MVE137: Probability and Statistical Learning Using Python

Giuseppe Durisi
Chalmers University of Technology

August 25, 2022

Contents

1 Introduction

The course will provide the participants with a solid foundation of probability theory and statistical learning. In particular, in this course the participants will become familiar with key probabilistic and statistical concepts in data science and will learn how to apply them to analyze data sets and draw meaningful conclusions from data. The course will cover both theoretical and practical aspects, with the objective of preparing the participants to apply the acquired knowledge to the real world. The participants will have the possibility to experiment and practice with the concepts taught in the course via Python programs and the Jupyter Notebook platform.

1.1 Learning Outcomes

- Explain probability concepts such as tail probability bounds, moment-generating functions and their applications, Markov chains, and central limit theorems.
- Explain statistical models and methods that are used for prediction in science and technology, such as regression- and classification-type statistical models.
- Select suitable statistical models to analyze existing data sets, apply sound statistical methods, and perform analyses using Python.
- Discuss the use of common Python libraries such as numpy, matplotlib, jupyter notebook, pandas, to perform data analysis.
- Design Python-programs that apply the probability and statistical learning concepts presented in the class, to draw meaningful conclusions from data.

1.2 Literature

The course will be mainly based on the following two references:

- M. Mitzenmacher and E. Upfal, "Proability and computing: randomization and probabilistic techniques in algorithms and data analysis". Cambridge, U.K.: Cambridge Univ. Press, 2017.
- T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction", 2nd ed. Springer, New York, NY, U.S.A., 2017.

The first book will be used in the first part of the course, which deals with probability. The lecture notes you are currently reading cover (roughly) the material in this book that is relevant for the course. These lecture notes can be used as a replacement of this book, although the book provides a larger

number of examples. The second part of the course will be based on the second reference, which is available electronically via Chalmers library. Additional references providing an introduction to Python, including instructions on how to install Python on your personal computer, are provided on the course website.

1.3 Prerequisites

A bachelor-level knowledge of probability and Python. For the students with no prior knowledge of Python, pointers to tutorials will be provided. The following book provides an excellent review of basic results in probability theory:

- R. D. Yates & D. J. Goodman, "Probability and Stochastic Processes", 3rd edition, Wiley, Singapore, 2015.

1.4 Some Comments on these Lecture Notes

These lecture notes will cover the first part of the course, which deals with probability theory. For the statistical-learning part, we will follow closely the material in the second reference that has been provided. Theoretical results in these notes are given mostly as theorems. Their proofs are not included in the lecture notes. Some of the proof will be detailed during the lectures. They can also be found in the textbook. Examples to clarify some of the theoretical results will be discussed during the lectures, the homework assignments, and the python laboratories.

2 Events and Probability

2.1 A Motivating Example

We will start by reviewing some basic definitions in probability theory. To introduce and motivate these definitions, we will consider throughout this chapter the following practical problem. Assume that you want to develop an algorithm, which will turn out to be *randomized*, to verify *polynomial identities*.

The setup is as follows: you have an expression involving a product of monomials, and you want reduce it to its canonical form. Here is a concrete example: Assume you have the following expression

$$F(x) = (x+1)(x-2)(x+3)(x-4)(x+5)(x-6). \tag{1}$$

We want to verify whether it can be rewritten in the following canonical form:

$$G(x) = x^6 + 3x^5 - 41x^4 - 153x^3 + 220x^2 + 1230x + 900. \tag{2}$$

One way to verify whether $F(x) = G(x)$, is to perform all the involved multiplications. Let us count how many multiplications are needed. The underlying assumption is that additions are cheap, and we can ignore their cost. You can verify that if you have d monomials, so that the resulting polynomial in canonical form has degree d, you need $d^2 + d - 2$ multiplications. So, roughly speaking, the number of multiplications scales quadratically with the degree of the polynomial.

We will discuss next a method, which involves the use of randomness, that allows us to verify probabilistically the correctness of our computation using a much smaller number of multiplications. Here is the idea. Let us denote by $F(x)$ the product of monomials and by $G(x)$ the canonical form, whose correctness we want to verify. Suppose that these two polynomials have degree d and suppose we have a method to generate an integer r uniformly over the interval $\{1, \dots, 100d\}$. Note that evaluating $F(r)$ requires only $d-1$ multiplications, while evaluating $G(r)$ requires at most additional $2d-1$ multiplications. So this can be done much more efficiently. Clearly, if $F(r) \neq G(r)$, we can conclude straight away that our computation is incorrect. However, it may be possible that $F(r) = G(r)$ for our chosen r, although the polynomial $F(x)$ is different from the polynomial $G(x)$. For this to happen, the chosen r must be a root of the equation $F(x) - G(x) = 0$. Since these two polynomials have degree d, the number of roots is no more than d. So what is the probability that we selected by chance a root of the polynomial and our verification method returns a wrong answer?

2.2 Probability Space

To answer this question, we need to introduce some definitions.

Definition 1 (Sample space) *The sample space Ω is the set of all possible outcomes of a random experiment.*

Throughout a large portion of these notes, we will think of Ω as a discrete set. Furthermore, we will think of all possible probabilistic *events* we may be interested in as subsets of the sample space Ω.

Definition 2 (σ-field) *The allowable events constitute a family of sets $\mathcal{F}$, usually referred to as σ-field. Each set in $\mathcal{F}$ is a subset of the sample space Ω.*

When Ω is a finite or countable set, we can think of $\mathcal{F}$ as the collection of all possible subsets of Ω. The construction of $\mathcal{F}$ turns out to be more delicate for uncountable sets, such as the set of real numbers. We will not get into the reasons behind this in this course. The reader may want to consult more advanced textbooks, such as [**?**], to understand the *measurability issues* involved with the extension of the notion of σ-fields to uncountable sets.

Since probabilistic events are subsets of Ω, we will use set-theory notation to describe combinations of events. For example, if $\mathcal{A}$ and $\mathcal{B}$ are events, we will use $\mathcal{A}\cup\mathcal{B}$, $\mathcal{A}\cap\mathcal{B}$, $\bar{\mathcal{A}}$, and $\mathcal{A}\setminus\mathcal{B}$, to denote $\mathcal{A}$ **or** $\mathcal{B}$, $\mathcal{A}$ **and** $\mathcal{B}$, **not** $\mathcal{A}$, and $\mathcal{A}$ **but not** $\mathcal{B}$, respectively. We also say that $\mathcal{A}$ and $\mathcal{B}$ are disjoint if their intersection is the empty set, which we denote by $\emptyset$.

Definition 3 (Probability measure and probability space) *A probability measure on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ that satisfies the following two properties:*

1. $\mathbb{P}[\Omega] = 1$
2. *If $\mathcal{A}_1, \mathcal{A}_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, i.e., $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for all pairs i, j satisfying $i \neq j$, then*

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} \mathcal{A}_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[\mathcal{A}_i]. \tag{3}$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a ***probability space****.*

We next state some important properties of probability measures, which we will use repeatedly during the course.

Lemma 4 (Basic properties of probability measures) *The following properties hold:*

- $\mathbb{P}[\emptyset] = 0$
- $\mathbb{P}[\bar{\mathcal{A}}] = 1 - \mathbb{P}[\mathcal{A}]$
- *If $A \subset B$, then $\mathbb{P}[\mathcal{B}] = \mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{B} \setminus \mathcal{A}] \geq \mathbb{P}[\mathcal{A}]$*
- $\mathbb{P}[\mathcal{A} \cup \mathcal{B}] = \mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{B}] - \mathbb{P}[\mathcal{A} \cap \mathcal{B}]$

- *If $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are events, then*

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} \mathcal{A}_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i] - \sum_{i<j} \mathbb{P}[\mathcal{A}_i \cap \mathcal{A}_j] + \sum_{i<j<k} \mathbb{P}[\mathcal{A}_i \cap \mathcal{A}_j \cap \mathcal{A}_k] - \ldots + (-1)^{n+1} \mathbb{P}[\mathcal{A}_1 \cap \mathcal{A}_2 \cap \ldots \cap \mathcal{A}_n]. \quad (4)$$

This last property is often referred to as ***inclusion-exclusion principle****.*

It follows from Lemma 4 that for every two events $\mathcal{A}$ and $\mathcal{B}$,

$$\mathbb{P}[\mathcal{A} \cup \mathcal{B}] \leq \mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{B}]. \quad (5)$$

This inequality, which is usually referred to as **union bound**, is tremendously useful despite its simplicity. It can be generalized to finite or countably infinite sequence of events as follows: given the sets $\mathcal{A}_1, \mathcal{A}_2, \ldots$, we have that

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} \mathcal{A}_i\right] \leq \sum_{i=1}^{\infty} \mathbb{P}[\mathcal{A}_i]. \quad (6)$$

Let us now go back to our motivating example, and let us use the definitions just introduced to answer our question. Let $\mathcal{E}$ be the event that our randomized verification algorithm returns the wrong answer, i.e., the algorithm tells that the equality is correct, even if it is not. The set $\mathcal{E}$ contains the roots of the polynomial $F(x) - G(x)$ that are in the set of integers $\{1, \ldots, 100d\}$, which is our sample space Ω. Now, since the polynomial $F(x) - G(x)$ has no more than d roots, we conclude that the event $\mathcal{E}$ contains at most d integers. Since we have assumed that the numbers in the set $\{1, \ldots, 100d\}$ are extracted uniformly at random, i.e, $\mathbb{P}[\{r\}] = 1/(100d)$ for all $r \in \{1, \ldots, 100d\}$, we conclude that

$$\mathbb{P}[\mathcal{E}] \leq \frac{d}{100d} = \frac{1}{100}. \quad (7)$$

So the verification algorithm we have just introduced identifies that an equality between polynomials in not correct 99% of the times.

2.3 Sampling, Independence, and Conditional Probability

Say that we want to improve the performance of our algorithm. One way to do so is to choose our random number r from a larger range of integers. For example, if we choose as sample space $\Omega = \{1, \ldots, 1000d\}$, then the success probability of our algorithm increases to 99.9%. However, this approach is unavoidably limited by the numerical precision available on the computer we use to run the algorithm.

An alternative, more promising approach, is to repeat the algorithm multiple times, using different randomly generated samples to test the identity. The motivation for this approach is that if the verification fails for one sample, then

we can conclude immediately that the identity is wrong. Such a property is sometimes referred to as **one-sided error**.

So the idea is to choose repeatedly a random number in $\Omega = \{1, \dots, 100d\}$. This strategy is generally referred to as **sampling** (do not confuse this with other notions of sampling you are familiar with; this has nothing to do with the sampling theorem!). Sampling can be performed in two ways, **with replacement or without replacement**. In sampling with replacement, we keep on selecting an outcome from Ω according to the underlying probability measure (in our case, the uniform distribution), regardless of the previous choices. So there is some chance that we will choose the same number r multiple times. In sampling without replacement, once we choose a number r, we do not allow the same number to be selected on subsequent runs of the algorithm. So the number chosen on a given run is sampled uniformly at random over the set of all previously unselected numbers.

2.3.1 Sampling with Replacement

When we run the algorithm k times, the probability that the algorithm fails is the probability that the algorithm declares that the two polynomials are identical in each of the runs. To understand how to compute the probability of this event, we need the following definition.

Definition 5 (Independence) *Events $\mathcal{A}$ and $\mathcal{B}$ are called independent if*

$$\mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[A]\,\mathbb{P}[B]. \tag{8}$$

More generally, a family $\{A_i, i \in \mathcal{I}\}$ is called independent if

$$\mathbb{P}\left[\bigcap_{i \in \mathcal{J}} \mathcal{A}_i\right] = \prod_{i \in \mathcal{J}} \mathbb{P}[\mathcal{A}_i] \tag{9}$$

for all finite subsets $\mathcal{J}$ of $\mathcal{I}$. If the family has the weaker property that

$$\mathbb{P}[\mathcal{A}_i \cap \mathcal{A}_j] = \mathbb{P}[A_i]\,\mathbb{P}[A_j] \tag{10}$$

*for all $i \neq j$, then it is called **pairwise independent**.*

Note that pairwise independent families are not necessarily independent.

In our setup, the choice of the sample at each run is independent of the choice at the previous iterations. Let us assume that the polynomials are not equivalent and let us determine the probability that the algorithm fails to realize so in k runs. Let $\mathcal{E}_i$ be the event that the algorithm fails at run i, i.e., that it selects a root of the polynomial $F(x) - G(x)$. The probability that the algorithm returns the wrong answer is then

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_k] = \prod_{i=1}^{k} \mathbb{P}[\mathcal{E}_i] \leq \left(\frac{1}{100}\right)^k. \tag{11}$$

Note that the probability of error decays exponentially with the number of trials.

2.3.2 Sampling without Replacement

Let us now look at sampling without replacement. When we perform sampling without replacement, the number that we choose on the ith run is *conditioned* on the numbers chosen in the previous runs. The next definition formalizes this concept.

Definition 6 (Conditional probability) *If $\mathbb{P}[\mathcal{B}] > 0$, the conditional probability that $\mathcal{A}$ occurs given that $\mathcal{B}$ occurs is*

$$\mathbb{P}[\mathcal{A} \,|\, \mathcal{B}] = \frac{\mathbb{P}[\mathcal{A} \cap \mathcal{B}]}{\mathbb{P}[\mathcal{B}]}. \tag{12}$$

The intuition behind this definition is as follows: if $\mathcal{B}$ occurs then $\mathcal{A}$ occurs only if $\mathcal{A} \cap \mathcal{B}$ occurs. So the conditional probability must be proportional to $\mathbb{P}[\mathcal{A} \cap \mathcal{B}]$. This means that $\mathbb{P}[\mathcal{A} \,|\, \mathcal{B}] = \alpha \, \mathbb{P}[\mathcal{A} \cap \mathcal{B}]$, for some constant α, which may depend on $\mathcal{B}$. Now since $\mathbb{P}[\Omega \,|\, \mathcal{B}] = 1$, we conclude that $\alpha \, \mathbb{P}[\Omega \cap \mathcal{B}] = \alpha \, \mathbb{P}[\mathcal{B}] = 1$. But this implies that $\alpha = 1/\mathbb{P}[\mathcal{B}]$.

Note that if $\mathcal{A}$ and $\mathcal{B}$ are independent and if $\mathbb{P}[\mathcal{B}] > 0$, then

$$\mathbb{P}[\mathcal{A} \,|\, \mathcal{B}] = \mathbb{P}[\mathcal{A}]. \tag{13}$$

Let us now go back to our example. As before, we need to evaluate $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_k]$. Since the $\{\mathcal{E}_i\}$ are not independent, we use (12) repeatedly and obtain

$$\begin{aligned} \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_k] &= \mathbb{P}[\mathcal{E}_k \,|\, \mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_{k-1}] \, \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_{k-1}] \qquad (14) \\ &= \mathbb{P}[\mathcal{E}_k \,|\, \mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_{k-1}] \times \\ &\quad \cdots \times \mathbb{P}[\mathcal{E}_3 \,|\, \mathcal{E}_1 \cap \mathcal{E}_2] \, \mathbb{P}[\mathcal{E}_2 \,|\, \mathcal{E}_1] \, \mathbb{P}[\mathcal{E}_1]. \qquad (15) \end{aligned}$$

To bound each term $\mathbb{P}[\mathcal{E}_j \,|\, \mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_{j-1}]$, where we assume that $j - 1 < d$, we note that if $\mathcal{E}_1, \ldots, \mathcal{E}_{j-1}$ occurred, this means that we found $j - 1$ out of the d roots of the polynomial $F(x) - G(x)$. So in run j we are left with $100d - (j-1)$ possible integers to choose and $d - (j-1)$ possible roots. Hence,

$$\mathbb{P}[\mathcal{E}_j \,|\, \mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_{j-1}] \leq \frac{d - (j-1)}{100d - (j-1)}. \tag{16}$$

This implies that the probability that the algorithm returns an erroneous answer after $k \leq d$ runs is

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \ldots \cap \mathcal{E}_k] \leq \prod_{j=1}^{k} \frac{d - (j-1)}{100d - (j-1)}. \tag{17}$$

One can show that the right-hand side of (17) is smaller or equal to $(1/100)^k$. So the bound on the error probability without replacement is better than the one with replacement. In practice, one often considers sampling with replacement

instead of sampling without replacement, because sampling with replacement is easier to code and to analyze theoretically.

Note that if we run the algorithm $d+1$ times without replacement, we are guaranteed to get the right answer. Indeed, assume that $F(x) \neq G(x)$. If, by chance, the first d runs of the algorithm generated the d roots of $F(x) - G(x)$, the $(d+1)$th run is bound to generate a number $r \in \Omega$ for which $F(r) - G(r) \neq 0$. However, running the algorithm dth times requires computing $d(d-1)$ multiplications, which is of the same order as the standard approach in which we perform all multiplications required to put in canonical form the product of monomials in (1).

2.4 Law of Total Probability and Bayes Theorem

We next present some additional results that we shall use throughout the course.

Theorem 7 (Law of Total Probability) *Let $\mathcal{E}_1, \dots, \mathcal{E}_n$ be a **partition** of Ω, i.e., a collection of sets satisfying $\mathcal{E}_i \cap \mathcal{E}_j = \emptyset$ when $i \neq j$, and $\bigcup_{i=1}^n \mathcal{E}_i = \Omega$. Assume also that $\mathbb{P}[\mathcal{E}_i] > 0$ for all i. Then the probability of an event $\mathcal{A}$ can be written as*

$$\mathbb{P}[\mathcal{A}] = \sum_{i=1}^{n} \mathbb{P}[\mathcal{A} \,|\, \mathcal{E}_i]\, \mathbb{P}[\mathcal{E}_i]. \tag{18}$$

The typical way we use this theorem is as follows: $\mathcal{A}$ involves many sources of randomness, and using the law of total probability, we can fix some of these sources of randomness and treat them as deterministic quantities. As we shall see, this is often convenient in theoretical analyses.

A related theorem is provided below.

Theorem 8 (Bayes' Theorem) *Assume that $\mathcal{E}_1, \dots, \mathcal{E}_n$ is a partition of Ω. Then*

$$\mathbb{P}[\mathcal{E}_j \,|\, \mathcal{B}] = \frac{\mathbb{P}[\mathcal{E}_j \cap \mathcal{B}]}{\mathbb{P}[\mathcal{B}]} = \frac{\mathbb{P}[B \,|\, \mathcal{E}_j]\, \mathbb{P}[\mathcal{E}_j]}{\sum_{i=1}^{n} \mathbb{P}[\mathcal{B} \,|\, \mathcal{E}_i]\, \mathbb{P}[\mathcal{E}_i]}. \tag{19}$$

2.5 Exercises

Exercise 1 (Coin flips) *We flip a fair coin 10 times. Find the probability of the following events.*

1. *The number of heads and the number of tails are equal.*
2. *There are more heads than tails.*
3. *The ith flip and the $(11-i)$th flips are the same for $i = 1, \dots, 5$.*
4. *We flip at least four consecutive heads.*

Exercise 2 (Families) *Jane has three children, each of which is equally likely to be a boy or a girl, independently. Define the following three events*

- $\mathcal{A} = \{$*all children are of the same sex*$\}$
- $\mathcal{B} = \{$*there is at most one boy*$\}$
- $\mathcal{C} = \{$*the family includes a boy and a girl*$\}$

1. *Show that $\mathcal{A}$ is independent of $\mathcal{B}$ and that $\mathcal{B}$ is independent of $\mathcal{C}$.*
2. *Is $\mathcal{A}$ independent of $\mathcal{C}$?*
3. *Do these results hold if boys and girls are not equally likely?*
4. *Do these results hold if Jane has four children?*

Exercise 3 (Virus test) *A medical company is assessing a newly developed test for a certain virus. The so called false negative rate is small: if you have contracted the virus, the probability that the test returns a positive result is* 0.999*. The false positive rate is also small: if you do not have the virus, the probability that the test returns a positive result is only* 0.005*. Assume that* 2% *of the population has contracted the virus. If a person chosen uniformly from the population is tested and the result comes back positive, what is the probability that the person has the virus?*

python*: Draw $N = 10^4$ pseudo random samples representing the infection status of N individuals. Use these samples and the statistics described above for the virus test to draw pseudo random samples representing a positive of negative virus test for each individual and verify your results from above.*

Exercise 4 (Birthdays) *Assume that m students born on independent days in* 1999 *are attending a lecture. Show that the probability that at least two of them share the same birthday is*

$$p = 1 - \frac{1}{365^m} \frac{365!}{(365-m)!}. \tag{20}$$

Verify that if there are at least $m = 23$ students, p is larger than $1/2$.

python*: Write a script drawing $m \in \{5, 23, 50\}$ birthdays n times to confirm the above equation. Plot the empirical (= based on you simulated data) and theoretical probability against the number of draws n for $1 \leq n \leq 2000$. How many draws do you need for the empirical probability to be representative of the theoretical probability for the different values of m? Interpret your observations.*

3 Discrete Random Variables and Expectation

3.1 Random Variables and Expectation

When studying random events, we are often interested in some value associated with the random event rather than in the random event itself. Assume for example that we roll two dice. The sample space consists of 36 events with equal probability: the ordered pair of numbers $\{(1,1),(1,2),\ldots,(6,5),(6,6)\}$. But when playing games, we are often interested in the sum of the two dice, which gives us 11 events with unequal probability, which are the 11 possible outcomes of the sum. Any such function from the sample space to the real numbers is called a *random variable*.

Definition 9 (Random variable) *A random variable X on a sample space Ω is a real-valued (measurable) function on Ω; that is $X : \Omega \to \mathbb{R}$. A discrete random variable is a random variable that takes on only a finite or countably infinite number of values.*

In this course, we will follow the convention of denoting random variables by capital letters, such as X and Y, while real numbers are usually denoted by lower-case letters. By saying that the random variable X takes values a, i.e., $X = a$, we refer to the set $\{s \in \Omega : X(s) = a\}$. As a consequence the probability that $X = a$ is given by

$$\mathbb{P}[X = x] = \sum_{\omega \in \Omega : X(\omega) = x} \mathbb{P}[\{\omega\}]. \tag{21}$$

For example, if the random variable X represents the sum of two dice, the event $X = 4$, corresponds to the set of equally probable "basic" events $\{(1,3),(2,2),(3,1)\}$. Hence, $\mathbb{P}[X = 4] = 3/36 = 1/12$, where we used that all basic events are equiprobable.

Definition 10 (Independence of random variables) *Two random variables X and Y are independent if and only if*

$$\mathbb{P}[(X = x) \cap (Y = y)] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y] \tag{22}$$

for all values x and y. Similarly, random variables $X_1, \ldots, X_k$ are mutually independent if and only if for any subset $\mathcal{I} \subset [1, \ldots, k]$, and any value x_i, $i \in \mathcal{I}$

$$\mathbb{P}\left[\bigcap_{i \in \mathcal{I}} (X_i = x_i)\right] = \prod_{i \in \mathcal{I}} \mathbb{P}[X_i = x_i]\,. \tag{23}$$

An important parameter of a random variable is its *expectation*, also called *mean*. It is a weighted average of the values assumed by the random variable, weighted with respect to the probability that the random variable assumes that value.

Definition 11 (Expectation) *The expectation of a discrete random variable X, denoted by $\mathbb{E}[X]$ is given by*

$$\mathbb{E}[X] = \sum_x x\, \mathbb{P}[X = x] \tag{24}$$

where the sum is over all values in the range of X.

The expectation of a random variable can be unbounded. Consider for example the random variable X that takes on the value 2^i with probability 2^{-i} for $i = 1, 2, \ldots$. This is a valid probability distribution, since $\sum_{i=1}^{\infty} 2^{-i} = 1$. The expected value of X is

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} 2^i 2^{-i} = \sum_{i=1}^{\infty} 1 = \infty. \tag{25}$$

3.1.1 Linearity of the Expectation

A fundamental property of the expectation is its *linearity*.

Theorem 12 (Linearity of Expectation) *For any finite collection of discrete random variables $X_1, X_2, \ldots, X_n$ with finite expectation*

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]\,. \tag{26}$$

Note that linearity of the expectation holds for any collection of random variables, even if they are not independent.

Assume for example we want to compute the expectation of the random variable X representing the sum of two dice. A direct computation reveals that

$$\mathbb{E}[X] = 2\frac{1}{36} + 3\frac{2}{36} + 4\frac{3}{36} + \cdots + 12\frac{1}{36} = 7. \tag{27}$$

An alternative, simpler way to obtain the same result is to use the linearity of the expectation. Let $X = X_1 + X_2$, where X_i, $i = 1, 2$, represents the outcome of die i. Then

$$\mathbb{E}[X_i] = \frac{1}{6}\sum_{j=1}^{6} j = \frac{7}{2}. \tag{28}$$

Hence, $\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 7$.

3.1.2 Jensen's Inequality

In this section we introduce an important inequality that involves expectation of functions of random variables. Here is a motivating example. Suppose we choose the length X of a side of a square uniformly at random from the range of integers $[1, 99]$. What is the expected value of the area? We can write this value

as $\mathbb{E}\left[X^2\right]$. Does it coincide with $\mathbb{E}[X]^2$, i.e., the area of the square with average side? A simple calculation using (214) reveals that $\mathbb{E}\left[X^2\right] = 9950/3 \approx 3316$. Note that this is larger than the area of the square with average side, which is $\mathbb{E}[X]^2 = 2500$. Hence, $\mathbb{E}\left[X^2\right] \geq (\mathbb{E}[X])^2$. This is a consequence of a more general theorem, known as Jensen's inequality, which states that for any convex function $f(\cdot)$, we have that $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Definition 13 (Convexity) *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be convex if, for all $x_1, x_2 \in \mathbb{R}$ and for all $0 \leq \lambda \leq 1$, we have that*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \tag{29}$$

A visual intepretation of the definition above is as follows: if you draw a straight line connecting two points on the graph of the function, the resulting line lies on or above the graph of the function. The following lemma provides a simple way to verify convexity.

Definition 14 (Convex differentiable function) *If f is a twice differentiable function, then f is convex if and only if $f''(x) \geq 0$ for all x.*

We are now ready to state Jensen's inequality.

Theorem 15 (Jensen's inequality) *If f is a convex function, then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]). \tag{30}$$

3.2 The Bernoulli and Binomial Random Variables

Suppose we run an experiment that succeeds with probability p and fails with probability $1-p$. We define a random variable Y such that $Y = 1$ if the experiment succeeds and $Y = 0$ if the experiment fails. The variable Y is usually referred to as *Bernoulli* or indicator random variable. Note that $\mathbb{E}[Y] = \mathbb{P}[Y = 1] = p$.

Let us now assume we perform n independent and identically distributed experiments and we let X be the number of successes. The random variable X then follows the so called *binomial distribution*.

Definition 16 (Binomial Random Variable) *A binomial random variable with parameters n and p, denoted by $B(n,p)$, is defined by the following probability distribution on $j = 0, 1, \ldots, n$:*

$$\mathbb{P}[X = j] = \binom{n}{j} p^j (1-p)^{n-j}. \tag{31}$$

It is a simple exercise to verify that this definition satisfies the requirement of a probability measure. Binomial random variables arise in many contexts. For example in communication systems, X may model the number of packets successfully delivered out of n transmitted packets, when the probability of reliable transmission is $p = 1 - \epsilon$, where ϵ is the packet error probability. In

such a context, we may be interested in the average number of successfully transmitted packet, the so called *goodput*. To determine this quantity, we use the linearity of expectation. Specifically, since X models the number of successes in n trials, where the probability of success is p, we can define a set of n Bernoulli random variables $X_1, \dots, X_n$ where $X_i = 1$ if the ith trial is successful and $X_i = 0$ otherwise, $i = 1, \dots, n$. Then $X = X_1 + \cdots + X_n$ and $\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = np$.

3.3 Conditional Expectation

Definition 17 (Conditional Expectation) *The conditional expectation of the random variable Y given $Z = z$ is defined as*

$$\mathbb{E}[Y \,|\, Z = z] = \sum_y y \, \mathbb{P}[Y = y \,|\, Z = z]. \tag{32}$$

The definition is identical to that of the ordinary expectation, with the only difference being that the probability term is replaced by a conditional-probability term. One can also define the conditional expectation of Y given an event $\mathcal{E}$ as follows:

$$\mathbb{E}[Y \,|\, \mathcal{E}] = \sum_y y \, \mathbb{P}[Y = y \,|\, \mathcal{E}]. \tag{33}$$

Here is an example. We roll two dice independently. Let X_1 and X_2 be the number shown on the first and second die, respectively. Also, let X be their sum. We may be interested in the expected value of X given that $X_1 = 2$, which we can compute as follows

$$\mathbb{E}[X \,|\, X_1 = 2] = \sum_x x \, \mathbb{P}[X = x \,|\, X_1 = 2] = \sum_{x=3}^{8} x \frac{1}{6} = \frac{11}{2}. \tag{34}$$

Similarly, we may want to compute $\mathbb{E}[X_1 \,|\, X = 5]$:

$$\mathbb{E}[X_1 \,|\, X = 5] = \sum_{x=1}^{4} x \, \mathbb{P}[X_1 = x \,|\, X = 5] = \sum_{x=1}^{4} x \frac{\mathbb{P}[X = 5 \cap X_1 = x]}{\mathbb{P}[X = 5]} = \frac{5}{2}. \tag{35}$$

The ordinary expectation can be expressed conveniently as an average sum of conditional expectations as illustrated in the following lemma.

Lemma 18 *For every random variables X and Y,*

$$\mathbb{E}[X] = \sum_y \mathbb{P}[Y = y] \, \mathbb{E}[X \,|\, Y = y] \,. \tag{36}$$

Perhaps confusingly, the term "conditional expectation" is sometimes used also to refer to the random variable introduced in the next definition.

Definition 19 (Conditional Expectation as Random Variable) *By $\mathbb{E}[Y \,|\, Z]$, we denote a random variable $f(Z)$ that takes on the value $\mathbb{E}[Y \,|\, Z = z]$ when $Z = z$.*

It is important to emphasize that $\mathbb{E}[Y \mid Z]$ is not a real number. It is a function of the random variable Z. Hence, it is a random variable itself. Coming back to the previous example of rolling two dices, we may be interested in the random variable $\mathbb{E}[X \mid X_1]$, which can be specified as follows

$$\mathbb{E}[X \mid X_1] = \sum_x x\, \mathbb{P}[X = x \mid X_1] = \sum_{x=X_1+1}^{X_1+6} x \frac{1}{6} = X_1 + \frac{7}{2}. \tag{37}$$

Since $\mathbb{E}[Y \mid Z]$ is a random variable, we may compute its expectation. In the previous example, we found that $\mathbb{E}[X \mid X_1] = X_1 + 7/2$. Hence, $\mathbb{E}[\mathbb{E}[X \mid X_1]] = \mathbb{E}[X_1] + 7/2 = 7$ in agreement with our previous computations. More generally, we have the following result, which is equivalent to Lemma 18.

Theorem 20

$$\mathbb{E}[\mathbb{E}[Y \mid Z]] = \mathbb{E}[Y]\,. \tag{38}$$

3.4 The Geometric Distribution

Suppose we perform a sequence of independent experiments, each one succeeding with probability p, until its first success. The distribution of the number of experiments we conducted follows a *geometric distribution*.

Definition 21 (Geometric Distribution) *A geometric random variable X with parameter p is specified by the following probability distribution on $n = 1, 2, \ldots$:*

$$\mathbb{P}[X = n] = (1-p)^{n-1} p. \tag{39}$$

In words, for the random variable X to take value n, there must have been $n-1$ failures followed by a success. In our example of packet transmission, if packets are received correctly with probability p, the number of packets transmitted after the last correctly received packet, up to and including the next correctly received packet, follows a geometric distribution.

Geometric random variables are said to be *memoryless*, because the probability that one experiences the first success n trials from now is independent of the number of failures one has experienced. Formally, the following result holds:

Lemma 22 (Memoryless Property of Geometric Distribution)

$$\mathbb{P}[X = n + k \mid X > k] = \mathbb{P}[X = n]. \tag{40}$$

We now compute the mean of a geometric random variable. We shall do so using two different approaches. For the first approach, we shall use the following alternative way to compute the expectation of a random variable that takes values in the set of natural numbers.

Lemma 23 (Expectation of Integer-Valued Random Variable) *Let X be a discrete random variable that takes only nonnegative integer values. Then*

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i]. \tag{41}$$

Now, for a geometric random variable with parameter p, we have that

$$\mathbb{P}[X \geq i] = \sum_{n=i}^{\infty} \mathbb{P}[X = i] = \sum_{n=i}^{\infty} (1-p)^{n-1} p = (1-p)^{i-1} \tag{42}$$

where the last step follows from (215). Hence,

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i] = \sum_{i=1}^{\infty} (1-p)^{i-1} = \frac{1}{p}. \tag{43}$$

For the second approach, we shall use the definition of conditional expectation. Recall that X corresponds to the number of experiments conducted until an experiment succeeds and that each experiment succeeds with probability p. Let $Y = 0$ if the first experiment fails and $Y = 1$ otherwise. Then, it follows from Lemma 18 that

$$\mathbb{E}[X] = \mathbb{P}[Y = 0]\, \mathbb{E}[X \,|\, Y = 0] + \mathbb{P}[Y = 1]\, \mathbb{E}[X \,|\, Y = 1] \tag{44}$$

$$= (1-p)\, \mathbb{E}[X \,|\, Y = 0] + p\, \mathbb{E}[X \,|\, Y = 1]\,. \tag{45}$$

Now if $Y = 1$ we have that $X = 1$ and $\mathbb{E}[X \,|\, Y = 1] = 1$. Otherwise, $X > 1$. Let now Z denote the number of remaining experiments until the first success. Then $\mathbb{E}[X \,|\, Y = 0] = \mathbb{E}[Z + 1] = \mathbb{E}[Z] + 1$. But by the memoryless property of geometric random variables, Z is also a geometric random variable with parameter p. Hence, $\mathbb{E}[Z] = \mathbb{E}[X]$. So we have that

$$\mathbb{E}[X] = (1-p)(\mathbb{E}[X] + 1) + p \tag{46}$$

from which it follows that $\mathbb{E}[X] = 1/p$.

3.4.1 Example: the Coupon Collector Problem

Suppose you buy boxes of cereals. Each box contains one of n different coupons. Once you obtain every type of coupon, you qualify for a price. Assume that the coupons in each box is chosen independently and uniformly at random from the n possible different coupons. Assume also that you do not collaborate with other people to collect coupons. How many boxes of cereals must you buy before you obtain at least one of every type of coupon?

Let X be the number of boxes open until every coupon is collected. We will compute $\mathbb{E}[X]$. To do so it is convenient to let X_i be the number of boxes bought when you had exactly $i-1$ coupons. It follows then that $X = \sum_{i=1}^{n} X_i$. Now the probability of obtaining a new coupon when opening a box, given that you have collected already $i-1$ coupons is

$$p_i = 1 - \frac{i-1}{n}. \tag{47}$$

As a consequence, we conclude that X_i is a geometric random variable with parameter p_i. Hence,

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}. \tag{48}$$

Now, by the linearity of the expectation,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i}. \tag{49}$$

The function $H(n) = \sum_{i=1}^{n} 1/i$ is known as *harmonic number*. It can be shown that $\log(n) \leq H(n) \leq \log(n)+1$. Hence, the expected number of boxes to open to collect all coupons scales with n roughly as $n\log(n)$.

3.5 Exercises

Exercise 5 (Rolling Dice) *Suppose that we independently roll two standard six-sided dice. Let X_1 and X_2 be the numbers shown on the first die and the second die, respectively. Let also X be the sum of the two numbers. Compute*

- $\mathbb{E}[X \,|\, X_1$ *is even*$]$
- $\mathbb{E}[X \,|\, X_1 = X_2]$
- $\mathbb{E}[X_1 \,|\, X = 9]$
- $\mathbb{E}[X_1 - X_2 \,|\, X = k]$ *for k in the range* $[2, 12]$.

Exercise 6 (A Large(?) Family) *Alice and Bob decide to have children until they have their first girl or they have $k \geq 1$ children. Assume that each child is a boy or girl independently with probability $1/2$. Assume also that there are no multiple birds. What is the expected number of female children that they have? What is the expected number of male children?*

Exercise 7 (Group Testing) *A blood test is to be performed on n individuals. Each person can be tested separately, but this is expensive. An alternative strategy is to pool and analyze together the samples of k people. If the test is negative, this one test suffices for the group of k people. If the test is positive, then each of the k persons must be tested individually, which results in $k+1$ total tests for the k people. Suppose we create n/k disjoint groups of k people (where k divides n) and use the pooling method. Assume that each person has positive result on the test independently with probability p.*

- *What is the probability that the test for a pooled sample of k people will be positive? What is the expected number of tests necessary?* **`python`***: Confirm your calculations by drawing $n = 10^2$ pseudo random test results and dividing the test results into groups of $k = 10$ with $p \in \{0.02, 0.2, 0.6\}$. Average your results over 10^4 iterations.*
- *Describe how to find the best value of k. Derive an approximation assuming p is close to 0.* **`python`***: Find the optimal integer k for $p \in \{0.02, 0.2, 0.6\}$ by assuming $k < 100$. Compare with your approximation.*

- *For which values of p is pooling better than just testing every individual? Again, derive an approximation assuming p is small and plug in your approximation of the best value of k.* **`python`***: Using your results from the previous parts, test if pooling is better or not for $n = 10^2$, $0.01 \leq p \leq 0.5$ and the best integer value of k (assume $k < 100$). Compare with your approximation.*

Exercise 8 (Gambler's Ruin) *Consider a simplified version of the game of roulette, in which you wager x Swedish kronas on either red or black. The wheel is spun and you receive your original wager plus another x kronas if the ball lands on your color. If the ball does not land on your color, you loose the wager. Assume for simplicity that each color occurs independently with probability $1/2$. The following gambling strategy used to be popular. On the first spin, bet 1 krona. If you loose, bet 2 krona on the next spin. In general, if you have lost on the first $k-1$ spins, bet 2^{k-1} krona on the kth spin. Argue that by following this strategy, you will eventually win 1 krona. Now let X be the random variable denoting your maximum loss before winning. Show that $\mathbb{E}[X]$ is unbounded.*

4 Moments and Deviations

In the previous chapter, we focused on the expectation of a random variable. In this chapter, we shall introduce simple techniques for bounding the *tail distribution* of a random variable. This is the probability that the random variable assumes values that are far from its expectation. In the context of data science, these bounds are extremely useful in determining for example the failure probability of algorithms, the number of samples to be collected for such algorithms to perform adequately, and to estimate their run time.

4.1 Markov's Inequality

We next present a fundamental inequality, which is often too weak to obtain useful results, but it is a stepping stone in developing more sophisticated bounds.

Theorem 24 (Markov's Inequality) *Let X be a random variable that assumes only nonnegative values. Then for all $a > 0$*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{50}$$

Here is an example we shall use throughout this section. Suppose we want to compute the probability of obtaining more than $3n/4$ heads in a sequence of n fair coin flips. Let $X_i = 1$ if the ith coin flip is head and $X_i = 0$ otherwise. Let also $X = \sum_{i=1}^{n} X_i$. We have that $\mathbb{E}[X] = n/2$. Applying Markov's inequality, we obtain

$$\mathbb{P}[X \geq 3n/4] \leq 2/3. \tag{51}$$

As we shall see, this bound is quite loose.

4.2 Variance and Moments of a Random Variable

Markov's bound can be improved upon if more information about the distribution of the random variable is available. Such additional information is typically expressed in terms of moments. The expectation is the first moment of a random variable. More generally, we define the moments of a random variable as follows.

Definition 25 (Moments of a Random Variable) *The kth moment of a random variable X is $\mathbb{E}\left[X^k\right]$.*

A significant improvement on Markov's inequality can be obtained when the second moment $\mathbb{E}\left[X^2\right]$ is also available. Given first and second moments of a random variable, one can compute the *variance* and the *standard deviation*, which provide a measure of how far the random variable deviates from its expectation.

Definition 26 (Variance and Standard Deviation) *The variance of a random variable X is defined as*

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2. \tag{52}$$

The standard deviation is defined as

$$\sigma[X] = \sqrt{\mathbb{V}\mathrm{ar}[X]}. \tag{53}$$

To gain some intuition about these quantities, note that if a random variable X is a constant, i.e., it assumes always the same value, then variance and standard deviations are zero. If a random variable X takes on the value $k\,\mathbb{E}[X]$ with probability $1/k$ and 0 with probability $1-1/k$ (here, $k>1$), then $\mathbb{V}\mathrm{ar}[X] = (k-1)(\mathbb{E}[X])^2$. This confirms the intuition that variance and standard deviation take on large value when the random variable assume values far from its expectation (k large in the second example).

We next study the variance of the sum of random variables. To do so, we first introduce the following quantity.

Definition 27 (Covariance) *The covariance of two random variables X and Y is*

$$\mathbb{C}\mathrm{ov}[X,Y] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]\,. \tag{54}$$

Theorem 28 (Variance of Sum of Random Variables) *For every two random variables X and Y,*

$$\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\,\mathbb{C}\mathrm{ov}[X,Y]\,. \tag{55}$$

This theorem can be easily extended to finite sums of random variables. It turns out that the variance of the sum of two random variables is equal to the sum of the variances when the random variables are independent. To prove this result, we first need the following intermediate fact.

Theorem 29 (Expectation of Product of Independent RVs) *If X and Y are two independent random variables then*

$$\mathbb{E}[X\cdot Y] = \mathbb{E}[X]\cdot\mathbb{E}[Y]\,. \tag{56}$$

Please note that independence is crucial for this result to hold. It is relatively easy to construct examples of dependent random variables for which Theorem 29 does not hold. Suppose for example that Y is a Bernoulli random variable with parameter $1/2$ and that $Z=Y$. Then $\mathbb{E}[Y]=\mathbb{E}[Z]=1/2$ but $\mathbb{E}[Y\cdot Z]=1/2$. So $\mathbb{E}[Y\cdot Z]\neq\mathbb{E}[Y]\cdot\mathbb{E}[Z]$.

Theorem 30 (Independent Random Variables are Uncorrelated) *If X and Y are independent random variables, then*

$$\mathbb{C}\mathrm{ov}[X,Y] = 0 \tag{57}$$

and

$$\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] \tag{58}$$

This result can be easily extended by induction to the sum of any finite number of random variables.

4.2.1 Example: Variance of a Binomial Random Variable

Recall that a binomial random variable X with parameters n and p can be represented as the sum of n independent Bernoulli random variables with parameter p. Since a Bernoulli random variable Y with parameter p has variance $\mathbb{V}\mathrm{ar}[Y] = p(1-p)$, we conclude that

$$\mathbb{V}\mathrm{ar}[X] = np(1-p). \tag{59}$$

4.3 Chebyshev's Inequality

Using the expectation and the variance of a random variable, one can derive the following tail bound.

Theorem 31 (Chebyshev's Inequality) *For every $a > 0$*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \frac{\mathbb{V}\mathrm{ar}[X]}{a^2} \tag{60}$$

Chebyshev's inequality is sometimes expressed also in the following useful variants, in which the variation from the expectation is expressed as a constant factor of the standard deviation of the random variable.

Corollary 32 *For every $t > 1$,*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t \cdot \sigma[X]] \leq \frac{1}{t^2} \tag{61}$$

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t \cdot \mathbb{E}[X]] \leq \frac{\mathbb{V}\mathrm{ar}[X]}{t^2(\mathbb{E}[X])^2} \tag{62}$$

Let us now go back to our coin-flipping example and this time use Chebyshev's inequality to bound the probability of obtaining more than $3n/4$ heads in a sequence of n fair coin flips. Let $X_i = 1$ if the ith coin flip is head and $X_i = 0$ otherwise. Let also $X = \sum_{i=1}^{n} X_i$. Note that $\mathbb{E}[X_i]$ and $\mathbb{V}\mathrm{ar}[X_i] = 1/4$, $i = 1, \ldots, n$, and, because of independence $\mathbb{V}\mathrm{ar}[X] = n/4$. Applying Chebyshev's inequality, we conclude that

$$\mathbb{P}\left[X \geq \frac{3n}{4}\right] \leq \mathbb{P}\left[\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right] \leq \frac{4}{n}. \tag{63}$$

Note how Chebyshev's inequality is significantly better than Markov's inequality for large n. Indeed, Markov's inequality gave as a probability of $2/3$, independent of n. With Chebyshev's inequality, we obtained a probability that vanishes with n. We shall see in the next chapter that this estimate can be further improved.

4.4 Example: Coupon Collector Problem

We now apply Markov's and Chebyshev's inequalities to the coupon collector problem Recall that the number of boxes to open to collect all coupons has expectation $\mathbb{E}[X] = nH(n)$ where $H(n) = \sum_{i=1}^{n} 1/i$. It then follows from Markov's inequality that

$$\mathbb{P}[X \geq 2nH(n)] \leq \frac{1}{2}. \tag{64}$$

To use Chebyshev's inequality, we need to find the variance of X. Recall that $X = \sum_{i=1}^{n} X_i$ where the X_i are geometric random variables with parameter $(n-i+1)/n$. Furthermore, the X_i are independent because the number of boxes to open to collect the ith coupon does not depend on how long it took to collect the previous $i-1$ coupons. Hence,

$$\mathbb{V}\mathrm{ar}[X] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i]\,. \tag{65}$$

So we need to find the variance of a geometric random variable. Let Y be a geometric random variable with parameter p. We have already seen that $\mathbb{E}[Y] = 1/p$. To compute $\mathbb{E}\left[Y^2\right]$ we use the conditional expectation trick we have already used to compute $\mathbb{E}[Y]$. Specifically, let $X = 1$ if the first experiment succeeds and $X = 0$ otherwise. Then,

$$\mathbb{E}\left[Y^2\right] = \mathbb{P}[X=0]\,\mathbb{E}\left[Y^2 \,|\, X=0\right] + \mathbb{P}[X=1]\,\mathbb{E}\left[Y^2 \,|\, X=1\right]. \tag{66}$$

If $X = 1$, then $Y = 1$ and $\mathbb{E}\left[Y^2 \,|\, X=1\right] = 1$. If $X = 0$ then $Y > 1$. In this case, let Z be the number of remaining experiments until the first success. Then

$$\mathbb{E}\left[Y^2\right] = (1-p)\,\mathbb{E}\left[(Z+1)^2\right] + p = (1-p)\,\mathbb{E}\left[Z^2\right] + 2(1-p)\,\mathbb{E}[Z] + 1. \tag{67}$$

By the memoryless property of geometric random variables, Z is also a geometric random variable with parameter p. Hence, $\mathbb{E}[Z] = 1/p$. Furthermore, $\mathbb{E}\left[Z^2\right] = \mathbb{E}\left[Y^2\right]$. Hence,

$$\mathbb{E}\left[Y^2\right] = \frac{2-p}{p^2}. \tag{68}$$

To summarize,

$$\mathbb{V}\mathrm{ar}[Y] = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}. \tag{69}$$

We state this useful result in the following lemma.

Lemma 33 (Variance of a Geometric RV) *The variance of a geometric random variable with parameter p is $(1-p)/p^2$.*

Let us now go back to the coupon collector problem. To simplify the final expression, we shall use the upper bound $\mathbb{V}\mathrm{ar}\left[Y^2\right] \leq 1/p^2$ instead of the exact result in Lemma 33. Then we have

$$\mathbb{V}\mathrm{ar}[Y] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i] \leq \sum_{i=1}^{n} \left(\frac{n}{n-i+1}\right)^2 = n^2 \sum_{i=1}^{n} \frac{1}{i^2} \leq \frac{\pi^2 n^2}{6} \tag{70}$$

where in the last step we used (218). We can now apply Chebyshev's inequality and conclude that

$$\mathbb{P}[X \geq 2nH(n)] \leq \mathbb{P}[|X - nH(n)| \geq nH(n)] \leq \frac{n^2\pi^2/6}{(nH(n))^2} = \frac{\pi^2}{6H^2(n)}. \tag{71}$$

So this tail bound decays roughly as $1/\ln^2(n)$, which is much better than what Markov's inequality gave us. Nevertheless, this bound is still weak. Remember that $H(n)$ grows roughly as $\log n$. Fix a $c > 0$ and consider the probability of not getting a coupon after $n(\ln n + c)$ steps. This probability is

$$\left(1 - \frac{1}{n}\right)^{n(\ln n + c)} < e^{-(\ln n + c)} = \frac{1}{e^c n}. \tag{72}$$

By a union bound, the probability that some coupon has not been collected after $n(\ln n + c)$ steps is at most $1/e^c$. In particular, if we set $c = \ln n$, we see that the probability that $X > 2n \ln n$ is at most $1/n$. This bound is significantly better that the one given by Chebyshev's inequality.

4.5 Exercises

Exercise 9 (Rolling a die many times) *Suppose we roll a standard fair die* 100 *times. Let* X *be the sum of the numbers that appear over the* 100 *rolls. Use Chebyshev's inequality to bound* $\mathbb{P}[|X - 350| \geq 50k]$

Exercise 10 (Stock Market) *A simple model for the stock market suggests that each day a stock with price* q *will increase by a factor* $r > 1$ *to* qr *with probability* p *and will fall to* q/r *with probability* $1 - p$*. Assume that we start with a stock with price* 1*. Find a formula for the expected value and the variance of the price of the stock after* d *days.*

python: *Let* $r = 1.1$*,* $p = 0.5$*, and* $q = 1$*. Simulate* 200 *stocks over* 50 *days. Estimate the expected value and the variance of the price of the stock for* $0 \leq d \leq 50$ *and compare with your theoretical results. Plot your results.*

Exercise 11 (Random Bits) *Suppose that we flip a fair coin* n *times to obtain* n *random bits. Consider all* $m = \binom{n}{2}$ *pairs of these bits and order them in some way. Let* Y_i *the ex-or of the* i*th pair of bits and let* $Y = \sum_{i=1}^{m} Y_i$ *be the number of* Y_i *that equal* 1*.*

- *Show that each* Y_i *is a Bernoulli random variable with parameter* $1/2$*.*

- *Argue that the Y_i are not mutually independent.*
- *Show that the Y_i satisfy the property that $\mathbb{E}[Y_i Y_j] = \mathbb{E}[Y_i]\,\mathbb{E}[Y_j]$.*
- *Find $\mathbb{V}\mathrm{ar}[Y]$*
- *Use Chebyshev's inequality to establish a bound on $\mathbb{P}[|Y - \mathbb{E}[Y]| \geq n]$*

5 Chernoff and Hoeffding Bounds

In this chapter, we introduce tail bounds commonly called Chernoff and Hoeffding bounds. These bounds are much more powerful than the one we have seen in the previous chapter, since they give exponentially decreasing bounds. The bounds are derived applying Markov's inequality to the so called *moment generating function* of a random variable.

5.1 Moment Generating Functions

Definition 34 (Moment Generating Functions) *The moment generating function of a random variable X is defined as*

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]. \tag{73}$$

Throughout these notes, we will assume that the moment generating function exists in a neighborhood of zero. This is required for all properties we shall state next to hold. The first important property is that all moments of a random variable can be easily obtained from the moment generating function.

Theorem 35 (Moments from Moment Generating Function) *Let X be a random variable with moment generating function $M_X(t)$. Let $M_X^{(n)}(t)$ denote the derivative of order n of $M_X(t)$. Then*

$$\mathbb{E}[X^n] = M_X^{(n)}(0). \tag{74}$$

Consider for example a geometric random variable with parameter p. Then for $t < -\ln(1-p)$,

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] && (75)\\ &= \sum_{k=1}^{\infty}(1-p)^{k-1}pe^{tk} && (76)\\ &= \frac{p}{1-p}\sum_{k=1}^{\infty}(1-p)^k e^{tk} && (77)\\ &= \frac{p}{1-p}\left(\frac{1}{1-(1-p)e^t}-1\right) && (78)\\ &= \frac{pe^t}{1-(1-p)e^t} && (79) \end{aligned}$$

You can verify that the expression for the mean and the second moment we have obtained in the previous chapters can be reobtained by evaluating the first and the second derivative of $M_X(t)$ at $t = 0$.

Another useful property of the moment generating function is that it uniquely define the distribution of a random variable.

Theorem 36 (Uniqueness of Moment Generating Function) *Let X and Y be two random variables. If $M_X(t) = M_Y(t)$ for all $t \in (-\delta, \delta)$ for some $\delta > 0$, then X and Y have the same distribution.*

An additional useful property is that the moment generating function of the sum of independent random variables can be easily obtained from the moment generating function of each random variable.

Theorem 37 (Moment Generating Function of Sum of Independent RVs) *If X and Y are independent random variables then*

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t). \tag{80}$$

Recall now that the probability distribution of a sum of random variables is the convolution of the two probability distributions. So the moment-generating function transforms convolutions into products, similar to the Fourier and the Laplace transform you have seen in signal-theory courses. This is one of the reasons why it is convenient to work with moment-generating functions.

5.2 Deriving and Applying Chernoff Bounds

Let X be a random variable with moment generating function $M_X(t)$. From Markov's inequality, we can derive the following useful inequality: for all $t > 0$

$$\mathbb{P}[X \geq a] = \mathbb{P}[e^{tX} \geq e^{ta}] \leq \frac{M_X(t)}{e^{ta}}. \tag{81}$$

In particular,

$$\mathbb{P}[X \geq a] \leq \min_{t>0} \frac{M_X(t)}{e^{ta}}. \tag{82}$$

Similarly, for $t < 0$,

$$\mathbb{P}[X \leq a] = \mathbb{P}\big[e^{tX} \geq e^{ta}\big] \leq \frac{M_X(t)}{e^{ta}} \tag{83}$$

$$\mathbb{P}[X \leq a] \leq \min_{t<0} \frac{M_X(t)}{e^{ta}} \tag{84}$$

The optimal value of t depends on the distribution. Often, rather than choosing this optimal value, one selects a value of t for which the bound has a convenient form. Bounds derived using this approach are collectively referred to as *Chernoff bounds*.

5.2.1 Chernoff Bounds for the Sum of Bernoulli random variables

One of the most commonly used application of the Chernoff bound is to bound the tail distribution of a sum of independent but not necessarily identically distributed Bernoulli random variables. Specifically, let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli random variables with $\mathbb{P}[X_i = 1] = p_i$, $i = 1, \ldots, n$.

Let $X = \sum_{i=1}^n X_i$ and let also $\mu = \sum_{i=1}^n p_i = \mathbb{E}[X]$. For a given $\delta > 0$, we are interested in bounding $\mathbb{P}[X \geq (1+\delta)\mu]$ and $\mathbb{P}[X \leq (1-\delta)\mu]$. In words, we are interested in bounding the probability that X deviates from its expectation μ by $\delta\mu$ or more. To compute a Chernoff bound, we start by computing the moment-generating function of each X_i.

$$M_{X_i}(t) = \mathbb{E}\left[e^{tX_i}\right] = p_i e^t + (1-p_i) = 1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)}. \tag{85}$$

Here, in the last step, we used that $1 + y \leq e^y$ for all y. The last bound is useful because X is a sum of independent random variables. Indeed, the moment-generating function of X can be compactly upper-bounded as

$$M_X(t) = \prod_{i=1}^n M_{X_i}(t) \leq \prod_{i=1}^n e^{p_i(e^t-1)} = e^{\sum_{i=1}^n p_i(e^t-1)} = e^{\mu(e^t-1)} \tag{86}$$

We are now ready to state a Chernoff bound for this setup. We start with a bound for deviations above the mean.

Theorem 38 (Chernoff bound upper deviations) *Let $X_i, \dots, X_n$ be independent Bernoulli random variables such that $\mathbb{P}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. Then the following Chernoff bounds hold:*

- *For every $\delta > 0$,*

$$\mathbb{P}[X \geq (1+\delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu. \tag{87}$$

- *For every $0 < \delta \leq 1$*

$$\mathbb{P}[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/3}. \tag{88}$$

We obtain similar results when bounding the deviations below the mean.

Theorem 39 (Chernoff bound lower deviations) *Let $X_i, \dots, X_n$ be independent Bernoulli random variables such that $\mathbb{P}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. Then for every $0 < \delta < 1$*

$$\mathbb{P}[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\mu. \tag{89}$$

and

$$\mathbb{P}[X \leq (1+\delta)\mu] \leq e^{-\mu\delta^2/2}. \tag{90}$$

The following two-sided form of Chernoff bound is often used. It follows directly from (88) and (90).

Corollary 40 (Chernoff, two-sided) *Let $X_1, \dots, X_n$ be independent Bernoulli random variables with $\mathbb{P}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. For $0 < \delta < 1$,*

$$\mathbb{P}[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu\delta^2/3}. \tag{91}$$

5.2.2 Example: Coin Flips

Let us go back to our running example involving coin flips. Recall that we are interested in the probability of observing more than $3n/4$ heads in a sequence of n independent fair coin flips. It follows from Theorem 40 that

$$\mathbb{P}\left[X \geq \frac{3n}{4}\right] \leq \mathbb{P}\left[\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right] \leq 2e^{-n/24}. \tag{92}$$

Note that this bound decays exponentially with n. In contrast, the probability we obtained using Chebyshev's bound decays with n only as $1/n$. We can determine which kind of deviations we need to allow, to obtain a probability that decays as $1/n$ using Chernoff bound. Again, an application of Theorem 40 reveals that

$$\mathbb{P}\left[\left|X - \frac{n}{2}\right| \geq \frac{1}{2}\sqrt{6n\ln(n/2)}\right] \leq \frac{4}{n}. \tag{93}$$

So the deviations from the mean compatible with a probability that decays as $1/n$ are of order $\sqrt{n\ln n}$.

5.2.3 Application: Estimating a Parameter

Suppose we are interested in determining the probability that a particular gene mutation occurs in the population. To do so we can collect DNA samples from each individual and determine by a test whether it carries the mutation. Since such tests may be expensive, we are interested in determining the minimum number of tests required to determine the probability we are interested in with a given accuracy. To be concrete, let p be the unknown probability value we are trying to estimate. Assume we collect n samples and that our test reveals that $X = \tilde{P}n$ of these samples have the mutation. Our intuition is that if n is chosen sufficiently large, then $\tilde{P}$ should be close to p. We express this intuition through the concept of confidence interval.

Definition 41 (Confidence Interval) *A $1-\gamma$ confidence interval for the parameter p is an interval $[\tilde{P}-\delta, \tilde{P}+\delta]$ such that*

$$\mathbb{P}\Big[p \in [\tilde{P}-\delta, \tilde{P}+\delta]\Big] \geq 1-\gamma. \tag{94}$$

Note that instead of specifying a single value for the unknown parameter p, we provide an interval that is likely to contain the parameter. Obviously, we would like both the interval size 2δ and the error probability γ to be small. Next, we use Chernoff bound to derive a tradeoff between these two parameter and the number of samples n. Note that $X = n\tilde{P}$ has a binomial distribution with parameters n and p. Hence, $\mathbb{E}[X] = np$. If $p \notin [\tilde{P}-\delta, \tilde{P}+\delta]$, then one of the two following events must have occurred:

- If $p < \tilde{P} - \delta$, then $X = n\tilde{P} > n(p+\delta) = \mathbb{E}[X]\,(1+\delta/p)$.
- If $p > \tilde{P} + \delta$, then $X = n\tilde{P} < n(p-\delta) = \mathbb{E}[X]\,(1-\delta/p)$.

We can apply Chernoff bound in (88) and (90) to obtain

$$\mathbb{P}\Big[p \notin [\tilde{P}-\delta, \tilde{P}+\delta]\Big] \leq \mathbb{P}\left[X < np\left(1-\frac{\delta}{p}\right)\right] + \mathbb{P}\left[X > np\left(1+\frac{\delta}{p}\right)\right] \tag{95}$$

$$\leq e^{-np(\delta/p)^2/2} + e^{-np(\delta/p)^2/3} \tag{96}$$

$$= e^{-n\delta^2/(2p)} + e^{-n\delta^2/(3p)}. \tag{97}$$

The bound just obtained is not useful yet, because p is actually unknown. A simple solution is to use that $p \leq 1$, by which we obtain

$$\mathbb{P}\Big[p \notin [\tilde{P}-\delta, \tilde{P}+\delta]\Big] \leq e^{-n\delta^2/2} + e^{-n\delta^2/3}. \tag{98}$$

To obtain the desired tradeoff, it suffices to set $\gamma = e^{-n\delta^2/2} + e^{-n\delta^2/3}$. By doing so, we can determine the number of samples n required for a confidence interval of size 2δ to hold with probability at least $1-\gamma$.

5.3 Tighter Bounds for Some Special Cases

We first consider the case when we have a sum of independent random variables where each random variable takes on the value 1 or -1 with equal probability. The crucial observation is that for such random variables, we can obtain a more accurate bound on the moment generating function than the one we used in (85).

Theorem 42 (Chernoff bound for symmetric uniform binary RVs.) *Let $X_1, \dots, X_n$ be independent random variables with*

$$\mathbb{P}[X_i = 1] = \mathbb{P}[X_i = -1] = \frac{1}{2}, \quad i = 1, \dots, n. \tag{99}$$

Let $X = \sum_{i=1}^n X_i$. Then for every $a > 0$,

$$\mathbb{P}[|X| \geq a] \leq 2e^{-a^2/(2n)}. \tag{100}$$

Applying the transformation $Y_i = (X_i + 1)/2$ we can apply this result to i.i.d. Bernoulli random variables, obtaining the following result, which improves upon Theorem 38 for the special case of Bernoulli random variables with parameter $1/2$.

Corollary 43 *Let $Y_1, \dots, Y_n$ be independent random variables with*

$$\mathbb{P}[Y_i = 1] = \mathbb{P}[Y_i = 0] = \frac{1}{2}. \tag{101}$$

Let $Y = \sum_{i=1}^n Y_i$ and $\mu = \mathbb{E}[Y] = n/2$.

- *For every $a > 0$,*

$$\mathbb{P}[Y \geq \mu + a] \leq e^{-2a^2/n} \tag{102}$$

- *For every* $\delta > 0$,

$$\mathbb{P}[Y \geq (1+\delta)\mu] \leq e^{-\delta^2 \mu}. \tag{103}$$

Note the improvement in (106) with respect to (88): the constant in the exponent is 1 and not 1/3. We can establish a similar result for the lower tail.

Corollary 44 *Let* $Y_1, \dots, Y_n$ *be independent random variables with*

$$\mathbb{P}[Y_i = 1] = \mathbb{P}[Y_i = 0] = \frac{1}{2}. \tag{104}$$

Let $Y = \sum_{i=1}^n Y_i$ *and* $\mu = \mathbb{E}[Y] = n/2$.

- *For every* $0 < a < \mu$,

$$\mathbb{P}[Y \geq \mu - a] \leq e^{-2a^2/n} \tag{105}$$

- *For every* $0 < \delta < 1$,

$$\mathbb{P}[Y \geq (1-\delta)\mu] \leq e^{-\delta^2 \mu}. \tag{106}$$

5.4 Application: Set Balancing

Set balancing refers to the problem of dividing a set into two subsets of the same characteristics. It arises, for example, in the design of experiments. Specifically, assume you have a group of m subjects. We associate to each subject a binary vector of dimension n, which describes which of some selected n features are possessed by the individual. Features may be for example being young, tall, etc. . . , and the corresponding entry of the feature vector is equal to 1 if the individual possesses the feature, and equal to 0 otherwise.

The task is to divide the subjects into two sets, a *treatment group* $\mathcal{T}$ and a *control group* $\mathcal{C}$ such that for each feature, the number of individuals possessing that feature in $\mathcal{T}$ is roughly equal to the number of individuals possessing the same feature in $\mathcal{C}$. We collect all feature vectors (which are column vectors) in a binary matrix $\mathbf{A}$ of dimension $n \times m$. The partition is described by an m dimensional vector $\mathbf{b} = [b_1 \dots b_m]^{\mathrm{T}}$ with entries in $\{-1, 1\}$. We use the convention that $b_i = 1$ means that the ith individual is assigned to $\mathcal{T}$, whereas $b_i = -1$ means that it is assigned to $\mathcal{C}$.

We can describe the imbalance in the partition by the vector $\mathbf{c} = \mathbf{A}\mathbf{b}$. Let $\mathbf{c} = [c_1 \dots c_n]^{\mathrm{T}}$. If $c_i > 0$ this means that there are more subjects with feature i in $\mathcal{T}$. Conversely, $c_i < 0$ this means that there are more subjects with feature i in $\mathcal{C}$.

It follows that we are interested in designing a partition that minimizes

$$\max_{i \in [1,\dots,n]} |c_i| \,. \tag{107}$$

It turns out that, when the number of features is large, a simple algorithm that assigns each individual to one of the two sets uniformly at random is close

to optimal. Specifically, assume that we randomly and independently choose the entries of $\mathbf{b}$ using the probability distribution $\mathbb{P}[b_i = 1] = \mathbb{P}[b_i = -1] = 1/2$, $i = 1, \dots, n$. Note that with this choice, we ignore completely the entries of the matrix $\mathbf{A}$.

Using the Chernoff bound in Theorem (42), we can show that

$$\mathbb{P}\left[\max_{i\in[1,\dots,n]} |c_i| > \sqrt{4m \ln n}\right] \leq \frac{2}{n}. \tag{108}$$

So this tells us that, roughly speaking, when the number of features n is large, the imbalance does not exceed $\sqrt{4m \ln n}$ with high probability. This bound is fairly tight. Indeed, for the case $m = n$, one can show that there exist matrices $\mathbf{A}$ for which the optimal imbalance (i.e., the one obtained by selecting $\mathbf{b}$ optimally) grows with n as $\sqrt{n}$. Our bound predicts that it would grow (with high probability) no faster than $\sqrt{n \ln n}$.

5.5 Hoeffding Bounds

Hoeffding bound extends Chernoff bound technique to general random variables with bounded range. Further extensions to random variables with unbounded range are known, but will not be reviewed in this course. They will be covered in the follow-up course *EEN100: statistics and machine learning in high dimensions*.

Theorem 45 (Hoeffding Bound) *Let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu$, and $\mathbb{P}[\alpha \leq X_i \leq \beta] = 1$. Then*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq a\right] \leq 2e^{-2na^2/(\beta-\alpha)^2} \tag{109}$$

Note that if we set $\alpha = -1$ and $\beta = 1$, we recover the result in Theorem 42. This result relies on the following lemma, which establishes a bound on the moment generating function of a zero-mean random variable with bounded range.

Lemma 46 (Hoeffding's Lemma) *Let X be a random variable satisfying*

$$\mathbb{P}[\alpha \leq X \leq \beta] = 1 \text{ and } \mathbb{E}[X] = 0. \tag{110}$$

Then for every $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2(\beta-\alpha)^2/8}. \tag{111}$$

5.6 Exercises

Exercise 12 (Checkers) *Alice and Bob play checkers often. Alice is a better player, so the probability that she wins a game is 0.6, independent of every other games. They decide to play a tournament of n games, where n is odd. Bound the probability that Alice loses the tournament using a Chernoff bound.*

Exercise 13 (Tail bounds) *We have a standard six-sided die. Let X be the number of times that a 6 occurs over n throws of the die. Let p the probability of the event $X \geq n/4$. Compare the best upper bounds on p that you can obtain using Markov's inequality Chebyshev's inequality, and Chernoff bounds.*

***python**: Simulate the dice throws for $10^2 \leq n \leq 150$ over 1000 instances and estimate p for the different values of n. Calculate the actual value of p. Compare the simulations and the bounds with the actual value of p. Plot your results. Which gives a better approximation for p for different values of n?*

Exercise 14 (An unexpected outcome) *A casino is testing a new class of simple slot machines. In each game, the player puts in 1 dollar and the slot machine is supposed to return either 3 dollars to the player with probability $4/25$ or 100 dollars with probability $1/200$ or nothing with all remaining probability. Each game is independent of the other games. The casino has been surprised to find in testing that the machines have lost 10000 dollars over the first million game. Derive a Chernoff bound for the probability of this event. You may want to write a **python** script to solve this exercise.*

***python**: Write a simulation estimating the probability of this event. Based on your Chernoff bound, how many instances of a million games do you think you need to generate?*

6 Balls and Bins

In this chapter, we analyze one of the most basic random processes: m balls are thrown randomly in n bins. Specifically, each ball lands in a bin chosen independently and uniformly at random. We use the techniques developed in the previous chapters to analyze this random process, which turns out to appear in many applications that are relevant for data science.

6.1 Example: the Birthday Paradox

Let us go back to an example we have seen in one of the exercises. You are sitting in a lecture room and you notice that there are 30 people in the room. How likely is it that two people in the room share the same birthday? To answer this question, let us introduce some modeling assumptions to simplify the problem. Let us assume that the birthday of a person is a random variable that is independent and uniformly distributed over the 365 days of the year. One way to answer this question is to count the configurations where people do not share a birthday. There are a total of 365^{30} possible ways of assigning birthdays to 30 people. Out of these configurations, there are a total of $\binom{365}{30}30! = \prod_{j=0}^{29} 365-j$ configurations where no birthday is repeated. So the probability that no two people share the same birthday is

$$\frac{\binom{365}{30}30!}{365^{30}} = 0.2937. \tag{112}$$

So there is about a 70% chance that two people in the room have the same birthday.

In general, if there are m people and n possible birthdays, the probability that all people have all different birthdays is

$$\frac{\binom{n}{m}m!}{n^m} = \prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right). \tag{113}$$

Assume now that m is small compared to n. Then for $j = 1, \dots, m-1$, we can approximate $(1 - j/n)$ by $e^{-j/n}$. Doing so we obtain

$$\begin{aligned}\prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right) &\approx \prod_{j=1}^{m-1} e^{-j/n} && (114)\\ &= e^{-\sum_{j=1}^{m-1} j/n} && (115)\\ &= e^{-m(m-1)/(2n)} && (116)\\ &\approx e^{-m^2/(2n)}. && (117)\end{aligned}$$

Using this approximation, we conclude that the value of m for which the probability that all people have different birthdays is p satisfies

$$\frac{m^2}{2n} = \ln\frac{1}{p} \tag{118}$$

or, equivalently,

$$m = \sqrt{2n \ln(1/p)}. \tag{119}$$

If we now set $n = 365$, and $p = 0.2937$ we find $m = 29.91$, which shows that this approximation is very accurate for this set of parameters.

6.2 Balls into Bins

The birthday paradox is an example of a more general mathematical framework that is often formulated in terms of balls and bins. Specifically, we have m balls that are thrown into n bins. The location of each ball is chosen independently and uniformly at random from the n bins. We are in general interested in the distribution of the balls. The birthday paradox is about determining whether there is a bin with two balls. But we may be interested in other questions. For example, how many balls are in the fullest bin? How many of the bins are empty? We will investigate some of these questions in the next sections.

6.3 The Poisson Distribution

Assume we have m balls and n bins. We will now compute the probability that a given bin is empty, as well as the expected number of empty bins. For the first bin to be empty, it must have been missed by all m balls. This occurs with probability

$$\left(1 - \frac{1}{n}\right)^m \approx e^{-m/n}. \tag{120}$$

Of course, the same holds for any of the other bins. Let $X_i = 1$ if bin i is empty and zero otherwise. Then $\mathbb{E}[X_i] = (1 - 1/n)^m$. Let now X denote the number of empty bins. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\left(1 - \frac{1}{n}\right)^m \approx ne^{-m/n}. \tag{121}$$

We can generalize the argument just provided to determine the expected fraction of bins with r balls. Indeed, the probability that a given bin has r balls is

$$\binom{m}{r}\left(\frac{1}{n}\right)^r\left(1 - \frac{1}{n}\right)^{m-r} = \frac{1}{r!}\frac{m(m-1)\cdots(m-r+1)}{n^r}\left(1 - \frac{1}{n}\right)^{m-r} \tag{122}$$

Assume that m is large compared to r and that n is large compared to m. Then

$$\frac{m(m-1)\cdots(m-r+1)}{n^r} \approx \left(\frac{m}{n}\right)^r \tag{123}$$

and

$$\left(1 - \frac{1}{n}\right)^{m-r} \approx e^{-m/n}. \tag{124}$$

Hence, the probability p_r that a given bin has r balls is approximately

$$p_r \approx \frac{e^{-m/n}(m/n)^r}{r!} \tag{125}$$

and the expected number of bins with r balls is approximately np_r.

This calculation leads us to consider the following distribution.

Definition 47 (Discrete Poisson distribution) *A discrete Poisson random variable X with parameter μ has the following distribution on $j = 0, 1, 2, \ldots$:*

$$\mathbb{P}[X = j] = \frac{e^{-\mu}\mu^j}{j!}. \tag{126}$$

It follows from (219) that (126) satisfies the properties of a probability distribution.

It is not difficult to verify that the expectation of this random variable is μ. In the context of throwing m balls into n bins, the distribution of the number of balls in a bin is approximately Poisson with $\mu = m/n$, which is the average number of balls per bin, as one may expect.

In the next lemma, we provide an important property of the Poisson distribution.

Lemma 48 (Sum of Poisson RVs) *The sum of a finite number of independent Poisson random variables is a Poisson random variable whose parameter is the sum of the parameters.*

This result can be easily established by noting that the moment-generating function of a Poisson random variable with parameter μ is

$$M_X(t) = e^{\mu(e^t - 1)}. \tag{127}$$

Using the moment-generating function, one can also verify that $\mathbb{E}\left[X^2\right] = \mu(\mu + 1)$, which implies that $\mathbb{V}\mathrm{ar}[X] = \mu$.

6.3.1 Limit of the Binomial Distribution

The Poisson distribution is the limit distribution of the binomial distribution with parameters n and p when n is large and p is small. More specifically, we have the following limiting result

Theorem 49 (Poisson as Limit Distribution) *Let X_n be a binomial random variable with parameters n and p, where p is a function of n and $\lim_{n\to\infty} np = \mu$ is a constant independent of n. Then for every fixed k,*

$$\lim_{n\to\infty} \mathbb{P}[X_n = k] = \frac{e^{-\mu}\mu^k}{k!}. \tag{128}$$

This theorem applies to the balls and bins scenario and validates the accuracy of our approximations. Consider indeed the situation where we have m balls and n bins, where n is a function of m. Assume that $\lim_{m\to\infty} m/n = \mu$. Then X_m is a binomial random variable with parameters m and $1/n$. Theorem 49 states that

$$\lim_{m\to\infty} \mathbb{P}[X_m = r] = \frac{e^{-\mu}\mu^k}{r!} \tag{129}$$

which matches our approximation in (125).

Binomial distribution of this kind arise frequently, and are often modeled by Poisson distributions. Consider for example the number of spelling or grammatical mistakes in these lecture notes. One model for such mistakes is that each word is likely to have an error with some very small probability p. The number of errors is then a binomial random variable with large n and small p, which can be treated as a Poisson random variable.

6.4 The Poisson Approximation

One difficulty in analyzing the balls and bins problem is that it contains dependencies. For example, if we know that the first bin is empty, then it is less likely that the second bin will be empty, because the m balls need now to be distributed across the remaining $n-1$ bins. And obviously, if we know the number of balls in the first $n-1$ bins, then we know as well the number of balls in the last bin. So the random variables that arise when we want to compute the loads of the various bins are not independent, which prevents us from using the Chernoff bound techniques we have studied in the previous chapter. In this section, we ask ourselves if there is a way to circumvent these dependencies. It turns out that the answer is positive.

We have already shown that when we throw m balls in n bins, the distribution of the number of balls in a given bin is well-approximated by a Poisson random variable with parameter m/n. In this section, we show that the joint distribution of the number of balls in all the bins is, under certain assumptions, well-approximated by assuming that the load in each bin is an *independent* Poisson random variable with mean m/n. In other words, we can treat bin loads as independent random variables. It turns out that this is a good approximation when we are concerned with sufficiently rare events.

To introduce this result, we need the following definitions. Assume that m balls are thrown in n bins, and let X_i be the random variable describing the number of balls in the ith bin, where $i = 1, \dots, n$. Let $Y_1, \dots, Y_n$ be independent Poisson random variables with mean m/n. We next derive a useful relation between these two sets of random variables.

Note that the difference between throwing m balls randomly, and assigning to each bin a number of balls distributed as a Poisson random variable is that, in the first case, we know that we have m balls in total, whereas in the second case, we know only that m is the expected number of balls in all the bins. So in order to establish a relation between the two sets of random variables, we need to condition in the second case on the total number of balls being m.

Theorem 50 *The conditional joint distribution of $Y_1, \dots, Y_n$ given $\sum_{i=1}^n Y_i = m$ coincides with the joint distribution of $X_1, \dots, X_n$.*

Equipped with this relationship between the two distributions, we can prove the following result.

Theorem 51 *Let $f(x_1, \dots, x_n)$ be an arbitrary nonnegative function. Then*

$$\mathbb{E}[f(X_1, \dots, X_n)] \leq e\sqrt{m}\,\mathbb{E}[f(Y_1, \dots, Y_n)]\,. \tag{130}$$

In particular, we can take as f in Theorem 51 the indicator function that some event occurs. Then the theorem gives us a bound on the probability of this event. If the event is sufficiently rare, so that the multiplicative term $e\sqrt{m}$ is not significant, then the theorem gives us an easy-to-compute bound on its probability. In particular, every event that occurs with sufficiently low probability according to the Poisson approximation, occurs with low probability also in the exact model. The bound we reported in Theorem 51 is simple but loose: it can be improved in many practially relevant cases, although this is outside the scope of this course.

6.5 Exercises

Exercise 15 (Approximations Involving e) *For what value of n is $(1 + 1/n)^n$ within 1% of e? And within 0.0001% of e? Similarly, for what value of n is $(1 - 1/n)^n$ within 1% of e? And within 0.0001% of e?*

Exercise 16 (Errors in these notes) *Let X be a Poisson random variable with mean μ representing the number of errors on a page of these notes. Each error is independently a grammatical error with probability p or a spelling error with probability $1-p$. If Y and Z are random variables representing the number of grammatical and spelling errors on a page of this book, prove that Y and Z are Poisson random variables with means μp and $\mu(1-p)$, respectively. Also prove that Y and Z are independent.*

***python**: Generate X, Y, Z according to the description above with $\mu = 10$ and $p = 0.3$. Generate 10^4 instances and verify that what you've proven holds.*

Exercise 17 (Balls in Bins) *Consider the probability that every bin receives exactly one ball when n balls are thrown randomly into n bins.*

- *Given an upper bound on this probability using the Poisson approximation*
- *Determine the exact probability of this event*
- *Show that these two probabilities differ by a multiplicative factor that equals the probability that a Poisson random variable with parameter n takes on the value n.*

7 Discrete-Time Markov Chains

So far, we dealt mainly with sequences of independent random variables. In this chapter we consider a discrete-value random sequence $X_0, X_1, X_2, \dots$ that is not an i.i.d. random sequence. We will focus on specific sequences, called *Markov chains*, in which X_{n+1} depends on the values $X_0, X_1, \dots, X_{n-1}$ only through X_n. To keep the notation simple, we will restrict our attention to the case where each X_n takes value on the set of positive and negative integers. Also, it is convenient to think of n as a discrete-time variable.

7.1 Fundamental Definitions

Definition 52 (Discrete Time Markov Chain) *A discrete time Markov chain* $\{X_n\}_{n=0}^{\infty}$ *is a discrete value random sequence satisfying*

$$\mathbb{P}[X_{n+1} = j \,|\, X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0] = \mathbb{P}[X_{n+1} = j \,|\, X_n = i] = P_{ij} \tag{131}$$

for all $n, j, i, i_{n-1}, \dots, i_0$.

Note in particular that we have assumed that the probability P_{ij} does not depend on n. This property is sometimes referred to as *homogeinity* and will be assumed throughout the chapter.

The variable X_n is usually referred to as the *state* of the system at time n, and the set of values it can take is referred to as *state space*. The quantity P_{ij} describes the probability that the next state will be j given that the current state is i. It is usually referred to as transition probability and it satisfies the following properties: $P_{ij} \geq 0$ for all i and j, and $\sum_{j=0}^{\infty} P_{ij} = 1$ for all i.

It is often convenient to represent a Markov chain by a graph, where the nodes represent the elements in the state space and directed arcs (i, j) are drawn between all pairs of states (i, j) for which $P_{ij} > 0$.

Consider for example a simple two-state Markov chain, which can be used to model systems that alternate between ON and OFF states. The assumption is that after a unit of time in the OFF state, the system turns ON with probability p. Similarly, after a unit of time in the ON state, the system turns OFF with probability q. Let us denote by 0 the OFF state and by 1 the ON state. Then the Markov chain has transition probabilities $P_{00} = 1 - p$, $P_{01} = p$, $P_{10} = q$ and $P_{11} = 1 - q$ and can be graphically represented as illustrated in Fig. 1.

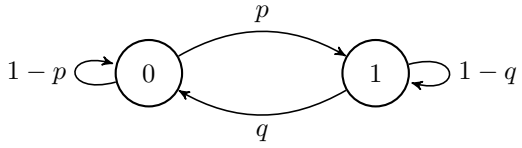

Figure 1: Markov chain describing an ON-OFF system.

The number of states in a Markov chain does not need to be finite (although we will assume so in the first part of this chapter). Consider for example the

case of a discrete random walk: a person's position is marked by an integer on the real line. Each unit of time, the person moves randomly to the right with probability p or to the left with probability $1-p$. The transition probabilities are $P_{i,i+1} = p$ and $P_{i,i-1} = 1-p$. The corresponding Markov Chain is depicted in Fig. 2 below

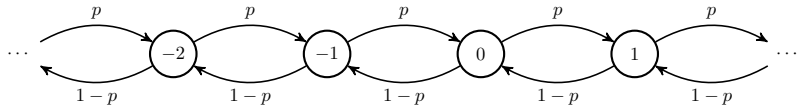

Figure 2: Markov chain describing a 1-dimensional discrete random walk.

The graphical representation of Markov chain encourages the use of the following terminology. A state transition is sometimes refer to as a *hop*, because a transition from i to j corresponds to hopping from i to j on the Markov chain. The state sequence resulting from a sequence of hops will be frequently called a *path*.

We will start our analysis of Markov chains by focusing on the case when the number of states is finite and equal to $K+1$. In this case, we can represent the one step transition probability by the matrix

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0K} \\ P_{10} & P_{11} & \cdots & \vdots \\ \vdots & & \ddots & \\ P_{K0} & \cdots & & P_{KK} \end{bmatrix}. \tag{132}$$

Note that $\mathbf{P}$ has nonnegative elements. Furthermore, each of its rows sums up to 1. This matrix is usually referred to as *state transition matrix* or *stochastic matrix*.

7.2 Dynamics of Discrete-Time Markov Chains

To describe the dynamics of Markov chains, we need to be able to predict future states X_{m+n} of the system when the current state is X_m. This information is contained in the n-step transition probabilities.

Definition 53 (n-step transition probabilities) *For a finite-state Markov chain, the n-step transition probabilities are given by the matrix $\mathbf{P}(n)$ whose elements are*

$$P_{ij}(n) = \mathbb{P}[X_{n+m} = j \,|\, X_m = i]\,. \tag{133}$$

Note that $\mathbf{P}(1) = \mathbf{P}$. The following theorem provides an efficient way to evaluate $\mathbf{P}(n)$, based on the observation that to go from i to j in $n+m$ steps, the system will need to be in some state k after n steps.

Theorem 54 (Chapman-Kolmogorov equation) *For a finite Markov chain, the n-step transition probability satisfies*

$$P_{ij}(n+m) = \sum_{k=0}^{K} P_{ik}(n) P_{kj}(m). \tag{134}$$

Equivalently,

$$\mathbf{P}(n+m) = \mathbf{P}(n)\mathbf{P}(m). \tag{135}$$

The following result is an immediate consequence of Theorem 54.

Theorem 55 *For a finite Markov chain with transition matrix* $\mathbf{P}$*, the n-step transition matrix is*

$$\mathbf{P}(n) = \mathbf{P}^n. \tag{136}$$

As an example of application of this result, let us consider the Markov chain in Fig. 1. The state transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}. \tag{137}$$

To compute $\mathbf{P}^n$, it is convenient to identify the eigenvalues of $\mathbf{P}$ and to diagonalize the matrix. The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 1-(p+q)$. The associated left eigenvectors must satisfy $\mathbf{s}_i\mathbf{P} = \lambda \mathbf{s}_i$ (note that they are row vectors) and are given by $\mathbf{s}_1 = \begin{bmatrix} q/(p+q) & p/(p+q) \end{bmatrix}$ and $\mathbf{s}_2 = \begin{bmatrix} -1 & 1 \end{bmatrix}$. Let now

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{138}$$

Then we have that

$$\mathbf{P} = \mathbf{S}^{-1}\mathbf{D}\mathbf{S}. \tag{139}$$

It then follows that

$$\mathbf{P}^n = \mathbf{S}^{-1}\mathbf{D}^n\mathbf{S} = \frac{1}{p+q}\begin{bmatrix} q & p \\ q & p \end{bmatrix} + \frac{\lambda_2^n}{p+q}\begin{bmatrix} p & -q \\ -q & p \end{bmatrix}. \tag{140}$$

When working with Markov chains, we are often interested in the value of the state probabilities $p_j(n) = \mathbb{P}[X_n = j]$ rather than of the entire n-step transition matrix. We collect these probabilities in the so called *state probability vector* $\mathbf{p}(n) = [p_0(n) \dots p_K(n)]$. Note again that this vector is defined as a row vector, since this will turn out convenient, given the convention we used to define the transition matrix $\mathbf{P}$. It is rather easy to see that the state probability vector satisfies

$$\mathbf{p}(n) = \mathbf{p}(n-1)\mathbf{P} = \mathbf{p}(0)\mathbf{P}^n. \tag{141}$$

7.3 Limiting State Probabilities

When analyzing Markov chains, we are often interested in the behavior of the state probability vector $\mathbf{p}(n)$ in the asymptotic limit $n \to \infty$. In this regime, we can assume that the system is in steady state and we are interested in the fraction of time the system spends in each of the states.

Definition 56 (Limiting State Probabilities) *For a finite Markov chain with initial state probability vector* $\mathbf{p}(0)$*, the limiting state probabilities, if they exist, are defined by the vector*

$$\boldsymbol{\pi} = \lim_{n\to\infty} \mathbf{p}(n). \tag{142}$$

To gain some insights into this definition, let us go back to the Markov chain of Fig. 1. Let $\mathbf{p}(0) = \begin{bmatrix} p_0 & p_1 \end{bmatrix}$. It follows from (140) that

$$\mathbf{p}(n) = \frac{1}{p+q}\begin{bmatrix} q & p \end{bmatrix} + \frac{\lambda_2^n}{p+q}\begin{bmatrix} p_0 p - p_1 q & -p_0 p + p_1 q \end{bmatrix} \tag{143}$$

Since $\lambda_2^n = 1 - (p+q)$, we conclude that if $0 < p + q < 2$, then

$$\boldsymbol{\pi} = \frac{1}{p+q}\begin{bmatrix} q & p \end{bmatrix}. \tag{144}$$

This is an example of a well-behaved system, in which the limiting state probability exists, and it does not depend on the initial-state probability vector $\mathbf{p}(0)$.

Limiting-state probability vectors must satisfy certain conditions.

Theorem 57 *If a finite Markov chain with transition matrix* $\mathbf{P}$ *and initial-state probability* $\mathbf{p}(0)$ *has a limiting state probability vector* $\boldsymbol{\pi} = \lim_{n\to\infty} \mathbf{p}(n)$*, then*

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}. \tag{145}$$

Closely related to limiting state probabilities are *stationary probabilities*.

Definition 58 (Stationary Probability Vector) *For a finite Markov chain with transition probability* $\mathbf{P}$*, a state probability vector* $\boldsymbol{\pi}$ *is stationary if* $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$*.*

Note that if we initialize the system with a stationary probability $\boldsymbol{\pi}$, i.e., if we set $\mathbf{p}(0) = \boldsymbol{\pi}$, then the state probabilities never change: $\mathbf{p}(n) = \boldsymbol{\pi}$ for all n. It can actually be proven that in this case the Markov chain $\{X_n\}$ is a stationary process.

Let us go back now to the concept of limiting state distribution. For a well-behaved system, we would like to think of such state distribution as the fraction of time the system spends on each state when the system is observed for a sufficiently long time. Intuitively, this should not depend on the initial conditions, as their effect should wear off as time goes by. It turns out that for this to occur, the Markov chain should be "well-behaved" in a sense we shall make precise next.

First note that for a general finite Markov chain, there are three distinct possibilities:

- $\lim_{n\to\infty} \mathbf{p}(n)$ exists and is independent of $\mathbf{p}(0)$.
- $\lim_{n\to\infty} \mathbf{p}(n)$ exists but it depends on $\mathbf{p}(0)$.
- $\lim_{n\to\infty} \mathbf{p}(n)$ does not exist.

We will see that the first case corresponds to a Markov chain with a unique stationary probability vector. The second case occurs when the Markov chain admits multiple stationary probability vectors. Finally, the third case occurs when there is no stationary probability vector. The two-state Markov chain in Fig. 1 can be used to illustrate these possibilities. As we have already seen, a stationary distribution exists, and is independent of $\mathbf{p}(0)$ when $0 < p + q < 2$. If $p = q = 0$, then the transition matrix $\mathbf{P}$ becomes the identity matrix $\mathbf{I}_2$. But then $\mathbf{p}(n) = \mathbf{p}(0)$ for every n and the initial conditions dictate completely the stationary distribution. Finally if $p + q = 2$, then $\lambda_2 = -1$ and

$$\mathbf{P}^n = \frac{1}{2}\begin{bmatrix} 1+(-1)^n & 1-(-1)^n \\ 1-(-1)^n & 1+(-1)^n \end{bmatrix}. \tag{146}$$

In this case, the limit $\lim_{n\to\infty} \mathbf{P}^n$ does not actually exist, since $\mathbf{P}^n$ has a periodic behavior. So we have no limiting state distribution.

7.4 State Classification

In the example of Fig. 1, we have seen that the Markov Chain does not have unique limiting state probabilities when either the chain disconnects into two separate chains, or when the chain causes periodic behavior in the state transitions. We will now generalize these observations to general Markov chains and identify the structural properties that determine them. To do so, we will introduce next some definitions.

Definition 59 (Accessibility) *State j is accessible from state i, which we write as $i \to j$ if $P_{ij}(n) > 0$ for some $n > 0$.*

Note that in the Markov chain graph $i \to j$ if there is a path from i to j.

Definition 60 (Communicating States) *States i and j communicate, which we write as $i \leftrightarrow j$, if $i \to j$ and $j \to i$.*

We adopt the convention that a state always communicates with itself. Hence, for every state i there is a set of states that communicate with i. Note that if both j and k communicate with i, then j and k must communicate. Thus, associated with any state i there is a set of states that all communicate with each other. We formalize this in the following definition.

Definition 61 (Communicating Class) *A communicating class is a nonempty subset of states $\mathcal{C}$ such that if $i \in \mathcal{C}$, then $j \in \mathcal{C}$ if and only if $i \leftrightarrow j$.*

So a set of states $\mathcal{C}$ is not a communicating class, if there is a state $j \notin \mathcal{C}$ that communicates with a state $i \in \mathcal{C}$.

In the example of Fig. 1, we have seen that we cannot determine the limiting state probabilities when the sequence of states has a periodic behavior. The periodicity is defined by $P_{ii}(n)$, which is the probability that the systems is in state i at time n given that it is in state i at time 0.

Definition 62 (Periodic and Aperiodic States) *State i has period d if d is the largest integer such that $P_{ii}(n) = 0$ whenever n is not divisible by d. If $d = 1$, then the state i is called aperiodic.*

We have the following important result

Theorem 63 *Communicating states all have the same period.*

To analyze the long-term behavior of a Markov chain, we need to differentiate between states that may be visited repeatedly, and states that are visited only a finite number of times.

Definition 64 (Transient and Recurrent States) *In a finite Markov chain, a state i is transient if there exists a state j such that $i \to j$ but $j \not\to i$. If no such state j exists, then the state i is recurrent.*

Theorem 65 *If i is recurrent and $i \leftrightarrow j$, then j is recurrent.*

Theorem 65 implies that if any state in a communicating class is recurrent, then all states in the communicating class must be recurrent. Similarly, if any state is transient, then all states in the communicating class must be transient. In other words, recurrence and transience are properties of the communicating class.

In the next theorem, we prove that the expected number of visits to a transient sate must be finite, since the system will eventually enter a state from which the transient state is no longer reachable.

Theorem 66 *If state i is transient, then the number N_i of visits to state i over all time has expected value $\mathbb{E}[N_i] < \infty$.*

This in particular implies that in a finite-state Markov chain not all states can be transient, otherwise, we will run out of states to visit. This is not true if a Markov chain has infinitely many states, as in the example in Fig. 2.

So for finite Markov chains, we can partition the set of states into a set of transient states $\mathcal{T}$, and r sets of recurring communicating classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_r$. So in terms of evolution of the system state, we have the following possibilities:

- If the system starts in the recurrent class $\mathcal{C}_\ell$, the system stays forever in $\mathcal{C}_\ell$.
- If the system starts in a transient state, the system passes through transient states for a finite period of time, until the system lands in a recurrent class $\mathcal{C}_\ell$. Then the system stays in $\mathcal{C}_\ell$ forever.

Because of this behavior, it makes sense to treat each recurrent class as an individual system. This leads to the following definition.

Definition 67 (Irreducible Markov Chain) *A Markov chain is irreducible if there is only one communicating class.*

7.5 Limit Theorems for Irreducible Finite Markov Chains

For irreducible, aperiodic, finite Markov chain, the limiting n-step probability admits a simple characterization, which is enabled by the following theorem.

Theorem 68 *For an irreducible, aperiodic, finite Markov chain with states $\{0, 1, 2, \dots, K\}$, the limiting n-step transition matrix is*

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \pi_0 & \pi_1 & \dots & \pi_K \end{bmatrix} = \begin{bmatrix} \pi_0 & \pi_1 & \dots & \pi_K \\ \pi_0 & \pi_1 & \dots & \pi_K \\ \vdots & & \ddots & \vdots \\ \pi_0 & \pi_1 & \dots & \pi_K \end{bmatrix} \tag{147}$$

where $\boldsymbol{\pi} = \begin{bmatrix} \pi_0 & \pi_1 & \dots & \pi_K \end{bmatrix}$ is the unique vector satisfying

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}, \quad \sum_{k=0}^{K} \pi_k = 1 \tag{148}$$

with $\pi_k \geq 0$, $k = 0, 1, \dots, K$.

Here comes an important note for the computation of $\boldsymbol{\pi}$. Although the system of equations $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ has $K + 1$ equations with $K + 1$ unknown, these are not sufficient to determine $\boldsymbol{\pi}$. To see why, observe that if $\boldsymbol{\pi}$ satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, then for any constant c, the vector $\mathbf{x} = c\boldsymbol{\pi}$ also satisfies $\mathbf{x} = \mathbf{x}\mathbf{P}$. This means that one of the equations in the system of equations is redundant. The additional constraint $\sum_{k=0}^{K} \pi_k = 1$ guarantees that the solution is unique.

Note that Theorem 68 implies that the limiting state probability does not depend on the initial state probability vector $\mathbf{p}(0)$. We state this observation in the following theorem.

Theorem 69 *For an irreducible, aperiodic, finite Markov chain with transition matrix $\mathbf{P}$ and initial state probability vector $\mathbf{p}(0)$, we have that*

$$\lim_{n\to\infty} \mathbf{p}(n) = \boldsymbol{\pi}. \tag{149}$$

Solving the system of equations in Theorem 68 to obtain the stationary probabilities $\boldsymbol{\pi}$ is not always straightforward. In the next theorem, we present a useful alternative approach to compute $\boldsymbol{\pi}$. It relies on the following observation. Let us assume that we partition the state space into to exclusive subsets $\mathcal{S}$ and $\mathcal{S}'$. One way to do so, is to cut the graph of the Markov chain into two subgraphs. We then count the number of crossings over time between $\mathcal{S}$ and $\mathcal{S}'$

and between $\mathcal{S}'$ and $\mathcal{S}$. The key observation is that this cumulative number of crossings cannot differ by more than 1, because we cannot make two $\mathcal{S} \to \mathcal{S}'$ crossings without a $\mathcal{S}' \to \mathcal{S}$ crossing. This observation is made precise in the following theorem.

Theorem 70 *Consider an irreducible, aperiodic finite Markov chain with transition probabilities P_{ij} and stationary probabilities π_i. Then for every partition of the state space into mutually exclusive subsets $\mathcal{S}$ and $\mathcal{S}'$, we have that*

$$\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}'}\pi_i P_{ij} = \sum_{j\in\mathcal{S}'}\sum_{i\in\mathcal{S}}\pi_j P_{ji}. \tag{150}$$

Note that this is theorem also holds for the countably infinite chains which are introduced below.

7.6 Countably Infinite Chains

We next look at what happens if the number of state in the chain is countably infinite. We shall focus for simplicity on the case of a single communicating class. As for finite-state Markov chains, Chapman-Kolmogorov equations in Theorem 54 hold, which means that we can compute the n-step transition probabilities iteratively from the initial-state probabilities.

As in the case of finite chains, we are interested in obtaining the limiting state probabilities $\pi_j = \lim_{n\to\infty} p_j(n)$. It turns out that to determine under which circumstances a limiting state probability exists, we need to develop a new definition for transient states, and we need to distinguish between two types of recurrent states.

Definition 71 (First return time) *Given that the system is in state i at an arbitrary time, T_{ii} is the time (expressed in number of state transitions) until the system first returns to state i.*

Using the definition of T_{ii}, we can define transient and recurrent states for countably infinite chains.

Definition 72 (Transient and Recurrent States) *For a countably infinite Markov chain, state i is recurrent if*

$$\mathbb{P}[T_{ii} < \infty] = 1. \tag{151}$$

Otherwise the state is transient.

Note that this definition applies also to finite Markov chain. Verifying that (151) holds for finite Markov chain is relatively straightforward. It can easily be inspected from the graph of the Markov chain. For countably infinite Markov chains, this verification is more complicated, as shown in the next example. Consider the Markov chain in Fig. 3. Note that in this chain, for each state i we have that $P_{i,0} = 1/(i+1)$ and $P_{i,i+1} = i/(i+1)$. We want to determine whether

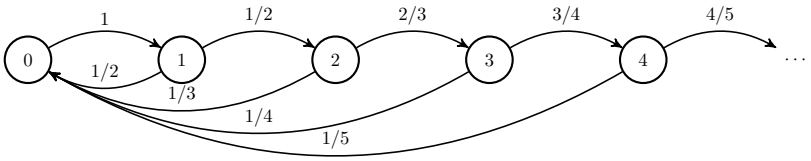

Figure 3: An infinitely countable Markov chain.

state 0 is transient or recurrent. Assume that the system starts in state 0 at time 0. We have that $T_{00} > n$ if the system reaches state n before returning to state 0. This occurs with probability

$$\mathbb{P}[T_{00} > n] = 1 \times \frac{1}{2} \times \frac{2}{3} \times \cdots \times \frac{n-1}{n} = \frac{1}{n}. \tag{152}$$

Since this probability vanishes as $n \to \infty$, we conclude that state 0 is recurrent.

For a transient state, let $p = 1 - \mathbb{P}[T_{ii} < \infty]$. Let us also denote by N_{ii} the number of returns to state i over time. Then N_{ii} is a shifted geometric random variable with probability mass function

$$P_{N_{ii}}(n) = (1-p)^n p, \quad n = 0, 1, 2, \ldots. \tag{153}$$

Using (216), one can verify that $\mathbb{E}[N_{ii}] = (1-p)/p$. This means that if a state is transient, the expected number of return visits $\mathbb{E}[N_{ii}]$ is finite. This observation leads to the following theorem.

Theorem 73 *State i is recurrent if and only if $\mathbb{E}[N_{ii}] = \infty$.*

It turns out that countably infinite chains permit two kinds of recurrent states depending on how long it takes in average to revisit a state.

Definition 74 (Positive Recurrence and Null Recurrence) *A recurrent state i is positive recurrent if $\mathbb{E}[T_{ii}] < \infty$. Otherwise, the state is null recurrent.*

If a state is recurrent, then the system will visit it again with probability 1. The distinguishing property of positive recurrent states is that the expected time to revisit the state is finite. As an example let us go back to the example in Fig. 3. Note that, for $n > 1$,

$$\mathbb{P}[T_{00} = n] = \mathbb{P}[T_{00} > n-1] - \mathbb{P}[T_{00} > n] = \frac{1}{n(n-1)}. \tag{154}$$

Then,

$$\mathbb{E}[T_{00}] = \sum_{n=2}^{\infty} n\, \mathbb{P}[T_{00} = n] = \sum_{n=2}^{\infty} \frac{1}{n-1} = \infty. \tag{155}$$

So we conclude that state 0 is null recurrent.

It turns out that positive recurrence, null recurrence, and transience are class properties.

Theorem 75 *For a communicating class of a Markov chain, either all states are transient, or all states are null recurrent, or all states are positive recurrent.*

The existence of a limiting state distribution for a countably infinite irreducible Markov chain requires all its states to be positive recurrent. Instead of verifying the conditions given in Definitions 72 and 74, it turns out sufficient to check whether a stationary distribution exists.

Theorem 76 *An irreducible Markov chain with transition probabilities P_{ij} is positive recurrent if and only if there exist stationary probabilities π_i satisfying*

$$\sum_{j=0}^{\infty} \pi_j = 1 \quad \pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij},\ j = 0, 1 \dots \tag{156}$$

Furthermore, if such a solution exists, then it is unique.

One can also show that if the state probabilities π_j exists, then $\pi_j = 1/\,\mathbb{E}[T_{jj}]$.

7.7 Exercises

Exercise 18 (Four State Markov Chain) *Consider a Markov chain with state space $\{0, 1, 2, 3\}$ and transition matrix*

$$\mathbf{P} = \begin{bmatrix} 0 & 3/10 & 1/10 & 3/5 \\ 1/10 & 1/10 & 7/10 & 1/10 \\ 1/10 & 7/10 & 1/10 & 1/10 \\ 9/10 & 1/10 & 0 & 0 \end{bmatrix}. \tag{157}$$

- *Find the stationary distribution of the Markov chain.*
- *Find the probability of being in state 3 after 32 steps if the chain begins in state 0.*
- *Find the probability of being in state 3 after 128 steps if the chain begins at a state chosen uniformly at random from the four states.*
- *Suppose that the chain begins in state 0. What is the smallest value of n for which $\max_j \left|P^n_{0,j} - \pi_j\right| \leq 0.01$?*

Exercise 19 (Loyalty Coffee Card) *Each morning on your way to Chalmers, you go to a coffee shop with probability p and get a cup of coffee. With each cup purchased, you get your "loyalty card" punched. After 7 punches, you redeem your club card on your next visit for one free cup of coffee and then receive a new unpunched card. Let X_n denote the number of punches on your card when you wake up on day n. What is $\pi_k = \lim_{n\to\infty} \mathbb{P}[X_n = k]$?*

Exercise 20 (Game Board) *A circular game board has K spaces numbered $0, 1, \dots K-1$. Starting at space 0 at time $n = 0$, a player rolls a fair six-sided die to move a token. Hence, given the current token position X_n, the next token*

position is $X_{n+1} = (X_n + R_n) \mod K$ *where* R_n *is the result of the player's nth roll. Find the stationary probability vector.*

python*: Fix* $K > 20$ *and initialise* $2K$ *different state probability vectors* $\boldsymbol{p}_i(0)$, $i \in \{0, \cdots, 2K-1\}$. *Calculate* $\boldsymbol{p}_i(10^3)$, $i \in \{0, \cdots, 2K-1\}$ *and see if it is close to the stationary probability vector. Plot your results.*

Exercise 21 (Convenience Store) *In each one-second interval at a convenience store, a new customer arrives with probability* p*, independent of the number of customers in the store. The clerk gives each arriving customer a friendly "Hello". In each unit of time in which there is no arrival, the clerk can provide a unit of service to a waiting customer. Given that a customer has received a unit of service, the customer departs with probability* q*. When the store is empty, the clerk is idle. Sketch a Markov chain for the number of customers in the store. Under which conditions on* p *and* q *do limiting state probabilities exist? Under those conditions, find the limiting state probabilities.*

python*: Let* $p = 0.3$ *and* $q = 0.5$*. Initialise* 100 *different probability state vectors* $\boldsymbol{p}_i(0)$, $i \in \{0, \cdots, 99\}$ *and run them through the Markov chain for* 1000 *steps. You need to limit the size of* $\boldsymbol{p}_i(n)$ *i.e. how many states you model. Pick this termination based on* p, q *and the limiting state probability that you calculated above such that at convergence you will miss less than* 10^{-8} *of the total probability. Plot the final state probability vectors* $\boldsymbol{p}_i(1000)$ *and compare them to your limiting state probability.*

8 Continuous Distributions, Poisson Process, and Continuous-Time Markov Chains

In this section, we review the concept of continuous random variables and review some important distributions, such as the uniform and the exponential distribution. Then we introduce the so-called Poisson process and continuous-time Markov chains.

8.1 Continuous Random Variables

The probability distribution of a random variable can be described in terms of its cumulative distribution function $F(x)$ defined as

$$F(x) = \mathbb{P}[X \leq x] \tag{158}$$

for all $x \in \mathbb{R}$. We say that a random variable is *continuous* if its distribution function $F(x)$ is a continuous function of x. In this case, we have that $\mathbb{P}[X = x] = 0$ for all x and also that $\mathbb{P}[X \leq x] = \mathbb{P}[X < x]$. If there exists a function $f(x)$ such that for all $-\infty < a < \infty$,

$$F(a) = \int_{-\infty}^{a} f(t)\, dt, \tag{159}$$

then $f(x)$ is called the *probability density function* of $F(x)$ and $f(x) = F'(x)$.

Using $f(x)$ we can express the probability of an interval $[a, b]$ as

$$\mathbb{P}[a \leq X \leq b] = \int_{a}^{b} f(x)dx. \tag{160}$$

Furthermore, the moments of X can be computed as

$$\mathbb{E}\left[X^k\right] = \int_{-\infty}^{\infty} x^k f(x)dx. \tag{161}$$

More generally, for any function $g(\cdot)$, we have that

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx. \tag{162}$$

In the following lemma, we provide a continuous analog to Lemma 23.

Lemma 77 (Expectation as Integral of Complementary Distribution Function) *Let X be a continuous random variable that takes on only nonnegative values. Then*

$$\mathbb{E}[X] = \int_{0}^{\infty} \mathbb{P}[X \geq x]\, dx. \tag{163}$$

8.1.1 Joint Distribution and Conditional Probability

The notion of distribution function can be easily generalized to multiple random variables.

Definition 78 (Joint Distribution Function) *The joint distribution function of X and Y is*

$$F(x,y) = \mathbb{P}[X \le x, Y \le y]\,. \tag{164}$$

The variables X and Y have joint density function $f(\cdot,\cdot)$ if, for all x and y,

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv. \tag{165}$$

Hence,

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\,\partial y}. \tag{166}$$

The random variables X and Y are independent if for all x and y,

$$\mathbb{P}[(X \le x) \cap (Y \le y)] = \mathbb{P}[X \le x]\,\mathbb{P}[Y \le y]\,. \tag{167}$$

So two random variables are independent if their joint distribution function satisfies $F(x,y) = F_X(x)F_Y(y)$, where $F_X(x)$ and $F_Y(y)$ are the marginal distributions of the two random variables. By taking derivatives it follows that if X and Y are independent, then $f(x,y) = f(x)f(y)$.

Conditional probabilities for random variables need to be defined carefully. Recall that when defining conditional probabilities in Definition 6 we required the probability of the event we conditioned with respect to to be positive. But what if X and Y are continuous random variables and we are interested in $\mathbb{P}[X \le x \,|\, Y = y]$? Then we have a problem, because $\mathbb{P}[Y = y] = 0$. The way to solve this issue is let Y belong to a small interval containing y and then let the size of the interval go to zero:

$$\mathbb{P}[X \le x \,|\, Y = y] = \lim_{\delta \to 0} \mathbb{P}[X \le x \,|\, y \le Y \le y + \delta]\,. \tag{168}$$

This choice leads to the following definition:

$$\mathbb{P}[X \le x \,|\, Y = y] = \int_{-\infty}^{x} \frac{f(u,y)}{f_Y(y)}du. \tag{169}$$

The value

$$f_{X\,|\,Y}(x,y) = \frac{f(u,y)}{f_Y(y)} \tag{170}$$

is called conditional density function. Using this quantity, we can evaluate the conditional expectation as follows:

$$\mathbb{E}[X \,|\, Y = y] = \int_{-\infty}^{\infty} x f_{X\,|\,Y}(x,y)dx. \tag{171}$$

8.2 The Uniform Distribution

We say that a random variable X that assumes values in the interval $[a, b]$ is uniformly distributed if all subintervals of equal lengths have the same probability. The density function of such random variable is

$$f(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{if } x > b. \end{cases} \tag{172}$$

The distribution function is given by

$$F(X) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & \text{if } x > b. \end{cases} \tag{173}$$

The expectation of X is $\mathbb{E}[X] = (b + a)/2$, and the variance is $\mathbb{V}\text{ar}[X] = (b - a)^2/12$.

8.2.1 Properties of the uniform distribution

Lemma 79 *Let X be a uniform random variable on $[a, b]$. Then, for $c \le d$ where $c, d \in [a, b]$,*

$$\mathbb{P}[X \le c \,|\, X \le d] = \frac{c - a}{d - a}. \tag{174}$$

This means that the conditional distribution of X, given that $X \le d$ coincides with that of a uniform random variable on $[a, d]$.

Here is another fact about the uniform distribution: if n points are uniformly distributed, then they are roughly equispaced.

Lemma 80 *Let $X_1, X_2, \ldots, X_n$ be independent uniform random variable over $[0, 1]$. Let $Y_1, Y_2, \ldots, Y_n$ be the same values as $X_1, X_2, \ldots, X_n$ sorted in increasing order. Then*

$$\mathbb{E}[Y_k] = \frac{k}{n + 1}. \tag{175}$$

8.3 The Exponential Distribution

Definition 81 *An exponential distribution with parameter θ is given by the following probability distribution*

$$F(x) = \begin{cases} 1 - e^{-\theta x}, & \text{if } x \ge 0 \\ 0, & \text{otherwise.} \end{cases} \tag{176}$$

The density function is

$$f(x) = \theta e^{-\theta x}, \quad \text{for } x \ge 0. \tag{177}$$

The mean is $\mathbb{E}[X] = 1/\theta$ and the variance is $\mathbb{V}\text{ar}[X] = 1/\theta^2$.

8.3.1 Properties of the Exponential Distribution

Lemma 82 *For an exponential random variable with parameter θ,*

$$\mathbb{P}[X > s + t \,|\, X > t] = \mathbb{P}[X > s]\,. \tag{178}$$

In words, the exponential distribution satisfies the memoryless property. It can be seen as a continuous version of the geometric distribution. It models the time until the first success in a memoryless continuous time process.

The minimum of several exponential random variables also exhibits some interesting properties.

Lemma 83 *If $X_1, X_2, \ldots, X_n$ are independent exponentially distributed random variables with parameters $\theta_1, \theta_2, \ldots \theta_n$, respectively, then $\min\{X_1, X_2, \ldots, X_n\}$ is exponentially distributed with parameter $\sum_{i=1}^n \theta_i$ and*

$$\mathbb{P}[\min\{X_1, X_2, \ldots, X_n\} = X_i] = \frac{\theta_i}{\sum_{j=1}^n \theta_j}. \tag{179}$$

Here is an example of the application of this lemma. Suppose that an airline ticket counter has n service agents, where the time that agent i takes per customer has an exponential distribution with parameter θ_i. You stand ahead of the line at time T_0 and all the n agents are busy. What is the average time you wait for an agent? Here is the solution. Because service times are exponentially distributed, the time remaining for agent i to be free is exponential with parameter θ_i (memoryless property). Therefore the time until the first agent becomes free is exponential with parameter $\sum_{j=1}^n \theta_j$. So the expected waiting time is $1/\sum_{j=1}^n \theta_j$. Lemma 83 also tells us that the probability that the ith agent is the first agent to become free is $\theta_i/\sum_{j=1}^n \theta_j$.

8.4 The Poisson Process

A counting process $N(t)$ is a random process that starts at time 0 and counts the occurrences of events. These events are often referred to as arrivals, because counting process are traditionally used to model the arrivals of customers at service facilities. We can use an i.i.d. Bernoulli process $X_1, X_2, \ldots$ to model a simple counting process. Specifically, consider a small step size of Δ seconds such that it is reasonable to assume that there is only at most one arrival in the interval $(n\Delta, (n+1)\Delta]$. We assume that an arrival occurs if $X_n = 1$ and that $X_n = 0$ implies no arrivals within this interval. For an average arrival rate $\lambda > 0$ arrivals/second, we can choose Δ so that $\lambda\Delta \ll 1$. Under these assumptions, the number of arrivals N_m before time $T = m\Delta$ follows a binomial distribution with parameters m and $\lambda\Delta = \lambda T/m$. If we now let $\Delta \to 0$ or equivalently $m \to \infty$, it follows from Theorem 49 that the number of arrivals before time T becomes a Poisson distribution with parameter λT.

More generally, for any interval $(t_0, t_1]$, the number of arrivals would have a Poisson distribution with parameter $\lambda(t_1 - t_0)$. Furthermore, the number of

arrivals in an interval in another nonoverlapping interval is independent from the number of arrivals in $(t_0, t_1]$, because the two counting random variables depend on independent Bernoulli random variables. We call this process the *Poisson process*.

Definition 84 (Poisson Process) *A counting process $N(t)$ is a Poisson process of rate λ if*

- *The number of arrivals $N(t_1) - N(t_0)$ in any interval (t_0, t_1) is a Poisson random variable with parameter $\lambda(t_1 - t_0)$.*
- *For every pair of nonoverlapping intervals $(t_0, t_1]$ and $(t_0', t_1']$, the number of arrivals in each interval $N(t_1) - N(t_0)$ and $N(t_1') - N(t_0')$, respectively, are independent random variables.*

The parameter λ is called rate of the process, since the expected number of arrivals per unit time is $\mathbb{E}[N(t)]/t = \lambda$.

8.4.1 Interarrival Distribution

Let X_1 be the time of the first arrival in the Poisson process and let X_n be the interval of time between the $(n-1)$th and the nth arrival. The random variables X_n are usually referred to as *interarrival times*. It turns out that for a Poisson process these arrival times have all the same distribution and this distribution is exponential.

Theorem 85 (Interarrival Times in Poisson Processes) *Let X_i, $i = 1, 2, \ldots$ denote the interarrival times in a Poisson process rate λ. Then the $\{X_i\}$ are i.i.d. exponential random variables with parameter λ.*

8.4.2 Combining and Splitting Poisson Processes

Theorem 86 (Combining Poisson processes) *Let $N_1(t)$ and $N_2(t)$ be independent Poisson processes with parameters λ_1 and λ_2, respectively. Then $N_1(t) + N_2(t)$ is a Poisson process with parameter $\lambda_1 + \lambda_2$. Furthermore, each event of the process $N_1(t) + N_2(t)$ arises from the process $N_1(t)$ with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$.*

Theorem 87 (Splitting Poisson processes) *Suppose that we have a Poisson process $N(t)$ with rate λ. Each event is independently labeled as being type 1 with probability p or type 2 with probability $1-p$. Then the type-1 events form a Poisson process $N_1(t)$ of rate λp, the type-2 events form a Poisson process $N_2(t)$ of rate $\lambda(1-p)$ and the two Poisson processes are independent.*

8.4.3 Conditional Arrival Time Distribution

Another consequence of the fact that the distribution of the interarrival time is exponential is the following. If we condition on exactly one event occurring within an interval, then the actual time at which that event occurs is uniformly

distributed over that interval. To prove this, consider a Poisson process where $N(t) = 1$, and consider the time X_1 of the single event that falls in the interval $(0, t]$. Then

$$\mathbb{P}[X_1 < s \,|\, N(t) = 1] = \frac{\mathbb{P}[(X_1 < s) \cap (N(t) = 1))]}{\mathbb{P}[N(t) = 1]} \tag{180}$$

$$= \frac{\mathbb{P}[(N(s) = 1) \cap (N(t) - N(s) = 0)]}{\mathbb{P}[N(t) = 1]} \tag{181}$$

$$= \frac{(\lambda s e^{-\lambda s}) e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \tag{182}$$

$$= \frac{s}{t}. \tag{183}$$

Here the last but one step follows from the independence of $N(s)$ and $N(t) - N(s)$.

8.5 Continuous Time Markov Process

In this section, we generalize the concept of discrete-time Markov chains to the case when the transition can occur at any time, rather than only at discrete time instants. These systems are described by a discrete-value stochastic process $X(t)$. As before, we will assume for simplicity that the state space is $\{0, 1, 2, \dots\}$. To define such processes, we will first assume that within a small time Δ, only one transition can occur, and then we will let $\Delta \to 0$. Specifically, for a given $\Delta \ll 1$, we assume that the continuous-time Markov chain $X(t)$ satisfies

$$\mathbb{P}[X(t + \Delta) = j \,|\, X(t) = i] = q_{ij}\Delta \tag{184}$$

and

$$\mathbb{P}[X(t + \Delta) = i \,|\, X(t) = i] = 1 - \sum_{j \neq i} q_{ij}\Delta. \tag{185}$$

Note in particular that

$$\mathbb{P}[X(t + \Delta) \neq i \,|\, X(t) = i] = \sum_{j \neq i} q_{ij}\Delta. \tag{186}$$

So this means that in every infinitesimal time interval, the system goes out of the present state i according to a Bernoulli distribution with parameter $\sum_{j \neq i} q_{ij}\Delta$. Let

$$\nu_i = \sum_{j \neq i} q_{ij}. \tag{187}$$

We see that the continuous-time Markov chain is then closely related to the Poisson process: indeed, the number of state changes over an interval $T = m\Delta$, given that the system is in state i is a binomial distribution with parameters m and $\nu_i\Delta = \nu_i T/m$. So as we let $\Delta \to 0$, which implies $m \to \infty$, the number of state changes is a Poisson random variable with parameter $\nu_i T$. Furthermore, the time until the next transition will be an exponential random variable with

parameter ν_i. We call ν_i the *departure rate* of state i. Note in particular that because of the memoryless property of the exponential distribution, the amount of time the system has already spent in state i has no influence on the time it will take until the next transition. This is aligned with the requirement that for a continuous-time Markov chain, $X(t)$ summarizes the state history prior to time t.

We can further interpret the state transitions for a continuous-time Markov chain in terms of the sum of independent Poisson processes. Specifically, when the system enters state i at time 0, we start a Poisson process $N_{ik}(t)$ of rate q_{ik} for every $k \neq i$. If the process $N_{ij}(t)$ is the first to have an arrival, then the system transition to state j. The system then resets and starts a Poisson process $N_{jk}(t)$ for each state $k \neq j$. It then follows from Theorem 86, that when the system is in state i the time until a transition is an exponential random variable with parameter ν_i. Furthermore, let D_i the event that the system departs state i in the time interval $(t, t+\Delta]$. Let also D_{ij} the probability that when departing state i, the system went to state j. Then

$$\mathbb{P}[D_{ij} \,|\, D_i] = \frac{\mathbb{P}[D_{ij}]}{\mathbb{P}[D_i]} = \frac{q_{ij}\Delta}{\nu_i \Delta} = \frac{q_{ij}}{\nu_i}. \tag{188}$$

So another way to interpret the transition probability is as follows. Starting from state i, the system spends an amount of time in state i described by an exponential random variable with parameter ν_i. Then the system moves to state j with probability $P_{ij} = q_{ij}/\nu_i$. Note that ν_i has the function of normalizing the probabilities q_{ij}, when summed for all $j \neq i$. In other words, the transition probabilities P_{ij} can be viewed as the transition probabilities of a discrete-time Markov chain, if we ignore the time spent in each state. The resulting discrete-time Markov chain is sometimes referred to as *embedded* discrete-time Markov chain.

The state classification we introduced for discrete-time Markov chains in Section 7 applies also to continuous-time Markov chains, via the corresponding embedded discrete-time Markov chain.

8.5.1 Limiting State Distribution

To summarize, so far we have introduced the probabilities $\{q_{ij}\}$, which we can interpret as transition rates from i to j, and the probabilities $\{\nu_i\}$, which we can interpret as departure rates from i. It will be useful for the following calculation, to define the following additional rates:

$$r_{ij} = \begin{cases} q_{ij} & i \neq j \\ -\nu_i & i = j. \end{cases} \tag{189}$$

We will first investigate the problem of how to calculate the probability that the system is in state j, which we will denote as

$$p_j(t) = \mathbb{P}[X(t) = j]\,. \tag{190}$$

In the discrete-time case, the evolution of the state probabilities is governed by the Chapman-Kolmogorov equations. In the continuous-time case, since we let $\Delta \to 0$, these equations are replaced by continuous-time differential equations.

Theorem 88 (Evolution of State Probabilities) *For a continuous-time Markov chain, the state probabilities $p_j(t)$ evolve according to the differential equations*

$$\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = \sum_i r_{ij} p_i(t), \quad j = 0, 1, 2, \ldots \tag{191}$$

As for the discrete-time case, we are interested in systems in which the state probabilities converge to constant values, i.e., they are in steady state. This happens when

$$\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = 0, \quad \text{for all } j. \tag{192}$$

In this case, the limiting state distribution is also a stationary distribution, since if $p_j(t) = p_j$ for all j, then its derivative is 0 and $p_j(t)$ never changes. In the next theorem we provide the conditions under which a limiting state probability exists, and explain how to compute it.

Theorem 89 (Limiting State Probabilities) *For an irreducible, positive recurrent continuous-time Markov chain, the state probabilities satisfy*

$$\lim_{t\to\infty} p_j(t) = p_j, \quad j = 0, 1, 2, \ldots \tag{193}$$

where the limiting state probabilities are the unique solutions to

$$\sum_i r_{ij} p_i = 0, \quad i = 0, 1, 2, \ldots \tag{194}$$

and

$$\sum_j p_j = 1. \tag{195}$$

Note that since $r_{jj} = -v_j$ and $r_{ij} = q_{ij}$, it follows from Theorem 89 that

$$p_j v_j = \sum_{i \neq j} p_i q_{ij} \tag{196}$$

So the transition rate out of state j is the sum of the average rates of transitions from i to j summed over all states $i \neq j$. So the limiting state probabilities balance the average transition rate into a state with the average transition rate out of the state.

8.6 Birth-Death Processes and Queuing Systems

One of the simplest form of continuous-time Markov chain is the birth-death process we define next.

Definition 90 (Birth-Death Process) *A continuous-time Markov chain is a birth-death process if the transition rates satisfy $q_{ij} = 0$ for $|i - j| > 1$.*

In words, when the Birth-Death Process is in state i, it can transition only to state $i - 1$ or $i + 1$. And this holds for all the states. Such processes earn their names because the state can represent the size of a population. A transition from i to $i + 1$ is a birth, whereas a transition from i to $i - 1$ is a death.

Queuing systems are often modeled as birth-death processes in which the population is the number of customers in the queue. For queuing systems, one often uses a specific terminology and notation. For example, the transition probability $q_{i,i-1}$ is usually denoted by μ_i and called *service rate*, because a customer departs only after being served. Similarly, $q_{i,i+1}$ is denoted by λ_i and called the *arrival rate*, since a transition from i to $i+1$ corresponds to the arrival of a customer. Throughout, we will assume that $\mu_i > 0$ for all states i that are reachable from state 0. This results in an irreducible chain.

For this kind of processes the limiting state probabilities are easy to compute. We start by using Theorem 89 to characterize the stationary probabilities.

Theorem 91 (Stationary probabilities) *For a birth-death queue with arrival rates λ_i and service rates μ_i, the stationary probabilities p_i satisfy*

$$p_{i-1}\lambda_{i-1} = p_i\mu_i \tag{197}$$

and

$$\sum_{i=1}^{\infty} p_i = 1. \tag{198}$$

This theorem implies that the limiting state probabilities can be computed as follows.

Theorem 92 (Limiting State Probabilities) *For a birth-death queue with arrival rates λ_i and service rates μ_i, let $\rho_i = \lambda_i/\mu_{i+1}$. The limiting state probabilities p_i, if they exist, satisfy*

$$p_i = \frac{\prod_{j=0}^{i-1} \rho_j}{1 + \sum_{k=1}^{\infty} \prod_{j=0}^{k-1} \rho_j}. \tag{199}$$

Whether the limiting state probabilities depends on whether the chain in positive recurrent or not. For birth-death process, this depends on whether the sum $\sum_{k=1}^{\infty} \prod_{j=0}^{k-1} \rho_j$ converges.

We next describe several common queue models. In queuing theory, one uses a naming convention of the form A/S/n/m for common types of queues. In this notation, "A" describes the arrival process "S" describes the service time, "n" the number of servers and "m" the maximum number of customers that can be in the queue. For example A = M indicates that the arrival process is *memoryless*, which implies that the number of arrivals is Poisson distributed. Other possibilities are A = D, which denotes a *deterministic* arrival

process in which the inter-arrival times are constant, and A = G indicates a *general* arrival process. In all cases, a common assumption is that the arrival process is independent of the service process. Similarly, S = M corresponds to memoryless, i.e., exponential, service times, S = D is for deterministic service times, and S = G denotes a general service time distribution. When the number of customers in the system is less than the number of servers n, an arriving customer is immediately assigned to a server. When the number m of customers is finite, new arrivals are blocked, i.e., discarded, when the queue has already m customers. If m is left unspecified, then it is assumed to be infinite. We next study some of these queues.

8.6.1 The M/M/1 Queue

We assume that the arrivals are a Poisson process of rate λ. The service time of a customer is an exponential random variable with parameter μ. The queue has only one server. Hence, the service rate of every state i is $\mu_i = \mu$. Assume that $\mu > \lambda$, i.e., that the service rate is larger than the arrival rate. Then, the limiting state probabilities are given by

$$p_n = (1-\rho)\rho^n, \quad n = 0, 1, 2, \dots \tag{200}$$

where $\rho = \lambda/\mu$.

Note that if $\lambda > \mu$, the new customers arrive faster than they depart. In this case, all states are transient and the queue backlog grows without bound. Note that it is a typical property of queuing systems that the system is stable, i.e., the queue has an embedded positive recurrent Markov chain and the limiting state probability exists, as long as the system service rate is greater than the arrival rate when the system is busy.

8.6.2 The M/M/∞ Queue

Let us assume now that the number of servers is ∞. Then each customer is immediately served without waiting. So when n customers are in the system, all n customers are in service and the system service rate is $\mu_n = n\mu$. For this system, the limiting state probabilities are

$$p_n = \rho^n e^{-\rho}/n! \tag{201}$$

Note that the condition $\lambda < \mu$ is not necessary in this case, since even when μ is very small, a sufficiently large backlog of n customers will yield $\mu n > \lambda$.

8.7 Exercises

Exercise 22 (Subway station) *A subway station carries trains from both the blue line (B) and the red line (R). Red line trains and blue line trains arrive as independent Poisson processes with rates $\lambda_R = 0.05$ and $\lambda_B = 0.15$ trains/min, respectively. You arrive at a random time t and wait until a red train arrives.*

Let N denote the number of blue line trains that pass through the station while you are waiting. What is $\mathbb{P}[N = n]$?

Exercise 23 (Runners) *Ten runners compete in a race starting at time $t = 0$. The runners' finishing time $R_1, \dots, R_{10}$ are i.i.d. exponential random variables with expected value $1/\mu = 10$ minutes.*

- *What is the probability that the last runner will finish in less than 20 minutes?*
- *Let X_1 be the finishing time of the winning runner. What is the distribution of X_1?*
- *Find the density function of $Y = R_1 + \cdots + R_{10}$.*
- *Let $X_1, \dots, X_{10}$ be the runners' interarrival times at the finishing line. Find the joint density function of these random variables.*

***python**: Generate 10^4 instances of the variables $R_1, \dots, R_{10}$ and use these to verify your answers'. When confirming that a variable has a certain distribution, you may want to plot the empirical pdf and compare it with the theoretical pdf.*

Exercise 24 (Tool Booths on a Highway) *A set of c tool booths at the entrance to a highway can be modeled as a queue with c servers. Assume that the service times are independent exponential random variables with mean $\mu = 1$ second. Sketch a continuous-time Markov-chain for the system. What is the maximum arrival rate such that the limiting state probability exists?*

Exercise 25 (Grocery store) *Consider a grocery store with two queues. At either queue, a customer has an exponential service time with an expected value of 3 minutes. Customers arrive at the two queues as a Poisson process of rate λ customers per minute. Consider the following possibilities:*

- *Customers choose a queue at random so each queue has a Poisson arrival process of rate $\lambda/2$.*
- *Customers wait in a combined line. When a customer completes service at either queue, the customer at the front of the line goes into service.*

For each system, calculate the limiting state probabilities. Under which system is the expected number of queuing clients smaller?

***python**: Let $\lambda = 0.5$ customers per minute. Initialise 500 different probability state vectors $\boldsymbol{p}_i(t = 0)$, $i \in \{0, \cdots, 499\}$. Terminate the size of the probability state vectors in such a way that at convergence you will miss less than 10^{-8} of the total probability. Propagate the state vectors through the system for 1000 minutes by taking small steps in time of $\Delta t = 0.01$ minutes. Compare your final state vectors $\boldsymbol{p}_i(t = 1000)$ with your calculated limiting state probability. Estimate the expected number of queuing clients using your final state vectors under both systems and compare to the calculated values from above.*

9 The Normal Distribution

The normal or Gaussian distribution plays a central role in probability theory and statistics. Empirically, many real-world observable quantities are often well approximated by the normal distribution. Furthermore, the central limit theorem states that under very general conditions, the distribution of the average of a large number of independent random variables converges to the normal distribution.

9.1 The Standard Normal Distribution

The standard normal distribution, denoted by $\mathcal{N}(0,1)$, is a continuous distribution on the real numbers. Its density function is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{202}$$

The distribution function $\Phi(z)$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} \mathrm{d}x. \tag{203}$$

Since the density $\phi(z)$ is symmetric with respect to $z = 0$, it follows that $\mathbb{E}[Z] = 0$. It is also possible to verify via integration by part that $\mathbb{V}\mathrm{ar}[X] = 1$.

9.2 The General Univariate Normal Distribution

The univariate normal distribution is characterized by two parameters μ and σ corresponding to the mean and the standard deviation and is denoted by $\mathcal{N}(\mu, \sigma^2)$. Its density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{204}$$

Note that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ follows a standard normal distribution. So we can get a random variable following a general univariate normal distribution by applying a linear transformation to a random variable following a standard normal distribution.

9.3 The Moment Generating Function

The moment generating function of $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$M_x(t) = e^{t^2\sigma^2/2 + \mu t}. \tag{205}$$

Using the moment generating function, it is easy to verify that if $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent random variables, then $X+Y$ is distributed according to $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_1^2)$.

Using the moment generating function, we can also obtain the following large deviation bound: let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\mathbb{P}[|(X-\mu)/\sigma| \geq a] \leq 2e^{-a^2/2}. \tag{206}$$

9.4 The Central Limit Theorem

The central limit theorem is one of the most fundamental results in probability theory. It states that, under various mild conditions, the distribution of the average of a large number of independent random variables converges to the normal distribution, regardless the distribution of each of the random variables. The convergence is in *distribution*, which means the following.

Definition 93 (Convergence in Distribution) *A sequence of distributions $F_1, F_2, \ldots$ converges in distribution to a distribution F, denoted as $F_n \xrightarrow{D} F$, if for every $a \in \mathbb{R}$ at which F is continuous, we have that*

$$\lim_{n\to\infty} F_n(a) = F(a). \tag{207}$$

We provide here a basic version of the central limit theorem, for the average of i.i.d. random variables with finite mean and variance

Theorem 94 (Central Limit Theorem) *Let $X_1, \ldots, X_n$ be n i.i.d. random variables with mean μ and variance σ^2.*

Let $A_n = (X_1 + \cdots + X_n)/n$. Then for every a, b,

$$\lim_{n\to\infty} \mathbb{P}\left[a \leq \frac{A_n - \mu}{\sigma/\sqrt{n}} \leq b\right] \xrightarrow{D} \Phi(b) - \Phi(a). \tag{208}$$

Note that the variance of A_n is σ^2/n and vanishes as n grows large. So to obtain the desired convergence, we need to normalize by $\sigma/\sqrt{n}$ as done in (208). To prove this result, we will use the following result

Theorem 95 (Lévy's Continuity Theorem) *Let $Y_1, Y_2, \ldots$ be a sequence of random variables where Y_i has distribution F_i and moment generating function M_i. Let Y be a random variable with distribution F and moment generating function M. If $\lim_{n\to\infty} M_n(t) = M(t)$ for all t, then $F_n \xrightarrow{D} F$ for all t for which $F(t)$ is continuous.*

The central limit theorem can be improved under a variety of condition. For example, in the following version, we do not require the random variables to be identically distributed.

Theorem 96 (More General Central Limit Theorem) *Let $X_1, \ldots, X_n$ be n independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{V}\mathrm{ar}[X_i] = \sigma_i^2$. Assume that*

- $\mathbb{P}[|X_i| < M] = 1$ *for all i and some $M < \infty$.*

- $\lim_{n\to\infty} \sum_{i=1}^{n} \sigma_i^2 = \infty$.

Then

$$\mathbb{P}\left[a \leq \frac{\sum_{i=1}^{n}(X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \leq b\right] \xrightarrow{D} \Phi(b) - \Phi(a). \tag{209}$$

Furthermore, under more stringent result one can get a uniform convergence result.

Theorem 97 (Berry-Esséen Theorem) *Let $X_1, \ldots, X_n$ be n i.i.d. random variables with finite mean μ and finite variance σ^2. Let $\rho = \mathbb{E}\left[|X_i - \mu|^2\right]$ and $A_n = (X_1 + \cdots + X_n)/n$. Then there is a constant C such that for all a*

$$\left|\mathbb{P}\left[\frac{A_n - \mu}{\sigma/\sqrt{n}} \leq a\right] - \Phi(a)\right| \leq C\frac{\rho}{\sigma^3\sqrt{n}} \tag{210}$$

9.5 Exercises

Exercise 26 (Box-Muller method) *The Box-Muller method is a well-known method to generate Gaussian random variables from uniform random variables. It works as follows. Let U and V to uniform random variables on $(0, 1)$ and set*

$$X = \sqrt{-2\ln V}\cos(2\pi U) \tag{211}$$

and

$$Y = \sqrt{-2\ln V}\sin(2\pi U). \tag{212}$$

Why does this method work? Hint: *express the joint cumulative function of two independent standard Gaussian random variable in polar coordinates.*

python: *Write a program that generates 10^5 variables X and Y according to this method. Plot the corresponding distribution functions and compare them with that of a standard Gaussian random variable. Also determine how many sampled value x satisfy $|x| \leq k$ for $k = 1, 2, 3, 4$. Do your result seem reasonable? Explain.*

10 A Collection of Useful Results

In this section, we collect some basic results we will use repeatedly during the first part of the course.

$$\sum_{k=1}^{n} k = \frac{(n+1)n}{2}. \tag{213}$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \tag{214}$$

$$\sum_{i=k}^{\infty} a^i = \frac{a^k}{1-a}, \quad 0 < a < 1. \tag{215}$$

$$\sum_{n=1}^{\infty} n a^n = \frac{a}{(1-a)^2}, \quad 0 < a < 1. \tag{216}$$

$$\sum_{k=1}^{\infty} k^2 x^k = \frac{x^2 + x}{(1-x)^3}. \tag{217}$$

$$\sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^2 = \frac{\pi^2}{6}. \tag{218}$$

$$e^x = \sum_{i=0}^{\infty} \frac{x^j}{j!}. \tag{219}$$

References