Introduction to Stochastic Processes - Lecture Notes

(with 33 illustrations)

Gordan Žitković
Department of Mathematics
The University of Texas at Austin

Contents

1 Probability review **4**
- 1.1 Random variables . . . 4
- 1.2 Countable sets . . . 5
- 1.3 Discrete random variables . . . 5
- 1.4 Expectation . . . 7
- 1.5 Events and probability . . . 8
- 1.6 Dependence and independence . . . 9
- 1.7 Conditional probability . . . 10
- 1.8 Examples . . . 12

2 Mathematica in 15 min **15**
- 2.1 Basic Syntax . . . 15
- 2.2 Numerical Approximation . . . 16
- 2.3 Expression Manipulation . . . 16
- 2.4 Lists and Functions . . . 17
- 2.5 Linear Algebra . . . 19
- 2.6 Predefined Constants . . . 20
- 2.7 Calculus . . . 20
- 2.8 Solving Equations . . . 22
- 2.9 Graphics . . . 22
- 2.10 Probability Distributions and Simulation . . . 23
- 2.11 Help Commands . . . 24
- 2.12 Common Mistakes . . . 25

3 Stochastic Processes **26**
- 3.1 The canonical probability space . . . 27
- 3.2 Constructing the Random Walk . . . 28
- 3.3 Simulation . . . 29
  - 3.3.1 Random number generation . . . 29
  - 3.3.2 Simulation of Random Variables . . . 30
- 3.4 Monte Carlo Integration . . . 33

4 The Simple Random Walk **35**
- 4.1 Construction . . . 35
- 4.2 The maximum . . . 36

5 Generating functions **40**
5.1 Definition and first properties . . . 40
5.2 Convolution and moments . . . 42
5.3 Random sums and Wald's identity . . . 44

6 Random walks - advanced methods **48**
6.1 Stopping times . . . 48
6.2 Wald's identity II . . . 50
6.3 The distribution of the first hitting time T_1 . . . 52
6.3.1 A recursive formula . . . 52
6.3.2 Generating-function approach . . . 53
6.3.3 Do we actually hit 1 sooner or later? . . . 55
6.3.4 Expected time until we hit 1? . . . 55

7 Branching processes **56**
7.1 A bit of history . . . 56
7.2 A mathematical model . . . 56
7.3 Construction and simulation of branching processes . . . 57
7.4 A generating-function approach . . . 58
7.5 Extinction probability . . . 61

8 Markov Chains **63**
8.1 The Markov property . . . 63
8.2 Examples . . . 64
8.3 Chapman-Kolmogorov relations . . . 70

9 The "Stochastics" package **74**
9.1 Installation . . . 74
9.2 Building Chains . . . 74
9.3 Getting information about a chain . . . 75
9.4 Simulation . . . 76
9.5 Plots . . . 76
9.6 Examples . . . 77

10 Classification of States **79**
10.1 The Communication Relation . . . 79
10.2 Classes . . . 81
10.3 Transience and recurrence . . . 83
10.4 Examples . . . 84

11 More on Transience and recurrence **86**
11.1 A criterion for recurrence . . . 86
11.2 Class properties . . . 88
11.3 A canonical decomposition . . . 89

12 Absorption and reward **92**
12.1 Absorption 92
12.2 Expected reward 95

13 Stationary and Limiting Distributions **98**
13.1 Stationary and limiting distributions 98
13.2 Limiting distributions 104

14 Solved Problems **107**
14.1 Probability review 107
14.2 Random Walks 111
14.3 Generating functions 114
14.4 Random walks - advanced methods 120
14.5 Branching processes 122
14.6 Markov chains - classification of states 133
14.7 Markov chains - absorption and reward 142
14.8 Markov chains - stationary and limiting distributions 148
14.9 Markov chains - various multiple-choice problems 156

Chapter 1

Probability review

The probable is what usually happens.

—Aristotle

It is a truth very certain that when it is not in our power to determine. what is true we ought to follow what is most probable

—Descartes - "Discourse on Method"

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

—Pierre Simon Laplace - "Théorie Analytique des Probabilités, 1812 "

Anyone who considers arithmetic methods of producing random digits is, of course, in a state of sin.

—John von Neumann - quote in "Conic Sections" by D. MacHale

I say unto you: a man must have chaos yet within him to be able to give birth to a dancing star: I say unto you: ye have chaos yet within you ...

—Friedrich Nietzsche - "Thus Spake Zarathustra"

1.1 Random variables

Probability is about random variables. Instead of giving a precise definition, let us just metion that a **random variable** can be thought of as an uncertain, numerical (i.e., with values in $\mathbb{R}$) quantity. While it is true that we do not know with certainty what value a random variable X will take, we usually know how to compute the probability that its value will be in some some subset of $\mathbb{R}$. For example, we might be interested in $\mathbb{P}[X \geq 7]$, $\mathbb{P}[X \in [2, 3.1]]$ or $\mathbb{P}[X \in \{1, 2, 3\}]$. The collection of all such probabilities is called the **distribution** of X. One has to be very careful not to confuse the random variable itself and its distribution. This point is particularly important when several random variables appear at the same time. When two random variables X and Y have the same distribution, i.e., when $\mathbb{P}[X \in A] = \mathbb{P}[Y \in A]$ for any set A, we say that X and Y are **equally distributed** and write $X \overset{(d)}{=} Y$.

1.2 Countable sets

Almost all random variables in this course will take only countably many values, so it is probably a good idea to review breifly what the word *countable* means. As you might know, the countable infinity is one of many different infinities we encounter in mathematics. Simply, a set is countable if it has the same number of elements as the set $\mathbb{N} = \{1, 2, \dots\}$ of natural numbers. More precisely, we say that a set A is **countable** if there exists a function $f : \mathbb{N} \to A$ which is bijective (one-to-one and onto). You can think f as the correspondence that "proves" that there exactly as many elements of A as there are elements of $\mathbb{N}$. Alternatively, you can view f as an *ordering* of A; it arranges A into a particular order $A = \{a_1, a_2, \dots\}$, where $a_1 = f(1)$, $a_2 = f(2)$, etc. Infinities are funny, however, as the following example shows

Example 1.1.

1. $\mathbb{N}$ itself is countable; just use $f(n) = n$.
2. $\mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$ is countable; use $f(n) = n - 1$. You can see here why I think that infinities are funny; the set $\mathbb{N}_0$ and the set $\mathbb{N}$ - which is its proper subset - have the same size.
3. $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, 3, \dots\}$ is countable; now the function f is a bit more complicated;

$$f(k) = \begin{cases} 2k + 1, & k \geq 0 \\ -2k, & k < 0. \end{cases}$$

   You could think that $\mathbb{Z}$ is more than "twice-as-large" as $\mathbb{N}$, but it is not. It is the same size.
4. It gets even weirder. The set $\mathbb{N} \times \mathbb{N} = \{(m, n) : m \in \mathbb{N}, n \in \mathbb{N}\}$ of all pairs of natural numbers is also countable. I leave it to you to construct the function f.
5. A similar argument shows that the set $\mathbb{Q}$ of all rational numbers (fractions) is also countable.
6. The set $[0, 1]$ of all real numbers between 0 and 1 is *not* countable; this fact was first proven by Georg Cantor who used a neat trick called the *diagonal argument*.

1.3 Discrete random variables

A random variable is said to be discrete if it takes at most countably many values. More precisely, X is said to be *discrete* if there exists a *finite* or *countable* set $S \subset \mathbb{R}$ such that $\mathbb{P}[X \in S] = 1$, i.e., if we know with certainty that the only values X can take are those in S. The smallest set S with that property is called the **support** of X. If we want to stress that the support corresponds to the random variable X, we write $_X$.

Some supports appear more often then the others:

1. If X takes only the values $1, 2, 3, \dots$, we say that X is $\mathbb{N}$-**valued**.
2. If we allow 0 (in addition to $\mathbb{N}$), so that $\mathbb{P}[X \in \mathbb{N}_0] = 1$, we say that X is $\mathbb{N}_0$-**valued**

3. Sometimes, it is convenient to allow discrete random variables to take the value $+\infty$. This is mostly the case when we model the waiting time until the first occurence of an event which may or may not ever happen. If it never happens, we will be waiting forever, and the waiting time will be $+\infty$. In those cases - when $S = \{1, 2, 3, \dots, +\infty\} = \mathbb{N} \cup \{+\infty\}$ - we say that the random variable is **extended $\mathbb{N}$-valued**. The same applies to the case of $\mathbb{N}_0$ (instead of $\mathbb{N}$), and we talk about the **extended $\mathbb{N}_0$-valued** random variables. Sometimes the adjective "extended" is left out, and we talk about $\mathbb{N}_0$-valued random variables, even though we allow them to take the value $+\infty$. This sounds more confusing that it actually is.

4. Occasionally, we want our random variables to take values which are not necessarily numbers (think about H and T as the possible outcomes of a coin toss, or the suit of a randomly chosen playing card). Is the collection of all possible values (like $\{H, T\}$ or $\{\heartsuit, \spadesuit, \clubsuit, \diamondsuit\}$) is countable, we still call such random variables discrete. We will see more of that when we start talking about Markov chains.

Discrete random variables are very nice due to the following fact: in order to be able to compute any conceivable probability involving a discrete random variable X, it is enough to know how to compute the probabilities $\mathbb{P}[X = x]$, for all $x \in S$. Indeed, if we are interested in figuring out how much $\mathbb{P}[X \in B]$ is, for some set $B \subseteq \mathbb{R}$ ($B = [3, 6]$, or $B = [-2, \infty)$), we simply pick all $x \in S$ which are also in B and sum their probabilities. In mathematical notation, we have

$$\mathbb{P}[X \in B] = \sum_{x \in S \cap B} \mathbb{P}[X = x].$$

For this reason, the distribution of any discrete random variable X is usually described via a table

$$X \sim \begin{pmatrix} x_1 & x_2 & x_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix},$$

where the top row lists all the elements of S (the support of X) and the bottom row lists their probabilities ($p_i = \mathbb{P}[X = x_i]$, $i \in \mathbb{N}$). When the random variable is $\mathbb{N}$-valued (or $\mathbb{N}_0$-valued), the situation is even simpler because we know what $x_1, x_2, \dots$ are and we identify the distribution of X with the sequence $p_1, p_2, \dots$ (or $p_0, p_1, p_2, \dots$ in the $\mathbb{N}_0$-valued case), which we call the **probability mass function (pmf)** of the random variable X. What about the extended $\mathbb{N}_0$-valued case? It is as simple because we can compute the probability $\mathbb{P}[X = +\infty]$, if we know all the probabilities $p_i = \mathbb{P}[X = i]$, $i \in \mathbb{N}_0$. Indeed, we use the fact that

$$\mathbb{P}[X = 0] + \mathbb{P}[X = 1] + \dots + \mathbb{P}[X = \infty] = 1,$$

so that $\mathbb{P}[X = \infty] = 1 - \sum_{i=1}^{\infty} p_i$, where $p_i = \mathbb{P}[X = i]$. In other words, if you are given a probability mass function $(p_0, p_1, \dots)$, you simply need to compute the sum $\sum_{i=1}^{\infty} p_i$. If it happens to be equal to 1, you can safely conclude that X never takes the value $+\infty$. Otherwise, the probability of $+\infty$ is positive.

The random variables for which $S = \{0, 1\}$ are especially useful. They are called **indicators**. The name comes from the fact that you should think of such variables as signal lights; if $X = 1$ an event of interest has happened, and if $X = 0$ it has not happened. In other words, X *indicates* the occurence of an event. The notation we use is quite suggestive; for example, if Y is the outcome of a coin-toss, and we want to know whether *Heads* (H) occurred, we write

$$X = \mathbf{1}_{\{Y=H\}}.$$

Example 1.2. Suppose that two dice are thrown so that Y_1 and Y_2 are the numbers obtained (both Y_1 and Y_2 are discrete random variables with $S = \{1, 2, 3, 4, 5, 6\}$). If we are interested in the probability the their sum is at least 9, we proceed as follows. We define the random variable Z - the sum of Y_1 and Y_2 - by $Z = Y_1 + Y_2$. Another random variable, let us call it X, is defined by $X = \mathbf{1}_{\{Z \geq 9\}}$, i.e.,

$$X = \begin{cases} 1, & Z \geq 9, \\ 0, & Z < 9. \end{cases}$$

With such a set-up, X signals whether the event of interest has happened, and we can state our original problem in terms of X : "Compute $\mathbb{P}[X = 1]$!". Can you compute it?

1.4 Expectation

For a discrete random variable X with support , we define the **expectation** $\mathbb{E}[X]$ of X by

$$\mathbb{E}[X] = \sum_{x \in} x\mathbb{P}[X = x],$$

as long as the (possibly) infinite sum $\sum_{x \in} x\mathbb{P}[X = x]$ *absolutely converges*. When the sum does not converge, or if it converges only conditionally, we say that the expectation of X **is not defined**. When the random variable in question is $\mathbb{N}_0$-valued, the expression above simplifies to

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i \times p_i,$$

where $p_i = \mathbb{P}[X = i]$, for $i \in \mathbb{N}_0$. Unlike in the general case, the absolute convergence of the defining series can fail in essentially one way, i.e., when

$$\lim_{n \to \infty} \sum_{i=0}^{n} ip_i = +\infty.$$

In that case, the expectation does not formally exist. We still write $\mathbb{E}[X] = +\infty$, but really mean that the defining sum diverges towards infinity.

Once we know what the expectation is, we can easily define several more common terms:

Definition 1.3. Let X be a discrete random variable.

- If the expectation $\mathbb{E}[X]$ exists, we say that X is **integrable**.
- If $\mathbb{E}[X^2] < \infty$ (i.e., if X^2 is integrable), X is called **square-integrable**.
- If $\mathbb{E}[|X|^m] < \infty$, for some $m > 0$, we say that X **has a finite m-th moment**.
- If X has a finite m-th moment, the expectation $\mathbb{E}[|X - \mathbb{E}[X]|^m]$ exists and we call it the **m-th central moment**.

It can be shown that the expectation $\mathbb{E}$ possesses the following properties, where X and Y are both assumed to be integrable:

1. $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$, for $\alpha, \beta \in \mathbb{R}$ (linearity of expectation).

2. $\mathbb{E}[X] \geq \mathbb{E}[Y]$ if $\mathbb{P}[X \geq Y] = 1$ (monotonicity of expectation).

Definition 1.4. Let X be a square-integrable random variable. We define the **variance** $\mathrm{Var}[X]$ by

$$\mathrm{Var}[X] = \mathbb{E}[(X - m)^2], \text{ where } m = \mathbb{E}[X].$$

The square-root $\sqrt{\mathrm{Var}[X]}$ is called the **standard deviation** of X.

Remark 1.5. Each square-integrable random variable is automatically integrable. Also, if the m-th moment exists, then all lower moments also exist.

We still need to define what happens with random variables that take the value $+\infty$, but that is very easy. We stipulate that $\mathbb{E}[X]$ *does not exist*, (i.e., $\mathbb{E}[X] = +\infty$) as long as $\mathbb{P}[X = +\infty] > 0$. Simply put, the expectation of a random variable is infinite if there is a positive chance (no matter how small) that it will take the value $+\infty$.

1.5 Events and probability

Probability is usually first explained in terms of the **sample space** or **probability space** (which we denote by Ω in these notes) and various *subsets* of Ω which are called events[1] Events typically contain all **elementary events**, i.e., elements of the probability space, usually denoted by ω. For example, if we are interested in the likelihood of getting an odd number as a sum of outcomes of two dice throws, we build a probability space

$$\Omega = \{(1,1), (1,2), \ldots, (6,1), (2,1), (2,2), \ldots, (2,6), \ldots, (6,1), (6,2), \ldots, (6,6)\}$$

and define the event A which contains of all pairs $(k,l) \in \Omega$ such that $k+l$ is an odd number, i.e.,

$$A = \{(1,2), (1,4), (1,6), (2,1), (2,3), \ldots, (6,1), (6,3), (6,5)\}.$$

One can think of events as very simple random variables. Indeed, if, for an event A, we define the random variable $\mathbf{1}_A$ by

$$\mathbf{1}_A = \begin{cases} 1, & A \text{ happened,} \\ 0, & A \text{ did not happen,} \end{cases}$$

we get the *indicator random variable* mentioned above. Conversely, for any indicator random variable X, we define the **indicated event** A as the set of all elementary events at which X takes the value 1.

What does all this have to do with probability? The analogy goes one step further. If we apply the notion of expectation to the indicator random variable $X = \mathbf{1}_A$, we get the probability of A:

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}[A].$$

Indeed, $\mathbf{1}_A$ takes the value 1 on A, and the value 0 on the complement $A^c = \Omega \setminus A$. Therefore, $\mathbb{E}[\mathbf{1}_A] = 1 \times \mathbb{P}[A] + 0 \times \mathbb{P}[A^c] = \mathbb{P}[A]$.

[1] When Ω is infinite, not all of its subsets can be considered events, due to very strange technical reasons. We will disregard that fact for the rest of the course. If you feel curious as to why that is the case, google Banach-Tarski paradox, and try to find a connection.

1.6 Dependence and independence

One of the main differences between random variables and (deterministic or non-random) quantities is that in the former case the whole is more than the sum of its parts. What do I mean by that? When two random variables, say X and Y, are considered in the same setting, you must specify more than just their distributions, if you want to compute probabilities that involve both of them. Here are two examples.

1. We throw two dice, and denote the outcome on the first one by X and the second one by Y.
2. We throw two dice, and denote the outcome of the first one by X, set $Y = 6 - X$ and forget about the second die.

In both cases, both X and Y have the same distribution

$$X, Y \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

The pairs (X, Y) are, however, very different in the two examples. In the first one, if the value of X is revealed, it will not affect our view of the value of Y. Indeed, the dice are not "connected" in any way (they are independent in the language of probability). In the second case, the knowledge of X allows us to say what Y is without any doubt - it is $6 - X$.

This example shows that when more than one random variable is considered, one needs to obtain external information about their relationship - not everything can be deduced only by looking at their distributions (pmfs, or ...).

One of the most common forms of relationship two random variables can have is the one of example (1) above, i.e., no relationship at all. More formally, we say that two (discrete) random variables X and Y are **independent** if

$$\mathbb{P}[X = x \text{ and } Y = y] = \mathbb{P}[X = x]\mathbb{P}[Y = y],$$

for *all* x and y in the respective supports $_X$ and $_Y$ of X and Y. The same concept can be applied to events, and we say that two events A and B are independent if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B].$$

The notion of independence is central to probability theory (and this course) because it is relatively easy to spot in real life. If there is no physical mechanism that ties two events (like the two dice we throw), we are inclined to declare them independent[2]. One of the most important tasks in probabilistic modelling is the identification of the (small number of) independent random variables which serve as building blocks for a big complex system. You will see many examples of that as we proceed through the course.

[2] Actually, true independence does not exist in reality, save, perhaps a few quantum-theoretic phenomena. Even with apparently independent random variables, dependence can sneak in the most sly of ways. Here is a funny example: a recent survey has found a large correlation between the sale of diapers and the sale of six-packs of beer across many Walmart stores throughout the country. At first these two appear independent, but I am sure you can come up with many an amusing story why they should, actually, be quite dependent.

1.7 Conditional probability

When two random variables are not independent, we still want to know how the knowledge of the exact value of one of the affects our guesses about the value of the other. That is what the conditional probability is for. We start with the definition, and we state it for events first: for two events A, B such that $\mathbb{P}[B] > 0$, the **conditional probability** $\mathbb{P}[A|B]$ **of** A **given** B is defined as:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

The conditional probability is *not defined* when $\mathbb{P}[B] = 0$ (otherwise, we would be computing $\frac{0}{0}$ - why?). Every statement in the sequel which involves conditional probability will be assumed to hold only when $\mathbb{P}[B] = 0$, without explicit mention.

The conditional probability calculations often use one of the following two formulas. Both of them use the familiar concept of partition. If you forgot what it is, here is a definition: a collection $A_1, A_2, \ldots, A_n$ of events is called a **partition of** Ω if a) $A_1 \cup A_2 \cup \ldots A_n = \Omega$ and b) $A_i \cap A_j = \emptyset$ for all pairs $i, j = 1, \ldots, n$ with $i \neq j$. So, let $A_1, \ldots, A_n$ be a partition of Ω, and let B be an event.

1. **The Law of Total Probability.**

$$\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B|A_i]\mathbb{P}[A_i].$$

2. **Bayes formula.** For $k = 1, \ldots, n$, we have

$$\mathbb{P}[A_k|B] = \frac{\mathbb{P}[B|A_k]\mathbb{P}[A_k]}{\sum_{i=1}^{n} \mathbb{P}[B|A_i]\mathbb{P}[A_i]}.$$

Even though the formulas above are stated for finite partitions, they remain true when the number of A_k's is countably infinite. The finite sums have to be replaced by infinite series, however.

Random variables can be substituted for events in the definition of conditional probability as follows: for two random variables X and Y, the **conditional probabilty** that $X = x$, **given** $Y = y$ (with x and y in respective supports $_X$ and $_Y$) is given by

$$\mathbb{P}[X = x|Y = y] = \frac{\mathbb{P}[X = x \text{ and } Y = y]}{\mathbb{P}[Y = y]}.$$

The formula above produces a different probability distribution for each y. This is called the **conditional distribution** of X, **given** $Y = y$. We give a simple example to illustrate this concept. Let X be the number of *heads* obtained when two coins are thrown, and let Y be the indicator of the event that the second coin shows *heads*. The distribution of X is Binomial:

$$X \sim \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix},$$

or, in the more compact notation which we use when the support is clear from the context $X \sim (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. The random variable Y has the Bernoulli distribution $Y = (\frac{1}{2}, \frac{1}{2})$. What happens

to the distribution of X, when we are told that $Y = 0$, i.e., that the second coin shows *heads*. In that case we have

$$\mathbb{P}[X = x|Y = 0] = \begin{cases} \frac{\mathbb{P}[X=0,Y=0]}{\mathbb{P}[Y=0]} = \frac{\mathbb{P}[\text{ the pattern is TT }]}{\mathbb{P}[Y=0]} = \frac{1/4}{1/2} = \boxed{\frac{1}{2}}, & x = 0 \\ \frac{\mathbb{P}[X=1,Y=0]}{\mathbb{P}[Y=0]} = \frac{\mathbb{P}[\text{ the pattern is HT }]}{\mathbb{P}[Y=0]} = \frac{1/4}{1/2} = \boxed{\frac{1}{2}}, & x = 1 \\ \frac{\mathbb{P}[X=2,Y=0]}{\mathbb{P}[Y=0]} = \frac{\mathbb{P}[\text{ well, there is no such pattern }]}{\mathbb{P}[Y=0]} = \frac{0}{1/2} = \boxed{0}, & x = 2 \end{cases}$$

Thus, the conditional distribution of X, given $Y = 0$, is $(\frac{1}{2}, \frac{1}{2}, 0)$. A similar calculation can be used to get the conditional distribution of X, but now given that $Y = 1$, is $(0, \frac{1}{2}, \frac{1}{2})$. The moral of the story is that the additional information contained in Y can alter our views about the unknown value of X using the concept of conditional probability. One final remark about the relationship between independence and conditional probability: suppose that the random variables X and Y are independent. Then the knowledge of Y should not affect how we think about X; indeed, then

$$\mathbb{P}[X = x|Y = y] = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} = \frac{\mathbb{P}[X = x]\mathbb{P}[Y = y]}{\mathbb{P}[Y = y]} = \mathbb{P}[X = x].$$

The conditional distribution does not depend on y, and coincides with the unconditional one.

The notion of independence for two random variables can easily be generalized to larger collections

Definition 1.6. Random variables $X_1, X_2, \ldots, X_n$ are said to be **independent** if

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots X_n = x_n] = \mathbb{P}[X_1 = x_1]\mathbb{P}[X_2 = x_2] \ldots \mathbb{P}[X_n = x_n]$$

for *all* $x_1, x_2, \ldots, x_n$.

An infinite collection of random variables is said to be **independent** if all of its finite subcollections are independent.

Independence is often used in the following way:

Proposition 1.7. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

1. *$g_1(X_1)$, ..., $g_n(X_n)$ are also independent for (practically) all functions $g_1, \ldots, g_n$,*
2. *if X_1, ..., X_n are integrable then the product $X_1 \ldots X_n$ is integrable and*

$$\mathbb{E}[X_1 \ldots X_n] = \mathbb{E}[X_1] \ldots \mathbb{E}[X_n], \text{ and}$$

3. *if X_1, ..., X_n are square-integrable, then*

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

Equivalently

$$\mathrm{Cov}[X_i, X_j] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] = 0,$$

for all $i \neq j \in \{1, 2, \ldots, n\}$.

Remark 1.8. The last statement says that independent random variables are uncorrelated. The converse is not true. There are uncorrelated random variables which are not independent.

When several random variables $(X_1, X_2, \dots X_n)$ are considered in the same setting, we often group them together into a **random vector**. The **distribution** of the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is the collection of all probabilities of the form

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n],$$

when $x_1, x_2, \dots, x_n$ range through all numbers in the appropriate supports. Unlike in the case of a single random variable, writing down the distributions of random vectors in tables is a bit more difficult. In the two-dimensional case, one would need an entire matrix, and in the higher dimensions some sort of a hologram would be the only hope.

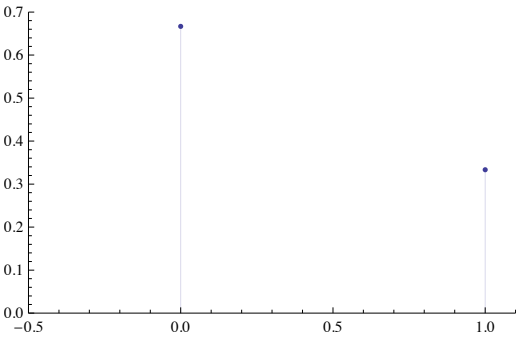
The distributions of the components $X_1, \dots, X_n$ of the random vector $\boldsymbol{X}$ are called the **marginal distributions** of the random variables $X_1, \dots, X_n$. When we want to stress the fact that the random variables $X_1, \dots, X_n$ are a part of the same random vector, we call the distribution of $\boldsymbol{X}$ the **joint distribution** of $X_1, \dots, X_n$. It is important to note that, unless random variables $X_1, \dots, X_n$ are a priori known to be independent, the joint distribution holds more information about $\boldsymbol{X}$ than all marginal distributions together.

1.8 Examples

Here is a short list of some of the most important discrete random variables. You will learn about generating functions soon.

Example 1.9.

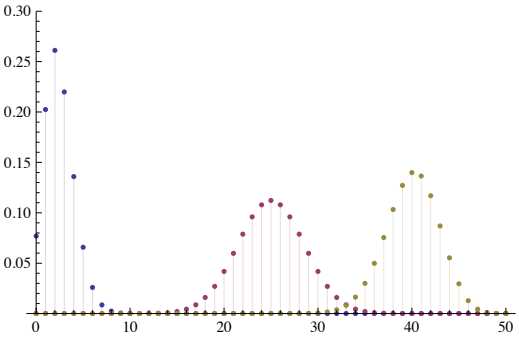
Bernoulli. *Success* (1) *of failure* (0) *with probability* p *(if success is encoded by* 1*, failure by* -1 *and* $p = \frac{1}{2}$*, we call it the* ***coin toss****).*



.parameters : $p \in (0, 1)$ $(q = 1 - p)$
.notation : $b(p)$
.support : $\{0, 1\}$
.pmf : $p_0 = p$ and $p_1 = q = 1 - p$
.generating function : $ps + q$
.mean : p
.standard deviation : $\sqrt{pq}$
.figure : the mass function a Bernoulli distribution with $p = 1/3$.

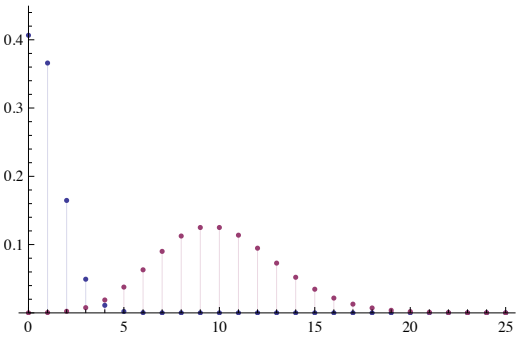
Binomial. *The number of successes in* n *repeti-*

tions of a Bernoulli trial with success probability p.



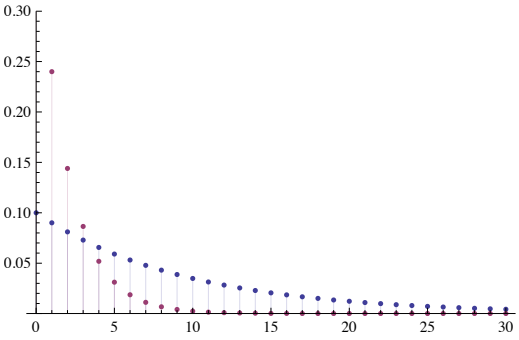
.parameters : $n \in \mathbb{N}$, $p \in (0,1)$ $(q = 1-p)$
.notation : $b(n,p)$
.support : $\{0,1,\ldots,n\}$
.pmf : $p_k = \binom{n}{k} p^k q^{n-k}$, $k = 0,\ldots,n$
.generating function : $(ps+q)^n$
.mean : np
.standard deviation : $\sqrt{npq}$
.figure : mass functions of three binomial distributions with $n = 50$ and $p = 0.05$ (blue), $p = 0.5$ (purple) and $p = 0.8$ (yellow).

Poisson. *The number of spelling mistakes one makes while typing a single page.*



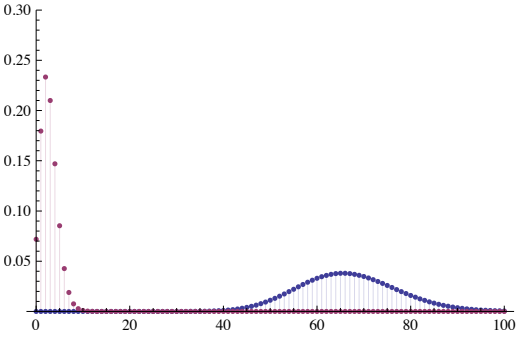
.parameters : $\lambda > 0$
.notation : $p(n,p)$
.support : $\mathbb{N}_0$
.pmf : $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$, $k \in \mathbb{N}_0$
.generating function : $e^{\lambda(s-1)}$
.mean : λ
.standard deviation : $\sqrt{\lambda}$
.figure : mass functions of two Poisson distributions with parameters $\lambda = 0.9$ (blue) and $\lambda = 10$ (purple).

Geometric. *The number of repetitions of a Bernoulli trial with parameter p until the first success.*



.parameters : $p \in (0,1)$, $q = 1-p$
.notation : $g(p)$
.support : $\mathbb{N}_0$
.pmf : $p_k = pq^{k-1}$, $k \in \mathbb{N}_0$
.generating function : $\frac{p}{1-qs}$
.mean : $\frac{q}{p}$
.standard deviation : $\frac{\sqrt{q}}{p}$
.figure : mass functions of two Geometric distributions with parameters $p = 0.1$ (blue) and $p = 0.4$ (purple).

Negative Binomial. *The number of failures it takes to obtain r successes in repeated independent Bernoulli trials with success probability p.*



.parameters : $r \in \mathbb{N}$, $p \in (0,1)$ $(q = 1-p)$
.notation : $g(n,p)$
.support : $\mathbb{N}_0$
.pmf : $p_k = \binom{-r}{k} p^r q^k$, $k = \in \mathbb{N}_0$
.generating function : $\left(\frac{p}{1-qs}\right)^r$
.mean : $r\frac{q}{p}$
.standard deviation : $\frac{\sqrt{qr}}{p}$
.figure : mass functions of two negative binomial distributions with $r = 100, p = 0.6$ (blue) and $r = 25, p = 0.9$ (purple).

Chapter 2

Mathematica in 15 min

Mathematica is a glorified calculator. Here is how to use it[1].

2.1 Basic Syntax

- Symbols `+, -, /, ^, *` are all supported by *Mathematica*. Multiplication can be represented by a space between variables. `a x + b` and `a*x + b` are identical.
- **Warning:** *Mathematica* is case-sensitive. For example, the command to exit is `Quit` and not `quit` or `QUIT`.
- Brackets are used around function arguments. Write `Sin[x]`, not `Sin(x)` or `Sin{x}`.
- Parentheses `( )` group terms for math operations: `(Sin[x]+Cos[y])*(Tan[z]+z^2)`.
- If you end an expression with a `;` (semi-colon) it will be executed, but its output will not be shown. This is useful for simulations, e.g.
- Braces `{ }` are used for lists:

```
In[1]:= A = {1, 2, 3}
Out[1]= {1, 2, 3}
```

- Names can refer to variables, expressions, functions, matrices, graphs, etc. A name is assigned using `name = object`. An expression may contain undefined names:

```
In[5]:= A = (a + b) ^ 3
Out[5]= (a + b)^3
In[6]:= A^2
Out[6]= (a + b)^6
```

[1] Actually, this is just a tip of the iceberg. It can do many many many other things.

- The percent sign % stores the value of the previous result

```
In[7]:= 5 + 3
Out[7]= 8
In[8]:= %^2
Out[8]= 64
```

2.2 Numerical Approximation

- N[expr] gives the approximate numerical value of expression, variable, or command:

```
In[9]:= N[Sqrt[2]]
Out[9]= 1.41421
```

- N[%] gives the numerical value of the previous result:

```
In[17]:= E + Pi
Out[17]= e + π
In[18]:= N[%]
Out[18]= 5.85987
```

- N[expr,n] gives n digits of precision for the expression expr:

```
In[14]:= N[Pi, 30]
Out[14]= 3.14159265358979323846264338328
```

- Expressions whose result can't be represented exactly don't give a value unless you request approximation:

```
In[11]:= Sin[3]
Out[11]= Sin[3]
In[12]:= N[Sin[3]]
Out[12]= 0.14112
```

2.3 Expression Manipulation

- Expand[expr] (algebraically) expands the expression expr:

```
In[19]:= Expand[(a + b) ^ 2]
Out[19]= a^2 + 2 a b + b^2
```

- `Factor[expr]` factors the expression `expr`

```
In[20]:= Factor[ a^2 - b^2]
Out[20]= (a - b) (a + b)

In[21]:= Factor[ x^2 - 5 x + 6]
Out[21]= (-3 + x) (-2 + x)
```

- `Simplify[expr]` performs all kinds of simplifications on the expression `expr`:

```
In[35]:= A = x / (x - 1) - x / (1 + x)
Out[35]= x/(-1 + x) - x/(1 + x)

In[36]:= Simplify[A]
Out[36]= 2 x/(-1 + x^2)
```

2.4 Lists and Functions

- If `L` is a list, its length is given by `Length[L]`. The n^{th} element of `L` can be accessed by `L[[n]]` (note the double brackets):

```
In[43]:= L = {2, 4, 6, 8, 10}
Out[43]= {2, 4, 6, 8, 10}

In[44]:= L[[3]]
Out[44]= 6
```

- Addition, subtraction, multiplication and division can be applied to lists element by element:

```
In[1]:= L = {1, 3, 4}; K = {3, 4, 2};
In[2]:= L + K
Out[2]= {4, 7, 6}
In[3]:= L / K
Out[3]= {1/3, 3/4, 2}
```

- If the expression `expr` depends on a variable (say `i`), `Table[expr,{i,m,n}]` produces a list of the values of the expression `expr` as `i` ranges from `m` to `n`

```
In[37]:= Table[i^2, {i, 0, 5}]
Out[37]= {0, 1, 4, 9, 16, 25}
```

- The same works with two indices - you will get a list of lists

```
In[40]:= Table[i^j, {i, 1, 3}, {j, 2, 3}]
Out[40]= {{1, 1}, {4, 8}, {9, 27}}
```

- It is possible to define your own functions in *Mathematica*. Just use the *underscore syntax* `f[x_]=expr`, where `expr` is some expression involving `x`:

```
In[47]:= f[x_] = x^2
Out[47]= x^2
In[48]:= f[x + y]
Out[48]= (x + y)^2
```

- To apply the function `f` (either built-in, like `Sin`, or defined by you) to each element of the list `L`, you can use the command `Map` with syntax `Map[f,L]`:

```
In[50]:= f[x_] = 3 * x
Out[50]= 3 x
In[51]:= L = {1, 2, 3, 4}
Out[51]= {1, 2, 3, 4}
In[52]:= Map[f, L]
Out[52]= {3, 6, 9, 12}
```

- If you want to add all the elements of a list L, use `Total[L]`. The list of the same length as L, but whose k^{th} element is given by the sum of the first k elements of L is given by `Accumulate[L]`:

```
In[8]:= L = {1, 2, 3, 4, 5}
Out[8]= {1, 2, 3, 4, 5}
In[9]:= Accumulate[L]
Out[9]= {1, 3, 6, 10, 15}
In[10]:= Total[L]
Out[10]= 15
```

2.5 Linear Algebra

- In *Mathematica*, matrix is a nested list, i.e., a list whose elements are lists. By convention, matrices are represented row by row (inner lists are row vectors).
- To access the element in the i^{th} row and j^{th} column of the matrix `A`, type `A[[i,j]]` or `A[[i]][[j]]`:

```
In[59]:= A = {{2, 1, 3}, {5, 6, 9}}
Out[59]= {{2, 1, 3}, {5, 6, 9}}
In[60]:= A[[2, 3]]
Out[60]= 9
In[61]:= A[[2]][[3]]
Out[61]= 9
```

- `Matrixform[expr]` displays `expr` as a matrix (provided it is a nested list)

```
In[9]:= A = Table[i * 2^j, {i, 2, 5}, {j, 1, 2}]
Out[9]= {{4, 8}, {6, 12}, {8, 16}, {10, 20}}
In[10]:= MatrixForm[A]
Out[10]//MatrixForm=
( 4   8  )
( 6   12 )
( 8   16 )
( 10  20 )
```

- Commands `Transpose[A]`, `Inverse[A]`, `Det[A]`, `Tr[A]` and `MatrixRank[A]` return the transpose, inverse, determinant, trace and rank of the matrix `A`, respectively.
- To compute the n^{th} power of the matrix `A`, use `MatrixPower[A,n]`

```
In[21]:= A = {{1, 1}, {1, 0}}
Out[21]= {{1, 1}, {1, 0}}
In[22]:= MatrixForm[MatrixPower[A, 5]]
Out[22]//MatrixForm=
( 8  5 )
( 5  3 )
```

- Identity matrix of order n is produced by `IdentityMatrix[n]`.
- If `A` and `B` are matrices of the same order, `A+B` and `A-B` are their sum and difference.

- If `A` and `B` are of compatible orders, `A.B` (that is a dot between them) is the matrix product of `A` and `B`.
- For a square matrix `A`, `CharacteristicPolynomial[A,x]` is the characteristic poynomial, $\det(xI - A)$ in the variable `x`:

```
In[40]:= A = {{3, 4}, {2, 1}}
Out[40]= {{3, 4}, {2, 1}}
In[42]:= CharacteristicPolynomial[A, x]
Out[42]= -5 - 4 x + x^2
```

- To get eigenvalues and eigenvectors use `Eigenvalues[A]` and `Eigenvectors[A]`. The results will be the list containing the eigenvalues in the `Eigenvalues` case, and the list of eigenvectors of `A` in the `Eigenvectors` case:

```
In[52]:= A = {{3, 4}, {2, 1}}
Out[52]= {{3, 4}, {2, 1}}
In[53]:= Eigenvalues[A]
Out[53]= {5, -1}
In[54]:= Eigenvectors[A]
Out[54]= {{2, 1}, {-1, 1}}
```

2.6 Predefined Constants

- A number of constants are predefined by *Mathematica*: `Pi`, `I` ($\sqrt{-1}$), `E` (2.71828...), `Infinity`. Don't use `I`, `E` (or `D`) for variable names - *Mathematica* will object.
- A number of standard functions are built into *Mathematica*: `Sqrt[]`, `Exp[]`, `Log[]`, `Sin[]`, `ArcSin[]`, `Cos[]`, etc.

2.7 Calculus

- `D[f,x]` gives the derivative of `f` with respect to `x`. For the first few derivatives you can use `f'[x]`, `f''[x]`, etc.

```
In[66]:= D[x^k, x]
Out[66]= k x^(-1+k)
```

- `D[f,{x,n}]` gives the n^{th} derivative of `f` with respect to `x`
- `D[f,x,y]` gives the mixed derivative of `f` with respect to `x` and `y`.

- `Integrate[f,x]` gives the *indefinite* integral of `f` with respect to `x`:

```
In[67]:= Integrate[Log[x], x]
Out[67]= -x + x Log[x]
```

- `Integrate[f,{x,a,b}]` gives the *definite* integral of `f` on the interval $[a, b]$ (`a` or `b` can be `Infinity` (∞) or `-Infinity` ($-\infty$)) :

```
In[72]:= Integrate[Exp[-2 * x], {x, 0, Infinity}]
Out[72]= 1/2
```

- `NIntegrate[f,{x,a,b}]` gives the numerical approximation of the definite integral. This usually returns an answer when `Integrate[]` doesn't work:

```
In[76]:= Integrate[1 / (x + Sin[x]), {x, 1, 2}]
Out[76]= Integrate[1/(x + Sin[x]), {x, 1, 2}]
In[77]:= NIntegrate[1 / (x + Sin[x]), {x, 1, 2}]
Out[77]= 0.414085
```

- `Sum[expr,{n,a,b}]` evaluates the (finite or infinite) sum. Use `NSum` for a numerical approximation.

```
In[80]:= Sum[1 / k^4, {k, 1, Infinity}]
Out[80]= π^4/90
```

- `DSolve[eqn,y,x]` solves (given the general solution to) an ordinary differential equation for function y in the variable x:

```
In[88]:= DSolve[y''[x] + y[x] == x, y[x], x]
Out[88]= {{y[x] → x + C[1] Cos[x] + C[2] Sin[x]}}
```

- To calculate using initial or boundary conditions use `DSolve[{eqn,conds},y,x]`:

```
In[93]:= DSolve[{y'[x] == y[x]^2, y[0] == 1}, y[x], x]
Out[93]= {{y[x] → 1/(1 - x)}}
```

2.8 Solving Equations

- Algebraic equations are solved with `Solve[lhs==rhs,x]`, where `x` is the variable with respect to which you want to solve the equation. Be sure to use == and not = in equations. *Mathematica* returns the list with all solutions:

```
In[81]:= Solve[ x^3 == x, x]
Out[81]= {{x → -1}, {x → 0}, {x → 1}}
```

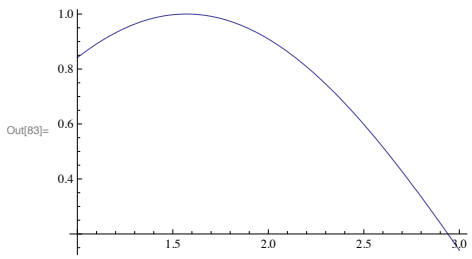
- `FindRoot[f,{x,x0}]` is used to find a root when `Solve[]` does not work. It solves for x numerically, using an initial value of x0:

```
In[82]:= FindRoot[Cos[x] == x, {x, 1}]
Out[82]= {x → 0.739085}
```

2.9 Graphics

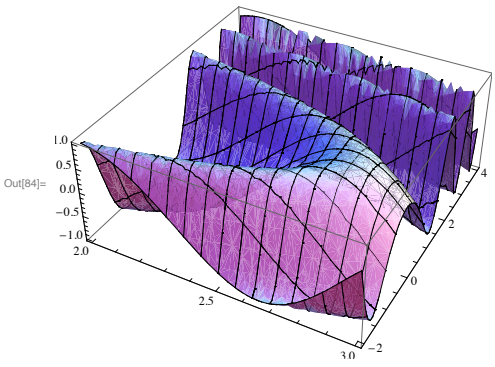
- `Plot[expr,{x,a,b}]` plots the expression `expr`, in the variable `x`, from a to b:

```
In[83]:= Plot[Sin[x], {x, 1, 3}]
```



- `Plot3D[expr,{x,a,b},{y,c,d}]` produces a 3D plot in 2 variables:

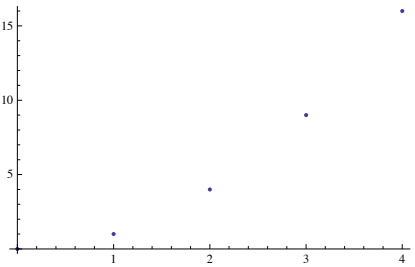
```
In[84]:= Plot3D[Sin[x^2 + y^2], {x, 2, 3}, {y, -2, 4}]
```



- If `L` is a list of the form L= $\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\}\}$, you can use the command `ListPlot[L]` to display a graph consisting of points (x_1, y_1), ..., (x_n, y_n):

```
In[11]:= L = Table[{i, i^2}, {i, 0, 4}]
Out[11]= {{0, 0}, {1, 1}, {2, 4}, {3, 9}, {4, 16}}
In[12]:= ListPlot[L]
```

Out[12]=



2.10 Probability Distributions and Simulation

- `PDF[distr,x]` and `CDF[distr,x]` return the pdf (pmf in the discrete case) and the cdf of the distribution `distr` in the variable x. `distr` can be one of:
  - `NormalDistribution[m,s]`,
  - `ExponentialDistribution[l]`,
  - `UniformDistribution[{a,b}]`,
  - `BinomialDistribusion[n,p]`,

  and many many others (see `?PDF` and follow various links from there).

- Use `ExpectedValue[expr,distr,x]` to compute the expectation $\mathbb{E}[f(X)]$, where `expr` is the expression for the function f in the variable x:

```
In[23]:= distr = PoissonDistribution[λ]
Out[23]= PoissonDistribution[λ]
In[25]:= PDF[distr, x]
Out[25]= e^(-λ) λ^x / x!
In[27]:= ExpectedValue[x^3, distr, x]
Out[27]= λ + 3 λ^2 + λ^3
```

- There is no command for the generating function, but you can get it by computing the characteristic function and changing the variable a bit `CharacteristicFunction[distr, - I Log[s]]`:

```
In[22]:= distr = PoissonDistribution[λ]

Out[22]= PoissonDistribution[λ]

In[23]:= CharacteristicFunction[distr, -I Log[s]]

Out[23]= e^((-1+s) λ)
```

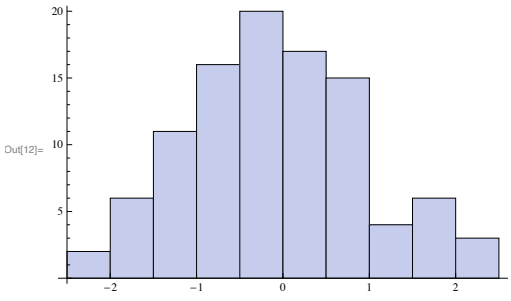
- To get a random number (unifomly distributed between 0 and 1) use `RandomReal[]`. A uniformly distributed random number on the interval $[a, b]$ can be obtained by `RandomReal[{a,b}]`. For a list of n uniform random numbers on $[a, b]$ write `RandomReal[{a,b},n]`.

```
In[2]:= RandomReal[]

Out[2]= 0.168904

In[3]:= RandomReal[{7, 9}]

Out[3]= 7.83027

In[5]:= RandomReal[{0, 1}, 3]

Out[5]= {0.368422, 0.961658, 0.692345}
```

- If you need a random number from a particular *continuous* distribution (normal, say), use `RandomReal[distr]` or `RandomReal[distr,n]` if you need n draws.
- When drawing from a *discrete* distribution use `RandomInteger` instead.
- If `L` is a list of numbers, `Histogram[L]` displays a histogram of `L` (you need to load the package *Histograms* by issuing the command `<<Histograms‘` before you can use it):

```
In[7]:= L = RandomReal[NormalDistribution[0, 1], 100];

In[10]:= << Histograms‘

In[12]:= Histogram[L]
```

Out[12]=



2.11 Help Commands

- `?name` returns information about `name`
- `??name` adds extra information about `name`
- `Options[command]` returns all options that may be set for a given command

- `?pattern` returns the list of matching names (used when you forget a command). pattern contains one or more asterisks `*` which match any string. Try `?*Plot*`

2.12 Common Mistakes

- *Mathematica* is case sensitive: `Sin` is not `sin`
- Don't confuse braces, brackets, and parentheses `{}`, `[]`, `()`
- Leave spaces between variables: write `a x^2` instead of `ax^2`, if you want to get ax^2.
- Matrix multiplication uses `.` instead of `*` or a space.
- Don't use `=` instead of `==` in `Solve` or `DSolve`
- If you are using an older version of *Mathematica*, a function might be defined in an external module which has to be loaded before the function can be used. For example, in some versions, the command `<<Graphics'` needs to be given before any plots can be made. The symbol at the end is *not* an apostrophe - it is the dash above the TAB key.
- Using `Integrate[]` around a singular point can yield wrong answers. (Use `NIntegrate[]` to check.)
- Don't forget the underscore `_` when you define a function.

Chapter 3

Stochastic Processes

Definition 3.1. Let $\mathcal{T}$ be a subset of $[0, \infty)$. A family of random variables $\{X_t\}_{t\in\mathcal{T}}$, indexed by $\mathcal{T}$, is called a **stochastic (or random) process**. When $\mathcal{T} = \mathbb{N}$ (or $\mathcal{T} = \mathbb{N}_0$), $\{X_t\}_{t\in\mathcal{T}}$ is said to be a **discrete-time process**, and when $\mathcal{T} = [0, \infty)$, it is called a **continuous-time process**.

When $\mathcal{T}$ is a singleton (say $\mathcal{T} = \{1\}$), the process $\{X_t\}_{t\in\mathcal{T}} \equiv X_1$ is really just a single random variable. When $\mathcal{T}$ is finite (e.g., $\mathcal{T} = \{1, 2, \ldots, n\}$), we get a random vector. Therefore, stochastic processes are generalizations of random vectors. The interpretation is, however, somewhat different. While the components of a random vector usually (not always) stand for different spatial coordinates, the index $t \in T$ is more often than not interpreted as time. Stochastic processes usually model the evolution of a random system in time. When $\mathcal{T} = [0, \infty)$ (**continuous-time processes**), the value of the process can change every instant. When $\mathcal{T} = \mathbb{N}$ (**discrete-time processes**), the changes occur discretely.

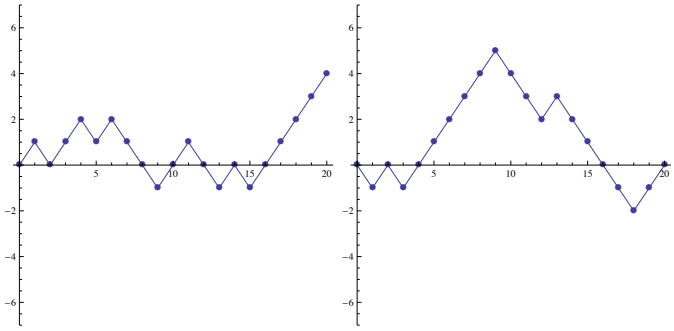
In contrast to the case of random vectors or random variables, it is not easy to define a notion of a density (or a probability mass function) for a stochastic process. Without going into details why exactly this is a problem, let me just mention that the main culprit is the infinity. One usually considers a family of (discrete, continuous, etc.) **finite-dimensional distributions**, i.e., the joint distributions of random vectors

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n}),$$

for all $n \in \mathbb{N}$ and all choices $t_1, \ldots, t_n \in \mathcal{T}$.

The notion of a stochastic processes is very important both in mathematical theory and its applications in science, engineering, economics, etc. It is used to model a large number of various phenomena where the quantity of interest varies discretely or continuously through time in a non-predictable fashion.

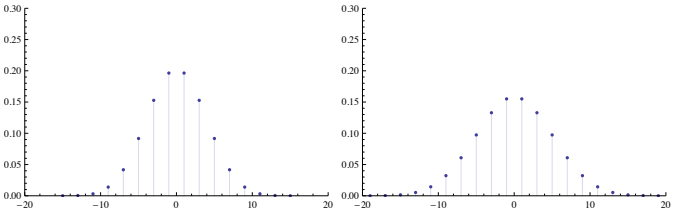
Every stochastic process can be viewed as a function of two variables - t and ω. For each fixed t, $\omega \mapsto X_t(\omega)$ is a random variable, as postulated in the definition. However, if we change our point of view and keep ω fixed, we see that the stochastic process is a function mapping ω to the real-valued function $t \mapsto X_t(\omega)$. These functions are called the **trajectories** of the stochastic process X.



Figures on the left show two different trajectories of a simple random walk[a], i.e., each one corresponds to a (different) frozen $\omega \in \Omega$, but t goes from 0 to 30.

[a] We will define the simple random walk later. For now, let us just say that is behaves as follows. It starts at $x = 0$ for $t = 0$. After that a fair coin is tossed and we move up (to $x = 1$) if *heads* is observed and down to $x = -1$ is we see *tails*. The procedure is repeated at $t = 1, 2, \ldots$ and the position at $t+1$ is determined in the same way, independently of all the coin tosses before (note that the position at $t = k$ can be any of the following $x = -k$, $x = -k + 2$, ..., $x = k - 2$, $x = k$).

Unlike with the figures above, the two pictures on the right show two *time-slices* of the same random process; in each graph, the time t is fixed ($t = 15$ vs. $t = 25$) but the various values random variables X_{15} and X_{25} can take are presented through the probability mass functions.



3.1 The canonical probability space

When one deals with infinite-index ($\#\mathcal{T} = +\infty$) stochastic processes, the construction of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to support a given model is usually quite a technical matter. This course does not suffer from that problem because all our models can be implemented on a special probability space. We start with the sample-space Ω:

$$\Omega = [0,1] \times [0,1] \times \cdots = [0,1]^{\infty},$$

and any generic element of Ω will be a sequence $\omega = (\omega_0, \omega_1, \omega_2, \ldots)$ of real numbers in $[0,1]$. For $n \in \mathbb{N}_0$ we define the mapping $\gamma_n : \Omega \to [0,1]$ which simply chooses the n-th coordinate :

$$\gamma_n(\omega) = \omega_n.$$

The proof of the following theorem can be found in advanced probability books:

Theorem 3.2. *There exists a σ-algebra $\mathcal{F}$ and a probability $\mathbb{P}$ on Ω such that*

1. *each γ_n, $n \in \mathbb{N}_0$ is a random variable with the uniform distribution on $[0,1]$, and*
2. *the sequence $\{\gamma_n\}_{n \in \mathbb{N}_0}$ is independent.*

Remark 3.3. One should think of the sample space Ω as a source of all the randomness in the system: the elementary event $\omega \in \Omega$ is chosen by a process beyond out control and the exact value of ω is assumed to be unknown. All the other parts of the system are possibly complicated, but deterministic, functions of ω (random variables). When a coin is tossed, only a single drop of randomness is needed - the outcome of a coin-toss. When several coins are tossed, more randomness is involved and the sample space must be bigger. When a system involves an infinite number of random variables (like a stochastic process with infinite $\mathcal{T}$), a large sample space Ω is needed.

3.2 Constructing the Random Walk

Let us show how to construct the simple random walk on the canonical probability space $(\Omega, \mathcal{F}, \mathbb{P})$ from Theorem 3.2. First of all, we need a definition of the simple random walk:

Definition 3.4. A stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ is called a **simple random walk** if

1. $X_0 = 0$,
2. the increment $X_{n+1} - X_n$ is independent of $(X_0, X_1, \ldots, X_n)$ for each $n \in \mathbb{N}_0$, and
3. the increment $X_{n+1} - X_n$ has the coin-toss distribution, i.e.

$$\mathbb{P}[X_{n+1} - X_n = 1] = \mathbb{P}[X_{n+1} - X_n = -1] = \tfrac{1}{2}.$$

For the sequence $\{\gamma_n\}_{n\in\mathbb{N}}$, given by Theorem 3.2, define the following, new, sequence $\{\xi_n\}_{n\in\mathbb{N}}$ of random variables:

$$\xi_n = \begin{cases} 1, & \gamma_n \geq \frac{1}{2} \\ -1, & \text{otherwise.} \end{cases}$$

Then, we set

$$X_0 = 0,\ X_n = \sum_{k=1}^{n} \xi_k,\ n \in \mathbb{N}.$$

Intuitively, we use each ξ_n to emulate a coin toss and then define the value of the process X at time n as the cumulative sum of the first n coin-tosses.

Proposition 3.5. *The sequence $\{X_n\}_{n\in\mathbb{N}_0}$ defined above is a simple random walk.*

Proof. (1) is trivially true. To get (2), we first note that the $\{\xi_n\}_{n\in\mathbb{N}}$ is an independent sequence (as it has been constructed by an application of a deterministic function to each element of an independent sequence $\{\gamma_n\}_{n\in\mathbb{N}}$). Therefore, the increment $X_{n+1} - X_n = \xi_{n+1}$ is independent of all the previous coin-tosses $\xi_1, \ldots, \xi_n$. What we need to prove, though, is that it is independent of all the previous values of the process X. These, previous, values are nothing but linear combinations of the coin-tosses $\xi_1, \ldots, \xi_n$, so they must also be independent of ξ_{n+1}. Finally, to get (3), we compute

$$\mathbb{P}[X_{n+1} - X_n = 1] = \mathbb{P}[\xi_{n+1} = 1] = \mathbb{P}[\gamma_{n+1} \geq \tfrac{1}{2}] = \tfrac{1}{2}.$$

A similar computation shows that $\mathbb{P}[X_{n+1} - X_n = -1] = \frac{1}{2}$. □

3.3 Simulation

Another way of thinking about sample spaces, and randomness in general, is through the notion of **simulation**. Simulation is what I did to produce the two trajectories of the random walk above; a computer tossed a fair coin for me 30 times and I followed the procedure described above to construct a trajectory of the random walk. If I asked the computer to repeat the process, I would get different 30 coin-tosses[1]. This procedure is the exact same one we imagine nature (or casino equipment) follows whenever a non-deterministic situation is involved. The difference is, of course, that if we use the random walk to model out winnings in a fair gamble, it is much cheaper and faster to use the computer than to go out and stake (and possibly loose) large amounts of money. Another obvious advantage of the simulation approach is that it can be repeated; a simulation can be run many times and various statistics (mean, variance, etc.) can be computed.

More technically, every simulation involves two separate inputs. The first one if the actual sequence of outcomes of coin-tosses. The other one is the structure of the model - I have to teach the computer to "go up" if *heads* shows and to "go down" if *tails* show, and to repeat the same procedure several times. In more complicated situations this structure will be more complicated. What is remarkable is that the first ingredient, the coin-tosses, will stay almost as simple as in the random walk case, even in the most complicated models. In fact, all we need is a sequence of so-called **random numbers**. You will see through the many examples presented in this course that if I can get my computer to produce an independent sequence of uniformly distributed numbers between 0 and 1 (these are the random numbers) I can simulate trajectories of *all* important stochastic processes. Just to start you thinking, here is how to produce a coin-toss from a random number: declare *heads* if the random number drawn is between 0 and 0.5, and declare *tails* otherwise.

3.3.1 Random number generation

Before we get into intricacies of simulation of complicated stochastic processes, let us spend some time on the (seemingly) simple procedure of the generation of a single random number. In other words, how do you teach a computer to give you a random number between 0 and 1? Theoretically, the answer is *You can't!*. In practice, you can get quite close. The question of what actually constitutes a random number is surprisingly deep and we will not even touch it in this course.

Suppose we have written a computer program, a random number generator (RNG) - call it `rand` - which produces a random number between 0 and 1 every time we call it. So far, there is nothing that prevents `rand` from always returning the same number 0.4, or from alternating between 0.3 and 0.83. Such an implementation of `rand` will, however, hardly qualify for an RNG since the values it spits out come in a predictable order. We should, therefore, require any candidate for a random number generator to produce a sequence of numbers which is as unpredictable as possible. This is, admittedly, a hard task for a computer having only deterministic functions in its arsenal, and that is why the random generator design is such a difficult field. The state of the affairs is that we speak of *good* or *less good* random number generators, based on some statistical properties of the produced sequences of numbers.

[1] Actually, I would get the exact same 30 coin-tosses with probability 0.000000001

One of the most important requirements is that our RNG produce **uniformly distributed numbers** in $[0,1]$ - namely - the sequence of numbers produced by `rand` will have to cover the interval $[0,1]$ evenly, and, in the long run, the number of random numbers in each subinterval $[a,b]$ of $[0,1]$ should be proportional to the length of the interval $b-a$. This requirement if hardly enough, because the sequence

$$0, 0.1, 0.2, \ldots, 0.8, 0.9, 1,\ 0.05, 0.15, 0.25, \ldots, 0.85, 0.95,\ 0.025, 0.075, 0.125, 0.175, \ldots$$

will do the trick while being perfectly predictable.

To remedy the inadequacy of the RNGs satisfying only the requirement of uniform distribution, we might require `rand` to have the property that the pairs of produced numbers cover the square $[0,1]\times[0,1]$ uniformly. That means that, in the long run, the proportion of pairs falling in a patch A of the square $[0,1]\times[0,1]$ will be proportional to its area. Of course, one could continue with such requirements and ask for triples, quadruples, ...of random numbers to be uniform in $[0,1]^3$, $[0,1]^4\ldots$. The highest dimension n such that the RNG produces uniformly distributed numbers in $[0,1]^n$ is called the **order** of the RNG. A widely-used RNG called the *Mersenne Twister*, has the order of 623.

Another problem with RNGs is that the numbers produced will start to repeat after a while (this is a fact of life and finiteness of your computer's memory). The number of calls it takes for a RNG to start repeating its output is called the **period** of a RNG. You might have wondered how is it that an RNG produces a different number each time it is called, since, after all, it is only a function written in some programming language. Most often, RNGs use a hidden variable called the **random seed** which stores the last output of `rand` and is used as an (invisible) input to the function `rand` the next time it is called. If we use the same `seed` twice, the RNG will produce the same number, and so the period of the RNG is limited by the number of possible seeds. It is worth remarking that the actual random number generators usually produce a "random" integer between 0 and some large number `RAND_MAX`, and report the result normalized (divided) by `RAND_MAX` to get a number in $[0,1)$.

3.3.2 Simulation of Random Variables

Having found a random number generator good enough for our purposes (the one used by Mathematica is just fine), we might want to use it to simulate random variables with distributions different from the uniform on $[0,1]$ (coin-tosses, normal, exponential, ...). This is almost always achieved through transformations of the output of a RNG, and we will present several methods for dealing with this problem. A typical procedure (see the Box-Muller method below for an exception) works as follows: a real (deterministic) function $f:[0,1]\to\mathbb{R}$ - called **the transformation function** - is applied to `rand`. The result is a random variable whose distribution depends on the choice of f. Note that the transformation function is by no means unique. In fact, $\gamma\sim U[0,1]$, then $f(\gamma)$ and $\hat{f}(\gamma)$, where $\hat{f}(x)=f(1-x)$, have the same distribution (why?).

What follows is a list of procedures commonly used to simulate popular random variables:

1. **Discrete Random Variables** Let X have a discrete distribution given by

$$X\sim\begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix}.$$

For discrete distributions taking an infinite number of values we can always truncate at a very large n and approximate it with a distribution similar to the one of X.

We know that the probabilities $p_1, p_2, \ldots, p_n$ add-up to 1, so we define the numbers $0 = q_0 < q_1 < \cdots < q_n = 1$ by

$$q_0 = 0,\ q_1 = p_1,\ q_2 = p_1 + p_2,\ \ldots\ q_n = p_1 + p_2 + \cdots + p_n = 1.$$

To simulate our discrete random variable X, we call `rand` and then return x_1 if $0 \leq$ `rand` $< q_1$, return x_2 if $q_1 \leq$ `rand` $< q_2$, and so on. It is quite obvious that this procedure indeed simulates a random variable X. The transformation function f is in this case given by

$$f(x) = \begin{cases} x_1, & 0 \leq x < q_1 \\ x_2, & q_1 \leq x < q_2 \\ \cdots & \\ x_n, & q_{n-1} \leq x \leq 1 \end{cases}$$

2. **The Method of Inverse Functions** The basic observation in this method is that , for any continuous random variable X with the distribution function F_X, the random variable $Y = F_X(X)$ is uniformly distributed on $[0, 1]$. By inverting the distribution function F_X and applying it to Y, we recover X. Therefore, if we wish to simulate a random variable with an invertible distribution function F, we first simulate a uniform random variable on $[0, 1]$ (using `rand`) and then apply the function F^{-1} to the result. In other words, use $f = F^{-1}$ as the transformation function. Of course, this method fails if we cannot write F^{-1} in closed form.

   Example 3.6. (Exponential Distribution) Let us apply the method of inverse functions to the simulation of an exponentially distributed random variable X with parameter λ. Remember that the density f_X of X is given by

   $$f_X(x) = \lambda \exp(-\lambda x),\ x > 0,\ \text{and so } F_X(x) = 1 - \exp(-\lambda x),\ x > 0,$$

   and so $F_X^{-1}(y) = -\frac{1}{\lambda}\log(1 - y)$. Since, $1-$`rand` has the same U$[0, 1]$-distribution as `rand`, we conclude that $f(x) = -\frac{1}{\lambda}\log(x)$ works as a transformation function in this case, i.e., that

   $$-\frac{\log(\ \texttt{rand}\ )}{\lambda}$$

   has the required Exp(λ)-distribution.

   Example 3.7. (Cauchy Distribution) The Cauchy distribution is defined through its density function

   $$f_X(x) = \frac{1}{\pi}\frac{1}{(1 + x^2)}.$$

   The distribution function F_X can be determined explicitly in this example:

   $$F_X(x) = \frac{1}{\pi}\int_{-\infty}^{x} \frac{1}{(1 + x^2)}\, dx = \frac{1}{\pi}\Big(\frac{\pi}{2} + \arctan(x)\Big),\ \text{ and so } F_X^{-1}(y) = \tan\Big(\pi\big(y - \frac{1}{2}\big)\Big),$$

   yielding that $f(x) = \tan(\pi(x - \frac{1}{2}))$ is a transformation function for the Cauchy random variable, i.e., $\tan(\pi(\texttt{rand} - 0.5))$ will simulate a Cauchy random variable for you.

3. **The Box-Muller method** This method is useful for simulating normal random variables, since for them the method of inverse function fails (there is no closed-form expression for the distribution function of a standard normal). Note that this method does not fall under that category of transformation function methods as described above. You will see, though, that it is very similar in spirit. It is based on a clever trick, but the complete proof is a bit technical, so we omit it.

   Proposition 3.8. *Let γ_1 and γ_2 be independent $U[0,1]$-distributed random variables. Then the random variables*

   $$X_1 = \sqrt{-2\log(\gamma_1)}\cos(2\pi\gamma_2),\ X_2 = \sqrt{-2\log(\gamma_1)}\sin(2\pi\gamma_2)$$

   are independent and standard normal (N(0,1)).

   Therefore, in order to simulate a normal random variable with mean $\mu = 0$ and variance $\sigma^2 = 1$, we produce call the function `rand` twice to produce two random numbers `rand1` and `rand2`. The numbers

   $$X_1 = \sqrt{-2\log(\texttt{rand1})}\cos(2\pi\,\texttt{rand2}),\ X_2 = \sqrt{-2\log(\texttt{rand1})}\sin(2\pi\,\texttt{rand2})$$

   will be two independent normals. Note that it is necessary to call the function `rand` twice, but we also get two normal random numbers out of it. It is not hard to write a procedure which will produce 2 normal random numbers in this way on every second call, return one of them and store the other for the next call. In the spirit of the discussion above, the function $f = (f_1, f_2) : (0,1] \times [0,1] \to \mathbb{R}^2$ given by

   $$f_1(x,y) = \sqrt{-2\log(x)}\cos(2\pi y),\ f_2(x,y) = \sqrt{-2\log(x)}\sin(2\pi y).$$

   can be considered a transformation function in this case.

4. **Method of the Central Limit Theorem** The following algorithm is often used to simulate a normal random variable:

   (a) Simulate 12 independent uniform random variables (`rands`) - $\gamma_1, \gamma_2, \ldots, \gamma_{12}$.

   (b) Set $X = \gamma_1 + \gamma_2 + \cdots + \gamma_{12} - 6$.

   The distribution of X is very close to the distribution of a unit normal, although not exactly equal (e.g. $\mathbb{P}[X > 6] = 0$, and $\mathbb{P}[Z > 6] \neq 0$, for a true normal Z). The reason why X approximates the normal distribution well comes from the following theorem

   Theorem 3.9. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables, all having the same (square-integrable) distribution. Set $\mu = \mathbb{E}[X_1]\,(= \mathbb{E}[X_2] = \ldots)$ and $\sigma^2 = \mathrm{Var}[X_1]\,(= \mathrm{Var}[X_2] = \ldots)$. The sequence of normalized random variables*

   $$\frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}},$$

   converges to the normal random variable (in a mathematically precise sense).

The choice of exactly 12 `rands` (as opposed to 11 or 35) comes from practice: it seems to achieve satisfactory performance with relatively low computational cost. Also, the standard deviation of a $U[0,1]$ random variable is $1/\sqrt{12}$, so the denominator $\sigma\sqrt{n}$ conveniently becomes 1 for $n = 12$. It might seem a bit wasteful to use 12 calls of `rand` in order to produce one draw from the unit normal. If you try it out, you will see, however, that it is of comparable speed to the Box-Muller method described above; while Box-Muller uses computationally expensive $\cos, \sin, \sqrt{}$ and $\log$, this method uses only addition and subtraction. The final verdict of the comparison of the two methods will depend on the architecture you are running the code on, and the quality of the implementation of the functions $\cos, \sin \ldots$.

5. **Other methods** There is a number of other methods for transforming the output of `rand` into random numbers with prescribed density (rejection method, Poisson trick, ...). You can read about them in the free online copy of *Numerical recipes in C* at

   `http://www.library.cornell.edu/nr/bookcpdf.html`

3.4 Monte Carlo Integration

Having described some of the procedures and methods used for simulation of various random objects (variables, vectors, processes), we turn to an application in probability and numerical mathematics. We start off by the following version of the Law of Large Numbers which constitutes the theory behind most of the Monte Carlo applications

Theorem 3.10. (Law of Large Numbers) *Let $X_1, X_2, \ldots$ be a sequence of identically distributed random variables, and let $g : \mathbb{R} \to \mathbb{R}$ be function such that $\mu = \mathbb{E}[g(X_1)]\,(= \mathbb{E}[g(X_2)] = \ldots)$ exists. Then*

$$\frac{g(X_1) + g(X_2) + \cdots + g(X_n)}{n} \to \mu = \int_{-\infty}^{\infty} g(x) f_{X_1}(x)\, dx, \text{ as } n \to \infty.$$

The key idea of Monte Carlo integration is the following

Suppose that the quantity y we are interested in can be written as $y = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$ for some random variable X with density f_X and come function g, and that $x_1, x_2, \ldots$ are random numbers distributed according to the distribution with density f_X. Then the average

$$\frac{1}{n}(g(x_1) + g(x_2) + \cdots + g(x_n)),$$

will approximate y.

It can be shown that the accuracy of the approximation behaves like $1/\sqrt{n}$, so that you have to quadruple the number of simulations if you want to double the precision of you approximation.

Example 3.11.

1. (**numerical integration**) Let g be a function on $[0,1]$. To approximate the integral $\int_0^1 g(x)\,dx$ we can take a sequence of n (U[0,1]) random numbers $x_1, x_2, \ldots,$

$$\int_0^1 g(x)\,dx \approx \frac{g(x_1)+g(x_2)+\cdots+g(x_n)}{n},$$

because the density of $X \sim U[0,1]$ is given by

$$f_X(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

2. (**estimating probabilities**) Let Y be a random variable with the density function f_Y. If we are interested in the probability $\mathbb{P}[Y \in [a,b]]$ for some $a < b$, we simulate n draws $y_1, y_2, \ldots, y_n$ from the distribution F_Y and the required approximation is

$$\mathbb{P}[Y \in [a,b]] \approx \frac{\text{number of } y_n\text{'s falling in the interval } [a,b]}{n}.$$

One of the nicest things about the Monte-Carlo method is that even if the density of the random variable is not available, but you can simulate draws from it, you can still preform the calculation above and get the desired approximation. Of course, everything works in the same way for probabilities involving random vectors in any number of dimensions.

3. (**approximating π**)
We can devise a simple procedure for approximating $\pi \approx 3.141592$ by using the Monte-Carlo method. All we have to do is remember that π is the area of the unit disk. Therefore, $\pi/4$ equals to the portion of the area of the unit disk lying in the positive quadrant, and we can write

$$\frac{\pi}{4} = \int_0^1 \int_0^1 g(x,y)\,dxdy,$$

where

$$g(x,y) = \begin{cases} 1, & x^2+y^2 \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

So, simulate n pairs $(x_i, y_i), i = 1 \ldots n$ of uniformly distributed random numbers and count how many of them fall in the upper quarter of the unit circle, i.e. how many satisfy $x_i^2 + y_i^2 \le 1$, and divide by n. Multiply your result by 4, and you should be close to π. How close? Well, that is another story ... Experiment!

Chapter 4

The Simple Random Walk

4.1 Construction

We have defined and constructed a random walk $\{X_n\}_{n\in\mathbb{N}_0}$ in the previous lecture. Our next task is to study some of its mathematical properties. Let us give a definition of a slightly more general creature.

Definition 4.1. A sequence $\{X_n\}_{n\in\mathbb{N}_0}$ of random variables is called a **simple random walk** (with parameter $p \in (0,1)$) if

1. $X_0 = 0$,
2. $X_{n+1} - X_n$ is independent of $(X_0, X_1, \ldots, X_n)$ for all $n \in \mathbb{N}$, and
3. the random variable $X_{n+1} - X_n$ has the following distribution

$$\begin{pmatrix} -1 & 1 \\ q & p \end{pmatrix}$$

where, as usual, $q = 1 - p$.

If $p = \frac{1}{2}$, the random walk is called **symmetric**.

The adjective *simple* comes from the fact that the size of each step is fixed (equal to 1) and it is only the direction that is random. One can study more general random walks where each step comes from an arbitrary prescribed probability distribution.

Proposition 4.2. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a simple random walk with parameter p. The distribution of the random variable X_n is discrete with support $\{-n, -n+2, \ldots, n-2, n\}$, and probabilities*

$$p_l = \mathbb{P}[X_n = l] = \binom{n}{\frac{l+n}{2}} p^{(n+l)/2} q^{(n-l)/2}, \ l = -n, -n+2, \ldots, n-2, n. \tag{4.1}$$

Proof. X_n is composed of n independent steps $\xi_k = X_{k+1} - X_k$, $k = 1, \ldots, n$, each of which goes either up or down. In order to reach level l in those n steps, the number u of up-steps and the number d of downsteps must satisfy $u - d = l$ (and $u + d = n$). Therefore, $u = \frac{n+l}{2}$ and $d = \frac{n-l}{2}$.

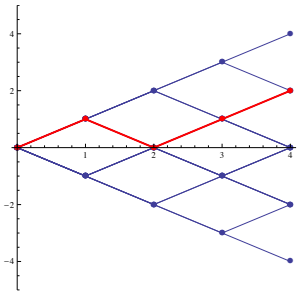
The number of ways we can choose these u up-steps from the total of n is $\binom{n}{\frac{n+l}{2}}$, which, with the fact the probability of any trajectory with exactly u up-steps is $p^u q^{n-u}$, gives the probability (4.1) above. Equivalently, we could have noticed that the random variable $\frac{n+X_n}{2}$ has the binomial $b(n,p)$-distribution. □

The proof of Proposition 4.2 uses the simple idea already hinted at in the previous lecture: view the random walk as a random trajectory in some space of trajectories, and, compute the required probability by simply counting the number of trajectories in the subset (event) you are interested in, and adding them all together, weighted by their probabilities. To prepare the ground for the future results, let C be the set of all possible trajectories:

$$C = \{(x_0, x_1, \ldots, x_n) \,:\, x_0 = 0,\ x_{k+1} - x_k = \pm 1, k \leq n-1\}.$$

You can think of the first n steps of a random walk simply as a probability distribution on the state-space C.

The figure on the right shows the superposition of all trajectories in C for $n = 4$ and a particular one - $(0,1,0,1,2)$ - in red.



4.2 The maximum

Now we know how to compute the probabilties related to the position of the random walk $\{X_n\}_{n\in\mathbb{N}_0}$ at a fixed future time n. A mathematically more interesting question can be posed about the maximum of the random walk on $\{0,1,\ldots,n\}$. A nice expression for this probability is available for the case of *symmetric* simple random walks.

Proposition 4.3. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a symmetric simple random walk, suppose $n \geq 2$, and let $M_n = \max(X_0, \ldots, X_n)$ be the maximal value of $\{X_n\}_{n\in\mathbb{N}_0}$ on the interval $0, 1, \ldots, n$. The support of M_n is $\{0,1,\ldots,n\}$ and its probability mass function is given by*

$$p_l = \mathbb{P}[M_n = l] = \binom{n}{\lfloor \frac{n+l+1}{2} \rfloor} 2^{-n},\ l = 0, \ldots, n.$$

Proof. Let us first pick a level $l \in \{0,1,\ldots,n\}$ and compute the auxilliary probability $q_l = \mathbb{P}[M_n \geq l]$ by counting the number of trajectories whose maximal level reached is at least l. Indeed, the symmetry assumption ensures that all trajectories are equally likely. More precisely, let $A_l \subset C_0(n)$ be given by

$$\begin{aligned} A_l &= \{(x_0, x_1, \ldots, x_n) \in C \,:\, \max_{k=0,\ldots,n} x_k \geq l\} \\ &= \{(x_0, x_1, \ldots, x_n) \in C \,:\, x_k \geq l, \text{ for at least one } k \in \{0,\ldots,n\}\}. \end{aligned}$$

Then $\mathbb{P}[M_n \geq l] = \frac{1}{2^n} \# A_l$, where $\# A$ denotes the number of elements in the set A. When $l = 0$, we clearly have $\mathbb{P}[M_n \geq 0] = 1$, since $X_0 = 0$.

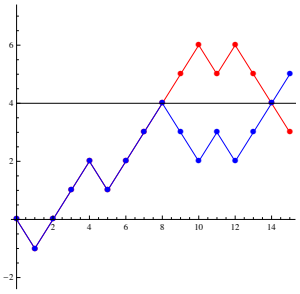
To count the number of elements in A_l, we use the following clever observation (known as the **reflection principle**):

Claim 4.4. For $l \in \mathbb{N}$, we have

$$\#A_l = 2\#\{(x_0, x_1, \ldots, x_n) : x_n > l\} + \#\{(x_0, x_1, \ldots, x_n) : x_n = l\}. \tag{4.2}$$

Proof Claim 4.4. We start by defining a bijective transformation which maps trajectories into trajectories. For a trajectory $(x_0, x_1, \ldots, x_n) \in A_l$, let $k(l) = k(l, (x_0, x_1, \ldots, x_n))$ be the smallest value of the index k such that $x_k \geq l$. In the stochastic-process-theory parlance, $k(l)$ is the **first hitting time of the set** $\{l, l+1, \ldots\}$. We know that $k(l)$ is well-defined (since we are only applying it to trajectories in A_l) and that it takes values in the set $\{1, \ldots, n\}$. With $k(l)$ at our disposal, let $(y_0, y_1, \ldots, y_n) \in C$ be a trajectory obtained from $(x_0, x_1, \ldots, x_n)$ by the following procedure:

1. do nothing until you get to $k(l)$:
   - $y_0 = x_0$,
   - $y_1 = x_1$, ...
   - $y_{k(l)} = x_{k(l)}$.
2. use the flipped values for the coin-tosses from $k(l)$ onwards:
   - $y_{k(l)+1} - y_{k(l)} = -(x_{k(l)+1} - x_{k(l)})$,
   - $y_{k(l)+2} - y_{k(l)+1} = -(x_{k(l)+2} - x_{k(l)+1})$,
     ...
   - $y_n - y_{n-1} = -(x_n - x_{n-1})$.



The picture on the right shows two trajectories: a blue one and its reflection in red, with $n = 15$, $l = 4$ and $k(l) = 8$. Graphically, $(y_0, \ldots, y_n)$ looks like $(x_0, \ldots, x_n)$ until it hits the level l, and then follows its reflection around the level l so that $y_k - l = l - x_k$, for $k \geq k(l)$. If $k(l) = n$, then $(x_0, x_1, \ldots, x_n) = (y_0, y_1, \ldots, y_n)$. It is clear that $(y_0, y_1, \ldots, y_n)$ is in C. Let us denote this transformation by

$$\Phi : A_l \to C, \ \Phi(x_0, x_1, \ldots, x_n) = (y_0, y_1, \ldots, y_n)$$

and call it the **reflection map**. The first important property of the reflexion map is that it is its own inverse: apply Φ to any $(y_0, y_1, \ldots, y_n)$ in A_l, and you will get the original $(x_0, x_1, \ldots, x_n)$. In other words $\Phi \circ \Phi = \mathrm{Id}$, i.e. Φ is an involution. It follows immediately that Φ is a bijection from A_l onto A_l.

To get to the second important property of Φ, let us split the set A_l into three parts according to the value of x_n:

1. $A_l^{>} = \{(x_0, x_1, \ldots, x_n) \in A_l : x_n > l\}$,
2. $A_l^{=} = \{(x_0, x_1, \ldots, x_n) \in A_l : x_n = l\}$, and
3. $A_l^{<} = \{(x_0, x_1, \ldots, x_n) \in A_l : x_n < l\}$,

So that

$$\Phi(A_l^{>}) = A_l^{<}, \ \Phi(A_l^{<}) = A_l^{>}, \text{ and } \Phi(A_l^{=}) = A_l^{=}.$$

We should note that, in the definition of $A_l^{>}$ and $A_l^{=}$, the a priori stipulation that $(x_0, x_1, \ldots, x_n) \in A_l$ is unncessary. Indeed, if $x_n \geq l$, you must already be in A_l. Therefore, by the bijectivity of Φ,

we have

$$\#A_l^< = \#A_l^> = \#\{(x_0, x_1, \ldots, x_n) : x_n > l\},$$

and so

$$\#A_l = 2\#\{(x_0, x_1, \ldots, x_n) : x_n > l\} + \#\{(x_0, x_1, \ldots, x_n) : x_n = l\},$$

just as we claimed. □

Now that we have (4.2), we can easily rewrite it as follows:

$$\mathbb{P}[M_n \geq l] = \mathbb{P}[X_n = l] + 2\sum_{j>l} \mathbb{P}[X_n = j] = \sum_{j>l} \mathbb{P}[X_n = j] + \sum_{j\geq l} \mathbb{P}[X_n = j].$$

Finally, we subtract $\mathbb{P}[M_n \geq l+1]$ from $\mathbb{P}[M_n \geq l]$ to get the expression for $\mathbb{P}[M_n = l]$:

$$\mathbb{P}[M_n = l] = \mathbb{P}[X_n = l+1] + \mathbb{P}[X_n = l].$$

It remains to note that only one of the probabilities $\mathbb{P}[X_n = l+1]$ and $\mathbb{P}[X_n = l]$ is non-zero, the first one if n and l have different parity and the second one otherwise. In either case the non-zero probability is given by

$$\binom{n}{\lfloor \frac{n+l+1}{2} \rfloor} 2^{-n}.$$

□

Let us use the reflection principle to solve a classical problem in combinatorics.

Example 4.5 (The Ballot Problem)**.** Suppose that two candidates, Daisy and Oscar, are running for office, and $n \in \mathbb{N}$ voters cast their ballots. Votes are counted by the same official, one by one, until all n of them have been processed (like in the old days). After each ballot is opened, the official records the number of votes each candidate has received so far. At the end, the official announces that Daisy has won by a margin of $m > 0$ votes, i.e., that Daisy got $(n+m)/2$ votes and Oscar the remaining $(n-m)/2$ votes. What is the probability that at no time during the counting has Oscar been in the lead?

We assume that the order in which the official counts the votes is completely independent of the actual votes, and that each voter chooses Daisy with probability $p \in (0,1)$ and Oscar with probability $q = 1-p$. For $k \leq n$, let X_k be the number of votes received by Daisy *minus* the number of votes received by Oscar in the first k ballots. When the $k+1$-st vote is counted, X_k either increases by 1 (if the vote was for Daisy), or decreases by 1 otherwise. The votes are independent of each other and $X_0 = 0$, so X_k, $0 \leq k \leq n$ is (the beginning of) a simple random walk. The probability of an up-step is $p \in (0,1)$, so this random walk is not necessarily symmetric. The ballot problem can now be restated as follows:

What is the probability that $X_k \geq 0$ for all $k \in \{0, \ldots, n\}$, given that $X_n = m$?

The first step towards understanding the solution is the realization that the exact value of p does not matter. Indeed, we are interested in the conditional probability $\mathbb{P}[F|G] = \mathbb{P}[F \cap G]/\mathbb{P}[G]$, where F denotes the family of all trajectories that always stay non-negative and G the family of those

that reach m at time n. Each trajectory in G has $(n+m)/2$ up-steps and $(n-m)/2$ down-steps, so its probability weight is always equal to $p^{(n+m)/2}q^{(n-m)/2}$. Therefore,

$$\mathbb{P}[F|G] = \frac{\mathbb{P}[F \cap G]}{\mathbb{P}[G]} = \frac{\#(F \cap G)\ p^{(n+m)/2}q^{(n-m)/2}}{\#G\ p^{(n+m)/2}q^{(n-m)/2}} = \frac{\#(F \cap G)}{\#G}. \tag{4.3}$$

We already know how to count the number of paths in G - it is equal to $\binom{n}{(n+m)/2}$ - so "all" that remains to be done is to count the number of paths in $G \cap F$.

The paths in $G \cap F$ form a portion of all the paths in G which don't hit the level $l = -1$, so that $\#(G \cap F) = \#G - \#H$, where H is the set of all paths which finish at m, but cross (or, at least, touch) the level $l = -1$ in the process. Can we use the reflection principle to find $\#H$? Yes, we do. In fact, you can convince yourself that the reflection of any path in H around the level $l = -1$ after its first hitting time of that level poduces a path that starts at 0 and ends at $-m-2$. Conversely, the same procedure applied to such a path yields a path in H. The number of paths from 0 to $-m-2$ is easy to count - it is equal to $\binom{n}{(n+m)/2+1}$. Putting everything together, we get

$$\mathbb{P}[F|G] = \frac{\binom{n}{k} - \binom{n}{k+1}}{\binom{n}{k}} = \frac{2k+1-n}{k+1}, \text{ where } k = \frac{n+m}{2}.$$

The last equality follows from the definition of binomial coefficients $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

The Ballot problem has a long history (going back to at least 1887) and has spurred a lot of research in combinatorics and probability. In fact, people still write research papers on some of its generalizations. When posed outside the context of probability, it is often phrased as "*in how many ways can the counting be performed ...*" (the difference being only in the normalizing factor $\binom{n}{k}$ appearing in (4.3) above). A special case $m = 0$ seems to be even more popular - the number of $2n$-step paths from 0 to 0 never going below zero is called the **Catalan number** and equals to

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

Can you derive this expression from (4.3)? If you want to test your understanding a bit further, here is an identity (called *Segner's recurrence formula*) satisfied by the Catalan numbers

$$C_n = \sum_{i=1}^{n} C_{i-1}C_{n-i}, n \in \mathbb{N}.$$

Can you prove it using the Ballot-problem interpretation?

Chapter 5

Generating functions

The path-counting method used in the previous lecture only works for computations related to the first n steps of the random walk, where n is given in advance. We will see later that most of the interesting questions do *not* fall into this category. For example, the distribution of *the time it takes* for the random walk to hit the level $l \neq 0$ is like that. There is no way to give an a-priori bound on the number of steps it will take to get to l (in fact, the expectation of this random variable is $+\infty$). To deal with a wider class of properties of random walks (and other processes), we need to develop some new mathematical tools.

5.1 Definition and first properties

The distribution of an $\mathbb{N}_0$-valued random variable X is completely determined by the sequence $\{p_n\}_{n\in\mathbb{N}_0}$ of numbers in $[0,1]$ given by

$$p_n = \mathbb{P}[X = n],\ n \in \mathbb{N}_0.$$

As a sequence of real numbers, $\{p_n\}_{n\in\mathbb{N}_0}$ can be used to construct a power series:

$$P_X(s) = \sum_{k=0}^{\infty} p_k s^k. \tag{5.1}$$

It follows from the fact that $\sum_n |p_n| \leq 1$ that the radius of convergence[1] of $\{p_n\}_{n\in\mathbb{N}_0}$ is at least equal to 1. Therefore, P_X is well defined for $s \in [-1,1]$, and, perhaps, elsewhere, too.

Definition 5.1. The function P_X given by $P_X(s) = \sum_{k=0}^{\infty} p_k s^k$ is called the **generating function** of the random variable X, or, more precisely, of its pmf $\{p_n\}_{n\in\mathbb{N}_0}$.

Before we proceed, let us find an expression for the generating functions of some of the popular $\mathbb{N}_0$-valued random variables.

Example 5.2.

[1] Remember, that the **radius of convergence** of a power series $\sum_{k=0}^{\infty} a_k x^k$ is the largest number $R \in [0,\infty]$ such that $\sum_{k=0}^{\infty} a_k x^k$ converges absolutely whenever $|x| < R$.

1. **Bernoulli** - $b(p)$ Here $p_0 = q$, $p_1 = p$, and $p_n = 0$, for $n \geq 2$. Therefore,

$$P_X(s) = ps + q.$$

2. **Binomial** - $b(n,p)$ Since $p_k = \binom{n}{k} p^k q^{n-k}$, $k = 0, \ldots, n$, we have

$$P_X(s) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} s^k = (ps+q)^n,$$

by the binomial theorem.

3. **Geometric** - $g(p)$ For $k \in \mathbb{N}_0$, $p_k = q^k p$, so that

$$P_X(s) = \sum_{k=0}^{\infty} q^k s^k p = p \sum_{k=0}^{\infty} (qs)^k = \frac{p}{1-qs}.$$

4. **Poisson** - $p(\lambda)$ Given that $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$, $k \in \mathbb{N}_0$, we have

$$P_X(s) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} s^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(s\lambda)^k}{k!} = e^{-\lambda} e^{s\lambda} = e^{\lambda(s-1)}.$$

Some of the most useful analytic properties of P_X are listed in the following proposition

Proposition 5.3. *Let X be an $\mathbb{N}_0$-valued random variable, let $\{p_n\}_{n\in\mathbb{N}_0}$ be its pmf, and let P_X be its generating function. Then*

1. *$P_X(s) = \mathbb{E}[s^X]$, $s \in [-1,1]$,*
2. *$P_X(s)$ is convex and non-decreasing with $0 \leq P_X(s) \leq 1$ for $s \in [0,1]$*
3. *$P_X(s)$ is infinitely differentiable on $(-1,1)$ with*

$$\frac{d^n}{ds^n} P_X(s) = \sum_{k=n}^{\infty} k(k-1)\ldots(k-n+1)s^{k-n}p_k, \ n \in \mathbb{N}. \tag{5.2}$$

In particular, $p_n = \frac{1}{n!}\frac{d^n}{ds^n}P_X(s)\big|_{s=0}$ and so $s \mapsto P_X(s)$ uniquely determines the sequence $\{p_n\}_{n\in\mathbb{N}_0}$.

Proof. Statement (1) follows directly from the formula $\mathbb{E}[g(X)] = \sum_{k=0}^{\infty} g(k)p_k$, applied to $g(x) = s^x$. As far as (3) is concerned, we only note that the expression (5.2) is exactly what you would get if you differentiated the expression (5.1) term by term. The rigorous proof of the fact this is allowed is beyond the scope of these notes. With (3) at our disposal, (2) follows by the fact that the first two derivatives of the function P_X are non-negative and that $P_X(1) = 1$. □

Remark 5.4.

1. If you know about moment-generating functions, you will notice that $P_X(s) = M_X(\log(s))$, for $s \in (0,1)$, where $M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$ is the moment-generating function of X.

2. Generating functions can be used with sequences $\{a_n\}_{n\in\mathbb{N}_0}$ which are not necessarily pmf's of random variables. The method is useful for any sequence $\{a_n\}_{n\in\mathbb{N}_0}$ such that the power series $\sum_{k=0}^{\infty} a_k s^k$ has a positive (non-zero) radius of convergence.

3. The name *generating function* comes from the last part of the property (3). The knowledge of P_X implies the knowledge of the whole sequence $\{p_n\}_{n\in\mathbb{N}_0}$. Put differently, P_X generates the whole distribution of X.

Remark 5.5. Note that the true radius of convergence varies from distribution to distribution. It is infinite in (1), (2) and (4), and equal to $1/q > 1$ in (3), in Example 5.2. For the distribution with pmf given by $p_k = \frac{C}{(k+1)^2}$, where $C = (\sum_{k=0}^{\infty} \frac{1}{(k+1)^2})^{-1}$, the radius of convergence is exactly equal to 1. Can you see why?

5.2 Convolution and moments

The true power of generating functions comes from the fact that they behave very well under the usual operations in probability.

Definition 5.6. Let $\{p_n\}_{n\in\mathbb{N}_0}$ and $\{q_n\}_{n\in\mathbb{N}_0}$ be two probability-mass functions. The **convolution** $p * q$ of $\{p_n\}_{n\in\mathbb{N}_0}$ and $\{q_n\}_{n\in\mathbb{N}_0}$ is the sequence $\{r_n\}_{n\in\mathbb{N}_0}$, where

$$r_n = \sum_{k=0}^{n} p_k q_{n-k}, n \in \mathbb{N}_0.$$

This abstractly-defined operation will become much clearer once we prove the following proposition:

Proposition 5.7. *Let X, Y be two* independent $\mathbb{N}_0$*-valued random variables with pmfs $\{p_n\}_{n\in\mathbb{N}_0}$ and $\{q_n\}_{n\in\mathbb{N}_0}$. Then the sum $Z = X + Y$ is also $\mathbb{N}_0$-valued and its pmf is the convolution of $\{p_n\}_{n\in\mathbb{N}_0}$ and $\{q_n\}_{n\in\mathbb{N}_0}$ in the sense of Definition 5.6.*

Proof. Clearly, Z is $\mathbb{N}_0$-valued. To obtain an expression for its pmf, we use the law of total probability:

$$\mathbb{P}[Z = n] = \sum_{k=0}^{n} \mathbb{P}[X = k]\mathbb{P}[Z = n|X = k].$$

However, $\mathbb{P}[Z = n|X = k] = \mathbb{P}[X + Y = n|X = k] = \mathbb{P}[Y = n - k|X = k] = \mathbb{P}[Y = n - k]$, where the last equality follows from independence of X and Y. Therefore,

$$\mathbb{P}[Z = n] = \sum_{k=0}^{n} \mathbb{P}[X = k]\mathbb{P}[Y = n - k] = \sum_{k=0}^{n} p_k q_{n-k}.$$

□

Corollary 5.8. *Let $\{p_n\}_{n\in\mathbb{N}_0}$ and $\{p_n\}_{n\in\mathbb{N}_0}$ be any two pmfs.*

1. *Convolution is commutative, i.e., $p * q = q * p$.*

2. *The convolution of two pmfs is a pmf, i.e.* $r_n \geq 0$, *for all* $n \in \mathbb{N}_0$ *and* $\sum_{k=0}^{\infty} r_k = 1$, *for* $r = p * q$.

Corollary 5.9. *Let* $\{p_n\}_{n\in\mathbb{N}_0}$ *and* $\{p_n\}_{n\in\mathbb{N}_0}$ *be any two pmfs, and let*

$$P(s) = \sum_{k=0}^{\infty} p_k s^k \text{ and } Q(s) = \sum_{k=0}^{\infty} q_k s^k$$

be their generating functions. Then the generating function $R(s) = \sum_{k=0}^{\infty} r_k s^k$, *of the convolution* $r = p * q$ *is given by*

$$R(s) = P(s)Q(s).$$

Equivalently, the generating function P_{X+Y} *of the sum of two independent* $\mathbb{N}_0$*-valued random variables is equal to the product*

$$P_{X+Y}(s) = P_X(s)P_Y(s),$$

of the generating functions P_X *and* P_Y *of* X *and* Y.

Example 5.10.

1. The binomial $b(n, p)$ distribution is a sum of n independent Bernoullis $b(p)$. Therefore, if we apply Corrolary 5.9 n times to the generating function $(q + ps)$ of the Bernoulli $b(p)$ distribution we immediately get that the generating function of the binomial is $(q + ps) \dots (q + ps) = (q + ps)^n$.

2. More generally, we can show that the sum of m *independent* random variables with the $b(n, p)$ distribution has a binomial distribution $b(mn, p)$. If you try to sum binomials with different values of the parameter p you will not get a binomial.

3. What is even more interesting, the following statement can be shown: Suppose that the sum Z of two independent $\mathbb{N}_0$-valued random variables X and Y is binomially distributed with parameters n and p. Then both X and Y are binomial with parameters n_X, p and n_y, p where $n_X + n_Y = n$. In other words, the only way to get a binomial as a sum of independent random variables is the trivial one.

Another useful thing about generating functions is that they make the computation of moments easier.

Proposition 5.11. *Let* $\{p_n\}_{n\in\mathbb{N}_0}$ *be a pmf of an* $\mathbb{N}_0$*-valued random variable* X *and let* $P(s)$ *be its generating function. For* $n \in \mathbb{N}$ *the following two statements are equivalent*

1. $\mathbb{E}[X^n] < \infty$,
2. $\frac{d^n P(s)}{ds^n}\Big|_{s=1}$ *exists (in the sense that the left limit* $\lim_{s\nearrow 1} \frac{d^n P(s)}{ds^n}$ *exists)*

In either case, we have

$$\mathbb{E}[X(X-1)(X-2)\dots(X-n+1)] = \frac{d^n}{ds^n} P(s)\Big|_{s=1}.$$

The quantities

$$\mathbb{E}[X], \quad \mathbb{E}[X(X-1)], \quad \mathbb{E}[X(X-1)(X-2)], \dots$$

are called **factorial moments** of the random variable X. You can get the classical moments from the factorial moments by solving a system of linear equations. It is very simple for the first few:

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X], \\
\mathbb{E}[X^2] &= \mathbb{E}[X(X-1)] + \mathbb{E}[X], \\
\mathbb{E}[X^3] &= \mathbb{E}[X(X-1)(X-2)]] + 3\mathbb{E}[X(X-1)] + \mathbb{E}[X], \dots
\end{aligned}$$

A useful identity which follows directly from the above results is the following:

$$\mathrm{Var}[X] = P''(1) + P'(1) - (P'(1))^2,$$

and it is valid if the first two derivatives of P at 1 exist.

Example 5.12. Let X be a Poisson random variable with parameter λ. Its generating function is given by

$$P_X(s) = e^{\lambda(s-1)}.$$

Therefore, $\frac{d^n}{ds^n} P_X(1) = \lambda^n$, and so, the sequence $(\mathbb{E}[X], \mathbb{E}[X(X-1)], \mathbb{E}[X(X-1)(X-2)], \dots)$ of factorial moments of X is just $(\lambda, \lambda^2, \lambda^3, \dots)$. It follows that

$$\begin{aligned}
\mathbb{E}[X] &= \lambda, \\
\mathbb{E}[X^2] &= \lambda^2 + \lambda, \ \mathrm{Var}[X] = \lambda \\
\mathbb{E}[X^3] &= \lambda^3 + 3\lambda^2 + \lambda, \dots
\end{aligned}$$

5.3 Random sums and Wald's identity

Our next application of generating function in the theory of stochastic processes deals with the so-called *random sums*. Let $\{\xi_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, and let N be a random time (a **random time** is simply an $\mathbb{N}_0 \cup \{+\infty\}$-value random variable). We can define the random variable

$$Y = \sum_{k=0}^{N} \xi_k \qquad \text{by} \qquad Y(\omega) = \begin{cases} 0, & N(\omega) = 0, \\ \sum_{k=1}^{N(\omega)} \xi_k(\omega), & N(\omega) \geq 1 \end{cases} \text{ for } \omega \in \Omega.$$

More generally, for an arbitrary stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ and a random time N (with $\mathbb{P}[N = +\infty] = 0$), we define the *random variable* X_N by $X_N(\omega) = X_{N(\omega)}(\omega)$, for $\omega \in \Omega$. When N is a constant ($N = n$), then X_N is simply equal to X_n. In general, think of X_N as a value of the stochasti process X taken at the time which is itself random. If $X_n = \sum_{k=1}^{n} \xi_k$, then $X_N = \sum_{k=1}^{N} \xi_k$.

Example 5.13. Let $\{\xi_n\}_{n\in\mathbb{N}}$ be the increments of a symmetric simple random walk (coin-tosses), and let N have the following distribution

$$N \sim \begin{pmatrix} 0 & 1 & 2 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

which is *independent* of $\{\xi_n\}_{n\in\mathbb{N}}$ (it is very important to specify the dependence structure between N and $\{\xi_n\}_{n\in\mathbb{N}}$ in this setting!). Let us compute the distribution of $Y = \sum_{k=0}^{N}\xi_k$ in this case. This is where we, typically, use the formula of total probability:

$$\begin{aligned}\mathbb{P}[Y=m] &= \mathbb{P}[Y=m|N=0]\mathbb{P}[N=0] + \mathbb{P}[Y=m|N=1]\mathbb{P}[N=1] + \mathbb{P}[Y=m|N=2]\mathbb{P}[N=2]\\ &= \mathbb{P}[\sum_{k=0}^{N}\xi_k = m|N=0]\mathbb{P}[N=0] + \mathbb{P}[\sum_{k=0}^{N}\xi_k = m|N=1]\mathbb{P}[N=1]\\ &\quad + \mathbb{P}[\sum_{k=0}^{N}\xi_k = m|N=2]\mathbb{P}[N=2]\\ &= \frac{1}{3}\left(\mathbb{P}[0=m] + \mathbb{P}[\xi_1 = m] + \mathbb{P}[\xi_1+\xi_2 = m]\right).\end{aligned}$$

When $m = 1$ (for example), we get

$$\mathbb{P}[Y=1] = \frac{0+\frac{1}{2}+0}{3} = 1/6.$$

Perform the computation for some other values of m for yourself.

What happens when N and $\{\xi_n\}_{n\in\mathbb{N}}$ are dependent? This will usually be the case in practice, as the value of the time N when we stop adding increments will typically depend on the behaviour of the sum itself.

Example 5.14. Let $\{\xi_n\}_{n\in\mathbb{N}}$ be as above - we can think of a situation where a gambler is repeatedly playing the same game in which a coin is tossed and the gambler wins a dollar if the outcome is *heads* and looses a dollar otherwise. A "smart" gambler enters the game and decides on the following tactic: *Let's see how the first game goes. If I lose, I'll play another 2 games and hopefully cover my losses, and If I win, I'll quit then and there.* The described strategy amounts to the choice of the random time N as follows:

$$N(\omega) = \begin{cases}1, & \xi_1 = 1,\\ 3, & \xi_1 = -1.\end{cases}$$

Then

$$Y(\omega) = \begin{cases}1, & \xi_1 = -1,\\ -1+\xi_2+\xi_3, & \xi_1 = 1.\end{cases}$$

Therefore,

$$\begin{aligned}\mathbb{P}[Y=1] &= \mathbb{P}[Y=1|\xi_1=1]\mathbb{P}[\xi_1=1] + \mathbb{P}[Y=1|\xi_1=-1]\mathbb{P}[\xi_1=-1]\\ &= 1\cdot\mathbb{P}[\xi_1=1] + \mathbb{P}[\xi_2+\xi_3=2]\mathbb{P}[\xi_1=-1]\\ &= \tfrac{1}{2}(1+\tfrac{1}{4}) = \tfrac{5}{8}.\end{aligned}$$

Similarly, we get $\mathbb{P}[Y=-1] = \frac{1}{4}$ and $\mathbb{P}[Y=-3] = \frac{1}{8}$. The expectation $\mathbb{E}[Y]$ is equal to $1\cdot\frac{5}{8} + (-1)\cdot\frac{1}{4} + (-3)\cdot\frac{1}{8} = 0$. This is not an accident. One of the first powerful results of the beautiful *martingale theory* states that no matter how smart a strategy you emply, you cannot beat a fair gamble.

We will return to the general (non-independent) case in the next lecture. Let us use generating functions to give a full description of the distribution of $Y = \sum_{k=0}^{N} \xi_k$ in this case.

Proposition 5.15. *Let $\{\xi_n\}_{n\in\mathbb{N}}$ be a sequence of independent $\mathbb{N}_0$-valued random variables, all of which share the same distribution with pmf $\{p_n\}_{n\in\mathbb{N}_0}$ and generating function $P_\xi(s)$. Let N be a random time independent of $\{\xi_n\}_{n\in\mathbb{N}}$. Then the generating function P_Y of the random sum $Y = \sum_{k=0}^{N} \xi_k$ is given by*

$$P_Y(s) = P_N(P_\xi(s)).$$

Proof. We use the idea from Example 5.13 and condition on possible values of N. We also use the following fact (Tonelli's theorem) without proof:

$$\text{If } \forall\, i,j \ a_{ij} \geq 0, \text{ then } \sum_{k=0}^{\infty}\sum_{i=0}^{\infty} a_{ij} = \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} a_{ij}. \tag{5.3}$$

$$\begin{aligned}
P_Y(s) &= \sum_{k=0}^{\infty} s^k \mathbb{P}[Y=k] \\
&= \sum_{k=0}^{\infty} s^k \Big(\sum_{i=0}^{\infty} \mathbb{P}[Y=k|N=i]\mathbb{P}[N=i] \Big) \\
&= \sum_{k=0}^{\infty} s^k \Big(\sum_{i=0}^{\infty} \mathbb{P}[\sum_{j=0}^{i} \xi_j = k]\mathbb{P}[N=i] \Big) \qquad \text{(by independence)} \\
&= \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} s^k \mathbb{P}[\sum_{j=0}^{i} \xi_j = k]\mathbb{P}[N=i] \qquad \text{(by Tonelli)} \\
&= \sum_{i=0}^{\infty} \mathbb{P}[N=i] \sum_{k=0}^{\infty} s^k \mathbb{P}[\sum_{j=0}^{i} \xi_j = k] \qquad \text{(by (5.3))}
\end{aligned}$$

By (iteration of) Corollary 5.9, we know that the generating function of the random variable $\sum_{j=0}^{i} \xi_j$ - which is exactly what the second sum above represents - is $(P_\xi(s))^i$. Therefore, the chain of equalities above can be continued as

$$\begin{aligned}
&= \sum_{i=0}^{\infty} \mathbb{P}[N=i](P_\xi(s))^i \\
&= P_N(P_\xi(s)).
\end{aligned}$$

□

Corollary 5.16 (Wald's Identity I)**.** *Let $\{\xi_n\}_{n\in\mathbb{N}}$ and N be as in Proposition 5.15. Suppose, also, that $\mathbb{E}[N] < \infty$ and $\mathbb{E}[\xi_1] < \infty$. Then*

$$\mathbb{E}[\sum_{k=0}^{N} \xi_k] = \mathbb{E}[N]\,\mathbb{E}[\xi_1].$$

Proof. We just apply the composition rule for derivatives to the equality $P_Y = P_N \circ P_\xi$ to get

$$P'_Y(s) = P'_N(P_\xi(s))P'_\xi(s).$$

After we let $s \nearrow 1$, we get

$$\mathbb{E}[Y] = P'_Y(1) = P'_N(P_\xi(1))P'_\xi(1) = P'_N(1)P'_\xi(1) = \mathbb{E}[N]\,\mathbb{E}[\xi_1].$$

□

Example 5.17. Every time Springfield Wildcats play in the Superbowl, their chance of winning is $p \in (0,1)$. The number of years between two Superbowls they get to play in has the Poisson distribution $p(\lambda)$, $\lambda > 0$. What is the expected number of years Y between the consequtive Superbowl wins?

Let $\{\xi_n\}_{n\in\mathbb{N}}$ be the sequence of independent $p(\lambda)$-random variables modeling the number of years between consecutive Superbowl appearances by the Wildcat. Moreover, let N be a geometric $g(p)$ random variable with success probability p. Then

$$Y = \sum_{k=0}^{N} \xi_k.$$

Indeed, every time the Wildcats lose the Superbowl, another ξ. years have to pass before they get another chance and the whole thing stops when they finally win. To compute the expectation of Y we use Corollary 5.16

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[\xi_k] = \frac{1-p}{p}\lambda.$$

Chapter 6

Random walks - advanced methods

6.1 Stopping times

The last application of generating functions dealt with sums evaluated between 0 and some random time N. An especially interesting case occurs when the value of N depends directly on the evolution of the underlying stochastic process. Even more important is the case where time's arrow is taken into account. If you think of N as the time you *stop* adding new terms to the sum, it is usually the case that you are not allowed (able) to see the values of the terms you would get if you continued adding. Think of an investor in the stock market. Her decision to stop and sell her stocks can depend only on the information available up to the moment of the decision. Otherwise, she would sell at the absolute maximum and buy at the absolute minimum, making tons of money in the process. Of course, this is not possible unless you are clairvoyant, so the mere mortals have to restrict their choices to so-called **stopping times**.

Definition 6.1. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a stochastic process. A random variable T taking values in $\mathbb{N}_0\cup\{+\infty\}$ is said to be a **stopping time** with respect to $\{X_n\}_{n\in\mathbb{N}_0}$ if for each $n\in\mathbb{N}_0$ there exists a function $G^n:\mathbb{R}^{n+1}\to\{0,1\}$ such that

$$\mathbf{1}_{\{T=n\}} = G^n(X_0, X_1, \ldots, X_n), \text{ for all } n\in\mathbb{N}_0.$$

The functions G^n are called the **decision functions**, and should be thought of as a black box which takes the values of the process $\{X_n\}_{n\in\mathbb{N}_0}$ observed up to the present point and outputs either 0 or 1. The value 0 means *keep going* and 1 means *stop*. The whole point is that the decision has to based only on the available observations and not on the future ones.

Example 6.2.

1. The simplest examples of stopping times are (non-random) *deterministic times*. Just set $T=5$ (or $T=723$ or $T=n_0$ for any $n_0\in\mathbb{N}_0\cup\{+\infty\}$), no matter what the state of the world $\omega\in\Omega$ is. The family of decision rules is easy to construct:

$$G^n(x_0,x_1,\ldots,x_n) = \begin{cases} 1, & n=n_0, \\ 0, & n\neq n_0. \end{cases}.$$

Decision functions G^n do not depend on the values of $X_0, X_1, \ldots, X_n$ *at all*. A gambler who stops gambling after 20 games, no matter of what the winnings of losses are uses such a rule.

2. Probably the most well-known examples of stopping times are *(first) hitting times*. They can be defined for general stochastic processes, but we will stick to simple random walks for the purposes of this example. So, let $X_n = \sum_{k=0}^{n} \xi_k$ be a simple random walk, and let T_l be the first time X hits the level $l \in \mathbb{N}$. More precisely, we use the following slightly non-intuitive but mathematically correct definition

$$T_l = \min\{n \in \mathbb{N}_0 \,:\, X_n = l\}.$$

The set $\{n \in \mathbb{N}_0 \,:\, X_n = l\}$ is the collection of all time-points at which X visits the level l. The earliest one - the minimum of that set - is the first hitting time of l. In states of the world $\omega \in \Omega$ in which the level l just never get reached, i.e., when $\{n \in \mathbb{N}_0 \,:\, X_n = l\}$ is an empty set, we set $T_l(\omega) = +\infty$. In order to show that T_l is indeed a stopping time, we need to construct the decision functions G^n, $n \in \mathbb{N}_0$. Let us start with $n = 0$. We would have $T_l = 0$ in the (impossible) case $X_0 = l$, so we always have $G^0(X_0) = 0$. How about $n \in \mathbb{N}$. For the value of T_l to be equal to exactly n, two things must happen:

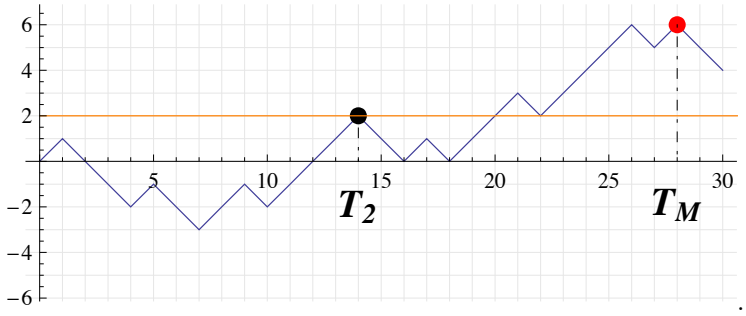
(a) $X_n = l$ (the level l must actually be hit at time n), and

(b) $X_{n-1} \neq l$, $X_{n-2} \neq l$, ..., $X_1 \neq l$, $X_0 \neq l$ (the level l has not been hit before).

Therefore,

$$G^n(x_0, x_1, \ldots, x_n) = \begin{cases} 1, & x_0 \neq l, x_1 \neq l, \ldots, x_{n-1} \neq l, x_n = l \\ 0, & \text{otherwise.} \end{cases}$$

The hitting time T_2 of the level $l = 2$ for a particular trajectory of a symmetric simple random walk is depicted below:



3. How about something that is *not* a stopping time? Let n_0 be an arbitrary time-horizon and let T_M be the last time during $0, \ldots, n_0$ that the random walk visits its maximum during $0, \ldots, n_0$ (see picture above). If you bought a stock at time $t = 0$, had to sell it some time before n_0 and had the ability to predict the future, this is one of the points you would choose to sell it at. Of course, it is impossible to decide whether $T_M = n$, for some $n \in 0, \ldots, n_0 - 1$ without the knowledge of the values of the random walk after n. More precisely, let us sketch the proof of the fact that T_M is not a stopping time. Suppose, to the contrary, that it is, and let G^n be the family of decision functions. Consider the following two trajectories: $(0, 1, 2, 3, \ldots, n-1, n)$ and $(0, 1, 2, 3, \ldots, n-1, n-2)$. The differ only in the direction of the

last step. They also differ in the fact that $T_M = n$ for the first one and $T_M = n-1$ for the second one. On the other hand, by the definition of the decision functions, we have

$$\mathbf{1}_{\{T_M=n-1\}} = G^{n-1}(X_0, \dots, X_{n-1}).$$

The right-hand side is equal for both trajectories, while the left-hand side equals to 0 for the first one and 1 for the second one. A contradiction.

6.2 Wald's identity II

Having defined the notion of a stopping time, let use try to compute something about it. The random variables $\{\xi_n\}_{n\in\mathbb{N}}$ in the statement of the theorem below are only assumed to be independent of each other and identically distributed. To make things simpler, you can think of $\{\xi_n\}_{n\in\mathbb{N}}$ as increments of a simple random walk. Before we state the main result, here is an extremely useful identity:

Proposition 6.3. *Let N be an $\mathbb{N}_0$-valued random variable. Then*

$$\mathbb{E}[N] = \sum_{k=1}^{\infty} \mathbb{P}[N \geq k].$$

Proof. Clearly, $\mathbb{P}[N \geq k] = \sum_{j\geq k} \mathbb{P}[N = j]$, so (note what happens to the indices when we switch the sums)

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbb{P}[N \geq k] &= \sum_{k=1}^{\infty}\sum_{j\geq k} \mathbb{P}[N = j] \\ &= \sum_{j=1}^{\infty}\sum_{k=1}^{j} \mathbb{P}[n = j] = \sum_{j=1}^{\infty} j\mathbb{P}[N = j] \\ &= \mathbb{E}[N].\end{aligned}$$

□

Theorem 6.4 (Wald's Identity II)**.** *Let $\{\xi_n\}_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables with $\mathbb{E}[|\xi_1|] < \infty$. Set*

$$X_n = \sum_{k=1}^{n} \xi_k, \ n \in \mathbb{N}_0.$$

If T is an $\{X_n\}_{n\in\mathbb{N}_0}$-stopping time such that $\mathbb{E}[T] < \infty$, then

$$\mathbb{E}[X_T] = \mathbb{E}[\xi_1]\mathbb{E}[T].$$

Proof. Here is another way of writing the sum $\sum_{k=1}^{T} \xi_k$:

$$\sum_{k=1}^{T} \xi_k = \sum_{k=1}^{\infty} \xi_k \mathbf{1}_{\{k\leq T\}}.$$

The idea behind it is simple: add all the values of ξ_k for $k \leq T$ and keep adding zeros (since $\xi_k \mathbf{1}_{\{k \leq T\}} = 0$ for $k > T$) after that. Taking expectation of both sides and switching $\mathbb{E}$ and $\sum$ (this can be justified, but the argument is technical and we omit it here) yields:

$$\mathbb{E}[\sum_{k=1}^{T} \xi_k] = \sum_{k=1}^{\infty} \mathbb{E}[\mathbf{1}_{\{k \leq T\}} \xi_k]. \tag{6.1}$$

Let us examine the term $\mathbb{E}[\xi_k \mathbf{1}_{\{k \leq T\}}]$ in some detail. We first note that

$$\mathbf{1}_{\{k \leq T\}} = 1 - \mathbf{1}_{\{k > T\}} = 1 - \mathbf{1}_{\{k-1 \geq T\}} = 1 - \sum_{j=0}^{k-1} \mathbf{1}_{\{T=j\}}.$$

Therefore,

$$\mathbb{E}[\xi_k \mathbf{1}_{\{k \leq T\}}] = \mathbb{E}[\xi_k] - \sum_{j=0}^{k-1} \mathbb{E}[\xi_k \mathbf{1}_{\{T=j\}}].$$

By the assumption that T is a stopping time, the indicator $\mathbf{1}_{\{T=j\}}$ can be represented as $\mathbf{1}_{\{T=j\}} = G^j(X_0, \ldots, X_j)$, and, because each X_i is just a sum of the increments, we can actually write $\mathbf{1}_{\{T=j\}}$ as a function of $\xi_1, \ldots, \xi_j$ only - say $\mathbf{1}_{\{T=j\}} = H^j(\xi_1, \ldots, \xi_j)$. By the independence of $(\xi_1, \ldots, \xi_j)$ from ξ_k (because $j < k$) we have

$$\mathbb{E}[\xi_k \mathbf{1}_{\{T=j\}}] = \mathbb{E}[\xi_k H^j(\xi_1, \ldots, \xi_j)] = \mathbb{E}[\xi_k]\mathbb{E}[H^j(\xi_1, \ldots, \xi_j)] = \mathbb{E}[\xi_k]\mathbb{E}[\mathbf{1}_{\{T=j\}}] = \mathbb{E}[\xi_k]\mathbb{P}[T = j].$$

Therefore,

$$\mathbb{E}[\xi_k \mathbf{1}_{\{k \leq T\}}] = \mathbb{E}[\xi_k] - \sum_{j=0}^{k-1} \mathbb{E}[\xi_k]\mathbb{P}[T = j] = \mathbb{E}[\xi_k]\mathbb{P}[T \geq k] = \mathbb{E}[\xi_1]\mathbb{P}[T \geq k],$$

where the last equality follows from the fact that all ξ_k have the same distribution.

Going back to (6.1), we get

$$\mathbb{E}[X_T] = \mathbb{E}[\sum_{k=1}^{T} \xi_k] = \sum_{k=1}^{\infty} \mathbb{E}[\xi_1]\mathbb{P}[T \geq k] = \mathbb{E}[\xi_1] \sum_{k=1}^{\infty} \mathbb{P}[T \geq k] = \mathbb{E}[\xi_1]\mathbb{E}[T],$$

where we use Proposition 6.3 for the last equality. □

Example 6.5 (Gambler's ruin problem)**.** . A gambler start with $x \in \mathbb{N}$ dollars and repeatedly plays a game in which he wins a dollar with probability $\frac{1}{2}$ and loses a dollar with probability $\frac{1}{2}$. He decides to stop when one of the following two things happens:

1. he goes bankrupt, i.e., his wealth hits 0, or
2. he makes enough money, i.e., his wealth reaches some level $a > x$.

The classical "Gambler's ruin" problem asks the following question: what is the probability that the gambler will make a dollars before he goes bankrupt?

Gambler's wealth $\{W_n\}_{n\in\mathbb{N}}$ is modelled by a simple random walk starting from x, whose increments $\xi_k = W_k - W_{k-1}$ are coin-tosses. Then $W_n = x + X_n$, where $X_n = \sum_{k=0}^{n} \xi_k$, $n \in \mathbb{N}_0$. Let T be the time the gambler stops. We can represent T in two different (but equivalent) ways. On the one hand, we can think of T as the smaller of the two hitting times T_{-x} and T_{a-x} of the levels $-x$ and $a-x$ for the random walk $\{X_n\}_{n\in\mathbb{N}_0}$ (remember that $W_n = x + X_n$, so these two correspond to the hitting times for the process $\{W_n\}_{n\in\mathbb{N}_0}$ of the levels 0 and a). On the other hand, we can think of T as the first hitting time of the two-element *set* $\{-x, a-x\}$ for the process $\{X_n\}_{n\in\mathbb{N}_0}$. In either case, it is quite clear that T is a stopping time (can you write down the decision functions?). We will see later that the probability that the gambler's wealth will remain strictly between 0 and a forever is zero, so $\mathbb{P}[T < \infty] = 1$. Moreover, we will prove that $\mathbb{E}[T] < \infty$.

What can we say about the random variable X_T - the gambler's wealth (minus x) at the *random* time T? Clearly, it is either equal to $-x$ or to $a-x$, and the probabilities p_0 and p_a with which it takes these values are exactly what we are after in this problem. We know that, since there are no other values X_T can take, we must have $p_0 + p_a = 1$. Wald's identity gives us the second equation for p_0 and p_a:

$$\mathbb{E}[X_T] = \mathbb{E}[\xi_1]\mathbb{E}[T] = 0 \cdot \mathbb{E}[T] = 0,$$

so

$$0 = \mathbb{E}[X_T] = p_0(-x) + p_a(a-x).$$

These two linear equations with two unknowns yield

$$p_0 = \frac{a-x}{a},\ p_a = \frac{x}{a}.$$

It is remarkable that the two probabilities are proportional to the amounts of money the gambler needs to make (lose) in the two outcomes. The situation is different when $p \neq \frac{1}{2}$.

6.3 The distribution of the first hitting time T_1

6.3.1 A recursive formula

Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a simple random walk, with the probability p of stepping up. Let $T_1 = \min\{n \in \mathbb{N}_0 : X_n = 1\}$ be the first hitting time of level $l = 1$, and let $\{p_n\}_{n\in\mathbb{N}_0}$ be its pmf, i.e., $p_n = \mathbb{P}[T_1 = n]$, $n \in \mathbb{N}_0$. The goal of this section is to use the powerful generating-function methods to find $\{p_n\}_{n\in\mathbb{N}_0}$. You cannot get from 0 to 1 in an even number of steps, so $p_{2n} = 0$, $n \in \mathbb{N}_0$. Also, $p_1 = p$ - you just have to go up on the first step. What about $n > 1$? In order to go from 0 to 1 in $n > 1$ steps (and not before!) the first step needs to be *down* and then you need to climb up from -1 to 1 in $n-1$ steps. Climbing from -1 to 1 is exactly the same as climbing from -1 to 0 and then climbing from 0 to 1. If it took j steps to go from -1 to 0 it will have to take $n-1-j$ steps to go from 1 to 2, where j can be anything from 1 to $n-2$, in order to finish the job in exactly $n-1$

steps. So, in formulas, we have

$$\mathbb{P}[T_1 = n] =$$
$$= q \sum_{j=1}^{n-2} \mathbb{P}[\text{“exactly } j \text{ steps to first hit 0 from } -1\text{” } \textit{and} \text{ “exactly } n-1-j \text{ steps to first hit 1 from 0”}]. \tag{6.2}$$

Taking j steps from -1 to 0 is exactly the same as taking j steps from 0 to 1, so

$$\mathbb{P}[\text{“exactly } j \text{ steps to first hit 0 from } -1\text{”}] = \mathbb{P}[T_1 = j] = p_j.$$

By the same token,

$$\mathbb{P}[\text{“exactly } n-1-j \text{ steps to first hit 1 from 0”}] = \mathbb{P}[T_1 = n-1-j] = p_{n-1-j}.$$

Finally, I claim that the two events are independent of each other[1]. Indeed, once we have reached 0, the future increments of the random walk behave exactly the same as the increments of a fresh random walk starting from zero - they are independent of everything. Equivalently, a knowledge of everything that happened until the moment the random walk hit 0 for the first time does not change our perception (and estimation) of what is going to happen later - in this case the likelihood of hitting 1 in exactly $n-1-j$ steps. This property is called the *regeneration property* or the *strong Lévy property* of random walks. More precisely (but still not entirely precise), we can make the following claim:

> Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a simple random walk and let T be any $\mathbb{N}_0$-valued stopping time. Define the process $\{Y_n\}_{n\in\mathbb{N}_0}$ by $Y_n = X_{T+n} - X_T$. Then $\{Y_n\}_{n\in\mathbb{N}_0}$ is also a simple random walk, and it is independent of X up to T.

In order to check your understanding, try to convince yourself that the requirement that T be a stopping time is necessary - find an example of a random time T which is not a stopping time where the statement above fails.

We can go back to the distribution of the hitting time T_1, and use our newly-found independence together with (6.2) to obtain the following recursion

$$p_n = q \sum_{j=1}^{n-2} p_j p_{n-j-1}, n > 1, \qquad p_0 = 0,\ p_1 = p. \tag{6.3}$$

6.3.2 Generating-function approach

This is where generating functions step in. We will use (6.3) to derive an equation for the generating function $P(s) = \sum_{k=0}^{\infty} p_k s^k$. The sum on the right-hand side of (6.3) looks a little bit like a convolution, so let us compare it to the following expansion of the square $P(s)^2$:

$$P(s)^2 = \sum_{k=0}^{\infty} (\sum_{i=0}^{k} p_i p_{k-i}) s^k.$$

[1] A demanding reader will object at this point, since my definitions of the two events are somewhat loose. I beg forgiveness.

The inner sum $\sum_{i=0}^{k} p_i p_{k-i}$ needs to be split into several parts to get an expression which matches (6.3):

$$\sum_{i=0}^{k} p_i p_{k-i} = \underbrace{p_0 p_k}_{0} + \sum_{i=1}^{k-1} p_i p_{k-i} + \underbrace{p_k p_0}_{0} = \sum_{i=1}^{(k+1)-2} p_i p_{(k+1)-i-1} = q^{-1} p_{k+1}, \quad k \geq 2.$$

Therefore, since the coefficients of $P(s)^2$ start at s^2, we have

$$qsP(s)^2 = qs \sum_{k=2}^{\infty} q^{-1} p_{k+1} s^k = \sum_{k=2}^{\infty} p_{k+1} s^{k+1} = P(s) - ps,$$

which is nothing but a quadratic equation for P. It admits two solutions (for each s):

$$P(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2qs}.$$

One of the two solutions is always greater than 1 in absolute value, so it cannot correspond to a value of a generating function, so we conclude that

$$P(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2qs}, \text{ for } |s| \leq \frac{1}{2\sqrt{pq}}.$$

It remains to extract the information about $\{p_n\}_{n \in \mathbb{N}_0}$ from P. The obvious way to do it is to compute higher and higher derivatives of P and $s = 1$. There is an easier way, though. The square root appearing in the formula for P is an expression of the form $(1 + x)^{1/2}$ and the (generalized) binomial formula can be used:

$$(1 + x)^{\alpha} = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^{\alpha}, \text{ where } \binom{\alpha}{k} = \frac{\alpha(\alpha - 1) \dots (\alpha - k + 1)}{k!}, \ k \in \mathbb{N}, \alpha \in \mathbb{R}.$$

Therefore,

$$P(s) = \frac{1}{2qs} - \frac{1}{2qs} \sum_{k=0}^{\infty} \frac{1}{2q} \binom{1/2}{k} (-4pqs^2)^k = \sum_{k=1}^{\infty} s^{2k-1} \frac{1}{2p} (4pq)^k (-1)^{k-1} \binom{1/2}{k}, \tag{6.4}$$

and so

$$p_{2k-1} = \frac{1}{2q} (4pq)^k (-1)^{k-1} \binom{1/2}{k}, p_{2k-2} = 0, \ k \in \mathbb{N}.$$

This expression can be simplified a bit further: the formula for $\binom{1/2}{k}$ evaluates to:

$$\binom{1/2}{k} = (-1)^{k-1} \frac{1}{4^{k-1}} \frac{1}{k} \binom{2k-3}{k-2}.$$

Thus,

$$p_{2k-1} = p^k q^{k-1} \frac{2}{k} \binom{2k-3}{k-2}.$$

This last expression might remind you of something related to the reflection principle. And it is! Can you derive the formula for p_{2k-1} from the reflection principle? How would you deal with the fact that the random walk here is not symmetric?

6.3.3 Do we actually hit 1 sooner or later?

What happens if we try to evaluate $P(1)$? We should get 1, right? In fact, what we get is the following:

$$P(1) = \frac{1-\sqrt{1-4pq}}{2q} = \frac{1-|p-q|}{2q} = \begin{cases} 1, & p \geq \frac{1}{2} \\ \frac{p}{q}, & p < \frac{1}{2} \end{cases}$$

Clearly, $P(1) < 1$ when $p < q$. The explanation is simple - the random walk may fail to hit the level 1 *at all*, if $p < q$. In that case $P(1) = \sum_{k=0}^{\infty} p_k = \mathbb{P}[T_1 < \infty] < 1$, or, equivalently, $\mathbb{P}[T_1 = +\infty] > 0$. It is remarkable that if $p = \frac{1}{2}$, the random walk *will* always hit 1 sooner or later, but this does not need to happen if $p < \frac{1}{2}$. What we have here is an example of a phenomenon known as *criticality*: many physical systems exhibit qualitatively different behavior depending on whether the value of certain parameter p lies above or below certain *critical value* $p = p_c$.

6.3.4 Expected time until we hit 1?

Another question that generating functions can help up answer is the following one: how long, on average, do we need to wait before 1 is hit? When $p < \frac{1}{2}$ $\mathbb{P}[T_1 = +\infty] > 0$, so we can immediately conclude that $\mathbb{E}[T_1] = +\infty$, by definition. The case $p \geq \frac{1}{2}$ is more interesting. Following the recipe from the lecture on generating functions, we compute the derivative of $P(s)$ and get

$$P'(s) = \frac{2p}{\sqrt{1-4pqs^2}} - \frac{1-\sqrt{1-4pqs^2}}{2qs^2}.$$

When $p = \frac{1}{2}$, we get

$$\lim_{s \nearrow 1} P'(s) = \lim_{s \nearrow 1} \left(\frac{1}{\sqrt{1-s^2}} - \frac{1-\sqrt{1-s^2}}{s^2} \right) = +\infty,$$

and conclude that $\mathbb{E}[T_1] = +\infty$.

For $p > \frac{1}{2}$, the situation is less severe:

$$\lim_{s \nearrow 1} P'(s) = \frac{1}{p-q}.$$

We can summarize the situation in the following table

	$\mathbb{P}[T_1 < \infty]$	$\mathbb{E}[T_1]$
$p < \frac{1}{2}$	$\frac{p}{q}$	$+\infty$
$p = \frac{1}{2}$	1	$+\infty$
$p > \frac{1}{2}$	1	$\frac{1}{p-q}$

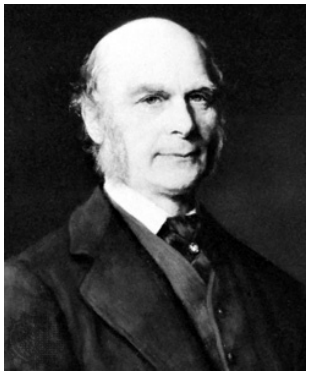
Chapter 7

Branching processes

7.1 A bit of history

In the mid 19th century several aristocratic families in Victorian England realized that their family names could become extinct. Was it just unfounded paranoia, or did something real prompt them to come to this conclusion? They decided to ask around, and Sir Francis Galton (a "polymath, anthropologist, eugenicist, tropical explorer, geographer, inventor, meteorologist, protogeneticist, psychometrician and statistician" and half-cousin of Charles Darwin) posed the following question (1873, *Educational Times*):

> *How many male children (on average) must each generation of a family have in order for the family name to continue in perpetuity?*



Sir Francis Galton

The first complete answer came from Reverend Henry William Watson soon after, and the two wrote a joint paper entitled *One the probability of extinction of families* in 1874. By the end of this lecture, you will be able to give a precise answer to Galton's question.

7.2 A mathematical model

The model proposed by Watson was the following:

1. A population starts with one individual at time $n = 0$: $Z_0 = 1$.
2. After one unit of time (at time $n = 1$) the sole individual produces Z_1 identical clones of itself and dies. Z_1 is an $\mathbb{N}_0$-valued random variable.
3. (a) If Z_1 happens to be equal to 0 the population is dead and nothing happens at any future time $n \geq 2$.

(b) If $Z_1 > 0$, a unit of time later, each of Z_1 individuals gives birth to a random number of children and dies. The first one has $Z_{1,1}$ children, the second one $Z_{1,2}$ children, etc. The last, Z_1^{th} one, gives birth to Z_{1,Z_1} children. We assume that the distribution of the number of children is the same for each individual in every generation and independent of either the number of individuals in the generation and of the number of children the others have. This distribution, shared by all $Z_{n,i}$ and Z_1, is called the *offspring distribution*. The total number of individuals in the second generation is now

$$Z_2 = \sum_{k=1}^{Z_1} Z_{1,k}.$$

(c) The third, fourth, etc. generations are produced in the same way. If it ever happens that $Z_n = 0$, for some n, then $Z_m = 0$ for all $m \geq n$ - the population is extinct. Otherwise,

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

Definition 7.1. A stochastic process with the properties described in (1), (2) and (3) above is called a **(simple) branching process**.

The mechanism that produces the next generation from the present one can differ from application to application. It is the offspring distribution alone that determines the evolution of a branching process. With this new formalism, we can pose Galton's question more precisely:

Under what conditions on the offspring distribution will the process $\{Z_n\}_{n\in\mathbb{N}_0}$ never go extinct, i.e., when does

$$\mathbb{P}[Z_n \geq 1 \text{ for all } n \in \mathbb{N}_0] = 1 \tag{7.1}$$

hold?

7.3 Construction and simulation of branching processes

Before we answer Galton's question, let us figure out how to simulate a branching process, for a given offspring distribution $\{p_n\}_{n\in\mathbb{N}_0}$ ($p_k = \mathbb{P}[Z_1 = k]$). The distribution $\{p_n\}_{n\in\mathbb{N}_0}$ is $\mathbb{N}_0$-valued - we have learned how to simulate such distributions in Lecture 3. We can, therefore, assume that a transformation function F is known, i.e., that the random variable $\eta = F(\gamma)$ is $\mathbb{N}_0$-valued with pmf $\{p_n\}_{n\in\mathbb{N}_0}$, where $\gamma \sim U[0,1]$.

Some time ago we assumed that a probability space with a sequence $\{\gamma_n\}_{n\in\mathbb{N}_0}$ of independent $U[0,1]$ random variables is given. We think of $\{\gamma_n\}_{n\in\mathbb{N}_0}$ as a sequence of random numbers produced by a computer. Let us first apply the function F to each member of $\{\gamma_n\}_{n\in\mathbb{N}_0}$ to obtain an independent sequence $\{\eta_n\}_{n\in\mathbb{N}_0}$ of $\mathbb{N}_0$-valued random variables with pmf $\{p_n\}_{n\in\mathbb{N}_0}$. In the case of a simple random walk, we would be done at this point - an accumulation of the first n elements of $\{\eta_n\}_{n\in\mathbb{N}_0}$ would give you the value X_n of the random walk at time n. Branching processes are a bit more complicated; the increment $Z_{n+1} - Z_n$ depends on Z_n: the more

individuals in a generation, the more offspring they will produce. In other words, we need a black box with two inputs - "randomness" and Z_n - which will produce Z_{n+1}. What do we mean by "randomness"? Ideally, we would need exactly Z_n (unused) elements of $\{\eta_n\}_{n\in\mathbb{N}_0}$ to simulate the number of children for each of Z_n members of generation n. This is exactly how one would do it in practice: given the size Z_n of generation n, one would draw Z_n simulations from the distribution $\{p_n\}_{n\in\mathbb{N}_0}$, and sum up the results to get Z_{n+1}. Mathematically, it is easier to be more wasteful. The sequence $\{\eta_n\}_{n\in\mathbb{N}_0}$ can be rearranged into a double sequence[1] $\{Z_{n,i}\}_{n\in\mathbb{N}_0,i\in\mathbb{N}}$. In words, instead of one sequence of independent random variables with pmf $\{p_n\}_{n\in\mathbb{N}_0}$, we have a sequence of sequences. Such an abundance allows us to feed the whole "row" $\{Z_{n,i}\}_{i\in\mathbb{N}}$ into the black box which produces Z_{n+1} from Z_n. You can think of $Z_{n,i}$ as the number of children the i^{th} individual in the n^{th} generation would have had he been born. The black box uses only the first Z_n elements of $\{Z_{n,i}\}_{i\in\mathbb{N}}$ and discards the rest:

$$Z_0 = 1, Z_{n+1} = \sum_{i=1}^{Z_n} Z_{n,i},$$

where all $\{Z_{n,i}\}_{n\in\mathbb{N}_0,i\in\mathbb{N}}$ are independent of each other and have the same distribution with pmf $\{p_n\}_{n\in\mathbb{N}_0}$. Once we learn a bit more about the probabilistic structure of $\{Z_n\}_{n\in\mathbb{N}_0}$, we will describe another way to simulate it.

7.4 A generating-function approach

Having defined and constructed a branching process with offspring distribution $\{Z_n\}_{n\in\mathbb{N}_0}$, let us analyze its probabilistic structure. The first question the needs to be answered is the following: *What is the distribution of Z_n, for $n \in \mathbb{N}_0$?* It is clear that Z_n must be $\mathbb{N}_0$-valued, so its distribution is completely described by its pmf, which is, in turn, completely determined by its generating function. While an explicit expression for the pmf of Z_n may not be available, its generating function can always be computed:

Proposition 7.2. *Let $\{Z_n\}_{n\in\mathbb{N}_0}$ be a branching process, and let the generating function of its offspring distribution $\{p_n\}_{n\in\mathbb{N}_0}$ be given by $P(s)$. Then the generating function of Z_n is the n-fold composition of P with itself, i.e.,*

$$P_{Z_n}(s) = \underbrace{P(P(\dots P(s)\dots))}_{n\ P\text{'s}},\ \text{for } n \geq 1.$$

Proof. For $n = 1$, the distribution of Z_1 is exactly $\{p_n\}_{n\in\mathbb{N}_0}$, so $P_{Z_1} = P(s)$. Suppose that the statement of the proposition holds for some $n \in \mathbb{N}$. Then

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

can be viewed as a random sum of Z_n independent random variables with pmf $\{p_n\}_{n\in\mathbb{N}_0}$, where the number of summands Z_n is independent of $\{Z_{n,i}\}_{i\in\mathbb{N}}$. By Proposition 5.16 in the lecture on

[1] Can you find a one-to-one and onto mapping from $\mathbb{N}$ into $\mathbb{N}\times\mathbb{N}$?

generating functions, we have seen that the generating function $P_{Z_{n+1}}$ of Z_{n+1} is a composition of the generating function $P(s)$ of each of the summands and the generating function P_{Z_n} of the random time Z_n. Therefore,

$$P_{Z_{n+1}}(s) = P_{Z_n}(P(s)) = \underbrace{P(P(\dots P(P(s))\dots))}_{n+1 \text{ P's}},$$

and the full statement of the Proposition follows by induction. □

Let us use Proposition 7.2 in some simple examples.

Example 7.3. Let $\{Z_n\}_{n\in\mathbb{N}_0}$ be a branching process with offspring distribution $\{p_n\}_{n\in\mathbb{N}_0}$. In the first three examples no randomness occurs and the population growth can be described exactly. In the other examples, more insteresting things happen.

1. $p_0 = 1$, $p_n = 0$, $n \in \mathbb{N}$:
   In this case $Z_0 = 1$ and $Z_n = 0$ for all $n \in \mathbb{N}$. This infertile population dies after the first generation.

2. $p_0 = 0, p_1 = 1, p_n = 0$, $n \geq 2$:
   Each individual produces exactly one child before he/she dies. The population size is always 1: $Z_n = 1$, $n \in \mathbb{N}_0$.

3. $p_0 = 0, p_1 = 0, \dots, p_k = 1, p_n = 0$, $n \geq k$, for some $k \geq 2$:
   Here, there are k kids per individual, so the population grows exponentially: $P(s) = s^k$, so $P_{Z_n}(s) = ((\dots(s^k)^k\dots)^k)^k = s^{k^n}$. Therefore, $Z_n = k^n$, for $n \in \mathbb{N}$.

4. $p_0 = p, p_1 = q = 1 - p, p_n = 0$, $n \geq 2$:
   Each individual tosses a (a biased) coin and has one child of the outcome is *heads* or dies childless if the outcome is *tails*. The generating function of the offspring distribution is $P(s) = p + qs$. Therefore,

   $$P_{Z_n}(s) = \underbrace{(p + q(p + q(p + q(\dots(p + qs)))))}_{n \text{ pairs of parentheses}}.$$

   The expression above can be simplified considerably. One needs to realize two things:

   (a) After all the products above are expanded, the resulting expression must be of the form $A + Bs$, for some A, B. If you inspect the expression for P_{Z_n} even more closely, you will see that the coefficient B next to s is just q^n.

   (b) P_{Z_n} is a generating function of a probability distribution, so $A + B = 1$.

   Therefore,

   $$P_{Z_n}(s) = (1 - q^n) + q^n s.$$

   Of course, the value of Z_n will be equal to 1 if and only if all of the coin-tosses of its ancestors turned out to be *heads*. The probability of that event is q^n. So we didn't need Proposition 7.2 after all.

This example can be interpreted alternatively as follows. Each individual has exactly one child, but its gender is determined at random - male with probability q and female with probability p. Assuming that all females change their last name when they marry, and assuming that all of them marry, Z_n is just the number of individuals carrying the family name after n generations.

5. $p_0 = p^2, p_1 = 2pq, p_2 = q^2, p_n = 0$, $n \geq 3$:
In this case each individual has exactly two children and their gender is female with probability p and male with probability q, independently of each other. The generating function P of the offspring distribution $\{p_n\}_{n\in\mathbb{N}}$ is given by $P(s) = (p+qs)^2$. Then

$$P_{Z_n} = \underbrace{(p+q(p+q(\dots p+qs)^2\dots)^2)^2}_{n \text{ pairs of parentheses}}.$$

Unlike the example above, it is not so easy to simplify the above expression.

Proposition 7.2 can be used to compute the mean and variance of the population size Z_n, for $n \in \mathbb{N}$.

Proposition 7.4. *Let $\{p_n\}_{n\in\mathbb{N}_0}$ be a pmf of the offpsring distribution of a branching process $\{Z_n\}_{n\in\mathbb{N}_0}$. If $\{p_n\}_{n\in\mathbb{N}_0}$ admits an expectation, i.e., if*

$$\mu = \sum_{k=0}^{\infty} kp_k < \infty,$$

then

$$\mathbb{E}[Z_n] = \mu^n. \tag{7.2}$$

If the variance of $\{p_n\}_{n\in\mathbb{N}_0}$ is also finite, i.e., if

$$\sigma^2 = \sum_{k=0}^{\infty}(k-\mu)^2 p_k < \infty,$$

then

$$\mathrm{Var}[Z_n] = \sigma^2\mu^n(1+\mu+\mu^2+\cdots+\mu^n) = \begin{cases}\sigma^2\mu^n\frac{1-\mu^{n+1}}{1-\mu}, & \mu \neq 1,\\ \sigma^2(n+1), & \mu = 1\end{cases} \tag{7.3}$$

Proof. Since the distribution of Z_1 is just $\{p_n\}_{n\in\mathbb{N}_0}$, it is clear that $\mathbb{E}[Z_1] = \mu$ and $\mathrm{Var}[Z_1] = \sigma^2$. We proceed by induction and assume that the formulas (7.2) and (7.3) hold for $n \in \mathbb{N}$. By Proposition 7.2, the generating function $P_{Z_{n+1}}$ is given as a composition $P_{Z_{n+1}}(s) = P_{Z_n}(P(s))$. Therefore, if we use the identity $\mathbb{E}[Z_{n+1}] = P'_{Z_{n+1}}(1)$, we get

$$P'_{Z_{n+1}}(1) = P'_{Z_n}(P(1))P'(1) = P'_{Z_n}(1)P'(1) = \mathbb{E}[Z_n]\mathbb{E}[Z_1] = \mu^n\mu = \mu^{n+1}.$$

A similar (but more complicated and less illuminating) argument can be used to establish (7.3). □

7.5 Extinction probability

We now turn to the central question (the one posed by Galton). We define **extinction** to be the following event:

$$E = \{\omega \in \Omega \,:\, Z_n(\omega) = 0 \text{ for some } n \in \mathbb{N}\}.$$

It is the property of the branching process that $Z_m = 0$ for all $m \geq n$ whenever $Z_n = 0$. Therefore, we can write E as an *increasing* union of sets E_n, where

$$E_n = \{\omega \in \Omega \,:\, Z_n(\omega) = 0\}.$$

Therefore, the sequence $\{\mathbb{P}[E_n]\}_{n\in\mathbb{N}}$ is non-decreasing and "continuity of probability" (see the very first lecture) implies that

$$\mathbb{P}[E] = \lim_{n\in\mathbb{N}} \mathbb{P}[E_n].$$

The number $\mathbb{P}[E]$ is called the **extinction probability**. Using generating functions, and, in particular, the fact that $\mathbb{P}[E_n] = \mathbb{P}[Z_n = 0] = P_{Z_n}(0)$ we get

$$\mathbb{P}[E] = \lim_{n\in\mathbb{N}} P_{Z_n}(0) = \lim_{n\in\mathbb{N}} \underbrace{P(P(\dots P(0)\dots))}_{n\ \mathrm{P's}}.$$

It is amazing that this probability can be computed, even if the explicit form of the generating function P_{Z_n} is not known.

Proposition 7.5. *The extinction probability $p = \mathbb{P}[E]$ is the smallest non-negative solution of the equation*

$$x = P(x), \text{ called the } \textbf{\textit{extinction equation}},$$

where P is the generating function of the offspring distribution.

Proof. Let us show first that $p = \mathbb{P}[E]$ is *a* solution of the equation $x = P(x)$. Indeed, P is a continuous function, so $P(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} P(x_n)$ for every convergent sequence $\{x_n\}_{n\in\mathbb{N}_0}$ in $[0,1]$ with $x_n \to x_\infty$. Let us take a particular sequence given by

$$x_n = \underbrace{P(P(\dots P(0)\dots))}_{n\ \mathrm{P's}}.$$

Then

1. $p = \mathbb{P}[E] = \lim_{n\in\mathbb{N}} x_n$, and
2. $P(x_n) = x_{n+1}$.

Therefore,

$$p = \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} P(x_n) = P(\lim_{n\to\infty} x_n) = P(p),$$

and so p solves the equation $P(x) = x$.

The fact that $p = \mathbb{P}[E]$ is the *smallest* solution of $x = P(x)$ on $[0,1]$ is a bit trickier to get. Let p' be another solution of $x = P(x)$ on $[0,1]$. Since $0 \leq p'$ and P is a non-decreasiing function, we have

$$P(0) \leq P(p') = p'.$$

We can apply the function P to both sides of the inequality above to get

$$P(P(0)) \leq P(P(p')) = P(p') = p'.$$

Continuing in the same way we get

$$P[E_n] = \underbrace{P(P(\dots P(0)\dots))}_{n\ \mathrm{P's}} \leq p',$$

we get $p = \mathbb{P}[E] = \lim_{n\in\mathbb{N}} \mathbb{P}[E_n] \leq \lim_{n\in\mathbb{N}} p' = p'$, so p is not larger then any other solution p' of $x = P(x)$. □

Example 7.6. Let us compute extinction probabilities in the cases from Example 7.3.

1. $p_0 = 1$, $p_n = 0$, $n \in \mathbb{N}$:
   No need to use any theorems. $\mathbb{P}[E] = 1$ in this case.

2. $p_0 = 0, p_1 = 1, p_n = 0$, $n \geq 2$:
   Like above, the situation is clear - $\mathbb{P}[E] = 0$.

3. $p_0 = 0, p_1 = 0, \dots, p_k = 1, p_n = 0$, $n \geq k$, for some $k \geq 2$:
   No extinction here - $\mathbb{P}[E] = 0$.

4. $p_0 = p, p_1 = q = 1 - p, p_n = 0$, $n \geq 2$:
   Since $P(s) = p + qs$, the extinction equation is $s = p + qs$. If $p = 0$, the only solution is $s = 0$, so no extinction occurs. If $p > 0$, the only solution is $s = 1$ - the extinction is guaranteed. It is interesting to note the jump in the extinction probability as p changes from 0 to a positive number.

5. $p_0 = p^2, p_1 = 2pq, p_2 = q^2, p_n = 0$, $n \geq 3$:
   Here $P(s) = (p + qs)^2$, so the extinction equation reads

   $$s = (p + qs)^2.$$

   This is a quadratic in s and its solutions are $s_1 = 1$ and $s_2 = \frac{p^2}{q^2}$, if we assume that $q > 0$. When $p < q$, the smaller of the two is s_2. When $p \geq q$, $s = 1$ is the smallest solution. Therefore

   $$\mathbb{P}[E] = \min(1, \frac{p^2}{q^2}).$$

Chapter 8

Markov Chains

8.1 The Markov property

Simply put, a stochastic process has the **Markov property** if its future evolution depends only on its current position, not on how it got there. Here is a more precise, mathematical, definition. It will be assumed throughout this course that any stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ takes values in a countable set S - the state space. Usually, S will be either $\mathbb{N}_0$ (as in the case of branching processes) or $\mathbb{Z}$ (random walks). Sometimes, a more general, *but still countable*, state space S will be needed. A generic element of S will be denoted by i or j.

Definition 8.1. A stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ talking values in a countable state space S is called a **Markov chain** (or said to have the **Markov property**) if

$$\mathbb{P}[X_{n+1}=i_{n+1}|X_n=i_n,X_{n-1}=i_{n-1},\dots,X_1=i_1,X_0=i_0]=\mathbb{P}[X_{n+1}=i_{n+1}|X_n=i_n], \qquad (8.1)$$

for all $n\in\mathbb{N}_0$, all $i_0,i_1,\dots,i_n,i_{n+1}\in S$, whenever the two conditional probabilities are well-defined, i.e., when $\mathbb{P}[X_n=i_n,\dots,X_1=i_1,X_0=i_0]>0$.

The Markov property is typically checked in the following way: one computes the left-hand side of (8.1) and shows that its value does not depend on $i_{n-1},i_{n-2},\dots,i_1,i_0$ (why is that enough?). A condition $\mathbb{P}[X_n=i_n,\dots,X_0=i_0]>0$ will be assumed (without explicit mention) every time we write a conditional expression like to one in (8.1).

All chains in this course will be *homogeneous*, i.e., the conditional probabilities $\mathbb{P}[X_{n+1}=j|X_n=i]$ will not depend on the current time $n\in\mathbb{N}_0$, i.e., $\mathbb{P}[X_{n+1}=j|X_n=i]=\mathbb{P}[X_{m+1}=j|X_m=i]$, for $m,n\in\mathbb{N}_0$.

Markov chains are (relatively) easy to work with because the Markov property allows us to compute all the probabilities, expectations, etc. we might be interested in by using only two ingredients.

1. **Initial probability** $\boldsymbol{a}^{(0)}=\{a_i^{(0)}:i\in S\}$, $a_i^{(0)}=\mathbb{P}[X_0=i]$ - the initial probability distribution of the process, and
2. **Transition probabilities** $p_{ij}=\mathbb{P}[X_{n+1}=j|X_n=i]$ - the mechanism that the process uses to jump around.

Indeed, if one knows all $a_i^{(0)}$ and all p_{ij}, and wants to compute a joint distribution $\mathbb{P}[X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]$, one needs to use the definition of conditional probability and the Markov property several times (the *multiplication theorem* from your elementary probability course) to get

$$\begin{aligned}\mathbb{P}[X_n = i_n, \dots, X_0 = i_0] &= \mathbb{P}[X_n = i_n | X_{n-1} = i_{n-1}, \dots, X_0 = i_0]\mathbb{P}[X_{n-1} = i_{n-1}, \dots, X_0 = i_0] \\ &= \mathbb{P}[X_n = i_n | X_{n-1} = i_{n-1}]\mathbb{P}[X_{n-1} = i_{n-1}, \dots, X_0 = i_0] \\ &= p_{i_{n-1}i_n}\mathbb{P}[X_{n-1} = i_{n-1}, \dots, X_0 = i_0]\end{aligned}$$

Repeating the same procedure, we get

$$\mathbb{P}[X_n = i_n, \dots, X_0 = i_0] = p_{i_{n-1}i_n} \times p_{i_{n-2}i_{n-1}} \times \cdots \times p_{i_0 i_1} \times a_{i_0}^{(0)}.$$

When S is finite, there is no loss of generality in assuming that $S = \{1, 2, \dots, n\}$, and then we usually organize the entries of $a^{(0)}$ into a row vector

$$\boldsymbol{a}^{(0)} = (a_1^{(0)}, a_2^{(0)}, \dots, a_n^{(0)}),$$

and the transition probabilities p_{ij} into a square matrix $\mathbf{P}$, where

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

In the general case (S possibly infinite), one can still use the vector and matrix notation as before, but it becomes quite clumsy in the general case. For example, if $S = \mathbb{Z}$, $\mathbf{P}$ is an infinite matrix

$$\mathbf{P} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \cdots & p_{-1\,-1} & p_{-1\,0} & p_{-1\,1} & \cdots \\ \cdots & p_{0\,-1} & p_{0\,0} & p_{0\,1} & \cdots \\ \cdots & p_{1\,-1} & p_{1\,0} & p_{1\,1} & \cdots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

8.2 Examples

Here are some examples of Markov chains - for each one we write down the transition matrix. The initial distribution is sometimes left unspecified because it does not really change anything.

1. Random walks Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a simple random walk. Let us show that it indeed has the Markov property (8.1). Remember, first, that $X_n = \sum_{k=1}^n \xi_k$, where ξ_k are *independent* coin-tosses. For a choice of $i_0, \dots, i_{n+1}$ (such that $i_0 = 0$ and $i_{k+1} - i_k = \pm 1$) we have

$$\begin{aligned}&\mathbb{P}[X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0] \\ =&\mathbb{P}[X_{n+1} - X_n = i_{n+1} - i_n | X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0] \\ =&\mathbb{P}[\xi_{n+1} = i_{n+1} - i_n | X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0] \\ =&\mathbb{P}[\xi_{n+1} = i_{n+1} - i_n],\end{aligned}$$

where the last equality follows from the fact that the increment ξ_{n+1} is idependent of the previous increments, and, therefore, of the values of $X_1, X_2, \ldots, X_n$. The last line above does not depend on $i_{n-1}, \ldots, i_1, i_0$, so X indeed has the Markov property.

The state space S of $\{X_n\}_{n\in\mathbb{N}_0}$ is the set $\mathbb{Z}$ of all integers, and the initial distribution $\boldsymbol{a}^{(0)}$ is very simple: we start at 0 with probability 1 (so that $a_0^{(0)} = 1$ and $a_i^{(0)} = 0$, for $i \neq 0$.). The transition probabilities are simple to write down

$$p_{ij} = \begin{cases} p, & j = i+1 \\ q, & j = i-1 \\ 0, & \text{otherwise.} \end{cases}$$

These can be written down in an infinite matrix,

$$\mathbf{P} = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \iddots & \\ \ldots & 0 & P & 0 & 0 & 0 & \ldots & \\ \ldots & q & 0 & p & 0 & 0 & \ldots & \\ \ldots & 0 & q & 0 & p & 0 & \ldots & \\ \ldots & 0 & 0 & q & 0 & p & \ldots & \\ \ldots & 0 & 0 & 0 & q & 0 & \ldots & \\ \ldots & 0 & 0 & 0 & 0 & q & \ldots & \\ \iddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

but it does not help our understanding much.

2. Branching processes Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a simple Branching process with the branching distribution $\{p_n\}_{n\in\mathbb{N}_0}$. As you surely remember, it is constructed as follows: $X_0 = 1$ and $X_{n+1} = \sum_{k=1}^{X_n} X_{n,k}$, where $\{X_{n,k}\}_{n\in\mathbb{N}_0,k\in\mathbb{N}}$ is a family of independent random variables with distribution $\{p_n\}_{n\in\mathbb{N}_0}$. It is now not very difficult to show that $\{X_n\}_{n\in\mathbb{N}_0}$ is a Markov chain

$$\begin{aligned} &\mathbb{P}[X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0] \\ =&\mathbb{P}[\sum_{k=1}^{X_n} X_{n,k} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0] \\ =&\mathbb{P}[\sum_{k=1}^{i_n} X_{n,k} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0] \\ =&\mathbb{P}[\sum_{k=1}^{i_n} X_{n,k} = i_{n+1}], \end{aligned}$$

where, just like in the random-walk case, the last equality follows from the fact that the random variables $X_{n,k}$, $k \in \mathbb{N}$ are independent of all $X_{m,k}$, $m < n$, $k \in \mathbb{N}$. In particular, they are independent of $X_n, X_{n-1}, \ldots, X_1, X_0$, which are obtained as combinations of $X_{m,k}$, $m < n$, $k \in \mathbb{N}$. The computation above also reveals the structure of the transition probabilities, p_{ij}, $i, j \in S = \mathbb{N}_0$:

$$p_{ij} = \mathbb{P}[\sum_{k=1}^{i} X_{n,k} = j].$$

There is little we can do to make the expression above more explicit, but we can remember generating functions and write $P_i(s) = \sum_{j=0}^{\infty} p_{ij} s^j$ (remember that each row of the transition matrix is a probability distribution). Thus, $P_i(s) = (P(s))^i$ (why?), where $P(s) = \sum_{k=0}^{\infty} p_k s^k$ is the generating function of the branching probability. Analogously to the random walk case, we have

$$a_i^{(0)} = \begin{cases} 1, & i = 1, \\ 0, & i \neq 1. \end{cases}$$

3. Gambler's ruin In Gambler's ruin, a gambler starts with $\$x$, where $0 \leq x \leq a \in \mathbb{N}$ and in each play wins a dollar (with probability $p \in (0,1)$) and loses a dollar (with probability $q = 1 - p$). When the gambler reaches either 0 or a, the game stops. The transition probabilities are similar to those of a random walk, but differ from them at the boundaries 0 and a. The state space is finite $S = \{0, 1, \ldots, a\}$ and the matrix $\mathbf{P}$ is, therefore, given by

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ q & 0 & p & 0 & \ldots & 0 & 0 & 0 \\ 0 & q & 0 & p & \ldots & 0 & 0 & 0 \\ 0 & 0 & q & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & p & 0 \\ 0 & 0 & 0 & 0 & \ldots & q & 0 & p \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix}$$

The initial distribution is deterministic:

$$a_i^{(0)} = \begin{cases} 1, & i = x, \\ 0, & i \neq 1. \end{cases}$$

4. Regime Switching Consider a system with two different states; think about a simple weather forcast (rain/no rain), high/low water level in a reservoire, high/low volatility regime in a financial market, high/low level of economic growth, etc. Suppose that the states are called 0 and 1 and the probabilities p_{01} and p_{10} of switching states are given. The probabilities $p_{00} = 1 - p_{01}$ and $p_{11} = 1 - p_{10}$ correspond to the system staying in the same state. The transition matrix for this Markov with $S = \{0, 1\}$ is

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11}. \end{bmatrix}$$

When p_{01} and p_{10} are large (close to 1) the system nervously jumps between the two states. When they are small, there are long periods of stability (staying in the same state).

5. Deterministically monotone Markov chain A stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ with state space $S = \mathbb{N}_0$ such that $X_n = n$ for $n \in \mathbb{N}_0$ (no randomness here) is called Deterministically monotone Markov chain (DMMC). The transition matrix looks something like this

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

6. Not a Markov chain Consider a frog jumping from a lotus leaf to a lotus leaf on in a small forest pond. Suppose that there are N leaves so that the state space can be described as $S = \{1, 2, \dots, N\}$. The frog starts on leaf 1 at time $n = 0$, and jumps around in the following fashion: at time 0 it chooses any leaf except for the one it is currently sitting on (with equal probability) and then jumps to it. At time $n > 0$, it chooses any leaf other than the one it is sitting on *and the one it visited immediately before* (with equal probability) and jumps to it. The position $\{X_n\}_{n\in\mathbb{N}_0}$ of the frog is not a Markov chain. Indeed, we have

$$\mathbb{P}[X_3 = 1 | X_2 = 2, X_1 = 3] = \frac{1}{N-2}, \text{ while } \mathbb{P}[X_3 = 1 | X_2 = 2, X_1 = 1] = 0.$$

A more dramatic version of this example would be the one where the frog remembers all the leaves it had visited before, and only chooses among the remaining ones for the next jump.

7. Making a non-Markov chain into a Markov chain How can we turn the process of Example 6. into a Markov chain. Obviously, the problem is that the frog has to remember the number of the leaf it came from in order to decide where to jump next. The way out is to make this information a part of the state. In other words, we need to change the state space. Instead of just $S = \{1, 2, \dots, N\}$, we set $S = \{(i_1, i_2) : i, j \in \{1, 2, \dots N\}\}$. In words, the state of the process will now contain not only the number of the current leaf (i.e., i) but also the numer of the leaf we came from (i.e., j). There is a bit of freedom with the initial state, but we simply assume that we start from $(1, 1)$. Starting from the state (i, j), the frog can jump to any state of the form (k, i), $k \neq i, j$ (with equal probabilities). Note that some states will never be visited (like (i, i) for $i \neq 1$), so we could have reduced the state space a little bit right from the start.

8. A more complicated example Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a simple symmetric random walk. The absolute-value process $Y_n = |X_n|$, $n \in \mathbb{N}_0$, is also a Markov chain. This processes is sometimes called the **reflected random walk**.

In order to establish the Markov property, we let $i_0, \dots, i_{n+1}$ be non-negative integers with $i_{k+1} - i_k = \pm 1$ for all $0 \leq k \leq n$ (the state space is $S = \mathbb{N}_0$). We need to show that the conditional probability

$$\mathbb{P}[|X_{n+1}| = i_{n+1} \Big| \, |X_n| = i_n, \dots, |X_0| = i_0] \tag{8.2}$$

does not depend on $i_{n-1}, \dots, i_0$. We write

$$\begin{aligned}\mathbb{P}[|X_{n+1}| = i_{n+1} \Big| \, |X_n| = i_n, \dots, |X_0| = i_0] = \mathbb{P}[X_{n+1} = i_{n+1} \Big| \, |X_n| = i_n, \dots, |X_0| = i_0] \\ + \mathbb{P}[X_{n+1} = -i_{n+1} \Big| \, |X_n| = i_n, \dots, |X_0| = i_0],\end{aligned} \tag{8.3}$$

and concentrate on the first conditional probability on the right-hand side, asumming that $i_n > 0$ (the case $i_n = 0$ is easier and is left to the reader who needs practice). Let us use the law of total probability (see Problem 1 in HW 6) with $A_1 = \{X_n = i_n\}$, $A_2 = \{X_n \neq i_n\}$ and $B = \{|X_n| = i_n, \dots, |X_0| = |i_0|\}$. Since $A_2 \cap B = \{X_n = -i_n\} \cap B$, we have

$$\begin{aligned}\mathbb{P}[X_{n+1} = i_{n+1}|B] &= \mathbb{P}[X_{n+1} = i_{n+1}|B \cap A_1] \, \mathbb{P}[A_1|B] \\ &\quad + \mathbb{P}[X_{n+1} = i_{n+1}|B \cap A_2] \, \mathbb{P}[A_2|B] \\ &= \mathbb{P}[X_{n+1} = i_{n+1}|B \cap \{X_n = i_n\}]\mathbb{P}[X_n = i_n|B] \\ &\quad + \mathbb{P}[X_{n+1} = i_{n+1}|B \cap \{X_n = -i_n\}]\mathbb{P}[X_n = -i_n|B]\end{aligned}$$

Conditionally on $X_n = i_n$, the probability that $X_{n+1} = i_{n+1}$ does not depend on the extra information B might contain. Therefore

$$\begin{aligned}\mathbb{P}[X_{n+1} = i_{n+1}|B \cap \{X_n = i_n\}] = \mathbb{P}[X_{n+1} = i_{n+1}|X_n = i_n], \text{ and, similarly,} \\ \mathbb{P}[X_{n+1} = i_{n+1}|B \cap \{X_n = -i_n\}] = \mathbb{P}[X_{n+1} = i_{n+1}|X_n = -i_n].\end{aligned}$$

How about the term $\mathbb{P}[X_n = i_n|B]$ (with $\mathbb{P}[X_n = -i_n|B]$ being completely analogous)? If we show that

$$\mathbb{P}[X_n = i_n|B] = \mathbb{P}[X_n = i_n \Big| \, |X_n| = i_n], \tag{8.4}$$

we would be making great progress. There is a rigorous way of doing this, but it is quite technical and not very illuminating. The idea is simple, though: for every path $(0, X_1(\omega), \dots, X_n(\omega))$, the flipped path $(0, -X_1(\omega), \dots, -X_n(\omega))$ is *equally likely* and gives exactly the same sequence of absolute values. Therefore, the knowledge of B does not permit us to distinguish between them. In particular, for every (possible) sequence of absolute values $|x_0| = i_0, |x_1| = i_1, \dots, |x_0| = i_n$ there are as many actual paths $(x_0, x_1, \dots, x_n)$ that end up in i_n as those that end up in $-i_n$. Therefore,

$$\mathbb{P}[X_n = i_n|B] = \mathbb{P}[X_n = i_n \Big| \, |X_n| = i_n] = \tfrac{1}{2}.$$

Similarly, $\mathbb{P}[X_n = -i_n|B] = \frac{1}{2}$.

Going back to the initial expression (8.4), we have

$$\mathbb{P}[X_{n+1} = i_{n+1}|B] = \tfrac{1}{2}\mathbb{P}[X_{n+1} = i_{n+1}|B \cap \{X_n = i_n\}] + \tfrac{1}{2}\mathbb{P}[X_{n+1} = i_{n+1}|B \cap \{X_n = -i_n\}].$$

Given that $i_n > 0$, the first conditional probability above equals to $\mathbb{P}[X_{n+1} = i_n]$ because of the Markov property; as far as X_{n+1} is concerned, the information $B \cap \{X_n = i_n\}$ is the same as just $\{X_n = i_n\}$). The second conditional probability is 0: it is impossible for X_{n+1} to be equal to $i_{n+1} > 0$ if $X_n = -i_n < 0$. Therefore, $\mathbb{P}[X_{n+1} = i_{n+1}|B] = 1/4$. An essential identical argument shows that $\mathbb{P}[X_{n+1} = -i_{n+1}|B] = 1/4$. Therefore $\mathbb{P}[|X_{n+1}| = i_{n+1}|B] = \frac{1}{2}$ - which is independent

of $i_{n-1}, \dots, i_1, i_0$ (it looks like it is also independent of i_n but it is not; this probability is equal to 0 unless $i_{n+1} - i_n = \pm 1$).

9. A more realistic example In a game of tennis, the scoring system is as follows: both players (let us call them Amélie and Björn) start with the score of 0. Each time Amélie wins a point, her score moves a step up in the following hierarchy

$$0 \mapsto 15 \mapsto 30 \mapsto 40.$$

Once Amélie reaches 40 and scores a point, three things can happen:

1. if Björn's score is 30 or less, Amélie wins the game.
2. if Björn's score is 40, Amélie's score moves up to "*advantage*", and
3. if Björn's score is "*advantage*", nothing happens to Amélie's score, but Björn's score falls back to 40.

Finally, if Amélie's score is "*advantage*" and she wins a point, she wins the game. The situation is entirely symmetric for Björn. We suppose that the probability that Amélie wins each point is $p \in (0, 1)$, independently of the current score.

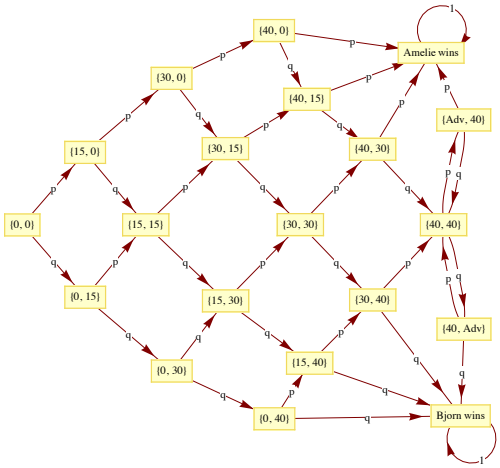


Figure. Markov chains with a finite number of states are usually represented by directed graphs (like the one in the figure above). The nodes are states, two states i, j are linked by a (directed) edge if the transition probability p_{ij} is non-zero, and the number p_{ij} is writen above the link. If $p_{ij} = 0$, no edge is drawn.

A situation like this is a typical example of a Markov chain in an applied setting. What are the states of the process? We obviously need to know both players' scores and we also need to know if one of the players has won the game. Therefore, a possible state space is the following:

$$S = \Big\{ \text{"Amelie wins"}, \text{"Bjorn wins"}, (0,0), (0,15), (0,30), (0,40), (15,0), (15,15), (15,30), (15,40), \\ (30,0), (30,15), (30,30), (30,40), (40,0), (40,15), (40,30), (40,40), (40,\text{Adv}), (\text{Adv},40) \Big\} \tag{8.5}$$

It is not hard to assign probabilities to transitions between states. Once we reach either "Amelie wins" or "Bjorn wins" the game stops. We can assume that the chain remains in that state forever, i.e., the state is absorbing. The initial distribution is quite simple - we aways start from the same state $(0, 0)$, so that $a^{(0)}_{(0,0)} = 1$ and $a^{(0)}_i = 0$ for all $i \in S \setminus \{0\}$.

How about the transition matrix? When the number of states is big ($\#S = 20$ in this case), transition matrices are useful in computer memory, but not so much on paper. Just for the fun of it, here is the transition matrix for our game-of-tennis chain, with the states ordered as in (8.5):

$$\mathbf{P} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & p & 0 & 0 & 0 \\
0 & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 \\
p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & 0 & 0 \\
p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & 0 \\
p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & p \\
0 & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 \\
p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0
\end{bmatrix}$$

Here is a question we will learn how to answer later:

Question 8.2. Does the structure of a game of tennis make is easier or harder for the better player to win? In other words, if you had to play against Roger Federer (I am rudely assuming that he is better than you), would you have a better chance of winning if you only played a point, or if you actually played the whole game?

8.3 Chapman-Kolmogorov relations

The transition probabilities p_{ij}, $i, j \in S$ tell us how a Markov chain jumps from a state to a state in one step. How about several steps, i.e., how does one compute the the probabilities like $\mathbb{P}[X_{k+n} = j | X_k = i]$, $n \in \mathbb{N}$? Since we are assuming that all of our chains are homogeneous (transition probabilities do not change with time), this probability does not depend on the time k, and we set

$$p_{ij}^{(n)} = \mathbb{P}[X_{k+n} = j | X_k = i] = \mathbb{P}[X_n = j | X_0 = i].$$

It is sometimes useful to have a more compact notation for this, last, conditional probability, so we write

$$\mathbb{P}_i[A] = \mathbb{P}[A | X_0 = i], \text{ for any event } A.$$

Therefore,

$$p_{ij}^{(n)} = \mathbb{P}_i[X_n = j].$$

For $n = 0$, we clearly have

$$p_{ij}^{(0)} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Once se have defined the **multi-step transition probabilities** $p_{ij}^{(n)}$, $i, j \in S$, $n \in \mathbb{N}_0$, we need to be able to compute them. This computation is central in various applications of Markov chains: they relate the small-time (one-step) behavior which is usually easy to observe and model to a long-time (multi-step) behavior which is really of interest. Before we state the main result in this direction, let us remember how matrices are multiplied. When A and B are $n \times n$ matrices, the product $C = AB$ is also an $n \times n$ matrix and its ij-entry C_{ij} is given as

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

There is nothing special about finiteness in the above definition. If A and B were infinite matrices $A = (A_{ij})_{i,j \in S}$, $B = (B_{ij})_{i,j \in S}$ for some countable set S, the same procedure could be used to define $C = AB$. Indeed, C will also be an "$S \times S$"-matrix and

$$C_{ij} = \sum_{k \in S} A_{ik} B_{kj},$$

as long as the (infinite) sum above converges absolutely. In the case of a typical transition matrix $\mathbf{P}$, convergence will not be a problem since $\mathbf{P}$ is a **stochastic matrix**, i.e., it has the following two properties (why?):

1. $p_{ij} \geq 0$, for all $i, j \in S$, and
2. $\sum_{j \in S} p_{ij} = 1$, for all $i \in S$ (in particular, $p_{ij} \in [0, 1]$, for all i, j).

When $\mathbf{P} = (p_{ij})_{i,j \in S}$ and $\mathbf{P}' = (p'_{ij})_{i,j \in S}$ are two $S \times S$-stochastic matrices, the series $\sum_{k \in S} p_{ik} p'_{kj}$ converges absolutely since $0 \leq p'_{kj} \leq 1$ for all $k, j \in S$ and so

$$\sum_{k \in S} \left| p_{ik} p'_{kj} \right| \leq \sum_{k \in S} p_{ik} \leq 1, \text{ for all } i, j \in S.$$

Moreover, a product C of two stochastic matrices A and B is always a stochastic matrix: the entries of C are clearly positive and (by Tonelli's theorem)

$$\sum_{j \in S} C_{ij} = \sum_{j \in S} \sum_{k \in S} A_{ik} B_{kj} = \sum_{k \in S} \sum_{j \in S} A_{ik} B_{kj} = \sum_{k \in S} A_{ik} \underbrace{\sum_{j \in S} B_{kj}}_{1} = \sum_{k \in S} A_{ik} = 1.$$

Proposition 8.3. *Let $\mathbf{P}^n$ be the n-th (matrix) power of the transition matrix $\mathbf{P}$. Then $p_{ij}^{(n)} = (\mathbf{P}^n)_{ij}$, for $i, j \in S$.*

Proof. We proceed by induction. For $n = 1$ the statement follows directly from the definition of the matrix $\mathbf{P}$. Supposing that $p_{ij}^{(n)} = (\mathbf{P}^n)_{ij}$ for all i, j, we have

$$\begin{aligned} p_{ij}^{(n+1)} &= \mathbb{P}[X_{n+1} = j | X_0 = i] \\ &= \sum_{k \in S} \mathbb{P}[X_1 = k | X_0 = i] \mathbb{P}[X_{n+1} = j | X_0 = i, X_1 = k] \\ &= \sum_{k \in S} \mathbb{P}[X_1 = k | X_0 = i] \mathbb{P}[X_{n+1} = j | X_1 = k] \\ &= \sum_{k \in S} \mathbb{P}[X_1 = k | X_0 = i] \mathbb{P}[X_n = j | X_0 = k] \\ &= \sum_{k \in S} p_{ik} p_{kj}^{(n)}. \end{aligned}$$

where the second equality follows from the law of total probability, the third one from the Markov property, and the fourth one from homogeneity. The last sum above is nothing but the expression for the matrix product of $\mathbf{P}$ and $\mathbf{P}^n$, and so we have proven the induction step. □

Using Proposition 8.3, we can write a simple expression for the distribution of the random variable X_n, for $n \in \mathbb{N}_0$. Remember that the initial dsitribution (the distribution of X_0) is denoted by $\boldsymbol{a}^{(0)} = (a_i^{(0)})_{i \in S}$. Analogously, we define the vector $\boldsymbol{a}^{(n)} = (a_i^{(n)})_{i \in S}$ by

$$a_i^{(n)} = \mathbb{P}[X_n = i], \ i \in S.$$

Using the law of total probability, we have

$$a_i^{(n)} = \mathbb{P}[X_n = i] = \sum_{k \in S} \mathbb{P}[X_0 = k] \mathbb{P}[X_n = i | X_0 = k] = \sum_{k \in S} a_k^{(0)} p_{ki}^{(n)}.$$

We usually interpret $\boldsymbol{a}^{(0)}$ as a (row) vector, so the above relationship can be expressed using vector-matrix multiplication

$$\boldsymbol{a}^{(n)} = \boldsymbol{a}^{(0)} \mathbf{P}^n.$$

The following corollary shows a simple, yet fundamental, relationship between different multi-step transition probabilities $p_{ij}^{(n)}$.

Corollary 8.4 (Chapman-Kolmogorov relations)**.** *For $n, m \in \mathbb{N}_0$ and $i, j \in S$ we have*

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)}.$$

Proof. The statement follows directly from the matrix equality

$$\mathbf{P}^{m+n} = \mathbf{P}^m \mathbf{P}^n.$$

□

It is usually difficult to compute $\mathbf{P}^n$ for a general transition matrix $\mathbf{P}$ and a large n. We will see later that it will be easier to find the limiting values $\lim_{n\to\infty} p_{ij}^{(n)}$. In the mean-time, here is a simple example where this can be done by hand

Example 8.5. In the setting of a Regime Switching chain (Example 4.), let us write a for p_{01} and b for p_{10} to simplify the notation, so that the transition matrix looks like this:

$$\mathbf{P} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

The characteristic equation $\det(\lambda I - \mathbf{P}) = 0$ of the matrix $\mathbf{P}$ is

$$\begin{aligned} 0 = \det(\lambda I - \mathbf{P}) &= \begin{vmatrix} \lambda - 1 + a & -a \\ -b & \lambda - 1 + b \end{vmatrix} = ((\lambda-1)+a)((\lambda-1)+b) - ab \\ &= (\lambda - 1)(\lambda - (1-a-b)). \end{aligned}$$

The eigenvalues are, therefore, $\lambda_1 = 1$ and $\lambda_2 = 1-a-b$. The eigenvectors are $v_1 = \binom{1}{1}$ and $v_2 = \binom{a}{-b}$, so that with $V = \begin{bmatrix} 1 & a \\ 1 & -b \end{bmatrix}$ and $D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & (1-a-b) \end{bmatrix}$ we have

$$\mathbf{P}V = VD, \text{ i.e., } \mathbf{P} = VDV^{-1}.$$

This representation is very useful for taking matrix powers:

$$\mathbf{P}^n = (VDV^{-1})(VDV^{-1})\dots(VDV^{-1}) = VD^nV^{-1} = V \begin{bmatrix} 1 & 0 \\ 0 & (1-a-b)^n \end{bmatrix} V^{-1}$$

Assuming $a+b>0$ (i.e., $\mathbf{P} \neq \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$), we have $V^{-1} = \frac{1}{a+b}\begin{bmatrix} b & a \\ 1 & -1 \end{bmatrix}$, and

$$\begin{aligned} \mathbf{P}^n = VD^nV^{-1} &= \begin{bmatrix} 1 & a \\ 1 & -b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-a-b)^n \end{bmatrix} \frac{1}{a+b}\begin{bmatrix} b & a \\ 1 & -1 \end{bmatrix} \\ &= \frac{1}{a+b}\begin{bmatrix} b & a \\ b & a \end{bmatrix} + \frac{(1-a-b)^n}{a+b}\begin{bmatrix} a & -a \\ b & -b \end{bmatrix} \\ &= \begin{bmatrix} \frac{b}{a+b} + (1-a-b)^n \frac{a}{a+b} & \frac{a}{a+b} - (1-a-b)^n \frac{a}{a+b} \\ \frac{b}{a+b} + (1-a-b)^n \frac{b}{a+b} & \frac{a}{a+b} - (1-a-b)^n \frac{b}{a+b} \end{bmatrix} \end{aligned}$$

The expression for $\mathbf{P}^n$ above tells us a lot about the structure of the multi-step probabilities $p_{ij}^{(n)}$ for large n. Note that the second matrix on the right-hand side above comes multiplied by $(1-a-b)^n$ which tends to 0 as $n\to\infty$, unless we are in the uninteresting situation $a=b=0$ or $(a=b=1)$. Therefore,

$$\mathbf{P}^n \sim \frac{1}{a+b}\begin{bmatrix} a & b \\ a & b \end{bmatrix} \text{ for large } n.$$

The fact that the rows of the right-hand side above are equal points to the fact that, for large n, $p_{ij}^{(n)}$ does not depend (much) on the initial state i. In other words, this Markov chain forgets its initial condition after a long period of time. This is a rule more than an exception, and we will study such phenomena in the following lectures.

Chapter 9

The "Stochastics" package

`Stochastics` is a *Mathematica* package (a collection of functions) for various calculations, simulation and visualization of Markov chains. Here is a short user's guide.

9.1 Installation

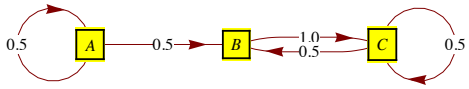
You can download the file `Stochastics.m` from the course web-site. It needs to be put in a directory/folder on your system that Mathematica knows about. The easiest way to do this is to type `$Path` in Mathematica. The output will be a list of folders. You simply copy the file `Stochastics.m` in any of the folders listed. Restart Mathematica, and that is it.

Every time you want to use functions from `Stochastics`, you issue the command `<<Stochastics‘` (note that the last symbol ‘ is not an apostrophe; it can be found above the Tab key on your keyboard).

If you want to get information about the syntax and usage of a certain command (say `BuildChain`), just type `?BuildChain`, and Mathematica will display a short help paragraph.

9.2 Building Chains

The first thing you need to learn is to tell Mathematica about the structure of the Markov Chain you want to analyze. As you know, two ingredients are needed for that: the initial distribution, and the set of transition probabilities. Since the transition matrix is usually quite sparse (has many zero elements), it is typically faster to specify those probabilities as a list of transitions of the form `{FromState,ToState,probability}`. That is exactly how `Stochastics` is set up. In order to store the information about a Markov Chain that Mathematica can do further computations with, you issue the command `MyChain=BuildChain[Triplets,InitialDistrubution]`, where `MyChain` is just the name of the variable (could be anything), `Triplets` is a list each of whose elements is a triplet of the from `{FromState,ToState,probability}`, and `InitialDistribution` is a list of pairs of the form `{State,probability}` where `probability`$= \mathbb{P}[X_0 =$`State`$]$. Here is an example for the simple Markov chain with three states A,B and C, the initial distribution $\mathbb{P}[X_0 = A] = 1/2$, $\mathbb{P}[X_0 = B] = 1/4$ and $\mathbb{P}[X_0 = C] = 1/4$, and the transition graph that looks like this:



The code I used to define it is

```
In[1]:= << Stochastics`

In[34]:= MyChain = BuildChain[
           {{A, B, 1/2}, {A, A, 1/2}, {B, C, 1}, {C, C, 1/2}, {C, B, 1/2}},
           {{A, 1/2}, {B, 1/4}, {C, 1/4}}
         ];
```

The "Tennis Chain" example we covered in class is already implemented in the package, so that you don't have to type it in yourself. All you need to do it issue the command `MyChain=TennisChain[p]`, where p is the probability of winning a rally for Amélie.

9.3 Getting information about a chain

Once you have built a chain (assume that its name is `MyChain`) you can get information about it using the following commands. We start by listing the simple ones:

- `States[MyChain]` returns the list of all states.
- `InitialDistribution[MyChain]` returns the list containing the list of initial probabilities. The order is the same as that returned by the command `States`.
- `TransitionMatrix[MyChain]` returns the transition matrix.
- `Probability[s1,s2,n,MyChain]` returns the n-step probability $\mathbb{P}[X_n = s2|X_0 = s1]$.
- `NumberOfStates[MyChain]` returns the number of states.
- `Classes[MyChain]` returns a list whose elements are the communication classes (lists themselves) of the chain.
- `Recurrence[MyChain]` returns a list of zeros or ones, one for each state. It is 1 if the state is recurrent and 0 otherwise.
- `TransientStates[MyChain]` returns the list of transient states.
- `RecurrentStates[MyChain]` returns the list of recurrent states.
- `ClassNumber[s,MyChain]` returns the number of the communication class the state `s` belongs to, i.e., the number of the element of `Classes[MyChain]` that `s` belongs to.

See Example 1 for a Mathematica notebook which illustrates these commands using `MyChain` defined above. The following commands help with more complicated (matrix) computations:

- `QMatrix[MyChain]`, `RMatrix[MyChain]` and `PMatrix[MyChain]` return the matrices Q, R and P from the canonical decomposition of `MyChain`.

- `CanonicalForm[MyChain]` returns the 2x2 block matrix representation of the canonical form of the chain `MyChain`
- `FundamentalMatrix[MyChain]` returns the fundamental matrix $(I-Q)^{-1}$ of the chain.

See Example 9.2 for a related Mathematica notebook.

9.4 Simulation

The functions in `Stochastics` can be used to perform repeated simulations of Markov chains:

- `SimulateFirst[MyChain]` outputs a single draw from the initial distribution of the chain.
- `SimulateNext[s,MyChain]` outputs a single draw from tomorrow's position of the chain if it is in the state `s` today.
- `SimulatePaths[nsim,nsteps,MyChain]` outputs a matrix with `nsim` rows and `nsteps` columns; each row is a simulated path of the chain of length `nsteps`.

See Example 9.3 for an illustration.

9.5 Plots

Finally, the package `Stochastics` can produce pretty pictures and animations of your chains:

- `PlotChain[MyChain,Options]` produces a graphical representation of the Markov Chain `MyChain`. The `Options` argument is optional and can be used to change the look and feel of the plot. Any option that the built-in function `GraphPlot` accepts can be used.
- `AnimateChain[nsteps,Mychain]` produces an animation of a single simulated trajectory of `MyChain`.

The picture in section 9.2 is the output of the command `PlotChain[MyChain]`.

9.6 Examples

Example 9.1. Simple commands

```
In[1]:= << Stochastics`

In[2]:= MyChain = BuildChain[
          {{A, B, 1 / 2}, {A, A, 1 / 2}, {B, C, 1}, {C, C, 1 / 2}, {C, B, 1 / 2}},
          {{A, 1 / 2}, {B, 1 / 4}, {C, 1 / 4}}
        ];

In[3]:= States[MyChain]

Out[3]= {A, B, C}

In[4]:= NumberOfStates[MyChain]

Out[4]= 3

In[6]:= InitialDistribution[MyChain]

Out[6]= {1/2, 1/4, 1/4}

In[7]:= TransitionMatrix[MyChain]

Out[7]= {{1/2, 1/2, 0}, {0, 0, 1}, {0, 1/2, 1/2}}

In[8]:= MatrixForm[TransitionMatrix[MyChain]]

Out[8]//MatrixForm=
        ( 1/2  1/2  0   )
        ( 0    0    1   )
        ( 0    1/2  1/2 )

In[9]:= Probability[A, B, 3, MyChain]

Out[9]= 3/8

In[10]:= Classes[MyChain]

Out[10]= {{A}, {B, C}}

In[11]:= Recurrence[MyChain]

Out[11]= {0, 1, 1}

In[12]:= TransientStates[MyChain]

Out[12]= {A}

In[13]:= RecurrentStates[MyChain]

Out[13]= {B, C}
```

Example 9.2. Canonical-form related commands (a continuation of Example 9.1)

```
In[14]:= QMatrix[MyChain]
```

$$\text{Out[14]}= \left\{\left\{\frac{1}{2}\right\}\right\}$$

```
In[15]:= RMatrix[MyChain]
```

$$\text{Out[15]}= \left\{\left\{\frac{1}{2}, 0\right\}\right\}$$

```
In[16]:= PMatrix[MyChain]
```

$$\text{Out[16]}= \left\{\{0, 1\}, \left\{\frac{1}{2}, \frac{1}{2}\right\}\right\}$$

```
In[19]:= M = CanonicalForm[MyChain]
```

$$\text{Out[19]}= \left\{\left\{\{0, 1\}, \left\{\frac{1}{2}, \frac{1}{2}\right\}\right\}, \{\{0\}, \{0\}\}\right\}, \left\{\left\{\left\{\frac{1}{2}, 0\right\}\right\}, \left\{\left\{\frac{1}{2}\right\}\right\}\right\}$$

```
In[20]:= MatrixForm[
          {
           {MatrixForm[M[[1, 1]]], MatrixForm[M[[1, 2]]]},
           {MatrixForm[M[[2, 1]]], MatrixForm[M[[2, 2]]]}
          }
         ]
```

Out[20]//MatrixForm=

$$\begin{pmatrix} \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \begin{pmatrix} \frac{1}{2} & 0 \end{pmatrix} & \begin{pmatrix} \frac{1}{2} \end{pmatrix} \end{pmatrix}$$

```
In[21]:= FundamentalMatrix[MyChain]
```

$$\text{Out[21]}= \{\{2\}\}$$

Example 9.3. Simulations (a continuation of Example 9.1)

```
In[22]:= SimulateFirst[MyChain]
Out[22]= A

In[23]:= SimulateFirst[MyChain]
Out[23]= C

In[24]:= SimulateNext[A, MyChain]
Out[24]= A

In[28]:= SimulateNext[A, MyChain]
Out[28]= B

In[36]:= MatrixForm[SimulatePaths[3, 10, MyChain]]
```

Out[36]//MatrixForm=

$$\begin{pmatrix} C & B & C & B & C & C & B & C & B & C \\ A & A & B & C & C & B & C & B & C & B \\ A & A & A & B & C & B & C & C & B & C \end{pmatrix}$$

Chapter 10

Classification of States

There will be a lot of definitions and some theory before we get to examples. You might want to peek into the last part (examples) as notions are being introduced; it will help your understanding.

10.1 The Communication Relation

Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a Markov chain on the state space $\mathcal{S}$. For a given set B of states, define the **hitting time** $\tau(B)$ of B as

$$\tau_B = \min\{n \in \mathbb{N}_0 \,:\, X_n \in B\}. \tag{10.1}$$

We know that τ_B is, in fact, a stopping time with respect to $\{X_n\}_{n\in\mathbb{N}_0}$. When B consists of only one element $B = \{i\}$, we simply write τ_i for $\tau_{\{i\}}$; τ_i is the first time the Markov chain $\{X_n\}_{n\in\mathbb{N}_0}$ "hits" the state i. As always, we allow τ_B to take the value $+\infty$; it means that no state in B is ever hit.

The hitting times are important both for immediate applications of $\{X_n\}_{n\in\mathbb{N}_0}$, as well as for better understanding of the structure of Markov chains.

Example 10.1. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be the chain which models a game of tennis (Example 9., in Section 2. of Lecture 8). The probability of winning for (say) Amélie can be phrased in terms of hitting times:

$$\mathbb{P}[\text{ Amélie wins }] = \mathbb{P}[\tau_{i_A} < \tau_{i_B}],$$

where $i_A =$ "Amélie wins" and $i_B =$"Björn wins" (the two absorbing states of the chain). We will learn how to compute such probabilities in the subsequent lectures.

Having introduced the hitting times τ_B, let us give a few more definitions. Recall that the notation $\mathbb{P}_i[A]$ is used to mean $\mathbb{P}[A|X_0 = i]$ (for any event A). In practice, we use $\mathbb{P}_i$ to signify that we are starting the chain from the state i, i.e., $\mathbb{P}_i$ corresponds to a Markov chain whose transition matrix is the same as the one of $\{X_n\}_{n\in\mathbb{N}_0}$, but the initial distribution is given by $\mathbb{P}_i[X_0 = j] = 0$ if $j \neq i$ and $\mathbb{P}_i[X_0 = i] = 1$.

Definition 10.2. A state $i \in \mathcal{S}$ is said to **communicate** with the state $j \in \mathcal{S}$, denoted by $i \to j$ if

$$\mathbb{P}_i[\tau_j < \infty] > 0.$$

Intuitively, i communicates with j is there is a non-zero chance that the Markov chain X will eventually visit j if it starts from i. Sometimes we also say that j is **a consequent of** i, or that j **is accessible from** i.

Example 10.3. In the tennis example, every state is accessible from $(0,0)$ (the fact that $p \in (0,1)$ is important here), but $(0,0)$ is not accessible from any other state. The consequents of $(40,40)$ are $(40,40)$ itself, $(40, Adv)$, $(Adv, 40)$, "Amélie wins" and "Björn wins".

Before we examine some properties of the relation $\to$, here is a simple but useful characterization. Before we give it, we recall the following fact about probability: let $\{A_n\}_{n\in\mathbb{N}_0}$ be an increasing sequence of events, i.e., $A_n \subseteq A_{n+1}$, for all $n \in \mathbb{N}_0$. Then

$$\mathbb{P}[\cup_{n\in\mathbb{N}_0} A_n] = \lim_n \mathbb{P}[A_n],$$

and the sequence inside the limit is non-decreasing.

Proposition 10.4. *$i \to j$ if and only if $p_{ij}^{(n)} > 0$ for some $n \in \mathbb{N}_0$.*

Proof. The event $A = \{\tau_j < \infty\}$ can be written as an *increasing union*

$$A = \cup_{n\in\mathbb{N}} A_n, \text{ where } A_n = \{\tau_j \le n\}.$$

Therefore,

$$\mathbb{P}_i[\tau_j < \infty] = \mathbb{P}_i[A] = \lim_n \mathbb{P}_i[A_n] = \lim_n \mathbb{P}_i[\tau_j \le n],$$

and the sequence $\mathbb{P}_i[\tau_j \le n]$, $n \in \mathbb{N}$ is *non-decreasing*. In particular,

$$\mathbb{P}_i[\tau_j < \infty] \ge \mathbb{P}_i[A_n], \text{ for all } n \in \mathbb{N}. \tag{10.2}$$

Suppose, first, that $p_{ij}^{(n)} > 0$ for some n. Since τ_j is the *first time* j is visited, we have

$$\mathbb{P}_i[A_n] = \mathbb{P}_i[\tau_j \le n] \ge \mathbb{P}_i[X_j = n] = p_{ij}^{(n)} > 0.$$

By (10.2), we have $\mathbb{P}_i[\tau_j < \infty] > 0$, and so, $i \to j$.

Conversely, suppose that $i \to j$, i.e., $\mathbb{P}_i[A] > 0$. Since $\mathbb{P}_i[A] = \lim_n \mathbb{P}_i[A_n]$, we must have $\mathbb{P}_i[A_n] > 0$ for some n_0 (and then all larger n), i.e., $\mathbb{P}_i[\tau_j \le n_0] > 0$. When $\tau_j \le n_0$, we must have $X_0 = j$ or $X_1 = j$ or $\dots$ or $X_n = j$, i.e.,

$$\{\tau_j \le n_0\} \subseteq \cup_{k=0}^{n_0} \{X_k = j\},$$

and so

$$0 < \mathbb{P}_i[\tau_j \le n_0] \le \mathbb{P}_i[\cup_{k=0}^{n_0} \{X_k = j\}] \le \sum_{k=0}^{n_0} \mathbb{P}_i[X_k = j].$$

Therefore, $\mathbb{P}_i[X_n = j]$ for at least one $n \in \{0, 1, \dots, n_0\}$. In other words, $p_{ij}^{(n)} > 0$, for at least one $n \in \{0, 1, \dots, n_0\}$. □

Proposition 10.5. *For all $i, j, k \in \mathcal{S}$, we have*

1. $i \to i$,
2. $i \to j, j \to k \Rightarrow i \to k$.

Proof.

1. If we start from state $i \in \mathcal{S}$ we are already there (note that 0 is allowed as a value for τ_B in (10.1)), i.e., $\tau_i = 0$ when $X_0 = i$.
2. Using Proposition 10.4, it will be enough to show that $p_{ik}^{(n)} > 0$ for some $n \in \mathbb{N}$. By the same Proposition, we know that $p_{ij}^{(n_1)} > 0$ and $p_{jk}^{(n_2)} > 0$ for some $n_1, n_2 \in \mathbb{N}_0$. By the Chapman-Kolmogorov relations, with $n = n_1 + n_2$, we have
$$p_{ik}^{(n)} = \sum_{l \in \mathcal{S}} p_{il}^{(n_1)} p_{lk}^{(n_2)} \geq p_{ij}^{(n_1)} p_{jk}^{(n_2)} > 0.$$

□

Remark 10.6. The inequality $p_{ik}^{(n)} \geq p_{il}^{(n_1)} p_{lk}^{(n_2)}$ is valid for all $i, l, k \in \mathcal{S}$, as long as $n_1 + n_2 = n$. It will come in handy later.

10.2 Classes

Definition 10.7. We say that the states i and j in $\mathcal{S}$ **intercommunicate**, denoted by $i \leftrightarrow j$ if $i \to j$ and $j \to i$. A set $B \subseteq \mathcal{S}$ of states is called **irreducible** if $i \leftrightarrow j$ for all $i, j \in \mathcal{S}$.

Unlike the relation of communication, the relation of intercommunication is symmetric. Moreover, we have the following three properties (the result follows directly from Proposition 10.4, so we omit it):

Proposition 10.8. *The relation $\leftrightarrow$ is an* equivalence relation *of $\mathcal{S}$, i.e., for all $i, j, k \in \mathcal{S}$, we have*

1. $i \leftrightarrow i$ *(*reflexivity*)* ,
2. $i \leftrightarrow j \Rightarrow j \leftrightarrow i$ *(*symmetry*), and*
3. $i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k$ *(*transitivity*).*

The fact that $\leftrightarrow$ is an equivalence relation allows us to split the state-space $\mathcal{S}$ into equivalence classes with respect to $\leftrightarrow$. In other words, we can write
$$\mathcal{S} = S_1 \cup S_2 \cup S_3 \cup \dots,$$
where $S_1, S_2, \dots$ are mutually exclusive (disjoint) and all states in a particular S_n intercommunicate, while no two states from different equivalence classes S_n and S_m do. The sets $S_1, S_2, \dots$ are called **classes** of the chain $\{X_n\}_{n \in \mathbb{N}_0}$. Equivalently, one can say that classes are *maximal irreducible sets*, in the sense that they are irreducible and no class is a subset of a (strictly larger) irreducible set. A cookbook algorithm for class identification would involve the following steps:

1. Start from an arbitrary state (call it 1).

2. Identify *all* states j that communicate with it - don't forget that always $i \leftrightarrow i$, for all i.

3. That is your first class, call it C_1. If there are no elements left, then there is only one class $C_1 = \mathcal{S}$. If there is an element in $\mathcal{S} \setminus C_1$, repeat the procedure above starting from that element.

The notion of a class is especially useful in relation to another natural concept:

Definition 10.9. A set $B \subseteq \mathcal{S}$ of states is said to be **closed** if $i \not\to j$ for all $i \in B$ and all $j \in \mathcal{S} \setminus B$. A state $i \in \mathcal{S}$ such that the set $\{i\}$ is closed is called **absorbing**.

Here is a nice characterization of closed sets:

Proposition 10.10. *A set B of states is closed if and only if $p_{ij} = 0$ for all $i \in B$ and all $j \in B^c = \mathcal{S} \setminus B$.*

Proof. Suppose, first, that B is closed. Then for $i \in B$ and $j \in B^c$, we have $i \not\to j$, i.e., $p_{ij}^{(n)} = 0$ for all $n \in \mathbb{N}$. In particular, $p_{ij} = 0$.

Conversely, suppose that $p_{ij} = 0$ for all $i \in B$, $j \in B^c$. We need to show that $i \not\to j$ (i.e. $p_{ij}^{(n)} = 0$ for all $n \in \mathbb{N}$) for all $i \in B$, $j \in B^c$. Suppose, to the contrary, that there exist $i \in B$ and $j \in B^c$ such that $p_{ij}^{(n)} > 0$ for some $n \in \mathbb{N}$. Since $p_{ij} = 0$, we must have $n > 1$. Out of all $n > 1$ such that $p_{ik}^{(n)} > 0$ *for some* $k \in B^c$, we pick the smallest one (let us call it n_0), so that that $p_{ik}^{(n_0-1)} = 0$ for all $k \in B^c$. By the Chapman-Kolmogorov relation, we have

$$\begin{aligned} p_{ij}^{(n)} &= \sum_{k \in \mathcal{S}} p_{ik}^{(n-1)} p_{kj} \\ &= \sum_{k \in B} p_{ik}^{(n-1)} p_{kj} + \sum_{k \in B^c} p_{ik}^{(n-1)} p_{kj}. \end{aligned} \tag{10.3}$$

The terms in the first sum in the second line of (10.3) are all zero because $p_{kj} = 0$ ($k \in B$ and $j \in B^c$), and the terms of the second one are also all zero because $p_{ik}^{(n-1)} = 0$ for all $k \in B^c$. Therefore, $p_{ij}^{(n)} = 0$ - a contradiction. □

Intuitively, a set of states is closed if it has the property that the chain $\{X_n\}_{n \in \mathbb{N}_0}$ stays in it forever, once it enters it. In general, if B is closed, it does not have to follow that $\mathcal{S} \setminus B$ is closed. Also, a class does not have to be closed, and a closed set does not have to be a class. Here are some examples:

Example 10.11. Consider the *tennis* chain of the previous lecture and consider the following three sets of states:

1. $B = \{$"Amélie wins"$\}$: closed and a class, but $\mathcal{S} \setminus B$ is not closed

2. $B = \mathcal{S} \setminus \{(0,0)\}$: closed, but not a class, and

3. $B = \{(0,0)\}$: class, but not closed.

There is a relationship between classes and the notion of closedness:

Proposition 10.12. *Every closed set B is a union of classes.*

Proof. Let $\hat{B}$ be the union of all classes C such that $C \cap B \neq \emptyset$. In other words, take all the elements of B and throw in all the states which intercommunicate with them. I claim that $\hat{B} = B$. Blearly, $B \subset \hat{B}$, so we need to show that $\hat{B} \subseteq B$. Suppose, to the contrary, that there exists $j \in \hat{B} \setminus B$. By construction, j intercommunicates with some $i \in B$. In particular $i \to j$. By closedness of B, we must have $j \in B$. This is a contradiction with the assumptions that $j \in \hat{B} \setminus B$. □

Example 10.13. A converse of Proposition 10.12 is not true. Just take the set $B = \{(0,0),(0,15)\}$ in the "tennis" example. It is a union of classes, but it is not closed.

10.3 Transience and recurrence

It is often important to know whether a Markov chain will ever return to its initial state, and if so, how often. The notions of transience and recurrence address this questions.

Definition 10.14. The **(first) visit time** to state j, denoted by $\tau_j(1)$ is defined as

$$\tau_j(1) = \min\{n \in \mathbb{N} : X_n = j\}.$$

As usual $\tau_j(1) = +\infty$ if $X_n \neq j$ for all $n \in \mathbb{N}$.

Note that the definition of the random variable $\tau_j(1)$ differs from the definition of τ_j in that the minimum here is take over the set $\mathbb{N}$ of natural numbers, while the set of non-negative integers $\mathbb{N}_0$ is used for τ_j. When $X_0 \neq j$, the hitting time τ_j and the visit time $\tau_j(1)$ coincide. The important difference occurs when $X_0 = j$. In that case $\tau_j = 0$ (we are already there), but it is always true that $\tau_j(1) \geq 1$. It can even happen that $\mathbb{P}_i[\tau_i(1) = \infty] = 1$.

Definition 10.15. A state $i \in \mathcal{S}$ is said to be

1. **recurrent** if $\mathbb{P}_i[\tau_i(1) < \infty] = 1$,
2. **positive recurrent** if $\mathbb{E}_i[\tau_i(1)] < \infty$ ($\mathbb{E}_i$ means expectation when the probability is $\mathbb{P}_i$),
3. **null recurrent** if it is recurrent, but not positive recurrent,
4. **transient** if it is not recurrent.

A state is recurrent if we are sure we will come back to it eventually (with probability 1). It is positive recurrent if the time between two consecutive visits has finite expectation. Null recurrence means the we will return, but the waiting time may be very long. A state is transient is there is a positive chance (however small) that the chain will never return to it.

Remember that the **greatest common denominator (GCD)** of a set A of natural numbers if the largest number d such that d divides each $k \in A$, i.e., such that each $k \in A$ is of the form $k = ld$ for some $l \in \mathbb{N}$.

Definition 10.16. A **period** $d(i)$ of a state $i \in \mathcal{S}$ is the greatest common denominator of the **return-time set** $R(i) = \{n \in \mathbb{N} : p_{ii}^{(n)} > 0\}$ of the state i. When $R(i) = \emptyset$, we set $d(i) = 1$. A state $i \in \mathcal{S}$ is called **aperiodic** if $d(i) = 1$.

Example 10.17. Consider the Markov chain with three states and the transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

The return set for each state $i \in \{1, 2, 3\}$ is given by

$$R(i) = \{3, 6, 9, 12, \dots\},$$

so $d(i) = 3$ for all $i \in \{1, 2, 3\}$. However, if we change the probabilities a bit:

$$\hat{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix},$$

the situation changes drastically:

$$\begin{aligned} R(1) &= \{3, 4, 5, 6, \dots\}, \\ R(2) &= \{2, 3, 4, 5, 6, \dots\}, \\ R(3) &= \{1, 2, 3, 4, 5, 6, \dots\}, \end{aligned}$$

so that $d(i) = 1$ for $i \in \{1, 2, 3\}$.

10.4 Examples

Random walks: $p \in (0, 1)$.

- **Communication and classes**. Clearly, it is possible to go from any state i to either $i+1$ or $i-1$ in one step, so $i \to i+1$ and $i \to i-1$ for all $i \in \mathcal{S}$. By transitivity of communication, we have $i \to i+1 \to i+2 \to \cdots \to i+k$. Similarly, $i \to i-k$ for any $k \in \mathbb{N}$. Therefore, $i \to j$ for all $i, j \in \mathcal{S}$, and so, $i \leftrightarrow j$ for all $i, j \in \mathcal{S}$, and the whole $\mathcal{S}$ is one big class.
- **Closed sets**. The only closed set is $\mathcal{S}$ itself.
- **Transience and recurrence** We studied transience and recurrence in the lectures about random walks (we just did not call them that). The situation highly depends on the probability p of making an up-step. If $p > \frac{1}{2}$, there is a positive probability that the first step will be "up", so that $X_1 = 1$. Then, we know that there is a positive probability that the walk will never hit 0 again. Therefore, there is a positive probability of never returning to 0, which means that the state 0 is transient. A similar argument can be made for any state i and any probability $p \neq \frac{1}{2}$. What happens when $p = \frac{1}{2}$? In order to come back to 0, the walk needs to return there from its position at time $n = 1$. If it went up, the we have to wait for the walk to hit 0 starting from 1. We have shown that this *will* happen sooner or later, but that the expected time it takes is infinite. The same argument works if $X_1 = -1$. All in all, 0 (and all other states) are null-recurrent (recurrent, but not positive recurrent).
- **Periodicity**. Starting from any state $i \in \mathcal{S}$, we can return to it after $2, 4, 6, \dots$ steps. Therefore, the return set $R(i)$ is always given by $R(i) = \{2, 4, 6, \dots\}$ and so $d(i) = 2$ for all $i \in \mathcal{S}$.

Gambler's ruin: $p \in (0,1)$.

- **Communication and classes**. The winning state a and the losing state 0 are clearly absorbing, and form one-element classes. The other $a-1$ states intercommunicate among each other, so they form a class of their own. This class is not closed (you can - and will - exit it and get absorbed sooner or later).
- **Transience and recurrence**. The absorbing states 0 and a are (trivially) positive recurrent. All the other states are transient: starting from any state $i \in \{1, 2, \dots, a-1\}$, there is a positive probability (equal to p^{a-i}) of winning every one of the next $a-i$ games and, thus, getting absorbed in a before returning to i.
- **Periodicity**. The absorbing states have period 1 since $R(0) = R(a) = \mathbb{N}$. The other states have period 2 (just like in the case of a random walk).

Deterministically monotone Markov chain

- **Communication and classes**. A state i communicates with the state j if and only if $j \geq i$. Therefore $i \leftrightarrow j$ if and only if $i = j$, and so, each $i \in \mathcal{S}$ is in a class by itself.
- **Closed sets**. The closed sets are precisely the sets of the form $B = i, i+1, i+2, \dots$, for $i \in \mathbb{N}$.
- **Transience and recurrence** All states are transient.
- **Periodicity**. The return set $R(i)$ is empty for each $i \in \mathcal{S}$, so $d(i) = 1$, for all $i \in \mathcal{S}$.

A game of tennis

- **Communication and classes**. All the states except for those in $E = \{(40, Adv), (40, 40), (Adv, 40),$ Amélie wins, Björn wins$\}$ intercommunicate only with themselves, so each $i \in \mathcal{S} \setminus E$ is in a class by iteself. The winning states *Amélie wins* and *Björn wins* are absorbing, and, so, also form classes with one element. Finally, the three states in $\{(40, Adv), (40, 40), (Adv, 40)\}$ intercommunicate with each other, so they form the last class.
- **Periodicity**. The states i in $\mathcal{S} \setminus E$ have have the property that $p_{ii}^{(n)} = 0$ for all $n \in \mathbb{N}$, so $d(i) = 1$. The winning states are absorbing so $d(i) = 1$ for $i \in \{$Amélie wins, Björn wins$\}$. Finally, the return set for the remaining three states is $\{2, 4, 6, \dots\}$ so their period is 2.

Chapter 11

More on Transience and recurrence

11.1 A criterion for recurrence

The definition of recurrence from the previous lecture is conceptually simple, but it gives us no clue about how to actually go about deciding whether a particular state in a specifis Markov chain is recurrent. A criterion stated entirely in terms of the transition matrix P would be nice. Before we give it, we need to introduce some notation. For two (not necessarily different) states, $i, j \in \mathcal{S}$, let $f_{ij}^{(n)}$ be the probability that it will take exactly n steps for the first visit to j (starting from i) to occur, i.e.,

$$f_{ij}^{(n)} = \mathbb{P}_i[\tau_j(1) = n] = \mathbb{P}[X_n = j, X_{n-1} \neq j, X_{n-2} \neq j, \ldots, X_2 \neq j, X_1 \neq j | X_0 = i].$$

Let f_{ij} denote the probability that j will be reached from i *eventually*, i.e.,

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}.$$

Clearly, we have the following equivalence: i is recurrent if and only if $f_{ii} = 1$.

The reason the quantities $f_{ij}^{(n)}$ are useful lies in the following recursive relationship:

Proposition 11.1. *For $n \in \mathbb{N}$ and $i, j \in \mathcal{S}$, we have*

$$p_{ij}^{(n)} = \sum_{m=1}^{n} p_{jj}^{(n-m)} f_{ij}^{(m)}.$$

Proof. In preparation for the rest of the proof, let us reiterate that the event $\{\tau_j(1) = m\}$ can be written as

$$\{\tau_j(1) = m\} = \{X_m = j, X_{m-1} \neq j, X_{m-2} \neq j, \ldots, X_1 \neq j\}.$$

Using the law of total probability, we can split the event $\{X_n = j\}$ according to the value of the random variable $\tau_j(1)$ (the values of $\tau_j(1)$ larger than n do not appear in the sum since X_n cannot

be equal to j in those cases):

$$\begin{aligned} p_{ij}^{(n)} &= \mathbb{P}[X_n = j | X_0 = i] = \sum_{m=1}^{n} \mathbb{P}[X_n = j | \tau_j(1) = m, X_0 = i] \, \mathbb{P}[\tau_j(1) = m | X_0 = i] \\ &= \sum_{m=1}^{n} \mathbb{P}[X_n = j | X_m = j, X_{m-1} \neq j, X_{m-2} \neq j, \ldots, X_1 \neq j, X_0 = i] \; f_{ij}^{(m)} \\ &= \sum_{m=1}^{n} \mathbb{P}[X_n = j | X_m = j] \; f_{ij}^{(m)} = \sum_{m=1}^{n} p_{jj}^{(n-m)} f_{ij}^{(m)} \end{aligned}$$

□

An important corollary to Proposition 11.1 is the following characterization of recurrence:

Proposition 11.2. *A state $i \in \mathcal{S}$ is recurrent if and only if*

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = +\infty.$$

Proof. For $i = j$, Proposition 11.1 states that

$$p_{ii}^{(n)} = \sum_{m=1}^{n} p_{ii}^{(n-m)} f_{ii}^{(m)}.$$

Summing over all n from 1 to $N \in \mathbb{N}$, we get

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{n=1}^{N} \sum_{m=1}^{n} p_{ii}^{(n-m)} f_{ii}^{(m)} = \sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{1}_{\{m \leq n\}} p_{ii}^{(n-m)} f_{ii}^{(m)} \tag{11.1}$$

We set $s_N = \sum_{n=1}^{N} p_{ii}^{(n)}$ for $N \in \mathbb{N}$, remember that $p_{ii}^{(0)} = 1$, and interchange the order of summation to get

$$s_N = \sum_{m=1}^{N} \sum_{n=m}^{N} p_{ii}^{(n-m)} f_{ii}^{(m)} = \sum_{m=1}^{N} f_{ii}^{(m)} \sum_{n=0}^{N-m} p_{ii}^{(n)} = \sum_{m=1}^{N} f_{ii}^{(m)} (1 + s_{N-m})$$

The sequence $\{s_N\}_{N \in \mathbb{N}}$ is non-decreasing, so

$$s_N \leq \sum_{m=1}^{N} f_{ii}^{(m)} (1 + s_N) \leq (1 + s_N) f_{ii}.$$

Therefore,

$$f_{ii} \geq \lim_N \frac{s_N}{1 + s_N}.$$

If $\sum_{n=1}^{\infty} p_{ii}^{(n)} = +\infty$, then $\lim_{N \to \infty} \frac{s_N}{1+s_N} = 1$, so $f_{ii} \geq 1$. On the other hand, $f_{ii} = \mathbb{P}_i[\tau_i(1) < \infty]$, so $f_{ii} \leq 1$. Therefore, $f_{ii} = 1$, and, so, the state i is recurrent.

Conversely, suppose that i is recurrent, i.e., $f_{ii} = 1$. We can repeat the procedure from above (but with $+\infty$ instead of N and using Tonelli's theorem to interchange the order of the sums) to get that

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \sum_{m=1}^{\infty} f_{ii}^{(m)}(1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}) = 1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}.$$

This can happen only if $\sum_{m=1}^{\infty} p_{ii}^{(n)} = +\infty$. □

11.2 Class properties

Certain properties of states are shared between all elements in a class. Knowing which properties share this feature is useful for a simple reason - if you can check them for a single class member, you know automatically that all the other elements of the class share it.

Definition 11.3. A property is called a **class property** it holds for all states in its class, whenever it holds for any one particular state in the that class.

Put differently, a property is a class property if and only if either all states in a class have or none does.

Proposition 11.4. *Transience and recurrence are class properties.*

Proof. Suppose that the state i is recurrent, and that j is in its class, i.e., that $i \leftrightarrow j$. Then, there exist natural numbers m and k such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(k)} > 0$. By the Chapman-Kolmogorov relations, for each $n \in \mathbb{N}$, we have

$$p_{jj}^{(n+m+k)} = \sum_{l_1 \in \mathcal{S}} \sum_{l_2 \in \mathcal{S}} p_{jl_1}^{(k)} p_{l_1 l_2}^{(n)} p_{l_2 m}^{(m)} \geq p_{ji}^{(k)} p_{ii}^{(n)} p_{ij}^{(m)}.$$

In other words, there exists a positive constant c (take $c = p_{ji}^{(k)} p_{ij}^{(m)}$), independent of n, such that

$$p_{jj}^{(n+m+k)} \geq c p_{ii}^{(n)}.$$

Therefore, by recurrence of i we have $\sum_{n=1}^{\infty} p_{ii}^{(n)} = +\infty$, and

$$\sum_{n=1}^{\infty} p_{jj}^{(n)} \geq \sum_{n=m+k+1}^{\infty} p_{jj}^{(n)} = \sum_{n=1}^{\infty} p_{jj}^{(n+m+k)} \geq c \sum_{n=1}^{\infty} p_{ii}^{(n)} = +\infty,$$

and so, j is recurrent. Therefore, recurrence is a class property.

Since transience is just the opposite of recurrence, it is clear that transience is also a class property. □

Proposition 11.5. *Period is a class property, i.e., all elements of a class have the same period.*

Proof. Let $d = d(i)$ be the period of the state i, and let $j \leftrightarrow i$. Then, there exist natural numbers m and k such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(k)} > 0$. By Chapman-Kolmogorov,

$$p_{ii}^{(m+k)} \geq p_{ij}^{(m)} p_{ji}^{(k)} > 0,$$

and so

$$m + k \in R(i). \tag{11.2}$$

Similarly, for any $n \in R(j)$,

$$p_{ii}^{(m+k+n)} \geq p_{ij}^{(m)} p_{jj}^{(n)} p_{ji}^{(k)} > 0,$$

so

$$m + k + n \in R(i). \tag{11.3}$$

By (11.2), $d(i)$ divides $m + k$, and, by (11.3), $d(i)$ divides $m + k + n$. Therefore, $d(i)$ divides n, for each $n \in R(j)$, and so, $d(i) \leq d(j)$, because $d(j)$ is the greatest common divisor of $R(j)$. The same argument with roles of i and j switched shows that $d(j) \leq d(i)$. Therefore, $d(i) = d(j)$. □

11.3 A canonical decomposition

Now that we know that transience and recurrence are class properties, we can introduce the notion of **canonical decomposition** of a Markov chain. Let $\mathcal{S}_1, \mathcal{S}_2, \ldots$ be the collection of all classes; some of them contain recurrent states and some transient ones. Proposition 11.4 tells us that if there is one recurrent state in a class, than all states in the class must be recurrent. This, it makes sense to call the whole class **recurrent**. Similarly, the classes which are not recurrent consists entirely of transient states, so we call them **transient**. There are at most countably many states, so the number of all classes is also at most countable. In particular, there are only countably (or finitely) many recurrent classes, and we usually denote them by $C_1, C_2, \ldots$. Transient classes are denoted by $T_1, T_2, \ldots$. There is no particular rule in the choice of indices $1, 2, 3, \ldots$ for particular classes. The only point is that they can be enumerated because there are at most countably many of them.

The distinction between different transient classes is usually not very important, so we pack all transient states together in a set $T = T_1 \cup T_2 \cup \ldots$.

Definition 11.6. Let $\mathcal{S}$ be the state space of a Markov chain $\{X_n\}_{n \in \mathbb{N}_0}$. Let $C_1, C_2, \ldots$ be its recurrent classes, and $T_1, T_2, \ldots$ the transient ones, and let $T = T_1 \cup T_2 \cup \ldots$ be their union. The decomposition

$$\mathcal{S} = T \cup C_1 \cup C_2 \cup C_3 \cup \ldots,$$

is called the **canonical decomposition** of the (state space of the) Markov chain $\{X_n\}_{n \in \mathbb{N}_0}$.

The reason that recurrent classes are important is simple - they can be interpreted as Markov chains themselves. In order for such an interpretation to be possible, we need to make sure that the Markov chain stays in a recurrent class if it starts there. In other words, we have the following important proposition:

Proposition 11.7. *Recurrent classes are closed.*

Proof. Suppose, contrary to the statement of the Proposition, that there exist two states $i \neq j$ such that

1. i is recurrent,
2. $i \to j$, and
3. $j \not\to i$.

The idea of the proof is the following: whenever the transition $i \to j$ occurs (perhaps in several steps), then the chain will never return to i, since $j \not\to i$. By recurrence of i, that is not possible. Therefore $i \not\to j$ - a contradiction.

More formally, the recurrence of i means that (starting from i, i.e., under $\mathbb{P}_i$) the event

$$A = \{X_n = i, \text{ for infinitely many } n \in \mathbb{N}\}$$

has probability 1, i.e., $\mathbb{P}_i[A] = 1$. The law of total probability implies that

$$1 = \mathbb{P}_i[A] = \sum_{n\in\mathbb{N}} \mathbb{P}_i[A|\tau_j(1) = n]\mathbb{P}_i[\tau_j(1) = n] + \mathbb{P}_i[A|\tau_j(1) = +\infty]\,\mathbb{P}_i[\tau_j(1) = +\infty]. \tag{11.4}$$

On the event $\tau_j(1) = n$, the chain can visit the state i at most $n-1$ times, because it cannot hit i after it visits j (remember $j \not\to i$). Therefore, $\mathbb{P}[A|\tau_j(1) = n] = 0$, for all $n \in \mathbb{N}$. It follows that

$$1 = \mathbb{P}_i[A|\tau_j(1) = +\infty]\,\mathbb{P}_i[\tau_j(1) = +\infty].$$

Both of the terms on the right-hand side above are probabilities whose product equals 1, which forces both of them to be equal to 1. In particular, $\mathbb{P}_i[\tau_j(1) = +\infty]$, or, phrased differently, $i \not\to j$ - a contradiction. □

Together with the canonical decomposition, we introduce the **canonical form** of the transition matrix P. The idea is to order the states in $\mathcal{S}$ with the canonical decomposition in mind. We start from all the states in C_1, followed by all the states in C_2, etc. Finally, we include all the states in T. The resulting matrix looks like this

$$P = \begin{bmatrix} P_1 & 0 & 0 & \dots & 0 \\ 0 & P_2 & 0 & \dots & 0 \\ 0 & 0 & P_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Q_1 & Q_2 & Q_3 & \dots & \dots \end{bmatrix},$$

where the entries should be interpreted as matrices: P_1 is the transition matrix within the first class, i.e., $P_1 = (p_{ij}, i \in C_1, j \in C_1)$, etc. Q_k contains the transition probabilities from the transient states to the states in the (recurrent) class C_k. Note that Proposition 11.7 implies that each P_k is a stochastic matrix, or, equivalently, that all the entries in the row of P_k outside of P_k are zeros.

We finish the discussion of canonical decomposition with an important result and one of its consequences.

Proposition 11.8. *Suppose that the state space $\mathcal{S}$ is finite. Then there exists at least one recurrent state.*

Proof. The transition matrix P is stochastic, and so are all its powers P^n. In particular, we have

$$1 = \sum_{j\in\mathcal{S}} p_{ij}^{(n)}, \text{ for all } i \in \mathcal{S}.$$

Summing the above equality over all $n \in \mathbb{N}$ and switching the order of integration gives

$$+\infty = \sum_{n\in\mathbb{N}} 1 = \sum_{n\in\mathbb{N}}\sum_{j\in\mathcal{S}} p_{ij}^{(n)} = \sum_{j\in\mathcal{S}}\sum_{n\in\mathbb{N}} p_{ij}^{(n)}.$$

We conclude that $\sum_{n\in\mathbb{N}} p_{ij}^{(n)} = +\infty$ for at least one state j (this is where the finiteness of $\mathcal{S}$ is crucial).

We claim that j is a recurrent state. Suppose, to the contrary, that it is transient. If we sum the equality from Proposition 11.1 over $n \in \mathbb{N}$, we get

$$\sum_{n\in\mathbb{N}} p_{ij}^{(n)} = \sum_{n\in\mathbb{N}}\sum_{m=1}^{n} p_{jj}^{(n-m)} f_{ij}^{(m)} = \sum_{m\in\mathbb{N}}\sum_{n=m}^{\infty} p_{jj}^{(n-m)} f_{ij}^{(m)} = \sum_{m\in\mathbb{N}} f_{ij}^{(m)} \sum_{n=0}^{\infty} p_{jj}^{(n)} = f_{ij}\sum_{n=0}^{\infty} p_{jj}^{(n)}. \quad (11.5)$$

Transience of j implies that $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$ and, by definition, $f_{ij} \leq 1$, for any i, j. Therefore, $\sum_{n\in\mathbb{N}} p_{ij}^{(n)} < \infty$ - a contradiction. □

Remark 11.9. If $\mathcal{S}$ is not finite, it is not true that recurrent states must exist. Just remember the Deterministically-monotone Markov chain example, or the random walk with $p \neq \frac{1}{2}$. All states are transitive there.

In a finite state-space case, we have the following dichotomy:

Corollary 11.10. *A class of a Markov chain on a finite state space is recurrent if and only if it is closed.*

Proof. We know that recurrent classes are closed. In order to show the converse, we need to prove that transient classes are not closed. Suppose, to the contrary, the there exists a finite state-space Markov chain with a closed transient class T. Since T is closed, we can see it as a state space of the restricted Markov chain. This, new, Markov chain has a finite number of states so there exists a recurrent state. This is a contradiction with the assumption that T consists only of transient states. □

Remark 11.11. Again, finiteness is necessary. For a random walk on $\mathbb{Z}$, all states intercommunicate. In particular, there is only one class $\mathbb{Z}$ itself and it it trivially closed. If $p \neq \frac{1}{2}$, however, all states are transient, and, so, $\mathbb{Z}$ is a closed and transient class.

Chapter 12

Absorption and reward

Note: Even though it is not by any means necessary, we are assuming that, from this point onwards, all Markov chains have finite state spaces.

12.1 Absorption

Remember the "Tennis" example from a few lectures ago and the question we asked there, namely, how does the probability of winning a single point affect the probability of winning the overall game? An algorithm that will help you answer that question will be described in this lecture.

The first step is to understand the structure of the question asked in the light of the canonical decomposition of the previous lecture. In the "Tennis" example, all the states except for the winning ones are transient, and there are two one-element recurrent classes $\{$*Amélie wins*$\}$ and $\{$*Björn wins*$\}$. The chain starts from a transient state $(0,0)$, moves around a bit, and, eventually, gets absorbed in one of the two. The probability we are interested in is not the probability that the chain will eventually get absorbed. This probability is always 1 (this is true in every finite Markov chain, but we do not give a proof of this). We are, instead, interested in the probability that the absorption will occurr in a particular state - the state $\{$*Amélie wins*$\}$ in the "Tennis" example.

A more general version of the problem above is the following: let $i \in \mathcal{S}$ be any state, and let j be a recurrent state. If the set of all recurrent states is denoted by C, and if τ_C is the first hitting time of the set C, then X_{τ_C} denotes the first recurrent state visited by the chain. Equivalently, X_{τ_C} is the value of X at (random) time τ_C; its value is the name of the state in which it happens to find itself the first time it hits the set of all recurrent states. For any two states $i, j \in \mathcal{S}$, the **absorption probability** u_{ij} is defined as

$$u_{ij} = \mathbb{P}_i[X_{\tau_C} = j] = \mathbb{P}_i[\text{ the first recurrent state visited by } X \text{ is } j].$$

When j, is *not* a recurrent state, then $u_{ij} = 0$; j cannot possibly be the first recurrent state we hit - it is not even recurrent. Whwn $i = j$ is a recurrent state, then $u_{ii} = 1$ - we are in i right from the start. The situation $i \in T$, $j \in C$ is the interesting one.

In many calculations related to Markov chains, the method of *first-step decomposition* works miracles. Simply, we cut the probability space according to what happened in the first step and

use the law of total probability (assuming $i \in T$, $j \in C$)

$$u_{ij} = \mathbb{P}_i[X_{\tau_C} = j] = \sum_{k \in \mathcal{S}} \mathbb{P}[X_{\tau_C} = j | X_0 = i, X_1 = k] \mathbb{P}[X_1 = k | X_0 = i]$$
$$= \sum_{k \in \mathcal{S}} \mathbb{P}[X_{\tau_C} = j | X_1 = k] p_{ik}$$

The conditional probability $\mathbb{P}[X_{\tau_C} = j | X_1 = k]$ is an absorption probability, too. If $k = j$, then $\mathbb{P}[X_{\tau_C} = j | X_1 = k] = 1$. If $k \in C \setminus \{j\}$, then we are already in C, but in a state different from j, so $\mathbb{P}[X_{\tau_C} = j | X_1 = k] = 0$. Therefore, the sum above can be written as

$$u_{ij} = \sum_{k \in T} p_{ik} u_{kj} + p_{ij},$$

which is a system of linear equations for the family $(u_{ij}, i \in T, j \in C)$. Linear systems are typically better understood when represented in the matrix form. Let U be a $T \times C$-matrix $U = (u_{ij}, i \in T, j \in C)$, and let Q be the portion of the transition matrix P corresponding to the transitions from T to T, i.e. $Q = (p_{ij}, i \in T, j \in T)$, and let R contain all transitions from T to C, i.e., $R = (p_{ij})_{i \in T, j \in C}$. If P_C denotes the matrix of all transitions from C to C, i.e., $P_C = (p_{ij}, i \in C, j \in C)$, then the canonical form of P looks like this:

$$P = \begin{bmatrix} P_C & 0 \\ R & Q \end{bmatrix}.$$

The system (12.1) now becomes:

$$U = QU + R, \text{ i.e., } (I - Q)U = R.$$

If the matrix $I - Q$ happens to be invertible, we are in business, because we then have an explicit expression for U:

$$U = (I - Q)^{-1} R.$$

So, is $I - Q$ invertible? It is when the state space $\mathcal{S}$ is finite, but you will not see the proof in these notes. When the inverse $(I - Q)^{-1}$ exists, it is called the **fundamental matrix** of the Markov chain.

Example 12.1. Before we turn to the "Tennis" example, let us analyze a simpler case of Gambler's ruin with $a = 3$. The states 0 and 3 are absorbing, and all the others are transient. Therefore $C_1 = \{0\}$, $C_2 = \{3\}$ and $T = T_1 = \{1, 2\}$. The transition matrix P in the canonical form (the rows and columns represent the states in the order $0, 3, 1, 2$)

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1-p & 0 & 0 & p \\ 0 & p & 1-p & 0 \end{bmatrix}$$

Therefore,

$$R = \begin{bmatrix} 1-p & 0 \\ 0 & p \end{bmatrix} \text{ and } Q = \begin{bmatrix} 0 & p \\ 1-p & 0 \end{bmatrix}.$$

The matrix $I - Q$ is a 2×2 matrix so it is easy to invert:

$$(I-Q)^{-1} = \frac{1}{1-p+p^2}\begin{bmatrix} 1 & p \\ 1-p & 1 \end{bmatrix}.$$

So

$$U = \frac{1}{1-p+p^2}\begin{bmatrix} 1 & p \\ 1-p & 1 \end{bmatrix}\begin{bmatrix} 1-p & 0 \\ 0 & p \end{bmatrix} = \begin{bmatrix} \frac{1-p}{1-p+p^2} & \frac{p^2}{1-p+p^2} \\ \frac{(1-p)^2}{1-p+p^2} & \frac{p}{1-p+p^2} \end{bmatrix}.$$

Therefore, for example, if the initial "wealth" is 1, the probability of getting rich before bankruptcy is $p^2/(1-p+p^2)$.

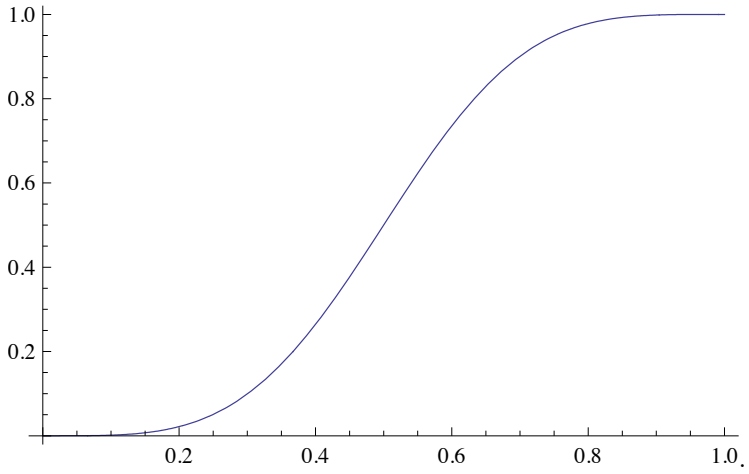
We have already calculated this probability in one of the homework problems (HW4, Problem 4.4). There, we obtained that the desired probability equals $(1-\frac{q}{p})/\big(1-(\frac{q}{p})^3\big)$. You can check that these two expressions are really the same.

Example 12.2. In the "Tennis" example, the transition matrix is 20×20, with only 2 recurrent states (each in its own class). The matrix given in Lecture 8 is already in the canonical form (the recurrent states correspond to the first two rows/columns). In order to get $(I-Q)^{-1}$, we need to invert an 18×18 matrix. This is a job for a computer, and we use *Mathematica* (`P` is the full transition matrix and 3 is order of the state $(0,0)$ in the internal representation). After we remove the absorbing states, the number of $(0,0)$ becomes 1 (in the *Mathematica* internal order) and that is why we are extracting the element in the first row and first column in U to get the result):

```
In[114]:= Q = P[[3 ;; 20, 3 ;; 20]]; R = P[[3 ;; 20, 1 ;; 2]];

In[115]:= F = Inverse[IdentityMatrix[18] - Q];

In[116]:= U = F.R;

In[117]:= U[[1, 1]]

Out[117]= p^4 + 4 p^4 q + 10 p^4 q^2 - 20 p^5 q^3 / (-1 + 2 p q)
```

.

If we plot this probability against the value of p, we get the following picture:



.

12.2 Expected reward

Suppose that each time you visit a transient state $j \in T$ you receive a *reward* $g(j) \in \mathbb{R}$. The name "reward" is a bit misleading since the negative $g(j)$ corresponds more to a fine than to a reward; it is just a name, anyway. Can we compute the expected total reward before absorption

$$v_i = \mathbb{E}_i[\sum_{n=0}^{\tau_C} g(X_n)]?$$

And if we can, what is it good for? Many things, actually, as the following two special cases show:

1. If $g(j) = 1$ for all $j \in T$, then v_i is the expected time until absorption. We will calculate $v_{(0,0)}$ for the "Tennis" example to compute the expected duration of a tennis game.
2. If $g(k) = 1$ and $g(j) = 0$ for $j \neq k$, then v_i is the expected number of visits to the state k before absorption. In the "Tennis" example, if $k = (40, 40)$, the value of $v_{(0,0)}$ is the expected number of times the score $(40, 40)$ is seen in a tennis game.

We compute v_i using the first-step decomposition:

$$\begin{aligned} v_i &= \mathbb{E}[\sum_{n=0}^{\tau_C} g(X_n)|X_0 = i] = g(i) + \mathbb{E}[\sum_{n=1}^{\tau_C} g(X_n)|X_0 = i] \\ &= g(i) + \sum_{k\in\mathcal{S}} \mathbb{E}[\sum_{n=1}^{\tau_C} g(X_n)|X_0 = i, X_1 = k]\mathbb{P}[X_1 = k|X_0 = i] \\ &= g(i) + \sum_{k\in\mathcal{S}} p_{ik}\mathbb{E}[\sum_{n=1}^{\tau_C} g(X_n)|X_1 = k] \end{aligned} \tag{12.1}$$

If $k \in T$, then the homogeneity of the chain implies that

$$\mathbb{E}[\sum_{n=1}^{\tau_C} g(X_n)|X_1 = k] = \mathbb{E}[\sum_{n=0}^{\tau_C} g(X_n)|X_0 = k] = v_k.$$

When $k \notin T$, then

$$\mathbb{E}[\sum_{n=1}^{\tau_C} g(X_n)|X_1 = k] = 0,$$

because we have "arrived" and no more rewards are going to be collected. Therefore, for $i \in T$ we have

$$v_i = g(i) + \sum_{k\in T} p_{ik}v_k.$$

If we organize all v_i and all $g(i)$ into column vectors $v = (v_i, i \in T)$, $g = (g(i), i \in T)$, we get

$$v = Qv + g, \text{ i.e., } v = (I - Q)^{-1}g.$$

Having derived the general forumla for various rewards, we can give an interpretation of the fundamental matrix itself. Let us pick a transient state j and use the reward function g given by

$$g(k) = \mathbf{1}_{\{k=j\}} = \begin{cases} 1, & k = j \\ 0, & k \neq j. \end{cases}$$

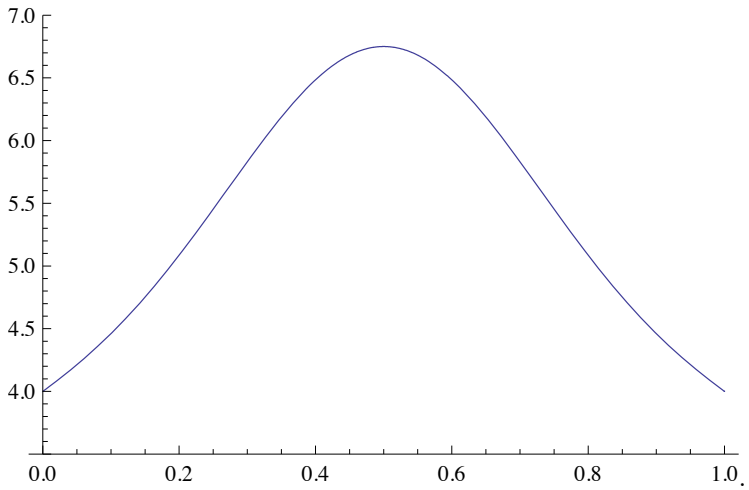
By the discussion above, the i^{th} entry in $v = (I-Q)^{-1}g$ is the expected reward when we start from the state i. Given the form of the reward function, v_i is the expected number of visits to the state j when we start from i. On the other hand, as the product of the matrix $(I-Q)^{-1}$ and the vector $g = (0, 0, \ldots, 1, \ldots, 0)$, v_i is nothing but the (i,j)-entry in $(I-Q)^{-1}$:

Proposition 12.3. *For two transient states i and j, the (i,j)-th entry in the fundamental matrix $(I-Q)^{-1}$ is the expteced number of visits to the state j prior to absorbtion if the chain starts from the state i (the time 0 is counted, if $i = j$).*

Example 12.4. We continue the analysis of the "Tennis" chain from Example 12.2. We set $g(i) = 1$ for all transient states i to find the expected duration of a tennis game.

```
In[87]:= G = Table[{1}, {i, 1, 18}];

In[86]:= f[p_] = Simplify[(F.G)[[1, 1]] //. {q → 1 - p}]

Out[86]= -(4 (-1 + p - p^2 - 6 p^3 + 18 p^4 - 18 p^5 + 6 p^6)) / (1 - 2 p + 2 p^2)
```

We can plot this as a function of p to get the following graph:



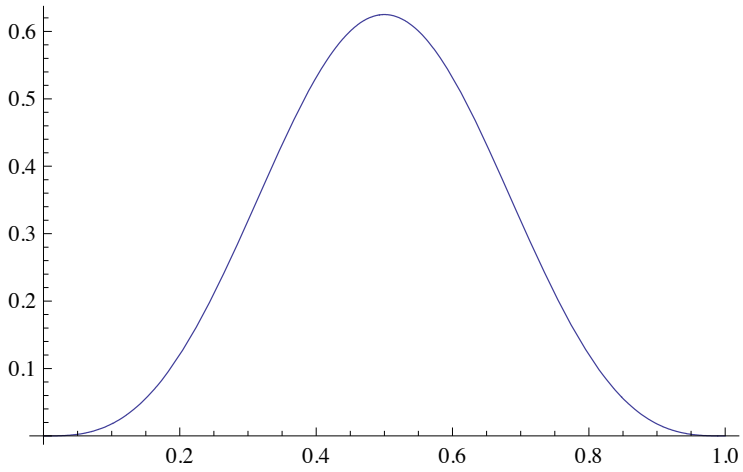
When $p = 0$, the course of the game is totally predictable and Amélie wins in 4 points. The same holds when $p = 1$, only it is Björn who wins with probabilty 1 this time. In between, we see that the expected game-length varies between 4 and about 7 (actually, the exact number is 6.75), and it longest when $p = \frac{1}{2}$.

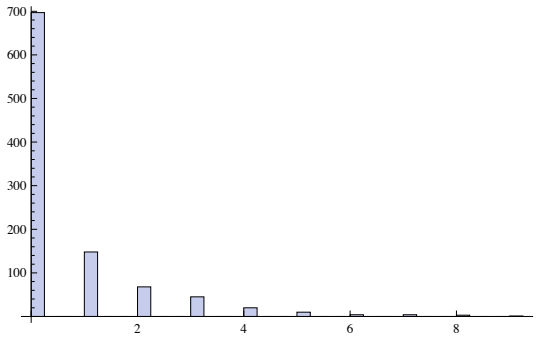
How about the expected number of deuces (scores $(40, 40)$)? We can compute that too by setting $g(i) = 0$ if $i \neq (40, 40)$ and $g((40, 40)) = 1$ (the number 16 used below is just the *Mathematica*

internal number of the state (40,40) among the transient states):

```
In[117]:= G = Table[{If[i == 16, 1, 0]}, {i, 1, 18}]

Out[117]= {{0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {0}, {1}, {0}, {0}}

In[118]:= f[p_] = Simplify[(F.G)[[1, 1]] //. {q → 1 - p}]

Out[118]= -(20 (-1 + p)^3 p^3 / (1 - 2 p + 2 p^2))
```

The plot of the obtained expressions, as a function of p, looks like this





Therefore, the expected number of deuces varies between 0 and a bit more than 0.6 (the exact number is 0.625 and corresponds to the case $p = \frac{1}{2}$). When asked, people would usually give a higher estimate for this probability. The reason is that the distribution of the number of deuces looks something like the picture on the left (a histogram of the number of deuces in a simulation of 1000 tennis games with $p = \frac{1}{2}$). We see that most of the games have no deuces. However, in the cases where a deuce happens, it is quite possible it will be repeated.

From a psychological point of view, we tend to forget all the games without deuces and focus on those with at least 1 deuce when we make predictions. In fact, the expected number of deuces *given that the game contains at least one deuce* is approximatly equal to 2.1.

Chapter 13

Stationary and Limiting Distributions

Transitions between different states of a Markov chain describe *short-time* behavior of the chain. In most models used in physical and social sciences, systems change states many times per second. In a rare few, the time scale of the steps can be measured in hours or days. What is of interest, however, is the long-term behavior of the system, measured in thousands, millions, or even billions of steps. Here is an example: for a typical liquid stock traded on the New York Stock Excahenge, there is a trade every few seconds, and each trade changes the price (state) of the stock a little bit. What is of interest to an investor is, however, the distribution of the stock-price in 6 months, in a year or, in 30 years - just in time for retirement. A back-of-an-envelope calculation shows that there are, approximately, 50 million trades in 30 years. So, a grasp of very-long time behavior of a Markov chain is one of the most important achievments of probability in general, and stochastic-process theory in particular. We only scratch the surface in this lecture.

13.1 Stationary and limiting distributions

Definition 13.1. A stochastic process $\{X_n\}_{n\in\mathbb{N}_0}$ is said to be **stationary** if the random vectors

$$(X_0, X_1, X_2, \ldots, X_k) \text{ and } (X_m, X_{m+1}, X_{m+2}, \ldots, X_{m+k})$$

have the same (joint) distribution for all $m, k \in \mathbb{N}_0$.

For stationary processes, all random variables $X_0, X_1, \ldots$ have the same distribution (just take $k = 0$ in the definition). That condition is, however, only necessary. The pairs (X_0, X_1) and (X_m, X_{m+1}) should be equally distributed as random vectors, the same for triplets, etc. Intuitively, a stochastic process is stationary if, statistically, it does not evolve. Its probabilistic behavior today is the same as its probabilistic behavior in a billion years. It is somethimes useful to think about stationarity in the following way; if a system is let to evolve for a long time, it will reach an equilibrium state and fluctuate around it forever. We can expect that such a system will look similar a million years from now and a billion years from now. It might, however, not resemble its present state at all. Think about a glass of water in which we drop a tiny drop of ink. Imediately after that, the glass will be clear, with a tiny black speck. The ink starts to diffuse and the spack starts to grow immediately. It won't be long before the whole glass is of uniform black color - the ink has permeated every last "corner" of the glass. After that, nothing much happens. The ink

will never spontaneously return to its initial state[1]. Ink is composed of many small particles which do not interact with each other too much. They do, however, get bombarded by the molecules of water, and this bombardment makes them behave as random walks which simply bounce back once they hit the glass wall (this phenomenon is called **diffusion**). Each ink particle will wander off in its own direction, and quite soon, they will be "everywhere". Eventally, the distribution of the ink in the glass becomes very close to uniform and no amount of further activity will change that - you just cannot get more "random" then the uniform distribution in a glass of water.

Let us get back to mathematics and give two simple examples; one of a process which is not stationary, and the other of a typical stationary process.

Example 13.2. The simple random walk is not stationary. Indeed, X_0 is a constant, while X_1 takes two values with equal probabilities, so they cannot have the same distribution.

For an example of a stationary process, take a regime switching chain $\{X_n\}_{n\in\mathbb{N}_0}$ with $p_{01} = p_{10} = 1$, and the initial distribution $\mathbb{P}[X_0 = 0] = \mathbb{P}[X_0 = 1] = \frac{1}{2}$. Then $X_n = X_0$ if n is even, and $X_n = 1-X_0$ if n is odd. Moreover, X_0 and $1-X_0$ have the same distribution (Bernoulli with $p = \frac{1}{2}$), and, so $X_0, X_1, \ldots$ all have the same distribution. How about k-tuples? Why do $(X_0, X_1, \ldots, X_k)$ and $(X_m, X_{m+1}, \ldots, X_{m+k})$ have the same distribution? For $i_0, i_1, \ldots, i_k \in \{0, 1\}$, by the Markov property, we have

$$\mathbb{P}[X_0 = i_0, X_1 = i_i, \ldots, X_k = i_k] = \mathbb{P}[X_0 = i_0]p_{i_0i_1}p_{i_1i_2}\ldots p_{i_{k-1}i_k} = \tfrac{1}{2}p_{i_0i_1}p_{i_1i_2}\ldots p_{i_{k-1}i_k}.$$

In the same manner,

$$\mathbb{P}[X_m = i_0, X_1 = i_i, \ldots, X_{m+k} = i_k] = \mathbb{P}[X_m = i_0]p_{i_0i_1}p_{i_1i_2}\ldots p_{i_{k-1}i_k} = \tfrac{1}{2}p_{i_0i_1}p_{i_1i_2}\ldots p_{i_{k-1}i_k},$$

so the two distributions are identical.

The second example above is quite instructive. We took a Markov chain and gave it an initial distribution with the property that X_0 and X_m have the same distribution for all $m \in \mathbb{N}_0$. Magically, the whole process became stationary. This is not a coincidence; we can play the same trick with any Markov chain, as long as the initial distribution with the above property can be found. Actually, such a distribution is so important that it even has a name:

Definition 13.3. A distribution $\pi = (\pi_i)_{i\in\mathcal{S}}$ on the state space $\mathcal{S}$ of a Markov chain with transition matrix P is called a **stationary distribution** if

$$\mathbb{P}[X_1 = i] = \pi_i \text{ for all } i \in \mathcal{S}, \text{ whenever } \mathbb{P}[X_0 = i] = \pi_i, \text{ for all } i \in \mathcal{S}.$$

In words, π is called a stationary distribution if the distribution of X_1 is equal to that of X_0 when the distribution of X_0 is π. Here is a hands-on characterization:

Proposition 13.4. *A vector $\pi = (\pi_i, i \in \mathcal{S})$ with $\sum_{i\in\mathcal{S}} \pi_i = 1$ is a stationary distribution if and only if*

$$\pi P = \pi,$$

when π is interpreted as a row vector. In that case the Markov chain with initial distribution π and transition matrix P is stationary and the distribution of X_m is π for all $m \in \mathbb{N}_0$.

[1] It will, actually, but it will take an unimaginably long time.

Proof. Suppose, first, that π is a stationary distribution, and let $\{X_n\}_{n\in\mathbb{N}_0}$ be a Markov chain with initial distribution $\boldsymbol{a}^{(0)} = \pi$ and transition matrix P. Then,

$$\boldsymbol{a}^{(1)} = \boldsymbol{a}^{(0)}P = \pi P.$$

By the assumption, the distribution $\boldsymbol{a}^{(1)}$ of X_1 is π. Therefore, $\pi = \pi P$.

Conversely, suppose that $\pi = \pi P$. We can directly prove more than just the fact that π is a stationary distribution. We can prove that the process $\{X_n\}_{n\in\mathbb{N}_0}$ is stationary. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a Markov chain with initial distribution π and transition matrix P. We need to show that $\{X_n\}_{n\in\mathbb{N}_0}$ is stationary. In order to do that, we first note that all random variables X_m, $m \in \mathbb{N}_0$, have the same distribution. Indeed, the distribution $\boldsymbol{a}^{(m)}$ of X_m is given by

$$\boldsymbol{a}^{(m)} = \boldsymbol{a}^{(0)}P^m = \pi P^m = (\pi P)P^{m-1} = \pi P^{m-1} = \cdots = \pi.$$

Next, we pick $m, k \in \mathbb{N}_0$ and a $k+1$-tuple $i_0, i_1, \ldots, i_k$ of elements of $\mathcal{S}$. By the Markov property, we have

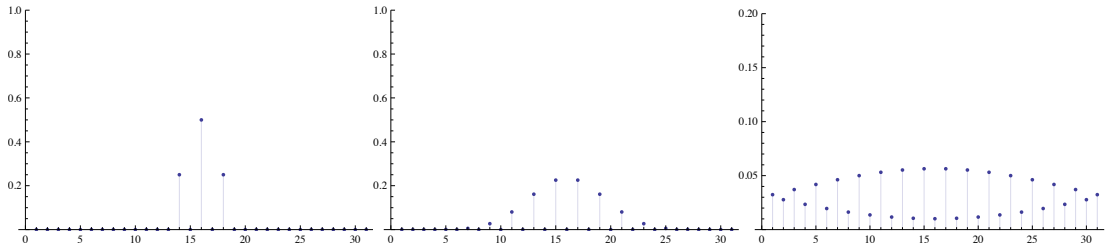
$$\mathbb{P}[X_m = i_0, X_{m+1} = i_1, \ldots, X_{m+k} = i_k] = \mathbb{P}[X_m = i_0]p_{i_0i_1}p_{i_1i_2}\cdots p_{i_{k-1}i_k} = \pi_{i_0}p_{i_0i_1}p_{i_1i_2}\cdots p_{i_{k-1}i_k}.$$

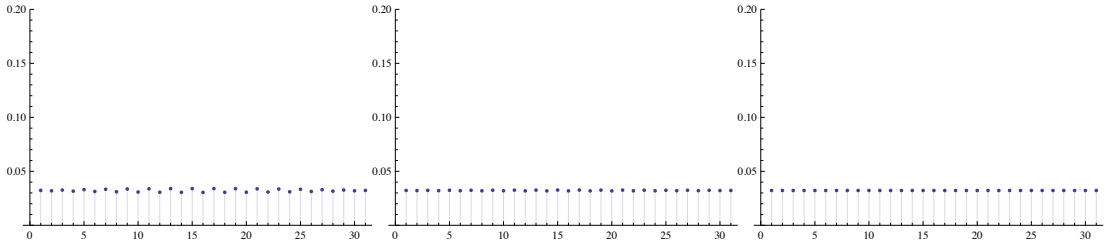
This last expression does not depend on m, so we can conclude that $\{X_n\}_{n\in\mathbb{N}_0}$ is stationary. □

Example 13.5 (A model of diffusion in a glass)**.** Let us get back to the story about the glass of water and let us analyze a simplified model of that phenomenon. Our glass will be represented by the set $\{0, 1, 2, \ldots, a\}$, where 0 and a are the positions adjacent to the walls of the glass. The ink particle performs a simple random walk inside the glass. Once it reaches the state 0 it either takes a step to the right to 1 (with probability $\frac{1}{2}$) or tries to go left (also with probability $\frac{1}{2}$). The passage to the left is blocked by the wall, so the particle ends up staying where it is. The same thing happens at the other wall. All in all, we get a Markov chain with the following transition matrix

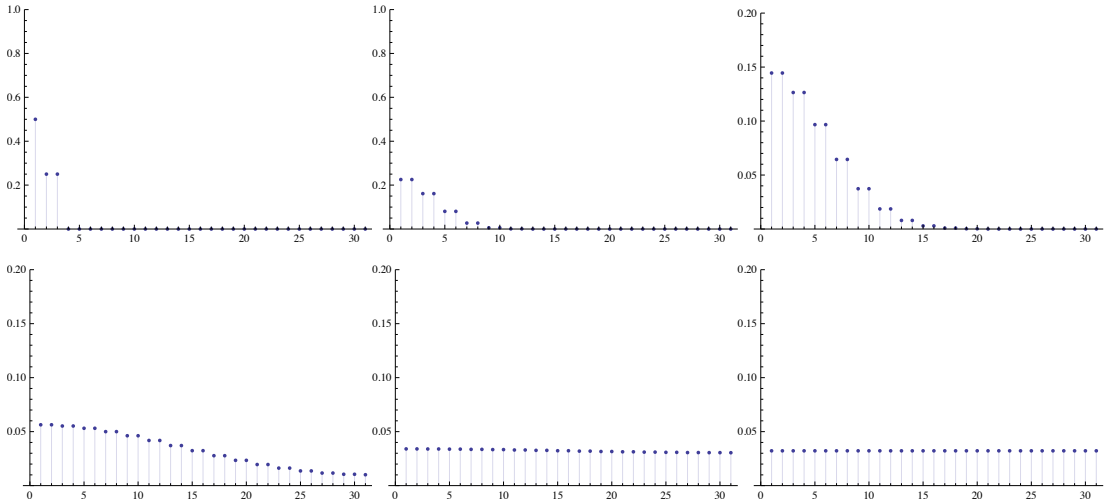
$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & \ldots & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \ldots & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \ldots & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 & \ldots & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 & \ldots & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Let us see what happens when we start the chain with a distribution concentrated in $a/2$ (assuming that a is even); a graphical representation of the distribution of X_3, X_{12}, X_{30}, X_{200}, X_{700} and X_{5000} when $a = 30$ represents the behavior of the system very well (the y axis is on a different scales on the first two plots):





How about if we start from a different initial distribution? Here are the same plots in that case:



As you can see, the distribution changes rapidly at first, but then, once it has reached the "equilibrium" it changes remarkably little (compare X_{700} and X_{5000}). Also, the "equilibrium" distribution is very uniform and *does not depend on the initial distribution*; this is exactly what you would expect from a long-term distribution of ink in a glass.

Let us show that the uniform distribution $\pi = (1/(a+1), 1/(a+1), \ldots, 1/(1+a)) = (1/31, 1/31, \ldots, 1/31)$ on $\{0, 1, 2, \ldots, a\}$ is indeed the (unique) stationary distribution. By Proposition 13.4, we need to solve the equation $\pi = P\pi$. If we write out this system of equations line by line, we get

$$\begin{aligned}
\pi_0 &= \tfrac{1}{2}(\pi_0 + \pi_1) \\
\pi_1 &= \tfrac{1}{2}(\pi_0 + \pi_2) \\
\pi_2 &= \tfrac{1}{2}(\pi_1 + \pi_3) \\
\pi_3 &= \tfrac{1}{2}(\pi_2 + \pi_4) \\
&\vdots \\
\pi_{a-1} &= \tfrac{1}{2}(\pi_{a-2} + \pi_a) \\
\pi_a &= \tfrac{1}{2}(\pi_{a-1} + \pi_a)
\end{aligned} \tag{13.1}$$

The first equation yields $\pi_0 = \pi_1$. Using that in the second one, we get $\pi_1 = \pi_2$, etc. Finally, we know that π is a probability distribution so that $\pi_0 + \pi_1 + \cdots + \pi_a = 1$. Therefore, $\pi_i = 1/(a+1)$ for all $i \in \mathcal{S}$, i.e., the uniform distribution on $\mathcal{S}$ is the only stationary distribution.

Can there be more than one stationary distribution? Can there be none? Sure, here is an example:

Example 13.6. For $P = I$, any distribution is stationary, and there are infinitely many.

A simple example where no stationary distribution exists can be constructed on an infinite state space. Take the Deterministically Monotone Markov Chain. The transition "matrix" looks like the identity matrix, with the diagonal of ones shifted to the right. Therefore, the system of equations $\pi = \pi P$ reads

$$\pi_1 = \pi_2, \pi_2 = \pi_3, \ldots, \pi_n = \pi_{n+1}, \ldots,$$

and so, for π to be a stationary distribution, we must have $\pi_n = \pi_1$ for all $n \in \mathbb{N}$. Now, if $\pi_1 = 0$, π is not a distribution (it sums to 0, not 1). But if $\pi_1 > 0$, then the sum is $+\infty$, so π is not a distribution either. Intuitively, the chain never stabilizes, it just keeps moving to the right ad infinitum.

The example with many stationary distributions can be constructed on any state space, but the other one, where no stationary distribution exists, had to use an infinite one. Was that necessary? Yes. Before we show this fact, let us analyze the relation between stationary distributions and the properties of recurrence and transience. Here is our first result:

Proposition 13.7. *Suppose that the state space $\mathcal{S}$ of a Markov chain is finite and let $\mathcal{S} = C_1 \cup C_2 \cup \cdots \cup C_m \cup T$ be its canonical decomposition, Then the following two statements are equivalent:*

1. *π is a stationary distribution, and*
2. *$\pi^{C_k} = \pi^{C_k} P_k$, $k = 1, \ldots, m$, and $\pi^T = (0, 0, \ldots, 0)$,*

where

$$P = \begin{bmatrix} P_1 & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & 0 \\ 0 & 0 & P_m & 0 \\ & R & & Q \end{bmatrix},$$

is the canonical form of the transition matrix, $\pi^{C_k} = (\pi_i, i \in C_k)$, $k = 1, 2, \ldots, m$ and $\pi^T = (\pi_i, i \in T)$.

Proof. We write the equation $\pi = \pi P$ coordinatewise as $\pi_j = \sum_{i \in \mathcal{S}} \pi_i p_{ij}$ and, by distinguishing the cases $i \in C_k$, $k \in \{1, 2, \ldots, m\}$, and $i \in T$, we get the following sytem of matrix equations (alternatively, just write the system $\pi = \pi P$ in the block-matrix form according to the cannonical decomposition above):

$$\pi^{C_k} = \pi_{C_K} P_{C_k} + \pi^T R, \ k = 1, \ldots, m, \text{ and } \pi^T = \pi^T Q.$$

The last equality can be read as follows: π^T is in a row null-space of $I - Q$. We know, however, that $I - Q$ admits an inverse, and so it is a regular square matrix. Its row null-space (as well as its column null-space) must be trivial, and, consequently, $\pi^T = 0$.

Having established that $\pi^T = 0$, we can de-couple the system of equations above and write it as

$$\pi^{C_k} = \pi_{C_K} P_k, \ k = 1, \ldots, m, \text{ and } \pi^T = (0, 0, \ldots, 0),$$

which is exactly what we needed to prove.

The other implication - the proof of which consists of a verification of the fact that each distribution from *(2)* above is indeed a stationaty distribution - is left to the reader. □

The moral of the story of Proposition 13.7 is the following: in order to compute the stationary distribution(s), classify the states and find the canonical decomposition of the state space. Then, set $\pi_i = 0$ for any transient state i. What remains are recurrent classes, and you can analyize each one separately. Note, however, that π^{C_k} does not need to be a real distribution on C_k, since $\sum_{i \in C_k} \pi_i^{C_k}$ does not need to equal 1. However, unless $\pi^{C_k} = (0, 0, \ldots, 0)$, we can always multiply all its elements by a constant to make the sum equal to 1.

Thanks to the results of Proposition 13.7, we can focus on Markov chains with only one class:

Definition 13.8. A Markov chain is said to be **irreducible** if there is only one class.

When a Markov chain is finite and irreducible, all of its states must be recurrent. Indeed, a finite Markov chain has at least one recurrent state, and recurrence is a class property.

Now that we have described the structure of the set of all stationary distributions, we still need to to tackle the question of existence: are there any stationary distributions? In addition to giving an affirmative answer to this question, we will also show how to construct a stationary distribution and how to interpret it probabilistically.

Proposition 13.9. *Let $\{X_n\}_{n \in \mathbb{N}_0}$ be an irreducible Markov chain with a finite state space. Then, the following two statements hold:*

1. *All states are positive recurrent.*
2. *Let i be a fixed (but arbitrary) recurrent state. Let*

$$\nu_j = \mathbb{E}_i \left[\sum_{n=0}^{\tau_i(1)-1} \mathbf{1}_{\{X_n = j\}} \right], \ j \in \mathcal{S}$$

be the expected number of visits to state j in between two consecutive visits to state i. Then the vector π, given by $\pi_j = \frac{1}{m_i}\nu_j$, $j \in \mathcal{S}$ (where $m_i = \mathbb{E}_i[\tau_i(1)]$) is a stationary distribution.

Remark 13.10. Even though it is not exceedingly hard, the proof of this proposition is a bit technical, so we omit it. It is important, however, to understand what the propositions states:

1. the expected number of visits to the state j in between two consecutive visits to the state i can be related to a stationary distribution of the Markov chain by $\nu_j = m_i \pi_j$, and
2. when we set $j = i$, ν_i counts the number of visits to i between two consecutive visits to i, which is always equal to 1 (the first visit is counted and the last one is not). Therefore, $\nu_i = 1$, and so $\pi_i = \frac{1}{m_i}$.

Proposition 13.9 and Remark 13.10 are typically used in the following way: one computes the unique stationary distribution π by solving the equation $\pi = \pi P$ and then draws conclusions about m_i or the ν_j's.

Note also that the computation above is not a special case of a reward computation of the previous lecture. There, you need to start from a transient state and finish in a recurrent state, while here your starting and final state are the same.

13.2 Limiting distributions

Example 13.5 shows vividly how the distribution of ink quickly reaches the uniform equilibrium state. It is no coincidence that this "limiting" distribution happens to be a stationaty distribution. Before we make this claim more precise, let us define rigorously what we mean by a limiting distribution:

Definition 13.11. A distribution $\pi = (\pi_i, i \in \mathcal{S})$ on the state space $\mathcal{S}$ of a Markov chain with transition matrix P is called a **limiting distribution** if

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j,$$

for all $i, j \in \mathcal{S}$.

Note that for π to be a limiting distribution, all the limits in Definition 13.11 must exist. Once they do, π is automatically a distribution: $\pi_j \geq 0$ (as a limit of non-negative numbers) and

$$\sum_{j\in\mathcal{S}} \pi_j = \sum_{j\in\mathcal{S}} \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{j\in\mathcal{S}} p_{ij}^{(n)} = \lim_{n\to\infty} 1 = 1.$$

Also, note that the independence on the initial state i is built into the definition of the limiting distribution: the sequence $\{p_{ij}^{(n)}\}_{n\in\mathbb{N}}$ must tend to the same limit π_j for all $i \in \mathcal{S}$. Moreover, it follows immediately there can be at most one limiting distribution.

The connection with stationary distributions is spelled out in the following propositions:

Proposition 13.12. *Suppose that a Markov chain with transition matrix P admits a limiting distribution $\pi = (\pi_i, i \in \mathcal{S})$. Then π is a stationary distribution.*

Proof. To show that π is a stationary distribution, we need to verify that it satisfies $\pi = \pi P$, i.e., that

$$\pi_j = \sum_{i\in\mathcal{S}} \pi_i p_{ij}.$$

We use the Chapman-Kolmogorov relation $p_{ij}^{(n+1)} = \sum_{k\in\mathcal{S}} p_{ik} p_{kj}^{(n)}$ and start from the observation that $\pi_j = \lim_{n\to\infty} p_{ij}^{(n+1)}$ to get exactly what we need:

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n+1)} = \lim_{n\to\infty} \sum_{k\in\mathcal{S}} p_{ik}^{(n)} p_{kj} = \sum_{k\in\mathcal{S}} (\lim_{n\to\infty} p_{ik}^{(n)}) p_{kj} = \sum_{k\in\mathcal{S}} \pi_k p_{kj}.$$

□

Example 13.13. Limiting distributions don't need to exist, even when there are stationary ones. Here are two examples:

1. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a Regime-switching chain with

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Then

$$p_{ij}^{(n)} = \begin{cases} \frac{1}{2}(1+(-1)^n), & i = j \\ \frac{1}{2}(1+(-1)^{n+1}), & i \neq j. \end{cases}$$

Clearly, the sequence $p_{ij}^{(n)}$ is alternating between 0 and 1, so it cannot admit a limit. A stationary distribution exists: the system $\pi = \pi P$ becomes

$$\pi_0 = 0 \times \pi_0 + 1 \times \pi_1,$$
$$\pi_1 = 1 \times \pi_0 + 0 \times \pi_1.$$

2. In the previous example the limiting distribution did not exist which implies that $\pi_0 = \pi_1$. In order for π to be a distribution on $\{0, 1\}$, we must have $\pi_0 = \pi_1 = \frac{1}{2}$. So, $(\frac{1}{2}, \frac{1}{2})$ is a stationary distribution.

3. In the previous example the limiting distribution did not exists because the limits of the sequences $p_{ij}^{(n)}$ did not exist. Another reason for the non-existence of limiting distributions can be dependence on the initial conditions: the limits $\lim_n p_{ij}^{(n)}$ may exist for all i, j, but their values can depend on the intial state i. The simplest example is a Markov chain with two states $i = 1, 2$, where $p_{11} = p_{22} = 1$. There are two recurrent (and therefore closed) classes, and the chain remains in the state it starte in forever. Therefore, $\lim_n p_{12}^{(n)} = 0$ and $\lim_n p_{22}^{(n)} = 1$, so no limiting distribution exists despite the fact that all limits $\lim_n p_{ij}^{(n)}$ do.

Proposition 13.14. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a finite-state, irreducible and aperiodic Markov chain. Then the limiting distribution exists.*

We conclude the discussion of limiting distributions with a version of the Law of Large Numbers (LLN) for Markov chains. Before we state it, we recall the classical LLN for independent variables:

Theorem 13.15. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a sequence of independent and identically distributed random variables, such that $\mathbb{E}[|X_0|] < \infty$. Then*

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} X_k = \mathbb{E}[X_0],$$

almost surely [2].

When $\{X_n\}_{n\in\mathbb{N}_0}$ is a Markov chain, two problems occur. First, the random variables $X_0, X_1, \ldots$ are not independent. Second, X_k takes its values in the state space $\mathcal{S}$ which does not necessarily consist of numbers, so the expression $X_0 + X_1$ or $\mathbb{E}[X_0]$ does not make sense all the time. To deal with the second problem, we pick a numerical function $f : \mathcal{S} \to \mathbb{R}$ (give each state a value) and form sums of the form $f(X_0) + f(X_1) + \cdots + f(X_{n-1})$. Independence is more subtle, but it can be replaced with stationarity (loosely speaking), as in the following proposition (we skip the proof):

[2] *Almost surely* means *with probability one*, and states that this statement is true all the time for all practical purposes. It may fail, but only in extremely exceptional cases such as the one where you toss a fair coin infinitely many times and get *tails* every single time. That means never, really.

Proposition 13.16. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be an irreducible Markov chain with a finite state space $\mathcal{S}$, let $\pi = (\pi_i, i \in \mathcal{S})$ be the (unique) stationary distribution, and let $f : \mathcal{S} \to \mathbb{R}$ be an arbitrary function. Then*

$$\lim_n \frac{\sum_{k=0}^{n-1} f(X_k)}{n} = \sum_{i\in\mathcal{S}} f(i)\pi_i,$$

almost surely, no matter what initial distribution we choose.

Example 13.17. Let $\{X_n\}_{n\in\mathbb{N}_0}$ be an irreducible finite Markov chain, and let the function $f : \mathcal{S} \to \mathbb{R}$ be given by

$$f(j) = \mathbf{1}_{(i=j)} = \begin{cases} 1, & j = i, \\ 0, & j \neq i. \end{cases}$$

Then, by the Law of Large Numbers, we gave

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{\{X_k=i\}} = \sum_{j\in\mathcal{S}} \mathbf{1}_{(i=j)}\pi_j = \pi_i,$$

so that π_i can be interpreted as a long-run percentage of time that the Markov chain $\{X_n\}_{n\in\mathbb{N}_0}$ spends in the state i.