Contents

PREFACE

This book is intended for students who have completed proof-based courses in Advanced Calculus and Linear Algebra. In addition to the standard topics (such as the Poisson approximation to the binomial, law of large numbers, central limit theorem, Markov chains, and simple linear regression), several other topics and results that are accessible at this level and that fit into a one semester course are covered:

- first moment method with some applications, such as cliques in the Erdős–Rényi random graph, and an upper bound on the typical longest increasing subsequence of a random permutation;
- second moment method with applications to Bernstein's polynomials, cliques in the Erdős–Rényi random graph, and the Hardy–Ramanujan theorem;
- Hoeffding's inequality, and the Johnson–Lindenstrauss lemma;
- the Hoeffding–Chernoff inequality, and the generalization ability of classification algorithms;
- Azuma's inequality with several examples, such as the chromatic number of the Erdős–Rényi random graph, max-cut in sparse random graphs, and the Hamming distance on the hypercube.

A knowledge of Lebesgue integration is not assumed, although, when discussing continuous distributions, I tried to give some idea of why learning about it is something to look forward to. A number of exercises are included throughout and at the end of each section.

In the second edition, a number of explanations were clarified and some mistakes were corrected.

Dmitry Panchenko,
Toronto, Canada

Chapter 1
Introduction

1.1 Example: Balancing Vectors

We will begin this chapter with an example that on the surface appears unrelated to Probability. We will then go through various steps of this example and reformulate them using probabilistic terminology and notation, introducing such notions as the *probability space, probability measure, random variable, expectation, independence, sample space, change of measure, distribution*. The general definitions will appear in the following sections, but an illustration of these concepts on a simple example will, hopefully, make things clearer later on.

Example 1.1.1 (Balancing vectors). Let us consider $n \geq 2$ vectors on the unit sphere in $\mathbb{R}^n$,

$$v_1, \ldots, v_n \in \mathbb{R}^n,\ |v_i| = 1 \text{ for all } i = 1, \ldots, n.$$

We denote the length of a vector $v \in \mathbb{R}^n$ by $|v|$. Consider the following question. Among all linear combinations of these vectors $v = \pm v_1 \pm \ldots \pm v_n$ with ± 1 coefficients, can we find a choice of signs such that

$$|v| \leq \sqrt{n}?$$

Can we always choose the signs in such a way that $|v| \geq \sqrt{n}$? Notice that if vectors v_i are orthogonal then, clearly, $|v| = \sqrt{n}$ for all choices of signs, but here we do not assume that they are orthogonal. We will now show that the answer to both questions is yes.

Let us consider n variables $\varepsilon_1, \ldots, \varepsilon_n$ each taking two possible values ± 1, which will represent possible choices of signs above. Then the question can be rephrased as follows. Can we find

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \big\{-1, 1\big\}^n \tag{1.1}$$

such that the vector $v(\varepsilon) := \varepsilon_1 v_1 + \ldots + \varepsilon_n v_n$ has length

$$|v(\varepsilon)| \leq \sqrt{n}?$$

We will answer this question by computing the average of $|v(\varepsilon)|^2$ over all 2^n choices of signs:

$$\frac{1}{2^n} \sum_{\varepsilon \in \{-1,1\}^n} |v(\varepsilon)|^2 = \frac{1}{2^n} \sum_{\varepsilon \in \{-1,1\}^n} |\varepsilon_1 v_1 + \ldots + \varepsilon_n v_n|^2. \tag{1.2}$$

First of all, if we rewrite in terms of scalar products,

$$|v(\varepsilon)|^2 = |\varepsilon_1 v_1 + \ldots + \varepsilon_n v_n|^2 = \sum_{i,j=1}^{n} \varepsilon_i \varepsilon_j (v_i, v_j),$$

then the average can be written as

$$\begin{aligned} &\frac{1}{2^n} \sum_{\varepsilon \in \{-1,1\}^n} \sum_{i,j=1}^{n} \varepsilon_i \varepsilon_j (v_i, v_j) \\ &= \sum_{i,j=1}^{n} (v_i, v_j) \Big[\frac{1}{2^n} \sum_{\varepsilon \in \{-1,1\}^n} \varepsilon_i \varepsilon_j \Big], \end{aligned} \tag{1.3}$$

where we simply interchanged the order of summation. If $i = j$ then $\varepsilon_i \varepsilon_j = 1$ and

$$\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\varepsilon_i\varepsilon_j = \frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n} 1 = \frac{2^n}{2^n} = 1.$$

For $i \neq j$, notice that for each choice of $\varepsilon_i, \varepsilon_j \in \{-1,+1\}$ there are 2^{n-2} choices of the remaining $n-2$ coordinates, which do not affect the value of $\varepsilon_i\varepsilon_j$. Therefore,

$$\begin{aligned}\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\varepsilon_i\varepsilon_j &= \frac{2^{n-2}}{2^n}\sum_{\varepsilon_i,\varepsilon_j\in\{-1,1\}}\varepsilon_i\varepsilon_j = \frac{1}{2^2}\sum_{\varepsilon_i,\varepsilon_j\in\{-1,1\}}\varepsilon_i\varepsilon_j \\ &= \Big(\frac{1}{2}\sum_{\varepsilon_i\in\{-1,1\}}\varepsilon_i\Big)\times\Big(\frac{1}{2}\sum_{\varepsilon_j\in\{-1,1\}}\varepsilon_j\Big) = 0\times 0 = 0. \qquad (1.4)\end{aligned}$$

This means that the average in (1.3) is

$$\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}|v(\varepsilon)|^2 = \sum_{i=1}^n (v_i, v_i) = n,$$

because we assumed that all $|v_i| = 1$. Since the average is n, all $|v(\varepsilon)|^2$ can not be strictly bigger than n, because the average would be strictly bigger than n.This shows that there exists ε such that $|v(\varepsilon)|^2 \leq n$. By the same logic, there exists a choice of signs ε such that $|v(\varepsilon)|^2 \geq n$. □

Let us introduce several key concepts using the setting of the above example. In the probabilistic language, the set

$$\Omega := \big\{-1,1\big\}^n \qquad (1.5)$$

that appeared in (1.1) consisting of vectors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is called a *probability space*, and elements of Ω are also called *outcomes*. To be more precise, in order for Ω to be called a probability space, it must have a *probability*, or *probability measure*, associated with it. In the above example, the measure implicitly appeared when we talked about averaging. Namely, if we let

$$\mathbb{P}(\varepsilon) = \frac{1}{2^n} \qquad (1.6)$$

represent the probability of an individual outcome ε and, for any subset $A \subseteq \Omega$, let

$$\mathbb{P}(A) = \frac{\operatorname{card}(A)}{2^n} \tag{1.7}$$

represent the probability of A, then $\mathbb{P}$ is an example of a probability measure. In this particular case, it is called a *uniform measure* on $\{-1,1\}^n$, because all outcomes are 'equally likely'. We can think of $\mathbb{P}$ as a function describing the chances of individual outcomes $\varepsilon_1,\ldots,\varepsilon_n$ if we choose them 'at random' in the everyday sense of the word, say, by flipping a coin.

Informally, probability (measure) is a function that assigns to subsets of Ω values in $[0,1]$ and satisfies certain basic properties. The general definition is more involved (we will say more in Chapter 4), but for now let us mention several properties that are obvious in the above setting, where Ω is a finite set:

(i) $\mathbb{P}(A) \in [0,1]$ for any set $A \subseteq \Omega$.
(ii) $\mathbb{P}(\Omega) = 1$.
(iii) for any disjoint sets $A_1, A_2 \subseteq \Omega$,

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

The last property is called *additivity*, or *finite additivity*, of probability, because it extends by induction to any finite number of disjoint sets $A_1,\ldots,A_n \subseteq \Omega$,

$$\mathbb{P}(A_1 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \ldots + \mathbb{P}(A_n). \tag{1.8}$$

The second property, $\mathbb{P}(\Omega) = 1$, is important because $1^2 = 1$, which allows us to take *products* of probability spaces, as well as leading to natural analogues of the Fubini theorem via the notion of conditional distributions, as we will see in Section 1.4 below.

Exercise 1.1.1. Show that the above properties (i), (ii), (iii) imply:

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, where $A^c = \Omega \setminus A$ is the complement of A in Ω.
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for any sets $A, B \subseteq \Omega$.
3. $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ for any sets $A, B \subseteq \Omega$.
4. $\mathbb{P}(A_1 \cup \ldots \cup A_n) \leq \sum_{i=1}^n \mathbb{P}(A_i)$ for any sets $A_1, \ldots, A_n \subseteq \Omega$.

The last inequality is called the *union bound*, and it will be used often.

Next, in the probabilistic language, the function

$$|v(\varepsilon)| = |\varepsilon_1 v_1 + \ldots + \varepsilon_n v_n|$$

in the above example is called a *random variable*. In fact, any function

$$f \colon \Omega \to \mathbb{R} \tag{1.9}$$

is called a random variable, so a 'random variable' is just another name for a 'function' in the probabilistic language. (Again, it is not quite correct that all functions are random variables, but we will discuss this more in Chapter 4 when we talk about continuous distributions). If we think about Ω as the set of possible outcomes ε of some random experiment (for example, flipping n coins) then the function $f(\varepsilon)$ can be thought of as some numerical attribute of this random outcome, which, hopefully, explains somewhat why functions are called random variables.

The average that appeared in (1.2) is called the *expectation* (or sometimes called *mathematical expectation*, or *expected value*) of the random variable $f(\varepsilon) = |v(\varepsilon)|^2$ and is denoted by $\mathbb{E}f$. More generally, given $f \colon \Omega \to \mathbb{R}$,

$$\mathbb{E}f = \sum_{\varepsilon \in \Omega} f(\varepsilon)\mathbb{P}(\varepsilon), \tag{1.10}$$

which is well defined because Ω is finite. Despite the fact that the expected value in the above example is just the average with weights $\mathbb{P}(\varepsilon) = 1/2^n$, it is also common to write it using integral notation,

$$\mathbb{E}f = \int_\Omega f(\varepsilon)\, d\mathbb{P}(\varepsilon), \tag{1.11}$$

to emphasize that its meaning is not much different than the familiar integrals in Calculus. For example, the interchange of summation in (1.3) is the first example of the *linearity of expectation*.

Exercise 1.1.2. Show that the definition (1.10) implies the following.

1. $\mathbb{E}(af_1 + bf_2) = a\mathbb{E}f_1 + b\mathbb{E}f_2$ for any $f_1, f_2 \colon \Omega \to \mathbb{R}$ and any $a, b \in \mathbb{R}$.
2. If $f \geq 0$ then $\mathbb{E}f \geq 0$. If $f_1 \geq f_2$ then $\mathbb{E}f_1 \geq \mathbb{E}f_2$.
3. Given a subset $A \subseteq \Omega$, consider an *indicator function* of A,

$$\mathrm{I}(\varepsilon \in A) = \begin{cases} 1, & \text{if } \varepsilon \in A \\ 0, & \text{if } \varepsilon \notin A. \end{cases}$$

Then,

$$\mathbb{E}\,\mathrm{I}(\varepsilon \in A) = \mathbb{P}(A). \tag{1.12}$$

Next, let us mention that the computation in the equation (1.4) was an illustration of another fundamental concept in Probability – *independence*. First of all, notice that $\varepsilon_1, \ldots,$ ε_n can be viewed as random variables on Ω, because ε_i is a function of ε. To emphasize this and to avoid confusion, let us give these functions a different name,

$$\pi_i(\varepsilon) = \varepsilon_i. \tag{1.13}$$

For any specific choice of signs $a_1, \ldots, a_n \in \{-1, +1\}$, let us consider an event

$$A = \Big\{\varepsilon \in \Omega \,:\, \pi_1(\varepsilon) = a_1, \ldots, \pi_n(\varepsilon) = a_n\Big\}.$$

Typically, a more concise notation will be used for sets like this,

$$A = \left\{ \pi_1 = a_1, \ldots, \pi_n = a_n \right\}.$$

In other words, we will often omit an explicit reference to the variable $\varepsilon \in \Omega$ when dealing with random variables (i.e. functions) on Ω. The above set A consists of one point $(a_1, \ldots, a_n)$, so it might look a bit strange to use such a convoluted way to describe it. However, in this language we can say what it means for the random variables $\pi_1, \ldots, \pi_n$ to be *independent*. It means that

$$\mathbb{P}\big(\pi_1 = a_1, \ldots, \pi_n = a_n\big) = \prod_{i=1}^{n} \mathbb{P}(\pi_i = a_i) \tag{1.14}$$

The chance of a bunch of things happening at the same time is the same as the product of each individual event

for any $a_1, \ldots, a_n$. We will discuss the meaning of this definition in more detail later in this chapter, but for now let us see why it holds in the balancing vectors example. By (1.6), the left hand side is $1/2^n$. What is $\mathbb{P}(\pi_i = a_i)$? The set

$$\{\pi_i = a_i\} = \{\varepsilon : \pi_i(\varepsilon) = a_i\}$$

consists of all the vectors ε with the i^{th} coordinate fixed, equal to a_i. There are 2^{n-1} such vectors so, according to our definition (1.7), the probability of this set is

$$\mathbb{P}(\pi_i = a_i) = \frac{2^{n-1}}{2^n} = \frac{1}{2},$$

which implies that (1.14) is satisfied.

Exercise 1.1.3. Show that (1.14) implies that

$$\mathbb{P}(\pi_1 = a_1, \pi_2 = a_2) = \mathbb{P}(\pi_1 = a_1)\mathbb{P}(\pi_2 = a_2).$$

Let us now show that the computation in the equation (1.4) can be rewritten as

$$\mathbb{E}\pi_i\pi_j = \mathbb{E}\pi_i\mathbb{E}\pi_j = 0 \cdot 0 = 0, \tag{1.15}$$

where the first equality is an important consequence of independence. By definition,

$$\mathbb{E}\pi_i\pi_j = \sum_{\varepsilon\in\Omega} \pi_i(\varepsilon)\pi_j(\varepsilon)\mathbb{P}(\varepsilon).$$

To make connection with (1.14), we will need to rewrite this sum in a different way (this is a very important idea of the *change of variables*). Let us first consider all possible values (a_i,a_j) that $(\pi_i(\varepsilon),\pi_j(\varepsilon))$ can take, then for each of these values consider the set

$$A(a_i,a_j) = \Big\{\varepsilon \in \Omega : \pi_i(\varepsilon) = a_i, \pi_j(\varepsilon) = a_j\Big\},$$

and partition the space Ω into a disjoint union

$$\Omega = \bigcup_{a_i,a_j} A(a_i,a_j).$$

Then the sum over ε can be written as sum over each set $A(a_i,a_j)$ and then the sum over these sets. In this particular case, $\pi_i(\varepsilon) = \varepsilon_i$ takes only two values ±1, so

$$\begin{aligned}
\mathbb{E}\pi_i\pi_j &= \sum_{\varepsilon\in\Omega} \pi_i(\varepsilon)\pi_j(\varepsilon)\mathbb{P}(\varepsilon)\\
&= \sum_{a_i,a_j=\pm1} a_i a_j \sum_{\varepsilon\in A(a_i,a_j)} \mathbb{P}(\varepsilon)\\
&= \sum_{a_i,a_j=\pm1} a_i a_j \mathbb{P}(A(a_i,a_j))\\
&= \sum_{a_i,a_j=\pm1} a_i a_j \mathbb{P}(\pi_i = a_i, \pi_j = a_j).
\end{aligned}$$

By independence, we can continue to write

$$\begin{aligned}
\mathbb{E}\pi_i\pi_j &= \sum_{a_i,a_j=\pm1} a_i a_j \mathbb{P}(\pi_i = a_i)\mathbb{P}(\pi_j = a_j)\\
&= \sum_{a_i=\pm1} a_i \mathbb{P}(\pi_i = a_i) \sum_{a_j=\pm1} a_j \mathbb{P}(\pi_j = a_j).
\end{aligned}$$

It remains to check that the factors on the right hand side are equal to $\mathbb{E}\pi_i$ and $\mathbb{E}\pi_j$. By a similar calculation,

$$\begin{aligned}\mathbb{E}\pi_i = \sum_{\varepsilon\in\Omega} \pi_i(\varepsilon)\mathbb{P}(\varepsilon) &= \sum_{a_i=\pm 1} a_i \sum_{\varepsilon:\pi_i(\varepsilon)=a_i} \mathbb{P}(\varepsilon) \\ &= \sum_{a_i=\pm 1} a_i \mathbb{P}(\pi_i = a_i), \qquad (1.16)\end{aligned}$$

as desired, so we checked the first equality in (1.15) from the definitions. For the second equality, as before, since there are 2^{n-1} different ε such that $\pi_i(\varepsilon) = \varepsilon_i = a_i$, the probability $\mathbb{P}(\pi_i = a_i) = 1/2$ and

$$\mathbb{E}\pi_i = (+1)\frac{1}{2} + (-1)\frac{1}{2} = 0.$$

Finally, let us mention some terminology associated with the last few equations, namely, the *sample space* and *change of variables*. For example, when we looked at the random variable $\pi_i(\varepsilon) = \varepsilon_i$ just now, we looked at all possible values a_i it could take (the range of π_i) and calculated the probabilities $\mathbb{P}(\pi_i = a_i)$ for all such values. In this case, the set of values was $\{-1, 1\}$ and the probabilities were both $1/2$. One can think of the pair,

$$\Omega' = \{-1, 1\},\ \mathbb{P}'(-1) = \frac{1}{2} \text{ and } \mathbb{P}'(1) = \frac{1}{2}, \qquad (1.17)$$

and a new probability space, and it is called the *sample space* of the random variable π_i. Probability measure $\mathbb{P}'$ is called the *distribution* (or sometimes *law*) of the random variable π_i. In other words, the sample space consists of all possible values of the random variable and their probabilities.

The process by which the probabilities $\mathbb{P}'$ were computed is called the *change of variables*, because $\mathbb{P}'(a_i)$ is just the probability $\mathbb{P}$ on the original space Ω of the set

$$\{\pi_i = a_i\} = \{\varepsilon : \pi_i(\varepsilon) = a_i\},$$

which can be also viewed as the pre-image $\pi_i^{-1}(a_i)$ of a_i. For this reason, we can write

$$\mathbb{P}'(a_i) = \mathbb{P}(\pi_i = a_i) = \mathbb{P}\Big(\pi_i^{-1}(a_i)\Big) =: \mathbb{P} \circ \pi_i^{-1}(a_i). \quad (1.18)$$

The distribution $\mathbb{P}' = \mathbb{P} \circ \pi_i^{-1}$ on the sample space is called the *image measure* of $\mathbb{P}$ by the map π_i. Notice that (1.16) can be rewritten as

$$\mathbb{E}\pi_i = \sum_{a_i = \pm 1} a_i \mathbb{P}'(a_i), \quad (1.19)$$

in terms of the distribution of π_i. This average with the weights given by $\mathbb{P}'$ can be taken as another definition of the expectation of the random variable, and (1.19) is called the *change of variables formula for the expectation*.

As a final comment, let us mention that independence is an analogue of the fact that the area of a rectangle is the product of the lengths of its sides,

$$\text{Area}([a,b] \times [c,d]) = \text{Length}([a,b]) \times \text{Length}([c,d]).$$

The consequence of this in Calculus is that

$$\iint_{[0,1]^2} f(x)g(y)\,dxdy = \int_0^1 f(x)\,dx \int_0^1 g(y)\,dy.$$

Similarly, $\mathbb{E}\pi_i\pi_j = \mathbb{E}\pi_i\mathbb{E}\pi_j$ above was the consequence of independence.

Exercise 1.1.4. What is the sample space and distribution of the random variable $f(\varepsilon) = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$ in the above example? Calculate its expectation using the linearity of expectation, and using the distribution.

Exercise 1.1.5. Suppose we have 50 vectors of length 3 and 50 vectors of length 5 in $\mathbb{R}^{200}$. If we do not know anything about these vectors except their lengths, can we find x such that there exist linear combinations v of these vectors with ± 1 coefficients such that $|v| \le x$ and such that $|v| \ge x$?

1.2 Discrete Probability Spaces and Distributions

In this section, we will discuss discrete probability spaces and distributions. Discrete spaces Ω consist of finitely many or countably many elements,

$$\Omega = \big\{\omega_1, \omega_2, \ldots, \omega_n, \ldots\big\}, \tag{1.20}$$

so they can be indexed by natural numbers. A probability measure on such a space is defined by assigning probabilities to individual outcomes,

$$\mathbb{P}(\omega_n) = p_n, \tag{1.21}$$

which satisfy the conditions

$$p_n \geq 0, \text{ and } \sum_{n\geq 1} p_n = 1. \tag{1.22}$$

For a subset $A \subseteq \Omega$, its probability is

$$\mathbb{P}(A) = \sum_{\omega\in A} \mathbb{P}(\omega). \tag{1.23}$$

The order of summation here does not matter, because the probabilities are nonnegative.

Exercise 1.2.1. If the series $\sum_{n\geq 1} a_n$ is absolutely convergent then, for any bijection $\pi\colon \mathbb{N} \to \mathbb{N}$, the series $\sum_{n\geq 1} a_{\pi(n)}$ gives the same answer. *Hint:* See the proof of Lemma 1.2 below.

For example, we can always sum in the order ω's were enumerated in (1.20) and, if ω_n is not in A, replace the term $\mathbb{P}(\omega_n)$ by zero in the summation,

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(\omega_n)\,\mathrm{I}(\omega_n \in A). \tag{1.24}$$

This implies all the properties in Exercise 1.1.1 (check!) or, for example, the *monotonicity* property of probability:

$$\text{if } A \subseteq B \text{ then } \mathbb{P}(A) \le \mathbb{P}(B). \tag{1.25}$$

In the last section, we denoted a generic random variable by f to emphasize that it is nothing but a real-valued function on our space. It is more common, however, to denote *generic* random variables by capital letters, for example, X, Y or Z. Of course, this is not required and in particular situations more suitable or evocative notation may be used (such as $\pi, g, z, \eta, h, \varepsilon$ etc.).

If the set Ω is finite then all the definitions in the previous section, as well as the basic properties of probability and expectation, stay the same. If the set Ω is countably infinite, the definition of expectation in (1.10),

$$\mathbb{E}X = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega), \tag{1.26}$$

is formally the same, but it is used only for random variables X that satisfy

$$\mathbb{E}|X| = \sum_{\omega\in\Omega} |X(\omega)|\mathbb{P}(\omega) < \infty. \tag{1.27}$$

Otherwise, $\mathbb{E}X$ is assumed undefined. Random variables that satisfy (1.27) are called *integrable*. One reason we have to assume the absolute convergence is to make sure that the definition of expectation does not depend on the order of summation, because the elements of Ω are not necessarily ordered. This follows from Exercise 1.2.1.

Since the order of summation does not matter, (1.26) makes sense and we can use basic properties of the series to derive basic properties of expectation, for example, the *linearity* and *monotonicity* of expectation.

Exercise 1.2.2. If $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}|X_2| < \infty$ then, for any $a, b \in \mathbb{R}$,

$$\mathbb{E}(aX_1 + bX_2) = a\mathbb{E}X_1 + b\mathbb{E}X_2. \tag{1.28}$$

Exercise 1.2.3. If $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$, and $X \le Y$, then $\mathbb{E}X \le \mathbb{E}Y$.

The entire Section 1.5 will be devoted to applications of the linearity of expectation. Notice that the representation (1.24) is a way of writing $\mathbb{P}(A) = \mathbb{E}\,\mathrm{I}(\omega \in A)$, as in (1.12) in the previous section.

Given a random variable $X\colon \Omega \to \mathbb{R}$, since Ω is countably infinite, the set of values $\{a_1, a_2, \ldots\}$ that X can take (range of X) is either finite or countably infinite. In particular, as in (1.18), we can define the *image measure* on this set by

$$\mathbb{P}'(a_n) = \mathbb{P}\big(\omega : X(\omega) = a_n\big) = \mathbb{P}\big(X^{-1}(a_n)\big). \tag{1.29}$$

Standard notation for this $\mathbb{P}'$ is $\mathbb{P} \circ X^{-1}$, and it is called the *distribution* of X.

If a probability space $\Omega \subseteq \mathbb{R}$ is a subset of the real line then the probability $\mathbb{P}$ can also be called a distribution, because it is the distribution of the identity function $X(\omega) = \omega$,

$$\mathbb{P}'(a_n) = \mathbb{P}\big(\omega : X(\omega) = \omega = a_n\big) = \mathbb{P}(a_n).$$

We will also often use the word *distribution* for $\mathbb{P}(X = a_n)$ when X is a vector consisting of several random variables. As in the equation (1.19), we can use the *change of variables* to rewrite the expectation of X in terms of its distribution.

Lemma 1.1 (Change of variables). *If $\mathbb{E}|X| < \infty$ then $\mathbb{E}X$ in (1.26) can be rewritten as*

$$\mathbb{E}X = \sum_{n\ge1} a_n \mathbb{P}(X = a_n) = \sum_{n\ge1} a_n \mathbb{P}'(a_n). \tag{1.30}$$

Proof. To show this, let us enumerate the points in the set $X^{-1}(a_n)$ in an arbitrary order,

$$X^{-1}(a_n) = \big\{\omega : X(\omega) = a_n\big\} = \big\{\omega_{nm} : 1 \le m \le M_n\big\}.$$

Some (or all) of these sets could be infinite, in which case $M_n = \infty$. Let us rewrite, using (1.23) and that $a_n = X(\omega_{nm})$ for all $m \geq 1$,

$$\begin{aligned}\sum_{n=1}^{\infty} a_n \mathbb{P}(X = a_n) &= \sum_{n=1}^{\infty} a_n \sum_{m=1}^{M_n} \mathbb{P}(\omega_{nm}) \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{M_n} a_n \mathbb{P}(\omega_{nm}) \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{M_n} X(\omega_{mn}) \mathbb{P}(\omega_{nm}).\end{aligned}$$

This looks very similar to the definition of the expectation (1.26), except that now the elements of Ω are enumerated by the double index (n,m) and we have the double sum. To finish the proof, it remains to apply the following lemma, which will be useful to us in other ways. □

Lemma 1.2. *For any bijection* $\pi \colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$,

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_{\pi(n,m)}$$

if either all $c_n \geq 0$ *or the series on either side is absolutely convergent.*

Here, for simplicity, we take the second sum over $1 \leq m < \infty$, but the same argument will work when $1 \leq m \leq M_n$.

Proof. Let us consider the case $c_n \geq 0$ first. In one direction, for any $K \geq 1$, there exists $N \geq 1$ such that

$$\sum_{k=1}^{K} c_k \leq \sum_{n=1}^{N} \sum_{m=1}^{N} c_{\pi(n,m)} \leq \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_{\pi(n,m)},$$

because π is a bijection and $1, \ldots, K$ must appear somewhere on the list of $\pi(n,m)$. Letting $K \to \infty$ proves that the left hand side is less than or equal to the right hand side. In the opposite direction, for any $N,M \geq 1$ there exists $K \geq 1$ such that

$$\sum_{n=1}^{N}\sum_{m=1}^{M} c_{\pi(n,m)} \leq \sum_{k=1}^{K} c_k \leq \sum_{k=1}^{\infty} c_k,$$

for example, by taking K to be the largest among $\pi(n,m)$ for $n \leq N, m \leq M$. Letting $M \to \infty$ first and then letting $N \to \infty$ proves the inequality in the other direction.

In the second case, suppose, for example, that $\sum_{n=1}^{\infty} c_n$ is absolutely convergent. Let us write each c_n as the difference $c_n = a_n - b_n$ where $a_n = |c_n|\,\mathrm{I}(c_n \geq 0)$ and $b_n = |c_n|\,\mathrm{I}(c_n \leq 0)$. By the first part of the proof, we know that

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_{\pi(n,m)}, \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} b_{\pi(n,m)}.$$

Since $\sum_{n=1}^{\infty}(a_n + b_n) = \sum_{n=1}^{\infty} |c_n| < \infty$, all the series above are convergent and, subtracting the two equations, we obtain the same equality for (c_n). □

The proof also shows that if $c_n \geq 0$ then one side is $+\infty$ only if the other side is $+\infty$.

In addition to showing that we can compute $\mathbb{E}X$ in terms of the distribution of X as in (1.30), this lemma immediately implies that, if the set Ω is countably infinite, we can extend finite additivity property (1.8) to *countable additivity*.

Exercise 1.2.4. If Ω is countably infinite and $A = \bigcup_{n\geq 1} A_n$ for disjoint $A_n \subseteq \Omega$ then

$$\mathbb{P}(A) = \sum_{n\geq 1} \mathbb{P}(A_n). \tag{1.31}$$

Moreover, if $\mathbb{E}|X| < \infty$ then

$$\mathbb{E}\big[X\,\mathrm{I}(\omega \in A)\big] = \sum_{n\geq 1} \mathbb{E}\big[X\,\mathrm{I}(\omega \in A_n)\big]. \tag{1.32}$$

One simple consequence of countable additivity is that the *tail probability* $\mathbb{P}(X \geq t)$ goes to zero as t goes to infinity,

$$\lim_{t\to+\infty} \mathbb{P}(X \geq t) = 0. \tag{1.33}$$

To see this, for integer $n \geq 0$, we can write the set $\{X \geq n\}$ as a disjoint union

$$\{X \geq n\} = \bigcup_{m \geq n} \{m \leq X < m+1\},$$

and, using countable additivity, we can write

$$\mathbb{P}(X \geq n) = \sum_{m=n}^{\infty} \mathbb{P}(m \leq X < m+1).$$

This means that $\mathbb{P}(X \geq n)$ is the tail of convergent series, so it must go to zero. For non-integer t, by the monotonicity property of probability, $\mathbb{P}(X \geq t) \leq \mathbb{P}(X \geq \lfloor t \rfloor)$, so the tail probability must go to zero.

It will sometimes be convenient to rewrite the expectation $\mathbb{E}X$ of a nonnegative random variable X in terms of its tail probability $\mathbb{P}(X \geq t)$. If X takes integer values $k \geq 0$ and if we write $k = \sum_{m=1}^{k} 1$ then

$$\begin{aligned}\mathbb{E}X &= \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \sum_{k=1}^{\infty} \sum_{m=1}^{k} \mathbb{P}(X = k) \\ &= \sum_{m=1}^{\infty} \sum_{k=m}^{\infty} \mathbb{P}(X = k) = \sum_{m=1}^{\infty} \mathbb{P}(X \geq m), \end{aligned} \tag{1.34}$$

where the interchange of the order of summation is justified by the previous lemma. For $X \geq 0$ that do not necessarily take only integer values, we have the following analogue.

Lemma 1.3. *For any nonnegative random variable $X \geq 0$,*

$$\mathbb{E}X = \int_0^{\infty} \mathbb{P}(X \geq t)\, dt. \tag{1.35}$$

In particular, one side is equal to $+\infty$ only if the other side is.

Proof. If we express $a_n = \int_0^\infty \mathrm{I}(t \le a_n)\,dt$ and plug into the formula (1.30), we can formally write

$$\begin{aligned}
\mathbb{E}X &= \sum_{n\ge 1} \int_0^\infty \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n)\,dt \\
&= \int_0^\infty \sum_{n\ge 1} \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n)\,dt \\
&= \int_0^\infty \sum_{a_n \ge t} \mathbb{P}(X = a_n)\,dt = \int_0^\infty \mathbb{P}(X \ge t)\,dt,
\end{aligned}$$

where in the last equality we used countable additivity of probability in (1.31). If you are familiar with the Lebesgue integration then the second step (interchanging the series and integral) is justified by the Monotone Convergence Theorem. If you are only familiar with the Riemann integral, we have to say a bit more. First, if we write the improper integral $\int_0^\infty$ as $\sum_{m\ge 1} \int_{m-1}^m$, we can use the previous lemma to interchange the order of summation,

$$\begin{aligned}
\mathbb{E}X &= \sum_{n\ge 1}\sum_{m\ge 1} \int_{m-1}^m \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n)\,dt \\
&= \sum_{m\ge 1}\sum_{n\ge 1} \int_{m-1}^m \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n)\,dt.
\end{aligned}$$

If we write the series over $n \ge 1$ as the limit of partial sums, we can continue to write

$$\begin{aligned}
\mathbb{E}X &= \sum_{m\ge 1} \lim_{N\to\infty} \int_{m-1}^m \sum_{n=1}^N \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n)\,dt \\
&= \sum_{m\ge 1} \lim_{N\to\infty} \int_{m-1}^m f_N(t)\,dt,
\end{aligned}$$

where we denoted

$$f_N(t) := \sum_{n=1}^N \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n).$$

By the countable additivity of probability in (1.31),

$$\lim_{N\to\infty} f_N(t) = \sum_{n=1}^{\infty} \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n) = \mathbb{P}(X \ge t).$$

Moreover, this convergence is uniform, because

$$\begin{aligned} \big|\mathbb{P}(X \ge t) - f_N(t)\big| &= \Big|\sum_{n>N} \mathrm{I}(t \le a_n)\mathbb{P}(X = a_n)\Big| \\ &\le \sum_{n>N} \mathbb{P}(X = a_n) \to 0 \qquad (1.36) \end{aligned}$$

as $N \to \infty$ as the tail of convergent series. This implies that

$$\lim_{N\to\infty} \int_{m-1}^{m} f_N(t)\,dt = \int_{m-1}^{m} \mathbb{P}(X \ge t)\,dt$$

and, therefore, we again showed that

$$\mathbb{E}X = \sum_{m\ge 1} \int_{m-1}^{m} \mathbb{P}(X \ge t)\,dt = \int_{0}^{\infty} \mathbb{P}(X \ge t)\,dt.$$

This finishes the proof. □

Exercise 1.2.5. For any nonnegative random variable $X \ge 0$ and $n > 0$, show that

$$\mathbb{E}X^n = \int_0^\infty n t^{n-1} \mathbb{P}(X \ge t)\,dt. \qquad (1.37)$$

Hint: use (1.35).

Exercise 1.2.6. Show that if, for some $c, \varepsilon > 0$ and $t_0 \ge 0$,

$$\mathbb{P}\big(|X| \ge t\big) \le \frac{c}{t^{2+\varepsilon}} \text{ for all } t \ge t_0,$$

then $\mathbb{E}X^2 < \infty$.

As we mentioned in the previous section (see equation (1.14)), random variables $X_1, \ldots, X_n$ that satisfy

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \tag{1.38}$$

for all possible values $x_1, \ldots, x_n$ that they can take are called *independent*. We will discuss independence and its meaning in great detail later in this chapter, in Section 1.4, and in this section simply use the definition (1.38). Next, let us give some examples of probability spaces and distributions.

Example 1.2.1 (Bernoulli distribution). A distribution on $\{0,1\}$ with the probabilities

$$\mathbb{P}(1) = p \text{ and } \mathbb{P}(0) = 1 - p$$

for some $p \in [0,1]$ is called the *Bernoulli distribution*, and is denoted $B(p)$. A random variable X on any probability space with this distribution is called a Bernoulli random variable. This means that

$$\mathbb{P}(X = 1) = p \text{ and } \mathbb{P}(X = 0) = 1 - p, \tag{1.39}$$

and is denoted by $X \sim B(p)$. This can also be written as

$$\mathbb{P}(X = x) = p^x (1-p)^{1-x} \text{ for } x \in \{0,1\}. \tag{1.40}$$

The expected value of the Bernoulli random variable X is

$$\mathbb{E}X = 1 \cdot p + 0 \cdot (1-p) = p. \tag{1.41}$$

Bernoulli distribution can be used to model a flip of a coin, or an outcome of a random experiment with two possible outcomes. The case $p = 1/2$ corresponds to a *fair coin* and $p \neq 1/2$ is a *biased coin*. □

Example 1.2.2 (Products of Bernoulli). Given

$$p_1, \ldots, p_n \in [0,1]$$

for some $n \geq 2$, the probability measure on the set of vectors $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ given by

$$\mathbb{P}(x_1,\ldots,x_n)=\prod_{i=1}^{n}p_i^{x_i}(1-p_i)^{1-x_i} \tag{1.42}$$

is called *product of Bernoulli* and is denoted by $\otimes_{i\le n}B(p_i)$. The probabilities add up to one because

$$\sum_{x\in\{0,1\}^n}\prod_{i=1}^{n}p_i^{x_i}(1-p_i)^{1-x_i}=\prod_{i=1}^{n}\sum_{x_i\in\{0,1\}}p_i^{x_i}(1-p_i)^{1-x_i}=1.$$

A random vector $X=(X_1,\ldots,X_n)$ consisting of n random variables on some probability space has this distribution if

$$\mathbb{P}(X=x)=\prod_{i=1}^{n}p_i^{x_i}(1-p_i)^{1-x_i} \tag{1.43}$$

for $x=(x_1,\ldots,x_n)\in\{0,1\}^n$. One way to construct such a random vector is to take the probability space $\Omega=\{0,1\}^n$ and simply let $X_i(x)=x_i$ be the i^{th} coordinate x_i of $x\in\Omega$. In other words, X can be defined as the identity function on its own *sample space* (see Remark 1.1 below). However, for the rest of this example, X can be defined on any probability space as long as it satisfies (1.43).

If we fix x_i and consider the set

$$A=\Big\{y\in\{0,1\}^n\,:\,y_i=x_i\Big\}$$

of all vectors with the ith coordinate fixed to x_i then, by the finite additivity of probability,

$$\begin{aligned}\mathbb{P}(X_i=x_i)&=\sum_{y\in A}\mathbb{P}(X=y)=\sum_{y\in A}\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}\\&=p_i^{x_i}(1-p_i)^{1-x_i}\prod_{j\neq i}\sum_{y_j\in\{0,1\}}p_j^{y_j}(1-p_j)^{1-y_j}\\&=p_i^{x_i}(1-p_i)^{1-x_i}.\end{aligned}$$

In other words, each X_i has Bernoulli distribution $B(p_i)$ and, therefore,

$$\mathbb{P}(X=x)=\prod_{i=1}^{n}\mathbb{P}(X_i=x_i). \tag{1.44}$$

If we recall the definition (1.38), we showed that a random vector $X=(X_1,\dots,X_n)$ with the distribution given by the product of Bernoulli (1.43) has independent coordinates $X_i \sim B(p_i)$. This is not surprising since the two definitions look almost identical, but we had to check that $\mathbb{P}(X_i=x_i)$ is the corresponding i^{th} factor in (1.43).

If all parameters $p_i=p\in[0,1]$ then $X_1,\dots,X_n$ also have the same distribution and are called *independent identically distributed*, or *i.i.d.* for short. In this case, we can rewrite

$$\mathbb{P}(X=x)=p^{\sum_{i\leq n}x_i}(1-p)^{n-\sum_{i\leq n}x_i}. \tag{1.45}$$

Instead of $\otimes_{i\leq n}B(p)$, we write $B(p)^{\otimes n}$. We will discuss in Section 1.4 why the definition in (1.38) corresponds to our intuitive notion of independence. □

Remark 1.1. Let us emphasize an important point. Random variables defined on *different* probability spaces can have the same distribution. For example, a Bernoulli random variable can be defined on its sample space $\{0,1\}$ or it can be defined as one of the coordinates X_i on the product space $\{0,1\}^n$ as in the previous example. In fact, a random variable defined on any space Ω and taking two values $\{0,1\}$ (i.e. an indicator of some set) is a Bernoulli random variable. Whenever we are interested to study properties of a particular distribution, we can work with a random variable with this distribution defined on any probability space. For example, when we derived the equation (1.44) from (1.43), the calculation depended only on the distribution of the vector X in (1.43) and not on the particular probability space on which X is defined. For this reason, it is quite common in Probability not to specify a particular probability space and use generic notation $\mathbb{P}$ and $\mathbb{E}$ for probability measures and expectations on any probability space.

On the other hand, it is often convenient to represent a distribution using a specific random variable, sometimes defined on a specific probability space. For example, in the example of the Binomial distribution below, we will see that it is much easier to compute its expectation using the linearity of expectation and representation via the sum of Bernoulli random variables. In the next section, we will see another important example when a specific probability space construction is also quite useful.

Example 1.2.3 (Erdős–Rényi random graph). Let us take $p \in [0,1]$ and consider a set

$$V = \{v_1, \ldots, v_n\}$$

of n elements, called *vertices*. For each pair of vertices v_i and v_j for $i \neq j$, we draw an *edge* between them with the probability p. We do this independently for different pairs of vertices, for example by flipping a coin for each edge. The random graph obtained in this way is denoted by $G(n,p)$. One can, for example, view this random graph as a model of a group of people, where any two people like each other with probability p and dislike each other with probability $1-p$.

Of course, mathematically this means that we consider i.i.d. Bernoulli random variables from the previous example,

$$(X_1, \ldots, X_m) \sim B(p)^{\otimes m} \text{ with } m = \binom{n}{2},$$

since $\binom{n}{2}$ is the number of distinct pairs of vertices, and we say that the edge number k is present if the corresponding $X_k = 1$. Although this model appears to be a special case of the previous example, the graph structure will allow us to consider various interesting functions of the edge indicators $X_1, \ldots, X_m$, for example, the number of triangles. □

Example 1.2.4 (Binomial distribution). Let us consider n i.i.d. Bernoulli random variables $X_1,\ldots,X_n \sim B(p)$. Then the distribution of their sum

$$S_n = X_1 + \ldots + X_n$$

is called a *binomial distribution*, denoted $B(n,p)$. The sum takes values in the set

$$\{0,1,2,\ldots,n\}.$$

What is the probability $\mathbb{P}(S_n = k)$? We can break the set $\{S_n = k\}$ into a disjoint union of sets $\{X = (X_1,\ldots,X_n) = x\}$ over all $x \in \{0,1\}^n$ with exactly k coordinates equal to 1 and sum their probabilities. For each such vector x, by (1.45),

$$\mathbb{P}(X = x) = p^k(1-p)^{n-k}.$$

Since there are

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

possible ways to choose which k coordinates are equal to 1,

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k(1-p)^{n-k}. \tag{1.46}$$

This distribution on $\{0,\ldots,n\}$ is called the binomial distribution $B(n,p)$. The name comes from the binomial formula,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

which for $a = p, b = 1-p$ implies that the probabilities in (1.46) add up to one. By linearity of expectation and (1.41),

$$\mathbb{E}S_n = \mathbb{E}X_1 + \ldots + \mathbb{E}X_n = np. \tag{1.47}$$

We can also compute the expectation using the distribution formula (1.46),

$$\mathbb{E}S_n = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}, \tag{1.48}$$

which we will leave as an exercise. Notice how much easier the computation using the linearity of expectation was. This will often be useful in situations when a random variable can be written as a sum of indicators of events with probabilities that can be computed more easily than the distribution of the sum. □

Exercise 1.2.7. Show that the right hand side of (1.48) is np.

Example 1.2.5 (Multinomial distribution). As an analogue of a coin flip, let us imagine rolling a die with $k \geq 2$ sides that have probabilities

$$p_1, \ldots, p_k > 0 \text{ such that } p_1 + \ldots + p_k = 1.$$

In other words, we consider any distribution on k outcomes. If we roll the die n times, or roll n dice, the analogue of the Binomial random variable is the number of times each side comes up, which is a vector that belongs to the set

$$\Omega = \Big\{ (n_1, \ldots, n_k) \, : \, n_j \geq 0 \text{ for } j \leq k, n_1 + \ldots + n_k = n \Big\}.$$

The probability measure on this set given by

$$\mathbb{P}(n_1, \ldots, n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k} \tag{1.49}$$

is called a *multinomial distribution*. The product $p_1^{n_1} \cdots p_k^{n_k}$ is the analogue of the product of Bernoulli, and the factor in front represents the number of different sequences of rolls that result in $n_1, \ldots, n_k$ outcomes of each type. This factor comes from the formula,

$$\frac{n!}{n_1!\cdots n_k!} = \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_k}{n_k},$$

where we choose n_1 dice with the first side up, then choose n_2 out of the remaining $n-n_1$ dice with the second side up, and so on. □

Example 1.2.6 (Geometric distribution). Let $0 < p \leq 1$ and let us imagine that we toss a $\{0,1\}$-valued coin until the first time 1 comes up. The number N of coin tosses can be any integer $k \geq 1$ and

$$\mathbb{P}(N = k) = (1-p)^{k-1}p, \tag{1.50}$$

by (1.45), because in terms of the i.i.d. Bernoulli random variables $X_1,\ldots,X_k$ the event $\{N = k\}$ can be expressed as

$$\{N = k\} = \{X_1 = 0,\ldots,X_{k-1} = 0, X_k = 1\}.$$

The sum of these probabilities is 1 by the geometric series formula.

Exercise 1.2.8. Show that the expectation of the geometric random variable (1.50) is

$$\mathbb{E}N = \frac{1}{p}. \tag{1.51}$$

We see that, for example, if a probability of 1 in a biased coin is $1/100$, on average we 'expect' to toss the coin 100 times until we see 1. □

Example 1.2.7 (Poisson distribution). Consider $\lambda > 0$. The distribution

$$\mathbb{P}(k) = \frac{\lambda^k}{k!}e^{-\lambda} \text{ for } k = 0,1,2,\ldots \tag{1.52}$$

on nonnegative integers is called a *Poisson distribution with mean* λ, and denoted by $\mathrm{Poiss}(\lambda)$. The probabilities add up to one by the Taylor series for exponential. The word ‘mean’ is just another name for ‘expectation’, so the parameter $\lambda > 0$ equals to the expectation of this distribution.

Exercise 1.2.9. Show that the mean of a Poisson random variable $X \sim \text{Poiss}(\lambda)$ is

$$\mathbb{E}X = \lambda. \tag{1.53}$$

The Poisson distribution arises as an approximation of the binomial $B(n,p)$ distribution in the regime when $p \approx \lambda/n$, as will be explained in the next section.

Let us prove a very important *stability property* of the Poisson distribution. Consider two random variables X_1 and X_2 defined on the same probability space Ω and suppose that X_1, X_2 are independent and have Poisson distributions with means λ_1 and $\lambda_2 > 0$:

$$\begin{aligned}\mathbb{P}(X_1 = n, X_2 = m) &= \mathbb{P}(X_1 = n)\mathbb{P}(X_2 = m) \\ &= \frac{\lambda_1^n}{n!}e^{-\lambda_1}\frac{\lambda_2^m}{m!}e^{-\lambda_2}.\end{aligned} \tag{1.54}$$

We can construct such pair X_1 and X_2 as the coordinates on $\Omega = \{0,1,\dots\} \times \{0,1,\dots\}$, in the same way we constructed the product of Bernoulli above. Then the following holds.

Lemma 1.4 (Stability of Poisson). *The sum $X = X_1 + X_2$ of two independent* Poiss(λ_1) *and* Poiss(λ_2) *random variables has* Poiss$(\lambda_1 + \lambda_2)$ *distribution.*

Proof. The sum can take only integer values $n \geq 0$, and it is equal to n, $X_1 + X_2 = n$, if and only if $X_1 = m$ and $X_2 = n - m$ for some $m = 0, 1, \dots, n$. Therefore,

$$\begin{aligned}
\mathbb{P}(X=n) &= \sum_{m=0}^{n} \mathbb{P}(X_1=m, X_2=n-m) \\
&= \sum_{m=0}^{n} \frac{\lambda_1^m}{m!} e^{-\lambda_1} \frac{\lambda_2^{n-m}}{(n-m)!} e^{-\lambda_2} \\
&= \frac{1}{n!} \Big(\sum_{m=0}^{n} \frac{n!}{m!(n-m)!} \lambda_1^m \lambda_2^{n-m} \Big) e^{-(\lambda_1+\lambda_2)} \\
&= \frac{(\lambda_1+\lambda_2)^n}{n!} e^{-(\lambda_1+\lambda_2)},
\end{aligned}$$

where the last equality is the binomial formula. □

One can show by induction that the following holds.

Exercise 1.2.10. Consider $k \geq 2$ independent Poisson random variables $X_1, \ldots, X_k$ with the parameters $\lambda_1, \ldots, \lambda_k > 0$,

$$\mathbb{P}(X_1 = n_1, \ldots, X_k = n_k) = \prod_{j=1}^{k} \frac{\lambda_j^{n_j}}{n_j!} e^{-\lambda_j}.$$

Show that their sum has $\text{Poiss}(\lambda_1 + \ldots + \lambda_k)$ distribution. *Hint:* To use the induction, show that if the random variables $X_1, \ldots, X_k$ are independent in the sense of (1.38) then $X_1 + X_2$, X_3, …, X_n are also independent.

Exercise 1.2.11. Show that if a random variable X satisfies $|X| \leq 10$ then $|\mathbb{E}X| \leq 10$.

Exercise 1.2.12. Find the event D such that

$$(D \cup A)^c \cup (D \cup A^c)^c = B.$$

Exercise 1.2.13. Prove that $\lim_{x \to +\infty} \mathbb{E}|X| \mathrm{I}(|X| \geq x) = 0$ if $\mathbb{E}|X| < \infty$. *Hint:* use (1.32).

Exercise 1.2.14. Three soccer players are trying out for a soccer team. They will get one chance to take a penalty kick to get a spot on the team. Each player has the probability $3/5$ of making the goal, independently of each other. What is the probability at least one of them will make the team?

Exercise 1.2.15. Suppose $\Omega = \{a,b,c,d\}$, $\mathbb{P}(\{a,b\}) = 0.6$, $\mathbb{P}(\{b,c\}) = 0.3$, and $\mathbb{P}(\{c,d\}) = 0.4$. Describe the set of all possible probabilities of a,b,c, and d.

Exercise 1.2.16. Four people play a game, where all of them simultaneously roll a six sided die. The winner is determined to be the first person to roll a one; then that person is eliminated from the game. They roll again if more than one person rolls a one. After that, the second place winner is determined to be the first person who rolls a one, then that person is eliminated from the game. The third place winner is determined to be the first person who rolls a one, then the game ends. What is the expected number of rolls until the game ends?

1.3 Poisson Approximation to the Binomial

Let us first notice how Poiss(λ) distribution arises as a limit of $B(n,p)$ when

$$p = \frac{\lambda}{n} \text{ and } n \to \infty. \tag{1.55}$$

If we rewrite the binomial probability of any fixed integer $k \geq 0$ with the choice of $p = \lambda/n$, we see that

$$\begin{aligned}
&\binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \Big(\frac{\lambda}{n}\Big)^k \Big(1 - \frac{\lambda}{n}\Big)^{n-k} \\
&= \frac{\lambda^k}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \cdot \Big(1 - \frac{\lambda}{n}\Big)^{n-k} \\
&\to \frac{\lambda^k}{k!} e^{-\lambda},
\end{aligned}$$

as $n \to \infty$, because k is fixed, k factors in the middle of the form $(n-\ell)/n$ converge to one, and the last factor converges to $e^{-\lambda}$. We will generalize this in two ways. First, we will prove an explicit error bound that works simultaneously for any set of outcomes rather than for one fixed k. Moreover, instead of the Binomial distribution, we will consider sums of independent Bernoulli $B(p_i)$ with possibly different p_i. As we alluded to in the previous section, our argument will be based on a special *coupling* construction on some specific probability space.

Given $\lambda > 0$ and a set $A \subseteq \{0,1,2,\ldots\}$, let

$$\text{Poiss}_\lambda(A) = \sum_{k \in A} \frac{\lambda^k}{k!} e^{-\lambda} \tag{1.56}$$

be the Poiss(λ) probability of the set A. Our main result will be the following.

Theorem 1.1. *Consider independent Bernoulli $X_i \sim B(p_i)$ random variables with $p_i \in (0,1)$ for $i \le n$ defined on the same probability space, and let*

$$S_n = X_1 + \ldots + X_n \text{ and } \lambda = p_1 + \ldots + p_n.$$

Then, for any subset $A \subseteq \{0,1,2,\ldots\}$,

$$\big|\mathbb{P}(S_n \in A) - \mathrm{Poiss}_\lambda(A)\big| \le \sum_{i=1}^n p_i^2. \tag{1.57}$$

When all $p_i = p \in (0,1)$, the sum S_n has binomial $B(n,p)$ distribution, while $\lambda = np$ and the equation (1.57) becomes

$$\big|\mathbb{P}(S_n \in A) - \mathrm{Poiss}_\lambda(A)\big| \le np^2 = \frac{\lambda^2}{n}. \tag{1.58}$$

If the right hand side is small, for example, if $n \to \infty$ and λ is fixed, then binomial $B(n,p)$ distribution is approximated by the Poisson Poiss(λ) distribution in the strong sense that the probabilities of all events are close.

Example 1.3.1. The chances of winning the jackpot in Lotto Max are $p = 1/28,633,528$. If $n = 20,000,000$ tickets are sold then the number of jackpot winners has the Binomial $B(n,p)$ distribution. (Of course, we ignore people's preferences for their birthdays, or many people using the same fortune cookie.) Then, if

$$\lambda = np = \frac{20,000,000}{28,633,528} \approx 0.69848,$$

we can approximate the probability that no one wins a jackpot by

$$\mathrm{Poiss}_\lambda(0) = e^{-\lambda} \approx 0.49734,$$

and the probability of exactly one winning ticket by

$$\mathrm{Poiss}_\lambda(1) = \lambda e^{-\lambda} \approx 0.34738.$$

The error of this approximation is bounded by

$$np^2 = \frac{\lambda^2}{n} \approx 2.44 \times 10^{-8}$$

for each outcome, but also for any collection of outcomes. For example, the same error is valid if we approximate the probability of at most one winning ticket by $e^{-\lambda} + \lambda e^{-\lambda} \approx 0.84472256$. □

Example 1.3.2 (Sparse Erdős–Rényi graph). Let $G(n,p)$ be the Erdős–Rényi random graph with $p = \frac{\lambda}{n}$ for a fixed λ and large n. This is called an example of a *sparse Erdős–Rényi random graph*. If N is the number of edges connected to a particular vertex, say v_1, then, since $N \sim B(n-1,p)$, its distribution can be approximated by $\text{Poiss}(\lambda)$. □

Proof (Theorem 1.1). First, we will construct (on the same probability space) independent Bernoulli $X_i \sim B(p_i)$ random variables and independent Poisson $Y_i \sim \text{Poiss}(p_i)$ random variables coupled in a special way. Later on we will use the stability property of Poisson,

$$S_n' = Y_1 + \ldots + Y_n \sim \text{Poiss}(\lambda), \tag{1.59}$$

proved in Lemma 1.4 and Exercise 1.2.10 at the end of last section. It is important to point out right away that $\mathbb{P}\big(S_n \in A\big)$ does not depend on the particular probability space on which the random variables $X_1, \ldots, X_n$ are defined and depends only on their distribution

$$\mathbb{P}(X_1 = x_n, \ldots, X_n = x_n) = \prod_{i \le n} p_i^{x_i}(1-p_i)^{1-x_i},$$

so the construction below on a particular probability space implies the result in full generality.

First, to explain what we have in mind, let us show how to construct a pair $X \sim B(p)$ and $Y \sim \text{Poiss}(p)$ on the same probability space in such a way that X and Y are close in some sense when p is small. Let us start with the usual sample space of the Poisson random variable $Y \sim \text{Poiss}(p)$, namely,

$$\Omega = \{0, 1, 2, \ldots\} \text{ with } \mathbb{P}(k) = \frac{p^k}{k!} e^{-p}.$$

Notice that $\mathbb{P}(0) = e^{-p}$ is bigger than the probability $1 - p$ that a Bernoulli random variable X is equal to 0,

$$1 - p < e^{-p} \text{ for } p \neq 0,$$

because $1 - x$ is the tangent line to e^{-x} at zero. This means that we can split e^{-p} into two positive numbers,

$$e^{-p} = (1 - p) + (e^{-p} - 1 + p),$$

one of them being the probability that $X = 0$. This suggests an idea that, in order to accommodate the Bernoulli random variable X on the same probability space, we can split the outcome 0 into two outcomes, say -1 and 0, and enlarge the probability space,

$$\Omega_+ = \{-1, 0, 1, 2, \ldots\}. \tag{1.60}$$

We assign new probabilities to -1 and 0 and denote them

$$\mathbb{P}_p(-1) = 1 - p,\ \mathbb{P}_p(0) = e^{-p} - 1 + p. \tag{1.61}$$

All the other probabilities will be untouched and denoted

$$\mathbb{P}_p(k) = \frac{p^k}{k!} e^{-p} \text{ for } k \geq 1. \tag{1.62}$$

The way the probability space $(\Omega_+, \mathbb{P}_p)$ was constructed makes it obvious what will be X and what will be Y. For $\omega \in \Omega_+$, we define a Bernoulli random variable

$$X = X(\omega) = \begin{cases} 0, \text{ if } \omega = -1, \\ 1, \text{ if } \omega \geq 0, \end{cases} \tag{1.63}$$

and we define a Poisson random variable

$$Y = Y(\omega) = \begin{cases} 0, \text{ if } \omega = -1 \text{ or } 0, \\ \omega, \text{ if } \omega \geq 1. \end{cases} \tag{1.64}$$

What do we accomplish by this construction? Notice that $X(\omega) = Y(\omega)$ if and only if $\omega = -1$ or 1. Therefore,

$$\begin{aligned} \mathbb{P}\big(X = Y\big) &= 1 - p + pe^{-p} \\ &\geq 1 - p + p(1-p) = 1 - p^2, \end{aligned}$$

where we again used that $1 - p \leq e^{-p}$. This implies that

$$\mathbb{P}(X \neq Y) \leq p^2 \tag{1.65}$$

and, when p is small, the square makes it of an even smaller order, so X and Y are equal with high probability.

Armed with this coupling construction for one pair (X, Y), how do we construct independent $X_i \sim B(p_i)$ for $i \leq n$ and independent $Y_i \sim \text{Poiss}(p_i)$ on the same probability space, so that each pair is coupled as we just described? Of course, we will use the product space construction that we have already seen for products of Bernoulli (i.e. defining random variables in terms of coordinates on the product space), which will be generalized in the next section. We will take

$$\Omega = \big(\Omega_+\big)^n, \tag{1.66}$$

denote its elements by $\omega = (\omega_1, \ldots, \omega_n)$, and let

$$\mathbb{P}(\omega) = \prod_{i=1}^n \mathbb{P}_{p_i}(\omega_i). \tag{1.67}$$

We will then define X_i and Y_i in terms of the coordinate ω_i,

$$X_i(\omega) = X(\omega_i) \text{ and } Y_i(\omega) = Y(\omega_i), \tag{1.68}$$

where X and Y were defined in (1.63) and (1.64). One can check that: (a) $X_i \sim B(p_i)$, and (b) they are independent over $i \le n$, by using the same calculation as in the Example 1.2.2 about the products of Bernoulli. For example, to calculate

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

for any $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, we can sum $\mathbb{P}(\omega)$ over all $\omega \in \{X_1 = x_1, \ldots, X_n = x_n\}$. If we denote

$$A(0) := \{-1\},\ A(1) := \{0,1,2,\ldots\},$$

then the definition (1.63) means that

$$X_i = x_i \Longleftrightarrow \omega_i \in A(x_i),$$

so the sum should be taken over $\omega = (\omega_1, \ldots, \omega_n)$ in

$$A(x) := A(x_1) \times \cdots \times A(x_n).$$

Since the probabilities are positive, Lemma 1.2 allows us to sum over one coordinate ω_i at a time and, because $\mathbb{P}(\omega)$ is given by the product (1.67), we will get

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) &= \sum_{\omega \in A(x)} \mathbb{P}(\omega) \\
= \sum_{\omega \in A(x)} \prod_{i=1}^{n} \mathbb{P}_{p_i}(\omega_i) &= \prod_{i=1}^{n} \sum_{\omega_i \in A(x_i)} \mathbb{P}_{p_i}(\omega_i) \\
= \prod_{i=1}^{n} \mathbb{P}_{p_i}\big(A(x_i)\big) &= \prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i},
\end{aligned}$$

where the last equality follows from the definition of $\mathbb{P}_{p_i}$. This distribution appeared in Example 1.2.2, which means that the random variables $X_1, \ldots, X_n$ are independent and each X_i has Binomial distribution $B(p_i)$.

A similar calculation for $Y_1, \ldots, Y_n$ also shows that each $Y_i \sim \text{Poiss}(p_i)$ and they are independent (check!).

If the sum S_n of Bernoulli random variables and the sum S'_n of Poisson random variables are not equal then at least one of the summands X_i is not equal to Y_i, which means that

$$\big\{S_n \neq S'_n\big\} \subseteq \bigcup_{i\leq n} \big\{X_i \neq Y_i\big\}.$$

By the union bound, we get that

$$\mathbb{P}\big(S_n \neq S'_n\big) \leq \sum_{i=1}^{n} \mathbb{P}\big(X_i \neq Y_i\big).$$

By construction and (1.65),

$$\mathbb{P}\big(X_i \neq Y_i\big) \leq p_i^2 \tag{1.69}$$

(exercise below asks to fill in the details) and, therefore,

$$\mathbb{P}(S_n \neq S'_n) \leq \sum_{i=1}^{n} p_i^2.$$

Finally, using that (we leave it as an exercise below)

$$\big|\mathbb{P}\big(S_n \in A\big) - \mathbb{P}\big(S'_n \in A\big)\big| \leq \mathbb{P}\big(S_n \neq S'_n\big)$$

and using (1.59) finishes the proof. □

The Poisson distribution is often a good model for the number of occurrences of certain events, when there is a large number of opportunities and a small probability for an event to occur at a given moment, such as: a number of shark attacks in a given year, number of wrong number phone calls, number of goals scored by a hockey player in a game, etc. For example, in a given month, there are many phone calls made to numbers relatively close to a given phone number and there is always a small chance of misdial, so the total will be a sum of many Bernoulli $B(p_i)$ with small p_i. We might not know the numbers n and p_i, but we may be able to estimate the expected number $\lambda = \sum_{i\leq n} p_i$ from previous experience and, as a result, estimate the probabilities using the Poiss(λ) distribution.

For example, in his 9 seasons with Edmonton Oilers, Wayne Gretzky had $1,669$ points in 696 games at a rate of $\lambda = 2.397988$ per game. The table below shows the number of games broken down by the actual number of points, as well as the corresponding Poisson approximation

$$696 \times \frac{\lambda^k}{k!} e^{-\lambda}.$$

The agreement is quite remarkable.

Points	0	1	2	3	4	5	6	7	8	9
Games	69	155	171	143	79	57	14	6	2	0
Poisson	63.27	151.71	181.9	145.4	87.17	41.81	16.71	5.72	1.72	0.46

Exercise 1.3.1. Prove that, for any two variables X, Y on the same (discrete) probability space,

$$\big|\mathbb{P}\big(X \in A\big) - \mathbb{P}\big(Y \in A\big)\big| \le \mathbb{P}\big(X \ne Y\big)$$

for any subset $A \subseteq \mathbb{R}$.

Exercise 1.3.2. Fill in the details in the proof of (1.69) above. *Hint:* if you are not sure what to do, take a look at Example 1.4.1 in the next section.

Exercise 1.3.3. An emperor orders preparations for a massive banquet for all of his 250 nobles. The Chancellor prepares enough seats for 245 nobles knowing that the probability a noble will not come is 0.05. Use the Poisson approximation to compute the probability that the Chancellor will get to keep his head? (Or what is the probability that there will be enough seats?) Estimate the error of approximation.

Exercise 1.3.4. Although many large and ferocious creatures wander around the African savannah, none are quite as imposing as the Hippopotamus. In fact, many smaller creatures are often trampled by this massive, lumbering neighbour. On average, three black mambas are trampled by Hippopotami per year. What is the probability that no black mambas will be squashed this year?

Exercise 1.3.5. The Kicker for the Dallas Cowboys scores an average of 2 Field goals per game. Over this players 150 game career, what is the probability that in at least one game he scored exactly 6 Field goals?

Exercise 1.3.6. When you bet on black in Roulette, your chances are $18/38$. Suppose also that bets over \$250 are not allowed by the casino. You decide to play the following strategy: you start with a \$1 bet and double the bet until either you win (the same amount as the bet) or the bet exceeds \$250; then you start again with a \$1 bet and repeat. We will call this sequence of bets in the strategy until restart with a \$1 bet 'one round'. If you play 1000 rounds of this strategy, what is the probability that your total winnings/losses are ≥ 0? Compute the exact formula and then compare it with Poisson approximation.

1.4 Independence, Conditional Distributions

Two events $A,B \subseteq \Omega$ on a (discrete) probability space $(\Omega,\mathbb{P})$ are called *independent* if

$$\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{1.70}$$

This definition is motivated by the following related concept. If $\mathbb{P}(B) > 0$ then the *conditional probability of* A *given* B is defined as

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}. \tag{1.71}$$

It represents the proportion (in the sense of probability) of the event A inside B. From the two definitions it is clear that if $\mathbb{P}(B) > 0$ then A and B are independent if and only if

$$\mathbb{P}(A \mid B) := \mathbb{P}(A). \tag{1.72}$$

In other words, if we know that an outcome ω is in the set B, this information does not alter the chances that ω is in A. This justifies the term 'independent events'. We use the more symmetric definition (1.70), because it makes sense even if one or both events have zero probability.

Exercise 1.4.1. If A and B are independent, show that A^c and B are also independent. *Hint:* write $B = (B\cap A)\cup(B\cap A^c)$ and use that $B\cap A$ and $B\cap A^c$ are disjoint.

Given $n \geq 2$, events $A_i \subseteq \Omega$ for $i \leq n$ on $(\Omega,\mathbb{P})$ are called *independent* if, for any subset of indices $I \subseteq \{1,\ldots,n\}$,

$$\mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) = \prod_{i\in I}\mathbb{P}(A_i). \tag{1.73}$$

As in the above exercise, this implies that

$$\mathbb{P}\Big(\bigcap_{i\in I} A_i^*\Big) = \prod_{i\in I}\mathbb{P}(A_i^*), \tag{1.74}$$

where each A_i^* is either A_i or A_i^c, which can be proved by induction on the number of $* = c$. For example, if we know (1.74) for all intersections with at most one $* = c$ then

$$\begin{aligned}\mathbb{P}\big(A_1^c \cap A_2 \cap A_3^c\big) &= \mathbb{P}\big(A_1^c \cap A_2\big) - \mathbb{P}\big(A_1^c \cap A_2 \cap A_3\big)\\ &= \mathbb{P}\big(A_1^c\big)\mathbb{P}\big(A_2\big) - \mathbb{P}\big(A_1^c\big)\mathbb{P}\big(A_2\big)\mathbb{P}\big(A_3\big)\\ &= \mathbb{P}\big(A_1^c\big)\mathbb{P}\big(A_2\big)\big(1 - \mathbb{P}\big(A_3\big)\big)\\ &= \mathbb{P}\big(A_1^c\big)\mathbb{P}\big(A_2\big)\mathbb{P}\big(A_3^c\big).\end{aligned}$$

The general case is similar. In its turn, (1.74) implies that

$$\mathbb{P}(A_i \mid B) = \mathbb{P}(A_i) \tag{1.75}$$

for any event B given by an intersection of A_j^* over any subset of indices $j \neq i$. This means that any information that an outcome ω belongs to or does not belong to some sets A_j does not affect the probability that ω belongs to another set A_i. This definition is stronger than *pairwise independence*, which is requiring any pair of sets (A_i, A_j) to be independent.

Exercise 1.4.2. Consider a regular tetrahedron die painted blue, red and green on three sides and painted in all three colours on the fourth side. If the die is equally likely to land on any side, show that the appearances of these colours on the side it lands on are pairwise-independent but not independent.

Random variables $X_1, \ldots, X_n$ defined on $(\Omega, \mathbb{P})$ are called *independent* if

$$\mathbb{P}(X_1 = a_1, \ldots, X_n = a_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = a_i) \tag{1.76}$$

for all possible values $a_1, \ldots, a_n$ that these random variables can take. This definition is equivalent to asking that, for arbitrary subsets of outcomes A_i,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i), \qquad (1.77)$$

which follows from (1.76) by summing over outcomes a_i of X_i in the set A_i. If we take the set $A_i = \mathbb{R}$ then the event $\{X_i \in \mathbb{R}\} = \Omega$ has probability one and can be omitted on both sides of the equation. This means that the definition of independence of n random variables $X_1, \ldots, X_n$ automatically includes independence for any subfamily of these random variables, for example, X_1, X_3, X_4.

In the proof of Theorem 1.1 in the previous section, we constructed independent random variables with Bernoulli and Poisson distributions. The same construction can be used to define independent random variables $X_1, \ldots, X_n$ on the same probability space Ω with arbitrary distributions.

Example 1.4.1 (Product space construction). Let $(\Omega_i, \mathbb{P}_i)$ for $i \leq n$ be arbitrary discrete probability spaces. Then the space

$$\Omega = \Omega_1 \times \cdots \times \Omega_n \qquad (1.78)$$

with the probability measure $\mathbb{P}$ on it given by

$$\mathbb{P}(\omega) = \mathbb{P}(\omega_1, \ldots, \omega_n) = \prod_{i=1}^{n} \mathbb{P}_i(\omega_i) \qquad (1.79)$$

is called the *product* of the above probability spaces. Here, we denoted the elements of Ω by $\omega = (\omega_1, \ldots, \omega_n)$. If we consider any subsets $A_i \subseteq \Omega_i$ and consider the rectangle

$$A = A_1 \times \cdots \times A_n$$

in the product space Ω then its probability equals

$$\begin{aligned}\mathbb{P}(A) &= \sum_{\omega\in A}\mathbb{P}(\omega) = \sum_{\omega\in A}\prod_{i=1}^{n}\mathbb{P}_i(\omega_i)\\ &= \prod_{i=1}^{n}\sum_{\omega_i\in A_i}\mathbb{P}_i(\omega_i) = \prod_{i=1}^{n}\mathbb{P}_i\big(A_i\big),\end{aligned} \tag{1.80}$$

which is the product of probabilities $\mathbb{P}_i(A_i)$ of its sides.

One consequence of this is that any random variables $X_1,\ldots,X_n\colon\Omega\to\mathbb{R}$ on this product space such that each X_i depends only on the coordinate ω_i,

$$X_i(\omega) = f_i(\omega_i)\ \text{ for some }\ f_i\colon\Omega_i\to\mathbb{R},$$

are independent, because the event $\{X_1 = x_1,\ldots,X_n = x_n\}$ is a rectangle with sides $\{f_i = x_i\}$, so

$$\mathbb{P}(X_1 = x_1,\ldots,X_n = x_n) = \prod_{i=1}^{n}\mathbb{P}_i\big(f_i = x_i\big). \tag{1.81}$$

On the other hand, the event

$$\big\{X_i = x_i\big\} = \big\{X_i = x_i \text{ and } X_j\in\mathbb{R} \text{ for } j\neq i\big\}$$

is a rectangle with the side $\{f_i = x_i\}$ on the i^{th} coordinate and sides Ω_j on the coordinates $j\neq i$, because no constraints are imposed on these ω_j. Since $\mathbb{P}_j(\Omega_j) = 1$, by (1.80),

$$\mathbb{P}(X_i = x_i) = \mathbb{P}_i\big(f_i = x_i\big). \tag{1.82}$$

Together, the equations (1.81) and (1.82) show that

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

which proves the claim that $X_1,\ldots,X_n$ are independent. If we want these random variables to have specific prescribed distributions then we can choose each space $(\Omega_i,\mathbb{P}_i)$ to be the sample space of X_i with $X_i(\omega_i) = \omega_i$. □

We have already used independence in important ways in the previous sections. Let us now discuss various additional key consequences of the definition.

Let us partition random variables $X_1,\ldots,X_n$ into m disjoint groups,

$$\{1,\ldots,n\} = \bigcup_{k=1}^{m} I_k,$$

and consider random variables Y_k for $k \le m$ given by some functions

$$Y_k = f_k\big((X_i)_{i\in I_k}\big) \tag{1.83}$$

of random variables X_i that belong to the group k. Then the following intuitive statement holds. A special example of this already appeared in Exercise 1.2.10.

Lemma 1.5 (Grouping Lemma). *If the random variables $X_1,\ldots,X_n$ are independent then $Y_1,\ldots,Y_m$ defined in (1.83) are also independent.*

Proof. Let us consider the event $\{Y_1 \in A_1,\ldots,Y_m \in A_m\}$ for some sets $A_k \subseteq \mathbb{R}$. Let

$$B_k = f_k^{-1}(A_k) = \Big\{(x_i)_{i\in I_k} : f_k\big((x_i)_{i\in I_k}\big) \in A_k\Big\}$$

be the set of all vectors that are mapped by f_k into A_k. If we use the notation $x^k = (x_i)_{i\in I_k}$ and $X^k = (X_i)_{i\in I_k}$ for the coordinates inside the k^{th} group then

$$\big\{Y_k \in A_k\big\} = \big\{f_k(X^k) \in A_k\big\} = \big\{X^k \in B_k\big\}.$$

If we denote $B = B_1 \times \cdots \times B_m$ then

$$\mathbb{P}\big(Y_1 \in A_1,\ldots,Y_m \in A_m\big) = \sum_{x\in B} \mathbb{P}\big(X^1 = x^1,\ldots,X^m = x^m\big).$$

By independence, each term

$$\mathbb{P}\big(X^1 = x^1,\ldots,X^m = x^m\big) = \mathbb{P}\big(X^1 = x^1\big)\cdots\mathbb{P}\big(X^m = x^m\big)$$

(why?), so the above sum can be rewritten as

$$\begin{aligned}
&\sum_{x\in B} \mathbb{P}\big(X^1 = x^1\big)\cdots\mathbb{P}\big(X^m = x^m\big) \\
&= \sum_{x^1\in B_1} \mathbb{P}\big(X^1 = x^1\big)\cdots \sum_{x^m\in B_m} \mathbb{P}\big(X^m = x^m\big) \\
&= \mathbb{P}\big(X^1 \in B_1\big)\cdots\mathbb{P}\big(X^m \in B_m\big) \\
&= \mathbb{P}\big(Y_1 \in A_1\big)\cdots\mathbb{P}\big(Y_m \in A_m\big).
\end{aligned}$$

This proves that

$$\mathbb{P}\big(Y_1 \in A_1,\ldots,Y_m \in A_m\big) = \mathbb{P}\big(Y_1 \in A_1\big)\cdots\mathbb{P}\big(Y_m \in A_m\big),$$

so $Y_1,\ldots,Y_m$ are independent. □

Example 1.4.2 (Stability of Poisson, revisited). Recall the stability property of the Poisson distribution at the end of Section 1.2. From the results of Section 1.3, we now know that the Poisson distribution is a certain limit of the Binomial distribution. Let us take some $\lambda_1,\lambda_2 > 0$ and, for simplicity, suppose that their ratio is rational,

$$\frac{\lambda_1}{\lambda_2} = \frac{r_1}{r_2}$$

for some integers $r_1,r_2 \geq 1$. Take large $n \geq 1$ and set

$$p := \frac{\lambda_1}{nr_1} = \frac{\lambda_2}{nr_2} \in (0,1).$$

By the results in the previous section, we know that if

$$m_1 = nr_1, m_2 = nr_2, m = m_1 + m_2,$$

and $X_1,\ldots,X_m \sim B(p)$ are i.i.d. Bernoulli then

$$\begin{aligned}
Y &= X_1 + \ldots + X_{m_1} \sim B(m_1,p) \approx \text{Poiss}(\lambda_1), \\
Z &= X_{m_1+1} + \ldots + X_m \sim B(m_2,p) \approx \text{Poiss}(\lambda_2),
\end{aligned}$$

while their sum

$$Y+Z \sim B(m,p) \approx \mathrm{Poiss}(mp) = \mathrm{Poiss}(\lambda_1+\lambda_2).$$

By the Grouping lemma, Y and Z are independent, so this approximation of the Poisson by independent flips of a coin gives an intuitive explanation for the stability property of Poisson. □

Next, we will show the following.

Lemma 1.6. *If random variables X and Y are independent and $\mathbb{E}|X| < \infty, \mathbb{E}|Y| < \infty$ then $\mathbb{E}|XY| < \infty$ and*

$$\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y. \tag{1.84}$$

Proof. To prove this, we will essentially apply Lemma 1.2 that allows us to rewrite

$$\mathbb{E}XY = \sum_{\omega\in\Omega} X(\omega)Y(\omega)\mathbb{P}(\omega)$$

in an arbitrary order, or as a double summation. For possible values a_n and b_m of the random variables X and Y, let us consider the event

$$\Omega_{nm} = \big\{\omega : X(\omega) = a_n, Y(\omega) = b_m\big\}$$

and write, using Lemma 1.2 and independence of X and Y,

$$\begin{aligned}\sum_{\omega\in\Omega} X(\omega)Y(\omega)\mathbb{P}(\omega) &= \sum_{n,m}\sum_{\omega\in\Omega_{nm}} a_n b_m \mathbb{P}(\omega) \\ &= \sum_{n,m} a_n b_m \mathbb{P}(X=a_n, Y=b_m) \\ &= \sum_{n,m} a_n b_m \mathbb{P}(X=a_n)\mathbb{P}(Y=b_m) \\ &= \sum_{n} a_n \mathbb{P}(X=a_n)\sum_{m} b_m \mathbb{P}(Y=b_m) \\ &= \mathbb{E}X\mathbb{E}Y.\end{aligned} \tag{1.85}$$

We can apply Lemma 1.2, because all the series on the right hand side are absolutely convergent by our assumption that $\mathbb{E}|X| < \infty, \mathbb{E}|Y| < \infty$. This finishes the proof. □

Of course, this lemma can be extended by induction, using grouping lemma, to show that

$$\mathbb{E}\prod_{i\le n} X_i = \prod_{i\le n} \mathbb{E}X_i, \tag{1.86}$$

when $X_1,\ldots,X_n$ are independent and all $\mathbb{E}|X_i| < \infty$.

Exercise 1.4.3. If $X_1,\ldots,X_n$ are i.i.d. Bernoulli $B(p)$, compute $\mathbb{E}(X_1+\ldots+X_n)^2$.

We can repeat the calculation (1.85) for any function $f(X,Y)$ instead of the product XY to obtain the following.

Theorem 1.2 (Fubini's Theorem). *If the random variables X and Y are independent and $\mathbb{E}|f(X,Y)| < \infty$ then*

$$\mathbb{E}f(X,Y) = \sum_n \Big[\sum_m f(a_n,b_m)\mathbb{P}(Y=b_m)\Big]\mathbb{P}(X=a_n), \tag{1.87}$$

where the sum is over possible values a_n and b_m of X and Y.

In other words, we can first fix $X = a_n$ and average over the distribution of Y and then average over a_n with respect to the distribution of X. This is an analogue of the formula

$$\iint_{[0,1]^2} f(x,y)\,dxdy = \int_0^1 \Big[\int_0^1 f(x,y)\,dy\Big]dx$$

in Calculus.

Remark 1.2. In the calculation (1.85), the fact that X and Y are random variables and not, for example, random vectors consisting of several coordinates each was never used. This means that we can apply it to random vectors, as long as they are independent of each other. Fubini's theorem will sometimes be used in this way.

One can also rewrite (1.87) by expressing the averages over the distributions of X and Y in terms of averages over the points ω on our probability space Ω. We need two copies of ω because we have two averages. We leave this simple observation as an exercise.

Exercise 1.4.4. Show that if the random variables X and Y are independent and $\mathbb{E}|f(X,Y)| < \infty$ then

$$\mathbb{E}f(X,Y) = \sum_{\omega \in \Omega} \Big[\sum_{\omega' \in \Omega} f\big(X(\omega), Y(\omega')\big) \mathbb{P}(\omega') \Big] \mathbb{P}(\omega).$$

What is the analogue of Fubini's theorem when X and Y are not independent? As in the above calculation (1.85), we can still write

$$\begin{aligned} \mathbb{E}f(X,Y) &= \sum_{\omega \in \Omega} f\big(X(\omega), Y(\omega)\big) \mathbb{P}(\omega) \\ &= \sum_{n,m} f(a_n, b_m) \mathbb{P}(X = a_n, Y = b_m), \end{aligned}$$

but we have to stop here if we do not have independence. However, if we multiply and divide by $\mathbb{P}(X = a_n)$, we can write this as

$$\sum_n \Big[\sum_m f(a_n, b_m) \frac{\mathbb{P}(X = a_n, Y = b_m)}{\mathbb{P}(X = a_n)} \Big] \mathbb{P}(X = a_n),$$

which looks quite similar to Fubini's theorem. Recall that, by (1.71),

$$\frac{\mathbb{P}(X = a_n, Y = b_m)}{\mathbb{P}(X = a_n)} = \mathbb{P}\big(Y = b_m \mid X = a_n\big) \tag{1.88}$$

is the conditional probability of the event $Y = b_m$ given that $X = a_n$. This gives us a generalization of the Fubini theorem for non-independent random variables.

Theorem 1.3 (Fubini's Theorem II). *If $\mathbb{E}|f(X,Y)| < \infty$ then the expectation of $f(X,Y)$ can be written as*

$$\mathbb{E}f(X,Y) \qquad (1.89)$$
$$= \sum_n \Big[\sum_m f(a_n,b_m)\mathbb{P}\big(Y = b_m \mid X = a_n\big) \Big] \mathbb{P}(X = a_n),$$

where the sum is over possible values a_n, b_m of X and Y.

For any fixed a_n, the conditional probability

$$\mathbb{P}\big(Y = b_m \mid X = a_n\big) = \frac{\mathbb{P}(X = a_n, Y = b_m)}{\mathbb{P}(X = a_n)} \qquad (1.90)$$

in (1.88) viewed as a function on the set of values $\{b_m\}$ is called the *conditional distribution of Y given $X = a_n$*. It is a probability measure because the sum over all values b_m is obviously equal to one.

The expectation of a function $g(Y)$ of Y with respect to this conditional distribution, denoted

$$\mathbb{E}\big[g(Y) \,|\, X = a_n\big] := \sum_m g(b_m)\mathbb{P}\big(Y = b_m \mid X = a_n\big),$$

is called the *conditional expectation* of $g(Y)$ given $X = a_n$. Notice that the sum over m inside the bracket in (1.89) is a conditional expectation

$$\mathbb{E}\big[f(a_n,Y) \,|\, X = a_n\big] := \sum_m f(a_n,b_m)\mathbb{P}\big(Y = b_m \mid X = a_n\big)$$

of $f(a_n,Y)$ given $X = a_n$. This conditional expectation can also be denoted by $\mathbb{E}\big[f(X,Y) \,|\, X = a_n\big]$, because X is fixed to be a_n.

In this terminology, we can describe the Fubini formula (1.89) as a two-step process. We first fix the value of $X = a_n$ and average $f(a_n,Y)$ over possible values of Y with respect to the conditional distribution (1.90). This average (called conditional expectation) depends on a_n, so it can be viewed as some function $h(a_n)$, and we average it with respect to the distribution of X,

$$\mathbb{E}f(X,Y) = \sum_n h(a_n)\mathbb{P}(X = a_n) = \mathbb{E}h(X).$$

Quite often, by nature of the problem, the distribution $\mathbb{P}(X = a_n)$ and the conditional distribution (1.90) are known or defined first, and then the *joint distribution* of (X,Y) is computed as

$$\mathbb{P}(X = a_n, Y = b_m) = \mathbb{P}\big(Y = b_m \,|\, X = a_n\big)\mathbb{P}(X = a_n). \quad (1.91)$$

In this case, the representation of $\mathbb{E}f(X,Y)$ in (1.89) is very natural, because, for a fixed value of $X = a_n$, we average first with respect to known conditional distribution and then average with respect to the distribution of X, which is called the *marginal* distribution of X. This type of construction of the pair (X,Y) is sometimes called a *two-stage experiment*. Let us consider an example.

Example 1.4.3. Suppose we want to predict the number of shark attacks in Florida over the next summer. This depends on the number of people going swimming, which depends on the weather. When the weather is typical, the average number of shark attacks is 10; when the weather is colder than normal, the average number of shark attacks is 8; when the weather is warmer than normal, the average number of shark attacks is 12. The forecast predicts warmer than usual summer with probability 60%, typical weather with probability 30%, and colder weather with probability 10%.

First, let us write down a mathematical model for this problem. We have two sources of randomness. The first one is the weather, which determines the average number Λ of shark attacks and, according to the forecast,

$$\mathbb{P}(\Lambda = 8) = 0.1,\ \mathbb{P}(\Lambda = 10) = 0.3,\ \mathbb{P}(\Lambda = 12) = 0.6.$$

The second random variable is the number N of shark attacks, which can be modelled by the Poisson distribution $\mathrm{Poiss}(\lambda)$ once we know the mean λ. Since the average Λ is random itself, what we are modelling here is precisely the conditional distribution

$$\mathbb{P}\big(N = k \mid \Lambda = \lambda\big) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

By (1.91), the joint distribution of (Λ, N) equals

$$\mathbb{P}(N = k, \Lambda = \lambda) = \begin{cases} 0.1 \cdot \frac{8^k}{k!} e^{-8}, & \text{if } \lambda = 8, \\ 0.3 \cdot \frac{10^k}{k!} e^{-10}, & \text{if } \lambda = 10, \\ 0.6 \cdot \frac{12^k}{k!} e^{-12}, & \text{if } \lambda = 12. \end{cases}$$

We can use Fubini's Theorem 1.3 to compute expectation of any (integrable) function of (Λ, N) and, in particular,

$$\mathbb{E}N = 0.1 \cdot 8 + 0.3 \cdot 10 + 0.6 \cdot 12 = 9.8,$$

because the conditional distribution of N for a fixed $\Lambda = \lambda$ is $\mathrm{Poiss}(\lambda)$, so its conditional expectation equals to λ. We can similarly compute the probabilities of events, for example,

$$\begin{aligned}\mathbb{P}(N \geq 10) = 0.1 \sum_{k=10}^{\infty} \frac{8^k}{k!} e^{-8} + 0.3 \sum_{k=10}^{\infty} \frac{10^k}{k!} e^{-10} \\ + 0.6 \sum_{k=10}^{\infty} \frac{12^k}{k!} e^{-12} \approx 0.645,\end{aligned}$$

because, it terms of (Λ, N) the event $N \geq 10$ can be written as $(\Lambda, N) \in \{8, 10, 12\} \times \{10, 11, \ldots\}$. □

Remark 1.3. Notice how in this example we talked about Λ and N as random variables without defining a probability space Ω on which they are defined and specifying how $\Lambda = \Lambda(\omega)$ and $N = N(\omega)$ are defined as functions on this space. This is because we are only interested in their distributions or, in other words, in probabilities of values that these random variables model, so the precise probability space is not important. In the situations like this, we can always think of the sample space as our probability space and random variables as the coordinates on this space. For example, when we consider $\mathbb{R}^2$ and denote its elements by (x,y), we can think of x as the first coordinate on $\mathbb{R}^2$ but also as a function $f(x,y)=x$. In our example, the sample space was

$$\{8,10,12\} \times \{0,1,2,\ldots\}$$

and Λ and N can be viewed as the coordinates on this space.

On the other hand, the advantage of not mentioning the precise probability space on which these random variables are defined is that it does not really matter. For example, if Λ was a function of not only the weather forecast but also economic forecast (think of beach vacations) and came out as an output of some complicated model on another probability space, for the calculations regarding the shark attacks only the distribution on the values $\{8,10,12\}$ would matter. □

Example 1.4.4 (Colouring property of Poisson). Let us consider one more two-stage experiment. Let N be a $\text{Poiss}(\lambda)$ random variable. Given $N=n$, we flip n coins with Bernoulli $B(p)$ distribution and we let N_1 and N_2 be the number of coins taking values 1 and 0 correspondingly. Another way to say it is that the conditional distribution of N_1 given $N=n$ is binomial $B(n,p)$,

$$\mathbb{P}\big(N_1 = k \mid N = n\big) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $1 \le k \le n$, and $N_2 = N - N_1$.

Lemma 1.7 (Poisson colouring). *The random variables N_1 and N_2 are independent and have distributions* $\text{Poiss}(p\lambda)$ *and* $\text{Poiss}((1-p)\lambda)$.

If a chicken lays Poisson number of eggs and we colour them red or blue by tossing a coin then knowing the number of red eggs gives no information about the number of blue eggs and, moreover, both are Poisson. Also, notice that their means agree with the linearity of expectations, since

$$\lambda = \mathbb{E}N = \mathbb{E}N_1 + \mathbb{E}N_2 = p\lambda + (1-p)\lambda.$$

Proof. By (1.91), the joint distribution of (N_1, N) is

$$\mathbb{P}\big(N_1 = k, N = n\big) = \binom{n}{k} p^k (1-p)^{n-k} \times \frac{\lambda^n}{n!} e^{-\lambda}.$$

Since $N_1 + N_2 = N$, we can also write this as

$$\mathbb{P}\big(N_1 = k, N_2 = n-k\big) = \binom{n}{k} p^k (1-p)^{n-k} \times \frac{\lambda^n}{n!} e^{-\lambda}.$$

Making the change of variables $m = n - k$,

$$\mathbb{P}\big(N_1 = k, N_2 = m\big) = \binom{k+m}{k} p^k (1-p)^{m} \times \frac{\lambda^{k+m}}{(k+m)!} e^{-\lambda}.$$

The right hand side can be rewritten as

$$\begin{aligned} &\frac{(k+m)!}{k!m!} p^k (1-p)^m \times \frac{\lambda^{k+m}}{(k+m)!} e^{-\lambda} \\ &= \frac{(p\lambda)^k}{k!} e^{-p\lambda} \times \frac{(p\lambda)^m}{m!} e^{-(1-p)\lambda}, \end{aligned}$$

where we recognize the Poisson probabilities $\mathrm{Poiss}_{p\lambda}(k)$ and $\mathrm{Poiss}_{(1-p)\lambda}(m)$. Hence,

$$\mathbb{P}\big(N_1 = k, N_2 = m\big) = \mathrm{Poiss}_{p\lambda}(k) \times \mathrm{Poiss}_{(1-p)\lambda}(m).$$

Summing both sides over $m \geq 0$ or over $k \geq 0$ gives that

$$\begin{aligned}\mathbb{P}(N_1 = k) &= \mathrm{Poiss}_{p\lambda}(k),\\ \mathbb{P}(N_2 = m) &= \mathrm{Poiss}_{(1-p)\lambda}(m),\end{aligned}$$

so these random variables have distributions $\mathrm{Poiss}(p\lambda)$ and $\mathrm{Poiss}((1-p)\lambda)$. This implies that

$$\mathbb{P}\big(N_1 = k, N_2 = m\big) = \mathbb{P}\big(N_1 = k\big)\mathbb{P}\big(N_2 = m\big),$$

which means that N_1 and N_2 are independent. This finishes the proof. □

Two-stage constructions can be extended to any number of stages. For example, consider the following exercise.

Exercise 1.4.5. Show that

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(C \mid A \cap B)\mathbb{P}(B \mid A)\mathbb{P}(A),$$

when the events A and $A \cap B$ have positive probabilities.

Once you have done this exercise, it should be clear that the joint distribution formula (1.91) can be extended to longer vectors of random variables $(X_1, \ldots, X_n)$,

$$\begin{aligned}&\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n\big)\\ &= \prod_{k=0}^{n-1} \mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_1 = x_1, \ldots, X_k = x_k\big), \qquad (1.92)\end{aligned}$$

where the term for $k = 0$ is just $\mathbb{P}(X_1 = x_1)$. In other words, if we know the distribution of the next outcome X_{k+1} once the preceding outcomes $X_1 = x_1, \ldots, X_k = x_k$ are revealed, this allows us to reconstruct the joint probabilities of the entire sequence recursively.

Example 1.4.5 (Markov chains). A sequence $X_1, \ldots, X_n$ is called *a Markov chain* if

$$\begin{aligned}&\mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_1 = x_1, \ldots, X_k = x_k\big)\\ &= \mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_k = x_k\big), \qquad (1.93)\end{aligned}$$

which means that the conditional distribution of the next outcome X_{k+1} given the preceding outcomes $X_1 = x_1, \dots,$ $X_k = x_k$ depends only on the most recent outcome $X_k = x_k$. This property is called a *Markov property* of the sequence, also called a *memoryless property*, in the sense that we do not need to remember the entire past and only need to know the most recent outcome to know the chances of the next outcome. The conditional distribution in (1.93) is also called *transition probability*. Markov chain is called *homogeneous* if transition probabilities $\mathbb{P}(X_{k+1} = a \mid X_k = b)$ do not depend on the index k. The entire Chapter 5 will be devoted to homogeneous Markov chains. □

Exercise 1.4.6. If X and Y are independent, fill in the blanks in their joint distribution

$X \backslash Y$	1	2	3
1	?	?	0.1
2	0.24	0.16	0.4

and compute $\mathbb{P}(Y = 2 \mid X = 1)$.

Exercise 1.4.7 (Bayes' Formula). Show that, for possible values a and b of the random variables X and Y,

$$\mathbb{P}\big(Y = b \mid X = a\big) = \frac{\mathbb{P}(X = a \mid Y = b)\,\mathbb{P}(Y = b)}{\sum_m \mathbb{P}(X = a \mid Y = b_m)\,\mathbb{P}(Y = b_m)}.$$

Exercise 1.4.8. If X_1, X_2, and X_3 are independent Bernoulli $B(1/2)$ random variables, find the conditional distribution of $X_1 + X_2 + X_3$ given $X_1 X_2 X_3 = 0$.

Exercise 1.4.9. Charlie met a pretty girl at the local coffee shop Friday night. She wanted to see Charlie again, so she wrote her phone number on his hand; her number was 854-2564. When Charlie got home he realized that his hands were sweating more than usual, and two of the numbers wore off. Unfortunately, Charlie could only remember that the girl's phone number had a lot of fours and fives, so when he calls her the next day he dials either four or five in place of the numbers that wore away, and he flips his lucky coin twice to pick the numbers (heads - four, tails - five). What is the probability that he guessed right and calls the girl?

Exercise 1.4.10. In a Probability class, 20% of the students are failing. Of the entire class, 60% of students are both passing and going to tutorials. Given that a student is passing, what is the probability that she is going to tutorials?

Exercise 1.4.11. In the Example 1.4.3, what is the conditional distribution $\mathbb{P}(\Lambda = \lambda \mid N = 15)$? In other words, if we did not follow the weather in Florida over the summer and later saw on the news that there were 15 shark attacks, how should the probabilities of normal, colder and warmer weather be updated given this information?

Exercise 1.4.12. Bill drives to work 50% of the time and walks the rest of the time. If Bill drives, he is speeding 5mph 60% of the time, 10mph 10% of the time and not speeding 30% of the time. He is always late when he walks or drives within speed limit, he is always on time when he is speeding 10mph, and he is late 50% of the time when he is speeding only 5mph. If he is running late, what is the probability he is walking.

Exercise 1.4.13. Generalize the colouring property of the Poisson to more than two colours using the multinomial distribution in Example 1.2.5 in Section 1.2.

Exercise 1.4.14. Show that the Markov property in (1.93) is equivalent to

$$\begin{aligned}&\mathbb{P}\big(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, X_{k+1} = x_{k+1} \mid X_k = x_k\big)\\&= \mathbb{P}\big(X_1 = x_1, \ldots, X_{k-1} = x_{k-1} \mid X_k = x_k\big)\\&\quad \times \mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_k = x_k\big).\end{aligned}$$

This can be expressed by saying that the past outcomes $X_1 = x_1, \dots, X_{k-1} = x_{k-1}$ and the future outcome $X_{k+1} = x_{k+1}$ are *conditionally independent* given the present $X_k = x_k$.

Exercise 1.4.15. If $X_1, \dots, X_n$ are independent and all $X_i \geq 0$, show that

$$\mathbb{E} \max_{i \leq n} X_i = \int_0^\infty \Big[1 - \prod_{i=1}^n \mathbb{P}(X_i < t) \Big] \, dt.$$

Hint: use Lemma 1.3.

1.5 Applications of the Linearity of Expectation

When we introduced the binomial distribution in Section 1.2, we saw that computing its expectation was much easier using the linearity of expectation. Similarly, whenever a random variable N takes integer values $n \geq 0$ and can be represented as a sum of indicators of some events $A_i \subseteq \Omega$,

$$N = N(\omega) = \sum_{i=1}^{n} \mathrm{I}(\omega \in A_i),$$

then its expectation can be computed using the linearity of expectation,

$$\mathbb{E}N = \sum_{i=1}^{n} \mathbb{E}\,\mathrm{I}(\omega \in A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

It is often easier to compute the probabilities $\mathbb{P}(A_i)$ than to compute the probabilities $\mathbb{P}(N = n)$ and to use the formula $\mathbb{E}N = \sum_{n \geq 1} n\mathbb{P}(N = n)$.

Example 1.5.1 (Functions on a finite set). Let us consider the set Ω of all functions

$$\omega\colon \big\{1,\ldots,n\big\} \to \big\{1,\ldots,n\big\}.$$

Choosing a function at random from Ω means that we assign them equal probabilities

$$\mathbb{P}(\omega) = \frac{1}{n^n}.$$

Let $N = \operatorname{card} R(\omega)$ be the cardinality of the range of ω,

$$R(\omega) = \big\{\omega(1),\ldots,\omega(n)\big\}.$$

One can also think of this as throwing n balls into n boxes at random and N being the number of non-empty boxes. If we want to compute the distribution $\mathbb{P}(N = k)$ then we need to count how many functions have the range of cardinality exactly k. On the other hand, if we only want to know $\mathbb{E}N$ we can represent

$$N = \sum_{i=1}^{n} \mathrm{I}\big(i \in R(\omega)\big),$$

so that, by symmetry,

$$\mathbb{E}N = \sum_{i=1}^{n} \mathbb{P}\big(i \in R(\omega)\big) = n\mathbb{P}\big(1 \in R(\omega)\big).$$

To calculate the probability that $1 \in R(\omega)$, we can write it as

$$\mathbb{P}\big(1 \in R(\omega)\big) = 1 - \mathbb{P}\big(1 \notin R(\omega)\big) = 1 - \frac{(n-1)^n}{n^n},$$

because the number of functions $\omega \colon \{1,\ldots,n\} \to \{2,\ldots,n\}$ is $(n-1)^n$. Therefore,

$$\mathbb{E}N = n\Big(1 - \Big(1 - \frac{1}{n}\Big)^n\Big) \sim n\Big(1 - \frac{1}{e}\Big),$$

where $\sim$ means that the ratio goes to one. □

To compute the distribution $\mathbb{P}(N = k)$ in the previous example, we need to use *the inclusion-exclusion principle*, which is a generalization of Exercise 1.1.1, part 2.

Lemma 1.8 (Inclusion-exclusion principle). *For any events $A_1,\ldots,A_n \subseteq \Omega$,*

$$\begin{aligned}\mathbb{P}\big(\cup_{i\le n} A_i\big) = {} & \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}\big(A_i \cap A_j\big) \\ & + \sum_{i<j<k} \mathbb{P}\big(A_i \cap A_j \cap A_k\big) + \ldots \\ & \ldots + (-1)^{n-1}\mathbb{P}\big(\cap_{i\le n} A_i\big). \qquad (1.94)\end{aligned}$$

Proof. Let us begin by writing

$$(1-x_1)\cdots(1-x_n) = 1-\sum_i x_i + \sum_{i<j} x_i x_j$$
$$-\sum_{i<j<k} x_i x_j x_k + \ldots + (-1)^n x_1\cdots x_n,$$

where we multiplied out the product and wrote all possible ways to choose factors x_i and factors 1.

Exercise 1.5.1. Prove the above identity by induction on n.

To continue the proof, let us regroup the terms in the above identity,

$$1-(1-x_1)\cdots(1-x_n) = \sum_i x_i - \sum_{i<j} x_i x_j$$
$$+\sum_{i<j<k} x_i x_j x_k + \ldots + (-1)^{n-1} x_1\cdots x_n.$$

Let us take $x_i = \mathrm{I}(\omega\in A_i)$ and rewrite the left hand side as

$$1-\prod_{i\le n}\big(1-\mathrm{I}(\omega\in A_i)\big) = 1-\prod_{i\le n}\mathrm{I}(\omega\in A_i^c)$$
$$= 1-\mathrm{I}\Big(\omega\in\bigcap_{i\le n}A_i^c\Big) = \mathrm{I}\Big(\omega\in\Big(\bigcap_{i\le n}A_i^c\Big)^c\Big)$$
$$= \mathrm{I}\Big(\omega\in\bigcup_{i\le n}A_i\Big).$$

Therefore, taking expectation of the left hand side we get $\mathbb{P}(\cup_{i\le n}A_i)$. On the other hand, taking expectation of the right hand side, by linearity of expectation we get the right hand side of (1.94), and this finishes the proof. □

Example 1.5.2 (Functions on a finite set, continued). Let us now compute the probability $\mathbb{P}(N=k)$. First of all, there are $\binom{n}{k}$ ways to choose k values out of $\{1,\ldots,n\}$ that will constitute the range $R(\omega)$ when $N(\omega)=k$. Then we need to count how many functions are there with this range. By symmetry, we can assume that the range is $\{1,\ldots,k\}$. Let us denote the set of n^k functions

$$\omega\colon \{1,\ldots,n\} \to \{1,\ldots,k\}$$

by Ω_k. Then the question is how many surjective functions are there in Ω_k, i.e. whose range covers all the k values. Again, we can think of this as the number of ways to place n balls into k boxes so that no box is left empty. It will be easier to count the complement, namely, all functions $\omega \in \Omega_k$ whose range misses at least one of the values. Let us consider the event

$$A_i = \Big\{\omega \in \Omega_k \,:\, i \notin R(\omega)\Big\},$$

where $R(\omega)$ is the range of ω, and let $\mathbb{P}_k$ be the uniform probability on Ω_k,

$$\mathbb{P}_k(A) = \frac{\mathrm{card}(A)}{n^k}. \tag{1.95}$$

The set of all functions $\omega \in \Omega_k$ whose range misses at least one of the values is just $A_1 \cup \ldots \cup A_k$, and we can apply the inclusion-exclusion principle to this union. For any $\ell \leq k$ and any indices $1 \leq i_1 < \ldots < i_\ell \leq k$,

$$\mathbb{P}_k\big(A_{i_1} \cap \ldots \cap A_{i_\ell}\big) = \frac{(k-\ell)^n}{k^n},$$

because the number of functions in Ω_k that are not allowed to take values $i_1 < \ldots < i_\ell$ is $(k-\ell)^n$. There are $\binom{k}{\ell}$ choices of such indices so, by inclusion-exclusion principle,

$$\mathbb{P}_k\big(A_1 \cup \ldots \cup A_k\big) = \sum_{\ell=1}^{k} (-1)^{\ell-1} \binom{k}{\ell} \frac{(k-\ell)^n}{k^n}.$$

By (1.95), to get the cardinality, we need to multiply by k^n, and recalling that we counted over the complement, we get that the number of surjective functions in Ω_k is

$$k^n - \sum_{\ell=1}^{k}(-1)^{\ell-1}\binom{k}{\ell}(k-\ell)^n = \sum_{\ell=0}^{k}(-1)^{\ell}\binom{k}{\ell}(k-\ell)^n.$$

Finally, multiplying by the number $\binom{n}{k}$ of ways to choose k values out of $\{1,\ldots,n\}$, we get that the number of functions in Ω with $N(\omega)=k$ is

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

and, therefore, we finally get

$$\mathbb{P}(N=k) = \frac{1}{n^n}\binom{n}{k}\sum_{\ell=0}^{k}(-1)^{\ell}\binom{k}{\ell}(k-\ell)^n.$$

Using this formula together with $\mathbb{E}N = \sum_{k=0}^{n} k\mathbb{P}(N=k)$, it is much harder to get to the simple answer for $\mathbb{E}N$ we obtained above by the linearity of expectation. □

In many problems it is not feasible to compute the distribution of N, but we can obtain useful information about the tail probabilities $\mathbb{P}(N \geq k)$ by computing the expected number of events that 'witness' that $N \geq k$. Let us look at a couple of examples.

Example 1.5.3 (Longest increasing subsequence of a random permutation). Let S_n be the set of all permutations σ of n elements, i.e. all bijections

$$\sigma\colon \{1,\ldots,n\} \to \{1,\ldots,n\}.$$

Let $\mathbb{P}$ be the uniform probability on S_n such that $\mathbb{P}(\sigma) = 1/n!$. If, for some $i_1 < \ldots < i_k$,

$$\sigma(i_1) < \ldots < \sigma(i_k),$$

we call this an *increasing subsequence of length k.* Let us define $L = L(\sigma)$ to be the maximum length among all increasing subsequences of σ. For example, the permutation $(1,5,2,4,3)$ has $L = 3$ with subsequences $(1,2,4)$ and $(1,2,3)$ having length 3. To compute the distribution of L or even $\mathbb{E}L$ is too difficult, and for a number of years it was an open problem to show that

$$\lim_{n\to\infty} \frac{\mathbb{E}L}{\sqrt{n}} = 2.$$

In this example we will show that, for most permutations, $L \le 3\sqrt{n}$ and, as a result, the expectation $\mathbb{E}L \le 4\sqrt{n}$.

Let $N_k = N_k(\sigma)$ be the number of increasing subsequences of length k in the permutation σ. In other words, N_k is the number of different $i_1 < \ldots < i_k$ on which σ is increasing. Let us compute the expectation $\mathbb{E}N_k$. If

$$I_k = \Big\{ i = (i_1,\ldots,i_k) : 1 \le i_1 < \ldots < i_k \le n \Big\}$$

is the set of all possible choices of k different indices then we can write

$$N_k = \sum_{i\in I_k} \mathrm{I}\big(\sigma(i_1) < \ldots < \sigma(i_k)\big)$$

and, by the linearity of expectation and symmetry,

$$\begin{aligned}\mathbb{E}N_k &= \sum_{i\in I_k} \mathbb{P}\big(\sigma(i_1) < \ldots < \sigma(i_k)\big)\\ &= \binom{n}{k} \mathbb{P}\big(\sigma(1) < \ldots < \sigma(k)\big).\end{aligned}$$

The last probability equals

$$\mathbb{P}\big(\sigma(1) < \ldots < \sigma(k)\big) = \frac{1}{k!}$$

which can be argued by symmetry, since all possible orders among $\sigma(1),\ldots,\sigma(k)$ are equally likely. Or, we can simply count the number of permutations σ such that $\sigma(1) < \ldots <$

$\sigma(k)$, which equals

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

because there are $\binom{n}{k}$ ways to choose values $\sigma(1),\dots,\sigma(k)$ (and one way to arrange them in the increasing order) and $(n-k)!$ ways to arrange the remaining $n-k$ values. Dividing by $n!$ we again get that the probability is $\frac{1}{k!}$. This implies

$$\mathbb{E}N_k = \binom{n}{k}\frac{1}{k!}. \tag{1.96}$$

We can already use this to get some information about the length L of the longest increasing subsequence, so we will come back to this a little later in this section. □

Example 1.5.4 (Cliques in Erdős–Rényi random graph). If $V = \{v_1,\dots,v_n\}$ is the set of vertices of a graph G, a subset of vertices $W \subseteq V$ is called a *clique* if all vertices in W are connected by edges in the graph. The *clique number* $\omega(G)$ of the graph G is the size k of the largest clique.

In this example we will consider the Erdős–Rényi random graph $G = G(n,p)$ from Example 1.2.3 in Section 1.2 and show that its clique number is *typically* (meaning, with probability close to 1) not bigger than of order $\mathscr{O}(\log n)$. We will do this by counting the expected number of cliques of a given size.

Let N_k be the number of cliques of size k in G and let us compute $\mathbb{E}N_k$. There are $\binom{n}{k}$ subsets of size k and, for a given subset W of k vertices to be a clique, all of the $\binom{k}{2}$ edges between these vertices must be present, which happens with probability

$$\mathbb{P}(W \text{ is a clique}) = p^{\binom{k}{2}}.$$

As in the previous example, representing N_k as the sum of indicators over all such W and using the linearity of expectation, we get

$$\mathbb{E}N_k = \binom{n}{k} p^{\binom{k}{2}}. \tag{1.97}$$

We will extract some information from this formula a little later in this section. □

To extract information from the formulas we proved in the last two examples, we will need one result. It is absolutely fundamental and will be the starting point of many much more sophisticated results later.

Theorem 1.4 (Chebyshev's inequality). *For any random variable X and any $x > 0$,*

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}X\,\mathrm{I}(X \geq x)}{x}. \tag{1.98}$$

Moreover, if $X \geq 0$ then, for any $x > 0$,

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}X}{x}. \tag{1.99}$$

Proof. We can write

$$X\,\mathrm{I}(X \geq x) \geq x\,\mathrm{I}(X \geq x),$$

because the indicator $\mathrm{I}(X \geq x)$ is non-zero only when $X \geq x$. Taking expectation of both sides and using the monotonicity property of expectation, we get

$$\mathbb{E}X\,\mathrm{I}(X \geq x) \geq x\mathbb{E}\,\mathrm{I}(X \geq x) = x\mathbb{P}(X \geq x).$$

Dividing both sides by x finishes the proof of (1.98). If $X \geq 0$ then $X\,\mathrm{I}(X \geq x) \leq X$ and, taking expectations,

$$\mathbb{E}X\,\mathrm{I}(X \geq x) \leq \mathbb{E}X.$$

In this case, (1.98) implies (1.99). □

The proof of Chebyshev's inequality (1.99) looks almost trivial, but its historical significance and its usefulness can not be overstated.

In this section, we will use it for integer-valued random variables $N \geq 0$, for which $\{N > 0\} = \{N \geq 1\}$. Using Chebyshev's inequality with $x = 1$,

$$\mathbb{P}(N > 0) = \mathbb{P}(N \geq 1) \leq \mathbb{E}N. \tag{1.100}$$

This tells us that if the expectation $\mathbb{E}N$ is very small then $\mathbb{P}(N = 0) = 1 - \mathbb{P}(N > 0)$ is close to 1, so for most outcomes $\omega \in \Omega$, we have $N = N(\omega) = 0$.

Another result we will need is Stirling's formula,

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{1.101}$$

as $n \to \infty$, where as usual $\sim$ means that the ratio of two sides goes to one. We will not prove this here in order not to get sidetracked by a somewhat technical result. However, let us mention that it is very easy to show that

$$n! \sim c\sqrt{n}\left(\frac{n}{e}\right)^n$$

for some constant $c > 0$, which was discovered by Abraham de Moivre, and it requires more work to show that $c = \sqrt{2\pi}$, which was proved by James Sterling. There are a number of different proofs of this fact. For our applications below, the factor $\sqrt{2\pi n}$ in the Stirling's formula will be irrelevant, so we can just as well use the formula with some unknown c. Also, there exist precise estimates in Stirling's formula, for example, for $n \geq 1$,

$$e^{\frac{1}{12n+1}}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq e^{\frac{1}{12n}}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n. \tag{1.102}$$

Now we go back to our two previous examples.

Example 1.5.5 (Longest increasing subsequence, part 2). We showed in (1.96) that the expectation of the number N_k of increasing subsequences of length k in a random uniform permutation of $\{1,\ldots,n\}$ is

$$\mathbb{E}N_k = \binom{n}{k}\frac{1}{k!} = \frac{n!}{(k!)^2(n-k)!} \le \frac{n^k}{(k!)^2}. \tag{1.103}$$

By Stirling's formula, $k! \ge (\frac{k}{e})^k$, so

$$\mathbb{E}N_k \le n^k\left(\frac{e}{k}\right)^{2k} = \left(\frac{e\sqrt{n}}{k}\right)^{2k}. \tag{1.104}$$

By Chebyshev's inequality (1.100),

$$\mathbb{P}(N_k > 0) \le \left(\frac{e\sqrt{n}}{k}\right)^{2k},$$

and if we take $k = 3\sqrt{n}$ then

$$\mathbb{P}\Big(N_{3\sqrt{n}} > 0\Big) \le \left(\frac{e}{3}\right)^{6\sqrt{n}} \le e^{-\sqrt{n}/2},$$

where in the last inequality we used that $(e/3)^6 \le e^{-1/2}$ to simplify the expression. First of all, this means that it is very unlikely that there exists even one increasing subsequence of length $3\sqrt{n}$ in a random permutation. Of course, in this case the length L of the longest increasing subsequence is smaller than $3\sqrt{n}$, so we showed that

$$\mathbb{P}\big(L \ge 3\sqrt{n}\big) \le e^{-\sqrt{n}/2}. \tag{1.105}$$

From this, we can also estimate the expectation $\mathbb{E}L$. Since L never exceeds n,

$$\begin{aligned} L &= L\Big[\mathrm{I}\big(L < 3\sqrt{n}\big) + \mathrm{I}\big(L \ge 3\sqrt{n}\big)\Big] \\ &\le 3\sqrt{n} + n\mathrm{I}\big(L \ge 3\sqrt{n}\big). \end{aligned}$$

Taking expectation of both sides, we get

$$\mathbb{E}L \le 3\sqrt{n} + n\mathbb{P}\big(L \ge 3\sqrt{n}\big) \le 3\sqrt{n} + ne^{-\sqrt{n}/2} \le 4\sqrt{n},$$

although the inequality (1.105) is, of course, a much stronger statement. □

Next, we will come back to the example of cliques in the Erdős–Rényi random graph. We will have to carry out some tedious calculations, but it is a good idea to see at least one example which show that the linearity of expectation can yield non-trivial information when it is not clear right away. These calculations will also prepare us for Section 2.4 in the next chapter, where we will obtain further information using the second moment method.

Example 1.5.6 (Cliques in Erdős–Rényi graph, part 2). Let us recall the formula in (1.97), which states that

$$f(k) := \binom{n}{k} p^{\binom{k}{2}} = \frac{n!}{k!(n-k)!} p^{k(k-1)/2} \tag{1.106}$$

is the expectation of the number N_k of cliques of size k in the Erdős–Rényi random graph. For a fixed k, this number goes to infinity when $n \to \infty$, so we expect many cliques of small size. If we find $k = k(n)$ for which this expectation becomes small, Chebyshev's inequality will tell us that it is unlikely to have cliques of that size. Since

$$f(k+1) := \frac{n!}{(k+1)!(n-k-1)!} p^{k(k+1)/2},$$

it is easy to check that

$$\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1} p^k. \tag{1.107}$$

Notice that this ratio is decreasing in k, it is bigger than 1 for small k, and eventually it becomes smaller than 1. This means that $f(k)$ is first increasing and then decreasing. Since

$$f(1) = n \text{ and } f(n) = p^{n(n-1)/2} \ll 1,$$

there is a unique point k_0 such that

$$f(k_0) \geq 1 > f(k_0+1). \tag{1.108}$$

This is the *transition point* where the expected number of cliques of size k becomes smaller than 1 and, heuristically, the graph is unlikely to have cliques of size bigger than k_0. We will make this more precise below, but first we would like to understand how k_0 looks like. We can not write down the exact formula for k_0, but we can accurately estimate its location.

Lemma 1.9 (Clique number). *For large n, we have*

$$k_0 = k_0(n) \sim \frac{2}{\log(1/p)} \log n. \tag{1.109}$$

As usual, $\sim$ means that the ratio of two sides goes to 1. In fact, we will give precise explicit bounds in the proof.

Proof. We will use simple bounds for binomial coefficients:

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

which imply the bounds for $f(k)$:

$$\left(\frac{n}{k}-1\right)^k p^{k(k-1)/2} \leq f(k) \leq n^k p^{k(k-1)/2}. \tag{1.110}$$

The right hand side is smaller than 1 if $np^{(k-1)/2} < 1$. Taking logarithms, this is the same as $k > b_n$, where we introduce the notation

$$b_n := c_p \log n + 1, \text{ where } c_p := \frac{2}{\log(1/p)}. \tag{1.111}$$

This means that $f(k) \geq 1$ can hold only if $k \leq b_n$, which proves that $k_0 \leq b_n$. Similarly, taking logarithms, the left hand side of (1.110) is ≥ 1 if and only if

$$k \leq c_p \log\left(\frac{n}{k}-1\right) + 1.$$

Since $f(k_0+1)<1$, k_0+1 violates this inequality, so

$$k_0+1 \geq c_p \log\Big(\frac{n}{k_0+1}-1\Big)+1$$

and, therefore,

$$k_0 \geq c_p \log\Big(\frac{n}{k_0+1}-1\Big).$$

Since we have already shown that $k_0 \leq b_n$, this implies that

$$k_0 \geq c_p \log\Big(\frac{n}{b_n+1}-1\Big).$$

Thus, we obtained explicit bounds on k_0,

$$c_p \log\Big(\frac{n}{b_n+1}-1\Big) \leq k_0 \leq b_n. \tag{1.112}$$

It is easy to see that the ratio of the two sides goes to 1, because we can rewrite the left hand side as

$$\begin{aligned} c_p \log\Big(\frac{n}{b_n+1}-1\Big) &= c_p \log n - c_p \log(b_n+1) \\ &\quad + c_p \log\Big(1-\frac{b_n+1}{n}\Big) \end{aligned}$$

and notice that the last two terms are much smaller than the first when $n \to \infty$. This finishes the proof of (1.109). □

To continue with our example, we will now check that as soon as the expectation $f(k)$ of the number N_k of cliques of size k in (1.106) crosses level 1 as defined in (1.108), i.e. for $k>k_0$, the expectation $f(k)$ becomes small very quickly. By (1.107),

$$f(k+1) \leq np^k f(k).$$

Also notice that, for any $\varepsilon>0$,

$$k \geq \frac{(2-\varepsilon)\log n}{\log(1/p)} \Longleftrightarrow p^k \leq \frac{1}{n^{2-\varepsilon}}. \tag{1.113}$$

Since, by the above lemma, for any $\varepsilon > 0$, k_0 satisfies this condition for large enough n, we get that, for $k \geq k_0$,

$$f(k+1) \leq np^k f(k) \leq \frac{1}{n^{1-\varepsilon}} f(k).$$

Repeating this recursively $m \geq 1$ times, we get

$$f(k+m) \leq \frac{1}{n^{m(1-\varepsilon)}} f(k).$$

In particular, since $f(k_0+1) < 1$,

$$f(k_0+m+1) \leq \frac{1}{n^{m(1-\varepsilon)}}.$$

By Chebyshev's inequality (1.100),

$$\mathbb{P}\big(N_{k_0+m+1} > 0\big) \leq \frac{1}{n^{m(1-\varepsilon)}}. \tag{1.114}$$

This means that, it is unlikely that the random graph has cliques of size k_0+2, and it gets even more unlikely to have cliques of size k_0+1+m as m grows. □

Remark 1.4. To summarize, we used together the linearity of expectation and Chebyshev's inequality to show that the longest increasing subsequence in a random permutation is typically not bigger than $3\sqrt{n}$ and the clique number $\omega(G(n,p))$ of the Erdős–Rényi random graph is typically not bigger than $\frac{2\log n}{\log(1/p)}$. In both cases, we were counting the number of subsets in a certain configuration and, as the size of subsets got bigger, there was a competition between the large number of such subsets and the small probability that the subset appears in a special configuration. Eventually, the small probability dominated, which allowed us to conclude that typically there are no such large subsets. □

Exercise 1.5.2. Give an example of integer valued random variable $N \geq 0$ such that $\mathbb{P}(N > 0) = e^{-10}$ and $\mathbb{E}N = e^{10}$.

Exercise 1.5.3. Suppose there were n pairs of animals in Noah's ark and m animals died. Compute the expectation of the number of complete pairs left.

Exercise 1.5.4. Let us say that four vertices $\{v_1, v_2, v_3, v_4\}$ in a graph G form a square if there are exactly 4 edges present among 6 possible edges between them and these edges form a cycle. Compute the expected number of squares in the Erdős–Rényi random graph $G(n, p)$.

Exercise 1.5.5. Six geese are flying overhead. Three hunters each picks one at random, kills it with probability 0.3 and misses with probability 0.7. What is the expected number of geese killed.

Exercise 1.5.6. *(Matching problem)* After a party on a rainy evening, n gentlemen are not in a position to recognize their umbrellas and everyone takes one at random. What is the probability that at least one takes his own umbrella? *Hint:* use the inclusion-exclusion principle.

Exercise 1.5.7. What is the expected number of cycles in a random permutation of $\{1, \ldots, n\}$? *Hint:* write $N = \sum_{i=1}^{n} \frac{1}{X_i}$, where $X_i =$ length of the cycle including i.

Exercise 1.5.8. In a permutation $(\pi_1, \ldots, \pi_n)$ of $\{1, \ldots, n\}$, we say that π_k is *a record* if $\pi_k > \pi_i$ for $i = 1, \ldots, k-1$. What is the expected number of records in a random permutation?

Chapter 2
Second Moment Calculations

So far, we have only considered discrete probability spaces and distributions, and in the next two chapters we implicitly assume that all distributions are discrete. However, many of the results in these chapters rely only on basic properties of probability and also apply to non-discrete probability spaces. In Chapter 4, we will introduce and study some continuous distributions, and it will be clear that many general results in Chapter 2 and Chapter 3 still apply.

2.1 Variance and Covariance

For $k \geq 1$, the expectation $\mathbb{E}X^k$ is called *the k^{th} moment of X*, which is defined if $\mathbb{E}|X|^k < \infty$. In particular, $\mathbb{E}X^2$ is called the second moment, and all results in this chapter will be based on the calculations of second moments of various random variables. A quantity related to the second moment $\mathbb{E}X^2$ is the *variance of X*,

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2, \tag{2.1}$$

which is the second moment of the random variable $X - \mathbb{E}X$. The square root of variance, $\sqrt{\mathrm{Var}(X)}$, is called the *standard deviation*.

Subtracting the expected value $\mathbb{E}X$ from a random variable X is called *centering*, because the expectation of $X - \mathbb{E}X$ is equal to zero,

$$\mathbb{E}(X - \mathbb{E}X) = \mathbb{E}X - \mathbb{E}X = 0.$$

Any random variable whose expectation is zero, $\mathbb{E}X = 0$, is called *centred*.

If we denote $\mu = \mathbb{E}X$, we can rewrite the variance as

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)\\ &= \mathbb{E}X^2 - 2\mu^2 + \mu^2 = \mathbb{E}X^2 - \mu^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2,\end{aligned}$$

so the variance can be computed in two ways,

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2. \tag{2.2}$$

Of course, we need to assume that both $\mathbb{E}X^2$ and $\mathbb{E}X$ are well defined. In fact, it is enough to assume that $\mathbb{E}X^2 < \infty$, because it implies that $\mathbb{E}|X| < \infty$. Indeed,

$$|X| = |X|\mathrm{I}(|X| < 1) + |X|\mathrm{I}(|X| \geq 1) \leq 1 + X^2,$$

because $|X| \leq X^2$ when $|X| \geq 1$, and taking expectations of both sides gives $\mathbb{E}|X| < \infty$. Another way to see this is using the following inequality that will be useful to us later on.

Lemma 2.1 (Jensen's inequality). *If $f\colon \mathbb{R} \to \mathbb{R}$ is convex and $\mathbb{E}X$ is well-defined then*

$$f(\mathbb{E}X) \leq \mathbb{E}f(X). \tag{2.3}$$

If f is concave, the inequality is reversed.

Proof. Let $\mu = \mathbb{E}X$. Since $f(x)$ is convex, its tangent line at $x = \mu$ (or subdifferential line) is below $f(x)$, so

$$f(x) \geq f(\mu) + f'(\mu)(x - \mu).$$

Plugging in the random variable X gives

$$f(X) - f(\mu) - f'(\mu)(X - \mu) \geq 0.$$

Therefore, taking expectations,

$$\mathbb{E}f(X) - f(\mu) - f'(\mu)(\mathbb{E}X - \mu) = \mathbb{E}f(X) - f(\mu) \geq 0,$$

which is exactly (2.3). It is possible that this expectation is undefined, but, because we are taking the expectation of a non-negative random variable, the only way it is undefined is if $\mathbb{E}f(X) = +\infty$. In this case, (2.3) also holds. The proof for concave f is similar. □

Note that the above proof works for functions f defined on any interval, as long as X takes values on the same interval. For example, for nonnegative $X \geq 0$, we could consider a convex of concave function f on $[0, \infty)$.

Example 2.1.1. Suppose that $\mathbb{E}X^2 < \infty$. Take $Y = X^2$, so that $\mathbb{E}Y$ is well-defined. Since $f(x) = \sqrt{x}$ is concave on $[0, \infty)$,

$$\sqrt{\mathbb{E}Y} \geq \mathbb{E}\sqrt{Y}.$$

In other words, $\sqrt{\mathbb{E}X^2} \geq \mathbb{E}|X|$. This is another way to see that $\mathbb{E}X^2 < \infty$ implies that $\mathbb{E}|X| < \infty$. □

We leave some simple examples of computing variance as an exercise.

Exercise 2.1.1. Check that:

(a) If $X \sim B(p)$ then $\mathrm{Var}(X) = p(1 - p)$.
(b) If $X \sim \mathrm{Poiss}(\lambda)$ then $\mathrm{Var}(X) = \lambda$.
Hint: compute $\mathbb{E}X(X - 1)$ first.
(c) If $\mathbb{E}X^2 < \infty$ then $\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var} X$ for all $a, b \in \mathbb{R}$.

Another quantity related to the second moment is called *covariance* and it involves two random variables. Given two random variables X and Y on the same probability space, their covariance is defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y, \quad (2.4)$$

where the last equality can be checked by multiplying out

$$(X-\mathbb{E}X)(Y-\mathbb{E}Y)=XY-(\mathbb{E}X)Y-(\mathbb{E}Y)X+(\mathbb{E}X)(\mathbb{E}Y)$$

and taking the expectation of both sides. Of course, in the definition of covariance we assume that all the expectations are well defined, namely, $\mathbb{E}|XY|,\mathbb{E}|X|,\mathbb{E}|Y|<\infty$. One way to ensure this is to assume that the second moments are finite, $\mathbb{E}X^2,\mathbb{E}Y^2<\infty$, because

$$|XY|\le\frac{1}{2}(X^2+Y^2).$$

In this case, we also have the following.

Lemma 2.2 (Cauchy-Schwarz inequality). *For any random variables X and Y on the same probability space,*

$$\mathbb{E}|XY|\le\big(\mathbb{E}X^2\big)^{1/2}\big(\mathbb{E}Y^2\big)^{1/2}. \tag{2.5}$$

Proof. The proof follows by a standard argument. If we let $a=\mathbb{E}X^2$, $b=\mathbb{E}XY$, and $c=\mathbb{E}Y^2$, then, for all $t\in\mathbb{R}$,

$$0\le\mathbb{E}(tX-Y)^2=at^2-2bt+c.$$

Quadratic function is nonnegative if it has at most one root, so the discriminant $D=4b^2-4ac\le0$. Therefore, $b^2\le ac$ and $|\mathbb{E}XY|^2\le\mathbb{E}X^2\mathbb{E}Y^2$. Using this for $|X|$ and $|Y|$ instead of X and Y proves (2.5). □

Example 2.1.2 (Uncorrelated random variables). If two random variables X and Y are independent and their expectations are well defined, $\mathbb{E}|X|,\mathbb{E}|Y|<\infty$, then

$$\mathrm{Cov}(X,Y)=0. \tag{2.6}$$

This follows from Lemma 1.6 in Section 1.4. Of course, the random variables X and Y do not have to be independent for (2.6) to hold, and random variables with zero covariance are called *uncorrelated*.

For example, consider two independent random variables ε and Z such that

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2},$$

$$\mathbb{P}(Z = \pm 2) = \mathbb{P}(Z = \pm 3) = \frac{1}{4},$$

and let $X = Z, Y = \varepsilon Z$. Then

$$\mathbb{E}X = \mathbb{E}Z = 0, \mathbb{E}Y = \mathbb{E}\varepsilon\mathbb{E}Z = 0,$$

and

$$\mathrm{Cov}(X,Y) = \mathbb{E}XY = \mathbb{E}\varepsilon Z^2 = \mathbb{E}\varepsilon\mathbb{E}Z^2 = 0,$$

so X and Y are uncorrelated. But they are not independent because $|X| = |Y|$, so knowing the value of one determines the other one up to a sign, while independence means that the chances of all outcomes should not be affected. □

Example 2.1.3 (Variance of the sum). Consider random variables $X_1, \ldots, X_n$ on the same probability space such that all $\mathbb{E}X_i^2 < \infty$. If $S_n = \sum_{i=1}^n X_i$ then multiplying out,

$$\begin{aligned}\big(S_n - \mathbb{E}S_n\big)^2 &= \Big(\sum_{i=1}^n (X_i - \mathbb{E}X_i)\Big)^2 \\ &= \sum_{i,j=1}^n (X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j),\end{aligned}$$

and taking expectations on both sides gives

$$\mathrm{Var}(S_n) = \sum_{i,j=1}^n \mathrm{Cov}(X_i, X_j). \tag{2.7}$$

When $i = j$, $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$ and, when $i \neq j$, the terms (i,j) and (j,i) are the same, so covariance of the sum can be also written as

$$\operatorname{Var}(S_n) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j). \tag{2.8}$$

If the random variables $X_1, \dots, X_n$ are uncorrelated then

$$\operatorname{Var}(S_n) = \sum_{i=1}^{n} \operatorname{Var}(X_i). \tag{2.9}$$

In particular, by (2.6), this holds when they are independent. Moreover, if $X_1, \dots, X_n$ are i.i.d. then

$$\operatorname{Var}(S_n) = n \operatorname{Var}(X_1), \tag{2.10}$$

because all their variances are equal. □

Example 2.1.4 (Variance of the Binomial). The variance of the binomial random variable $X \sim B(n, p)$ can be computed by definition, but it is much easier if we recall that the sum of n i.i.d. Bernoulli $B(p)$ random variables is binomial, so

$$\operatorname{Var}(X) = np(1-p),$$

by (2.10) and Exercise 2.1.1 (a). □

Exercise 2.1.2. Show that $|\operatorname{Cov}(X,Y)| \le [\operatorname{Var}(X)\operatorname{Var}(Y)]^{1/2}$.

Exercise 2.1.3. Let $G = G(n, p)$ be the Erdős-Rényi random graph. Let N be the number of vertices among $\{v_3, \dots, v_n\}$ that are connected by edges to both v_1 and v_2. Compute $\operatorname{Var}(N)$.

Exercise 2.1.4. Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. random variables such that $\mathbb{P}(\varepsilon_i = \pm 1) = \frac{1}{2}$. Let $X = \sum_{i<j} \varepsilon_i \varepsilon_j$. Compute $\operatorname{Var}(X)$.

Exercise 2.1.5. Compute variance of the geometric random variable N with the distribution $\mathbb{P}(N = k) = (1-p)^{k-1} p$ for $k \ge 1$.

Exercise 2.1.6. Suppose that the random variables X_1 and X_2 are independent, $\mathbb{E}X_j = \mu_j$ and $\operatorname{Var}(X_j) = \sigma_j^2$ for $j = 1, 2$. What is the variance of $X_1 X_2$? Is it well-defined?

Exercise 2.1.7. If $\mathbb{E}|X|^p < \infty$ for $p > 0$, show that $\mathbb{E}|X|^q < \infty$ for $0 \le q \le p$.

Exercise 2.1.8. Suppose that a random variable X takes four values $-2, -1, 1, 2$ with equal probabilities $\frac{1}{4}$ and let $Y = X^2$. Show that X and Y are uncorrelated but not independent.

Exercise 2.1.9. Suppose that the random variables $X_1, \ldots, X_n$ are i.i.d. and $\mathbb{P}(X_i = \pm 1) = \frac{1}{2}$. For each i, let us pick an index $e(i) \in \{1, \ldots, n\} \setminus \{i\}$ uniformly at random, independently over $i \le n$, and independently of $X_1, \ldots, X_n$. Compute the variance of $X_{e(1)} + \ldots + X_{e(n)}$.

2.2 Classical Law of Large Numbers

When we flip a fair coin 100 times, we expect the number of Heads to be close to 50 or, in other words, the proportion of Heads to be close to its probability 0.5. In this section, we will make a quantitative mathematical statement of this kind, called the *Law of Large Numbers*. In the next chapter, we will strengthen it using more sophisticated methods, but a simple argument based on the second moment calculations that we will use here is still very useful in many situations where 'simple' calculations is all one can do.

Let us recall Chebyshev's inequality in Theorem 1.4 in Section 1.5. A straightforward consequence is the following inequality, which is also called Chebyshev's inequality.

Theorem 2.1 (Chebyshev's inequality II). *If X has finite variance then*

$$\mathbb{P}\big(|X-\mathbb{E}X|\geq x\big)\leq\frac{\mathrm{Var}(X)}{x^2}, \tag{2.11}$$

for any $x>0$.

Proof. Consider $Y=(X-\mathbb{E}X)^2$, which is a non-negative random variable. Then $|X-\mathbb{E}X|\geq x$ if and only if $Y\geq x^2$ and, by Theorem 1.4,

$$\mathbb{P}\big(Y\geq x^2\big)\leq\frac{\mathbb{E}Y}{x^2}=\frac{\mathrm{Var}(X)}{x^2}.$$

This finishes the proof. □

By definition, variance $\mathrm{Var}(X)$ measures deviations of a random variable X from its expectation $\mathbb{E}X$ by computing the expectation of the square of this deviation $(X-\mathbb{E}X)^2$. Small variance indicates that X is typically close to its expected value $\mathbb{E}X$, and the inequality (2.11) gives precise meaning to this statement. Namely, the probability that this deviation exceeds $x>0$ is smaller than $\mathrm{Var}(X)/x^2$. To state our first application, given random variables $X_1,\dots,X_n$, let us denote their average by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.12}$$

Theorem 2.2 (Law of large numbers). *If random variables $X_1,\dots,X_n$ are i.i.d., $\mu = \mathbb{E}X_1$ and $\sigma^2 = \mathrm{Var}(X_1) < \infty$ then, for any $\varepsilon > 0$,*

$$\mathbb{P}\big(|\overline{X}_n - \mu| \geq \varepsilon\big) \leq \frac{\sigma^2}{n\varepsilon^2}. \tag{2.13}$$

Proof. By the properties of variance proved in the previous section,

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n\,\mathrm{Var}(X_1)}{n^2} = \frac{\sigma^2}{n}.$$

Then Chebyshev's inequality (2.11) with $X = \overline{X}_n$ and $x = \varepsilon$ becomes exactly (2.13) because $\mathbb{E}\overline{X}_n = \mathbb{E}X_1 = \mu$. □

Because $\sigma^2/n\varepsilon^2 \to 0$ as $n \to \infty$, Chebyshev's inequality (2.11) implies that the average $\overline{X}_n$ differs from its expectation μ by more than an arbitrarily small $\varepsilon > 0$ only with small probability when n gets large. For example, the variance of a fair coin $B(1/2)$ equals $\sigma^2 = 1/4$ and the expectation is $\mu = 0.5$. Therefore,

$$\mathbb{P}\big(|\overline{X}_n - 0.5| \geq \varepsilon\big) \leq \frac{1}{4n\varepsilon^2}. \tag{2.14}$$

For example, the probability that the proportion of Heads differs from 0.5 by more than, say $\varepsilon = 0.01$, is bounded by

$$\mathbb{P}\big(|\overline{X}_n - 0.5| \geq 0.01\big) \leq \frac{2500}{n}.$$

Later on we will prove quantitatively stronger inequalities, but qualitatively this statement is already quite useful as we will see in the next section. Also, because the calculations in the above proof of the Law of Large Numbers were quite simple, assumptions can be easily relaxed, which we leave as an exercise below.

Remark 2.1. Law of large numbers if the first manifestation of a very general idea in Probability and Analysis called the *"concentration of measure phenomenon"*. In the law of large numbers, the average function

$$f(x_1,\ldots,x_n) = \frac{x_1+\ldots+x_n}{n}$$

was shown to concentrate near a constant (its expectation), under certain assumptions on the coordinates, for example, independence. It turns out that many other functions with large number of variables exhibit this behaviour. Later we will see other examples of interesting non-linear functions that also concentrate.

Exercise 2.2.1. Suppose that random variables $X_1,\ldots,X_n$ are uncorrelated, $\mu = \mathbb{E}\overline{X}_n$ and $\mathrm{Var}(X_i) \le \sigma^2$ for all $i \le n$. Then show (2.13) still holds.

Exercise 2.2.2. In the setting of the previous exercise, show that, for any $\delta > 0$,

$$\mathbb{P}\Big(|\overline{X}_n - \mu| \ge \frac{\sigma}{\sqrt{n}}\sqrt{\frac{1}{\delta}}\Big) \le \delta.$$

Exercise 2.2.3. Suppose that X_k for $k \ge 1$ are independent, but not identically distributed, random variables such that

$$\mathbb{P}(X_k = \pm k) = \frac{1}{2k\log(2k)}, \mathbb{P}(X_k = 0) = 1 - \frac{1}{k\log(2k)}.$$

Show that $\lim_{n\to\infty}\mathbb{P}(|\overline{X}_n| \ge \varepsilon) = 0$ for any $\varepsilon > 0$, where $\overline{X}_n$ is the average $\frac{1}{n}\sum_{k=1}^{n} X_k$.

Exercise 2.2.4. If $X \sim \mathrm{Poiss}(\lambda)$, show that, for any $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|\frac{X}{\lambda}-1\Big|\geq\varepsilon\Big)\leq\frac{1}{\lambda\varepsilon^2}.$$

Exercise 2.2.5. *(Law of Large Numbers for U-statistics)* Let $X_1,\ldots,X_n$ be i.i.d. such that $\mathbb{E}X_1=\mu$ and $\sigma^2=\mathrm{Var}(X_1)<\infty$, and consider the random variable

$$U:=\binom{n}{2}^{-1}\sum_{1\leq i<j\leq n}X_iX_j.$$

Prove an explicit bound on $\mathbb{P}(|U-\mu^2|\geq\varepsilon)$ and show that it goes to zero as $n\to\infty$.

Exercise 2.2.6. Suppose that $\varphi\colon\mathbb{R}\to\mathbb{R}$ is a strictly positive nondecreasing function, and suppose that $\varphi(X)$ is integrable. Prove that, for any $x\in\mathbb{R}$,

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

Exercise 2.2.7. Suppose that the random variables X_1, $\ldots$, X_n are i.i.d. and $\mathbb{E}|X_1|^p<\infty$ for some $p>0$. Show that

$$\lim_{n\to\infty}\mathbb{P}\big(\max_{i\leq n}|X_i|\geq\varepsilon n^{1/p}\big)=0$$

for any $\varepsilon>0$. *Hint:* Use the union bound, then Chebyshev's inequality in the form (1.98), and then Exercise 1.2.13.

2.3 Bernstein Polynomials

The Weierstrass approximation theorem states that any continuous function on a closed interval $[a,b]$ can be approximated uniformly by polynomials. In this section, we will use the Law of Large Numbers from the previous section to give one explicit construction of approximating polynomials called Bernstein's polynomials. One can make a change of variables to scale the interval $[a,b]$ to $[0,1]$, so we will only consider the interval $[0,1]$.

If $f\colon [0,1] \to \mathbb{R}$ is a continuous function on $[0,1]$ then the *Bernstein polynomial of order n* associated to this function is defined by

$$B_n(x) := \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} x^k (1-x)^{n-k}. \tag{2.15}$$

Our goal will be to show that $B_n(x)$ uniformly approximates $f(x)$ on the interval $[0,1]$.

Theorem 2.3. *If $f\colon [0,1] \to \mathbb{R}$ is continuous then*

$$\lim_{n\to\infty} \max_{x\in[0,1]} \big|f(x) - B_n(x)\big| = 0. \tag{2.16}$$

Proof. Given $p \in [0,1]$, let $X_1,\ldots,X_n$ be i.i.d. Bernoulli $B(p)$ random variables and let

$$S_n = X_1 + \ldots + X_n \ \text{ and } \ \overline{X}_n = \frac{S_n}{n}.$$

Since the sum S_n has binomial distribution $B(n,p)$,

$$\mathbb{E}f(\overline{X}_n) = \mathbb{E}f\Big(\frac{S_n}{n}\Big) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big)\binom{n}{k} p^k (1-p)^{n-k}$$

is Bernstein's polynomial $B_n(x)$ evaluated at $x = p$. The statement of the theorem is that this expectation is close to $f(p)$ uniformly over $p \in [0,1]$. First,

$$\begin{aligned}\big|B_n(p)-f(p)\big| &= \big|\mathbb{E}f(\bar{X}_n)-f(p)\big| \\ &= \big|\mathbb{E}\big[f(\bar{X}_n)-f(p)\big]\big| \le \mathbb{E}\big|f(\bar{X}_n)-f(p)\big|,\end{aligned}$$

because $|\mathbb{E}X| \le \mathbb{E}|X|$, which follows from the monotonicity property of expectation,

$$-|X| \le X \le |X| \Longrightarrow -\mathbb{E}|X| \le \mathbb{E}X \le \mathbb{E}|X|,$$

or from Jensen's inequality (2.3). Take any $\varepsilon > 0$ and write

$$\begin{aligned}\big|f(\bar{X}_n)-f(p)\big| = \big|f(\bar{X}_n)-f(p)\big|\,\mathrm{I}\big(|\bar{X}_n-p|\le\varepsilon\big) \\ +\big|f(\bar{X}_n)-f(p)\big|\,\mathrm{I}\big(|\bar{X}_n-p|>\varepsilon\big).\end{aligned}$$

Since $|\bar{X}_n-p|\le\varepsilon$ in the first indicator, we can bound the first term by the modulus of continuity of f,

$$\Delta(\varepsilon) := \max_{|x-y|\le\varepsilon}\big|f(x)-f(y)\big|.$$

For the second term, we use that a continuous function f on $[0,1]$ is bounded by some constant, $|f|\le C$, so

$$\big|f(\bar{X}_n)-f(p)\big|\,\mathrm{I}\big(|\bar{X}_n-p|>\varepsilon\big) \le 2C\,\mathrm{I}\big(|\bar{X}_n-p|>\varepsilon\big).$$

Putting two bounds together,

$$\big|f(\bar{X}_n)-f(p)\big| \le \Delta(\varepsilon)+2C\,\mathrm{I}\big(|\bar{X}_n-p|>\varepsilon\big),$$

and taking expectations on both sides,

$$\mathbb{E}\big|f(\bar{X}_n)-f(p)\big| \le \Delta(\varepsilon)+2C\mathbb{P}\big(|\bar{X}_n-p|>\varepsilon\big).$$

The probability in the last term is where we finally use Chebyshev's inequality (2.13) and, since the variance of $B(p)$ is $p(1-p)\le 1/4$, we showed that

$$\big|B_n(p)-f(p)\big| \le \Delta(\varepsilon)+\frac{C}{2n\varepsilon^2}. \tag{2.17}$$

This upper bound does not depend on p, so maximizing over p on the left hand side and then taking the limit,

$$\lim_{n\to\infty}\max_{p\in[0,1]}\big|f(p)-B_n(p)\big|\leq\Delta(\varepsilon).$$

A continuous function on $[0,1]$ is uniformly continuous, so the modulus of continuity $\Delta(\varepsilon)\downarrow 0$ as $\varepsilon\downarrow 0$. Since the upper bound was proved for any $\varepsilon>0$, this finishes the proof. □

Exercise 2.3.1. If $f(x)=x^2$ on $[0,1]$, compute its Bernstein polynomials $B_n(x)$ and show that

$$|B_n(x)-x^2|\leq\frac{1}{4n}.$$

Exercise 2.3.2. Suppose that a differentiable function f on $[0,1]$ is bounded by some constant, $|f|\leq C$, and its derivative is also bounded, $|f'|\leq D$. Show that

$$\max_{x\in[0,1]}\big|B_n(x)-f(x)\big|\leq\frac{3D^{2/3}C^{1/3}}{2n^{1/3}}.$$

Hint: start with (2.17).

Exercise 2.3.3. *(Multivariate Bernstein polynomials)* Consider a continuous $f\colon[0,1]^m\to\mathbb{R}$ and show that

$$\sum_{0\leq k_1,\ldots,k_m\leq n}f\Big(\frac{k_1}{n},\ldots,\frac{k_m}{n}\Big)\prod_{i\leq m}\binom{n}{k_i}x_i^{k_i}(1-x_i)^{n-k_i}\\ \to f(x_1,\ldots,x_m)$$

as $n\to\infty$, uniformly on $[0,1]^m$. *Hint:* consider independent Binomial random variables $B(n,x_i)$, or find a way to use induction.

Exercise 2.3.4. Suppose that $f\colon[0,\infty)\to\mathbb{R}$ is continuous and uniformly bounded, and define

$$P_n(x) := \sum_{k=0}^{\infty} f\left(\frac{k}{n}\right) \frac{(nx)^k}{k!} e^{-nx}.$$

Show that

$$\lim_{n\to\infty} \max_{x\in[a,b]} \left| f(x) - P_n(x) \right| = 0$$

for any finite interval $0 \leq a < b < \infty$. *Hint:* In the proof of (2.16), take $X_1, \ldots, X_n$ to be independent Poisson $\Pi(\lambda)$ instead of Bernoulli $B(p)$.

2.4 Cliques in the Erdős–Rényi Random Graph

In this section we will continue discussing the example of cliques in the Erdős–Rényi graph from Section 1.5. The main purpose of this section is to illustrate that such a simple tool as Chebyshev's inequality can still be quite useful when a random variable is complicated but computing or estimating its variance is doable, even if it requires some effort. When things get complicated, simple things is all one can do.

Let us recall what we did in Section 1.5. We considered the number N_k of cliques of size k in the Erdős–Rényi random graph $G(n,p)$. We showed that its expectation is

$$f(k) = \mathbb{E}N_k = \binom{n}{k} p^{\binom{k}{2}}. \tag{2.18}$$

Then we considered the unique transition point k_0 such that

$$f(k_0) \geq 1 > f(k_0+1), \tag{2.19}$$

which means that at $k = k_0$ there is a transition when the expected number of cliques of size $k > k_0$ becomes smaller than 1. We showed that, for $m \geq 1$,

$$\mathbb{P}\big(N_{k_0+m+1} > 0\big) \leq \frac{1}{n^{m(1-\varepsilon)}}, \tag{2.20}$$

which means that it is very unlikely to have cliques of size k_0+2 and bigger. We also analyzed the formula for $f(k)$ and showed that

$$k_0 = k_0(n) \sim \frac{2\log n}{\log(1/p)} \tag{2.21}$$

in the sense that their ratio goes to 1 when n gets large.

In this section, we would like to understand what happens for $k < k_0$. First, we will see from the calculations we have already done in Section 1.5 that the expected number of such cliques is large. Then we will carry out the second moment calculation and use Chebyshev's inequality to see that N_k

concentrates, similarly to the law of large numbers, which means that the number of such cliques is large not only on average but typically (with probability close to 1).

First, let us look at the expected number $f(k)$ of cliques of size k for $k < k_0$. More precisely, let us take small $\varepsilon > 0$ and consider k in the range

$$\frac{(2-\varepsilon)\log n}{\log(1/p)} \leq k < k_0. \tag{2.22}$$

Exponentiating the first inequality, it is equivalent to

$$p^k \leq \frac{1}{n^{2-\varepsilon}},$$

which we have already seen in (1.113). If we recall the formula (1.107) for the ratio of two consecutive expectations, for such k, this ratio is small:

$$\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1} p^k \leq np^k \leq \frac{1}{n^{1-\varepsilon}}.$$

Since $f(k_0) \geq 1$, this implies that $f(k_0 - 1) \geq n^{1-\varepsilon}$, and $f(k_0 - 2) \geq n^{2(1-\varepsilon)}$, and, by induction, for any fixed $m \geq 1$,

$$f(k_0 - m) \geq n^{m(1-\varepsilon)}. \tag{2.23}$$

This shows that the expected number of cliques of size k in the range (2.22) is large. Next, we will compute the variance of N_k and apply Chebyshev's inequality.

Remark 2.2. A common way to use Chebyshev's inequality (2.11) when the expectation $\mathbb{E}X > 0$ is positive is to take $x = \delta \mathbb{E}X$ for small $\delta > 0$, so

$$\mathbb{P}\big(|X - \mathbb{E}X| \geq \delta \mathbb{E}X\big) \leq \frac{\mathrm{Var}(X)}{\delta^2 (\mathbb{E}X)^2}. \tag{2.24}$$

If the variance is much smaller that $(\mathbb{E}X)^2$,

$$\mathrm{Var}(X) \ll (\mathbb{E}X)^2, \tag{2.25}$$

then the probability in (2.24) is small. Since we can rewrite the opposite inequality $|X - \mathbb{E}X| < \delta \mathbb{E}X$ as

$$1 - \delta \le \frac{X}{\mathbb{E}X} \le 1 + \delta,$$

the probability of the complement of (2.24),

$$\mathbb{P}\Big(1 - \delta \le \frac{X}{\mathbb{E}X} \le 1 + \delta\Big) \ge 1 - \frac{\mathrm{Var}(X)}{\delta^2 (\mathbb{E}X)^2}, \tag{2.26}$$

is close to 1. Especially when $\mathbb{E}X$ is large, this tells us that the random variable X is close to its expectation when viewed on the right scale, because typical deviations of X from $\mathbb{E}X$ are small relative to $\mathbb{E}X$. This concentration around expectation, assuming (2.25) holds, can be viewed as an analogue of the law of large numbers. □

We already know that $f(k) = \mathbb{E}N_k$ is large for k in the range (2.22). If we can show

$$\mathrm{Var}(N_k) \ll (\mathbb{E}N_k)^2 = f(k)^2$$

then we can conclude that N_k is also typically large, by (2.26). Let

$$V_k = \big\{W \subseteq V \,:\, \mathrm{card}(W) = k\big\}$$

be the collection of all subsets of vertices W of size k, and let us represent

$$N_k = \sum_{W \in V_k} \mathrm{I}_W, \ \text{ where } \ \mathrm{I}_W := \mathrm{I}(W \text{ is a clique}). \tag{2.27}$$

In Section 2.1 we saw that the variance of the sum can be written as

$$\mathrm{Var}(N_k) = \sum_{W,W' \in V_k} \mathrm{Cov}(\mathrm{I}_W, \mathrm{I}_{W'}). \tag{2.28}$$

If the subsets W and W' do not intersect or only share one vertex then they do not have any common edges and, hence, the indicators $\mathrm{I}_W, \mathrm{I}_{W'}$ are functions of disjoint sets of edges. Since all edges are independent of each other in the Erdős–Rényi random graph, by the Grouping Lemma in Section 1.4, I_W and $\mathrm{I}_{W'}$ are independent and, therefore, their covariance is equal to zero. This means that in (2.28) we only need to sum over pairs (W,W') such $i = \mathrm{card}(W \cap W') \geq 2$. For such pairs, we will use that

$$\begin{aligned}\mathrm{Cov}(\mathrm{I}_W, \mathrm{I}_{W'}) &= \mathbb{E}\,\mathrm{I}_W\,\mathrm{I}_{W'} - \mathbb{E}\,\mathrm{I}_W\,\mathbb{E}\,\mathrm{I}_{W'} \\ &\leq \mathbb{E}\,\mathrm{I}_W\,\mathrm{I}_{W'} = \mathbb{P}(W,\, W' \text{ are cliques}).\end{aligned}$$

The probability that W and W' are both cliques depends on how many vertices they have in common. If the cardinality $i = \mathrm{card}(W \cap W') \geq 2$ then

$$\mathbb{P}(W,\, W' \text{ are cliques}) = p^{\binom{k}{2}} p^{\binom{k}{2} - \binom{i}{2}},$$

because $\binom{k}{2}$ edges that must be present in W include $\binom{i}{2}$ edges in W', which leaves $\binom{k}{2} - \binom{i}{2}$ additional edges in W'. Next, how many pairs $W, W' \in V_k$ are there such that $i = \mathrm{card}(W \cap W') \geq 2$? There are

$$\binom{n}{k}\binom{k}{i}\binom{n-k}{k-i}$$

such pairs, because there are $\binom{n}{k}$ ways to select vertices in W, then there are $\binom{k}{i}$ ways to select i vertices in W that will be shared with W', and then there are $\binom{n-k}{k-i}$ ways to select $k-i$ vertices in W' that are not shared and are chosen from $n-k$ remaining vertices. Summing over intersection sizes $i = \mathrm{card}(W \cap W') \geq 2$ in (2.28),

$$\begin{aligned}\operatorname{Var}(N_k) &\le \sum_{i=2}^{k} \binom{n}{k}\binom{k}{i}\binom{n-k}{k-i} p^{\binom{k}{2}} p^{\binom{k}{2}-\binom{i}{2}} \\ &= f(k)^2 \sum_{i=2}^{k} \frac{\binom{k}{i}\binom{n-k}{k-i}}{\binom{n}{k}} p^{-\binom{i}{2}}, \qquad (2.29)\end{aligned}$$

where we used the formula (2.18). If we denote the terms in the last sum by

$$a(i) := \frac{\binom{k}{i}\binom{n-k}{k-i}}{\binom{n}{k}} p^{-\binom{i}{2}} \qquad (2.30)$$

then Chebyshev's inequality (2.24) implies

$$\mathbb{P}\big(|N_k - \mathbb{E}N_k| \ge \delta \mathbb{E}N_k\big) \le \frac{1}{\delta^2} \sum_{i=2}^{k} a(i), \qquad (2.31)$$

and it remains to show that the sum on the right hand side is small for k in the range (2.22).

The rest of the calculation is just a tedious analysis of the sequence $a(i)$ and does not really contain any clever idea. It simply requires experimentation and time. We will proceed in two steps:

1. We will check that endpoints $a(2)$ and $a(k)$ are both small.
2. We will check that other $a(i)$ are dominated by $a(2)$ and $a(k)$.

Step 1. Cancelling out common factors in the factorials,

$$\begin{aligned}a(2) &= \frac{\binom{k}{2}\binom{n-k}{k-2}}{\binom{n}{k}} p^{-\binom{2}{2}} \\ &= \frac{k^2(k-1)^2}{2} \cdot \frac{(n-2k+3)\cdots(n-k)}{(n-k+1)\cdots n} \cdot \frac{1}{p} \\ &\le \frac{k^4}{2} \cdot \frac{(n-k)^{k-2}}{(n-k)^k} \cdot \frac{1}{p} = \frac{k^4}{2} \cdot \frac{1}{(n-k)^2} \cdot \frac{1}{p} \le c_p \frac{(\log n)^4}{n^2},\end{aligned}$$

for some constant c_p that depends on p, because k is of order $\mathscr{O}(\log n)$ in the range (2.22). This shows that $a(2)$ is small. Next,

$$a(k) = \frac{1}{\binom{n}{k}} p^{-\binom{k}{2}} = \frac{1}{f(k)},$$

which, as we checked in (2.23), is also small in the range (2.22). □

Step 2. When analyzing the binomial coefficients and factorials, it is often helpful to study the ratio of two neighbours to understand their behaviour. This is what we will do here. For $2 \le i \le k-1$, let us consider the ratio of two consecutive numbers $a(i)$,

$$b(i) := \frac{a(i+1)}{a(i)} = \frac{(k-i)^2}{(i+1)(n-2k+i+1)} p^{-i}. \tag{2.32}$$

This looks much simpler than the original sequence $a(i)$ and we can derive some of its basic properties.

Lemma 2.3. *Suppose that k is in the range (2.22) and take $2 \le i \le k-1$. Then the following statements hold.*

(a) For $i \le \frac{1}{3}\frac{\log n}{\log(1/p)}$ the sequence $b(i) < 1$.
(b) For $i \ge \frac{3}{2}\frac{\log n}{\log(1/p)}$ the sequence $b(i) > 1$.
(c) For $\frac{1}{3}\frac{\log n}{\log(1/p)} < i < \frac{3}{2}\frac{\log n}{\log(1/p)}$ the sequence $b(i)$ is strictly increasing.

Proof. (a) First of all, in the range (2.22), $n-2k+i+1 \ge n/2$ and

$$b(i) \le \frac{2k^2}{n} p^{-i} \le a_p \frac{(\log n)^2}{n} p^{-i}, \tag{2.33}$$

for some constant a_p. Since

$$p^{-i} \le n^{1/3} \iff i \le \frac{1}{3}\frac{\log n}{\log(1/p)}, \tag{2.34}$$

in this range of i we get

$$b(i) \le a_p \frac{(\log n)^2}{n^{2/3}} < 1.$$

(b) Next, for $2 \le i \le k-1$, we can bound $b(i)$ from below by

$$b(i) = \frac{(k-i)^2}{(i+1)(n-2k+i+1)} p^{-i} \ge \frac{1}{kn} p^{-i}.$$

Since

$$p^{-i} \ge n^{3/2} \Longleftrightarrow i \ge \frac{3}{2} \frac{\log n}{\log(1/p)}, \tag{2.35}$$

in this range of i and for k in the range (2.22) we get

$$b(i) \ge \frac{n^{1/2}}{k} > 1. \tag{2.36}$$

(c) In the intermediate range of i in between (2.34) and (2.35),

$$\frac{1}{3} \frac{\log n}{\log(1/p)} < i < \frac{3}{2} \frac{\log n}{\log(1/p)}, \tag{2.37}$$

and for k in the range (2.22), both i and $k-i$ are of order $\mathscr{O}(\log n)$. Therefore,

$$\frac{b(i+1)}{b(i)} \approx \frac{1}{p} > 1,$$

so in this range the sequence $b(i)$ is strictly increasing. □

These three properties imply that there is a unique i_0 such that $b(i) \le 1$ for $i < i_0$ and $b(i) > 1$ for $i \ge i_0$. This implies that $a(i) \le a(3)$ for $i \le i_0$ and $a(i) \le a(k-1)$ for $i > i_0$. As a result,

$$\sum_{i=2}^{k} a(i) \le a(2) + a(k) + k\big(a(3) + a(k-1)\big).$$

Using (2.33) one more time, we see that

$$a(3) \le \frac{a_p}{p^2} \frac{(\log n)^2}{n} a(2),$$

and, using (2.36), we see that

$$a(k-1) \le \frac{k}{\sqrt{n}} a(k) \le \frac{k_0}{\sqrt{n}} a(k) \le b_p \frac{\log n}{\sqrt{n}} a(k),$$

for some constant b_p. Finally, adding up these inequalities and using that $k = \mathscr{O}(\log n)$,

$$\sum_{i=2}^{k} a(i) \le \big(a(2) + a(k)\big)\Big(1 + d_p \frac{(\log n)^2}{\sqrt{n}}\Big) \le 2\big(a(2) + a(k)\big).$$

This finishes Step 2. □

Combining the two steps, we can summarize what we proved as follows.

Theorem 2.4. *Let $\varepsilon, \delta > 0$. For k in the range*

$$\frac{(2-\varepsilon)\log n}{\log(1/p)} \le k < k_0 \sim \frac{2\log n}{\log(1/p)}, \tag{2.38}$$

where k_0 was defined in (2.19), the number N_k of cliques of size k satisfies

$$\mathbb{P}\Big(|N_k - f(k)| \ge \delta f(k)\Big) \le \frac{2}{\delta^2}\Big(c_p \frac{(\log n)^4}{n^2} + \frac{1}{f(k)}\Big). \tag{2.39}$$

Since we saw in (2.23) that $f(k_0 - m) \ge n^{m(1-\varepsilon)}$, the right hand side goes to zero as n goes to infinity. As we mentioned before, this means that, with probability close to 1, the ratio $N_k/f(k)$ is close to 1, so the number of cliques of size k is typically large, just as its expectation $f(k)$. Together with (2.20), this shows that there is a transition occurring at $k = k_0$ where the typical number of cliques goes from being large for smaller k to zero for larger k.

Exercise 2.4.1. Let $\omega(G)$ be the size of the largest clique in the graph G. Show that

$$\mathbb{E}\omega(G(n,p)) \sim \frac{2\log n}{\log(1/p)}.$$

Exercise 2.4.2. A subset W of vertices in a graph is called an *independent set* if there are no edges between any vertices in W. What can you say about the typical size of the largest independent set in the Erdős–Rényi random graph $G(n,p)$ for large n? *Hint:* how do independent sets relate to cliques if we switch present and absent edges?

Exercise 2.4.3. Suppose there were n pairs of animals in Noah's ark and m animals died. Compute the variance of the number of complete pairs of animals left.

Exercise 2.4.4. Compute the variance of the number N_3 of triangles in the Erdős–Rényi random graph.

Exercise 2.4.5. *(Matching problem)* After a party on a rainy evening, n gentlemen are not in a position to recognize their umbrellas and everyone takes one at random. What is the expectation and variance of the number of correct umbrellas taken?

Exercise 2.4.6. If $\mathbb{E}X^2 < \infty$, show that

$$\mathbb{P}(X=0) \le \frac{\mathrm{Var}(|X|)}{\mathbb{E}(X^2)}.$$

Hint: rewrite this in a way that follows from the Cauchy-Schwarz inequality.

2.5 Hardy–Ramanujan Theorem

In this section we will give an example of the second moment calculation from Number Theory. For integer $n \geq 1$, let $\omega(n)$ be the number of distinct prime factors of n,

$$\omega(n) = \mathrm{card}\big\{p \leq n : p \text{ is prime}, p \mid n\big\}. \tag{2.40}$$

Hardy–Ramanujan theorem states the following.

Theorem 2.5 (Hardy–Ramanujan). *For any sequence $\psi(n)$ such that $\psi(n) \to \infty$ as $n \to \infty$, the proportion of numbers $N \in \{1, \ldots, n\}$ that satisfy*

$$\big|\omega(N) - \log\log n\big| \leq \psi(n)\sqrt{\log\log n} \tag{2.41}$$

goes to 1.

In other words, one can take $\psi(n)$ that goes to infinity arbitrarily slowly, and for most natural numbers N in between 1 and n, the number $\omega(N)$ of distinct prime divisors of N is close to $\log\log n$ in the sense that their difference is smaller than $\psi(n)\sqrt{\log\log n}$, which is relatively small compared to $\log\log n$. Of course, $\log\log n$ itself grows very slowly, so the effect becomes noticeable only for very large n.

The proof of the Hardy–Ramanujan theorem will rely on Chebyshev's inequality and the following result that we will prove below.

Theorem 2.6 (Mertens' first and second theorem). *The following holds as $n \to \infty$,*

$$\sum_{p \leq n} \frac{\log p}{p} = \log n + \mathscr{O}(1), \tag{2.42}$$

$$\sum_{p \leq n} \frac{1}{p} = \log\log n + \mathscr{O}(1), \tag{2.43}$$

where the sum is over prime numbers $p \leq n$.

First, let us prove the Hardy–Ramanujan theorem.

Proof (Theorem 2.5). Let N be a uniform random variable on the set $\{1,\ldots,n\}$,

$$\mathbb{P}(N=k)=\frac{1}{n} \text{ for } k=1,\ldots,n.$$

Then, what we want to prove is that

$$\mathbb{P}\Big(\big|\omega(N)-\log\log n\big|>\psi(n)\sqrt{\log\log n}\Big)\to 0.$$

As usual, we will represent $\omega(N)$, which counts the number of prime divisors of N, as the sum of indicators

$$\omega(N)=\sum_{p\le n}\mathrm{I}(p\mid N),$$

where the sum is over prime numbers $p\le n$.

First of all, let $m=\sqrt{n}$ and consider a modified sum

$$\omega_m(N)=\sum_{p\le m}\mathrm{I}(p\mid N),$$

where we do not count prime divisor(s) of N bigger that $\sqrt{n}$. There could be no more than one such prime divisor because, if there were two, their product would exceed n. This means that $|\omega_m(N)-\omega(N)|\le 1$ and it is enough to prove that

$$\mathbb{P}\Big(\big|\omega_m(N)-\log\log n\big|>\psi(n)\sqrt{\log\log n}\Big)\to 0.$$

Expectation of one indicator $\mathrm{I}(p\mid N)$ in the above sum is

$$\mathbb{E}\,\mathrm{I}(p\mid N)=\frac{1}{n}\sum_{k=1}^{n}\mathrm{I}(p\mid k)=\frac{1}{n}\Big\lfloor\frac{n}{p}\Big\rfloor,$$

because the number of $k\le n$ divisible by p equals $\lfloor\frac{n}{p}\rfloor$. Therefore,

$$\mathbb{E}\omega_m(N) = \sum_{p\le m} \mathbb{E}\,\mathrm{I}(p \mid N) = \sum_{p\le m} \frac{1}{n}\left\lfloor \frac{n}{p} \right\rfloor .$$

Since $\lfloor \frac{n}{p} \rfloor \le \frac{n}{p} \le \lfloor \frac{n}{p} \rfloor + 1$, the last sum

$$\sum_{p\le m} \frac{1}{n}\left\lfloor \frac{n}{p} \right\rfloor = \sum_{p\le m} \frac{1}{p} + \mathscr{O}(1)\,.$$

(Or even $\mathscr{o}(1)$.) Using the second Mertens' theorem (2.43),

$$\begin{aligned}\mathbb{E}\omega_m(N) &= \sum_{p\le m} \frac{1}{p} + \mathscr{O}(1) &(2.44)\\ &= \log\log\sqrt{n} + \mathscr{O}(1) = \log\log n + \mathscr{O}(1)\,.\end{aligned}$$

Next, let us compute the variance of $\omega_m(N)$. Using the formula (2.8) in Section 2.1 for the variance of the sum in terms of the sum of covariances,

$$\begin{aligned}\mathrm{Var}\big(\omega_m(N)\big) = &\sum_{p\le m} \mathrm{Var}\Big(\mathrm{I}(p \mid N)\Big) &(2.45)\\ &+ \sum_{1\le p\ne q\le m} \mathrm{Cov}\Big(\mathrm{I}(p \mid N), \mathrm{I}(q \mid N)\Big).\end{aligned}$$

First of all, since the indicator $\mathrm{I}(p \mid N)$ is a Bernoulli $B(x)$ random variable with

$$x = \mathbb{E}\,\mathrm{I}(p \mid N) = \frac{1}{n}\left\lfloor \frac{n}{p} \right\rfloor = \frac{1}{p} + \mathscr{O}\left(\frac{1}{n}\right),$$

and the variance of Bernoulli $B(x)$ is $x(1-x)$, the variance of this indicator is

$$\mathrm{Var}\Big(\mathrm{I}(p \mid N)\Big) = \frac{1}{p}\Big(1 - \frac{1}{p}\Big) + \mathscr{O}\left(\frac{1}{n}\right).$$

Then the sum of variances is

$$\begin{aligned}\sum_{p\le m}\operatorname{Var}\Big(\mathrm{I}(p\mid N)\Big) &= \sum_{p\le m}\frac{1}{p}-\sum_{p\le m}\frac{1}{p^2}+\mathscr{O}(1)\\ &= \log\log n+\mathscr{O}(1)\,, \qquad (2.46)\end{aligned}$$

where for the first sum we again used the second Mertens' theorem (2.43), and the second sum is bounded by

$$\sum_{j=1}^{\infty}\frac{1}{j^2}<\infty.$$

Next, let us compute the covariances in (2.45),

$$\begin{aligned}&\operatorname{Cov}\Big(\mathrm{I}(p\mid N),\mathrm{I}(q\mid N)\Big)\\ &=\mathbb{E}\,\mathrm{I}\big(p\mid N,q\mid N\big)-\mathbb{E}\,\mathrm{I}(p\mid N)\mathbb{E}\,\mathrm{I}(q\mid N).\end{aligned}$$

The last two expectations we already computed, and the first expectation is

$$\mathbb{E}\,\mathrm{I}\big(p\mid N,q\mid N\big)=\frac{1}{n}\sum_{k=1}^{n}\mathrm{I}\big(p\mid k,q\mid k\big)=\frac{1}{n}\Big\lfloor\frac{n}{pq}\Big\rfloor,$$

because the number of $k\le n$ divisible by both prime numbers p and q equals to the number of $k\le n$ divisible by pq, which equals $\lfloor\frac{n}{pq}\rfloor$. Therefore,

$$\begin{aligned}\operatorname{Cov}\Big(\mathrm{I}(p\mid N),\mathrm{I}(q\mid N)\Big) &= \frac{1}{n}\Big\lfloor\frac{n}{pq}\Big\rfloor-\frac{1}{n}\Big\lfloor\frac{n}{p}\Big\rfloor\cdot\frac{1}{n}\Big\lfloor\frac{n}{q}\Big\rfloor\\ &\le\frac{1}{n}\frac{n}{pq}-\frac{1}{n}\Big(\frac{n}{p}-1\Big)\cdot\frac{1}{n}\Big(\frac{n}{q}-1\Big)\\ &=\frac{1}{n}\Big(\frac{1}{p}+\frac{1}{q}\Big)-\frac{1}{n^2}\le\frac{1}{n}\Big(\frac{1}{p}+\frac{1}{q}\Big).\end{aligned}$$

This shows that the events that a random integer $N\le n$ is divisible by a prime p or prime q are almost uncorrelated. In particular, the sum of covariances in (2.45) is bounded by

$$\begin{aligned}
&\sum_{1\le p\ne q\le m} \mathrm{Cov}\Big(\mathrm{I}(p \mid N), \mathrm{I}(q \mid N)\Big) \\
&\le \sum_{1\le p\ne q\le m} \frac{1}{n}\Big(\frac{1}{p}+\frac{1}{q}\Big) \\
&\le \frac{m}{n}\sum_{p\le m}\frac{1}{p}+\frac{m}{n}\sum_{q\le m}\frac{1}{q} \\
&= \frac{2}{\sqrt{n}}\Big(\log\log n+\mathscr{O}(1)\Big)=o(1)\,, \qquad (2.47)
\end{aligned}$$

where in the last line we again used the second Mertens' theorem (2.43) and the fact that $m=\sqrt{n}$. Combining (2.46) and (2.47), the variance of the sum in (2.45) is

$$\mathrm{Var}\big(\omega_m(N)\big)=\log\log n+\mathscr{O}(1)\,. \qquad (2.48)$$

By Chebyshev's inequality (2.11),

$$\begin{aligned}
&\mathbb{P}\Big(\big|\omega_m(N)-\log\log n\big|>\psi(n)\sqrt{\log\log n}\Big) \\
&\le \frac{\log\log n+\mathscr{O}(1)}{\psi^2(n)\log\log n}\to 0,
\end{aligned}$$

since $\psi(n)\to\infty$. This finishes the proof. □

It remains to prove Mertens' theorem. In the proof, we will need the following result.

Lemma 2.4. *For any integer $n\ge 1$,*

$$\prod_{p\le n} p\le 2^{4n}, \qquad (2.49)$$

where the product is taken over primes $p\le n$.

One can prove a better inequality using the prime number theorem, but here we will use only elementary arguments.

Proof. Let us consider the following binomial coefficient

$$\binom{2n}{n} = \frac{(n+1)\cdots(2n)}{1\cdots n}.$$

All the primes between $n+1 \le p \le 2n$ are in the numerator and, since this binomial coefficient is integer, other factors in the numerator will cancel with $n!$ in the denominator, which implies that

$$\prod_{n+1\le p\le 2n} p \le \binom{2n}{n}.$$

This binomial coefficient is smaller than 2^{2n} because it is just one term in

$$2^{2n} = (1+1)^{2n} = \sum_{k=0}^{2n} \binom{2n}{k} 1^k 1^{2n-k},$$

so we showed that $\prod_{n+1\le p\le 2n} p \le 2^{2n}$. For $n = 2^k$, this gives

$$\prod_{2^k+1\le p\le 2^{k+1}} p \le 2^{2^{k+1}}.$$

Multiplying these equations over $k = 0,\ldots,m$, we get

$$\prod_{p\le 2^{m+1}} p \le 2^{2+2^2+\ldots+2^{m+1}} \le 2^{2^{m+2}}.$$

Given $n \ge 1$, let us take m such that $2^m < n \le 2^{m+1}$. Then, this equation implies

$$\prod_{p\le n} p \le \prod_{p\le 2^{m+1}} p \le 2^{2^{m+2}} \le 2^{4n},$$

which finishes the proof. □

Finally, we can prove Mertens' theorem.

Proof (Theorem 2.6). (Part 1) Let us consider the factorial $n! = 1\cdot 2\cdots(n-1)\cdot n$ and let us count how many times the factor p appears in this product, for a given prime number $p \le n$. First of all, each the following numbers

$$p, 2p, 3p, \ldots, \left\lfloor \frac{n}{p} \right\rfloor p$$

contains a factor p. This gives $\lfloor \frac{n}{p} \rfloor$ factors of p, but some of these numbers also contain at least two factors of p, that is, they are divisible by p^2. These numbers are

$$p^2, 2p^2, 3p^2, \ldots, \left\lfloor \frac{n}{p^2} \right\rfloor p^2.$$

Since the first factor of p was already counted, this gives additional $\lfloor \frac{n}{p^2} \rfloor$ factors of p. However, some of these are divisible by p^3, giving us another $\lfloor \frac{n}{p^3} \rfloor$ factors of p, and if we continue in the same fashion, the total number of factors of p in $n!$ is

$$f(n,p) := \left\lfloor \frac{n}{p} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \left\lfloor \frac{n}{p^3} \right\rfloor + \ldots + \left\lfloor \frac{n}{p^k} \right\rfloor, \tag{2.50}$$

where p^k is the last power of p less than or equal to n. If we write $n!$ using prime number decomposition, we showed that $n! = \prod_{p \le n} p^{f(n,p)}$ and, taking logarithms,

$$\log(n!) = \sum_{p \le n} f(n,p) \log p.$$

To see where we are going, notice that by Stirling's formula (1.101) in Section 1.5,

$$\log(n!) = n \log n + \mathscr{O}(n),$$

and dividing both sides by n, we get

$$\log n + \mathscr{O}(1) = \sum_{p \le n} \frac{f(n,p)}{n} \log p. \tag{2.51}$$

Since

$$\frac{f(n,p)}{n} = \frac{1}{n} \left\lfloor \frac{n}{p} \right\rfloor + \frac{1}{n} \left\lfloor \frac{n}{p^2} \right\rfloor + \ldots + \frac{1}{n} \left\lfloor \frac{n}{p^k} \right\rfloor,$$

this is starting to look like (2.42), because the first term in this sum looks like $1/p$. More precisely, the difference can be bounded by

$$\left|\frac{f(n,p)}{n}-\frac{1}{p}\right| \leq \frac{1}{n}+\frac{1}{p^2}+\ldots+\frac{1}{p^k} \leq \frac{1}{n}+\frac{2}{p^2},$$

using the geometric series formula. Adding up over $p \leq n$,

$$\left|\sum_{p\leq n}\frac{f(n,p)}{n}\log p-\sum_{p\leq n}\frac{\log p}{p}\right| \leq \frac{1}{n}\sum_{p\leq n}\log p+\sum_{p\leq n}\frac{\log p}{p^2}.$$

By the previous lemma, the first term

$$\frac{1}{n}\sum_{p\leq n}\log p=\frac{1}{n}\log\prod_{p\leq n}p \leq \frac{1}{n}\log 2^{4n} \leq 4\log 2.$$

The second term is bounded by $\sum_{j=1}^{\infty}\log j/j^2 < \infty$, so the sum of two terms is bounded by a constant. Therefore, (2.51) implies that

$$\log n+\mathscr{O}(1)=\sum_{p\leq n}\frac{\log p}{p}+\mathscr{O}(1)$$

and finishes the proof of the first equation (2.42).

(Part 2) The second equation (2.43) follows from the first one. Equation (2.42) can also be written for all real $x \geq 2$,

$$\sum_{p\leq x}\frac{\log p}{p}=\log x+\mathscr{O}(1)\,, \tag{2.52}$$

instead of only integer $x=n$, because rounding to the nearest integer changes both sides by at most a constant. Next, we will use that, for $p \leq n$,

$$\int_p^n \frac{dt}{t\log^2 t}=-\frac{1}{\log t}\Big|_p^n=\frac{1}{\log p}-\frac{1}{\log n}.$$

Using this, we can write

$$\begin{aligned}\sum_{p\le n}\frac{1}{p} &= \sum_{p\le n}\frac{\log p}{p}\cdot\frac{1}{\log p} = \sum_{p\le n}\frac{\log p}{p}\Big(\frac{1}{\log n}+\int_p^n\frac{dt}{t\log^2 t}\Big)\\ &= \frac{1}{\log n}\sum_{p\le n}\frac{\log p}{p}+\sum_{p\le n}\frac{\log p}{p}\int_p^n\frac{dt}{t\log^2 t},\end{aligned}$$

By (2.52), the first term is $\mathscr{O}(1)$. The second term can be rewritten as

$$\begin{aligned}\sum_{p\le n}\frac{\log p}{p}\int_2^n \mathrm{I}(p\le t)\frac{dt}{t\log^2 t} &= \int_2^n\sum_{p\le n}\frac{\log p}{p}\mathrm{I}(p\le t)\frac{dt}{t\log^2 t}\\ &= \int_2^n\sum_{p\le t}\frac{\log p}{p}\frac{dt}{t\log^2 t}.\end{aligned}$$

Using (2.52) again, this equals

$$\begin{aligned}&\int_2^n\big(\log t+\mathscr{O}(1)\big)\frac{dt}{t\log^2 t} = \int_2^n\frac{dt}{t\log t}+\mathscr{O}(1)\\ &= \log\log t\Big|_2^n+\mathscr{O}(1) = \log\log n+\mathscr{O}(1).\end{aligned}$$

This finishes the proof of (2.43). □

Exercise 2.5.1. If we denote $\tilde{\mathrm{I}}(p\mid N)=\mathrm{I}(p\mid N)-\mathbb{E}\,\mathrm{I}(p\mid N)$, compute the 3-point correlation

$$\mathbb{E}\Big[\tilde{\mathrm{I}}(p\mid N)\,\tilde{\mathrm{I}}(q\mid N)\,\tilde{\mathrm{I}}(r\mid N)\Big]$$

for three distinct prime numbers p,q, and r.

Chapter 3
Exponential Inequalities

3.1 Hoeffding Inequality

Using Chebyshev's inequality, we showed that if i.i.d. random variables $X_1, \dots, X_n$ have finite variance,

$$\sigma^2 = \mathbb{E}X_1^2 < \infty,$$

then their average

$$\bar{X}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is close to the expectation $p = \mathbb{E}X_1$, with high probability. This is known as the law of large numbers. In this chapter we will prove several quantitatively stronger results under stronger assumptions on the random variables. More precisely, instead of only assuming finite variance we will work with random variables such that

$$\mathbb{E}e^{\lambda X} < \infty \text{ for all } \lambda \in \mathbb{R}. \tag{3.1}$$

For example, this holds if a random variable X is bounded by a constant. To take advantage of this, we will be using the exponential form of Chebyshev's inequality, known as Markov's inequality.

Lemma 3.1 (Markov's inequality). *For any $\lambda \geq 0$,*

$$\mathbb{P}\big(X \geq t\big) \leq e^{-\lambda t}\mathbb{E}e^{\lambda X}, \tag{3.2}$$

assuming that $\mathbb{E}e^{\lambda X} < \infty$.

Proof. For positive $\lambda > 0$, the inequality $X \geq t$ is equivalent to $e^{\lambda X} \geq e^{\lambda t}$, so

$$\mathbb{P}\big(X \geq t\big) = \mathbb{P}\big(e^{\lambda X} \geq e^{\lambda t}\big) \leq \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda t}},$$

by the usual Chebyshev's inequality, since $e^{\lambda X}$ is positive. If $\lambda = 0$ then the right hand side of (3.2) is equal to 1, so the inequality still holds. □

First, we will consider independent flips of a fair coin. It will be convenient to rescale the usual values $\{0,1\}$ of the Bernoulli random variable to $\{-1,1\}$. In other words, we consider i.i.d. random variables $\varepsilon_1,\ldots,\varepsilon_n$ such that

$$\mathbb{P}\big(\varepsilon_n = 1\big) = \mathbb{P}\big(\varepsilon_n = -1\big) = \frac{1}{2}. \tag{3.3}$$

These are usually called *Rademacher* random variables. The first classical exponential inequality that we will prove is the following.

Theorem 3.1 (Hoeffding's inequality). *For any $t \geq 0$ and any constant $a_1,\ldots,a_n \in \mathbb{R}$,*

$$\mathbb{P}\Big(\sum_{i=1}^{n}\varepsilon_i a_i \geq t\Big) \leq \exp\Big(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}\Big) \tag{3.4}$$

and

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\varepsilon_i a_i\Big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}\Big). \tag{3.5}$$

Proof. We begin by using Markov's inequality (3.2),

$$\mathbb{P}\Big(\sum_{i=1}^{n} \varepsilon_i a_i \geq t\Big) \leq e^{-\lambda t}\mathbb{E}e^{\lambda \sum_{i=1}^{n} \varepsilon_i a_i} = e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}e^{\lambda \varepsilon_i a_i},$$

where in the last step we used Lemma 1.6 and (1.86). One factor in this product equals

$$\mathbb{E}e^{\lambda \varepsilon_i a_i} = \frac{1}{2}e^{\lambda a_i} + \frac{1}{2}e^{-\lambda a_i} = \cosh(\lambda a_i).$$

Next, we will use the inequality $\cosh(x) \leq e^{x^2/2}$, which can be seen by comparing the Taylor series,

$$\cosh(x) = \frac{e^x + e^{-x}}{2} = \sum_{k=0}^{\infty} \frac{1}{(2k)!}x^{2k} \leq \sum_{k=0}^{\infty} \frac{1}{2^k k!}x^{2k} = e^{x^2/2}.$$

This implies that $\mathbb{E}e^{\lambda \varepsilon_i a_i} \leq e^{\lambda^2 a_i^2/2}$ and

$$\mathbb{P}\Big(\sum_{i=1}^{n} \varepsilon_i a_i \geq t\Big) \leq \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{n} a_i^2\Big).$$

This holds for any $\lambda \geq 0$, so we can minimize over $\lambda \geq 0$. Taking derivative in λ, the critical point is equal to

$$-t + \lambda \sum_{i=1}^{n} a_i^2 = 0 \Longleftrightarrow \lambda = \frac{t}{\sum_{i=1}^{n} a_i^2},$$

which is nonnegative, because $t \geq 0$. Plugging back this λ, we get (3.4).

To prove (3.5), let us notice that, by symmetry, $-\varepsilon_i$ has the same distribution as ε_i, because (3.3) implies that

$$\mathbb{P}\big(-\varepsilon_n = 1\big) = \mathbb{P}\big(-\varepsilon_n = -1\big) = \frac{1}{2}. \tag{3.6}$$

This means that we can apply (3.4) to $(-\varepsilon_i)$, which reads

$$\mathbb{P}\Big(\sum_{i=1}^{n} \varepsilon_i a_i \leq -t\Big) \leq \exp\Big(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}\Big). \tag{3.7}$$

Since

$$\Big\{|\sum_{i=1}^{n}\varepsilon_i a_i| \geq t\Big\} = \Big\{\sum_{i=1}^{n}\varepsilon_i a_i \geq t\Big\}\bigcup\Big\{\sum_{i=1}^{n}\varepsilon_i a_i \leq -t\Big\},$$

the union bound in the Exercise 1.1.1 implies (3.5). □

Example 3.1.1 (Law of large numbers for a fair coin). If in the inequality (3.5) we take $a_i = \frac{1}{n}$, we get that

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Big| \geq t\Big) \leq 2e^{-nt^2/2}.$$

Since the random variables $X_i := (\varepsilon_i + 1)/2$ take values 0 and 1 with probability $1/2$ each, $X_1, \ldots, X_n$ are i.i.d. Bernoulli $B(1/2)$ corresponding to independent flips of a fair coin. Since $\varepsilon_i = 2X_i - 1$, we can rewrite the above inequality as

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{2}\Big| \geq \frac{t}{2}\Big) \leq 2e^{-nt^2/2}.$$

Making the change of variables $t = 2\varepsilon$, we get

$$\mathbb{P}\big(|\bar{X}_n - 0.5| \geq \varepsilon\big) \leq 2e^{-2n\varepsilon^2}. \tag{3.8}$$

Compare this with the application of Chebyshev's inequality in (2.14) in Section 2.2,

$$\mathbb{P}\big(|\bar{X}_n - 0.5| \geq \varepsilon\big) \leq \frac{1}{4n\varepsilon^2}. \tag{3.9}$$

We can see that the dependence on n is dramatically better in Hoeffding's inequality, which is crucial for its many applications. For the sake of illustration, if we take $n = 10,000$ and $\varepsilon = 0.02$, the two inequalities give

$$\mathbb{P}\big(|\bar{X}_{10,000} - 0.5| \geq 0.02\big) \leq 0.00067 \text{ vs. } 0.0625.$$

When we get to the Central Limit Theorem, we will see that the dependence on n and ε in Hoeffding's inequality is, in fact, nearly optimal. □

Example 3.1.2 (Higher moments of Rademacher sums). The following consequence of Hoeffding's inequality will be crucial in the application in the next section. Let us consider $a_1,\ldots,a_n \in \mathbb{R}$ such that

$$a_1^2+\ldots+a_n^2=1 \tag{3.10}$$

and let us consider the random variable

$$X := \varepsilon_1 a_1+\ldots+\varepsilon_n a_n. \tag{3.11}$$

By symmetry, its odd moments are equal to zero,

$$\mathbb{E}X^{2k+1}=\mathbb{E}\big(\varepsilon_1 a_1+\ldots+\varepsilon_n a_n\big)^{2k+1}=0$$

(prove this carefully). Its second moment is equal to

$$\mathbb{E}X^2=a_1^2+\ldots+a_n^2=1. \tag{3.12}$$

Now we will see how Hoeffding's inequality allows us to control even moments $\mathbb{E}X^{2k}$. By Lemma 1.3 and Exercise 1.2.5 in Section 1.2,

$$\mathbb{E}X^{2k}=\int_0^\infty 2kt^{2k-1}\mathbb{P}(|X|\geq t)\,dt \leq \int_0^\infty 4kt^{2k-1}e^{-t^2/2}\,dt,$$

where we used Hoeffding's inequality $\mathbb{P}(|X|\geq t)\leq 2e^{-t^2/2}$. Making the change of variables $t=\sqrt{2u}$,

$$\mathbb{E}X^{2k}\leq k2^{k+1}\int_0^\infty u^{k-1}e^{-u}\,du.$$

The last integral is the Gamma function $\Gamma(k)=(k-1)!$, which can be easily checked by induction on k, using integration by parts (see Section 4.4). Therefore, we showed that, for $k\geq 1$,

$$\mathbb{E}X^{2k} = \mathbb{E}\big(\varepsilon_1 a_1 + \ldots + \varepsilon_n a_n\big)^{2k} \le 2^{k+1}k!, \tag{3.13}$$

if the condition $a_1^2 + \ldots + a_n^2 = 1$ holds.

Let us now consider the random variable

$$Y = X^2 - 1 = \big(\varepsilon_1 a_1 + \ldots + \varepsilon_n a_n\big)^2 - 1. \tag{3.14}$$

In the next section, we will need the following consequence of the estimate (3.13), which will be used in a very important geometric application.

Lemma 3.2. *If (3.10) holds then, for any* $0 \le \lambda \le \frac{1}{4}$,

$$\mathbb{E}e^{\lambda Y} \le e^{16\lambda^2} \text{ and } \mathbb{E}e^{-\lambda Y} \le e^{16\lambda^2}. \tag{3.15}$$

Proof. Let us start with the following simple observation: for all $x \in \mathbb{R}$,

$$e^x \le 1 + x + \frac{x^2}{2} + \sum_{k=3}^{\infty} \frac{(x_+)^k}{k!}, \tag{3.16}$$

where $x_+ = \max(x, 0)$. If $x \ge 0$ then $x_+ = x$ and the inequality is actually an equality, since the right hand side becomes the Taylor series of e^x. If $x \le 0$ then $x_+ = 0$ and the inequality becomes $e^x \le 1 + x + x^2/2$, which we leave as a simple exercise below. Using (3.16), we can write

$$\mathbb{E}e^{\lambda Y} \le 1 + \lambda \mathbb{E}Y + \frac{\lambda^2}{2}\mathbb{E}Y^2 + \sum_{k=3}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(Y_+)^k.$$

First of all, $\mathbb{E}Y = \mathbb{E}X^2 - 1 = 0$ by (3.12). Next,

$$\mathbb{E}Y^2 = \mathbb{E}X^4 - 2\mathbb{E}X^2 + 1 = \mathbb{E}X^4 - 1 \le \mathbb{E}X^4.$$

Also, since $Y_+ = \max(X^2 - 1, 0) \le X^2$, we have $\mathbb{E}(Y_+)^k \le \mathbb{E}X^{2k}$ and, therefore,

$$\mathbb{E}e^{\lambda Y} \le 1 + \frac{\lambda^2}{2}\mathbb{E}X^4 + \sum_{k=3}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}X^{2k} = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}X^{2k}.$$

Here, we finally use the moment estimate (3.13),

$$\begin{aligned}\mathbb{E}e^{\lambda Y} &\le 1+\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}2^{k+1}k!\\ &= 1+\sum_{k=2}^{\infty}\lambda^k 2^{k+1} = 1+\frac{8\lambda^2}{1-2\lambda} \le 1+16\lambda^2.\end{aligned}$$

since this is a geometric series and $2\lambda \le \frac{1}{2}$ by assumption. To finish the proof of the first inequality in (3.15), it remains to use that $1+x \le e^x$. The proof of the second inequality is almost the same and we leave it as an exercise. □

Exercise 3.1.1. Estimate the probability that in $100,000$ flips of a fair coin, the number of Heads will deviate from $50,000$ by more that 500.

Exercise 3.1.2. Show that, for $k \ge 1$,

$$\mathbb{E}\big(\varepsilon_1 a_1+\ldots+\varepsilon_n a_n\big)^{2k} \le 2^{k+1}k!(a_1^2+\ldots+a_n^2)^k$$

for any $a_1,\ldots,a_n \in \mathbb{R}$.

Exercise 3.1.3. Show that

$$e^x \le 1+x+\frac{x^2}{2} \text{ for } x \le 0.$$

Exercise 3.1.4. Prove the second inequality in (3.15).

Exercise 3.1.5 (*). Suppose that $X_1, X_1', \ldots, X_n, X_n'$ are independent and, for all $i \le n$, X_i and X_i' have the same distribution. Prove that

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - X_i') > \big(2t\sum_{i=1}^{n}(X_i - X_i')^2\big)^{1/2}\Big) \le e^{-t}.$$

Hint: think about a way to introduce Rademacher random variables ε_i into the problem and then use Hoeffding's inequality.

3.2 Johnson–Lindenstrauss Lemma

In this section, we will give one classical application of the results in the previous section. Let us consider $N \geq 1$ and a set

$$V = \big\{v_1, \ldots, v_m\big\} \subseteq \mathbb{R}^N \tag{3.17}$$

of m points in $\mathbb{R}^N$. The dimension N here can be arbitrarily large, and we should think of the number of points m as also being large. The Johnson–Lindenstrauss lemma states that there exists a linear map from $\mathbb{R}^N$ into Euclidean space $\mathbb{R}^n$ of possibly much lower dimension n that preserves the distances between all the points in the set V up to a small relative error. This is called a *low-distortion embedding*. It was discovered in a work on functional analysis, but it found many applications to computational algorithms in various fields as a preprocessing step to reduce the dimensionality of high-dimensional data. Here is the precise statement.

Theorem 3.2 (Johnson–Lindenstrauss lemma). *Given m points $V = \{v_1, \ldots, v_m\}$ in $\mathbb{R}^N$, any $\varepsilon \in (0,1)$, and*

$$n > \frac{128}{\varepsilon^2} \log m, \tag{3.18}$$

there exists a linear map $f\colon \mathbb{R}^N \to \mathbb{R}^n$ such that

$$\sqrt{1-\varepsilon} \leq \frac{\|f(v_k) - f(v_\ell)\|}{\|v_k - v_\ell\|} \leq \sqrt{1+\varepsilon} \tag{3.19}$$

for all $1 \leq k < \ell \leq m$, where $\|\cdot\|$ denotes the Euclidean norm.

The dimension n in (3.18) that we are allowed to choose depends on the distortion parameter ε and the number of points m, but it does not depend on N. The dependence on m is logarithmic, so the dimension can be relatively small even when the number of points is very large. The constant 128 is not optimized here and can, for example, be improved to 8.

There are various proofs of the Johnson–Lindenstrauss lemma, but most of them use probabilistic constructions. We will consider a random matrix

$$\mathscr{E} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} & \dots & \varepsilon_{1N} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} & \dots & \varepsilon_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \varepsilon_{n3} & \dots & \varepsilon_{nN} \end{bmatrix}, \tag{3.20}$$

where the entries ε_{ij} are i.i.d. Rademacher random variables, and show that the linear map

$$f(x) = \frac{1}{\sqrt{n}}\mathscr{E}x \tag{3.21}$$

satisfies the low-distortion property (3.19) with positive probability. This means that there exists at least one matrix with ±1 entries that satisfies (3.19), if the dimension n is properly chosen. This is another example of application the probabilistic method idea that we saw in Section 1.1.

First of all, let us denote, for $1 \le k < \ell \le m$,

$$a_{k\ell} = \frac{v_k - v_\ell}{\|v_k - v_\ell\|}. \tag{3.22}$$

Squaring (3.19), it can be rewritten for f in (3.21) as

$$1 - \varepsilon \le \frac{1}{n}\|\mathscr{E}a_{k\ell}\|^2 \le 1 + \varepsilon \tag{3.23}$$

for all $1 \le k < \ell \le m$. Since we rescaled so that the length $\|a_{k\ell}\| = 1$, all the vectors $a_{k\ell}$ belong to the unit sphere in $\mathbb{R}^N$, $a_{k\ell} \in S^{N-1}$.

For the moment, let us consider an arbitrary vector

$$a = (a_1, \dots, a_N) \in S^{N-1},$$

let us denote the i^{th} coordinate of $\mathscr{E}a$ by

$$X_i(a) := (\mathscr{E}a)_i = \varepsilon_{i1}a_1 + \ldots + \varepsilon_{iN}a_N, \tag{3.24}$$

and let us denote

$$Y_i(a) := X_i^2(a) - 1 = \big(\varepsilon_{i1}a_1 + \ldots + \varepsilon_{iN}a_N\big)^2 - 1. \tag{3.25}$$

With this notation, (3.23) can be rewritten as

$$1 - \varepsilon \le \frac{1}{n}\sum_{i=1}^{n} X_i^2(a) \le 1 + \varepsilon,$$

or

$$\Big|\frac{1}{n}\sum_{i=1}^{n} Y_i(a)\Big| \le \varepsilon.$$

This looks like the law of large numbers for $Y_1(a), \ldots, Y_n(a)$, because $\mathbb{E}Y_i(a) = 0$, and in the last section we prepared what will be needed to control the probability of this event. Namely, we showed in Lemma 3.2 that, for any $0 \le \lambda \le \frac{1}{4}$,

$$\mathbb{E}e^{\lambda Y_i(a)} \le e^{16\lambda^2} \text{ and } \mathbb{E}e^{-\lambda Y_i(a)} \le e^{16\lambda^2}. \tag{3.26}$$

In addition, observe that, by the Grouping Lemma 1.5 in Section 1.4, the random variables $Y_1(a), \ldots, Y_n(a)$ are independent, because they are functions of different rows of the matrix $\mathscr{E}$, which are all independent by construction. This independence will be crucial in the proof below.

As a consequence of the estimate (3.26) on the exponential moments of $Y_i(a)$, we obtain the following exponential inequality.

Lemma 3.3. *For any $\varepsilon \in (0,1)$ and $a = (a_1, \ldots, a_N) \in S^{N-1}$,*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} Y_i(a)\Big| \ge \varepsilon\Big) \le 2e^{-n\varepsilon^2/64}. \tag{3.27}$$

Proof. In one direction, using Markov's inequality, for any $0 \le \lambda \le \frac{1}{4}$,

$$\begin{aligned}
\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i(a) \geq \varepsilon\Big) &= \mathbb{P}\Big(\sum_{i=1}^{n} Y_i(a) \geq n\varepsilon\Big) \\
&\leq e^{-n\lambda\varepsilon}\mathbb{E}e^{\lambda\sum_{i=1}^{n} Y_i(a)} \\
\big(\text{by independence}\big) \quad &= e^{-n\lambda\varepsilon}\prod_{i=1}^{n}\mathbb{E}e^{\lambda Y_i(a)} \\
\big(\text{using (3.26)}\big) \quad &\leq e^{-n\lambda\varepsilon+16n\lambda^2}.
\end{aligned}$$

Since this upper bound holds for any $0 \leq \lambda \leq \frac{1}{4}$, we can minimize it over such λ. Critical point is equal to $\lambda = \frac{\varepsilon}{32}$ and, since it belongs to $(0, \frac{1}{4})$, plugging it in we get

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i(a) \geq \varepsilon\Big) \leq e^{-n\varepsilon^2/64}.$$

Exactly the same calculation with $-Y_i(a)$ instead of $Y_i(a)$ gives

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i(a) \leq -\varepsilon\Big) = \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} (-Y_i(a)) \geq \varepsilon\Big) \leq e^{-n\varepsilon^2/64},$$

and the claim (3.27) follows by the union bound, just like in the proof of Hoeffding's inequality. □

We are now ready to prove the Johnson–Lindenstrauss Lemma.

Proof (Theorem 3.2). Recall that we would like to show that there exists a matrix $\mathscr{E}$ such that (3.23) holds for all $1 \leq k < \ell \leq m$ or, in other words,

$$\Big|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{k\ell})\Big| \leq \varepsilon$$

holds for all $1 \leq k < \ell \leq m$. The complement of this is

$$\begin{aligned}
&\left\{\exists k < \ell \text{ such that } \left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{k\ell})\right| \geq \varepsilon\right\} \\
&= \bigcup_{1\leq k<\ell\leq m} \left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{k\ell})\right| \geq \varepsilon\right\}.
\end{aligned}$$

By the union bound in Exercise 1.1.1 and previous lemma, the probability of this event is bounded by the sum of probabilities

$$\begin{aligned}
\sum_{1\leq k<\ell\leq m} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{k\ell})\right| \geq \varepsilon\right) &\leq \sum_{1\leq k<\ell\leq m} 2e^{-n\varepsilon^2/64} \\
&\leq m^2 e^{-n\varepsilon^2/64}.
\end{aligned}$$

Therefore, we showed that

$$\mathbb{P}\left(\forall\, 1 \leq k < \ell \leq m,\ \left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{k\ell})\right| \leq \varepsilon\right) \geq 1 - m^2 e^{-n\varepsilon^2/64}. \tag{3.28}$$

This probability is strictly positive if $m^2 e^{-n\varepsilon^2/64} < 1$, which is the same as

$$n > \frac{128}{\varepsilon^2}\log m.$$

This is precisely the assumption in (3.18). The fact that the probability is positive means that there exists $\mathscr{E}$ such that the low-distortion condition (3.23) is satisfied. □

Remark 3.1. In the above proof, the logarithmic dependence on the number of points came from the inequality

$$m^2 e^{-n\varepsilon^2/64} < 1,$$

which can be viewed as a competition between the large number of pairs (of order m^2) and the exponentially small estimate $e^{-n\varepsilon^2/64}$ on the probability that a given pair violates the low-distortion condition. This exponential dependence on n is exactly what resulted in the logarithmic dependence on m.

By increasing the dimension n, we can make sure that the random matrix $\frac{1}{\sqrt{n}}\mathscr{E}$ yields a low-distortion embedding not just with positive probability, but with high probability close to 1. We leave this observation as an exercise.

Exercise 3.2.1. Show that, for $\delta \in (0,1)$, if the dimension

$$n > \frac{128}{\varepsilon^2} \log \frac{m}{\sqrt{\delta}}$$

then the map $f(x) = \frac{1}{\sqrt{n}}\mathscr{E}x$ in the above proof satisfies (3.19) with probability at least $1-\delta$.

Suppose that we also want to control the distortion of the areas of triangles formed by any three points v_i, v_j and v_k. This can be done as follows. First, the areas of all triangles in the same (two-dimensional) plane are distorted by the same factor under a linear transformation. For each three points, we can add another point v_{ijk} in the same plane, which forms an *isosceles right* triangle with v_i, v_j. This means that, instead of m points, we now have $m + \binom{m}{3} \leq m^3$ points on the list. We can now find f as in Lemma 3.2 for this new enlarged list of points. Since the dependence on m is logarithmic, we lose a factor of 3 in the bound (3.18) for n. It remains to solve the following exercise to see that this procedure allows us to control the distortion of areas.

Exercise 3.2.2. Suppose that three points v_i, v_j and v_k in the setting of Lemma 3.2 form an isosceles right triangle Δ. Show that

$$\sqrt{1-5\varepsilon} \leq \frac{\text{Area}(f(\Delta))}{\text{Area}(\Delta)} \leq 1+\varepsilon.$$

3.3 Hoeffding–Chernoff Inequality

The proof of Hoeffding's inequality strongly relied on the symmetry of the Rademacher random variables and it does not apply, for example, to flips of a biased coin $B(p)$ with $p \neq 1/2$. In this section, we will prove another inequality, called the Hoeffding–Chernoff Inequality, which applies to any bounded random variables. Bounded random variables can be rescaled to take values in $[0,1]$, so it is enough to consider this case.

Let $X_1,\dots,X_n$ be independent random variables taking values in $[0,1]$, with common expectation $p = \mathbb{E}X_i \in [0,1]$. They do not need to have the same distribution. Consider the function of $p \in (0,1)$ and $q \in [0,1]$,

$$D(q\|p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}, \tag{3.29}$$

called the *Kullback-Leibler divergence*, or *relative entropy*. If $q=0$ or $q=1$, we understand that $0\cdot\log 0 = 0$, so the function is continuous in q. Notice that D is not symmetric in its arguments, $D(q\|p) \neq D(p\|q)$. First, let us check that this function is nonnegative.

Lemma 3.4. *For any* $p,q \in (0,1)$, $D(q\|p) \geq 0$. *It is equal to* 0 *only if* $p=q$.

Proof. Since $-\log x$ is convex and $-\log x \geq 1-x$ (its tangent line at $x=1$) with equality only at $x=1$,

$$\begin{aligned} D(q\|p) &= q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \\ &= q\Big(-\log\frac{p}{q}\Big) + (1-q)\Big(-\log\frac{1-p}{1-q}\Big) \\ &\geq q\Big(1-\frac{p}{q}\Big) + (1-q)\Big(1-\frac{1-p}{1-q}\Big) = 0, \end{aligned}$$

with equality only if $p=q$. □

Theorem 3.3 (Hoeffding–Chernoff's inequality). *Suppose that the random variables $X_1, \ldots, X_n$ are independent, each $X_i \in [0,1]$, and $p = \mathbb{E}X_i$. Then*

$$\mathbb{P}\big(\overline{X}_n \geq p+t\big) \leq e^{-nD(p+t\|p)} \tag{3.30}$$

for any $0 \leq t \leq 1-p$, *and*

$$\mathbb{P}\big(\overline{X}_n \leq p-t\big) \leq e^{-nD(1-p+t\|1-p)} \tag{3.31}$$

for any $0 \leq t \leq p$.

The dependence on n in these inequalities is exponential, just like in Hoeffding's inequality, which is much better than in Chebyshev's inequality. Also, notice that the probability on the left hand side of (3.30) is zero when $p+t > 1$, because the average $\overline{X}_n$ can never exceed 1, so $t \leq 1-p$ is not really a constraint there. Similarly, $t \leq p$ in (3.31) is not really a constraint.

The factor $D(p+t\|p)$ in the exponent in the upper tail bound (3.30) is not the same as $D(1-p+t\|1-p)$ in the exponent in the lower tail bound (3.31), so these bound are not symmetric. Heuristically, they should not be the same since, for example, when $p < 1/2$, there is 'less room' for the deviations below p between 0 and p than above p between p and 1. However, one can simplify both bounds and obtain more intuitive (although slightly worse) bounds

$$\mathbb{P}\big(\overline{X}_n \geq p+t\big) \leq e^{-2nt^2}, \tag{3.32}$$

and

$$\mathbb{P}\big(\overline{X}_n \leq p-t\big) \leq e^{-2nt^2}, \tag{3.33}$$

which we will leave as a simple exercise below.

Proof (Theorem 3.3). Using the convexity of e^t, we can write for $x \in [0,1]$:

$$e^{\lambda x} = e^{x\cdot\lambda + (1-x)\cdot 0} \le xe^{\lambda} + (1-x)e^{0} = 1 - x + xe^{\lambda}.$$

Since each $X_i \in [0,1]$, this implies that

$$\begin{aligned}\mathbb{E}e^{\lambda X_i} &\le 1 - \mathbb{E}X_i + \mathbb{E}X_i\, e^{\lambda}\\ &= 1 - p + pe^{\lambda}.\end{aligned}$$

Using this together with Markov's inequality, for any $\lambda \ge 0$,

$$\begin{aligned}\mathbb{P}\big(\overline{X}_n \ge p+t\big) &= \mathbb{P}\Big(\sum_{i=1}^{n} X_i \ge n(p+t)\Big)\\ &\le e^{-\lambda n(p+t)}\mathbb{E}e^{\lambda\sum_{i=1}^{n} X_i}\\ \big(\text{by independence}\big) &= e^{-\lambda n(p+t)}\prod_{i=1}^{n}\mathbb{E}e^{\lambda X_i}\\ &\le e^{-\lambda n(p+t)}\big(1-p+pe^{\lambda}\big)^{n}.\end{aligned}$$

This upper bound holds for any $\lambda \ge 0$ and we can minimize it over $\lambda \ge 0$. Equivalently, we can minimize its logarithm,

$$-\lambda n(p+t) + n\log\big(1-p+pe^{\lambda}\big) \longrightarrow \underset{\lambda\ge 0}{\text{minimize}}\,.$$

To find the critical point, we set the derivative equal to zero

$$-n(p+t) + \frac{npe^{\lambda}}{1-p+pe^{\lambda}} = 0$$

and solve for λ. We find that

$$e^{\lambda} = \frac{(1-p)(p+t)}{p(1-p-t)}.$$

This is ≥ 1 because $p+t \ge p$ and $1-p \ge 1-p-t$, which implies that $\lambda \ge 0$ as required in the application of Markov's inequality. Therefore, we can plug this back into the upper bound,

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^{n} X_i \geq n(p+t)\Big)\\
&\leq \Big[\Big(\frac{p(1-p-t)}{(1-p)(p+t)}\Big)^{p+t}\Big(1-p+\frac{(1-p)(p+t)}{1-p-t}\Big)\Big]^n\\
&= \Big[\Big(\frac{p}{p+t}\Big)^{p+t}\Big(\frac{1-p}{1-p-t}\Big)^{1-p-t}\Big]^n\\
&= \exp\Big[-n\Big((p+t)\log\frac{p+t}{p}+(1-p-t)\log\frac{1-p-t}{1-p}\Big)\Big],
\end{aligned}$$

which is precisely $e^{-nD(p+t\|p)}$. This finishes the proof of (3.30).

To prove (3.31), we can consider the random variables $Z_i = 1 - X_i \in [0,1]$ with $\mathbb{E}Z_i = 1-p$. If we apply (3.30) to these random variables, we get

$$\mathbb{P}\big(\bar{Z}_n \geq 1-p+t\big) \leq e^{-nD(1-p+t\|1-p)}.$$

Since

$$\bar{Z}_n = 1-\bar{X}_n \geq 1-p+t \Longleftrightarrow \bar{X}_n \leq p-t,$$

this proves (3.31). □

Let us describe one typical application of the Hoeffding–Chernoff inequality.

Example 3.3.1 (The generalization error of classification algorithms). The goal of a classification algorithm is to classify data. For example, given an image as an input, an algorithm outputs a label describing the image content, for example, ‘dog’ or ‘cat’. We can think of an image as a pair (X,Y), where X is the vector of pixel values and Y is one of possible labels. In applications, the algorithms will be given X and will output Y as some function $Y = f(X)$. To find a good candidate $f(x)$, however, requires that the algorithm has access to *training data*, which is a collection of examples (X_i,Y_i) for $i \leq n$, for which the label is known. Given the training data, an algorithm will attempt to find a function $f(x)$ such that the proportion of incorrectly predicted labels is as small as possible,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\big(Y_i \neq f(X_i)\big) \longrightarrow \text{minimize}\,.$$

Sometimes, a different *loss function* $L(Y, f(X))$ may be used instead of the indicator function above,

$$E_n(f) := \frac{1}{n}\sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) \longrightarrow \text{minimize}\,.$$

This average is called the *empirical error* of the classifier f, and different algorithms will minimize this error over different classes of functions $\mathscr{F}$ to find a good candidate f. Suppose an algorithm succeeds in finding a function f with small empirical error on the training examples provided. Does this mean that it will be good at classifying future examples with unknown labels?

One typical framework to study this question is to assume that the examples are coming from some (unknown) probability distribution $\mathbb{P}$ over a population of possible examples Ω, and that the training data (X_i, Y_i) for $i \leq n$ consists of i.i.d. observations from this distribution $\mathbb{P}$. Then, the ability to classify future examples, called the *generalization ability*, is measured by the *generalization error*, which is the expectation of the loss function

$$E(f) := \mathbb{E} L\big(Y, f(X)\big).$$

Does the small value of $E_n(f)$ mean that $E(f)$ is also small?

For a fixed function f this looks exactly like the law of large numbers, but the problem is that the choice of the function f itself depends on the training data, so we can not use the law of large numbers directly. Think of the example when an algorithm is allowed to use all possible functions f. Then the algorithm simply memorizes the training examples and there is no guarantee that it will have small generalization error on the future data, just like a student memorizing a bunch of examples but lacking understanding. This is called *overfitting* in Statistics.

Suppose that the algorithm chooses from a large, but finite, number of possible classifiers

$$\mathscr{F} = \{f_1, f_2, \ldots, f_N\}.$$

Let us also suppose that the loss function L takes values in $[0,1]$, for example, an indicator function above. Then, for a fixed $f \in \mathscr{F}$, the second inequality in (3.33) implies that

$$\mathbb{P}\big(E_n(f) \le E(f) - \varepsilon\big) \le e^{-2n\varepsilon^2}.$$

This shows that the complement $E(f) < E_n(f) + \varepsilon$ will have probability close to one, at least for large n. This means that, with high confidence, if the training error $E_n(f)$ was small, we expect the generalization error $E(f)$ also to be small.

Now we want to make the same statement simultaneously for all $f \in \mathscr{F}$. Since

$$\begin{aligned}&\Big\{\exists f \in \mathscr{F} \text{ such that } E_n(f) \le E(f) - \varepsilon\Big\}\\ &= \bigcup_{f \in \mathscr{F}} \Big\{E_n(f) \le E(f) - \varepsilon\Big\},\end{aligned}$$

the union bound implies that

$$\begin{aligned}&\mathbb{P}\Big(\exists f \in \mathscr{F}, E_n(f) \le E(f) - \varepsilon\Big)\\ &\le \sum_{f \in \mathscr{F}} \mathbb{P}\big(E_n(f) \le E(f) - \varepsilon\big) \le Ne^{-2n\varepsilon^2}.\end{aligned}$$

Then the complement of this event satisfies

$$\mathbb{P}\Big(\forall f \in \mathscr{F}, E_n(f) \ge E(f) - \varepsilon\Big) \ge 1 - Ne^{-2n\varepsilon^2},$$

or, written differently,

$$\mathbb{P}\Big(\forall f \in \mathscr{F}, E(f) \leq E_n(f) + \varepsilon\Big) \geq 1 - Ne^{-2n\varepsilon^2}. \quad (3.34)$$

Let us denote the last term by δ,

$$\delta := Ne^{-2n\varepsilon^2}. \quad (3.35)$$

If ε and δ are both small then, with probability close to 1, no matter what $f \in \mathscr{F}$ the classification algorithm selected, its generalization error $E(f)$ will be bounded by $E_n(f) + \varepsilon$, so it will be small if the training error $E_n(f)$ is small. Another way to use this is to solve (3.35) for n,

$$n = \frac{1}{2\varepsilon^2} \log \frac{N}{\delta}, \quad (3.36)$$

and to say that we need n training examples in order to be $(1-\delta) \times 100\%$ confident that the generalization error $E(f)$ is smaller that $E_n(f) + \varepsilon$. The exponential dependence on n in the Hoeffding–Chernoff inequality was crucial to obtain the logarithmic dependence on N and δ in (3.36) of the size n of the training set.

Of course, classification algorithms usually optimize over infinite function sets $\mathscr{F}$. In this case, one can often find a finite subset of functions that approximate all functions in $\mathscr{F}$ within a given error, and then apply the Hoeffding–Chernoff inequality to this set. There are many other ideas involved in the analysis of the generalization ability of classification algorithms, including various analogues of the Hoeffding–Chernoff inequality. □

Exercise 3.3.1. (*Variational representation*) Prove that, for $p, q \in (0,1)$,

$$D(q\|p) = \sup_{a,b \in \mathbb{R}} \Big[aq + b(1-q) - \log\Big(e^a p + e^b (1-p)\Big)\Big].$$

Exercise 3.3.2. Prove that the function $f(p,q) = D(q\|p)$ is convex on $(0,1)^2$. *Hint:* use the previous exercise.

Exercise 3.3.3. Prove (3.33) by showing that $D(p+t\|p) \geq 2t^2$ for $t \leq 1-p$. *Hint:* compare second derivatives.

Exercise 3.3.4. Prove that for $0 < p \leq 1/2$ and $0 \leq t < p$,

$$D(1-p+t\|1-p) \geq \frac{t^2}{2p(1-p)},$$

so

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p-t\Big) \leq \exp\Big(-\frac{nt^2}{2p(1-p)}\Big).$$

Hint: compare second derivatives.

Exercise 3.3.5. Suppose that the random variables $X_1,\ldots,X_n$ are independent, $X_i \in [a,b]$ and $\mu = \mathbb{E}X_i$. How can we use the Hoeffding–Chernoff inequality to bound $\mathbb{P}(\overline{X}_n \geq \mu + t)$ for $0 \leq t \leq b-\mu$ and $\mathbb{P}(\overline{X}_n \leq \mu - t)$ for $0 \leq t \leq \mu - a$?

Exercise 3.3.6. Suppose that the loss function $L \in [0,1]$ and the class of functions $\mathscr{F}$ utilized by a classification algorithm are such that the following approximation property holds for some constant $V > 0$. For any $\varepsilon > 0$, we can find $N = (\frac{1}{\varepsilon})^V$ functions $f_1,\ldots,f_N$ in $\mathscr{F}$ such that, for any $f \in \mathscr{F}$,

$$\sup_{x,y}\big|L(y,f(x)) - L(y,f_j(x))\big| \leq \varepsilon$$

for some $j \leq N$. What should the training data size n be in order to guarantee with probability at least $1-\delta$ that the generalization error $E(f) \leq E_n(f) + \varepsilon$ for all $f \in \mathscr{F}$?

3.4 Azuma Inequality

Until now we have considered sums of independent random variables. In this section, we will prove a concentration inequality that deals with non-linear functions

$$Z = f(X_1, \ldots, X_n) \tag{3.37}$$

of independent random variables, or random vectors, X_i. For applications (we will see several in the next section), it will be convenient to suppose that each entry X_i can itself be a vector consisting of several random variables, as long as

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i), \tag{3.38}$$

for any possible (vector) values $x_1, \ldots, x_n$ that these vectors $X_1, \ldots, X_n$ can take.

We will assume the following *stability condition* on the function f:

$$\big| f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n) \big| \le a_i \tag{3.39}$$

for all $i \le n$, for some constants $a_1, \ldots, a_n$. This means that changing the i^{th} coordinate of the function f while keeping all the other coordinates fixed can change its value by not more than a_i. In particular, the function f is bounded, which we assume throughout this section. Our main result is the following.

Theorem 3.4 (Azuma's inequality). *If (3.39) holds then*

$$\mathbb{P}\big(f - \mathbb{E}f \ge t\big) \le \exp\Big(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}\Big) \tag{3.40}$$

for any $t \ge 0$.

In the proof we will finally use the Fubini Theorem 1.2 from Section 1.4 in an essential way. It states that if the random variables X are Y and independent then

$$\mathbb{E}f(X,Y) = \sum_k \Big[\sum_\ell f(a_k, b_\ell)\mathbb{P}(Y = b_\ell)\Big]\mathbb{P}(X = a_k), \quad (3.41)$$

where the sum is over possible values a_k and b_ℓ of X and Y, which means that we can first fix the value of $X = a_n$ and average with respect to the distribution of Y, and then average the result over the distribution of X. The proof of Fubini's theorem was very simple, based on the rearrangement of the summation over possible outcomes, and the same proof works for random vectors without any changes. For example, if we split $(X_1, \ldots, X_n)$ into two vectors $(X_1, \ldots, X_i)$ and $(X_{i+1}, \ldots, X_n)$, for simplicity of notation denote $a = (x_1, \ldots, x_i)$ and $b = (x_{i+1}, \ldots, x_n)$, and use that the independence in (3.38) implies that

$$\begin{aligned} &\mathbb{P}\big((X_1, \ldots, X_n) = (a, b)\big) \qquad (3.42)\\ &= \mathbb{P}\big((X_1, \ldots, X_i) = a\big)\mathbb{P}\big((X_{i+1}, \ldots, X_n) = b\big), \end{aligned}$$

then the Fubini theorem allows to write $\mathbb{E}f(X_1, \ldots, X_n)$ as

$$\sum_k \sum_\ell f(a_k, b_\ell)\mathbb{P}\big((X_{i+1}, \ldots, X_n) = b_\ell\big)\mathbb{P}\big((X_1, \ldots, X_i) = a_k\big),$$

where the sum is over all possible values of $a = (x_1, \ldots, x_i)$ and $b = (x_{i+1}, \ldots, x_n)$. It will be convenient to simplify the notation, namely, denote the average over the coordinates $(x_{i+1}, \ldots, x_n)$ by $\mathbb{E}_i$. In other words,

$$\begin{aligned} &\mathbb{E}_i f(X_1, \ldots, X_n) \qquad (3.43)\\ &= \sum_\ell f((X_1, \ldots, X_i), b_\ell)\mathbb{P}\big((X_{i+1}, \ldots, X_n) = b_\ell\big), \end{aligned}$$

which is a function of $(X_1, \ldots, X_i)$. Averaging these remaining coordinates with respect to their (marginal) distribution $\mathbb{P}((X_1, \ldots, X_i) = a_k)$ is the second step in the Fubini theorem, so it can be written in this more compact notation as

$$\mathbb{E}f(X_1,\ldots,X_n)=\mathbb{E}\big[\mathbb{E}_i f(X_1,\ldots,X_n)\big]. \tag{3.44}$$

Notice that we can also use the Fubini theorem in the following way,

$$\mathbb{E}_{i-1}f(X_1,\ldots,X_n)=\mathbb{E}_{i-1}\big[\mathbb{E}_i f(X_1,\ldots,X_n)\big], \tag{3.45}$$

because averaging in the coordinates $(x_i,\ldots,x_n)$ can also be done in two steps, first averaging over $(x_{i+1},\ldots,x_n)$ and then averaging over x_i.

We will need the following observation in the proof of Azuma's inequality.

Lemma 3.5. *If a random variable X satisfies $|X|\le 1$ and $\mathbb{E}X=0$ then, for any $\lambda\ge 0$,*

$$\mathbb{E}e^{\lambda X}\le e^{\lambda^2/2}.$$

Proof. We can write λX as a convex combination

$$\lambda X=\frac{1+X}{2}\lambda+\frac{1-X}{2}(-\lambda),$$

because $(1+X)/2,(1-X)/2\in[0,1]$ and their sum is equal to 1. By the convexity of e^x,

$$e^{\lambda X}\le\frac{1+X}{2}e^{\lambda}+\frac{1-X}{2}e^{-\lambda},$$

and, taking expectations and using that $\mathbb{E}X=0$, we get

$$\mathbb{E}e^{\lambda X}\le\frac{1}{2}e^{\lambda}+\frac{1}{2}e^{-\lambda}=\cosh(\lambda).$$

The result follows, because we checked in the proof of the Hoeffding inequality that $\cosh(\lambda)\le e^{\lambda^2/2}$. □

We are now ready to prove Azuma's inequality.

Proof (Theorem 3.4). The main idea of the proof will be to represent the quantity of interest, $f-\mathbb{E}f$, as a sum of n terms, where we will average one coordinate at a time (instead of averaging them all at once) and then compare two consecutive averages. The stability condition will allow us to control the difference of two consecutive averages and, moreover, the average of each increment will be zero. Then, we will apply the previous lemma.

Recalling the notation $\mathbb{E}_i$ in (3.43), we denote

$$Y_i = \mathbb{E}_i f(X_1,\ldots,X_n) - \mathbb{E}_{i-1} f(X_1,\ldots,X_n) \tag{3.46}$$

for $i = 1,\ldots,n$. As we mentioned above, this is the difference of two consecutive averages. What will be crucial to us is that the stability condition (3.39) implies that

$$|Y_i| \le a_i, \tag{3.47}$$

which we will leave as a simple exercise below. Intuitively, this should be clear, since the average $\mathbb{E}_{i-1}$ differs from $\mathbb{E}_i$ only in that $\mathbb{E}_{i-1}$ also averages the coordinate X_i while it remains fixed in $\mathbb{E}_i$, so the corresponding terms in the averages differ by not more than a_i. Notice that, since

$$\mathbb{E}_0 f(X_1,\ldots,X_n) = \mathbb{E} f(X_1,\ldots,X_n)$$

(because $\mathbb{E}_0$ is the average with respect to $X_1,\ldots,X_n$) and

$$\mathbb{E}_n f(X_1,\ldots,X_n) = f(X_1,\ldots,X_n)$$

(because $\mathbb{E}_n$ is the placeholder for empty average), by the telescoping sum,

$$Y_1 + \ldots + Y_n = f(X_1,\ldots,X_n) - \mathbb{E} f(X_1,\ldots,X_n). \tag{3.48}$$

This is called a *martingale-difference representation*. The name comes from the fact that, by (3.45),

$$\mathbb{E}_{i-1} Y_i = \mathbb{E}_{i-1} f(X_1,\ldots,X_n) - \mathbb{E}_{i-1} f(X_1,\ldots,X_n) = 0,$$

i.e. the average of each term Y_i with respect to the last coordinate X_i that it depends on is zero.

Let us now take $\lambda \geq 0$ and start with Markov's inequality

$$\mathbb{P}\big(f - \mathbb{E}f \geq t\big) = \mathbb{P}\Big(\sum_{i=1}^{n} Y_i \geq t\Big) \leq e^{-\lambda t}\mathbb{E}e^{\lambda Y_1 + \ldots + \lambda Y_n}.$$

Using (3.44) with $i = n-1$,

$$\begin{aligned}\mathbb{E}e^{\lambda Y_1 + \ldots + \lambda Y_n} &= \mathbb{E}\Big[\mathbb{E}_{n-1}e^{\lambda Y_1 + \ldots + \lambda Y_{n-1}}e^{\lambda Y_n}\Big] \\ &= \mathbb{E}\Big[e^{\lambda Y_1 + \ldots + \lambda Y_{n-1}}\mathbb{E}_{n-1}e^{\lambda Y_n}\Big], \qquad (3.49)\end{aligned}$$

because $\mathbb{E}_{n-1}$ is the average in X_n, and the terms $Y_1, \ldots, Y_{n-1}$ do not depend on the coordinate X_n.

Let $X = Y_n/a_n$, which we view as a function of X_n only, with all other coordinates fixed. Since we showed above that $|X| \leq 1$ and $\mathbb{E}_{n-1}X = 0$, Lemma 3.5 gives

$$\mathbb{E}_{n-1}e^{\lambda Y_n} = \mathbb{E}_{n-1}e^{\lambda a_n X} \leq e^{(\lambda a_n)^2/2}.$$

Plugging this into (3.49), we get

$$\mathbb{E}e^{\lambda Y_1 + \ldots + \lambda Y_n} \leq e^{(\lambda a_n)^2/2}\mathbb{E}e^{\lambda Y_1 + \ldots + \lambda Y_{n-1}}.$$

We can now proceed in exactly the same fashion to average X_{n-1}, X_{n-2}, and so on, removing one by one the terms Y_{n-1}, $Y_{n-2}, \ldots$, and in the end we get

$$\mathbb{E}e^{\lambda Y_1 + \ldots + \lambda Y_n} \leq e^{\sum_{i=1}^{n}(\lambda a_i)^2/2}.$$

This proves that

$$\mathbb{P}\Big(\sum_{i=1}^{n} Y_i \geq t\Big) \leq \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{n} a_i^2\Big).$$

Optimizing over $\lambda \geq 0$ yields the claim (3.40), so the proof is complete. □

Exercise 3.4.1. Show that (3.38) implies (3.42).

Exercise 3.4.2. Check carefully that (3.45) holds.

Exercise 3.4.3. Prove the equation (3.47)

Exercise 3.4.4. Among all the random variables X such that $|X| \leq 100$ and $\mathbb{E}X = 0$, which one has the largest exponential moment $\mathbb{E}e^X$?

3.5 Applications of Azuma Inequality

In this section, we will show several applications of Azuma's inequality.

Example 3.5.1 (Chromatic number of the Erdős–Rényi graph). Consider the Erdős–Rényi random graph $G(n,p)$ on n vertices. Let $\chi(G(n,p))$ be the *chromatic number* of this graph, which is the smallest number of colours needed to colour the vertices so that no two adjacent vertices share the same colour. Let us denote the set of vertices by $\{v_1,\ldots,v_n\}$ and let

$$e_{i,j} = \mathrm{I}\big(\text{the edge between } v_i \text{ and } v_j \text{ is present}\big).$$

Then, all $e_{i,j}$ are independent Bernoulli $B(p)$ random variables, by the definition of the Erdős–Rényi random graph. For $i = 2,\ldots,n$, let us denote by $X_i = (e_{1,i}, e_{2,i}\ldots, e_{i-1,i})$ the vector of indicators of edges between the vertex v_i and vertices $v_1,\ldots,v_{i-1}$. The vectors $X_2,\ldots,X_n$ are independent because they consist of indicators of disjoint sets of edges, and the chromatic number is a function

$$\chi(G(n,p)) = f(X_2,\ldots,X_n),$$

since these vectors include indicators of all edges in the graph. To apply Azuma's inequality, we need to determine the stability constants $a_2,\ldots,a_n$. Notice that if we replace X_i with another value $X_i' = (e_{1,i}', e_{2,i}'\ldots, e_{i-1,i}')$, this means that we modify some edges between v_i and $v_1,\ldots,v_{i-1}$. The chromatic number can not increase by more than 1, because we can always assign a new colour to the vertex v_i, so

$$f(X_2,\ldots,X_i',\ldots,X_n) - f(X_2,\ldots,X_i,\ldots,X_n) \le 1.$$

By the same logic,

$$f(X_2,\ldots,X_i,\ldots,X_n) - f(X_2,\ldots,X_i',\ldots,X_n) \le 1,$$

and this shows that the stability condition (3.39) holds with $a_i = 1$. Therefore, Azuma's inequality implies (when applied to both f and $-f$)

$$\mathbb{P}\Big(\big|\chi(G(n,p)) - \mathbb{E}\chi(G(n,p))\big| \geq t\Big) \leq 2e^{-\frac{t^2}{2(n-1)}}.$$

For example, if we take $t = \sqrt{2n\log n}$, we get

$$\mathbb{P}\Big(\big|\chi(G(n,p)) - \mathbb{E}\chi(G(n,p))\big| \leq \sqrt{2n\log n}\Big) \geq 1 - \frac{2}{n},$$

so, with high probability, the chromatic number will be within $\sqrt{2n\log n}$ from its expected value $\mathbb{E}\chi(G(n,p))$. It is known (but non-trivial) that this expected value

$$\mathbb{E}\chi(G(n,p)) \sim \frac{n}{2\log n}\log\frac{1}{1-p},$$

so the deviation $\sqrt{2n\log n}$ is of a smaller order compared to the expectation. This shows that the chromatic number is typically of the same order as its expectation, which gives us an example of the law of large numbers for a very non-trivial functional of independent random variables. □

Example 3.5.2 (Balls in boxes). Suppose that we throw n balls into m boxes at random, so that the probability that a ball lands in any given box is $1/m$, and independently of each other. Let N be the number of non-empty boxes. If $X_i \in \{1,\ldots,m\}$ is the box number in which the i^{th} ball lands then $X_1,\ldots,X_n$ are independent random variables and $N = \text{card}\{X_1,\ldots,X_n\}$ is the number of distinct boxes hit.

As in the Example 1.5.1 in Section 1.5, it is easy to compute, using the linearity of expectation, that

$$\mathbb{E}N = m\Big(1 - \Big(1 - \frac{1}{m}\Big)^n\Big). \tag{3.50}$$

When n is large and $m = \alpha n$ for some fixed $\alpha > 0$ then

$$\mathbb{E}N \sim n\alpha\big(1 - e^{-1/\alpha}\big).$$

Changing the value of one box X_i changes N by at most 1, so the stability constants are all equal to $a_i = 1$. As in the previous example,

$$\mathbb{P}\big(\big|N - \mathbb{E}N\big| \geq t\big) \leq 2e^{-\frac{t^2}{2n}}, \tag{3.51}$$

by Azuma's inequality, and

$$\mathbb{P}\big(\big|N - \mathbb{E}N\big| \leq \sqrt{2n\log n}\big) \geq 1 - \frac{2}{n}.$$

So the typical number of non-empty boxes in this case is close to its expectation, relatively speaking. □

Example 3.5.3 (Max-Cut of sparse random graph). This example is quite similar to the previous one, only balls and boxes have a different meaning and, instead of the number of non-empty boxes, we consider a much more complicated function.

Let $V = \{v_1, \ldots, v_n\}$ be the set of vertices of a graph and E be its set of edges. Max-Cut problem is to divide vertices into two groups in a way that maximizes the number of edges between the two groups. This is a very important problem in computer science that has many applications, for example, to layout of electronic circuitry, and as a reformulations of various combinatorial optimization problems. For example, imagine that we have a group of n people and edges represent people that dislike each other. Then our goal is to separate them into two groups in such a way that as many 'enemies' as possible are in the opposite groups.

Here we will consider a specific model of a random graph, and will show that the maximal number of edges between the two groups concentrates around its expectation. We will select edges randomly, but using a different procedure than in the Erdős–Rényi graph. We will take

$$m = dn$$

possible edges for some fixed $d > 0$ and, as usual, we think of n as being large. Then we place each of these m edges at random among the set of $\binom{n}{2}$ possible edges, independently of each other. Each pair of vertices (a possible location to place an edge) represents a box, and edges represent balls. The random variables $X_1, \dots, X_m$ that describe between which vertices each edge is placed take $\binom{n}{2}$ possible values and are all independent. This model produces what is called a *sparse* graph, because the number of edges $m = nd$ is small relative to $\binom{n}{2}$ and there is a fixed average number of edges per vertex, d. As in the above example, let

$$e_{i,j} = \mathrm{I}\big(\text{the edge between } v_i \text{ and } v_j \text{ is present}\big).$$

It is possible that more than one edge is placed between two vertices, in which case we keep one of them.

The function that we will consider on the above sparse random graph is called Max-Cut. It is defined as follows. If we want to split all vertices into two groups, one way to encode this is to assign each vertex one of the two labels $\{-1, 1\}$. Then all vertices with the same label belong to the same group. In other words, each vector

$$\sigma = \big(\sigma_1, \dots, \sigma_n\big) \in \{-1, 1\}^n$$

describes a possible *cut* of the graph into two groups. Let $E(\sigma)$ be the number of present edges connecting the vertices in opposite groups,

$$E(\sigma) = \mathrm{card}\Big\{i < j : e_{i,j} = 1, \sigma_i \neq \sigma_j\Big\}. \tag{3.52}$$

Another way to represent $E(\sigma)$ is as follows. Notice that $\sigma_i \neq \sigma_j$ only if $\sigma_i\sigma_j = -1$, otherwise, $\sigma_i\sigma_j = 1$. Therefore, $\mathrm{I}(\sigma_i \neq \sigma_j) = (1 - \sigma_i\sigma_j)/2$ and

$$E(\sigma) = \frac{1}{2} \sum_{i<j} e_{i,j} \big(1 - \sigma_i \sigma_j\big). \tag{3.53}$$

Then the value M of the *Max-Cut* corresponds to a way to cut the graph so that the number of edges between the two groups is as large as possible,

$$M = \max_{\sigma} E(\sigma). \tag{3.54}$$

Recall that M is a random variable that depends on the positions $X_1, \ldots, X_m$ of our possible $m = dn$ edges.

If we change the placement X_i of one edge, this can change the maximum M by at most 1, because, for each possible cut (configuration σ), moving one edge might increase or decrease the number of edges between the two groups by at most 1. This means that the stability condition in Azuma's inequality holds with $a_i = 1$ and, therefore,

$$\mathbb{P}\big(\big|M - \mathbb{E}M\big| \geq t\big) \leq 2e^{-\frac{t^2}{2m}} = 2e^{-\frac{t^2}{2dn}}. \tag{3.55}$$

If we take $t = \sqrt{2dn \log n}$, we get

$$\mathbb{P}\Big(\big|M - \mathbb{E}M\big| \leq \sqrt{2dn \log n}\Big) \geq 1 - \frac{2}{n}.$$

Is the deviation $\sqrt{2dn \log n}$ of smaller order than $\mathbb{E}M$? It turns out that there exists a cut such that at least half of all edges are between the vertices that belong to opposite groups. To see this, choose labels $\sigma_1, \ldots, \sigma_n$ to be i.i.d. Rademacher, which means that we assign each vertex to one of the two groups $\{-1, +1\}$ at random. Then

$$\begin{aligned} M &= \max_{\sigma} E(\sigma) \geq \mathbb{E}E(\sigma) \\ &= \frac{1}{2} \sum_{i<j} e_{i,j} \big(1 - \mathbb{E}\sigma_i \sigma_j\big) = \frac{1}{2} \sum_{i<j} e_{i,j}, \end{aligned}$$

which is exactly one half of all the present edges. Taking the expectations on both sides, we can write

$$\begin{aligned}\mathbb{E}M &\geq \frac{1}{2}\sum_{i<j}\mathbb{P}\big(\text{there is an edge between } v_i \text{ and } v_j\big)\\ &= \frac{1}{2}\binom{n}{2}\Big(1-\Big(1-\frac{1}{\binom{n}{2}}\Big)^{dn}\Big),\end{aligned}$$

where on the right hand side we have the analogue of (3.50). Using the inequality $1-x \leq e^{-x}$, we can write

$$\begin{aligned}\Big(1-\frac{1}{\binom{n}{2}}\Big)^{dn} \leq e^{-dn/\binom{n}{2}} \leq e^{-\frac{2d}{n}}&\\ &= 1-\frac{2d}{n}+\frac{2d^2}{n^2}+O\Big(\frac{1}{n^3}\Big),\end{aligned}$$

where in the second line we used Taylor's theorem for e^x at zero. Plugging this in the above inequality, we get

$$\begin{aligned}\mathbb{E}M &\geq \frac{1}{2}\binom{n}{2}\Big(\frac{2d}{n}-\frac{2d^2}{n^2}+O\Big(\frac{1}{n^3}\Big)\Big)\\ &= \frac{dn}{2}-\frac{d+d^2}{2}+O\Big(\frac{1}{n}\Big).\end{aligned}$$

This shows that (up to the smaller order terms) $\mathbb{E}M$ is at least $dn/2$. This means that the deviation $\sqrt{2dn\log n}$ in Azuma's inequality above is of a smaller order, so the typical Max-Cut value M is relatively close to its expectation.

By the way, it is a major open problem to compute the limit $\lim_{n\to\infty}\mathbb{E}M/n$ in terms of d. □

Example 3.5.4 (Hamming cube). For any $x,y \in \{0,1\}^n$, the number of coordinates of x and y that differ,

$$\rho(x,y) = \sum_{i=1}^{n} \mathrm{I}(x_i \neq y_i), \tag{3.56}$$

is called the *Hamming distance* between x and y. The function ρ is called the Hamming metric on $\{0,1\}^n$. We will now use Azuma's inequality to show that, given a subset $A \subseteq \{0,1\}^n$ that contains a positive proportion of all points,

$$\mathrm{card}(A) > \varepsilon 2^n \tag{3.57}$$

for some $\varepsilon > 0$, most points in $\{0,1\}^n$ are relatively close to the set A, namely, within the Hamming distance of order $\sqrt{n}$. In other words, for most points in $\{0,1\}^n$, only of order $\sqrt{n}$ coordinates are different from one of the points in A.

Given $t > 0$, let us denote by

$$B(A,t) = \Bigl\{ x \in \{0,1\}^n : \rho(x,y) \le t \text{ for some } y \in A \Bigr\} \tag{3.58}$$

the set of all points within Hamming distance t from A. For $\varepsilon, \delta \in (0,1)$, let us denote

$$t_{\varepsilon,\delta} = \sqrt{2\log\frac{1}{\varepsilon}} + \sqrt{2\log\frac{1}{\delta}}.$$

Then the following lemma shows that $t_{\varepsilon,\delta}\sqrt{n}$-neighbourhood of A contains at least $(1-\delta)$ proportion of all points in $\{0,1\}^n$.

Lemma 3.6. *For* $\varepsilon, \delta \in (0,1)$, *if* $\mathrm{card}(A) > \varepsilon 2^n$ *then*

$$\mathrm{card}\, B\bigl(A, t_{\varepsilon,\delta}\sqrt{n}\bigr) \ge (1-\delta)2^n. \tag{3.59}$$

Proof. Let $X = (X_1, \ldots, X_n)$ be a vector consisting of i.i.d. Bernoulli $B(1/2)$ random variables, which means that, for any subset $S \subseteq \{0,1\}^n$,

$$\mathbb{P}\bigl(X \in S\bigr) = \frac{\mathrm{card}(S)}{2^n}.$$

Let us consider a random variable $Z = \min_{y \in A} \rho(X,y)$ equal to the Hamming distance from X to the set A. Changing one coordinate X_i to X_i' can change Z by at most 1, and Azuma's inequality (applied to Z and $-Z$) implies that

$$\mathbb{P}\big(Z \le \mathbb{E}Z - t\sqrt{n}\big) \le e^{-t^2/2}, \tag{3.60}$$

$$\mathbb{P}\big(Z \ge \mathbb{E}Z + t\sqrt{n}\big) \le e^{-t^2/2}. \tag{3.61}$$

If in the first inequality we take $t_\varepsilon = \sqrt{2\log\frac{1}{\varepsilon}}$ then

$$\mathbb{P}\big(Z \le \mathbb{E}Z - t_\varepsilon\sqrt{n}\big) \le \varepsilon.$$

This direction of Azuma's inequality allows us to conclude that $\mathbb{E}Z \le t_\varepsilon\sqrt{n}$ because of the following observation. If $\mathbb{E}Z - t_\varepsilon\sqrt{n} > 0$ then the above probability would be bigger than $\mathbb{P}(Z = 0)$. However, since $Z = 0$ if and only if $X \in A$,

$$\mathbb{P}(Z = 0) = \mathbb{P}(X \in A) = \frac{\operatorname{card}(A)}{2^n} > \varepsilon,$$

by our assumption, which is a contradiction. Therefore,

$$\mathbb{E}Z \le t_\varepsilon\sqrt{n}.$$

The second inequality then implies

$$\mathbb{P}\big(Z \ge t_\varepsilon\sqrt{n} + t\sqrt{n}\big) \le \mathbb{P}\big(Z \ge \mathbb{E}Z + t\sqrt{n}\big) \le e^{-t^2/2}.$$

In this inequality, we take $t_\delta = \sqrt{2\log\frac{1}{\delta}}$ then

$$\mathbb{P}\big(Z \ge t_\varepsilon\sqrt{n} + t_\delta\sqrt{n}\big) \le \delta,$$

which implies that

$$\mathbb{P}(Z \le t_\varepsilon\sqrt{n} + t_\delta\sqrt{n}) \ge 1 - \delta.$$

Since this event is exactly the set $B(A, t_{\varepsilon,\delta}\sqrt{n})$ of all points within the Hamming distance $t_{\varepsilon,\delta}\sqrt{n}$ from A, this finishes the proof. □

Exercise 3.5.1 (Empirical process). Given a family $\mathscr{F}$ of functions $f\colon \mathbb{R} \to [0,1]$ and independent random variables $X_1, \ldots, X_n$, consider

$$Z := \frac{1}{\sqrt{n}} \sup_{f \in \mathscr{F}} \Big| \sum_{i=1}^{n} \big(f(X_i) - \mathbb{E} f(X_i)\big) \Big|.$$

Prove that $\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-t^2/2}$ for all $t \geq 0$.

Exercise 3.5.2. In a graph with n vertices and $m = dn$ edges, if d_i is the number of edges connected to the vertex v_i, show that $\frac{1}{n}\sum_{i=1}^{n} d_i = 2d$.

Exercise 3.5.3 (3-SAT). Suppose that a debate class has a large number of students, n. The professor creates $m = 2n$ debate teams, each composed of 3 members, and he does it by assigning each student to 6 different teams.

Each member of each team is assigned to defend a position from the platform of one of two major political parties, Party A or Party B. The professor does not know the party affiliations of the students, and we suppose that each student belongs to either Party A or Party B with probability $1/2$ each, independently of each other. Given their turn, each team will select only one of its members to defend his or her assigned position. However, if it turns out that all three members have been assigned positions of the opposing party (not the one to which they belong), the team will skip their turn.

Let N be the number of speeches delivered in class. What is the expectation $\mathbb{E}N$? Use Azuma's inequality to show that, with high probability, the number of speeches N will be relatively close to $\mathbb{E}N$.

Exercise 3.5.4. Suppose we throw n balls into m boxes at random, and let N be the number of boxes with at least two balls in them. Compute $\mathbb{E}N$. What is the limit $\lim_{n\to\infty} \mathbb{E}N/n$ when $m = \alpha n$ for a fixed $\alpha > 0$? What does Azuma's inequality say about $|N - \mathbb{E}N|$?

Chapter 4
Gaussian Distributions

In Example 3.1.1, we have seen that Hoeffding's inequality for the average

$$\overline{X}_n = \frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}$$

of i.i.d. $X_1, \ldots, X_n$ Bernoulli $B(1/2)$ random variables implies that, for $x \geq 0$,

$$\mathbb{P}\big(|\overline{X}_n - 1/2| > x\big) \leq 2e^{-2nx^2}.$$

If we make the change of variables $x = \frac{t}{2\sqrt{n}}$ and denote

$$Z_n := 2\sqrt{n}(\overline{X}_n - 1/2),$$

we can rewrite this inequality as

$$\mathbb{P}\big(|Z_n| \geq t\big) \leq 2e^{-t^2/2}.$$

The expectation and variance of Bernoulli $B(1/2)$ random variables equal

$$\mu = \mathbb{E}X_1 = \frac{1}{2},\ \sigma^2 = \mathrm{Var}(X_1) = \frac{1}{4}$$

and, with this notation, we can represent Z_n as

$$Z_n = \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}.$$

We mentioned in the Example 3.1.1 that the dependence on t in Hoeffding's inequality is nearly optimal, which will be demonstrated in this chapter once we prove the Central Limit Theorem (CLT). For Z_n as above, the CLT states that

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx, \tag{4.1}$$

for all $t \in \mathbb{R}$. In other words, instead of upper bounds on the probabilities, we will get precise asymptotic formulas in the limit $n \to \infty$. For example, (4.1) will imply

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|Z_n| \geq t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-t} e^{-x^2/2}\, dx + \frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} e^{-x^2/2}\, dx \\ &= 2\frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} e^{-x^2/2}\, dx,\end{aligned}$$

by symmetry of the integrand. For large t, one can show the following.

Exercise 4.0.1. As $t \to \infty$,

$$\frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} e^{-x^2/2}\, dx \sim \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}$$

in the sense that their ratio goes to 1. *Hint:* use L'Hospital's rule.

This shows that Hoeffding's inequality was quite precise for large t, because it was only missing a factor $1/t\sqrt{2\pi}$, which is large compared to $e^{-t^2/2}$. It is important to mention that, while it might look like the CLT is a more precise result than Hoeffding's inequality, it is an asymptotic statement, whereas Hoeffding's inequality holds for all n. For this reason, Hoeffding's inequality is often more useful but, on the other hand, it is difficult to argue that the Central Limit Theorem is a mathematically more beautiful statement.

4.1 Gaussian Distributions on $\mathbb{R}$

The function that appeared in the integral in (4.1),

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \tag{4.2}$$

is called the *density of standard Gaussian distribution*. This is our first example of a continuous probability space and a continuous distribution on the real line. In contrast with discrete probability spaces, the space of possible outcomes $\Omega = \mathbb{R}$ is not finite or countable, and the probability of a subset of outcomes $A \subseteq \mathbb{R}$ (an event) is not defined as a sum but as an integral

$$\gamma(A) := \int_A p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2}\,dx. \tag{4.3}$$

The probability measure γ is called the *standard Gaussian measure* on $\mathbb{R}$, and we use the notation γ instead of the usual $\mathbb{P}$ to save $\mathbb{P}$ for more generic situations. This definition means that probabilities of individual outcomes are all zero, $\gamma(\{x\}) = 0$, and probability of an interval $[a,b]$ is

$$\gamma([a,b]) = \int_a^b p(x)\,dx. \tag{4.4}$$

Compared to discrete distributions, not every subset $A \subseteq \Omega$ can be an event, because for some sets A the integral may be undefined. Allowed choices of A depend on whether we use the Lebesgue integral or Riemann integral. For now, we will work only with nice sets such as intervals of the form

$$[a,b], [a,b), (a,b], (a,b),$$

where a and b can be $\pm\infty$, or finite unions of intervals, in which case the Riemann integral is well defined.

Since the Lebesgue integral has better properties than Riemann integral, you can think of all the integrals as the Lebesgue integral if you are familiar with it. But even if we work with the Riemann integral, we can notice that at least some of the properties of probability measures that we observed in the setting of discrete probability spaces still hold here. For example,

$$\gamma(\mathbb{R}) = \int_{-\infty}^{\infty} p(x)\,dx = 1. \tag{4.5}$$

To see this, one can use the following replica trick,

$$\begin{aligned}\gamma(\mathbb{R})^2 &= \Big(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-x^2/2}\,dx\Big)^2\\ &= \frac{1}{2\pi}\iint_{\mathbb{R}^2} e^{-(x^2+y^2)/2}\,dxdy\\ &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\,dr d\theta\\ &= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2/2} r\,dr = 1,\end{aligned}$$

where we used polar coordinates as $dxdy = r\,dr\,d\theta$. We also have finite additivity,

$$\gamma(A_1\cup\ldots\cup A_n) = \gamma(A_1)+\ldots+\gamma(A_n) \tag{4.6}$$

for disjoint sets $A_1,\ldots,A_n$, as long as these sets are nice, for example, finite unions of intervals. This holds by the usual linearity of integral.

However, if we would like to have countable additivity as in Section 1.2 for discrete spaces, working with the Riemann integral is not enough. For example, if we take a sequence of one point sets $\{q_n\}$, where $Q = \{q_1, q_2, \ldots\}$ are rational numbers on the interval $[0,1]$ (enumerated in an arbitrary order) then the Riemann integral over one point is zero, but the Riemann integral of the indicator of Q is undefined (Q is not rectifiable), while the Lebesgue integral is well defined and equal to zero. Countable additivity is satisfied by the Lebesgue integral over the class of Lebesgue measurable sets, which can play the role of events on this probability space.

Let $g\colon \mathbb{R} \to \mathbb{R}$ be the identity function $g(x) = x$. On the space Ω with the measure γ, this function may be called a *standard Gaussian random variable*, or *random variable with the standard Gaussian distribution* (by analogy with the definitions on discrete spaces), because

$$\mathbb{P}(g \in A) = \gamma\big(x : g(x) \in A\big) = \gamma(A). \tag{4.7}$$

Exercise 4.1.1. Show that $g(x) = -x$ also has the standard Gaussian distribution.

Recall that, on discrete probability spaces, any function was called a random variable. However, on a continuous space such as $(\mathbb{R}, \gamma)$, just like not every set A can be called an event, a function $X\colon \mathbb{R} \to \mathbb{R}$ has to be nice enough to be called a random variable. The reason for this is that, if we want to calculate the probability that X takes values in some set A (for example, an interval),

$$\begin{aligned}\mathbb{P}(X \in A) &= \gamma(x : X(x) \in A) \\ &= \gamma\big(X^{-1}(A)\big) = \int_{X^{-1}(A)} p(x)\,dx,\end{aligned} \tag{4.8}$$

we need to be able to integrate over the pre-image $X^{-1}(A)$. If we are using the Lebesgue integral then random variables are (Borel) measurable functions. In general, the Lebesgue integral is much better suited for Probability, but, since we do not assume the familiarity with Lebesgue integration, we simply have to limit ourselves to a nicer class of functions. We will call a function $f\colon \mathbb{R} \to \mathbb{R}$ *nice* if

1. f is piecewise-continuous (and bounded on compacts);
2. the pre-image $f^{-1}(A)$ of any interval A is a rectifiable set.

We can be even more restrictive and require that $f^{-1}(A)$ for any interval A is a finite union of intervals.

For nice functions, we can calculate their distributions. Let us consider several examples.

Example 4.1.1. Let us first consider the function

$$X(x) = \mathrm{I}(x \geq 0) - \mathrm{I}(x < 0),$$

which takes two values $\{-1,+1\}$. Then

$$\mathbb{P}(X = +1) = \gamma(x : x \geq 0) = \int_0^\infty p(x)\,dx = \frac{1}{2},$$
$$\mathbb{P}(X = -1) = \gamma(x : x < 0) = \int_{-\infty}^0 p(x)\,dx = \frac{1}{2}.$$

This means that X is a familiar Rademacher random variable, now defined on a continuous probability space $(\mathbb{R}, \gamma)$. □

Example 4.1.2. Next, let us consider $X(x) = |x|$. Obviously, X takes values in $[0,\infty)$, but how do we describe probabilities $\mathbb{P}(X \in A)$? If $0 \leq a \leq b$,

$$\big\{X \in [a,b]\big\} = \big\{x : |x| \in [a,b]\big\} = [-b,-a] \cup [a,b]$$

and, therefore, by symmetry of the Gaussian density $p(x)$,

$$\mathbb{P}(X \in [a,b]) = 2\gamma([a,b]) = \int_a^b 2p(x)\,dx.$$

Since X can not take negative values, if we define

$$p_X(x) = 2p(x)\,\mathrm{I}(x \geq 0) = \begin{cases} 0, & x < 0, \\ 2p(x), & x \geq 0, \end{cases}$$

then, for any $a \leq b$, we can write

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p_X(x)\,dx.$$

If we add up over disjoint intervals, we get

$$\mathbb{P}(X \in A) = \int_A p_X(x)\,dx,$$

for any finite union of intervals A. The distribution of X is encoded by the function $p_X(x)$ via an integral, in the same way as the standard Gaussian distribution above. □

When the probabilities $\mathbb{P}(X \in A)$ are given by

$$\mathbb{P}(X \in A) = \int_A p_X(x)\,dx, \tag{4.9}$$

for an integrable function $p_X(x) \geq 0$ such that

$$\int_{\mathbb{R}} p_X(x)\,dx = 1,$$

we say that X has *continuous distribution with the density* $p_X(x)$ on the real line. More precisely, such distributions are called *absolutely continuous*, but for simplicity we will call them continuous. Notice that, if we replace the Gaussian density $p(x)$ in (4.3) with $p_X(x)$, we get another example (or as many examples as we like) of a non-discrete probability space, $(\mathbb{R}, \mathbb{P})$.

Example 4.1.3 (Exponential distribution). A probability distribution with the density

$$p_X(x) = \lambda e^{-\lambda x}\mathrm{I}(x \geq 0) \tag{4.10}$$

is called the *exponential distribution with parameter* λ. □

Example 4.1.4 (Uniform distribution). A probability distribution with the density

$$p_X(x) = \frac{1}{b-a}\mathrm{I}(a \leq x \leq b), \tag{4.11}$$

for $-\infty < a < b < \infty$, is called the *uniform distribution on the interval* $[a,b]$. □

The distribution $\mathbb{P}(X \in A)$ can be encoded by a function

$$F(x) = \mathbb{P}(X \leq x), \tag{4.12}$$

which is called the *cumulative distribution function (c.d.f.)* of X. For example,

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

If $F(x)$ is continuous at $x = a$ then

$$\begin{aligned}\mathbb{P}(X = a) &\leq \mathbb{P}(a - \varepsilon < X \leq a + \varepsilon)\\ &= F(a + \varepsilon) - F(a - \varepsilon)\end{aligned}$$

and, letting $\varepsilon \downarrow 0$ shows that $\mathbb{P}(X = a) = 0$. If the function $F(x)$ is differentiable and the derivative $F'(x) = p_X(x)$ is piecewise-continuous then, by the Fundamental Theorem of Calculus,

$$\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_a^b p_X(x)\,dx,$$

so $p_X(x)$ is the density of the distribution of X. To find the density, we can find the c.d.f. first and differentiate it.

Example 4.1.5. Let $X = g^2$, where $g(x) = x$ is a standard Gaussian random variable on $(\mathbb{R}, \gamma)$. For $x > 0$,

$$\begin{aligned}F(x) &= \mathbb{P}(X \leq x) = \mathbb{P}(g^2 \leq x)\\ &= \mathbb{P}(-\sqrt{x} \leq g \leq \sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} p(t)\,dt.\end{aligned}$$

Taking derivatives and using the symmetry of the Gaussian density $p(x)$,

$$\begin{aligned}F'(x) &= p(\sqrt{x})\frac{1}{2\sqrt{x}} - p(-\sqrt{x})\frac{-1}{2\sqrt{x}}\\ &= p(\sqrt{x})\frac{1}{\sqrt{x}} = \frac{1}{\sqrt{2\pi x}}e^{-x/2}.\end{aligned}$$

Since X does not take negative values, if we define

$$p_X(x) = \begin{cases} 0, & x \le 0, \\ \frac{1}{\sqrt{2\pi x}} e^{-x/2}, & x > 0, \end{cases}$$

then we showed that $p_X(x)$ is the density of the distribution of X. This distribution is called $\Gamma(\frac{1}{2}, \frac{1}{2})$ distribution, and we will come back to it in Section 4.4 □

Example 4.1.6. If $\mu \in \mathbb{R}$, $\sigma > 0$ and $X = \mu + \sigma g$, one can check as in the previous example that X has density

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{4.13}$$

This distribution, denoted $N(\mu, \sigma^2)$, is called the *Gaussian distribution with mean* μ *and variance* σ^2 (we will see why in a second). When $\mu = 0$, we denote $p_{0,\sigma}(x)$ by $p_\sigma(x)$. □

The *expectation* or *expected value* of a random variable $X\colon \mathbb{R} \to \mathbb{R}$ on the probability space $(\mathbb{R}, \gamma)$ is defined by

$$\mathbb{E}X = \int_{\mathbb{R}} X(x) p(x)\, dx, \tag{4.14}$$

assuming that $\mathbb{E}|X| < \infty$ (to make sure that the integrals of the positive and negative parts of X are well defined). For example, for $g(x) = x$,

$$\mathbb{E}g = \int_{\mathbb{R}} x p(x)\, dx = 0,$$

because the function $xp(x)$ is odd, and $\mathbb{E}|g| < \infty$. Also,

$$\begin{aligned}\mathbb{E}g^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, d(-e^{-x^2/2}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1,\end{aligned}$$

using integration by parts. The expectation and variance of the standard Gaussian random variable g is $\mathbb{E}g = 0$ and $\mathrm{Var}(g) = 1$. This implies that, for $X = \mu + \sigma g \sim N(\mu, \sigma^2)$,

$$\mathbb{E}X = \mu, \mathrm{Var}(X) = \mathbb{E}(\sigma g)^2 = \sigma^2,$$

so the meaning of the parameter μ, σ^2 in $N(\mu, \sigma^2)$ is the expectation and variance of X.

The process by which we computed the distribution of X in the examples above is called the *change of variables* in the sense of measures, where the function X does not have to be one-to-one (injective). When the distribution of X turned out to be continuous, we found that

$$\mathbb{P}(X \in A) = \int_{X^{-1}(A)} p(x)\,dx = \int_A p_X(t)\,dt, \tag{4.15}$$

for some *density function* $p_X(t)$. This change of variables formula can be extended to expectations.

Lemma 4.1 (Change of variables for integrals). *If (4.15) holds and* $\mathbb{E}|X| < \infty$ *then*

$$\mathbb{E}X = \int_{\mathbb{R}} X(x)p(x)\,dx = \int_{\mathbb{R}} t p_X(t)\,dt. \tag{4.16}$$

This is the analogue of the change of variables in the setting of discrete spaces in the equation (1.30) in Section 1.2.

Proof. First of all, if we write X as a sum of positive and negative parts,

$$X = X\,\mathrm{I}(X > 0) + X\,\mathrm{I}(X \le 0),$$

the density $p_X(t)$ on $[0, \infty)$ is the density of the positive part, and on $(-\infty, 0)$ it is the density of the negative part. If we can prove (4.16) separately for the positive and negative parts, the claim follows by adding them up. This means that it is enough to assume that $X > 0$ and prove that

$$\mathbb{E}X = \int_{\mathbb{R}} X(x)p(x)\,dx = \int_0^\infty t p_X(t)\,dt. \tag{4.17}$$

Let us take large $M > 0$ and divide the interval $(0, M]$ into many small subintervals

$$0 = y_0 < y_1 < \ldots < y_m = M,$$

where the length of each interval is less than $\varepsilon > 0$. Let us use the change of variables formula (4.15) for one interval $A = (y_i, y_{i+1}]$. If, for $i = 0, \ldots, m-1$, we denote

$$A_i := X^{-1}\Big((y_i, y_{i+1}]\Big) = \Big\{x \in \mathbb{R} : y_i < X(x) \leq y_{i+1}\Big\},$$

then (4.15) states that

$$\int_{A_i} p(x)\,dx = \int_{(y_i, y_{i+1}]} p_X(t)\,dt.$$

Since we restricted ourselves to nice functions, the sets A_i are rectifiable and the integrals are well defined. This equation can also be written as

$$\int_{\mathbb{R}} \mathrm{I}\big(y_i < X(x) \leq y_{i+1}\big)\,p(x)\,dx$$
$$= \int_{\mathbb{R}} \mathrm{I}(y_i < t \leq y_{i+1})\,p_X(t)\,dt.$$

If we multiply both sides by y_i and sum over i, we get

$$\int_{\mathbb{R}} \sum_{i=0}^{m-1} y_i\,\mathrm{I}\big(y_i < X(x) \leq y_{i+1}\big)\,p(x)\,dx$$
$$= \int_{\mathbb{R}} \sum_{i=0}^{m-1} y_i\,\mathrm{I}(y_i < t \leq y_{i+1})\,p_X(t)\,dt.$$

If we denote the sum in the last integral by

$$y_\varepsilon(t) := \sum_{i=0}^{m-1} y_i\,\mathrm{I}(y_i < t \leq y_{i+1}),$$

this equation can be written as

$$\int_{\mathbb{R}} y_\varepsilon\big(X(x)\big)\,p(x)\,dx = \int_{\mathbb{R}} y_\varepsilon(t)\,p_X(t)\,dt. \tag{4.18}$$

Notice that the step function $y_\varepsilon(t)$ is an approximation of the identity function on $(0,M]$,

$$\Big| y_\varepsilon(t) - t\,\mathrm{I}(0 < t \le M)\Big| \le \varepsilon,$$

because $y_\varepsilon(t) = y_i$ when $y_i < t \le y_{i+1} \le y_i + \varepsilon$. Therefore, letting $\varepsilon \downarrow 0$ in (4.18), we get

$$\begin{aligned} &\int_{\mathbb{R}} X(x)\,\mathrm{I}(0 < X(x) \le M)\,p(x)\,dx && (4.19)\\ &= \int_{\mathbb{R}} t\,\mathrm{I}(0 < t \le M)\,p_X(t)\,dt = \int_0^M t\,p_X(t)\,dt. \end{aligned}$$

By the Monotone Convergence Theorem for the Riemann (or Lebesgue) integral, letting $M \uparrow +\infty$ implies (4.17), because the limit function $X(x)$ on the left hand side was assumed to be integrable. □

This result can be generalized to functions $f(X)$ of the random variable X.

Lemma 4.2 (Change of variables for integrals). *If (4.15) holds and* $\mathbb{E}|f(X)| < \infty$ *then*

$$\mathbb{E}f(X) = \int_{\mathbb{R}} f(t)\,p_X(t)\,dt. \qquad (4.20)$$

The proof is identical, only instead of approximating the identity function by step functions, we approximate $f(t)$ by step functions of the form

$$f_\varepsilon(t) = \sum_{i=0}^{m-1} y_i\,\mathrm{I}(y_i < f(t) \le y_{i+1}).$$

Also, instead of $A = (y_i, y_{i+1}]$, now we apply the change of variables formula (4.15) to the set

$$B = f^{-1}(A) = \big\{t \in \mathbb{R} : y_i < f(t) \le y_{i+1}\big\}.$$

The set $X^{-1}(B)$ should be rectifiable and, since

$$X^{-1}(B) = X^{-1}(f^{-1}(A)) = (f \circ X)^{-1}(A),$$

this poses some restrictions on f. When we assumed that the random variable X is a nice function, we assumed that $X^{-1}(B)$ is rectifiable when B is a union of intervals. This means that we can consider f such that the pre-image of an interval is a finite union of intervals. In applications, many functions that we deal with satisfy this condition. However, this is another reason to learn about Lebesgue integration, where a composition $f \circ X$ of two nice (Borel measurable) functions is automatically nice (Borel measurable), and we do not need to make unnatural assumptions in the change of variables formula.

Exercise 4.1.2. Compute the expectation of a uniform and exponential random variable.

Exercise 4.1.3. What is the density of the random variable $X = g^3$?

Exercise 4.1.4. Compute the expectation $\mathbb{E}|X|$, where $X \sim N(0, \sigma^2)$.

Exercise 4.1.5. Show that the moments of the standard Gaussian distribution are $\mathbb{E}g^{2k-1} = 0$, and

$$\mathbb{E}g^{2k} = (2k-1)!! = 1 \cdot 3 \cdot 5 \cdots (2k-1)$$

for integer $k \geq 1$. *Hint:* use integration by parts.

Exercise 4.1.6. *(Gaussian integration by parts)* If g is the standard Gaussian random variable, prove that

$$\mathbb{E}gF(g) = \mathbb{E}F'(g)$$

and both sides are well defined if F is continuously differentiable and $|F'(x)| \leq 2 + e^{|x|}$. *Hint:* use integration by parts.

4.2 Gaussian Distributions on $\mathbb{R}^n$

If for $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ we denote its length by

$$|x| = (x_1^2 + \dots + x_n^2)^{1/2}$$

then

$$p_n(x) = \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2} \tag{4.21}$$

is called the density of the *standard Gaussian distribution on* $\mathbb{R}^n$. If we recall the standard Gaussian density $p(x)$ on the real line defined in (4.2), we see that

$$p_n(x) = p(x_1)p(x_2)\cdots p(x_n). \tag{4.22}$$

In other words, p_n is the product of one-dimensional Gaussian densities. We define the *standard Gaussian measure* γ_n on $\mathbb{R}^n$ by

$$\gamma_n(A) = \int_A p_n(x)\,dx = \int_A p_n(x_1, \dots, x_n)\,dx_1 \dots dx_n. \tag{4.23}$$

As in the previous section, we assume that the set A is nice, so that the integral is well defined. By Fubini's theorem,

$$\begin{aligned} \gamma_n(\mathbb{R}^n) &= \int_{\mathbb{R}^n} p(x_1)\cdots p(x_n)\,dx_1 \dots dx_n \\ &= \prod_{i=1}^n \int_{\mathbb{R}} p(x_i)\,dx_i = 1. \end{aligned}$$

In other words, Fubini's theorem tells us that a product of densities on $\mathbb{R}$ is a density on $\mathbb{R}^n$ in the sense that it satisfies basic properties of a probability measure.

This definition as a product corresponds to *independence* of random variables that played such an important role in all the previous chapters in the setting of discrete probability spaces. Specifically, in the case of the space $\Omega = \mathbb{R}^n$ with the probability measure γ_n, the coordinate functions

$$g_1(x) = x_1, \ldots, g_n(x) = x_n \tag{4.24}$$

can be viewed as *independent standard Gaussian random variables*, because, for any nice sets A_i for $i \le n$, by Fubini's theorem,

$$\begin{aligned}\mathbb{P}(g_1 \in A_1, \ldots, g_n \in A_n) &= \gamma_n(A_1 \times \cdots \times A_n) \qquad (4.25)\\ = \int_{A_1 \times \cdots \times A_n} p_n(x)\,dx &= \prod_{i=1}^{n} \int_{A_i} p(x_i)\,dx_i = \prod_{i=1}^{n} \gamma(A_i).\end{aligned}$$

If we take all sets equal to $\mathbb{R}$ except for one, A_i, we get that $\mathbb{P}(g_i \in A_i) = \gamma(A_i)$ and, therefore,

$$\mathbb{P}(g_1 \in A_1, \ldots, g_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(g_i \in A_i).$$

As before, this can be expressed by saying that $g_1, \ldots, g_n$ are independent random variables with the standard Gaussian distribution $N(0,1)$.

Given coordinate functions $g_i(x) = x_i$ in (4.24), consider a vector-valued function

$$g = (g_1, \ldots, g_n) \colon \mathbb{R}^n \to \mathbb{R}^n,$$

which is just the identity function $g(x) = x$ on $\mathbb{R}^n$. Given a rectifiable set $A \subseteq \mathbb{R}^n$, similarly to (4.25), we can write

$$\mathbb{P}(g \in A) = \gamma_n(A) = \int_A p_n(x)\,dx. \tag{4.26}$$

By analogy with continuous random variables, we say that a random vector g has *density* $p_n(x)$. Such random vectors are called *standard Gaussian random vectors*.

As in the previous section, in order to call a function

$$X \colon \mathbb{R}^n \to \mathbb{R}$$

a random variable, we can either assume that it is measurable if we use the Lebesgue integral or, otherwise, we have to restrict ourselves to nice functions. We assume that $X^{-1}(A)$ is a rectifiable set for any interval A, and that X is piecewise-continuous and bounded on compacts, so that all the integrals that we consider are well defined.

For any random variable $X\colon \mathbb{R}^n \to \mathbb{R}$ on the space $(\mathbb{R}^n, \gamma_n)$, its *expectation* is defined by

$$\mathbb{E}X = \int_{\mathbb{R}^n} X(x) p_n(x)\,dx, \tag{4.27}$$

assuming that $\mathbb{E}|X| < \infty$.

To compute the distribution of X, i.e. the probabilities

$$\begin{aligned}\mathbb{P}(X \in A) &= \gamma_n(x : X(x) \in A)\\ &= \gamma_n\big(X^{-1}(A)\big) = \int_{X^{-1}(A)} p_n(x)\,dx\end{aligned}$$

for intervals $A \subseteq \mathbb{R}$, we can try to compute the cumulative distribution function $F(x) = \mathbb{P}(X \le x)$ first and reconstruct the probabilities of other sets from there. As we have seen in the previous section, X might have a discrete distribution or continuous distribution or, in principle, it can have a more general distribution. If we find that its distribution has the density $p_X(t)$, this means that

$$\mathbb{P}(X \in A) = \int_{X^{-1}(A)} p_n(x)\,dx = \int_A p_X(t)\,dt, \tag{4.28}$$

for some function $p_X(t) \ge 0$ such that

$$\int_{\mathbb{R}} p_X(t)\,dt = 1.$$

This is again the *change of variables* formula. As in Lemma 4.2 in the last section, the change of variables formula in (4.28) can be extended to expectations.

Lemma 4.3 (Change of variables for integrals). *If (4.28) holds and* $\mathbb{E}|f(X)| < \infty$ *then*

$$\mathbb{E}f(X) = \int_{\mathbb{R}} f(t)p_X(t)\,dt. \tag{4.29}$$

The proof is identical to the proof of Lemma 4.2, as long as f satisfies certain conditions. In the proof, we needed that $(f \circ X)^{-1}(A) = X^{-1}(f^{-1}(A))$ is rectifiable for any interval A. Or, if we are using the Lebesgue integral, we can take f to be Borel measurable.

Remark 4.1. In this and the last section, we considered two probability spaces $(\Omega, \mathbb{P})$, namely, $(\mathbb{R}, \gamma)$ and $(\mathbb{R}^n, \gamma_n)$. In both settings, given a random variable $X \colon \Omega \to \mathbb{R}$, we said that if its distribution can be written as

$$\mathbb{P}(X \in A) = \int_A p_X(t)\,dt \tag{4.30}$$

for some density function $p_X(t)$ then

$$\mathbb{E}f(X) = \int_{\mathbb{R}} f(t)p_X(t)\,dt. \tag{4.31}$$

The proof of this does not depend much on what $(\Omega, \mathbb{P})$ is and in a more advanced Real Analysis or Measure Theory class this type of change of variables formula is proved in much more generality. □

Let us consider an example of the change of variables, which will be the most important example to us. It describes the *stability property* of Gaussian distributions.

Theorem 4.1 (Gaussian stability). *If* $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ *then the random variable*

$$X = (a, g) = a_1 g_1 + \ldots + a_n g_n$$

has the Gaussian distribution $N(0, |a|^2)$ *with variance* $|a|^2$.

Since $a_i g_i \sim N(0, a_i^2)$, this means that the sum of independent Gaussian random variables is also Gaussian with the variance equal to the sum of variances. We will see that this property will be at the centre of the proof of the Central Limit Theorem below.

Proof. We can assume that $a \neq 0$, since otherwise there is nothing to prove. We will now compute the probability

$$\mathbb{P}(X \leq t) = \mathbb{P}\big((a,g) \leq t\big) = \int_H p_n(x)\, dx,$$

where $t \in \mathbb{R}$ and H is a half-space $H = \{x : (a,x) \leq t\}$.

Let us denote $q_1 = a/|a|$, and let us take arbitrary vectors $q_2, \ldots, q_n$ so that $q_1, q_2, \ldots, q_n$ form an orthonormal basis of $\mathbb{R}^n$. Let Q be the $n \times n$ matrix with rows $q_1, \ldots, q_n$. Then Q is an orthogonal matrix, $\det Q = 1$ and $|Qx| = |x|$ for all $x \in \mathbb{R}^n$. If we make the change of variables $y = Qx$ then, since

$$y_1 = (q_1, x) = \frac{1}{|a|}(a,x),$$

the half-space H can be written as $H = \{y : |a| y_1 \leq t\}$ in the coordinates y. Since Q is orthogonal and $|y| = |Qx| = |x|$,

$$p_n(x) = \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-|y|^2/2} = p_n(y).$$

Since $\det Q = 1$, the Jacobian determinant in the change of variables equals to 1 and

$$\begin{aligned} \int_H p_n(x)\, dx &= \int_H p_n(y)\, dy \\ &= \int_{|a|y_1 \leq t} p(y_1) \cdots p(y_n)\, dy_1 \cdots dy_n \\ \text{(by Fubini)} \ &= \int_{|a|y_1 \leq t} p(y_1)\, dy_1 = \mathbb{P}\big(|a| g_1 \leq t\big), \end{aligned}$$

since the constraint $|a|y_1 \le t$ in the integral depends only on the first coordinate y_1, so the coordinates $y_2,\dots,y_n$ integrated to 1. We showed that

$$\mathbb{P}(X \le t) = \mathbb{P}(|a|g_1 \le t)$$

and, since we already know that $|a|g_1$ has the distribution $N(0,|a|^2)$, this finishes the proof. □

Example 4.2.1. If we use Lemma 4.3 above together with the Gaussian stability property, we get that, for any

$$a = (a_1,\dots,a_n) \in \mathbb{R}^n$$

and a function f that satisfies the conditions of Lemma 4.3,

$$\begin{aligned}&\mathbb{E}f\big((a,g)\big) = \mathbb{E}f(a_1g_1+\ldots+a_ng_n)\\&= \int_{-\infty}^{\infty} f(t)p_{|a|}(t)\,dt = \frac{1}{\sqrt{2\pi}|a|}\int_{-\infty}^{\infty} f(t)e^{-\frac{t^2}{2|a|^2}}\,dt.\end{aligned}$$

For example, in the proof of the central limit theorem, we will use this formula with a smooth bounded monotone f. Since the pre-image $f^{-1}(A)$ of an interval A is an interval and, since $X=(a,g)$, the pre-image $(f\circ X)^{-1}(A)$ is a region bounded by parallel hyperplanes, which is rectifiable.

As we emphasized in the Remark 4.1, the expectation $\mathbb{E}f(X)$ depends only on the distribution of X and not on the specific form of X or probability space on which it is defined, so another way to write the above formula is

$$\mathbb{E}f\big((a,g)\big) = \mathbb{E}f\big(|a|g_1\big), \tag{4.32}$$

because $|a|g_1$ also has the distribution $N(0,|a|^2)$. □

General Gaussian distributions on $\mathbb{R}^n$. Let us consider a linear transformation of the standard Gaussian vector,

$$X = (X_1,\dots,X_n)^T = Ag = A(g_1,\dots,g_n)^T, \tag{4.33}$$

where A is an $n \times n$ matrix. It turns out that the distribution of the random vector X depends on A only through the matrix

$$C = AA^T, \tag{4.34}$$

and it is called the *Gaussian distribution with covariance C*, denoted $N(0,C)$. Before we explain this, let us mention that the *covariance* of a random vector X is the matrix

$$\mathrm{Cov}(X) = \Big[\mathrm{Cov}(X_i, X_j)\Big]_{i,j \le n}, \tag{4.35}$$

assuming that all the entries are well defined. If $\mu = \mathbb{E}X = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)^T$ then the covariance matrix can be written

$$\mathrm{Cov}(X) = \mathbb{E}[(X-\mu)(X-\mu)^T].$$

Since $\mathbb{E}X_i = 0$ for $X = Ag$,

$$\begin{aligned}\mathrm{Cov}(Ag) &= \mathbb{E}Ag(Ag)^T = \mathbb{E}Agg^TA^T \\ &= A(\mathbb{E}gg^T)A^T = AA^T = C,\end{aligned}$$

where we used that $\mathbb{E}gg^T = [\mathbb{E}g_ig_j]_{i,j\le n} = I$ is the identity matrix.

Let us now describe what the distribution $N(0,C)$ looks like. Let us first consider the invertible case $\det(C) \neq 0$.

Lemma 4.4. *If $X = Ag$, $C = AA^T$ and $\det(C) \neq 0$ then the distribution of X has density*

$$p_C(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(C)}} e^{-\frac{1}{2}x^TC^{-1}x}. \tag{4.36}$$

Of course, as in (4.26) above, by analogy with (absolutely) continuous random variables, when we say that a random vector X has density $p_C(x)$, this means that

$$\mathbb{P}(X \in \Delta) = \int_\Delta p_C(x)\,dx \tag{4.37}$$

for any nice set Δ on $\mathbb{R}^n$.

Proof (of Lemma 4.4). For any nice set Δ on $\mathbb{R}^n$, for example a rectangle, we can write

$$\mathbb{P}(Ag \in \Delta) = \mathbb{P}(g \in A^{-1}\Delta) = \int_{A^{-1}\Delta} \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2}\,dx.$$

If we make the change of variables $y = Ax$, or $x = A^{-1}y$, then

$$\mathbb{P}(Ag \in \Delta) = \int_{\Delta} \frac{1}{(2\pi)^{n/2}} e^{-|A^{-1}y|^2/2} \frac{1}{|\det(A)|}\,dy.$$

We have

$$\det(C) = \det(AA^T) = \det(A)\det(A^T) = \det(A)^2,$$

and

$$\begin{aligned}|A^{-1}y|^2 &= (A^{-1}y)^T(A^{-1}y) = y^T(A^T)^{-1}A^{-1}y \\ &= y^T(AA^T)^{-1}y = y^TC^{-1}y.\end{aligned}$$

Therefore, we showed that

$$\begin{aligned}\mathbb{P}(Ag \in \Delta) &= \int_{\Delta} \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{\det(C)}} e^{-\frac{1}{2}y^TC^{-1}y}\,dy \\ &= \int_{\Delta} p_C(y)\,dy,\end{aligned}$$

which means that (4.36) is the density of the distribution $N(0,C)$. □

Example 4.2.2 (Orthogonal maps). In particular, if Q is an orthogonal matrix and $X = Qg$ then $C = QQ^T = I$ and Lemma 4.4 implies that

$$\mathbb{P}(X \in \Delta) = \int_{\Delta} \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2}\,dx. \tag{4.38}$$

In other words, $X = Qg$ has the density $p_I(x) = p_n(x)$, so X is also a standard Gaussian random vector. □

If $\det C = 0$, or $\det A = 0$, the range of the map Ax is a proper subspace of $\mathbb{R}^n$, so the distribution of $X = Ag$ can not have a density on $\mathbb{R}^n$. However, one can still see that it depends only on C and describe it constructively as follows. This is based on the following exercises.

Exercise 4.2.1. Show that a covariance matrix $\mathrm{Cov}(X)$ is symmetric and nonnegative definite.

Exercise 4.2.2. If $C = AA^T$ and $C = QDQ^T$ is an eigen-decomposition of C, for some orthogonal matrix Q and $D = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$, then $A = QD^{1/2}R$ for some orthogonal R and $D^{1/2} = \mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_n})$. *Hint:* use that $B := Q^TA$ satisfies $BB^T = D$ to represent it as $B = D^{1/2}R$.

In the notation of the last exercise, if $\tilde{g} = Rg$ and $q_1,\ldots,q_n$ are the column vectors of Q then

$$X = Ag = QD^{1/2}\tilde{g} = \sqrt{\lambda_1}\tilde{g}_1 q_1 + \ldots + \sqrt{\lambda_n}\tilde{g}_n q_n.$$

Therefore, in the orthonormal basis $q_1,\ldots,q_n$, the random vector X has coordinates $\sqrt{\lambda_1}\tilde{g}_1,\ldots,\sqrt{\lambda_n}\tilde{g}_n$. In this representation, only the matrix R in $\tilde{g} = Rg$ depends on A. However, since R is orthogonal, by the previous lemma, $\tilde{g}$ has the standard Gaussian distribution γ_n, which does not depend on the matrix A. More precisely, the coordinates $\sqrt{\lambda_1}\tilde{g}_1$, $\ldots$, $\sqrt{\lambda_n}\tilde{g}_n$ are independent and have Gaussian distributions with variances $\lambda_1,\ldots,\lambda_n$. When $\det(C) = 0$, some of its eigenvalues will be zero, say, $\lambda_{k+1} = \ldots = \lambda_n = 0$. Then the distribution will be concentrated on the subspace spanned by the first k vectors $q_1,\ldots,q_k$, and it will have the density

$$f(x_1,\ldots,x_k) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\lambda_i}} e^{-x_i^2/2\lambda_i}$$

on this subspace, where $x_1,\ldots,x_k$ are the coordinates in the basis $q_1,\ldots,q_k$. Notice that this description does not depend on the specific choice of A as long as $AA^T = C$, because the eigenvalues and eigenvectors depend only on C. □

Exercise 4.2.3. Consider the matrix

$$C = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}.$$

Find a two-by-two matrix A such that $Ag \sim N(0,C)$, where $g = (g_1,g_2)^T \sim \gamma_2$.

Exercise 4.2.4. Given $a_1,\ldots,a_n \in \mathbb{R}$, consider the matrix

$$C = \big[a_i a_j\big]_{1\le i,j\le n}.$$

Find an $n \times n$ matrix A such that $Ag \sim N(0,C)$, where $g = (g_1,\ldots,g_n)^T \sim \gamma_n$.

Exercise 4.2.5. Given $0 < t_1 < \ldots < t_n$, consider the matrix $C = \big[\min(t_i,t_j)\big]_{1\le i,j\le n}$. Show that the Gaussian distribution $N(0,C)$ has density on $\mathbb{R}^n$.

Exercise 4.2.6. Suppose that $X = (X_1,\ldots,X_n) \sim N(0,C)$, C is invertible, and $C_{1i} = 0$ for $i = 2,\ldots,n$. Prove that X_1 and $(X_2,\ldots,X_n)$ are independent, i.e.

$$\begin{aligned}&\mathbb{P}\big(X_1 \in A_1, X_2 \in A_2,\ldots,X_n \in A_n\big)\\ &\quad = \mathbb{P}\big(X_1 \in A_1\big)\mathbb{P}\big(X_2 \in A_2,\ldots,X_n \in A_n\big).\end{aligned}$$

Exercise 4.2.7. If $a_1^2 + \ldots + a_n^2 = 1$, compute the expectation $\mathbb{E}(a_1 g_1 + \ldots + a_n g_n)^k$ for integer $k \ge 1$.

Exercise 4.2.8. If (g_1,g_2) is a standard Gaussian random vector on $\mathbb{R}^2$, compute the density $p_X(t)$ of the distribution of $X = g_1^2 + g_2^2$.

4.3 Central Limit Theorem

Let us consider random variables $X_1,\dots,X_n$ defined on the same probability space $(\Omega,\mathbb{P})$, and let us suppose that they are independent and have distributions $\mathbb{P}_1,\dots,\mathbb{P}_n$. As usual, this means that

$$\begin{aligned}\mathbb{P}(X_1 \in A_1,\dots,X_n \in A_n) &= \mathbb{P}(X_1 \in A_1)\cdots\mathbb{P}(X_n \in A_n)\\ &= \mathbb{P}_1(A_1)\cdots\mathbb{P}_n(A_n).\end{aligned}$$

Is it possible to construct random variables $X_1,\dots,X_n$ with these prescribed properties? This can be done using the same product space construction as in Chapter 1.

In previous chapters we often worked with independent discrete random variables, but since in this chapter we have encountered continuous random variables, we can suppose that some of the random variables $X_1,\dots,X_n$ are discrete and some are continuous. If a random variable X_i has discrete distribution, let Ω_i be the set of all the values that X_i can take. If X_i has continuous distribution, let $\Omega_i = \mathbb{R}$. Then we can take the space Ω to be the *product space*

$$\Omega = \Omega_1 \times \cdots \times \Omega_n, \tag{4.39}$$

and define the probability $\mathbb{P}$ on Ω by

$$\mathbb{P}(A_1 \times \cdots \times A_n) = \mathbb{P}_1(A_1)\cdots\mathbb{P}_n(A_n), \tag{4.40}$$

where A_i can be any subset of Ω_i in the discrete case, and A_i is a union of intervals when $\Omega_i = \mathbb{R}$. Actually, this definition only specifies the probabilities of rectangles $A_1 \times \ldots \times A_n$ in the product space Ω, so how do we compute probabilities of more general sets? Before we address this, let us first mention that the random variables $X_1,\dots,X_n$ are defined on this space Ω as the coordinate functions

$$X_i(\omega) = X_i(\omega_1,\dots,\omega_n) = \omega_i. \tag{4.41}$$

Then

$$\begin{aligned}&\mathbb{P}(X_1 \in A_1,\ldots,X_n \in A_n) = \mathbb{P}(\omega_1 \in A_1,\ldots,\omega_n \in A_n)\\&= \mathbb{P}(A_1 \times \cdots \times A_n) = \mathbb{P}_1(A_1)\cdots\mathbb{P}_n(A_n),\end{aligned}$$

as we wished. Such construction of random variables as coordinates on the product space is the simplest way to construct independent random variables $X_1,\ldots,X_n$ with prescribed distributions $\mathbb{P}_1,\ldots,\mathbb{P}_n$.

To extend (4.40) to more general sets A, we can define the probability $\mathbb{P}(A)$ as an integral over the set A,

$$\mathbb{P}(A) = \int_A d\mathbb{P}_1(\omega_1)\cdots d\mathbb{P}_n(\omega_n), \tag{4.42}$$

where we understand the integral in the *Fubini sense* (integrating one coordinate at a time) and where each integral $\int_{\Omega_i}\ldots d\mathbb{P}_i(\omega_i)$ is understood as the sum $\sum\ldots\mathbb{P}_i(\omega_i)$ when $\mathbb{P}_i$ is discrete and as the integral $\int\ldots p_{X_i}(\omega_i)\,d\omega_i$ with respect to the density of X_i when $\mathbb{P}_i$ is continuous. Of course, the set A should be such that all these consecutive integrals are well defined.

More generally, an integral of a function $f\colon \Omega \to \mathbb{R}$ on the above product space $(\Omega,\mathbb{P})$,

$$\mathbb{E}f = \int_\Omega f(\omega_1,\ldots,\omega_n)\,d\mathbb{P}_1(\omega_1)\cdots d\mathbb{P}_n(\omega_n), \tag{4.43}$$

is also understood in the Fubini sense. We assume that the function is nice with respect to the continuous coordinates, so the integrals are well defined, and, as usual, assume that $\mathbb{E}|f| < \infty$.

Example 4.3.1. Suppose that ε is a Rademacher random variable, g is a standard Gaussian random variable, and ε and g are independent. Then, if we constructed them as the coordinates on the product space $\Omega = \{-1,1\} \times \mathbb{R}$, we can, for example, compute

$$\mathbb{E}f(g+\varepsilon)=\int_{\mathbb{R}}\frac{1}{2}\Big[f(x-1)+f(x+1)\Big]p(x)dx, \qquad (4.44)$$

where we first averaged over the values ±1 of ε and then integrated with respect to the standard Gaussian distribution of g.

What if the random variables ε and g were defined on a different probability space instead of the product space? Then we could still compute probabilities and expectations using the formula (4.43), as if ε and g were defined on the product space; this is an analogue of the change of variables formula that we proved for a single random variable (discrete or continuous), because we integrate over the range of these random variables with respect to their distributions. The idea of the proof is, basically, the same but in such generality it is better left for a more advanced class. □

Now that we have seen how to define independent random variables with arbitrary prescribed distributions on the same probability space, we are ready to prove the Central Limit Theorem. There are a number of different proofs of the CLT, but the proof below via *Lindeberg's method* is, perhaps, the most transparent. It clearly demonstrates that the Gaussian distribution emerges in the limit due of its stability property proved in Theorem 4.1 in the last section.

Again, let $X_1,\ldots,X_n$ be independent random variables (on the same space) with distributions $\mathbb{P}_1,\ldots,\mathbb{P}_n$. Denote

$$\mu_i=\mathbb{E}X_i \text{ and } \sigma_i^2=\mathrm{Var}(X_i)<\infty. \qquad (4.45)$$

In other words, we suppose that their variances are finite. We will also assume that

$$\frac{\mathbb{E}|X_i|^3}{\sigma_i^3}\le K<\infty \qquad (4.46)$$

for all $i\le n$ for some constant $K>0$. Let

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_i}{\sigma_i}. \tag{4.47}$$

When the random variables are i.i.d. then $\mu_i = \mu$, $\sigma_i^2 = \sigma^2$, and we can write (4.47) as

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}, \tag{4.48}$$

where $S_n = X_1 + \ldots + X_n$ and $\overline{X}_n = S_n/n$. In the i.i.d. case, the assumption (4.46) reduces to $\mathbb{E}|X_1|^3 < \infty$ and with some more work it can be removed. Here is the main result of this chapter.

Theorem 4.2 (Central Limit Theorem). *If the condition (4.46) holds then, for all $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx. \tag{4.49}$$

Since the right hand side is $\mathbb{P}(g \le t)$ for $g \sim N(0,1)$, this type of statement is called *convergence of distributions*, or convergence of the random variables Z_n to g *in distribution*.

Remark 4.2. In general, we say that a sequence of random variables X_n converges in distribution to a random variable X if their cumulative distribution functions $F_n(t) = \mathbb{P}(X_n \le t)$ converge to $F(t) = \mathbb{P}(X \le t)$ for all *points on continuity* t of the c.d.f. $F(t)$. Convergence is not required at the points of discontinuity, where $\mathbb{P}(X = t) > 0$. For example, consider ‘non-random’ random variables $X_n = 1/n$. They converge to $X = 0$ in the usual Calculus sense, so it seems natural that they should converge in distribution, but $\mathbb{P}(X_n \le 0) = 0$, which does not converge to $\mathbb{P}(X \le 0) = 1$. An equivalent definition of convergence in distribution is that

$$\lim_{n\to\infty} \mathbb{E} f(X_n) = \mathbb{E} f(X)$$

for all continuous bounded functions f on $\mathbb{R}$. Here, we will not discuss the general theory of convergence in distribution, although the argument that shows the equivalence of the two definitions we mentioned is quite simple and its variant will appear in the proof of the CLT. □

Proof (Theorem 4.2). Before we start the main argument of the proof, let us reduce the claim to a statement which will be more suitable to us.

First, since the probability of a set is also an expectation of its indicator, we can write

$$\mathbb{P}(Z_n \le t) = \mathbb{E}\,\mathrm{I}(Z_n \le t).$$

However, instead of working with an indicator, it will be convenient to work with a smooth function that approximates the indicator, because this will allow us to use the Taylor theorem. Given small $\varepsilon > 0$, let us consider a smooth non-increasing function $\varphi(x)$ such that (see Figure 4.1)

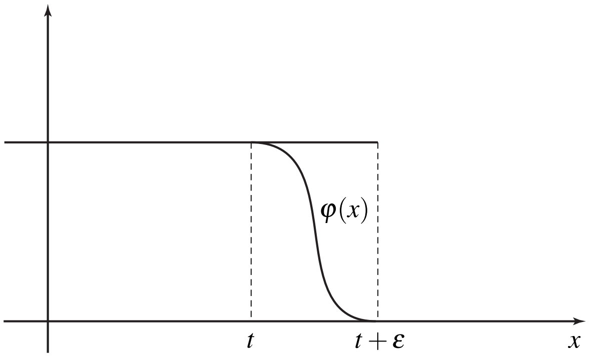


Fig. 4.1 Smooth approximation $\varphi(x)$ of an indicator function $\mathrm{I}(x \le t)$.

$$\mathrm{I}(x \le t) \le \varphi(x) \le \mathrm{I}(x \le t+\varepsilon), \tag{4.50}$$

and $|\varphi'''(x)| \le c$, for some constant c that may depend on ε. In other words, $\varphi(x)$ equals 1 for $x \le t$, $\varphi(x)$ equals 0 for $x \ge t+\varepsilon$, and smoothly decreases in between t and $t+\varepsilon$.

Such function is easy to construct, so we will leave it as an exercise. Then, in order to prove the theorem, all we need to show is

$$\lim_{n\to\infty} \mathbb{E}\varphi(Z_n) = \mathbb{E}\varphi(g) = \int_{\mathbb{R}} \varphi(x)p(x)\,dx, \tag{4.51}$$

where g is a standard Gaussian random variable and $p(x)$ is its density,

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The reason why this is enough is because, by (4.50),

$$\mathbb{E}\,\mathrm{I}(Z_n \le t) \le \mathbb{E}\varphi(Z_n) \le \mathbb{E}\,\mathrm{I}(Z_n \le t+\varepsilon).$$

The first inequality implies, assuming (4.51),

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}\,\mathrm{I}(Z_n \le t) &\le \lim_{n\to\infty} \mathbb{E}\varphi(Z_n) \\ &= \int_{\mathbb{R}} \varphi(x)p(x)\,dx \le \int_{-\infty}^{t+\varepsilon} p(x)\,dx,\end{aligned}$$

where the last inequality holds because $\varphi(x) \le \mathrm{I}(x \le t+\varepsilon)$. The second inequality implies, assuming (4.51),

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}\,\mathrm{I}(Z_n \le t+\varepsilon) &\ge \lim_{n\to\infty} \mathbb{E}\varphi(Z_n) \\ &= \int_{\mathbb{R}} \varphi(x)p(x)\,dx \ge \int_{-\infty}^{t} p(x)\,dx,\end{aligned}$$

where the last inequality holds because $\varphi(x) \ge \mathrm{I}(x \le t)$. Since t here is arbitrary, using this with $t-\varepsilon$ gives

$$\liminf_{n\to\infty} \mathbb{E}\,\mathrm{I}(Z_n \le t) \ge \int_{-\infty}^{t-\varepsilon} p(x)\,dx.$$

Together these two inequalities imply that $\mathbb{E}\,\mathrm{I}(Z_n \le t)$ is squeezed in the limit in between

$$\int_{-\infty}^{t-\varepsilon} p(x)\,dx \le \lim_{n\to\infty} \mathbb{E}\,\mathrm{I}(Z_n \le t) \le \int_{-\infty}^{t+\varepsilon} p(x)\,dx.$$

Since $\varepsilon > 0$ was arbitrary, letting $\varepsilon \downarrow 0$ yields (4.49).

We will now focus on proving (4.51). To simplify the notation, let us introduce

$$Y_i = \frac{X_i - \mu_i}{\sigma_i},$$

so that

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

All the assumptions on X_i's translate into

$$\mathbb{E}Y_i = 0, \mathbb{E}Y_i^2 = 1, \text{ and } \mathbb{E}|Y_i|^3 \le K < \infty. \tag{4.52}$$

Let us now suppose that the random variables $X_1,\ldots,X_n$ are constructed on the same probability space together with i.i.d. standard Gaussian random variables $g_1,\ldots,g_n$. The discussion at the beginning of this section explains how this can be done. We can just think of $g_1,\ldots,g_n$ as $X_{n+1},\ldots,X_{2n}$ and use the product space with $2n$ coordinates instead of n. Consider the random variable

$$g = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i. \tag{4.53}$$

By the stability property of Gaussians proved in Theorem 4.1, the distribution of g is standard Gaussian, $g \sim N(0,1)$, and, in particular, by the change of variables formula in Lemma 4.3,

$$\mathbb{E}\varphi\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i\Big) = \int_{\mathbb{R}} \varphi(x) p(x)\,dx.$$

In other words, this choice of g agrees with the notation in (4.51), where g also denoted standard Gaussian, and, in order to prove (4.51), it is enough to show that the difference

$$\mathbb{E}\varphi\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\Big) - \mathbb{E}\varphi\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\Big) \tag{4.54}$$

is getting small for large n.

The idea of Lindeberg's method is to compare

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \text{ and } g = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i$$

by 'taking small steps' and by replacing $Y_i/\sqrt{n}$ with $g_i/\sqrt{n}$ one by one, as follows. This is where the stability property of Gaussians was so crucial, allowing us to represent g as a sum of similar small increments. For $1 \le i \le n+1$, we define

$$T_i = \frac{1}{\sqrt{n}}\big(g_1 + \ldots + g_{i-1} + Y_i + \ldots + Y_n\big),$$

so that at the first step $T_1 = Z_n$ and at the last step $T_{n+1} = g$. Then we can write (4.54)

$$\mathbb{E}\varphi(Z_n) - \mathbb{E}\varphi(g) = \sum_{i=1}^{n}\big(\mathbb{E}\varphi(T_i) - \mathbb{E}\varphi(T_{i+1})\big)$$

as a telescoping sum of 'small increments' because T_i differs from T_{i+1} only in one term, $Y_i/\sqrt{n}$ instead of $g_i/\sqrt{n}$. They share all the other terms, which we will denote by

$$S_i = \frac{1}{\sqrt{n}}\big(g_1 + \ldots + g_{i-1} + Y_{i+1} + \ldots + Y_n\big),$$

so that the consecutive terms are

$$T_i = S_i + \frac{Y_i}{\sqrt{n}},\ T_{i+1} = S_i + \frac{g_i}{\sqrt{n}}.$$

The telescoping sum representation implies that

$$\big|\mathbb{E}\varphi(Z_n)-\mathbb{E}\varphi(g)\big|\le\sum_{i=1}^{n}\big|\mathbb{E}\varphi(T_i)-\mathbb{E}\varphi(T_{i+1})\big|.$$

Since $|\varphi'''(x)|\le c$, if we expand $\varphi(T_i)$ and $\varphi(T_{i+1})$ around S_i using the Taylor polynomials of order 2, we get

$$\Big|\varphi(T_i)-\varphi(S_i)-\varphi'(S_i)\frac{Y_i}{\sqrt{n}}-\varphi''(S_i)\frac{Y_i^2}{2n}\Big|\le\frac{c|Y_i|^3}{6n^{3/2}},$$

$$\Big|\varphi(T_{i+1})-\varphi(S_i)-\varphi'(S_i)\frac{g_i}{\sqrt{n}}-\varphi''(S_i)\frac{g_i^2}{2n}\Big|\le\frac{c|g_i|^3}{6n^{3/2}}.$$

Now we open the absolute values by writing $-b\le a\le b$ instead of $|a|\le b$ and take expectations. When we take the expectations of the first and second order terms,

$$\mathbb{E}\Big[\varphi'(S_i)\frac{Y_i}{\sqrt{n}}\Big],\mathbb{E}\Big[\varphi''(S_i)\frac{Y_i^2}{2n}\Big],\mathbb{E}\Big[\varphi'(S_i)\frac{g_i}{\sqrt{n}}\Big],\mathbb{E}\Big[\varphi''(S_i)\frac{g_i^2}{2n}\Big],$$

we notice that S_i is a function of the coordinates that do not include Y_i and g_i and, by independence and the Fubini formula (4.43),

$$\mathbb{E}\Big[\varphi'(S_i)\frac{Y_i}{\sqrt{n}}\Big]=\mathbb{E}\varphi'(S_i)\mathbb{E}\frac{Y_i}{\sqrt{n}}=0,$$

since $\mathbb{E}Y_i=0$, and

$$\mathbb{E}\Big[\varphi'(S_i)\frac{g_i}{\sqrt{n}}\Big]=\mathbb{E}\varphi'(S_i)\mathbb{E}\frac{g_i}{\sqrt{n}}=0,$$

since $\mathbb{E}g_i=0$. As we can see, because we subtracted the expectation in $Y_i=(X_i-\mu_i)/\sigma_i$, the first order terms in the Taylor expansions match. Similarly,

$$\mathbb{E}\Big[\varphi''(S_i)\frac{Y_i^2}{2n}\Big]=\mathbb{E}\varphi''(S_i)\mathbb{E}\frac{Y_i^2}{2n}=\frac{1}{2n}\mathbb{E}\varphi''(S_i)$$

since $\mathbb{E}Y_i^2=1$, and

$$\mathbb{E}\Big[\varphi''(S_i)\frac{g_i^2}{2n}\Big]=\mathbb{E}\varphi''(S_i)\mathbb{E}\frac{g_i^2}{2n}=\frac{1}{2n}\mathbb{E}\varphi''(S_i)$$

since $\mathbb{E}g_i^2=1$. Because we scaled by σ_i in $Y_i=(X_i-\mu_i)/\sigma_i$, the second order terms in the Taylor expansions match. The zeroth order terms were the same, $\mathbb{E}\varphi(S_i)$, and, as a result, the difference is only in the third order error terms,

$$\Big|\mathbb{E}\varphi(T_i)-\mathbb{E}\varphi(T_{i+1})\Big|\le\frac{c(\mathbb{E}|Y_i|^3+\mathbb{E}|g_i|^3)}{6n^{3/2}}.$$

By (4.52), $\mathbb{E}|Y_i|^3\le K$, and $\mathbb{E}|g_i|^3$ is the same for all $i\le n$, so it is just another constant and

$$\Big|\mathbb{E}\varphi(T_i)-\mathbb{E}\varphi(T_{i+1})\Big|\le\frac{c'}{n^{3/2}}$$

for some constant c'. These third order terms are of smaller order $1/n^{3/2}$ than the number of 'steps' in the telescoping sum, n, and we finally get

$$\big|\mathbb{E}\varphi(Z_n)-\mathbb{E}\varphi(g)\big|\le\sum_{i=1}^{n}\big|\mathbb{E}\varphi(T_i)-\mathbb{E}\varphi(T_{i+1})\big|\le\frac{c'}{\sqrt{n}}.$$

This finishes the proof. □

Example 4.3.2. Suppose that all we know about the checkout time at a grocery store is that it has a mean of $\mu=5$ minutes and a standard deviation of $\sigma=2$ minutes. Let us estimate the probability that a checker will serve at least 50 customers during her 4-hour shift. In other words, what is the probability that 50 customers will be served in under 240 minutes? Let T_i be the time it takes to serve the i^{th} customer, and $S_n=\sum_{i=1}^{n}T_i$. Then $S_{50}\le 240$ is equivalent to

$$Z_{50}=\frac{S_{50}-50\cdot 5}{2\sqrt{50}}\le\frac{240-50\cdot 5}{2\sqrt{50}}\approx-0.7.$$

By the central limit theorem, the probability of this event can be approximated by

$$\mathbb{P}(Z_{50} \le -0.7) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-0.7} e^{-x^2/2}\, dx \approx 0.242.$$

So, we can estimate the chance to be about 24%. □

Sometimes one is interested in a sequence of independent random variables $X_1, \ldots, X_n$ such that all their distributions also depend on n. In other words, for each n, we consider $X_{n1}, \ldots, X_{nn}$, which are independent but, otherwise, the whole sequence varies with n. This setting is called the *triangular array*. Let us denote

$$\mu_{ni} = \mathbb{E}X_{ni},\ \sigma_{ni}^2 = \mathrm{Var}(X_{ni}) < \infty,\ D_n^2 = \sum_{i=1}^n \sigma_{ni}^2. \tag{4.55}$$

Instead of (4.47), we will consider a different rescaling,

$$Z_n = \frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}(S_n)}} = \frac{1}{D_n} \sum_{i=1}^n \big(X_{ni} - \mu_{ni}\big). \tag{4.56}$$

For i.i.d. sequences, the two definitions of Z_n are the same. We still have $\mathbb{E}Z_n = 0$ and $\mathrm{Var}(Z_n) = 1$, but the variance of each summand, σ_{ni}^2/D_n^2, is not necessarily $1/n$ as was the case in (4.47). When does Z_n satisfy the CLT?

One condition that ensures this is

$$\lim_{n\to\infty} \frac{1}{D_n^3} \sum_{i=1}^n \mathbb{E}\big|X_{ni} - \mu_{ni}\big|^3 = 0, \tag{4.57}$$

which is called the *Lyapunov condition*.

Theorem 4.3 (CLT for triangular arrays). *The central limit theorem (4.49) holds for the triangular arrays that satisfy the Lyapunov condition (4.57).*

The proof is almost identical to the above, so we will leave it as an exercise. One small change is to consider $a_{ni} = \sigma_{ni}/D_n$

and, instead of (4.53), represent $g = \sum_{i=1}^n a_{ni} g_i$. At some point, when controlling one of the error terms in the Taylor expansion, one also has to use that

$$\sigma_{ni}^3 = \Big(\mathbb{E}\big|X_{ni} - \mu_{ni}\big|^2\Big)^{3/2} \le \mathbb{E}\big|X_{ni} - \mu_{ni}\big|^3,$$

which follows from Jensen's inequality.

Example 4.3.3 (Records in a random permutation). Given a permutation $(\pi_1, \ldots, \pi_n)$ of $\{1, \ldots, n\}$, we say that π_k is *a record* if it is greater than the preceding values,

$$\pi_k > \pi_i \text{ for } i = 1, \ldots, k-1.$$

For example, in a permutation $(3,5,2,6,4,1)$, there are three records: $\pi_1 = 3, \pi_2 = 5$ and $\pi_4 = 6$. We will show that the number N_n of records in a random permutation of $\{1, \ldots, n\}$ satisfies the central limit theorem:

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{N_n - \ln n}{\sqrt{\ln n}} \le t\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-x^2/2}\, dx. \tag{4.58}$$

Let us consider the indicator that π_k is a record,

$$r_k = \begin{cases} 1, & \text{if } \pi_k \text{ is a record,} \\ 0, & \text{otherwise.} \end{cases}$$

Then $N_n = \sum_{k=1}^n r_k$. In order to apply Theorem 4.3, we will prove that $r_1, \ldots, r_n$ are independent random variables and

$$r_k \sim B\Big(\frac{1}{k}\Big) \text{ for } k \le n. \tag{4.59}$$

This might be counterintuitive, because, observing many records at the beginning of the permutation might suggest that the chances of further records are affected, but this is not the case. We will check this carefully, but the basic idea is that, no matter what the values $\{\pi_1, \ldots, \pi_k\}$ are, only their order and not their values determines the records $(r_1, \ldots, r_k)$, so knowing the records in the first k positions gives us no information about the values.

Our goal is to show that, for any $x_1,\ldots,x_n \in \{0,1\}$,

$$\mathbb{P}\big(r_1 = x_1,\ldots,r_n = x_n\big) = \prod_{k=1}^{n}\Big(\frac{1}{k}\Big)^{x_k}\Big(1-\frac{1}{k}\Big)^{1-x_k}.$$

Equivalently, if we multiply both sides by $n!$ and denote by $\#(A)$ the number of permutations in the set A, we want to show that

$$\#\big(r_1 = x_1,\ldots,r_n = x_n\big) = \prod_{k=1}^{n}(k-1)^{1-x_k}. \qquad (4.60)$$

It is easier to explain the proof of this formula on a specific example. Suppose that $n = 6$ and the sequence of records is

$$(r_1,r_2,r_3,r_4,r_5,r_6) = (1,1,0,1,0,0).$$

First of all, since the last record is at $k = 4$, we must have $\pi_4 = n = 6$, because 6 will always be a record and there can be no records after 6. Second, once we know that $\pi_4 = 6$, the numbers π_5 and π_6 can be anything but 6. There are 5×4 ways to choose these numbers. Third, no matter what π_1,π_2 and π_3 are, the number of ways to arrange them that will result in $(r_1,r_2,r_3) = (1,1,0)$ is the same; in other words, the specific values do not matter, since we simply compare which one is bigger than the other. This is a good place to use induction. If we assume that we already proved the above formula for $n = 3$, the number of permutations of any three distinct numbers that result in $(r_1,r_2,r_3) = (1,1,0)$ is

$$\#(r_1 = 1, r_2 = 1, r_3 = 0) = 0^0 \times 1^0 \times 2^1 = 2.$$

Multiplying this by 4×5 ways to choose π_5 and π_6 gives $2\times 4\times 5$, which agrees with the formula (4.60). The proof of the general case is exactly the same.

To apply Theorem 4.3, we need to compute the mean and variance of $N_n = \sum_{k=1}^{n} r_k$, and check the Lyapunov condition (4.57). Since the mean and variance of the Bernoulli $B(p)$ random variable is p and $p(1-p)$, and the random variables $r_1, \ldots, r_n$ are independent,

$$\mathbb{E}N_n = \sum_{k=1}^{n} \frac{1}{k}, \ \mathrm{Var}(N_n) = \sum_{k=1}^{n} \frac{1}{k}\Big(1 - \frac{1}{k}\Big).$$

In particular, both quantities are within a constant of $\ln n$,

$$\big|\mathbb{E}N_n - \ln n\big| \le c, \ \big|\mathrm{Var}(N_n) - \ln n\big| \le c, \tag{4.61}$$

and $D_n^3 = \mathrm{Var}(N_n)^{3/2} \sim (\ln n)^{3/2}$. Next,

$$\sum_{k=1}^{n} \mathbb{E}\Big|r_k - \frac{1}{k}\Big|^3 = \sum_{k=1}^{n} \Big(\frac{k-1}{k}\Big)^3 \frac{1}{k} + \sum_{k=1}^{n} \Big(\frac{1}{k}\Big)^3 \frac{k-1}{k}.$$

The second sum is bounded by a constant, and the first sum is bounded by $\mathbb{E}N_n \le \ln n + 1$. This implies that the Lyapunov condition (4.57) is satisfied and, therefore,

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{N_n - \mathbb{E}N_n}{\sqrt{\mathrm{Var}(N_n)}} \le t\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\,dx.$$

Together with (4.61), this implies (4.58). □

Exercise 4.3.1. If the random variables $X_1, \ldots, X_n$ are i.i.d. Bernoulli $B(p)$, what is the limit $\lim_{n\to\infty} \mathbb{P}\big(|\overline{X}_n - p| \ge t_n\big)$ if $\lim_{n\to\infty} \sqrt{n} t_n = t \in [0, \infty)$?

Exercise 4.3.2. If $X \sim \mathrm{Poiss}(\lambda)$ is Poisson with mean $\lambda > 0$, what is the limit $\lim_{\lambda\to\infty} \mathbb{P}\big(|X - \lambda| \ge a\sqrt{\lambda}\big)$ for $a > 0$? *Hint:* recall the stability property of Poisson.

Exercise 4.3.3. 24% of the residents in a community are members of a minority group but among the 96 people called for jury duty only 13 are. Does this data indicate that minorities are less likely to be called for jury duty?

Exercise 4.3.4. Members of the $A\Sigma\Phi$ fraternity each drink a random number of beers with mean 6 and standard deviation 3. If there are 81 fraternity members, how much should they buy so that using the normal approximation they are 95% sure they will not run out?

Exercise 4.3.5. Prove Theorem 4.3.

Exercise 4.3.6. Consider a triangular array of i.i.d. Bernoulli $X_{n1}, \ldots, X_{nn} \sim B(p_n)$, where p_n varies with n. If $p_n \to 0$ but $np_n \to +\infty$, show that Z_n in (4.56) satisfies the CLT. Check that the condition (4.46) fails.

Exercise 4.3.7. Let $X_1, X_2, \ldots$ be independent random variables such that $X_i \sim B(p_i)$ for some $p_i \in [0,1]$. If the series $\sum_{i=1}^{\infty} p_i(1-p_i)$ diverges, show that Z_n in (4.56) satisfies the CLT.

Exercise 4.3.8. Let $X_1, X_2, \ldots$ be independent random variables such that $\mathbb{P}(X_i = \pm i^{\alpha}) = \frac{1}{2}$ for some $\alpha > -\frac{1}{2}$. Show that Z_n in (4.56) satisfies the CLT.

4.4 Distributions Related to Gaussian

In this section we will introduce several distributions related to Gaussian: Gamma distribution, chi-squared distribution, F-distribution, and Student's t-distribution. This will give us another chance to study various transformations of random variables (such as sums and ratios), but the main reason we discuss these distributions is their central role in the analysis of Simple Linear Regression in the next section.

Gamma distribution. Let us recall the definition of the Gamma function $\Gamma(\alpha)$ for $\alpha > 0$,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx. \tag{4.62}$$

If we divide both sides by $\Gamma(\alpha)$ and, given $\beta > 0$, make the change of variables $x = \beta y$, we get

$$\begin{aligned} 1 &= \int_0^\infty \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}\, dx \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}\, dy. \end{aligned}$$

Therefore, if we define

$$f(x;\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & x \le 0 \end{cases} \tag{4.63}$$

then $f(x;\alpha,\beta)$ is a probability density function, since it is nonnegative and integrates to one. The distribution with the density $f(x;\alpha,\beta)$ is called *Gamma distribution with parameters α and β*, and it is denoted by $\Gamma(\alpha,\beta)$.

Next, let us recall some properties of $\Gamma(\alpha)$. If $\alpha > 1$ then, using integration by parts,

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty x^{\alpha-1}e^{-x}dx = \int_0^\infty x^{\alpha-1}d(-e^{-x})\\ &= x^{\alpha-1}(-e^{-x})\Big|_0^\infty - \int_0^\infty (-e^{-x})(\alpha-1)x^{\alpha-2}dx\\ &= (\alpha-1)\int_0^\infty x^{(\alpha-1)-1}e^{-x}dx = (\alpha-1)\Gamma(\alpha-1).\end{aligned}$$

Since $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$, by induction, we get that

$$\Gamma(n) = (n-1)!$$

for integer $n \geq 1$. Using $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, one can easily check the following.

Exercise 4.4.1. Show that the kth moment of $X \sim \Gamma(\alpha,\beta)$ equals

$$\mathbb{E}X^k = \frac{(\alpha+k-1)\cdots\alpha}{\beta^k}.$$

In particular, the expectation, second moment and variance are

$$\mathbb{E}X = \frac{\alpha}{\beta},\ \mathbb{E}X^2 = \frac{(\alpha+1)\alpha}{\beta^2},\ \mathrm{Var}(X) = \frac{\alpha}{\beta^2}.$$

The most important property of Gamma distribution to us will be the following *stability property*.

Lemma 4.5. *If the random variables $X_1,\ldots,X_n$ are independent and $X_1 \sim \Gamma(\alpha_1,\beta),\ldots,X_n \sim \Gamma(\alpha_n,\beta)$ then their sum $X_1+\ldots+X_n$ has distribution $\Gamma(\alpha_1+\ldots+\alpha_n,\beta)$.*

Notice that the scaling parameter β is the same. This lemma is based on the classical *convolution* formula for the density of the sum of two independent random variables whose distributions have densities.

Lemma 4.6. *If X_1, X_2 are independent and their distributions have densities $f(x), g(x)$ then X_1+X_2 has density*

$$h(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\,dy. \tag{4.64}$$

Proof. Let us compute the cumulative distribution function of X_1+X_2. By independence and the Fubini representation (4.42) in the previous section,

$$\begin{aligned}
\mathbb{P}(X_1+X_2\le t) &= \iint_{\mathbb{R}^2} \mathrm{I}(x+y\le t) f(x)g(y)\,dxdy \\
&= \iint_{\mathbb{R}^2} \mathrm{I}(x\le t-y) f(x)g(y)\,dxdy \\
&= \int_{-\infty}^{\infty}\Big[\int_{-\infty}^{t-y} f(x)g(y)\,dx\Big]dy \\
\{x=z-y\}\ &= \int_{-\infty}^{\infty}\Big[\int_{-\infty}^{t} f(z-y)g(y)\,dz\Big]dy \\
\{\text{switch order}\}\ &= \int_{-\infty}^{t}\Big[\int_{-\infty}^{\infty} f(z-y)g(y)\,dy\Big]dz \\
&= \int_{-\infty}^{t} h(z)\,dz,
\end{aligned}$$

where $h(x)$ coincides with (4.64). This shows that h is the density of X_1+X_2 and finishes the proof. □

Using this convolution formula, we can prove Lemma 4.5.

Proof (Lemma 4.5). Let us start with $X_1\sim\Gamma(\alpha_1,\beta)$ and $X_2\sim\Gamma(\alpha_2,\beta)$ and compute the distribution of their sum X_1+X_2. Since

$$\begin{aligned}
&f(x-y;\alpha_1,\beta)f(y;\alpha_2,\beta) \\
&= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta x}(x-y)^{\alpha_1-1}y^{\alpha_2-1}
\end{aligned}$$

when both $x-y\ge 0$ and $y\ge 0$, and 0 otherwise, by (4.64), the density of X_1+X_2 is

$$h(x)=\frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta x}\int_0^x (x-y)^{\alpha_1-1}y^{\alpha_2-1}\,dy.$$

Making the change of variables $y=xz$, this equals

$$\frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1+\alpha_2-1} e^{-\beta x} \int_0^1 (1-z)^{\alpha_1-1} z^{\alpha_2-1}\, dz.$$

The factors $x^{\alpha_1+\alpha_2-1}e^{-\beta x}$ that depend on x look exactly like the ones in the Gamma density $f(x;\alpha_1+\alpha_2,\beta)$ and, because $h(x)$ is a density and must integrate to 1, we must have

$$h(x) = f(x;\alpha_1+\alpha_2,\beta) = \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)} x^{\alpha_1+\alpha_2-1} e^{-\beta x}.$$

This proves the claim in the case of two Gamma random variables. As a byproduct, we also showed that

$$\int_0^1 (1-z)^{\alpha_1-1} z^{\alpha_2-1}\, dz = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

This integral is known as the Beta function $\mathrm{B}(\alpha_1,\alpha_2)$, so

$$\mathrm{B}(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

For the sum of more than two, one can proceed by induction on n, although it is a good idea to check the details. □

Exercise 4.4.2. Complete the induction step at the end of the above proof. *Hint:* use the change of variables formula (think about Remark 4.1), or consider the proof of the convolution formula above.

One can also use the convolution formula to prove the Gaussian stability property in Theorem 4.1.

Exercise 4.4.3. Prove Theorem 4.1 using the convolution formula (4.64). *Hint:* Recall notation in Example 4.1.6, and check that

$$p_\sigma(x) = \int_{-\infty}^{\infty} p_{\sigma_1}(x-y) p_{\sigma_2}(y)\, dy,$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Then use induction.

Chi-squared distribution. If $g_1,\ldots,g_n$ are i.i.d. standard Gaussian random variables then the distribution of the sum

$$g_1^2+\ldots+g_n^2 \tag{4.65}$$

is called the *chi-squared (or chi-square) distribution with n degrees of freedom*, and it is denoted χ_n^2. In the Example 4.1.5 in Section 4.1, we showed that the square $X=g^2$ of a standard Gaussian random variable g has density

$$p_X(x)=\frac{1}{\sqrt{2\pi}}x^{\frac{1}{2}-1}e^{-\frac{x}{2}}\,\mathrm{I}(x>0).$$

Comparing with the definition of the Gamma distribution in (4.63), we can see that $p_X(x)$ is the density of $\Gamma(\frac{1}{2},\frac{1}{2})$. The constants $\sqrt{2}\Gamma(\frac{1}{2})$ and $\sqrt{2\pi}$ have no choice but to agree, so $\Gamma(\frac{1}{2})=\sqrt{\pi}$.

By the stability property of the Gamma distribution in Lemma 4.5, we see that

$$\chi_n^2=\Gamma\Bigl(\frac{n}{2},\frac{1}{2}\Bigr). \tag{4.66}$$

This is a good place to mention one more time that here we implicitly use the change of variables formula in Lemma 4.3 in Section 4.2. When we compute, for example, the c.d.f. $\mathbb{P}(g_1^2+\ldots+g_n^2\le t)$, by definition, we integrate

$$\int_{\mathbb{R}^n}\mathrm{I}(x_1^2+\ldots+x_n^2\le t)p(x_1)\cdots p(x_n)\,dx_1\cdots dx_n,$$

where $p(x)$ is the standard Gaussian density. The change of variables formula in Lemma 4.3 allows us to replace

$$\int\ldots p(x_i)\,dx_i \text{ by } \int\ldots f(t_i)\,dt_i,$$

where $f(t_i)$ is the $\Gamma(1/2,1/2)$ density $f(t_i;1/2,1/2)$, by making the change of variables $t_i=x_i^2$ and using that, for $g\sim N(0,1)$, the distribution of g^2 is $\Gamma(1/2,1/2)$. Once the integral is rewritten in terms of the Gamma densities,

$$\int_{\mathbb{R}^n} \mathrm{I}(t_1+\ldots+t_n \leq t) f(t_1)\cdots f(t_n)\, dt_1 \cdots dt_n,$$

only then we apply the stability property in Lemma 4.5.

F-**distribution.** Consider two independent random variables with the distributions

$$X \sim \chi_k^2 = \Gamma\Big(\frac{k}{2},\frac{1}{2}\Big) \text{ and } Y \sim \chi_m^2 = \Gamma\Big(\frac{m}{2},\frac{1}{2}\Big).$$

The distribution of the ratio

$$Z = \frac{X/k}{Y/m} \tag{4.67}$$

is called *F-distribution with degrees of freedom k and m*, and is denoted by $F_{k,m}$. (It is also known as the Fisher–Snedecor distribution.) In other words, if

$$g_1,\ldots,g_k,\tilde{g}_1,\ldots,\tilde{g}_m$$

are i.i.d. standard Gaussian, then

$$\frac{m}{k}\frac{g_1^2+\ldots+g_k^2}{\tilde{g}_1^2+\ldots+\tilde{g}_m^2}$$

has distribution $F_{k,m}$. We will show that the density of $F_{k,m}$ is

$$f_{k,m}(x) = \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})} k^{k/2} m^{m/2} x^{\frac{k}{2}-1}(m+kx)^{-\frac{k+m}{2}} \tag{4.68}$$

for $x > 0$, and zero otherwise. We will need the following lemma.

Lemma 4.7. *If two positive random variables $X,Y > 0$ have densities $f(x)$ and $g(x)$, then $Z = X/Y$ has density*

$$h(x) = \int_0^\infty f(xy)g(y)y\,dy \ \textit{for}\ x > 0. \tag{4.69}$$

Proof. For $t > 0$,

$$\begin{aligned}
\mathbb{P}(Z \leq t) = \mathbb{P}(X \leq tY) &= \int_0^\infty \Big[\int_0^{ty} f(x)g(y)\,dx\Big]dy \\
\{x = zy\} \ &= \int_0^\infty \Big[\int_0^{t} f(zy)g(y)y\,dz\Big]dy \\
\{\text{switch order}\} \ &= \int_0^t \Big[\int_0^{\infty} f(zy)g(y)y\,dy\Big]dz \\
&= \int_0^t h(z)dz,
\end{aligned}$$

where $h(x)$ coincides with (4.69). This finishes the proof. □

To compute the density of (4.67), let us first compute the density of $\frac{k}{m}Z = \frac{X}{Y}$. The densities of X and Y are

$$f(x) = \frac{(\frac{1}{2})^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}x^{\frac{k}{2}-1}e^{-\frac{1}{2}x} \ \text{ and } \ g(y) = \frac{(\frac{1}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})}y^{\frac{m}{2}-1}e^{-\frac{1}{2}y}$$

correspondingly, for $x > 0$ and $y > 0$. By the previous lemma, the density of X/Y is

$$\begin{aligned}
f_{X/Y}(x) &= \int_0^\infty \frac{(\frac{1}{2})^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}(xy)^{\frac{k}{2}-1}e^{-\frac{1}{2}xy}\frac{(\frac{1}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})}y^{\frac{m}{2}-1}e^{-\frac{1}{2}y}y\,dy \\
&= \frac{(\frac{1}{2})^{\frac{k+m}{2}}}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}x^{\frac{k}{2}-1}\int_0^\infty y^{\frac{k+m}{2}-1}e^{-\frac{1}{2}(x+1)y}\,dy.
\end{aligned}$$

If we make the change of variables $z = \frac{1}{2}(x+1)y$, we get

$$\begin{aligned}
f_{X/Y}(x) &= \frac{1}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}x^{\frac{k}{2}-1}(1+x)^{-\frac{k+m}{2}}\int_0^\infty z^{\frac{k+m}{2}-1}e^{-z}\,dz \\
&= \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}x^{\frac{k}{2}-1}(1+x)^{-\frac{k+m}{2}},
\end{aligned}$$

where we used (4.62). Finally, we obtain (4.68) using that

$$f_Z(x) = f_{X/Y}\left(\frac{kx}{m}\right)\frac{k}{m}.$$

Student's t-distribution. If $g_0, g_1, \dots, g_n$ are i.i.d. standard Gaussian then the distribution of

$$T = \frac{g_0}{\sqrt{\frac{1}{n}(g_1^2 + \ldots + g_n^2)}} \tag{4.70}$$

is called the *t-distribution with n degrees of freedom*, and it is denoted by t_n.

To compute the density $f_T(x)$ of T, we only need to notice that T^2 has $F_{1,n}$-distribution and that the distribution of T is symmetric, so $f_T(x) = f_T(-x)$. If we write, for $t > 0$,

$$2\mathbb{P}(0 \le T \le t) = \mathbb{P}(-t \le T \le t) = \mathbb{P}(T^2 \le t^2)$$

as integrals of the densities, we get

$$2\int_0^t f_T(x)\,dx = \int_0^{t^2} f_{1,n}(x)\,dx.$$

Making the change of variables $x = y^2$ on the right hand side,

$$2\int_0^t f_T(x)\,dx = \int_0^t f_{1,n}(y^2)2y\,dy.$$

Taking derivatives of both side with respect to t, for $t > 0$,

$$f_T(t) = f_{1,n}(t^2)t = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}\frac{1}{\sqrt{n}}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}. \tag{4.71}$$

Since $f_T(-t) = f_T(t)$, the same formula holds for all $t \in \mathbb{R}$. It is not difficult to show that

$$\lim_{n\to\infty} f_T(t) = p(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}, \tag{4.72}$$

but we will skip the technical details here. Intuitively this should be true, because the denominator in (4.70) is close to 1 by the law of large numbers, so the random variable T should be close to a standard Gaussian in distribution (see exercise below).

Exercise 4.4.4. If $X \sim \Gamma(\alpha, \beta)$, show that, for $t < \beta$,

$$\mathbb{E}e^{tX} = \Big(\frac{\beta}{\beta - t}\Big)^{\alpha}.$$

Exercise 4.4.5. If X and Y are independent random variables with the densities $p_X(x) = p_Y(x) = \mathrm{I}(0 \le x \le 1)$ (uniform on $[0,1]$), compute the density of their sum $X + Y$.

Exercise 4.4.6. If $X \sim \chi^2_n$ and $\mathbb{P}(X \le c_n) = p \in (0,1)$, prove that $\lim_{n\to\infty} \frac{c_n}{n} = 1$. *Hint:* Represent X as $g_1^2 + \ldots + g_n^2$ and use the law of large numbers.

Exercise 4.4.7. If $X \sim F_{k,m}$ and $Y \sim F_{m,k}$, show that, for any $c > 0$, $\mathbb{P}(X \le c) = \mathbb{P}(Y \ge 1/c)$.

Exercise 4.4.8. If $X \sim t_n$, show that $\mathbb{E}|X|^a < \infty$ for $a < n$ and and $\mathbb{E}|X|^a = \infty$ (undefined) for $a \ge n$.

Exercise 4.4.9. In the notation of (4.70), show that

$$\lim_{n\to\infty} \mathbb{P}(T \le t) = \mathbb{P}(g_0 \le t)$$

for all $t \in \mathbb{R}$. *Hint:* use the Law of Large Numbers for the denominator in (4.70).

4.5 Simple Linear Regression

Simple linear regression (SLR) is a classical statistical model that deals with the data of the form

$$(X_1,Y_1),\ldots,(X_n,Y_n), \tag{4.73}$$

where $X_i, Y_i \in \mathbb{R}$, and one expects or observes a roughly linear dependence of the variable Y on X. For example, here is a graph of carbon monoxide content (mg) vs. nicotine content (mg) in 25 brands of cigarettes.

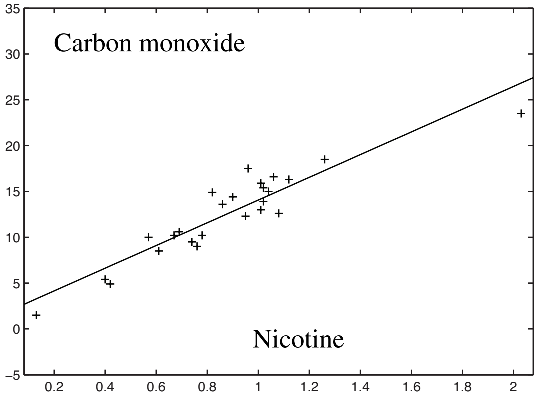


Fig. 4.2 Carbon monoxide content (mg) vs. nicotine content (mg). Solid line: least-squares line.

The SLR model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{4.74}$$

where the function

$$y = f(x) = \beta_0 + \beta_1 x \tag{4.75}$$

is called a *regression line* (whose parameters β_0 and β_1 are unknown) and where $\varepsilon_1,\ldots,\varepsilon_n$ represent *noise* modelled by i.i.d. Gaussian $N(0,\sigma^2)$ random variables with unknown variance σ^2. In this model, variable X is called the *predictor*, or *independent* variable, and Y is called the *response*, or *de-*

pendent variable. We will assume that not all X_i's are equal, because it does not make sense to use a linear function as a model in that case.

We have three unknown parameters, β_0, β_1, and σ^2, and we want to estimate them using the sample (4.73). Values $X_1, \dots, X_n$ can be either random or non random, but from the point of view of the SLR model they are treated as fixed and non random, and it is assumed that the randomness comes entirely from the noise variables ε_i. Another way to look at it is that each Y_i has Gaussian distribution $N(\beta_0 + \beta_1 X_i, \sigma^2)$ with the density

$$f_i\big(y; \beta_0, \beta_1, \sigma^2\big) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\beta_0-\beta_1 X_i)^2}$$

with three unknown parameters $\beta_0, \beta_1, \sigma^2$, and that $Y_1, \dots, Y_n$ are independent.

In Statistics, one of the most important methods of estimation of unknown parameters is the *Maximum Likelihood Estimation.* Consider the *likelihood function*

$$\ell(\beta_0, \beta_1, \sigma^2) := \prod_{i=1}^{n} f_i\big(Y_i; \beta_0, \beta_1, \sigma^2\big), \tag{4.76}$$

which is the joint density of the random vector $(Y_1, \dots, Y_n)$ evaluated at the observed values $Y_1, \dots, Y_n$. We maximize it over the unknown parameters,

$$\ell(\beta_0, \beta_1, \sigma^2) \to \underset{\beta_0, \beta_1, \sigma^2}{\text{maximize}}. \tag{4.77}$$

If the maximum is achieved on some $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$, these values are called the *maximum likelihood estimates (MLE)* of the unknown parameters. In our case,

$$\ell(\beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}.$$

First of all, no matter what σ is, we need to minimize

$$L := \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{4.78}$$

over β_0, β_1. The parameters that minimize L correspond to the famous *least-squares line*. Let us find critical points:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= -\sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i)) = 0, \\ \frac{\partial L}{\partial \beta_1} &= -\sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i))X_i = 0. \end{aligned}$$

If we introduce the notation

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,\ \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2,\ \overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

then the critical point conditions can be rewritten as

$$\begin{aligned} \beta_0 + \beta_1 \overline{X} &= \overline{Y}, \\ \beta_0 \overline{X} + \beta_1 \overline{X^2} &= \overline{XY}. \end{aligned}$$

Solving for β_0 and β_1 we obtain the MLE,

$$\hat{\beta}_1 := \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2},\ \hat{\beta}_0 := \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.79}$$

Since we assumed that not all X_i's are equal, the denominator

$$\sigma_x^2 := \overline{X^2} - \overline{X}^2 \neq 0. \tag{4.80}$$

It is called the *sample variance* of $X_1, \ldots, X_n$.

Then we can maximize $\ell(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)$ over σ^2. Ignoring the constant $(2\pi)^{n/2}$ and taking the logarithm, we would like to maximize

$$-n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Setting the derivative in σ to zero, we get the MLE

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2. \tag{4.81}$$

Interestingly, up to a factor $1/n$, the estimate $\hat{\sigma}^2$ turned out to be the minimum value of the function in (4.78) that we minimized in order to find $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\hat{\sigma}^2 = \frac{1}{n}\min_{\beta_0,\beta_1}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2. \tag{4.82}$$

This observation will play a crucial role below. In order to understand how these MLE estimates relate to the actual unknown parameters β_0, β_1 and σ^2, we will need to figure out their joint distribution.

Theorem 4.4. *The joint distribution of MLE* $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ *is described by the following properties.*

(a) The random vector $(\hat{\beta}_0 - \beta_0, \hat{\beta}_1 - \beta_1)$ *is Gaussian* $N(0,C)$*, where the covariance matrix*

$$C = \frac{\sigma^2}{n\sigma_x^2}\begin{bmatrix}\sigma_x^2 + \overline{X}^2 & -\overline{X} \\ -\overline{X} & 1\end{bmatrix}. \tag{4.83}$$

(b) $\hat{\sigma}^2$ *is independent of* $\hat{\beta}_0$ *and* $\hat{\beta}_1$.
(c) $\frac{n\hat{\sigma}^2}{\sigma^2}$ *has* χ^2_{n-2}*-distribution with* $n-2$ *degrees of freedom.*

Notice that, if the sample variance σ_x^2 stays bounded with n then the variances of $\hat{\beta}_0 - \beta_0$ and $\hat{\beta}_1 - \beta_1$ are of order $1/n$, by (4.83). Chebyshev's inequality tells us that the estimates $\hat{\beta}_j$ are close to unknown parameters β_j with probability close to one. However, we will do even better than Chebyshev's inequality by constructing precise confidence intervals for these unknown parameters below.

Proof. First of all, let us recenter the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and work with

$$\hat{\gamma}_0 = \hat{\beta}_0 - \beta_0,\ \hat{\gamma}_1 = \hat{\beta}_1 - \beta_1. \tag{4.84}$$

Let us also recall the SLR model equation (4.74), because it will be more convenient to work with the random variables ε_i instead of Y_i, since, by definition, we can think of them as

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T = \sigma g = \sigma(g_1, \ldots, g_n)^T,$$

where $g = (g_1, \ldots, g_n)^T \sim \gamma_n$ has the standard Gaussian distribution on $\mathbb{R}^n$. We leave it as a (simple) exercise to check that, with this recentering, the MLE equations (4.79) and (4.81) are equivalent to

$$\hat{\gamma}_1 = \frac{\overline{X\varepsilon} - \overline{X}\overline{\varepsilon}}{\sigma_x^2},\ \hat{\gamma}_0 = \bar{\varepsilon} - \hat{\gamma}_1 \overline{X} \tag{4.85}$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \hat{\gamma}_0 - \hat{\gamma}_1 X_i)^2. \tag{4.86}$$

Let us also record the analogue of (4.82),

$$\hat{\sigma}^2 = \frac{1}{n} \min_{\gamma_0, \gamma_1} \sum_{i=1}^{n} (\varepsilon_i - \gamma_0 - \gamma_1 X_i)^2. \tag{4.87}$$

The main idea will be to rewrite all the quantities above in terms of $Z = Ag$ for some orthogonal matrix A such that $\hat{\gamma}_0, \hat{\gamma}_1$ will be linear combinations of the first two coordinates Z_1, Z_2 of Z and $\hat{\sigma}^2$ will depend only on $Z_3, \ldots, Z_n$. Using the fact that an orthogonal transformation $Z = Ag$ of a standard Gaussian vector g on $\mathbb{R}^n$ is also standard Gaussian, all the claims will follow immediately from this representation in a new orthogonal basis.

Let us rewrite $\hat{\gamma}_0$ and $\hat{\gamma}_1$ as follows. Let us represent $\bar{\varepsilon}$ in terms of the scalar product

$$\bar{\varepsilon} = \frac{\varepsilon_1 + \ldots + \varepsilon_n}{n} = \frac{1}{\sqrt{n}}(\vec{a}_1, \varepsilon), \tag{4.88}$$

where the vector

$$\vec{a}_1 = (a_{11}, \ldots, a_{1n})^T = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)^T.$$

Similarly, let us represent

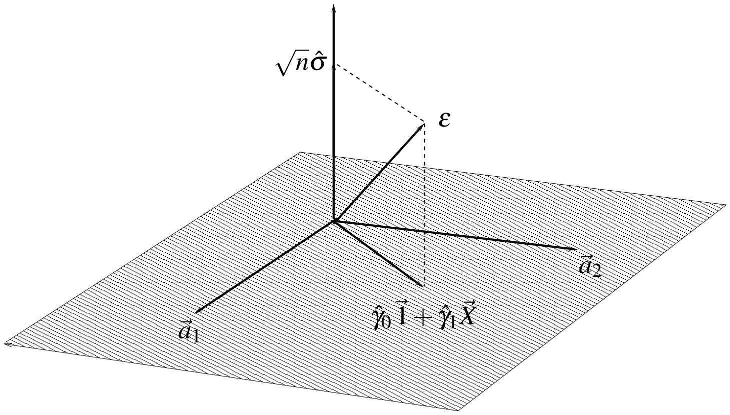


Fig. 4.3 A decomposition of ε into orthogonal components.

$$\hat{\gamma}_1 = \frac{\overline{X\varepsilon} - \overline{X}\bar{\varepsilon}}{\sigma_x^2} = \frac{1}{\sqrt{n}\sigma_x}(\vec{a}_2, \varepsilon), \tag{4.89}$$

where the vector

$$\vec{a}_2 = (a_{21}, \ldots, a_{2n})^T = \frac{1}{\sqrt{n}\sigma_x}\left(X_1 - \overline{X}, \ldots, X_n - \overline{X}\right)^T.$$

We can write

$$\hat{\gamma}_0 = \frac{1}{\sqrt{n}}(\vec{a}_1, \varepsilon) - \frac{\overline{X}}{\sqrt{n}\sigma_x}(\vec{a}_2, \varepsilon). \tag{4.90}$$

Notice that both vectors $\vec{a}_1$ and $\vec{a}_2$ have length 1, and they are orthogonal to each other,

$$|\vec{a}_1| = |\vec{a}_2| = 1,\ (\vec{a}_1, \vec{a}_2) = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{n\sigma_x} = 0.$$

Next, if we consider vectors

$$\vec{1} = (1, \dots, 1)^T \text{ and } \vec{X} = (X_1, \dots, X_n)^T,$$

then the equations (4.86) and (4.87) can be written as

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\left|\varepsilon - \hat{\gamma}_0\vec{1} - \hat{\gamma}_1\vec{X}\right|^2 \\ &= \frac{1}{n}\min_{\gamma_0,\gamma_1}\left|\varepsilon - \gamma_0\vec{1} - \gamma_1\vec{X}\right|^2.\end{aligned}$$

In particular, this implies that

$$\left(\hat{\gamma}_0, \hat{\gamma}_1\right) = \operatorname*{argmin}_{\gamma_0,\gamma_1}\left|\varepsilon - \gamma_0\vec{1} - \gamma_1\vec{X}\right|^2. \tag{4.91}$$

This equation states that $\hat{\gamma}_0\,\vec{1} + \hat{\gamma}_1\vec{X}$ is the closest point to ε in the subspace spanned by $\vec{1}$ and $\vec{X}$, so it must coincide with the orthogonal projection of ε onto this subspace. On the other hand, this subspace is also spanned by $\vec{a}_1$ and $\vec{a}_2$, $\mathrm{Span}\{\vec{1}, \vec{X}\} = \mathrm{Span}\{\vec{a}_1, \vec{a}_2\}$, because

$$\vec{a}_1 = \frac{1}{\sqrt{n}}\vec{1},\ \vec{a}_2 = \frac{1}{\sqrt{n}\sigma_x}\vec{X} - \frac{\overline{X}}{\sqrt{n}\sigma_x}\vec{1},$$

and

$$\vec{1} = \sqrt{n}\vec{a}_1,\ \vec{X} = \sqrt{n}\big(\overline{X}\,\vec{a}_1 + \sigma_x\vec{a}_2\big).$$

Since $\vec{a}_1, \vec{a}_2$ form an orthonormal basis of this subspace, the orthogonal projection of ε onto this subspace is

$$\hat{\gamma}_0\,\vec{1} + \hat{\gamma}_1\vec{X} = (\vec{a}_1, \varepsilon)\vec{a}_1 + (\vec{a}_2, \varepsilon)\vec{a}_2. \tag{4.92}$$

Another crucial consequence of this observation is that, by the Pythagorean theorem,

$$\begin{aligned}|\varepsilon|^2 &= |\varepsilon - \hat{\gamma}_0\vec{1} - \hat{\gamma}_1\vec{X}|^2 + (\vec{a}_1,\varepsilon)^2 + (\vec{a}_2,\varepsilon)^2\\ &= n\hat{\sigma}^2 + (\vec{a}_1,\varepsilon)^2 + (\vec{a}_2,\varepsilon)^2. \qquad (4.93)\end{aligned}$$

All the claims now follow from the fact that an orthogonal transformation of the standard Gaussian vector $(g_1,\dots,g_n)$ is standard Gaussian. Let us choose vectors $\vec{a}_3,\dots,\vec{a}_n$ so that $\vec{a}_1,\dots,\vec{a}_n$ is an orthonormal basis of $\mathbb{R}^n$ and the matrix A with rows $\vec{a}_1^T,\dots,\vec{a}_n^T$,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{bmatrix},$$

is orthogonal. By Lemma 4.4 in Section 4.2, we know that

$$Z = (Z_1,\dots,Z_n)^T = Ag \sim \gamma_n$$

and, since $\varepsilon = \sigma g$, we have $\sigma Z = A\varepsilon$. With this notation, the equations (4.89) and (4.90) become

$$\hat{\gamma}_0 = \frac{\sigma}{\sqrt{n}}Z_1 - \frac{\sigma\overline{X}}{\sqrt{n}\sigma_x}Z_2. \qquad (4.94)$$

$$\hat{\gamma}_1 = \frac{\sigma}{\sqrt{n}\sigma_x}Z_2, \qquad (4.95)$$

and, using that $\sigma^2|Z|^2 = |A\varepsilon|^2 = |\varepsilon|^2$, the equation (4.93) becomes

$$\sigma^2|Z|^2 = n\hat{\sigma}^2 + \sigma^2 Z_1^2 + \sigma^2 Z_2^2, \qquad (4.96)$$

which can be rewritten as

$$\frac{n\hat{\sigma}^2}{\sigma^2} = |Z|^2 - Z_1^2 - Z_2^2 = Z_3^2 + \ldots + Z_n^2. \qquad (4.97)$$

This representation proves claim (c), by the definition of χ^2_{n-2}-distribution. Claim (b) holds because the coordinates of Z are i.i.d., $\frac{n\hat{\sigma}^2}{\sigma^2}$ is a function of $Z_3,\ldots,Z_n$, while $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are functions of Z_1,Z_2. If we denote

$$B = \frac{\sigma}{\sqrt{n}\sigma_x}\begin{bmatrix}\sigma_x & -\bar{X}\\ 0 & 1\end{bmatrix},$$

the system of equations (4.94), (4.95) can be written as $(\hat{\gamma}_0,\hat{\gamma}_1)^T = B(Z_1,Z_2)^T$ and, therefore, $(\hat{\gamma}_0,\hat{\gamma}_1)^T \sim N(0,C)$ with the covariance

$$C = BB^T = \frac{\sigma^2}{n\sigma_x^2}\begin{bmatrix}\sigma_x^2+\bar{X}^2 & -\bar{X}\\ -\bar{X} & 1\end{bmatrix}.$$

This finishes the proof. □

Confidence intervals for parameters of SLR. Next, using Theorem 4.4, we will construct the *confidence intervals* for unknown parameters of the SLR model, β_0,β_1 and σ^2. For example, we showed that

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} = \Gamma\Big(\frac{n}{2}-1,\frac{1}{2}\Big),$$

and the density of χ^2_{n-2} is

$$f\Big(x;\frac{n}{2}-1,\frac{1}{2}\Big) = \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2}-1)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}.$$

For $n>0$, the density is 0 at $x=0$ and goes to 0 as $x\to\infty$, and it is easy to check that its unique maximum is at $x=n-2$. This means that the most unlikely outcomes x under this distribution (having smallest density) are near $x=0$ and $x=\infty$. To select more likely values in the middle, a common choice is to take a *confidence level* $\alpha<1$, say $\alpha=0.95$, and take constants c_1 and c_2 such that

$$\chi^2_{n-2}(0,c_1) = \frac{1-\alpha}{2} \text{ and } \chi^2_{n-2}(c_2,+\infty) = \frac{1-\alpha}{2}. \quad (4.98)$$

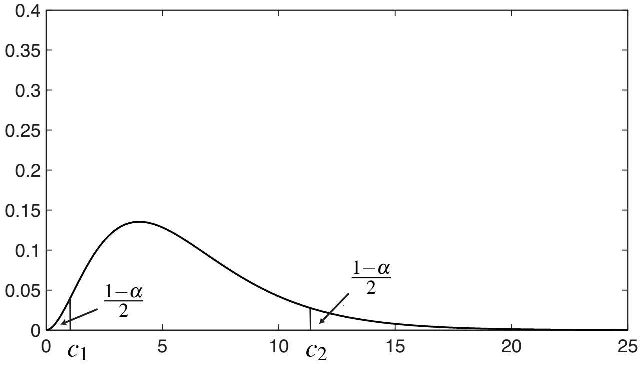


Fig. 4.4 Tails of chi-squared χ^2_{n-2}-distribution.

Then the interval $[c_1, c_2]$ has χ^2_{n-2} probability α,

$$\mathbb{P}\Big(c_1 \le \frac{n\hat{\sigma}^2}{\sigma^2} \le c_2\Big) = \alpha.$$

Solving this for σ^2, we find that the event

$$\frac{n\hat{\sigma}^2}{c_2} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{c_1} \tag{4.99}$$

occurs with probability α. This interval $[n\hat{\sigma}^2/c_2, n\hat{\sigma}^2/c_1]$ is called the α *confidence interval* for the parameter σ^2.

Similarly, we can find the α confidence intervals for β_1 and β_0. By Theorem 4.4,

$$(\hat{\beta}_1 - \beta_1)\Big(\frac{n\sigma_x^2}{\sigma^2}\Big)^{1/2} \sim N(0,1) \text{ and } \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2},$$

and the two random variables are independent. Therefore, by the definition of t-distribution, the ratio

$$(\hat{\beta}_1 - \beta_1)\Big(\frac{n\sigma_x^2}{\sigma^2}\Big)^{1/2} \Big/ \Big(\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}\Big)^{1/2}$$

has Student's t_{n-2}-distribution with $n-2$ degrees of freedom. Simplifying, we get

$$(\hat{\beta}_1 - \beta_1)\frac{\sqrt{n-2}\sigma_x}{\hat{\sigma}} \sim t_{n-2}.$$

Notice how the unknown parameter σ^2 cancelled out. The

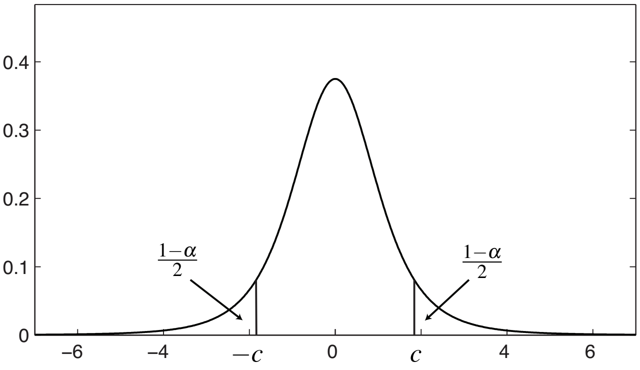


Fig. 4.5 Tails of t_{n-2}-distribution.

density of t-distribution in (4.71) is symmetric around $x = 0$, has maximum at $x = 0$, and is decreasing as $x \to \pm\infty$. This means that the most likely outcomes are around zero and, in order to construct α confidence interval, we select c such that $t_{n-2}(-c, c) = \alpha$. Then, with probability α,

$$-c \le (\hat{\beta}_1 - \beta_1)\frac{\sqrt{n-2}\sigma_x}{\hat{\sigma}} \le c, \tag{4.100}$$

and, solving for β_1, we obtain the α confidence interval

$$\left[\hat{\beta}_1 - c\frac{\hat{\sigma}}{\sqrt{n-2}\sigma_x}, \hat{\beta}_1 + c\frac{\hat{\sigma}}{\sqrt{n-2}\sigma_x}\right]. \tag{4.101}$$

Similarly, one can find the α confidence interval for β_0,

$$\left[\hat{\beta}_0 - c\sqrt{\frac{\hat{\sigma}^2}{n-2}\left(1+\frac{\overline{X}^2}{\sigma_x^2}\right)}, \hat{\beta}_0 + c\sqrt{\frac{\hat{\sigma}^2}{n-2}\left(1+\frac{\overline{X}^2}{\sigma_x^2}\right)}\right].$$

Simultaneous confidence set for the SLR regression line. Instead of constructing confidence intervals for β_0 and β_1 separately, we can also construct a simultaneous confidence set for (β_0, β_1) as follows. Let us recall the equations (4.94) and (4.95) in the proof of Theorem 4.4,

$$\hat{\gamma}_0 = \frac{\sigma}{\sqrt{n}} Z_1 - \frac{\sigma \overline{X}}{\sqrt{n}\sigma_x} Z_2, \; \hat{\gamma}_1 = \frac{\sigma}{\sqrt{n}\sigma_x} Z_2,$$

where $\hat{\gamma}_j = \hat{\beta}_j - \beta_j$. Solving them for Z_1, Z_2,

$$\begin{aligned} Z_1 &= \frac{\sqrt{n}}{\sigma}\left(\hat{\gamma}_0 + \overline{X}\hat{\gamma}_1\right) = \frac{\sqrt{n}}{\sigma}\left((\hat{\beta}_0 - \beta_0) + \overline{X}(\hat{\beta}_1 - \beta_1)\right), \\ Z_2 &= \frac{\sqrt{n}\sigma_x}{\sigma}\hat{\gamma}_1 = \frac{\sqrt{n}\sigma_x}{\sigma}(\hat{\beta}_1 - \beta_1). \end{aligned}$$

Random variables Z_1, Z_2 are independent standard Gaussian, also independent of

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Therefore, $Z_1^2 + Z_2^2$ has χ_2^2-distribution and, by the definition of F-distribution in the previous section,

$$\begin{aligned} F &:= \frac{Z_1^2 + Z_2^2}{2} \Big/ \frac{n\hat{\sigma}^2}{\sigma^2}\frac{1}{n-2} \\ &= \frac{n-2}{2\hat{\sigma}^2}\left[\sigma_x^2(\hat{\beta}_1 - \beta_1)^2 + \left((\hat{\beta}_0 - \beta_0) + \overline{X}(\hat{\beta}_1 - \beta_1)\right)^2\right] \end{aligned}$$

has $F_{2,n-2}$ distribution with degrees of freedom 2 and $n-2$. Given $\alpha \in (0,1)$, if we define c by $F_{2,n-2}(0,c) = \alpha$ then, with probability α,

$$\frac{n-2}{2\hat{\sigma}^2}\left[\sigma_x^2(\hat{\beta}_1 - \beta_1)^2 + \left((\hat{\beta}_0 - \beta_0) + \overline{X}(\hat{\beta}_1 - \beta_1)\right)^2\right] \leq c.$$

This inequality defines an ellipse centred at $(\hat{\beta}_0, \hat{\beta}_1)$, which is a simultaneous α confidence set for the pair (β_0, β_1).

Exercise 4.5.1. Check that (4.79) and (4.79) are equivalent to (4.85) and (4.85).

Exercise 4.5.2. Show that the constants c_1 and c_2 in (4.98) (that depend on n) satisfy $\lim_{n\to\infty} \frac{c_j}{n} = 1$. *Hint:* use (4.65) and Chebyshev's inequality.

Exercise 4.5.3. Check directly from the definitions that the equation (4.92) holds.

Exercise 4.5.4. How does c in (4.100) behave as $n \to \infty$? *Hint:* use (4.72) or Exercise 4.4.9.

Chapter 5
Finite State Markov Chains

5.1 Definitions and Basic Properties

In the Example 1.4.5 in Section 1.4, we discussed how the joint distribution of discrete random variables $X_1, \ldots, X_n$ can be constructed sequentially

$$\begin{aligned}&\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n\big)\\&= \prod_{k=0}^{n-1} \mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_1 = x_1, \ldots, X_k = x_k\big),\end{aligned} \tag{5.1}$$

where the first factor for $k = 0$ is just $\mathbb{P}(X_1 = x_1)$ and

$$\mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_1 = x_1, \ldots, X_k = x_k\big)$$

is the conditional probability that $X_{k+1} = x_{k+1}$ given $X_1 = x_1$, $\ldots$, $X_k = x_k$. A sequence of random variables $X_1, \ldots, X_n, \ldots$ is called *a Markov chain* if this conditional probability depends only on the last outcome $X_k = x_k$ in the condition,

$$\begin{aligned}&\mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_1 = x_1, \ldots, X_k = x_k\big)\\&= \mathbb{P}\big(X_{k+1} = x_{k+1} \mid X_k = x_k\big).\end{aligned} \tag{5.2}$$

This is called a *Markov property*, or *memoryless property*.

In this chapter, we will discuss only *finite state* Markov chains when all X_k's take values in a finite set

$$S = \{s_1, \dots, s_m\}. \tag{5.3}$$

Elements of this set are called *states*.

The conditional probabilities in (1.93) are called *transition probabilities*, and Markov chain is called *homogeneous* if the transition probabilities $\mathbb{P}(X_{k+1} = b \mid X_k = a)$ for $a, b \in S$ are the same for all k. A Markov chain is almost by default assumed to be homogeneous, and it is stated explicitly when it is not. Here we will discuss only homogeneous Markov chains.

The *transition matrix* P is the matrix

$$P := \left[p_{ij}\right]_{1 \le i,j \le n} = \left[p(s_i, s_j)\right]_{1 \le i,j \le n}, \tag{5.4}$$

whose entries are the transition probabilities

$$p_{ij} = p(s_i, s_j) = \mathbb{P}\big(X_{k+1} = s_j \mid X_k = s_i\big). \tag{5.5}$$

It is customary (and convenient in terms of notation) to start the Markov chain at 'time zero', i.e. $k = 0$. The distribution of X_0, which we will denote by μ,

$$\mu_i = \mu(s_i) := \mathbb{P}(X_0 = s_i) \text{ for } s_i \in S, \tag{5.6}$$

is called the *initial distribution* of the Markov chain. With this notation, the definition of the (homogeneous finite state) Markov chain can be rewritten as

$$\begin{aligned} &\mathbb{P}\big(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n\big) \\ &\quad = \mu(x_0) p(x_0, x_1) p(x_1, x_2) \cdots p(x_{n-1}, x_n). \end{aligned} \tag{5.7}$$

The initial distribution μ will vary depending on how we want to start the chain, so we can also express this definition entirely in terms of the transition matrix,

$$\begin{aligned}&\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n \mid X_0 = x_0\big)\\ &= p(x_0,x_1)p(x_1,x_2)\cdots p(x_{n-1},x_n).\end{aligned} \tag{5.8}$$

If we start the chain in the state x_0, then the probability that the chain will visit states $x_1, \ldots, x_n$ in the next n steps is the product of transition probabilities along the path. If we multiply by $\mu(x_0)$, we recover the previous equation.

It is also convenient to visualize transition probabilities by a *transition graph*. For example, the transition matrix

$$P = \begin{bmatrix} 0.25 & 0.25 & 0 & 0 & 0.5 & 0 & 0 \\ 0.4 & 0 & 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0.3 & 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0.75 & 0 & 0 \\ 0 & 0.9 & 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \end{bmatrix} \tag{5.9}$$

can be depicted as in Figure 5.1, with dots representing states, and arrows representing non-zero transition probabilities. Notice that, for example, $p_{12} \neq p_{21}$.

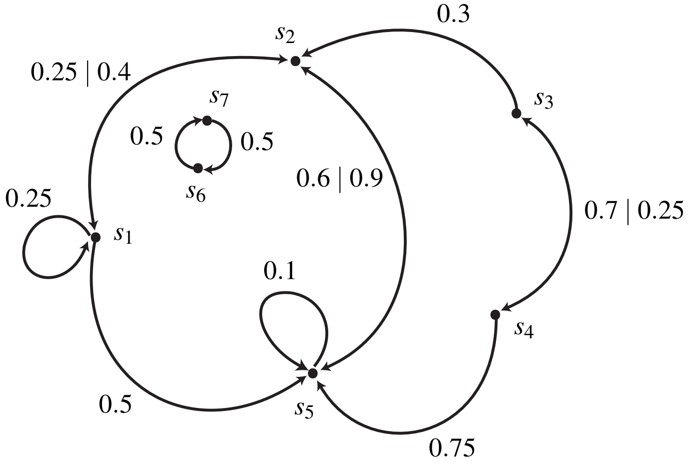


Fig. 5.1 Transition graph of a Markov chain with the transition matrix (5.9). States s_3, s_4 are inessential, and essential states are divided into two clusters, $\{s_1,s_2,s_5\}$ and $\{s_6,s_7\}$.

This graphical representation of transition probabilities suggests that we can think of Markov chains as random walks on the set of states, where at each step we pick one of the neighbours of the current state s_i with probabilities p_{ij} and move there. If $p_{ii} \neq 0$, we can end up staying at s_i.

A state s_i is called *inessential* if we can reach from it another state s_j with positive probability (not necessarily in one step), from which we cannot come back to s_i. For example, in Figure 5.1, the states s_3 and s_4 are inessential. Such states are also called *transient*, although this terminology is better suited for infinite state Markov chains. For finite state chains, once we reach s_j, we can never come back to s_i, so for the long-term behaviour of the Markov chain these states do not matter.

Two states s_i and s_j are called *communicating*, denoted $s_i \leftrightarrow s_j$, if we can go from each of them to the other with positive probability. If

$$S_i = \left\{ s \in S : s_i \leftrightarrow s \right\}$$

is the set of states communicating with s_i, and if s_i communicates with s_j then, obviously, $S_i = S_j$, because any state that communicates with s_i communicates with s_j and vice versa. This means that communicating states can be divided into disjoint *clusters*. For example, in Figure 5.1, there are two clusters of essential states, $\{s_1, s_2, s_5\}$ and $\{s_6, s_7\}$, and one cluster of inessential states, $\{s_3, s_4\}$. A state s_i is called *absorbing state* if $p_{ii} = 1$, because, once we reach this state, we can never leave. Transition probabilities on a given cluster of essential states define a Markov chain on that cluster.

Markov chain is called *irreducible* if all its states communicate with each other, which means that there are no inessential states and there is only one cluster of essential states. Below we will mostly study irreducible Markov chains.

The transition matrix P is much more useful than just as a representation of the transition probabilities. If we denote

$$p_{ij}(n) = \big(P^n\big)_{ij} \tag{5.10}$$

the (i, j)th element of P^n then the following holds.

Lemma 5.1 (n-step transition probabilities). *For all $n \geq 1$,*

$$\mathbb{P}(X_n = s_j \mid X_0 = s_i) = p_{ij}(n) \tag{5.11}$$

and, in particular,

$$\mathbb{P}(X_n = s_j) = \sum_{i=1}^{n} \mu_i p_{ij}(n), \tag{5.12}$$

where μ is the initial distribution defined in (5.6).

If we write the last equation in the vector form,

$$\big(\mathbb{P}(X_n = s_1), \ldots, \mathbb{P}(X_n = s_m)\big) = \mu P^n, \tag{5.13}$$

where $\mu = (\mu_1, \ldots, \mu_m)$, we can compute the distribution of the chain at time n by multiplying the initial distribution μ by P^n. It is important that we multiply the row vector μ by P^n on the right and, for example, $P^n \mu^T$ does not have the same meaning. The equation (5.11) also shows that

$$\sum_{j=1}^{m} p_{ij}(n) = 1,$$

so P^n can be viewed as an n-step transition matrix.

Proof. We start with $n = 2$. In order to compute

$$\mathbb{P}(X_2 = s_j \mid X_0 = s_i),$$

we can sum over possible outcomes of $X_1 = s_k$,

$$\begin{aligned}\mathbb{P}(X_2 = s_j \mid X_0 = s_i) &= \sum_{k=1}^{m} \mathbb{P}(X_1 = s_k, X_2 = s_j \mid X_0 = s_i)\\ &= \sum_{k=1}^{m} p(s_i, s_k)p(s_k, s_j) = \sum_{k=1}^{m} p_{ik}p_{kj},\end{aligned}$$

which is the (i, j)th entry of P^2. The general case follows by induction, using a similar calculation. Assuming that (5.11) holds,

$$\begin{aligned}&\mathbb{P}(X_{n+1} = s_j \mid X_0 = s_i)\\ &= \sum_{k=1}^{m} \mathbb{P}(X_n = s_k, X_{n+1} = s_j \mid X_0 = s_i)\\ &= \sum_{k=1}^{m} \mathbb{P}(X_{n+1} = s_j \mid X_n = s_k)\mathbb{P}(X_n = s_k \mid X_0 = s_i)\\ &= \sum_{k=1}^{m} p_{ik}(n)p_{kj} = p_{ij}(n+1).\end{aligned}$$

If we multiply this by $\mu_i = \mathbb{P}(X_0 = s_i)$ and sum over $i \leq m$, we get (5.12) for X_{n+1}, so the proof of the induction step is complete. □

The *period d_i of a state s_i* is defined by

$$d_i = \gcd\{n \geq 1 : p_{ii}(n) > 0\}, \tag{5.14}$$

the greatest common divisor all the times n when a walk starting at s_i can return to s_i. If $d_i = 1$ then the state is called *aperiodic*. We call a Markov chain *aperiodic* if all periods $d_i = 1$. It turns out that for irreducible chains this is the same as requiring only one period d_i to be equal to 1.

Lemma 5.2. *If a Markov chain is irreducible then all periods d_i are equal.*

Proof. Let us suppose that s_j can be reached from s_i in N steps and s_i can be reached from s_j in M steps with positive probabilities, $p_{ij}(N) > 0$ and $p_{ji}(M) > 0$. If $p_{jj}(n) > 0$ then

$$p_{ii}(N+M+n) \geq p_{ij}(N)p_{jj}(n)p_{ji}(M) > 0,$$

because we can reach s_j in N steps, then come back to s_j in n steps, and then reach s_i in M steps, so the probability to return to s_i in $N+M+n$ steps is positive. This is also true for $n=0$, because we can go to s_j and come back to s_i in $N+M$ steps. Since d_i is the period of s_i, by definition, d_i divides all numbers $N+M+n$ as above. In particular, it divides $N+M$, which implies that it divides all n such that $p_{jj}(n) > 0$. This means that d_i divides d_j, because d_j is the greatest common divisor of all such n. Similarly, d_j divides d_i, so they must be equal. □

The following property of irreducible aperiodic Markov chains will be very important in Section 5.3.

Lemma 5.3. *If a Markov chain is irreducible and aperiodic then there exists* $N \geq 1$ *such that, for all* $n \geq N$,

$$p_{ij}(n) > 0 \textit{ for all } i, j. \tag{5.15}$$

In other words, for large n, all the entries of P^n are strictly positive.

Proof. Let $T(s_1) = \{n \geq 1 : p_{11}(n) > 0\}$ be the set of all times the chain starting at s_1 can come back to s_1. Let $d \geq 1$ be the smallest positive integer that can be written as

$$d = a_1 n_1 + \ldots + a_k n_k,$$

for $k \geq 1$, $n_i \in T(s_1)$ and $a_i \in \mathbb{Z}$. Then d must divide all $n \in T(s_1)$, because if it does not divide some $n \in T(s_1)$ then

$$n = ad + r,$$

where the remainder $r \geq 1$ is strictly less than d. However,

$$r = n - ad = n - a_1 n_1 - \ldots - a_k n_k$$

is also a linear combination with integer coefficients of the times in $T(s_1)$, which contradicts that d was the smallest such number. This proves that d divides the period of s_1 and, since the chain is aperiodic, $d = 1$. We showed that

$$1 = a_1 n_1 + \ldots + a_k n_k, \tag{5.16}$$

for some $k \geq 1$, $n_i \in T(s_1)$ and $a_i \in \mathbb{Z}$. Some a_i of course can be negative.

Let us take the largest $|a_i|$ and, for certainty, suppose that $|a_1|$ is the largest. If

$$N_1 = |a_1| n_1 (n_1 + \ldots + n_k),$$

we will now show that all $n \geq N_1$ can be written as a linear combinations

$$n = c_1 n_1 + \ldots + c_k n_k,$$

with non-negative integer coefficients c_i. This will prove that $p_{11}(n) > 0$ for such n, because the chain starting from the state s_1 can come back to s_1 in n_i number of steps, so we just need to repeat these loops c_i times for all $i \leq k$ to see that the chain can come back in n steps with positive probability.

Let us first consider any n between N_1 and $N_1 + n_1$, which means that $n = N_1 + \ell$ for some $\ell \leq n_1$. By (5.16), we can write

$$\begin{aligned} n = N_1 + \ell = {} & |a_1| n_1 (n_1 + \ldots + n_k) \\ & + \ell (a_1 n_1 + \ldots + a_k n_k) \\ = {} & c_1 n_1 + \ldots + c_k n_k, \end{aligned}$$

where

$$c_i = |a_1| n_1 + \ell a_i \geq |a_1| \ell + \ell a_i = (|a_1| + a_i)\ell \geq 0$$

as we wished, because we assumed that $|a_1|$ is greater or equal to $|a_i|$. Since

$$N_1+n_1=(|a_1|n_1+1)n_1+|a_1|n_1(n_2+\ldots+n_k),$$

we can repeat the same argument for n in between N_1+n_1 and N_1+2n_1, and so on.

Similarly, for each state s_i, we can find some N_i such that $p_{ii}(n)>0$ for $n\geq N_i$. If we take $N'=\max(N_1,\ldots,N_m)$ then $p_{ii}(n)>0$ for all i and all $n\geq N'$. Since the chain is irreducible, we can always go from one state s_i to another state s_j in under m steps with positive probability (see exercise below), which implies that $p_{ij}(n)>0$ for all $n\geq N'+m$. □

Exercise 5.1.1. Let $U_1,U_2,\ldots$ be i.i.d. random variables with the uniform distribution on $\{1,\ldots,m\}$ and let

$$X_n=\max_{i\leq n}U_i.$$

Show that $X_1,X_2,\ldots$ is a Markov chain and find its transition probabilities. What are the essential states of this chain?

Exercise 5.1.2. Suppose that m while balls and m black balls are mixed together and divided evenly between two baskets. At each step two balls are chosen at random, one from each basket, and switched. We say that the system is in the state s_i if there are i white balls in the first basket. Find the transition probabilities of this chain.

Exercise 5.1.3. If a Markov chain on m states is irreducible, show that, for any $i,j\leq m$, $p_{ij}(k)>0$ for some $k\leq m$.

Exercise 5.1.4. If a Markov chain is irreducible and $p_{11}>0$, what is the period d_2 of the state s_2?

Exercise 5.1.5. If a Markov chain $(X_n)_{n\geq 0}$ is irreducible and has period d, show that $Y_n=X_{dn}$ for $n\geq 0$ is a Markov chain and that it is aperiodic. Does it have to be irreducible?

Exercise 5.1.6. Let $N\sim\mathrm{Poiss}(\lambda)$ be a Poisson random variable. Let us run N independent Markov chains starting from the state s_1, and let $N_n(i)$ be the number of these chains in the state s_i at time n. Show that $N_n(i)\sim\mathrm{Poiss}(\lambda p_{1i}(n))$.

5.2 Stationary Distributions

In Lemma 5.1 we showed that if the vector $\mu = (\mu_1, \ldots, \mu_m)$ describes the distribution of the chain X_0 at time zero, then μP^n is the distribution of X_n at time n. The distribution μ on the set of states is called *stationary* if

$$\mu = \mu P, \tag{5.17}$$

i.e. the distribution of X_1 is the same as that of X_0. Of course, this implies that $\mu = \mu P^n$, so the distribution of all X_n is the same. We will show that a stationary distribution always exists, and for an irreducible Markov chain it is unique. For irreducible chains, we will also derive a representation of the stationary distribution in terms of expected return times. To find a stationary distribution in practice, one can just solve a system of linear equations $\mu = \mu P$.

Lemma 5.4 (Existence). *For any finite state Markov chain, there exists at least one stationary distribution.*

Proof. Let us consider a sequence of matrices

$$A_n = \frac{1}{n}(1 + P + \ldots + P^n).$$

Since each matrix P^k is a transition matrix and its rows sum up to 1, the rows of A_n also sum up to one, so its entries are all between 0 and 1. By the Bolzano–Weierstrass theorem, there exists a convergent subsequence $A_{n_k} \to A$. Notice that the limit A is also a transition matrix because the entries must be nonnegative and rows add up to 1. If we consider

$$A_n P = \frac{1}{n}(P + \ldots + P^n + P^{n+1}),$$

we see that the difference goes to a zero matrix,

$$A_n P - A_n = \frac{1}{n}(-1 + P^{n+1}) \to 0.$$

This implies that

$$A = \lim_{k \to \infty} A_{n_k} = \lim_{k \to \infty} A_{n_k} P = AP,$$

which shows that $A = AP$. This means that each row a of this matrix satisfies $a = aP$. Since each row is a probability distribution on states (entries are nonnegative and add up to one), each row is a stationary distribution. □

Notice that if the stationary distribution is unique, all the rows of A in the above proof must be equal.

Next, we will show the following.

Lemma 5.5 (Uniqueness). *If a Markov chain is irreducible then the stationary distribution is unique.*

Proof. Previous lemma shows that $P - I$ is not invertible, since there exists a non-zero solution of $\mu(P - I) = 0$. Let us first show that, when P is the transition matrix of an irreducible Markov chain, the rank of $P - I$ is $m - 1$. Consider the linear system $(P - I)v^T = 0$ for $v = (v_1, \ldots, v_n)$. Let us show that irreducibility implies that

$$v_1 = v_2 = \ldots = v_n.$$

Consider the largest coordinate of v, let us say, v_1. The first equation in the system reads

$$\sum_{j=1}^{m} p_{1j} v_j = v_1.$$

On the other hand, since all $v_j \leq v_1$ and $p_{1j} v_j \leq p_{1j} v_1$,

$$v_1 = \sum_{j=1}^{m} p_{1j} v_j \leq \sum_{j=1}^{m} p_{1j} v_1 = v_1.$$

So the equality is actually an equality, and all $p_{1j} v_j = p_{1j} v_1$. This means that $v_j = v_1$ for all j such that $p_{1j} \neq 0$. In other words, all the states s_j that can be reached from the state

s_1 in one step must have $v_j = v_1$. Then, repeating the same calculation for these immediate neighbours we can see that all their neighbours also much have $v_j = v_1$. Since the chain is irreducible, we can reach all states in finitely many steps, which proves that all v_j's are equal.

This proves that the kernel of $P - I$ is one dimensional,

$$\ker(P-I) = \big\{c(1,\ldots,1) : c \in \mathbb{R}\big\}, \tag{5.18}$$

spanned by the vector $(1,\ldots,1)$, which means that the rank of $P - I$ equals to $m-1$. On the other hand, if we had more than one stationary distribution, the system $\mu = \mu P$ would have at least two linearly independent solutions, which would mean that the rank of $P - I$ would be not greater than $m-2$. This proves that the stationary distribution is unique. □

Next, we will derive a more descriptive formula for the stationary distribution of irreducible Markov chains. First, let us give a heuristic non-rigorous explanation and then check carefully that our intuitive guess is correct. In the proof of Lemma 5.4 we considered the sequence of matrices

$$A_n = \frac{1}{n}(1 + P + \ldots + P^n)$$

and showed that any limit A over some subsequence consists of rows which are stationary distributions. For irreducible chains, we just showed that the stationary distribution μ is unique, so we must have that

$$A = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \begin{bmatrix} \mu_1 & \cdots & \mu_m \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_m \end{bmatrix}.$$

In particular, this means that the limits over all convergent subsequences are equal to A and, therefore, the limit exists over the entire sequence,

$$\lim_{n\to\infty}\frac{1}{n}(1+P+\ldots+P^n)=\begin{bmatrix}\mu_1 & \cdots & \mu_m\\ \vdots & \ddots & \vdots\\ \mu_1 & \cdots & \mu_m\end{bmatrix}.$$

Since the (i,j)th entry of P^k is

$$(P^k)_{ij}=p_{ij}(k)=\mathbb{P}(X_k=s_j\mid X_0=s_i),$$

this can be written as

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\mathbb{P}(X_k=s_j\mid X_0=s_i)=\mu_j. \tag{5.19}$$

If we multiply both sides by $\mathbb{P}(X_0=s_i)$, sum over $i\le m$ and use that

$$\sum_{i=1}^{m}\mathbb{P}(X_k=s_j\mid X_0=s_i)\mathbb{P}(X_0=s_i)=\mathbb{P}(X_k=s_j),$$

we get

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\mathbb{P}(X_k=s_j)=\mu_j. \tag{5.20}$$

In other words, this equation holds no matter how we start the chain, i.e. no matter what the distribution of X_0 is. Since

$$\mathbb{P}(X_k=s_j)=\mathbb{E}\,\mathrm{I}(X_k=s_j),$$

by the linearity of expectation,

$$\lim_{n\to\infty}\mathbb{E}\frac{1}{n}\sum_{k=1}^{n}\mathrm{I}(X_k=s_j)=\mu_j. \tag{5.21}$$

Clearly, this equation can be interpreted by saying that μ_j is the *expected proportion of time the chain visits the state s_j*, if we looks at these proportions over long periods of time.

There is another interpretation of the last equation. Since it does not matter how we start the chain, let us start it in the state s_1. Then, let us divide the time between 1 and n into intervals between consecutive visits of the state s_1. Let us say that M visits of s_1 occurred between 1 and n, with intervals $T_1, T_2, \ldots, T_M$, so that

$$T_1 + \ldots + T_M \approx n$$

in the sense that their ratio is close to 1. Let $N_1(j), \ldots, N_M(j)$ be the number of times we visited s_j in between these consecutive returns to s_1, so that

$$N_1(j) + \ldots + N_M(j) \approx \sum_{k=1}^{n} \mathrm{I}(X_k = s_j).$$

Then

$$\frac{1}{n} \sum_{k=1}^{n} \mathrm{I}(X_k = s_j) \approx \frac{N_1(j) + \ldots + N_M(j)}{T_1 + \ldots + T_M}.$$

If we know that the Markov chain is visiting s_1, the future does not depend on the past because of the memoryless property. This suggests that what happens in between consecutive visits is independent of each other and the pairs

$$(N_1(j), T_1), \ldots, (N_M(j), T_M)$$

are actually independent and identically distributed. To make this statement rigorous, once needs to prove the so called *strong Markov property*, which we are not going to go into here. However, if we use it as a guiding intuition and divide both numerator and denominator above by M, the law of large numbers tells us that

$$\frac{N_1(j) + \ldots + N_M(j)}{T_1 + \ldots + T_M} \approx \frac{\mathbb{E}N_1(j)}{\mathbb{E}T_1}.$$

This suggests another representation,

$$\mu_j = \frac{\mathbb{E}N_1(j)}{\mathbb{E}T_1}, \tag{5.22}$$

where we assume that the chain starts in the state s_1,

$$T_1 = \min\{n \geq 1 \,:\, X_n = s_1\} \tag{5.23}$$

is the *first return time* to s_1, and

$$N_1(j) = \sum_{k=0}^{T_1-1} \mathrm{I}(X_k = s_j) \tag{5.24}$$

is the number of times the chain visits s_j before returning to s_1.

The first return time T_1 is an example of a *stopping time*. In general, an integer valued random variable $T \geq 0$ is called a stopping time if the event $\{T = n\}$ depends only on $X_0, \ldots,$ X_n. In other words, we decide whether to stop at time n based on what was observed up to time n. The return time T_1 to s_1 is a stopping time in this sense, because

$$\{T_1 = n\} = \{X_1 \neq s_1, \ldots, X_{n-1} \neq s_1, X_n = s_1\}.$$

Notice also that $N_1(1) = 1$, because the chain is in the state s_1 only at time zero before the return, so (5.22) implies that

$$\mu_1 = \frac{1}{\mathbb{E}T_1}. \tag{5.25}$$

Since it did not matter how we start the chain, if we start it at s_j and define

$$T_j = \min\{n \geq 1 \,:\, X_n = s_j\} \tag{5.26}$$

then the same logic suggests that

$$\mu_j = \frac{1}{\mathbb{E}T_j}. \tag{5.27}$$

In order to prove the representation (5.22), we need to check two things. First, we will need to check that $\mathbb{E}T_1 < \infty$. Because the number of visits to different states before time

T_1 adds up to T_1,

$$\sum_{j=1}^{m} N_1(j) = T_1,$$

and therefore $N_1(j) \le T_1$, proving $\mathbb{E}T_1 < \infty$ would also imply that $\mathbb{E}N_1(j) < \infty$. This would show that the numbers in (5.22) are well defined, and they define a distribution because

$$\sum_{j=1}^{m} \mu_j = \sum_{j=1}^{m} \frac{\mathbb{E}N_1(j)}{\mathbb{E}T_1} = \frac{\mathbb{E}T_1}{\mathbb{E}T_1} = 1.$$

After that we will need to check that $\mu P = \mu$, so it is the stationary distribution.

Lemma 5.6. *If a Markov chain is irreducible then $\mathbb{E}T_1 < \infty$, where T_1 defined in (5.23) is the first return time to the state s_1 for the chain that starts at s_1.*

In the proof we will use the Markov property (5.2) in the following way. Let us consider two vectors

$$X^1 = (X_1, \ldots, X_{t_1}),\ X^2 = (X_{t_1+1}, \ldots, X_{t_2})$$

consisting of the Markov chain up to time t_1 and in between t_1 and t_2. Let us consider some subsets

$$A \subseteq \{s_1, \ldots, s_m\}^{t_1},\ B \subseteq \{s_1, \ldots, s_m\}^{t_2 - t_1}$$

of possible outcomes for these two blocks of the Markov chain, and consider the conditional probability

$$\mathbb{P}(X^2 \in B \mid X^1 \in A).$$

We can not use the Markov property directly, because the condition is a set, but we can rewrite this as

$$\begin{aligned}\mathbb{P}(X^2 \in B \mid X^1 \in A) &= \frac{\mathbb{P}(X^2 \in B, X^1 \in A)}{\mathbb{P}(X^1 \in A)} \qquad (5.28)\\ &= \sum_{a \in A} \frac{\mathbb{P}(X^2 \in B, X^1 = a)}{\mathbb{P}(X^1 \in A)}\\ &= \sum_{a \in A} \frac{\mathbb{P}(X^2 \in B \mid X^1 = a)\mathbb{P}(X^1 = a)}{\mathbb{P}(X^1 \in A)}\\ &= \sum_{a \in A} \frac{\mathbb{P}(X^2 \in B \mid X_{t_1} = a_{t_1})\mathbb{P}(X^1 = a)}{\mathbb{P}(X^1 \in A)},\end{aligned}$$

where in the last step we used the Markov property to write the condition in terms of the last outcome $X_{t_1} = a_{t_1}$. This is very useful because, for example, if we can control the probability

$$\mathbb{P}(X^2 \in B \mid X_{t_1} = a_{t_1}) \leq p$$

for all possible outcomes $X_{t_1} = a_{t_1}$ then the above equation implies that

$$\mathbb{P}(X^2 \in B \mid X^1 \in A) \leq p \sum_{a \in A} \frac{\mathbb{P}(X^1 = a)}{\mathbb{P}(X^1 \in A)} = p. \qquad (5.29)$$

In other words, Markov property can be used in this way even if the condition is a set.

Proof (of Lemma 5.6). Since the chain is irreducible, if the chain is in the state s_i at time k, $X_k = s_i$, we can reach the state s_1 in under m steps, which means that

$$p_{i1}(\ell) = \mathbb{P}(X_{k+\ell} = s_1 \mid X_k = s_i) > 0$$

for some $\ell \leq m$. This implies that the probability that we do not reach the state s_1 in the next m steps is strictly smaller than 1,

$$\begin{aligned}&\mathbb{P}(X_{k+1} \neq s_1, \dots, X_{k+m} \neq s_1 \mid X_k = s_i)\\ &\leq \mathbb{P}(X_{k+\ell} \neq s_1 \mid X_k = s_i)\\ &= 1 - \mathbb{P}(X_{k+\ell} = s_1 \mid X_k = s_i) < 1.\end{aligned}$$

This is true for any s_i, so

$$\mathbb{P}(X_{k+1} \neq s_1, \dots, X_{k+m} \neq s_1 \mid X_k = s_i) \leq 1 - \varepsilon, \qquad (5.30)$$

for some small enough $\varepsilon > 0$, for all $i \leq m$.

Using Markov property, this will imply that the probability that the chain does not come back to s_1 in Nm steps (i.e. N cycles of m steps) will be smaller than $(1-\varepsilon)^N$,

$$\mathbb{P}(T_1 > Nm) \leq (1-\varepsilon)^N. \qquad (5.31)$$

To see this, let us write

$$\mathbb{P}(T_1 > Nm) = \mathbb{P}(X_1 \neq s_1, \dots, X_{(N-1)m} \neq s_1, \dots, X_{Nm} \neq s_1)$$

and let us decompose this sequence into two blocks

$$X^1 = (X_1, \dots, X_{(N-1)m}),\ X^2 = (X_{(N-1)m+1}, \dots, X_{Nm}).$$

If we take

$$A = \{s_2, \dots, s_m\}^{(N-1)m},\ B = \{s_2, \dots, s_m\}^m$$

to be the sets consisting of all the paths that avoid the state s_1 and have lengths $(N-1)m$ and m then

$$\begin{aligned}\mathbb{P}(T_1 > Nm) &= \mathbb{P}(X^2 \in B, X^1 \in A)\\ &= \mathbb{P}(X^2 \in B \mid X^1 \in A)\mathbb{P}(X^1 \in A).\end{aligned}$$

The equation (5.30) implies that, no matter what the value of $X_{(N-1)m}$ is, the conditional probability that all coordinates of X^2 avoid the state s_1 is smaller than $1-\varepsilon$. By the discussion before the proof and (5.29), the Markov property implies

$$\mathbb{P}(X^2 \in B \mid X^1 \in A) \le 1 - \varepsilon,$$

and, therefore,

$$\begin{aligned}\mathbb{P}(T_1 > Nm) &\le (1-\varepsilon)\mathbb{P}(X^1 \in A)\\ &= (1-\varepsilon)\mathbb{P}(T_1 > (N-1)m).\end{aligned}$$

By induction on N, the equation (5.31) follows.

Next, let us denote $\delta = (1-\varepsilon)^{1/m} < 1$ and consider any $k \ge 1$. Take $N \ge 0$ such that $Nm < k \le (N+1)m$. Then

$$\mathbb{P}(T_1 \ge k) \le \mathbb{P}(T_1 > Nm) \le (1-\varepsilon)^N = \delta^{Nm} \le \delta^{k-m}.$$

By the formula (1.34) in Section 1.2 and (5.31),

$$\mathbb{E}T_1 = \sum_{k=1}^{\infty} \mathbb{P}(T_1 \ge k) \le \sum_{k=1}^{\infty} \delta^{k-m} = \frac{\delta^{1-m}}{1-\delta} < \infty,$$

which finishes the proof. □

Recall the equations 5.22), (5.23), (5.24),

$$\mu_j = \frac{\mathbb{E}N_1(j)}{\mathbb{E}T_1}, \tag{5.32}$$

$$T_1 = \min\Big\{n \ge 1 : X_n = s_1\Big\}, \tag{5.33}$$

$$N_1(j) = \sum_{k=0}^{T_1-1} \mathrm{I}(X_k = s_j), \tag{5.34}$$

where we assumed that the chain starts in the state s_1. To check that our heuristic discussion above gives the correct answer, it remains to prove the following.

Theorem 5.1 (Representation of stationary distribution). *If a Markov chain is irreducible then μ defined in (5.32) is the stationary distribution.*

Proof. Showing that $\mu = \mu P$ is equivalent to showing

$$\mathbb{E}N_1(j) = \sum_{i=1}^{m} p_{ij}\mathbb{E}N_1(i), \tag{5.35}$$

because the two equations only differ by a factor of $1/\mathbb{E}T_1$. To prove (5.35), first let us consider $j \neq 1$. Because the chain starts at s_1, we have $X_0 = s_1 \neq s_j$ and, in the definition of $N_1(j)$, we can start the summation from $k = 1$,

$$N_1(j) = \sum_{k=0}^{T_1-1} \mathrm{I}(X_k = s_j) = \sum_{k=1}^{T_1-1} \mathrm{I}(X_k = s_j).$$

We can also write this as

$$\begin{aligned} N_1(j) &= \sum_{k=1}^{\infty} \mathrm{I}(X_k = s_j, k < T_1) \\ &= \sum_{k=1}^{\infty} \mathrm{I}(X_k = s_j, k-1 < T_1), \end{aligned} \tag{5.36}$$

where we used that

$$\{X_k = s_j, k < T_1\} = \{X_k = s_j, k-1 < T_1\},$$

because $X_k = s_j \neq s_1$, so it is impossible that $T_1 = k$. When $j = 1$, the equation (5.36) also holds, for a different reason. In this case, $N_1(1) = 1$ by definition, while the right hand side of (5.36) equals to 1 because $k-1 < T_1$ is the same as $k \leq T_1$ and, between times 1 and T_1, the chain visits s_1 once at $k = T_1$.

If we take expectations of both sides of (5.36),

$$\begin{aligned} \mathbb{E}N_1(j) &= \sum_{k=1}^{\infty} \mathbb{P}(X_k = s_j, k-1 < T_1) \qquad (5.37) \\ &= \sum_{k=1}^{\infty}\sum_{i=1}^{m} \mathbb{P}(X_{k-1} = s_i, X_k = s_j, k-1 < T_1), \end{aligned}$$

where we also partitioned the event into different outcomes $X_{k-1} = s_i$. The key observation we need to make is that the event

$$\{k-1 < T_1\} = \{X_1 \neq s_1, \ldots, X_{k-1} \neq s_1\}$$

depends only on the random variables X_1, …, X_{k-1} up to time $k-1$. Let us write one term in the above sum by conditioning on the first $k-1$ random variables,

$$\begin{aligned}&\mathbb{P}(X_{k-1} = s_i, X_k = s_j, k-1 < T_1)\\ &= \mathbb{P}(X_k = s_j \mid X_{k-1} = s_i, k-1 < T_1)\\ &\quad \times \mathbb{P}(X_{k-1} = s_i, k-1 < T_1).\end{aligned}$$

If we use the Markov property in the form (5.28), we can drop the condition $k-1 < T_1$ and rewrite this as

$$\begin{aligned}&\mathbb{P}(X_{k-1} = s_i, X_k = s_j, k-1 < T_1)\\ &= \mathbb{P}(X_k = s_j \mid X_{k-1} = s_i)\mathbb{P}(X_{k-1} = s_i, k-1 < T_1)\\ &= p_{ij}\mathbb{P}(X_{k-1} = s_i, k-1 < T_1).\end{aligned}$$

Plugging this back into (5.37) and rearranging terms,

$$\begin{aligned}\mathbb{E}N_1(j) &= \sum_{k=1}^{\infty}\sum_{i=1}^{m} p_{ij}\mathbb{P}(X_{k-1} = s_i, k-1 < T_1)\\ &= \sum_{i=1}^{m} p_{ij} \sum_{k=1}^{\infty} \mathbb{P}(X_{k-1} = s_i, k-1 < T_1)\\ \{\ell = k-1\} &= \sum_{i=1}^{m} p_{ij} \sum_{\ell=0}^{\infty} \mathbb{P}(X_\ell = s_i, \ell < T_1)\\ &= \sum_{i=1}^{m} p_{ij}\mathbb{E} \sum_{\ell=0}^{T_1-1} \mathrm{I}(X_\ell = s_i)\\ &= \sum_{i=1}^{m} p_{ij}\mathbb{E}N_1(i).\end{aligned}$$

This proves (5.35) and finishes the proof. □

Exercise 5.2.1. Show that if a Markov chain has two different stationary distributions then there exist infinitely many stationary distributions.

Exercise 5.2.2. Find the stationary distribution of Markov chain with the transition matrix

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \end{bmatrix}.$$

If the chain starts at s_2, what is the expectation of the first return time to s_2?

Exercise 5.2.3. If $\mu = (\mu_1, \ldots, \mu_m)$ is the stationary distribution of an irreducible Markov chain, show that $\mu_i \neq 0$ for all $1 \leq i \leq m$.

Exercise 5.2.4 (Another proof of uniqueness). Consider an irreducible Markov chain and suppose that there exist two stationary distributions μ^1 and μ^2. Let j be any minimizer

$$\frac{\mu_j^1}{\mu_j^2} = \min_{i \leq m} \frac{\mu_i^1}{\mu_i^2}.$$

Show that

$$\frac{\mu_j^1}{\mu_j^2} = \frac{\mu_i^1}{\mu_i^2}$$

for any i such that $p_{ij} > 0$. From this, derive that $\mu^1 = \mu^2$.

Exercise 5.2.5. If $p_{1j} = p_{2j}$ for all j and

$$Y_n = \begin{cases} s_j, & \text{if } X_n = s_j \text{ for } j \geq 3, \\ s^*, & \text{if } X_n = s_j \text{ for } j \leq 2, \end{cases}$$

show that $Y_1, Y_2, \ldots$ is a Markov chain on the new state space $S = \{s^*, s_3, \ldots, s_m\}$.

Exercise 5.2.6. If a Markov chain is irreducible, prove that all moments $\mathbb{E}T_1^p < \infty$ for $p \geq 1$, where T_1 is defined in (5.23). *Hint:* use Exercise 1.2.5 in Section 1.2.

Exercise 5.2.7. In Toronto, it rains (or snows) on about 38% of days per year. Let us consider a Markov chain model of weather with two states $s_1 =$ 'rain', $s_2 =$ 'no rain', and the transition matrix

$$P = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix},$$

for some $p, q \in (0,1)$, so that if it rains today then it will rain tomorrow with probability p, and if it does not rain today then it will not rain tomorrow with probability q. Find all p and q for which the expected proportion of rainy days over long periods of time is 0.38.

Exercise 5.2.8. Suppose that the state s_1 is inessential and let S_1 be the cluster of states that communicate with s_1. Suppose that the chain starts at s_1, and let T be the first time we exit the cluster S_1,

$$T = \min\{n \geq 1 : X_n \notin S\}.$$

Prove that $\mathbb{E}T < \infty$.

Exercise 5.2.9. Let us consider an irreducible Markov chain $X_0, X_1, \ldots$, and let $A \subseteq S$ be a non-empty subset of its states. If the chain starts at s_i then let

$$T_A(s_i) = \min\{n \geq 0 : X_n \in A\}$$

be the first time the chain visits one of the states in A and let $t_i = \mathbb{E}T_A(s_i)$ be its expectation. Prove that

(a) If $s_i \in A$ then $t_i = 0$.
(b) If $s_i \notin A$ then $t_i = 1 + \sum_{j=1}^{m} p_{ij} t_j$.
(c) The equations in parts (a) and (b) determine $t = (t_1, \ldots, t_m)$ uniquely. *Hint:* consider equation (5.18).

5.3 Convergence Theorem

In Lemma 5.4 in the last section we considered the sequence of matrices

$$A_n = \frac{1}{n}(1+P+\ldots+P^n)$$

and showed that any subsequential limit A consists of rows which are stationary distributions. For irreducible chains, we also showed that the stationary distribution μ is unique and, therefore,

$$\lim_{n\to\infty} \frac{1}{n}(1+P+\ldots+P^n) = \begin{bmatrix} \mu_1 & \cdots & \mu_m \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_m \end{bmatrix}.$$

Can we expect a stronger statement that

$$\lim_{n\to\infty} P^n = \begin{bmatrix} \mu_1 & \cdots & \mu_m \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_m \end{bmatrix}?$$

First, we need to check that this is, indeed, a stronger statement.

Exercise 5.3.1. If the limit $\lim_{n\to\infty} P^n$ exists then the limit $\lim_{n\to\infty} \frac{1}{n}(1+P+\ldots+P^n)$ exists and is the same.

It is easy to see that this can not be true in general. For example, if a Markov chain has two states s_1, s_2 and it has the transition probabilities $p_{12} = p_{21} = 1$ then,

$$P^{2n} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \; P^{2n+1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Another way to put it is that, if we start the chain in the state s_i, it can be back in s_i only at even times. The problem with this example is that this chain has period 2, which results in this periodic behaviour. However, if the chain is aperiodic then the convergence holds.

Theorem 5.2 (Convergence Theorem). *If a Markov chain is irreducible and aperiodic then*

$$\lim_{n\to\infty} P^n = \begin{bmatrix} \mu_1 & \cdots & \mu_m \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_m \end{bmatrix}, \tag{5.38}$$

where $\mu = (\mu_1, \ldots, \mu_m)$ is its stationary distribution.

Notice that this is equivalent to the following statement: if $\nu = (\nu_1, \ldots, \nu_m)$ is any distribution on the set of states then

$$\lim_{n\to\infty} \nu P^n = \mu. \tag{5.39}$$

In one direction, multiplying the right hand side of (5.38) by ν on the left, obviously, gives μ. In the other direction, if we take $\nu = (0, \ldots, 1, \ldots, 0)$ where the ith coordinate is 1, then

$$\lim_{n\to\infty} \nu P^n = \mu$$

is the i^{th} row of the limit $\lim_{n\to\infty} P^n$, so all the rows of this limit are equal to μ.

Recall that if ν is the initial distribution of the chain, i.e. the distribution of X_0, then νP^n is the distribution of X_n. Hence, the convergence theorem in the form (5.39) states that the distribution of X_n converges to the stationary distribution μ no matter what the initial distribution is.

Proof (Theorem 5.2). We will be proving the formula (5.39). Let us consider the L^1-distance between vectors on $\mathbb{R}^m$,

$$\|x - y\|_1 = \sum_{i=1}^{m} |x_i - y_i|.$$

Let us notice that multiplying by any transition matrix P on the right does not increase this distance, because

$$\begin{aligned}\|xP-yP\|_1 &= \sum_{j=1}^{m}\left|\sum_{i=1}^{m} p_{ij}(x_i-y_i)\right| \\ &\leq \sum_{j=1}^{m}\sum_{i=1}^{m} p_{ij}|x_i-y_i| \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij}|x_i-y_i| \\ &= \sum_{i=1}^{m}|x_j-y_j| = \|x-y\|_1.\end{aligned}$$

If we apply this to distributions ν and μ and use that $\mu = \mu P$, we get that

$$\|\nu P-\mu\|_1 \leq \|\nu-\mu\|_1.$$

This suggests that the distributions get closer to each other after one step but, in order to prove convergence, we would like to have a strict inequality and then iterate it.

This is where the assumption that our chain is aperiodic comes into play. By Lemma 5.3 in Section 5.1, for large enough N, all entries $p_{ij}(N)$ of P^N are strictly positive. Let $\varepsilon > 0$ be the smallest among its entries. If we apply the above calculation to the transition matrix P^N and distributions ν and μ, the improvement comes from the fact that

$$\sum_{i=1}^{m}\varepsilon(\nu_i-\mu_i) = \varepsilon(1-1) = 0$$

and, therefore,

$$\sum_{i=1}^{m} p_{ij}(N)(\nu_i-\mu_i) = \sum_{i=1}^{m}(p_{ij}(N)-\varepsilon)(\nu_i-\mu_i).$$

Notice that all $p_{ij}(N)-\varepsilon \geq 0$, because ε was the smallest among all entries, so

$$\begin{aligned}
\|\nu P^N - \mu P^N\|_1 &= \sum_{j=1}^{m} \Big| \sum_{i=1}^{m} p_{ij}(N)(\nu_i - \mu_i) \Big| \\
&= \sum_{j=1}^{m} \Big| \sum_{i=1}^{m} (p_{ij}(N) - \varepsilon)(\nu_i - \mu_i) \Big| \\
&\le \sum_{j=1}^{m} \sum_{i=1}^{m} (p_{ij}(N) - \varepsilon)|\nu_i - \mu_i| \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} (p_{ij}(N) - \varepsilon)|\nu_i - \mu_i| \\
&= (1 - m\varepsilon) \sum_{i=1}^{m} |\nu_j - \mu_j|.
\end{aligned}$$

Thus, $\|\nu P^N - \mu P^N\|_1 \le (1 - m\varepsilon)\|\nu - \mu\|_1$ and, by induction,

$$\|\nu P^{Nk} - \mu P^{Nk}\|_1 \le (1 - m\varepsilon)^k \|\nu - \mu\|_1.$$

For powers in between the multiples of N, of the form $Nk + \ell$ for $\ell \le N$, we use

$$\begin{aligned}
\|\nu P^{Nk+\ell} - \mu P^{Nk+\ell}\|_1 &\le (1 - m\varepsilon)^k \|\nu P^\ell - \mu P^\ell\|_1 \\
&\le (1 - m\varepsilon)^k \|\nu - \mu\|_1.
\end{aligned}$$

This proves that $\|\nu P^n - \mu P^n\|_1 \to 0$ for any two distributions ν and μ. When μ is the stationary distribution, $\mu P^n = \mu$, so $\|\nu P^n - \mu\|_1 \to 0$. This finishes the proof. □

Exercise 5.3.2. Given $p \in (0,1)$, consider a Markov chain with the transition probabilities

$$p_{i,i-1} = 1 - p, p_{i,i+1} = p \text{ for } i = 2, \ldots, m-1$$

and $p_{1,1} = 1 - p$, $p_{1,2} = p$, $p_{m,m-1} = 1 - p$, $p_{m,m} = p$. What is the limit $\lim_{n\to\infty} P^n$. *Hint:* solve the system $\mu = \mu P$ explicitly.

Exercise 5.3.3. Show that if a Markov chain is irreducible and aperiodic then, for any function $f\colon S \to \mathbb{R}$, the expected value $\mathbb{E}f(X_n)$ converges.

Exercise 5.3.4. Suppose that m people sit at the round table and all have plates with rice in front of them. At the same time, each person divides his of her rice in half and evens out one half with the half of the person on the right (i.e. they split the total of their halves evenly), and another half with the person on the left. If they keep repeating this, what will happen in the long run?

Exercise 5.3.5. Let us consider the L^∞ norm on $\mathbb{R}^m$,

$$||x-y||_\infty = \max_{i\le m} |x_i - y_i|.$$

If P is a Markov transition matrix P, show that

$$||Px-Py||_\infty \le ||x-y||_\infty.$$

5.4 Reversible Markov Chains

Let us consider a Markov chain with the transition matrix P. A probability distribution $\mu = (\mu_1, \dots, \mu_m)$ on the set of states is called *reversible for the chain* if

$$\mu_i p_{ij} = \mu_j p_{ji} \tag{5.40}$$

for all i and j. A Markov chain is called *reversible* if it has a reversible distribution. The equations (5.40) are called the *detailed balance equations*. If we start the chain with the distribution μ then (5.40) states that the first two outcomes (s_i, s_j) and (s_j, s_i) are equally likely. Moreover,

$$\begin{aligned}
&\mathbb{P}(X_0 = s_i, X_1 = s_j, X_2 = s_k) \\
&= \mu_i p_{ij} p_{jk} = p_{ji} \mu_j p_{jk} = p_{ji} p_{kj} \mu_k \\
&= \mathbb{P}(X_0 = s_k, X_1 = s_j, X_2 = s_i),
\end{aligned}$$

which means that the first three outcomes (s_i, s_j, s_k) and (s_k, s_j, s_i) are equally likely. The same calculation show that any sequence of outcomes and the same sequence in reverse order are equally likely, hence the name reversible. It is easy to check that a reversible distribution is stationary.

Lemma 5.7. *A reversible distribution μ is stationary.*

Proof. Summing (5.40) over i, we get

$$\sum_{i=1}^{m} \mu_i p_{ij} = \mu_j \sum_{i=1}^{m} p_{ji} = \mu_j,$$

which finishes the proof. □

There are a number of examples, where the equations (5.40) are easy to guess or easy to check, so this is one good way to find a stationary distribution. Let us consider some classical examples of reversible Markov chains.

Example 5.4.1 (Birth-and-death processes). Let us consider a Markov chain that in one step moves from a state s_i only to s_{i+1} or s_{i-1}. Of course, since our state space is finite, this means that from s_1 it only moves up to s_2 and from s_m it only moves down to s_{m-1}. The chain can also stay in each state s_i. Another way to say this is that

$$p_{ij} > 0 \text{ if } |i-j| = 1, \text{ and } p_{ij} = 0 \text{ if } |i-j| > 1.$$

The probabilities p_{ii} may be positive or zero. Such Markov chains are called *birth-and-death* chains, and they are all reversible. To find a reversible distribution, let us set $\mu_1 = x$ as an unknown variable. If (5.40) holds then

$$\mu_2 = \frac{\mu_1 p_{12}}{p_{21}} = \frac{p_{12}}{p_{21}} x.$$

Given μ_2, we can find μ_3,

$$\mu_3 = \frac{\mu_2 p_{23}}{p_{32}} = \frac{p_{12}}{p_{21}} \frac{p_{23}}{p_{32}} x,$$

and we can continue, by induction,

$$\mu_i = \frac{p_{12}}{p_{21}} \cdots \frac{p_{i-1,i}}{p_{i,i-1}} x.$$

Since all μ_i are of the form $c_i x$ for some positive $c_i > 0$, to find x we can use that the probabilities add up to 1, so

$$x = \frac{1}{c_1 + \ldots + c_m}.$$

Of course, the same calculation works for any chain with the transition graph that looks like a single path between two 'endpoint' vertices, and we can relabel the states of such chain to turn it into a formal birth-and-death chain. □

Example 5.4.2 (The Ehrenfest model of diffusion). One example of a birth-and-death Markov chain is Ehrenfest's model of diffusion. Suppose there are m particles inside two connected containers. The particles can move between the two containers, and suppose that the next particle that moves is chosen uniformly at random. This process can be described by a Markov chain with $m+1$ states s_i for $i = 0, \dots, m$, where the system is in the state i if there are i particles in the first container and $m-i$ in the second. The transition probabilities are given by

$$p_{i,i-1} = \frac{i}{m},\ p_{i,i+1} = \frac{m-i}{m}. \tag{5.41}$$

Since this is a birth-and-death chain, it is reversible and it turns out that its reversible distribution is the Binomial $B(m, \frac{1}{2})$ distribution,

$$\mu_i = \binom{m}{i} \frac{1}{2^m.}$$

We will leave it as an exercise to check that the detailed balance equations (5.40) are satisfied. □

Example 5.4.3 (Random walks on graphs). Let us consider a connected graph G on the set of vertices $\{s_1, \dots, s_m\}$. This means that all vertices are connected by a path on edges of the graph. Let N_i be the number of neighbours of s_i. Let us consider a Markov chain with the transition probabilities

$$p_{ij} = \begin{cases} \frac{1}{N_i}, & \text{if } s_j \text{ is a neighbour of } s_i, \\ 0, & \text{otherwise.} \end{cases} \tag{5.42}$$

In other words, in each state s_i the chain picks a neighbour uniformly at random and moves there. A distribution μ is reversible for this chain if

$$\mu_i \frac{1}{N_i} = \mu_j \frac{1}{N_j}.$$

Therefore, all the ratios $\frac{\mu_i}{N_i}$ must be equal to the same constant c, so $\mu_i = cN_i$. Since the probabilities must add up to 1, the constant $c = \frac{1}{N}$, where $N = N_1 + \ldots + N_m$, and

$$\mu_i = \frac{N_i}{N}. \tag{5.43}$$

Since the graph is connected, the Markov chain is irreducible and μ is the unique stationary distribution. □

The most important examples of reversible Markov chains appear in the context of *Markov Chain Monte Carlo (MCMC)* algorithms. Let us describe one such algorithm.

Example 5.4.4 (Metropolis algorithm). As in the previous example, let us consider a connected graph G on the set of vertices $S = \{s_1, \ldots, s_m\}$. Suppose that we are interested in some probability distribution μ on S, but μ is defined by a complicated model that makes the computation of its coordinates μ_i impractical. On the other hand, for any neighbours s_i and s_j on the graph, μ_i and μ_j are related in some simple way, so that we know the ratio

$$r_{ij} = \frac{\mu_i}{\mu_j}$$

without knowing μ_i and μ_j. Of course, given r_{ij}, the distribution μ can in principle be reconstructed by setting $\mu_1 = x$ as unknown parameter, then finding $\mu_i = r_{i1}x$ in terms of x for all its neighbours, and propagating this through the graph to express all $\mu_i = xc_i$ in terms of x. Then we find

$$x = \frac{1}{c_1 + \ldots + c_m},$$

because the probabilities must add up to one. However, in many applications the graph is so large (say $m = 2^{20}$) that it is not even feasible to add up m numbers.

If we only know the above ratios r_{ij}, how can we compute, for example, the expectation of some function $f \colon S \to \mathbb{R}$ on

S with respect to the distribution μ,

$$\mathbb{E}_\mu f := \sum_{i=1}^{m} f(s_i)\mu_i,$$

without computing μ? The idea of the Markov Chain Monte Carlo method is to combine the law of large numbers with the convergence theorem for Markov chains. Suppose that we can construct a Markov chain with the stationary distribution μ, whose transition probabilities depends only on the ratios r_{ij}. Then we can pick the initial state at random and run the chain for a large number of steps, n, to obtain a random variable X_n with the distribution close to μ. If we repeat this procedure N times, we will get N i.i.d. random variables $X_n^1, \dots, X_n^N$ with the distribution close to μ. By the law of large numbers,

$$\frac{1}{N}\sum_{k=1}^{N} f(X_n^k) \approx \mathbb{E} f(X_n^1) \approx \sum_{i=1}^{m} f(s_i)\mu_i.$$

In other words, this procedure allows us to approximate the expectation of f with respect to the distribution μ, when it is not feasible to compute μ directly.

Here is a classical construction of a Markov chain, called the *Metropolis* chain, with the stationary distribution μ and the transition probabilities that depend only on the ratios r_{ij}. Let us write $i \sim j$ to denote that s_i and s_j are neighbours. For each vertex s_i, we define

$$p_{ij} = \frac{1}{N_i}\min\Big(\frac{\mu_j N_i}{\mu_i N_j}, 1\Big) \tag{5.44}$$

if $j \sim i$ and define

$$p_{ii} = 1 - \sum_{j \sim i} p_{ij}. \tag{5.45}$$

All other $p_{ij} = 0$. Notice that $p_{ii} \geq 0$, because

$$\sum_{j\sim i} p_{ij} \le \sum_{j\sim i} \frac{1}{N_i} = 1,$$

so this definition makes sense. Notice also that the transition probabilities depend only on the ratios r_{ij}.

To show that μ is the stationary distribution of this Markov chain, we will check that μ satisfies the detailed balance equations (5.40). For $i = j$ there is nothing to check, and if $i \not\sim j$ then both sides are zero, so we only need to check this for $i \sim j$. If $\mu_j N_i \le \mu_i N_j$ then, by (5.44),

$$p_{ij} = \frac{1}{N_i} \text{ and } p_{ji} = \frac{1}{N_j}\frac{\mu_i N_j}{\mu_j N_i} = \frac{\mu_i}{\mu_j N_i},$$

which immediately implies (5.40). The case $\mu_j N_i \ge \mu_i N_j$ is exactly the same, so the Metropolis chain has the stationary distribution μ.

If at least one $p_{ii} > 0$ then the chain is aperiodic. If the chain is periodic, a slight modification will make it periodic without affecting the stationary distribution μ (see exercise below). Then the convergence theorem applies and we can use this chain to produce an i.i.d. sample with the distribution close to μ. □

Exercise 5.4.1. Show that if μ is a reversible distribution for P, it is also a reversible distribution for $\lambda I + (1-\lambda)P$ for any $\lambda \in [0,1]$.

Exercise 5.4.2. Show that if μ is a reversible distribution for P, it is also a reversible distribution for P^n.

Exercise 5.4.3. Show that a birth-and-death Markov chain is aperiodic if and only if at least one $p_{ii} > 0$.

Exercise 5.4.4. Show that a Markov chain with the following transition matrix is not reversible:

$$P = \begin{bmatrix} 0 & 0.75 & 0.25 \\ 0.25 & 0 & 0.75 \\ 0.75 & 0.25 & 0 \end{bmatrix}.$$

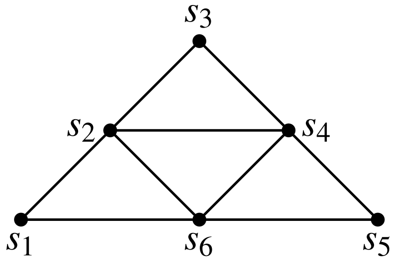
Exercise 5.4.5. Check that the Binomial distribution $B(m, \frac{1}{2})$ satisfies the detailed balance equations (5.40) for the Ehrenfest chain (5.41).

Exercise 5.4.6. Suppose that two containers contain a total of m while balls and m black balls. At each step a ball is chosen at random from all $2m$ balls and put into the other container. We say that a system is in the state s_i if there are i white balls in the first container. Find the transition probabilities of this chain and compute $\lim_{n\to\infty} P^n$.

Exercise 5.4.7. Consider the Ehrenfest chain (5.41) and let D_n be the difference of particles in the two containers at time n. This means that if the chain is in the state s_i at time n then $D_n = 2i - m$. Prove that

$$\mathbb{E}D_{n+1} = \frac{m-2}{m}\mathbb{E}D_n = \left(\frac{m-2}{m}\right)^n \mathbb{E}D_0.$$

Exercise 5.4.8. Consider a random walk on the following graph:



If we start the chain at s_1, what is the expected time of the first return to s_1? What is the expected number of visits to s_4 before the first return to s_1? *Hint:* recall the results in Section 5.2.

References

1. Alon, N.; Spencer, J.H.: The probabilistic method. Wiley Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., Hoboken, NJ, 2016.
2. Boucheron, S.; Lugosi, G.; Massart, P.: Concentration inequalities. A nonasymptotic theory of independence. Oxford University Press, Oxford, 2013.
3. Borovkov, A. A.: Probability theory. Universitext. Springer, London, 2013.
4. Durrett, R.: Elementary probability for applications. Cambridge University Press, Cambridge, 2009.
5. Feller, W.: An introduction to probability theory and its applications. Vol. I. John Wiley & Sons, Inc., New York-London-Sydney,1968.
6. Grimmett, G.R.; Stirzaker, D. R.: One thousand exercises in probability. Oxford University Press, Oxford, 2001.
7. Häggström, O.: Finite Markov chains and algorithmic applications. London Mathematical Society Student Texts, 52. Cambridge University Press, Cambridge, 2002.
8. Kingman, J. F. C.: Poisson processes. Oxford University Press, New York, 1993.
9. Levin, D. A.; Peres, Y.; Wilmer, E.L.: Markov chains and mixing times. American Mathematical Society, Providence, RI, 2009.
10. Rozanov, Y. A.: Probability theory. A concise course. Dover Publications, Inc., New York, 1977.

Index