrix-vector multiplication, which is straightforward to work out (either by hand or using a computer). This is frequently the case in problems involving Markov chains: the final answer can often be reduced to matrix manipulations.

Let us illustrate this computation in a numerical example. Consider a Markov chain on $D = \{a, b, c\}$ with transition matrix and reward vector

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}, \qquad f = \begin{pmatrix} f(a) \\ f(b) \\ f(c) \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}.$$

What is the expected reward accumulated after $n = 2$ steps? We can compute

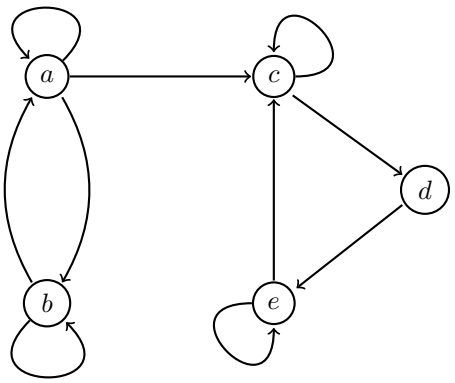
$$Pf = \begin{pmatrix} 1.5 \\ 3 \\ 2.5 \end{pmatrix}, \qquad P^2 f = P(Pf) = \begin{pmatrix} 2.25 \\ 2.75 \\ 2 \end{pmatrix}.$$

Therefore,

$$r = \begin{pmatrix} r(a) \\ r(b) \\ r(c) \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} + \begin{pmatrix} 1.5 \\ 3 \\ 2.5 \end{pmatrix} + \begin{pmatrix} 2.25 \\ 2.75 \\ 2 \end{pmatrix} = \begin{pmatrix} 4.75 \\ 7.75 \\ 8.5 \end{pmatrix}.$$

9.2 Classification of states

Consider a Markov chain with the following state diagram:



Recall that an arrow $i \to j$ in this diagram means that there is a positive probability that when the Markov chain is at the state i, it will go go the state j in the next time step. (We omitted the precise values of these probabilities from the diagram as they are irrelevant to the present discussion.)

Suppose this Markov chain starts at a or b. Then we can bounce back and forth between a and b, and we can also end up in the state c. However, the state c is a point of no return: once the Markov chain reaches c, there is no way to go back to the states a or b! Thus if we start in the states a or b, there may be a point in time after which we never return to these states again.

On the other hand, suppose we start in one of the states c, d, or e. No matter where the Markov chain goes next, it is always possible to get back to the starting state. Thus these states behave in a fundamentally different way than the states a and b. We call a and b *transient* states, while c, d, and e are *recurrent* states. More formally, these notions are defined as follows.

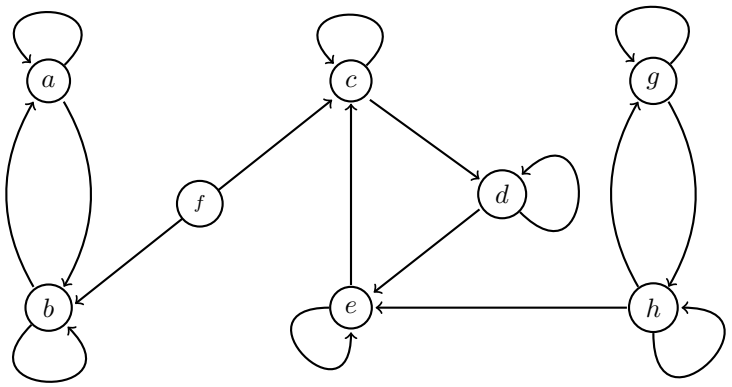
Definition 9.2.1. *We say that* there is a path from i to j, *denoted* $i \rightsquigarrow j$, *if there is a (directed) path in the state diagram that starts at i and ends at j. Equivalently, $i \rightsquigarrow j$ if and only if there is a nonzero probability of ending up at j sometime in the future given that we start at i.*

Definition 9.2.2. *A state i is* transient *if there exists a state j such that $i \rightsquigarrow j$ but $j \not\rightsquigarrow i$ (that is, there exists a "point of no return" j).*

Definition 9.2.3. *A state i is* recurrent *if for every state j with $i \rightsquigarrow j$, we also have $j \rightsquigarrow i$ ("no matter where we go, we can always get back").*

Note that recurrent is precisely the opposite of transient: thus every state is either recurrent or transient. The easiest way to find the recurrent and transient states is to stare at the state diagram; checking the above definitions is just common sense, as we did informally at the beginning of this section.

Example 9.2.4. Consider a Markov chain with the following state diagram:



The states f, g, h are transient: each of them has a path to a point of no return (which are?) On the other hand, a, b, c, d, e are recurrent.

In the previous example we see that not all recurrent states are equivalent. From a you can only reach b, and from b you can only reach a; thus a, b are recurrent, and a Markov chain that starts at either of these states will always remain in the set $\{a, b\}$. On the other hand, from c you can reach d, e, from d you can reach c, e, and from e you can reach c, d. Thus c, d, e are recurrent, and a Markov chain that starts at one of these states will always remain in the set $\{c, d, e\}$. Note that you can never go from a to c, say, even though both a and c are recurrent. It therefore makes sense to partition the set of all recurrent states into disjoint *recurrent classes*.

Definition 9.2.5. *Let i be a recurrent state. The set $\{j \in D : i \rightsquigarrow j\}$ is called the* recurrent class *that contains i. The collection of recurrent classes forms a partition of the set of all recurrent states.*

Example 9.2.6. In the Markov chain of Example 9.2.4, the recurrent classes are $\{a, b\}$ and $\{c, d, e\}$, while the set of transient states is $\{f, g, h\}$.

You might wonder at this point why we have insisted on the classification of states: the definitions are clear, but why do we care? It turns out that

the classification into recurrent classes and transient states greatly helps us understand how the Markov chain will behave on the long run. This is due to the following two important facts, which are not obvious from the way that we have defined recurrent and transient states.

a. If i is transient, then we *will* eventually never return to i. In particular, we can only visit transient states finitely often.

b. If i is recurrent, then we *will* eventually reach any state j such that $i \rightsquigarrow j$. In particular, we will revisit i and j infinitely often.

Example 9.2.7. Consider again the Markov chain of Example 9.2.4.

Suppose we start in the state a or b. Then the chain will always remain in the recurrent class $\{a, b\}$, but it will infinitely often switch back and forth between a and b. Similarly, if we start in c, d, or e, then we will forever remain in the recurrent class $\{c, d, e\}$, and visit each of these states infinitely often.

On the other hand, suppose we start in f. As this state is transient, we can visit it only a finitely often. Therefore, we must eventually leave f and end up in one of the recurrent classes $\{a, b\}$ or $\{c, d, e\}$. We do not know in advance where we will end up: either recurrent class is possible with some probability. However, once we ended up in one of the recurrent classes, we will remain there forever and bounce around inside the class infinitely often.

Similarly, if we start in g or h, we can only visit each of these states a finite number of times. Therefore, while we can bounce back and forth between g and h for some time, we will eventually leave these states forever and end up in the recurrent class $\{c, d, e\}$ (in this case we cannot end up in the recurrent class $\{a, b\}$ as there is no path from g, h to a, b).

As you can see in this example, the classification of states gives us a very good *qualitative* understanding of how a Markov chain will behave on the long run. Once this is understood, we might also want to answer quantitative questions: for example, what is the probability of ending up in a given recurrent class, and how long does it take on average to reach a recurrent class? We will revisit these quantitative questions in the following section.

It should be emphasized that the properties a. and b. above are far from obvious. For example, recall that a state i is called recurrent if there exists a "point of no return" j such that $i \rightsquigarrow j$ but $j \not\rightsquigarrow i$. However, the notation $i \rightsquigarrow j$ only means that there is a nonzero probability of ending up in j; on the other hand, point a. above claims that when this is the case, then we *will* eventually reach a point of no return with probability one! This is a Markov chain version of Murphy's law: if things *can* go wrong, they *will* go wrong. Clearly this is not obvious: there is something to prove here.

Rather than give a fully formal proof, let us sketch why these claims are true. Let us begin by considering a *transient* state i. Let p be the probability that the Markov chain never returns to i if we start the chain in the state i:

$$p = \mathbf{P}_i\{\text{Markov chain never returns to } i\}.$$

Because i is transient, by definition, $p > 0$. This means that there are two possibilities. With probability p, the chain never returns to i after time zero:

$$\mathbf{P}_i\left(\begin{array}{l} \vdash\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!\to \text{time} \\ i \ * \ * \ * \ * \ * \ * \ * \ * \ * \ * \ * \ * \ * \ * \ * \ \cdots \end{array} \right) = p,$$

where $*$ denotes any state but i. However, with probability $1 - p$, the chain will return to the state i after some time:

$$\mathbf{P}_i\left(\begin{array}{l} \vdash\!\!+\!\!+\!\!+\!\!+ \text{ - - - } \vdash\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!\to \text{time} \\ i \ * \ * \ * \ \cdots \ * \ * \ i \end{array} \right) = 1 - p.$$

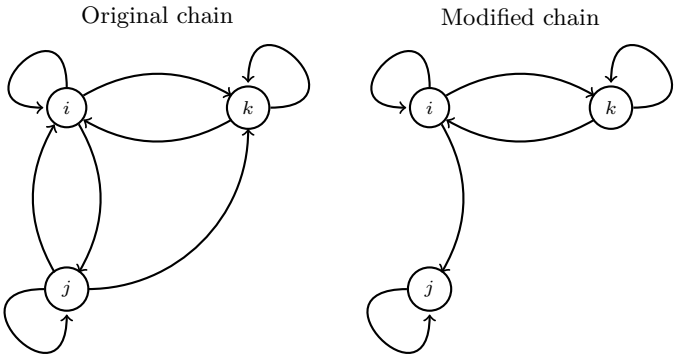
Now suppose the chain does return to the state i at some time T. By the Markov property, what happens after time T depends only on i: it is like we restarted the chain at the state i. Therefore, we again have a probability p of never returning to i again, and a probability $1 - p$ of returning to i again, independently of what happened in the past. In particular, we have

$$\mathbf{P}_i\left(\begin{array}{l} \vdash\!\!+\!\!+\!\!+\!\!+ \text{ - - - } \vdash\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!\to \text{time} \\ i \ * \ * \ * \ \cdots \ * \ * \ i \ * \ * \ * \ * \ * \ * \ * \ * \ \cdots \end{array} \right) = (1-p)p,$$

$$\mathbf{P}_i\left(\begin{array}{l} \vdash\!\!+\!\!+\!\!+\!\!+ \text{ - - - } \vdash\!\!+\!\!+\!\!+\!\!+\!\!+ \text{ - - - } \vdash\!\!+\!\!+\!\!+\!\!+\!\!+\!\!+\!\!\to \text{time} \\ i \ * \ * \ * \ \cdots \ * \ * \ i \ * \ * \ \cdots \ * \ i \end{array} \right) = (1-p)^2.$$

If we do return to i a second time, then again we have a probability p of never returning afterwards, and so on. In particular, we can view this as a sequence of Bernoulli trials with probability of success p: every time we get a failure, we return to i once more; but once we reach a success, we never return to i again. By Murphy's law, success will eventually happen with probability one, and so we will eventually never return to the transient state i with probability one. This is precisely the conclusion of claim a. above!

Now suppose i is a *recurrent* state, and let j be another state such that $i \rightsquigarrow j$. We would like to show that we *will* eventually reach j. To do this, let us consider a modified Markov chain which is identical to the original Markov chain, except that we have deleted all outgoing arrows from j. For example:



In words, we have modified the state j to be *absorbing*: in the modified chain, once we reach the state j we will stay there forever, while in the original chain we can move out of the state j again. On the other hand, we have only modified what happens to the chain *after* it has reached j. Before reaching j, both chains are exactly the same. Therefore, the probability that we will ever reach j is the same in both chains. How does this help? Notice that in the modified chain (in the above example), the states i and k have become transient, and the only remaining recurrent state is j! Therefore, it follows from claim a. that we *will* reach state j with probability one starting from the state i. This is precisely the conclusion of claim b. above!

We have now shown that if i is recurrent and $i \rightsquigarrow j$, then we will eventually reach state j. However, as i is recurrent and $i \rightsquigarrow j$, we must also have $j \rightsquigarrow i$; therefore, once we made it to state j, we will also eventually return to state i. Once we are back in state i, we will eventually return to state j, and so on. This shows that we will visit the states i and j (and all other states in the recurrent class) infinitely often, as claimed in b. above.

Remark 9.2.8. In the general theory of Markov chains, the properties in a. and b. above should actually be used as the *definition* of recurrent and transient states: a state i is recurrent if it is visited infinitely often with probability one, and is transient if it is visited finitely often with probability one. Fortunately, we have shown that these properties are equivalent to Definitions 9.2.2 and 9.2.3 for the Markov chains that we consider in this chapter. This is a very useful fact, because the properties of Definitions 9.2.2 and 9.2.3 can be easily read off from the state diagram, while the answer to the question whether a state is visited infinitey or finitely often is much less clear at the outset.

However, a word of caution is in order. In this course, we only consider Markov chains that take values in a *finite* state space D. In a more general setting, you might encounter Markov chains that take values in a countable state space, say $D = \mathbb{Z}^d$. In this case, it is *not* necessarily true that Definitions 9.2.2 and 9.2.3 are equivalent to recurrence or transience. For example, when

D is infinite, it can happen that $i \rightsquigarrow j$ for every $i, j \in D$, yet every state i is visited only finitely often (and thus every state is transient). This is because when D is infinite, space is "so big" that the Markov chain can get lost and never make it home, even if there is always a path back home.

When the Markov chain is finite, such issues can never arise. The definitions in this section are all correct in the finite case, and you should not worry about such problems for the purposes of this course. However, it is good to be aware (in case you every encounter such problems outside the context of this course) that these matters are not always obvious!

9.3 First step analysis

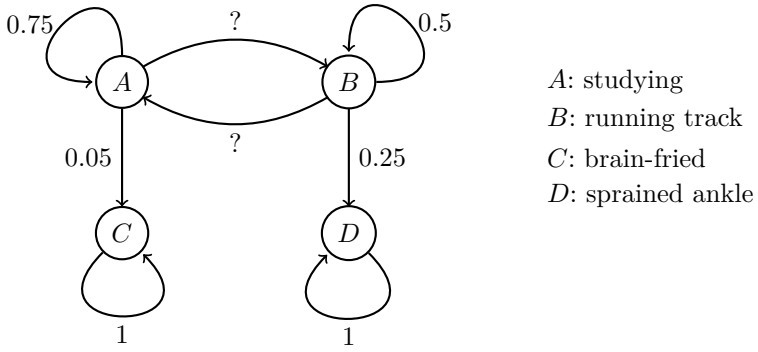
In the previous section, we saw how we can *qualitatively* predict the long-term behavior of a Markov chain by decomposing its states into recurrent classes and transient states: on the long run, the Markov chain always ends up in one of its recurrent classes, and once there it bounces around inside this class forever (visiting every state in the class infinitely often).

Such qualitative understanding is an essential first step in analyzing the behavior of a Markov chain. Once we understand it, however, we would also like to be able to answer *quantitative* questions. For example, suppose a Markov chain has two recurrent classes and some transient states. Assuming that we start in a given transient state, what is the probability that we end up in each recurrent class, and how long does it take before one of the recurrent classes is reached? As in the case of random walks and branching processes, the tool of choice to answer such questions is *first step analysis*. Once you understand how to use this powerful tool, you can apply it to all sorts of different problems. In this section, we will illustrate how to use first step analysis in the Markov chain setting in the context of a simple example.

Example 9.3.1 (The student-athlete). A student-athlete alternates between studying for her ORF 309 midterm and running track. Every hour, she decides what activity to engage in for the next hour. After an hour of study, she will study for another hour 75% of the time. After an hour of running track, she will run track for another hour 50% of the time.

Unfortunately, both running track and studying are risky activities. When you run track for an hour, the probability you will sprain your ankle is as high as 25%. On the other hand, studying for an hour will result in being brain-fried (a medical condition that prohibits further activity) with probability 5%. Once you sprain your ankle or end up brain-fried, you have to cease all other activity so you can recover: in particular, you can neither study nor run anymore. What fate will befall our student-athlete on the long run?

Our first aim should be to build a Markov chain model for the state of our student-athlete. From the data given in the problem, we immediately infer that the state of our student-athlete should follow the following state diagram:



The transition probabilities marked with "?" are not given explicitly in the problem statement. However, we know what they are anyway: for any Markov chain, the the probabilities on the outgoing arrows of any state must always sum to one (as *something* must happen in the next time step). We can therefore immediately fill in $P_{AB} = 0.2$ and $P_{BA} = 0.25$ in the above diagram.

Let us begin with the familiar qualitative analysis from the previous section. You can easily see in the above state diagram that the states A and B are transient, while the recurrent classes are $\{C\}$ and $\{D\}$. Therefore, no matter what state our student-athlete starts off in, she will always end up eventually wither the a sprained ankle or a fried brain. In order to plan ahead (e.g. to know how much aspirin to stock up on), she would like to know a bit more:

a. What is the probability she will end up eventually with a sprained ankle, as opposed to ending up brain-fried?

b. How long will it take, on average, before she is out of commission?

We will presently answer both these questions.

Let $\{X_n\}$ be the Markov chain with the above state diagram (that is, X_n is the state of our student-athlete after n hours), and define

$$T = \min\{n : X_n = C \text{ or } X_n = D\}.$$

That is, T is the time at which our student-athlete is first out of commission. Let us begin by computing the probability

$$s(i) := \mathbf{P}_i\{X_T = D\},$$

that she eventually ends up with a sprained ankle, given that she started in the state i. It is easy to see that $s(C) = 0$ and $s(D) = 1$: if you start off being brain-fried then you cannot sprain your ankle, and if you start off with a sprained ankle then you obviously end up with a sprained ankle. Thus it remains to compute the interesting quantities $s(A)$ and $s(B)$.

To this end, let us use the first step analysis argument. Suppose we start in the state A. Then we can hit C or D at the earliest after one step. After one step, we are in A with probability 0.75, in B with probability 0.2, and in C with probability 0.05, and the Markov chain now simply goes on from this new state. In particular, if we end up in A, then the probability of eventually ending up in D is $s(A)$; if we end up in B, then the probability of eventually ending up in D is $s(B)$; and if we end up in C, then the probability of eventually ending up in D is $s(C)$. We must therefore have the equation

$$s(A) = 0.75\, s(A) + 0.20\, s(B) + 0.05\, s(C),$$

and a similar equation holds for $s(B)$. To give a slightly more rigorous argument, note that for $i = A$ or B

$$\begin{aligned} s(i) &= \sum_{j\in\{A,B,C,D\}} \mathbf{P}_i\{X_T = D | X_1 = j\}\, \mathbf{P}_i\{X_1 = j\} \\ &= \sum_{j\in\{A,B,C,D\}} P_{ij}\, \mathbf{P}_j\{X_T = D\} \\ &= \sum_{j\in\{A,B,C,D\}} P_{ij} s(j). \end{aligned}$$

In the present case, this gives the system of equations

$$\begin{aligned} s(A) &= 0.75\, s(A) + 0.20\, s(B) + 0.05\, s(C), \\ s(B) &= 0.25\, s(A) + 0.50\, s(B) + 0.25\, s(D), \\ s(C) &= 0, \\ s(D) &= 1. \end{aligned}$$

Solving this system of linear equations is a matter of high-school algebra. You should check as an exercise that we obtain

$$s(A) = \frac{4}{6}, \qquad s(B) = \frac{5}{6}.$$

In particular, note that no matter whether she starts off studying or running, she is more likely to end up with a sports injury than being brain-fried.

Let us now compute the expected time until she is out of commission

$$\mu(i) := \mathbf{E}_i(T).$$

In this case, clearly $\mu(C) = \mu(D) = 0$: if you start off being out of commission, it takes zero time for this to happen. On the other hand, suppose you start off in the state A, say. If you go to the state A in the next time step, then the expected *remaining* time until you are out of commission is $\mu(A)$, so the expected *total* time will be $\mu(A) + 1$. The same reasoning holds for the remaining states. Our first-step analysis therefore shows that for $i = A$ or B

$$\mu(i) = \sum_{j \in \{A,B,C,D\}} P_{ij}(\mu(j) + 1) = 1 + \sum_{j \in \{A,B,C,D\}} P_{ij}\mu(j).$$

Substituting the transition probabilities of our chain gives

$$\begin{aligned}
\mu(A) &= 0.75\,\mu(A) + 0.20\,\mu(B) + 0.05\,\mu(C) + 1,\\
\mu(B) &= 0.25\,\mu(A) + 0.50\,\mu(B) + 0.25\,\mu(D) + 1,\\
\mu(C) &= 0,\\
\mu(D) &= 0.
\end{aligned}$$

Once again, solving this system of equations is an easy exercise, and we find

$$\mu(A) = \frac{28}{3} \approx 9.33, \qquad \mu(B) = \frac{20}{3} \approx 6.67.$$

Thus if you start off initially by studying, it will take much longer before you are out of commission than if you start off running track!

While we have only treated one example for sake of illustration, first step analysis can be used to solve many other problems involving Markov chains. Once you understand how first step analysis works, you will be able to apply this tool systematically to address interesting questions.

Remark 9.3.2. In this section, we have seen that we can easily solve quantitative problems by solving a linear system of equations. It is important to note, however, that we were able to formulate the right questions to ask precisely because we already understand the qualitative properties of our Markov chain! For example, suppose we had instead considered the first time $T_C = \min\{n : X_n = C\}$ that the Markov chain hits C. Due to the classification of states, we know that there is positive probability that the Markov chain never hits C, that is, $\mathbf{P}\{T_C = \infty\} > 0$. This certainly implies $\mathbf{E}(T_C) = \infty$, and thus there is no point in trying to compute $\mathbf{E}(T_C)$ using first step analysis. In general, both our qualitative and quantitative tools are extremely useful. First you should use the classification of states to understand qualitatively what is going on. Once you understand this, you are in a position formulate the right quantitative questions that can be answered by first step analysis.

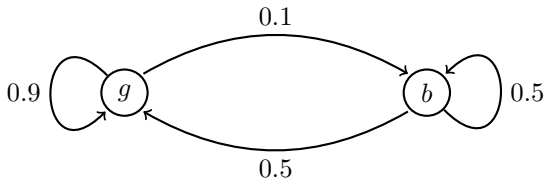
9.4 Steady-state behavior

Recall that the classification of states allows us to predict the qualitative behvior of a Markov chain in the long run: the Markov chain always ends up eventually in one of its recurrent classes, and once there it bounces around inside this class forever (visiting every state in the class infinitely often). In the previous section, we studied quantitatively what happens to the Markov chain before it reaches a recurrent class: what is the probability of ending up in each recurrent class, and how long does this take? On the other hand,

we will investigate in this section the long-term behavior of Markov chains *inside* a recurrent class. This question will lead us to investigate some very interesting and useful properties of Markov chains.

They sort of phenomena we will investigate in this section are best illustrated by means of a simple numerical example.

Example 9.4.1 (Mood swings). Recall the good/bad mood Markov chain of Example 9.1.4, which we provided here with some specific transition probabilities:



Suppose I am in a good mood today. What is the probability that I will be in a good mood a month from now, or a year from now? Similarly, if I am in a bad mood today, how do these numbers change?

The probability of being in a good mood after n days given that we start in a good or bad mood, respectively, is given by

$$\mathbf{P}_g\{X_n = g\} = (P^n)_{gg}, \qquad \mathbf{P}_b\{X_n = g\} = (P^n)_{bg}.$$

Computing these numbers is a matter of matrix multiplication, a straightforward but tedious business. Fortunately, this is precisely what computers were invented for. According to my computer, these probabilities for the first few days are as follows (rounded to three decimal digits):

n	$(P^n)_{gg}$	$(P^n)_{bg}$
0	1	0
1	0.9	0.5
2	0.86	0.7
3	0.844	0.780
4	0.838	0.812
5	0.835	0.825
6	0.834	0.830
7	0.833	0.832
8	0.833	0.833
9	0.833	0.833

We observe a surprising fact: no matter what is our initial mood, in the long run our probability of being in a good mood is always ≈ 0.833! That is, the Markov chain "forgets" its initial state and converges to a steady state.

The phenomenon observed in the previous example is extremely useful. It shows that to predict the behavior of the Markov chain on the long run, we do

not need to know the initial condition but only its behavior in steady state. A Markov chain with this property is said to be *ergodic*.

Definition 9.4.2. *A Markov chain* $\{X_n\}$ *is said to be* ergodic *if the limit*

$$\pi(j) := \lim_{n\to\infty} \mathbf{P}_i\{X_n = j\}$$

exists for every state j *and does not depend on the initial state* i*. Then* $\pi = (\pi(j))_{j\in D}$ *is called the* stationary probability *of the Markov chain.*

The above numerical computation strongly suggests that the good/bad mood Markov chain is ergodic. However, just because this phenomenon appears in this particular example does not mean this is always true! When are Markov chains ergodic, and how can we compute their steady-state probabilities? These questions will be answered in the remainder of this section.

The second question turns out to be easier, so let us start there. Suppose the Markov chain $\{X_n\}$ is ergodic. How can we compute $\pi(j)$? Let us compute $\pi(j)$ in two different ways. First, notice that by definition

$$\pi(j) = \lim_{n\to\infty} \mathbf{P}_i\{X_n = j\} = \lim_{n\to\infty} (P^n)_{ij}.$$

On the other hand, using $P^n = P^{n-1}P$, we can write

$$\pi(j) = \lim_{n\to\infty} \sum_k (P^{n-1})_{ik} P_{kj} = \sum_k \pi(k) P_{kj}.$$

If we view $\pi = (\pi(j))_{j\in D}$ as a row vector, we can rewrite this equation in matrix-vector form: the stationary probabilities must satisfy the equation

$$\pi = \pi P.$$

Together with the fact that $\pi(j)$ are probabilities (so that $\sum_j \pi(j) = 1$), this equation is all you need to compute stationary probabilities.

Example 9.4.3 (Mood swings revisited). Let us compute explicitly the stationary probabilities in Example 9.4.1. The equation $\pi = \pi P$ becomes

$$\begin{aligned} \pi(g) &= 0.9\,\pi(g) + 0.5\,\pi(b), \\ \pi(b) &= 0.1\,\pi(g) + 0.5\,\pi(b). \end{aligned}$$

These appear to be two equations with two unknowns. However, this is not quite true: if you look closely, you will notice that these two equations are

actually the *same* equation rearranged in two different ways! To solve the problem, we must also use that π is a probability vector:

$$\pi(g) + \pi(b) = 1.$$

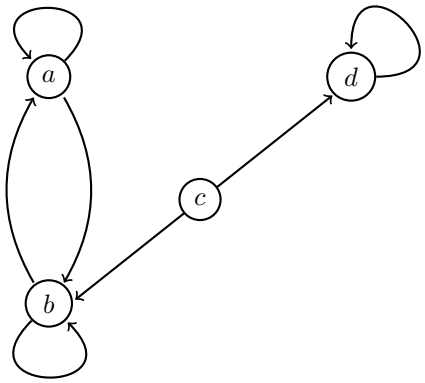
A simple algebra exercise shows that

$$\pi(g) = \frac{5}{6} \approx 0.833.$$

Thus we immediately see where the mysterious number 0.833 comes from!

The above example shows that the stationary probabilities $\pi(j)$ of an ergodic Markov chain are very easy to compute (much easier, in most cases, than the probabilities $\mathbf{P}_i\{X_n = j\}$ at any given time n!) But how do we know that the Markov chain is ergodic in the first place? Perhaps you might be tempted to guess that all Markov chains are ergodic, but this is definitely not true. Let us let us try to identify some things that can go wrong.

Example 9.4.4 (Multiple recurrent classes). Consider the following chain:



This chain has two recurrent classes, $\{a, b\}$ and $\{d\}$. Suppose we start the chain in a. Then the chain can never reach d, so

$$\lim_{n\to\infty} \mathbf{P}_a\{X_n = d\} = 0.$$

On the other hand, if we start in d, we always stay there. Therefore

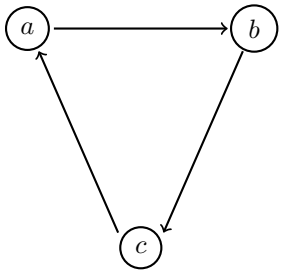
$$\lim_{n\to\infty} \mathbf{P}_d\{X_n = d\} = 1.$$

Thus the chain is not ergodic: the limiting probabilities depend on the initial state! There is nothing special about this example; the same argument shows:

A Markov chain with more than one recurrent class cannot be ergodic.

Your next best hope might be that a Markov chain with only one recurrent class is always ergodic. Unfortunately, this is not true either: there is an entirely different reason why the Markov chain can fail to be ergodic.

Example 9.4.5 (Walking in circles). Consider the following Markov chain:



This entire chain consists of one recurrent class $\{a, b, c\}$, so we cannot have the same problem as in the previous example. Nonetheless, this Markov chain is not ergodic. To see why, note that if we start at a then the chain *must* go to b at the next time, and if we start at b then the chain *must* go to c at the next time, etc. Therefore, the path of the Markov chain must look like this

$$abcabcabcabcabcabcabcabcabcabcabcabc\cdots$$

if we start at a, or like this

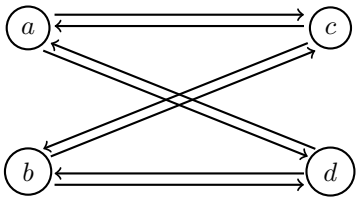
$$bcabcabcabcabcabcabcabcabcabcabcabca\cdots$$

if we start at b, etc. In particular, this implies that we have, for example,

$$\begin{aligned}
\mathbf{P}_a\{X_0 = a\} &= 1 & \mathbf{P}_b\{X_0 = a\} &= 0\\
\mathbf{P}_a\{X_1 = a\} &= 0 & \mathbf{P}_b\{X_1 = a\} &= 0\\
\mathbf{P}_a\{X_2 = a\} &= 0 & \mathbf{P}_b\{X_2 = a\} &= 1\\
\mathbf{P}_a\{X_3 = a\} &= 1 & \mathbf{P}_b\{X_3 = a\} &= 0\\
\mathbf{P}_a\{X_4 = a\} &= 0 & \mathbf{P}_b\{X_4 = a\} &= 0\\
\mathbf{P}_a\{X_5 = a\} &= 0 & \mathbf{P}_b\{X_5 = a\} &= 1\\
\mathbf{P}_a\{X_6 = a\} &= 1 & \mathbf{P}_b\{X_6 = a\} &= 0\\
\mathbf{P}_a\{X_7 = a\} &= 0 & \mathbf{P}_b\{X_7 = a\} &= 0\\
\mathbf{P}_a\{X_8 = a\} &= 0 & \mathbf{P}_b\{X_8 = a\} &= 1\\
&\vdots & &\vdots
\end{aligned}$$

In particular, notice that the Markov chains started at two different points a and b are forever out of sync: they never end up in the same place. Thus the Markov chain cannot be ergodic—it never forgets its initial state!

The above example is a bit special because the path of the Markov chain is actually nonrandom—once we specify its initial condition, we know exactly where it is going to be in the future. This is not the important feature of this example, however. Consider, for example, the following variant:



This chain is genuinely random: if we start in a, then we can go to both c or d (we cannot predict in advance which one); if we are in c, we can go to both a and b; etc. Moreover, the entire chain still consists of one recurrent class $\{a, b, c, b\}$. Nonetheless, we see the same behavior as in the previous example: if we start in $\{a, b\}$, then we *must* be in $\{c, d\}$ in the next time step; and if we start in $\{c, d\}$, then we *must* be in $\{a, b\}$ in the next time step. Therefore, for example, $\mathbf{P}_a\{X_n \in \{a, b\}\} = 1$ and $\mathbf{P}_c\{X_n \in \{a, b\}\} = 0$ whenever n is even, so the chain does not forget its initial state and thus cannot be ergodic.

What is evidently important in the above example is that the chain is *periodic*: it must alternate in consecutive time steps between the subsets $\{a, b\}$ and $\{c, d\}$. These are called the *periodic classes* of the chain.

Definition 9.4.6. *A recurrent class of a Markov chain is called* periodic *if it can be partitioned into $k \geq 2$* periodic classes *$A_1, \ldots, A_k$ such that $\mathbf{P}\{X_{n+1} \in A_r | X_n = i\} = 1$ whenever $i \in A_{r-1}$ (where we set $A_0 := A_k$). If a recurrent class cannot be partitioned in this way, it is called* aperiodic.

It is easy to identify periodicity if you draw the state diagram of a Markov chain! Whenever the Markov chain is periodic, precisely the same argument as we used in the examples above shows:

A periodic Markov chain cannot be ergodic.

We have now identified two different reasons why a Markov chain can fail to be ergodic: the presence of more than one recurrent class and periodicity. Having found this many problems, you might be inclined to expect that there is yet another way in which things can go wrong. Remarkably, this turns out not to be the case: we have identified the *only* possible obstructions to ergodicity!

Theorem 9.4.7 (Ergodic theorem). *A Markov chain is ergodic if and only if it has only one recurrent class and is aperiodic.*

An immediate consequence of Theorem 9.4.7 is that one can determine whether or not a Markov chain is ergodic just by looking at its state diagram! This is an extremely useful fact: once we have established ergodicity of a given Markov chain, we can compute its limiting probabilities by solving a linear equation as was shown at the beginning of this section.

In the rest of this section, we will try to sketch the proof of Theorem 9.4.7. The proof is somewhat tricky (and requires at the very end a simple result from number theory that we will skip for the purposes of this course), but contains a very intuitive probabilistic idea that is worth seeing once.

In order to show that the Markov chain is ergodic, we would like to argue that it forgets its initial state: that is, we would like to show that

$$\lim_{n\to\infty} |\mathbf{P}_i\{X_n = k\} - \mathbf{P}_j\{X_n = k\}| = 0$$

for all initial states i, j. To do this, we will use a romantic idea called *coupling*.

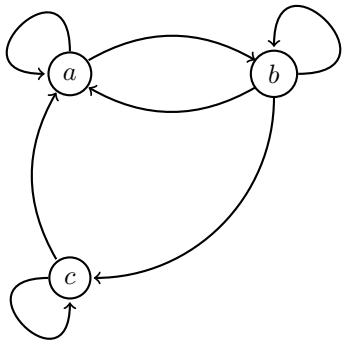
Recall that we can think of the state of a Markov chain as the location a frog who is jumping around on the state diagram. The idea behind coupling is to consider not one, but *two* frogs. A boy frog will be started at the point i, and a girl frog will be started at the point j. They now move as follows:

a. Initially, the boy and girl frog each jump to the next state independently according to the transition probabilities of the Markov chain. They keep doing this until they happen to land on the same state.

b. When the two frogs land on the same state for the first time, they fall madly in love. From this point onwards, they go everywhere together. That is, at each time, the next state is selected according to the transition probabilities of the Markov chain, and both frogs jump to the next state together.

The time at which the frogs first meet is called the *coupling time*.

For sake of illustration, let us consider the following simple chain:



Here is one possible trajectory of the boy frog $\{B_n\}$ and the girl frog $\{G_n\}$ for the above Markov chain, given that we start the boy frog at state a and the girl frog at state b (here the coupling time happens to be 6):

n	0	1	2	3	4	5	6	7	8	9	10	11
B_n	a	a	b	b	a	b	a	a	b	c	c	a
G_n	b	c	c	c	c	a	a	a	b	c	c	a

What is the point of this construction? Regardless of whether we are before or after the coupling time, each frog jumps to the next state with the same transition probabilities as those of the original Markov chain (the only difference is whether they jump together or separately). Therefore, if we forget about the girl frog, the boy frog simply moves according to the original Markov chain $\{X_n\}$ started at the point a. Similarly, if we forget about the boy frog, the girl frog moves according to the original chain started at the point b. That is,

$$\mathbf{P}_a\{X_n = k\} = \mathbf{P}\{B_n = k\}, \qquad \mathbf{P}_b\{X_n = k\} = \mathbf{P}\{G_n = k\}.$$

It follows that

$$\begin{aligned} |\mathbf{P}_a\{X_n = k\} - \mathbf{P}_b\{X_n = k\}| &= |\mathbf{E}(\mathbf{1}_{B_n=k} - \mathbf{1}_{G_n=k})| \\ &\le \mathbf{E}(\mathbf{1}_{B_n \neq G_n}) = \mathbf{P}\{B_n \neq G_n\}, \end{aligned}$$

because $|\mathbf{1}_{B_n=k} - \mathbf{1}_{G_n=k}| = \mathbf{1}_{B_n \neq G_n}$ (why?) This mathematical statement is very intuitive: the difference between the probabilities of the Markov chain started at two initial states is at most the probability that the boy frog and girl frog that started at those states have not yet met! Therefore, all we have to show is that the boy and girl frog *will* eventually meet, as in that case $\mathbf{P}\{B_n \neq G_n\} \to 0$ as $n \to \infty$ and thus the Markov chain is ergodic.

To show that the boy frog and girl do eventually meet, let us first notice that the joint distribution of (B_n, G_n) depends only on (B_{n-1}, G_{n-1}): if we know where the boy and girl frogs are in the previous time step, we know precisely with what probabilities they will jump to each new pair of states. Therefore, the pair (B_n, G_n) is itself a Markov chain in the state space $\{(i,j) : i,j \in \{a,b,c\}\}$. In order to show that the boy frog and girl frog eventually meet, it is enough to show that every state (i,j) in which the frogs are not together $i \neq j$ is a *transient* state of the Markov chain (B_n, G_n). But we know that once the boy frog and girl frog are coupled, they stay together forever: that is, $(k,k) \not\rightsquigarrow (i,j)$ when $i \neq j$. To show that (i,j) is transient, it is therefore enough to show that $(i,j) \rightsquigarrow (k,k)$ for some k.

Let k be any recurrent state of the original Markov chain. As there is only one recurrent class, we know that $i \rightsquigarrow k$ and $j \rightsquigarrow k$. If nonetheless $(i,j) \rightsquigarrow (k,k)$ were to fail, we would be in a very strange situation: even though both the boy frog and girl frog can (and will) end up in the state k infinitely often, they could never end up in this state at the same time! The only way in which this can happen is if the boy and girl frog are always out of

sync: that is, if the chain is periodic.[1] Therefore, if the original Markov chain has only one recurrent class and is aperiodic, then the boy and girl frogs will eventually meet and thus the Markov chain must be ergodic.

9.5 The law of large numbers revisited

Let $\{X_n\}$ be a Markov chain with values in the finite state space D. Suppose that each time that we visit the state i, we receive a reward of $f(i)$ dollars: that is, the reward received at time k is $f(X_k)$ dollars. Therefore, the average reward per unit time after n time steps is given by

$$\frac{f(X_1) + \ldots + f(X_n)}{n} = \frac{1}{n} \sum_{k=1}^{n} f(X_k).$$

In particular, the average daily reward over *all* time is

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k).$$

If the random variables $\{X_k\}$ were independent, then we could predict the long-term average reward in advance: the law of large numbers states that the above limit would then be equal to the expected reward $\mathbf{E}(f(X_k))$ (in particular, the long-time average reward is not random!) In the Markov chain setting, however, the random variables $\{X_k\}$ are generally very much dependent. Nonetheless, it turns out that there is an extremely useful extension of the law of large numbers for *ergodic* Markov chains.

Theorem 9.5.1 (Ergodic theorem II). *Let $\{X_n\}$ be an ergodic Markov chain with stationary probability $\pi = (\pi(j))_{j \in D}$. Then*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{j \in D} f(j)\, \pi(j).$$

In words, this ergodic theorem states that the average reward *over time* equals the average reward *over the state space* if we weight every point $i \in D$ according to its stationary probability $\pi(i)$. The equivalence of time-averages and space-averages is a very common idea in the study of dynamical systems in physics and in engineering: it is also known as the *ergodic principle*.

[1] To make this point completely rigorous, one needs a simple result from number theory that is beyond the level of this course. However, the idea is completely intuitive if you look at the examples of periodic classes.

Let us investigate an important special case of this result. Let $N_n(i)$ be the number of times the Markov chain visits the state i by time n, that is,

$$N_n(i) := \#\{1 \le k \le n : X_k = i\} = \sum_{k=1}^{n} \mathbf{1}_{X_k=i}.$$

Then

$$\frac{N_n(i)}{n} = \text{fraction of time spent at state } i \text{ by time } n.$$

Thus the fraction of *all* time that the Markov chain spends at the state i is

$$\lim_{n\to\infty} \frac{N_n(i)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_k=i} = \pi(i).$$

This gives us a new interpretation of the meaning of stationary probabilities.

> *The stationary probability $\pi(i)$ of an ergodic Markov chain is the fraction of time the Markov chain spends in the state i.*

This very intuitive idea can be very interesting in practice.

Example 9.5.2 (Mood swings). Consider the mood swings Markov chain of Example 9.4.1. We already computed its stationary probabilities as

$$\pi(\text{good mood}) = \frac{5}{6}, \qquad \pi(\text{bad mood}) = \frac{1}{6}.$$

Therefore, assuming the model is an accurate reflection of reality, you will spend one sixth of the time in a bad mood (the news media might report this as "scientists discover that students are in a bad mood once every six days").

Next, let us consider a more tricky application of the law of large numbers for Markov chains that is very important in practice.

Example 9.5.3 (Estimating transition probabilities). In the real world, when you model a certain phenomenon as a Markov chain, you may not necessarily know its transition probabilities in advance. For example, suppose you are trying to model a frog jumping between a number of different lily pads in a pond. It is natural to expect that the location X_n of the frog after n minutes is a Markov chain, as frogs are opportunistic creatures who decide where to jump next only on the basis of their current location. However, the psychology of the frog being a closed book to you (at least at the start of the investigation), you do not necessarily know with what probability the frog will jump to different locations. Therefore, we need a way of "measuring" the transition probabilities of the frog from data: that is, given a long sequence $X_1, X_2, X_3, \ldots$ of observations of the frog's location, we would like to infer the values of the transition probabilities. How should we go about doing that?

Here is an intuitive idea. The transition probability P_{ij} is the probability that we will jump to the state j if we are currently in the state i. Therefore, as probabilities are naturally interpreted as frequencies, you might estimate P_{ij} as follows: first, collect only those times at which you visit the state i; and then compute the fraction of those times in which you jumped to the state j in the next time step. That is, you might estimate P_{ij} by

$$\frac{N_n(i,j)}{N_n(i)} = \text{fraction of visits to } i \text{ that resulted in a jump to } j \text{ by time } n,$$

where $N_n(i,j)$ is the number of jumps from i to j that occured by time n:

$$N_n(i,j) := \#\{1 \le k \le n : X_k = i, X_{k+1} = j\} = \sum_{k=1}^{n} \mathbf{1}_{X_k=i, X_{k+1}=j}.$$

Using the ergodic theorem, we will now show that this intuitive procedure is indeed a correct way to estimate the transition probabilities of an ergodic Markov chain from observed data: that is, we will show that

$$\lim_{n\to\infty} \frac{N_n(i,j)}{N_n(i)} = P_{ij}.$$

In particular, this shows that our intuition is correct:

> *The transition probability P_{ij} of an ergodic Markov chain is the fraction of visits to i that result in a jump to j.*

This simple idea is used in many real-world applications of Markov chains.

Let us justify the above claim. By the ergodic theorem, we have

$$\lim_{n\to\infty} \frac{N_n(i,j)}{N_n(i)} = \lim_{n\to\infty} \frac{\frac{N_n(i,j)}{n}}{\frac{N_n(i)}{n}} = \frac{1}{\pi(i)} \lim_{n\to\infty} \frac{N_n(i,j)}{n}.$$

Thus the claim would be proved if we can show that

$$\lim_{n\to\infty} \frac{N_n(i,j)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_k=i, X_{k+1}=j} = \pi(i) P_{ij}.$$

But why should this be true? We would really like to use the ergodic theorem to compute this last limit, but the problem is that the function $\mathbf{1}_{X_k=i,X_{k+1}=j}$ is not of the form $f(X_k)$: it depends on the state of the Markov chain at *two* consecutive times, while the ergodic theorem allows us to compute the average of a function of the state of the Markov chain at a *single* time.

To solve this problem, let us define the random variable

$$\tilde{X}_n := (X_n, X_{n+1})$$

with values in the set $\{(i,j) : i,j \in D\}$. It is not hard to see that $\{\tilde{X}_n\}$ is again an ergodic Markov chain (why?). If we denote its stationary probability by $\tilde{\pi}$, then we obtain by the ergodic theorem

$$\lim_{n\to\infty} \frac{N_n(i,j)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\tilde{X}_k=(i,j)} = \tilde{\pi}(i,j).$$

It therefore remains to show that

$$\tilde{\pi}(i,j) = \pi(i)P_{ij}.$$

To compute $\tilde{\pi}$, note that the transition probability matrix $\tilde{P}$ of $\{\tilde{X}_n\}$ is

$$\tilde{P}_{(i,j),(k,l)} = \mathbf{P}\{X_{n+2} = l, X_{n+1} = k | X_{n+1} = j, X_n = i\} = \mathbf{1}_{k=j} P_{jl}.$$

To check that $\tilde{\pi}$ is stationary, we simply note that $\tilde{\pi}\tilde{P} = \tilde{\pi}$:

$$\begin{aligned}
(\tilde{\pi}\tilde{P})(k,l) &= \sum_{i,j} \tilde{\pi}(i,j)\tilde{P}_{(i,j),(k,l)} \\
&= \sum_{i,j} \pi(i)P_{ij}\mathbf{1}_{k=j}P_{jl} \\
&= \sum_{i} \pi(i)P_{ik}P_{kl} \\
&= \pi(k)P_{kl} = \tilde{\pi}(k,l),
\end{aligned}$$

where we have used $\pi P = \pi$. This justifies our original claim

$$\lim_{n\to\infty} \frac{N_n(i,j)}{N_n(i)} = \lim_{n\to\infty} \frac{\frac{N_n(i,j)}{n}}{\frac{N_n(i)}{n}} = \frac{\tilde{\pi}(i,j)}{\pi(i)} = P_{ij}.$$

In the rest of this section, we will try to sketch the proof of Theorem 9.5.1. As we did when we proved the law of large numbers, we first note that

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) = \sum_{i\in D} f(i)\,\frac{N_n(i)}{n}.$$

Therefore, it is enough to prove the special case

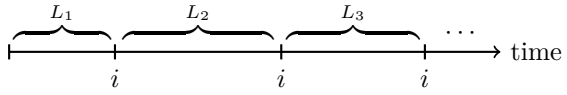
$$\lim_{n\to\infty} \frac{N_n(i)}{n} = \pi(i).$$

It is easy to see that this result holds *on average* when the Markov chain is ergodic: indeed, if we take the expectation, we find that

$$\lim_{n\to\infty} \mathbf{E}\left(\frac{N_n(i)}{n}\right) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{P}\{X_k = i\} = \pi(i),$$

because $\mathbf{P}\{X_k = i\} \to \pi(i)$ as $k \to \infty$. The claim of the ergodic theorem is however not just that $\mathbf{E}(N_n(i)/n) \to \pi(i)$, but that in fact $N_n(i)/n \to \pi(i)$: that is, we must still show that the limit fraction of time that we visit i is equal to its expectation or, in other words, that the limit is nonrandom!

To see why this is true, we are going to compute the limiting frequency of visits to i in a different way. Let L_1 be the amount of time we must wait until we first visit i; then L_2 is the amount of time we must subsequently wait until our next visit to i, etc. This is illustrated in the following figure:



At time $L_1 + L_2 + \cdots + L_k$, we have visited i exactly k times. Therefore, the fraction of time that we spend at the state i is given by

$$\text{fraction of time spent at } i = \lim_{n\to\infty} \frac{N_n(i)}{n} = \lim_{k\to\infty} \frac{k}{L_1 + L_2 + \cdots + L_k}.$$

Now notice that by the Markov property, whenever we visit the state i, the subsequent behavior of the Markov chain after that time depends only on i (and in particular is independent of any prior history of the process!) Therefore, the random variables $L_1, L_2, L_3, \ldots$ are independent, and $L_2, L_3, L_4, \ldots$ are in fact i.i.d.: their distribution is that of the time that the Markov chain that starts at the state i first returns to i. (L_1 might have a different distribution unless we assume that the Markov chain is also started at i; but this will make no difference when we average over all time.) Therefore,

$$\lim_{k\to\infty} \frac{L_1 + L_2 + \cdots + L_k}{k} = \mathbf{E}_i(L_1)$$

by the *ordinary* law of large numbers. In particular, we have shown that the limit of $N_n(i)/n$ is nonrandom, and thus the ergodic theorem is established.

Remark 9.5.4. The above argument has in fact established a rather interesting identity. Note that we proved two quite different expressions: we have

$$\lim_{n\to\infty} \frac{N_n(i)}{n} = \pi(i),$$

but on the other hand

$$\lim_{n\to\infty} \frac{N_n(i)}{n} = \lim_{k\to\infty} \frac{k}{L_1 + L_2 + \cdots + L_k} = \frac{1}{\mathbf{E}_i(L_1)}.$$

Thus we have obtained a surprising interpretation of the stationary probability $\pi(i)$: its inverse $1/\pi(i)$ is the expected time that the Markov chain first returns to the state i if it starts in i. This curious result is known as Kac's theorem, after the Polish-American mathematician Mark Kac (1914–1984).[2]

[2] His name is pronounced "kats".