PROBABILITY

Antti Knowles

Fall Semester 2025

Version 22 Sep 2025

FernUni.ch
UniDistance.ch

Contents

Preface

These are lecture notes for a one-semester course on probability theory. They are meant to be fully self-contained, assuming a basic knowledge of measure theory (which is reviewed briefly the beginning). For further reading, I recommend the three following standard references, on which these lecture notes are in part based:

- Rick Durrett: Probability: theory and examples
- Jean Jacod and Philip Protter: Probability essentials (Springer, 2004).
- Jean-François Le Gall: Intégration, probabilités et processus aléatoires (French).

There are of course many other excellent books on the subject.

These notes are liable (i.e. virtually certain) to contain typos. If you find any, please make sure you tell me!

Before getting into the subject proper, in this short preface we give a very brief overview of the subject's history and of its relation to the natural sciences, with which it has always had a close interaction. This is meant only for the curious reader, and does not constitute a part of the course itself.

Probability is the study of uncertain events – events whose outcome cannot be predicted with certainty. Examples of such events include

(i) I obtain heads when I flip a coin;
(ii) it rains in Brig tomorrow;
(iii) my kitchen light breaks in the next six months.

The classical view of how uncertainty arises in nature is based on nineteenth century physics (Newtonian mechanics and Maxwell's electrodynamics), where the state of a physical system at any time is a deterministic function of its initial state. In principle, therefore, the future state of any system is fully predictable, provided we have precise enough information about its current state. From this point of view, the uncertainty of a future event is simply an expression of a lack of knowledge about the present. In reality, however, this point of view is essentially useless for most systems of interest. This is because the complexity of the system and the sensitive dependence on the initial state means that the required precision in the knowledge of the initial state is not achievable by any conceivable means. A famous example is the impossibility of predicting the weather for more than two weeks into the future. A simpler example is the humble coin flip or toss of a die, whose outcome cannot be determined in advance no matter how accurately the initial toss is determined. The quantum revolution of the first half of the twentieth century went further: uncertainty is *inherent* in the laws of nature, and even simple physical systems behave in an intrinsically random fashion, no matter how accurately one determines the initial data (a famous example is the double slit experiment in quantum mechanics).

The historical development of probability was initially motivated by the desire for a theoretical understanding of gambling, in the sixteenth and seventeenth centuries. Today, probability theory has become one of the theoretical foundations of our modern society. It underpins statistics, machine learning, artificial intelligence, and computer science. It also constitutes the bedrock of any experimental discipline, and as such lies at the heart of the natural and social sciences.

Aside from its applications, probability theory is an area of pure mathematics, which has flourished in the past fifty years. Having shed its former reputation of an application-driven low-brow game of counting balls and boxes, it has become one of the most central and active areas of pure mathematics.

The study of probability can be roughly divided into two disciplines, which, while not wholly separate, have rather different goals and ways of thinking.

1. *Probability theory* – an area of mathematics, which develops a calculus for determining the probability of an event starting from a set of mathematical axioms. As a mathematical theory, it is purely a logical construct and detached from any interpretation in the real world. Its origins trace back to Blaise Pascal and Pierre de Fermat in the seventeenth century. It was put on a rigorous axiomatic basis by Kolmogorov in 1933, an achievement usually regarded as the beginning of modern probability theory.

   As we shall see, Kolmogorov's axioms build on measure theory. Thus, one could make the case that probability theory is nothing but a special case of analysis and measure theory. This point of view is however simplistic and often even misleading, since probability theory has developed its own very particular way of thinking, characterised by concepts such as independence, conditioning, and infinite product spaces.

2. *Interpretation of probability* – an area of epistemology and statistics, which aims to connect mathematical probability theory with random experiments. It strives to give meaning to probabilistic claims about real-world events, or in other words to give an *interpretation* of probability. There are several competing schools of thought, each with their strengths and weaknesses; which interpretation to adopt in a given situation is sometimes a matter of personal preference.

   For instance, returning to the example (i) above, what does the claim "the probability of obtaining heads when flipping a coin is 50%" mean? The most natural, and indeed oldest, interpretation is that of *frequentist probability*: the probability of a random event is the relative frequency of occurrence of the event when the experiment is repeated indefinitely and independently. The frequentist interpretation is independent of the observer, and it is the most prominent instance of an *objective interpretation* of probability.

   What about the example (ii) above? The frequentist interpretation fails here, because the event in question – it rains in Brig tomorrow – cannot be repeated independently: the current weather conditions are unique and we have no control over them. Nevertheless, we all believe that claims of the form "the probability that it rains in Brig tomorrow is 20%" somehow make sense, and indeed that is how weather forecasts are often formulated. For such events, a *subjective interpretation* of probability imposes itself, whereby the probability of an event corresponds to a *degree of belief* by a knowledgeable person, who incorporates expert knowledge (such as meteorology and weather models) and experimental data (such as the current and past weather conditions). The most popular version of subjective probability is *Bayesian probability*, where the expert knowledge is translated into a subjective *prior probability distribution* (an educated guess), which is then updated based on experimental data to obtain a *posterior probability distribution*. Different prior probability distributions will give

rise to different posterior probability distribution when given the same experimental data. This captures the subjective element of Bayesian probability. In everyday life this is clearly illustrated by the fact that we often use several different weather apps to check the weather forecast, since they typically give different probabilities for the same event[1].

As for the example (iii), a frequentist interpretation is possible if I have a large supply of identical copies of my kitchen light, which I can test individually and measure the proportion of lights that fail in the next six months. On the other hand, if my kitchen light is a unique sample (say an inherited antique piece), a subjective interpretation is required.

In most instances, if one is familiar with probability theory, simple common sense is sufficient to answer probabilistic questions about the real world. Nevertheless, aside from important philosophical questions it raises, the study of the interpretation of probability can be of great practical importance in several applied fields. This is typically discussed in more detail in classes on statistics.

Being a course on mathematics, this course is entirely devoted to mathematical probability theory. Henceforth, we shall wrap ourselves in the warm blanket of mathematical rigour and axiomatic deduction, without having to worry about tricky epistemological questions raised by interpretation[2].

[1] See `https://www.rmets.org/metmatters/what-does-30-chance-rain-mean` for an insightful and more detailed explanation of the meaning of probabilities in weather forecasts.

[2] A notable exception will be the final chapter on statistics, where we shall shed this blanket and jump into the cold pool of reality and empiricism.

CHAPTER 1

Recap of measure theory

WEEK 1

Since probability theory is founded on measure theory, in this preliminary chapter we give a review of the most important ingredients from measure theory. It is meant to be understandable for a reader who has learned some basic measure theory but may have forgotten some details or more technical aspects of it.

For full details and for proofs, we refer to Chapter 1 of the course *Calculus II* that you took last year.

Throughout these notes we use the following standard notations. We write $\mathbb{N} = \{0, 1, 2, \dots\}$ and $\mathbb{N}^* = \{1, 2, 3, \dots\}$. For a finite set X, we denote by $\#X$ the number of elements of X. For a set X and a subset $A \subset X$, we write $A^c := X \setminus A$ and denote by $\mathcal{P}(X)$ the collection of all subsets of X. We denote by $\mathbf{1}_A$ the *indicator function*[1] of the set A, defined through

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

We also use the notations $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$, which are common in probability theory. Moreover, we write $a_+ := a \vee 0$ and $a_- := (-a) \vee 0$ for the positive and negative parts of a real number a. Hence, for any $a \in \mathbb{R}$ we have $a = a_+ - a_-$ with $a_+, a_- \geqslant 0$. We often use the nonnegative reals augmented with ∞, denoted by $[0, \infty]$. They satisfy the obvious order relations as well as the convention $0 \cdot \infty = 0$.

Definition 1.1 Let X be a set. A *σ-algebra* (σ-field) on X is a collection $\mathcal{A}$ of subsets of X satisfying

(i) $X \in \mathcal{A}$;
(ii) $A \in \mathcal{A} \;\Rightarrow\; A^c \in \mathcal{A}$;
(iii) if $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$ then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

If $\mathcal{A}$ is a σ-algebra on X, then we say that any $A \in \mathcal{A}$ is a *measurable subset* of X, and call $(X, \mathcal{A})$ a *measurable space*.

The following construction plays a particularly prominent role in probability.

Definition 1.2 Let $\mathcal{C} \subset \mathcal{P}(X)$. Then

$$\sigma(\mathcal{C}) := \bigcap_{\substack{\mathcal{A} \text{ is a } \sigma\text{-algebra} \\ \mathcal{C} \subset \mathcal{A}}} \mathcal{A}$$

is the *σ-algebra generated by $\mathcal{C}$*.

[1] The term *indicator function* is used in probability theory, while the same object is usually called *characteristic function* in analysis. As we shall see in Section 4.3, the latter term is reserved for a very different object in probability theory.

The σ-algebra generated by $\mathcal{C}$ is indeed a σ-algebra as its name implies, because the intersection of σ-algebras is a σ-algebra.

Example 1.3

(i) Let $X = \mathbb{R}^d$ and $\mathcal{O}$ be the collection of open subsets of $\mathbb{R}^d$. (More generally, X can be a topological space whose collection of open sets is $\mathcal{O}$.) Then $\mathcal{B}(X) := \sigma(\mathcal{O})$ is the Borel σ-algebra of X.

(ii) Let $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ be measurable spaces. The *product σ-algebra* on $X_1 \times X_2$ is

$$\mathcal{A}_1 \otimes \mathcal{A}_2 := \sigma(A_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2).$$

Definition 1.4 A (positive) *measure* on a measurable space $(X, \mathcal{A})$ is a function $\mu\colon \mathcal{A} \to [0, \infty]$ satisfying $\mu(\emptyset) = 0$ and $\mu\big(\bigcup_{n\in\mathbb{N}} A_n\big) = \sum_{n\in\mathbb{N}} \mu(A_n)$ for any countable family $(A_n)_{n\in\mathbb{N}}$ of disjoint measurable subsets.

Example 1.5

(i) Let X be finite or countable, $\mathcal{A} = \mathcal{P}(X)$, and $\mu(A) := \#A$. This is the *counting measure* on X.

(ii) For $x \in X$ we define the *Dirac delta measure at x* through

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

(iii) The Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is defined as the unique measure λ satisfying $\lambda((a, b)) = b - a$ for all $a < b$. (Recall from your course on measure theory that the existence and uniqueness of λ is nontrivial. Later in this class we shall give a proof of uniqueness: see Theorem 3.10 below.)

A measurable space $(X, \mathcal{A})$ endowed with a measure μ is called a *measure space* and denoted by the triple $(X, \mathcal{A}, \mu)$.

Definition 1.6 Let $(X, \mathcal{A}, \mu)$ be a measure space. Then a property $P(x)$ depending on $x \in X$ holds *almost everywhere* if

$$\mu(\{x \in X : P(x) \text{ false}\}) = 0\,.$$

For example, on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ endowed with Lebesgue measure, the indicator function $\mathbf{1}_{\mathbb{Q}}$ equals 0 almost everywhere, or $\mathbf{1}_{\mathbb{Q}}(x) = 0$ for almost all x.

Definition 1.7 Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A function $f : X \to Y$ is *measurable* if for all $B \in \mathcal{B}$ we have $f^{-1}(B) \in \mathcal{A}$.

Here, f^{-1} denotes the preimage function on sets, i.e. $f^{-1}(B) := \{x \in X : f(x) \in B\}$.

Often, the σ-algebras $\mathcal{A}$ and $\mathcal{B}$ are clear from the context, and we do not even mention them explicitly.

The following definition allows one to transport measures between measurable spaces using measurable functions.

Definition 1.8 Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $f : X \to Y$ measurable, and μ a measure on $(X, \mathcal{A})$. Then we define the *pushforward* or *image measure* of μ under f, denoted by $f_*\mu$, as the measure on $(Y, \mathcal{B})$ defined by

$$f_*\mu(B) := \mu(f^{-1}(B)) \quad \text{for all } B \in \mathcal{B}\,.$$

We now recall the notation for the integral.

Definition 1.9 Let μ be a measure on $(X, \mathcal{A})$.

(i) Let $f : X \to [0, \infty]$. We use the notation

$$\int f \,\mathrm{d}\mu = \int f(x)\,\mu(\mathrm{d}x) \in [0, \infty]$$

for the integral of f with respect to μ (see the class on measure theory for its definition, which is also briefly reviewed below).

(ii) A function $f : X \to \mathbb{R}$ is called *integrable* if $\int |f| \,\mathrm{d}\mu < \infty$, in which case we define

$$\int f \,\mathrm{d}\mu := \int f_+ \,\mathrm{d}\mu - \int f_- \,\mathrm{d}\mu\,.$$

It is helpful to recall briefly the construction of the integral in Theorem 1.9 (i). It proceeds in two main steps.

1. We integrate a *simple function* (finite linear combination of indicator functions) of the form

$$f = \sum_{i=1}^{n} c_i \,\mathbf{1}_{A_i}\,, \tag{1.1}$$

where $c_i \in [0, \infty)$ and $A_i \in \mathcal{A}$ for all $i = 1, \dots, n$. By definition, the integral of this simple function is

$$\int f \,\mathrm{d}\mu := \sum_{i=1}^{n} c_i\,\mu(A_i)\,.$$

It is not hard to check that the left-hand side does not depend on the representation of the simple function f (if the sets A_i are not disjoint and the c_i are not distinct then the representation (1.1) of a simple function is not unique.)

2. Next, we note that an arbitrary measurable function $f : X \to [0, \infty]$ can be approximated monotonically from below by step functions f_n. For example, we can choose f_n to be equal to f rounded to the nearest multiple of 2^{-n} smaller than f, truncated at n, i.e.

$$f_n(x) := \left(2^{-n}\lfloor 2^n f(x)\rfloor\right) \wedge n\,,$$

where $\lfloor\cdot\rfloor$ denotes the integer part. (Plot this function!) Note that $(f_n(x))$ is a nondecreasing sequence for all $x \in X$. Then we *define* the integral of f through

$$\int f \,\mathrm{d}\mu := \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu\,,$$

where the limit exists in $[0, \infty]$ because it is the limit of a nondecreasing sequence. One can check that the left-hand side does not depend on the choice of the sequence f_n.

The preceding definition captures a basic idea of measure theory, which we shall consistently and often tacitly use in this class: one can define the integral of *any* function

provided that it is nonnegative, in which case the integral may be infinite. If the function is not nonnegative, then one has to impose that it is integrable for its integral to make sense. (Otherwise one might end up with expressions of the form $\infty - \infty$, which are ill-defined.)

The integral satisfies the three following convergence theorems, which are stated for some fixed measure space $(X, \mathcal{A}, \mu)$.

Proposition 1.10 (Monotone convergence, Beppo-Levi) *Let $f_1, f_2, \ldots : X \to [0, \infty]$ be a pointwise nondecreasing sequence of measurable functions. Then*

$$\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu = \int \lim_{n\to\infty} f_n \, \mathrm{d}\mu \, .$$

Proposition 1.11 (Fatou's lemma) *Let $f_1, f_2, \ldots : X \to [0, \infty]$ be a sequence of measurable functions. Then*

$$\liminf_{n\to\infty} \int f_n \, \mathrm{d}\mu \geqslant \int \liminf_{n\to\infty} f_n \, \mathrm{d}\mu \, .$$

Proposition 1.12 (Dominated convergence, Lebesgue) *Let $g, f, f_1, f_2, \ldots$ be measurable functions. Suppose that $f_n \to f$ almost everywhere, that g is integrable, and that $|f_n| \leqslant g$ almost everywhere for all n. Then*

$$\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu \, .$$

Next, we recall the notation of product measure. Its uniqueness is guaranteed by the following finiteness property. A measure μ on $(X, \mathcal{A})$ is *σ-finite* if there exists a countable decomposition $X = \bigcup_{n\in\mathbb{N}} X_n$ of X such that $\mu(X_n) < \infty$ for all $n \in \mathbb{N}$. (For instance Lebesgue measure on $\mathbb{R}$ is σ-finite but not finite.)

Definition 1.13 Let μ and ν be σ-finite measures on $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$, respectively. The *product measure* $\mu_1 \otimes \mu_2$ is the unique measure on $(X_1 \times X_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ satisfying

$$\mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1)\, \mu_2(A_2) \quad \text{for all } A_1 \in \mathcal{A}_1 \text{ and } A_2 \in \mathcal{A}_2.$$

For the proof of existence and uniqueness, we refer to the class on measure theory.

The following theorem states that product measures can be integrated successively over each component separately, provided the function is *nonnegative* or *integrable*.

Proposition 1.14 (Fubini-Tonelli) *Let μ_1 and μ_2 be σ-finite measures on $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$, respectively. Let $f : X_1 \times X_2 \to [0, \infty]$ be measurable. Then*

$$\begin{aligned} \int_{X_1\times X_2} f \, \mathrm{d}(\mu_1 \otimes \mu_2) &= \int_{X_1} \left(\int_{X_2} f(x_1, x_2)\, \mu_2(\mathrm{d}x_2) \right) \mu_1(\mathrm{d}x_1) \\ &= \int_{X_2} \left(\int_{X_1} f(x_1, x_2)\, \mu_1(\mathrm{d}x_1) \right) \mu_2(\mathrm{d}x_2) \, . \end{aligned} \tag{1.2}$$

The same identity holds if $f : X_1 \times X_2 \to \mathbb{R}$ is integrable with respect to $\mu_1 \otimes \mu_2$.

CHAPTER 2

Foundations of probability theory

2.1 Probability spaces

In this section we shall give a motivation of Kolmogorov's axioms of probability. We shall see that a mathematical formulation of probability theory rests on three core ingredients: (i) a set of *realisations*, (ii) a collection of *events*, and (iii) a *probability measure* that expresses probabilities of events.

A random experiment (such as the toss of a die) has a number of possible outcomes or realisations.

(i) We denote by Ω the *set of realisations*. Its elements (realisations of the randomness) are denoted by ω.

We consider two basic examples.

Example 2.1 Toss of a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. The realisation $\omega \in \Omega$ denotes the number shown by the die.

Example 2.2 A game of darts. A person throws a dart at a disc-shaped dartboard. Ω is the unit disc in the plane, $\Omega = \{\omega \in \mathbb{R}^2 : |\omega| \leqslant 1\}$. The realisation $\omega \in \Omega$ denotes where dart hits the dartboard.

These examples show that it makes sense to consider very general sets Ω, from finite to uncountable.

(ii) A collection $\mathcal{A} \subset \mathcal{P}(\Omega)$ is the collection of *events*, i.e. subsets of Ω whose probability can be determined.

Example 2.3 (Theorem 2.1 continued) The event $A = \{2, 4, 6\}$ is the event that I obtained an even number. The event $A = \{6\}$ is the event that I obtained a 6.

Example 2.4 (Theorem 2.2 continued) The event $A = \{\omega \in \mathbb{R}^2 : |\omega| \leqslant 1/20\}$ is the event that I hit the bull's eye of the dartboard.

(iii) A function $\mathbb{P} : \mathcal{A} \to [0, 1]$ determines the *probability* $\mathbb{P}(A)$ of an event $A \in \mathcal{A}$.

Example 2.5 (Theorem 2.1 continued) For a balanced die, we have $\mathbb{P}(\{2,4,6\}) = 1/2$ and $\mathbb{P}(\{6\}) = 1/6$.

Example 2.6 (Theorem 2.2 continued) If the dart hits any region of the dartboard with uniform probability, then we have $\mathbb{P}(\{\omega \in \mathbb{R}^2 : |\omega| \leqslant 1/20\}) = (1/20)^2$ (relative area of bull's eye).

That $\mathbb{P}(A) \in [0,1]$ reflects the fact that probabilities must be nonnegative and cannot exceed $1 = 100\%$. Moreover, we require $\mathbb{P}$ to satisfy the two following obvious properties.

- $\mathbb{P}(\Omega) = 1$. This just expresses that with probability 1 we always see some realisation.
- If A and B are disjoint events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. In other words, the probabilities of mutually exclusive events are additive.

The triple $(\Omega, \mathcal{A}, \mathbb{P})$ therefore looks rather similar to a measure space. Imposing that the additivity property for mutually exclusive events extends to countable families, we arrive at the following celebrated and fundamental definition.

Definition 2.7 (Kolmogorov, 1933) A *probability space* is a measure space $(\Omega, \mathcal{A}, \mathbb{P})$ satisfying $\mathbb{P}(\Omega) = 1$.

A measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$ satisfying $\mathbb{P}(\Omega) = 1$ is called a *probability measure*.

We give two examples that shall accompany us through much of this chapter.

Example 2.8 I throw a balanced die twice:

$$\Omega = \{1,2,\dots,6\}^2\,, \qquad \mathcal{A} = \mathcal{P}(\Omega)\,, \qquad \mathbb{P}(A) = \frac{\#A}{36}\,.$$

Example 2.9 Here is a more interesting (and more subtle) example. I throw a die repeatedly until I obtain a 6. Since I may have to throw the die an arbitrarily large number of times, I choose

$$\Omega = \{1,2,\dots,6\}^{\mathbb{N}^*}\,.$$

As a reminder, this is the set of sequences $\omega : \mathbb{N}^* \to \{1,2,\dots,6\}$. We use the notation $\omega = (\omega_k)_{k\in\mathbb{N}^*}$ for its elements.

The set Ω is uncountable, and as we shall see it is ill-advised to take $\mathcal{A}$ to be the full power set $\mathcal{P}(\Omega)$. To find the correct choice for $\mathcal{A}$, let us begin by noting that we certainly want to assign a probability to any event depending on a finite number of throws (such as "the first 10 throws are all smaller than 4"). Generally, such an event is called a *cylinder set*, and it is of the form

$$\{\omega \in \Omega : \omega_1 = i_1, \dots, \omega_n = i_n\}\,, \tag{2.1}$$

which is indexed by the parameters $n \in \mathbb{N}^*$ and $i_1, \dots, i_n \in \{1,2,\dots,6\}$. Hence, we define $\mathcal{A}$ to be the σ-algebra generated by the cylinder sets, i.e.

$$\mathcal{A} = \sigma\Big(\{\omega \in \Omega : \omega_1 = i_1, \dots, \omega_n = i_n\} : n \in \mathbb{N}^*, i_1, \dots, i_n \in \{1,2,\dots,6\}\Big)\,. \tag{2.2}$$

The σ-algebra $\mathcal{A}$ thus constructed is called the *cylinder* σ-algebra, and it plays a fundamental role in probability. It is the canonical σ-algebra on an infinite product space (such as Ω).
Clearly, the probability measure $\mathbb{P}$ on $\mathcal{A}$ should have the following value on any cylinder set:

$$\mathbb{P}\Big(\{\omega \in \Omega : \omega_1 = i_1, \dots, \omega_n = i_n\}\Big) = \left(\frac{1}{6}\right)^n . \tag{2.3}$$

In fact, we shall prove later that there exists a unique measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$ satisfying (2.3).

We conclude with a more difficult example, which is of great interest in mathematics and the sciences. It goes beyond the scope of this course, but we can nevertheless mention its basic mathematical structure.

Example 2.10 You will probably have heard of *Brownian motion*, which was first observed by the botanist Robert Brown in 1827. With a microscope, he observed a particle of pollen immersed in water and noticed that it underwent an erratic random motion. Brownian motion was famously studied by Albert Einstein in one of his groundbreaking papers of 1905, where he gave a theoretical explanation of its origin.
The random realisation is the entire trajectory of the particle, so that we choose

$$\Omega = C([0, \infty), \mathbb{R}^3)$$

to be the space of continuous paths $\omega = (\omega(t))_{t \geqslant 0}$ in $\mathbb{R}^3$. For the collection of events, as in the previous example, we choose the cylinder σ-algebra, which in this instance takes the form

$$\mathcal{A} = \sigma\Big(\{\omega \in \Omega : \omega(t) \in B\} : t \in [0, \infty), B \in \mathcal{B}(\mathbb{R}^3)\Big) .$$

(If you wish, you can think about the analogy between this definition and (2.2). It may help to consider intersections of cylinder sets $\{\omega \in \Omega : \omega(t) \in B\}$.) What about the probability measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$? Clearly, there are many possible choices, but one of them stands out by being by far the most natural one; it is called *Wiener measure*, an infinite-dimensional Gaussian measure which underlies the mathematical definition of Brownian motion. We shall not discuss it further in this course.

Remark 2.11 We conclude this section with an important remark. Since a probability measure $\mathbb{P}$ is a measure, we always have $\mathbb{P}\big(\bigcup_{n\in\mathbb{N}} A_n\big) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ for any countable family $(A_n)_{n\in\mathbb{N}}$ of disjoint events. From this it is easy[a] to deduce that, for any (not necessarily disjoint) family of events $(A_n)_{n\in\mathbb{N}}$ we have the bound

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) \leqslant \sum_{n\in\mathbb{N}} \mathbb{P}(A_n) . \tag{2.4}$$

An estimate of the form (2.4) is called a *union bound*. Union bounds are ubiquitous in probability, and we shall also use them throughout this course. Roughly, a union bound states that the union of unlikely events remains unlikely provided there are not too many of them.

[a] As a hint, you can consider the family $B_0 = A_0$, $B_1 = A_1 \setminus A_0$, $B_2 = A_2 \setminus (A_0 \cup A_1)$, ….

2.2 Conditional probability

Often one is interested in *conditional statements* statements, where instead of considering the probability of an event A, we are interested in the probability of the event *A knowing that the event B happened*. For instance, suppose I'd like to know the probability that my car breaks down today (event A). If I condition on the event B that I am driving 1000 km today, this will likely influence the answer.

The idea is that we have some extra knowledge in computing the probability of A: we know that B happened. This can change the probability of A dramatically, since now we only consider realisations within B and not in the whole space Ω. In the frequentist interpretation, we count the relative frequency of the occurrence of the event A, but only among those realisations that lie in B.

Definition 2.12 (Conditional probability) Suppose that B is an event satisfying $\mathbb{P}(B) > 0$. Then the conditional probability of an event A given B is

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

This is clearly the correct definition in light of the intuition above: we only consider the probabilities of realisations in B, and we normalize by $\mathbb{P}(B)$ to ensure that the following holds (check this carefully if you're not sure).

Remark 2.13 $\mathbb{P}(\cdot \mid B)$ is a probability measure for any event B satisfying $\mathbb{P}(B) > 0$.

Remark 2.14 Theorem 2.12 only makes sense if $\mathbb{P}(B) > 0$. For brevity, we shall usually omit the explicit mention of this condition, with the general convention that any statement involving the conditional probability $\mathbb{P}(A \mid B)$ is only valid provided that $\mathbb{P}(B) > 0$.
Moreover, we adopt the convention that

$$\mathbb{P}(A \mid B)\,\mathbb{P}(B) := 0 \quad \text{if } \mathbb{P}(B) = 0 .$$

(Recall also the convention $0 \cdot \infty = 0$ from Chapter 1.)

Example 2.15 Consider two throws of a balanced die from Theorem 2.8. Knowing that the sum of the throws is 4, what is the probability that on the first throw I obtained 2? Here,

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

We find

$$A \cap B = \{(2,2)\}$$

and hence

$$\mathbb{P}(A \cap B) = \frac{1}{36} , \qquad \mathbb{P}(B) = \frac{3}{36} .$$

We conclude that

$$\mathbb{P}(A \mid B) = \frac{1}{3} ,$$

which is different from

$$\mathbb{P}(A) = \frac{1}{6}\,.$$

Intuitively, this is not surprising: if we know that the sum of the throws was small, this should increase the odds that the first throw was a small number. Similarly, knowing that the sum of the throws is 4, the probability that on the first throw I obtained 4 (or more) is clearly zero.

Example 2.16 Consider the following two simple questions.

(1) I have two children, one of whom is a girl. What is the probability that the other one is also a girl?

(2) I have two children, the oldest of whom is a girl. What is the probability that the other one is also a girl?

To address them, for the purposes of this mathematical exercise, we make the simplifying assumption that children are either *boys* (B) or *girls* (G), each born with probability $1/2$. Thus, the probability space is

$$\Omega = \{(\mathrm{B},\mathrm{B}), (\mathrm{B},\mathrm{G}), (\mathrm{G},\mathrm{B}), (\mathrm{G},\mathrm{G})\}\,,$$

and each of the four realisations occurs with probability $1/4$.
For question (1), we have

$$A = \{(\mathrm{G},\mathrm{G})\}\,, \qquad B = \{(\mathrm{B},\mathrm{G}), (\mathrm{G},\mathrm{B}), (\mathrm{G},\mathrm{G})\}\,,$$

and therefore

$$\mathbb{P}(A \mid B) = \frac{1/4}{3/4} = \frac{1}{3}\,.$$

For question (2), we have

$$A = \{(\mathrm{G},\mathrm{G})\}\,, \qquad B = \{(\mathrm{B},\mathrm{G}), (\mathrm{G},\mathrm{G})\}\,,$$

and therefore

$$\mathbb{P}(A \mid B) = \frac{1/4}{2/4} = \frac{1}{2}\,.$$

More subtle apparent paradoxes easily arise from a careless use of conditional probabilities. For a famous, and a famously confusing and much debated, example, you can look up the Monty Hall problem online, e.g. on Wikipedia (we will not go into it here).

The following result is mathematically trivial, but it has profound consequences in statistics and the sciences.

Proposition 2.17 (Bayes' theorem) *Let A and B be events satisfying $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}\,.$$

Suppose that $\Omega = B_1 \cup \cdots \cup B_n$ with disjoint events $B_1, \ldots, B_n$; such events are called a partition *of Ω. Then the denominator can be expressed as*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid B_i)\,\mathbb{P}(B_i)\,.$$

Many mistakes in science, and popular reporting of science, arise from fallacies related to a misunderstanding or a misuse of conditional probabilities. A surprisingly common

mistake is to mix up $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$. One issue is that our intuition is bad at estimating conditional probabilities, which is why having a clear and rigorous formulation of the concept is so crucial.

Example 2.18 Here is a classical application of Bayes' theorem in medicine. A patient is tested for a disease. Suppose that

- in 1 % of cases the test is positive even though the patient is healthy;
- in 2 % of cases the test is negative even though the patient is sick.

We are interested in the questions

(1) If a patient tests positive, what is the probability that he is healthy?
(2) If a patient tests negative, what is the probability that he is sick?

It turns out that the answer depends greatly on the prevalence of the virus. Let us suppose that one in a thousand patients is sick.
This is where clear mathematical thinking and Bayes' theorem come in very handy. Introduce the events

$$\begin{aligned} S &= \{\text{patient is sick}\} \\ H &= \{\text{patient is healthy}\} = S^c \\ P &= \{\text{test is positive}\} \\ N &= \{\text{test is negative}\} = P^c \,. \end{aligned}$$

We know that

$$\mathbb{P}(P \mid H) = 0.01\,, \qquad \mathbb{P}(N \mid S) = 0.02\,, \qquad \mathbb{P}(S) = 0.001\,.$$

By Bayes' theorem, the answer to question (1) is

$$\begin{aligned} \mathbb{P}(H \mid P) &= \frac{\mathbb{P}(P \mid H)\mathbb{P}(H)}{\mathbb{P}(P)} \\ &= \frac{\mathbb{P}(P \mid H)\mathbb{P}(H)}{\mathbb{P}(P \mid S)\mathbb{P}(S) + \mathbb{P}(P \mid H)\mathbb{P}(H)} \\ &= \frac{0.01 \cdot (1 - 0.001)}{(1 - 0.02) \cdot 0.001 + 0.01 \cdot (1 - 0.001)} \approx 91\% \,. \end{aligned}$$

Thus, even though the patient was tested positive, the probability that he is healthy is more than 90%. This figure is perhaps higher than one would intuitively expect, and shows, first, the danger of relying on our intuition for questions of this type and, second, the usefulness of clear thinking combined with simple mathematics. A similar calculation gives the answer to question (2) as

$$\mathbb{P}(S \mid N) \approx 0.002\% \,.$$

Thus, a negative test is a very reliable sign of being healthy.
The assumption of one in a thousand patients being sick was crucial in the above calculations. If instead we consider a different population of patients, where the virus is far more prevalent, the conditional probabilities computed to answers questions (1) and (2) will change considerably.

2.3 Random variables

Informally, a random variable is a variable whose value depends on the realisation $\omega \in \Omega$.

Definition 2.19 A *random variable* is a measurable real-valued function on Ω. More generally, for a measurable space $(E, \mathcal{E})$, a *random variable* with values in E is a measurable function from Ω to E.

For instance we can speak about vector-valued random variables, with values in $E = \mathbb{R}^d$.

Example 2.20 (Theorem 2.8 continued) The sum of both values is the random variable $X : \Omega \to \mathbb{R}$ defined by

$$X((i,j)) := i + j\,,$$

with the notation $\omega = (i,j) \in \{1, 2, \dots, 6\}^2$.

Example 2.21 (Theorem 2.9 continued) Define the random variable $X : \Omega \to \mathbb{N}^* \cup \{\infty\}$ to be the number of throws required to obtain a 6 for the first time, i.e.

$$X(\omega) := \inf\{k : \omega_k = 6\}$$

with the convention that $\inf \emptyset = \infty$ (which happens if I never throw a 6).
To see that X is indeed a random variable, we have to check that it is measurable. To that end, we have to check that, for any $n \in \mathbb{N}^*$, the set $X^{-1}(\{n\})$ is a cylinder set of the form (2.1). Indeed,

$$X^{-1}(\{n\}) = \left\{\omega \in \Omega : \omega_1 \neq 6, \omega_2 \neq 6, \dots, \omega_{n-1} \neq 6, \omega_n = 6\right\},$$

as desired. Intuitively, that X is a random variable is clear since the event "X equals n" clearly depends only on the first n throws, and $\mathcal{A}$ is constructed precisely so that such events are measurable.

Definition 2.22 The *law* of a random variable with values in E is the measure

$$\mathbb{P}_X := X_* \mathbb{P}$$

on $(E, \mathcal{E})$. (Recall Theorem 1.8.)

We sometimes use the equivalence relation $\overset{\mathrm{d}}{=}$ on random variables, i.e. equality in law, defined by

$$X \overset{\mathrm{d}}{=} Y \quad \Longleftrightarrow \quad \mathbb{P}_X = \mathbb{P}_Y\,. \tag{2.5}$$

Clearly, $\mathbb{P}_X$ is a probability measure on $(E, \mathcal{E})$. Hence, any random variable X with values in E gives rise to a new probability space $(E, \mathcal{E}, \mathbb{P}_X)$. The intuition is that this space is in general smaller than the original space, and it contains only information captured by the random variable X. If all we care about is the value of X, we can completely forget the original probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and only work on the smaller space $(E, \mathcal{E}, \mathbb{P}_X)$, which is often much simpler.

For instance, in Theorem 2.8, if we only care about the value of $X = i + j$ (and not, say, which of the two throws produced the larger value), we can work on the space $E = X(\Omega) = \{2, 3, \dots, 12\}$ instead of on the original larger space $\Omega = \{1, 2, \dots, 6\}^2$. You

can easily check that the probability measure $\mathbb{P}_X$ on E is given by

$$\mathbb{P}_X(\{k\}) = \frac{(k-1) \wedge (13-k)}{36}\,.$$

In general, for any $B \in \mathcal{E}$, we have

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}) =: \mathbb{P}(X \in B)\,,$$

where the notation on the right-hand side is being defined by this equation. This quantity is the probability that X lies in B.

Probability theory uses its own shorthand notation for events and probabilities determined by a random variable X:

$$\{\omega : X(\omega) \in B\} \equiv \{X \in B\}\,,$$
$$\mathbb{P}\big(\{\omega : X(\omega) \in B\}\big) \equiv \mathbb{P}(X \in B)\,.$$

In addition, inside $\mathbb{P}$, intersection of events is often denoted with a comma instead of the symbol $\cap$. For instance, we write

$$\mathbb{P}(\{X \in A\} \cap \{Y \in B\}) \equiv \mathbb{P}(X \in A, Y \in B)\,. \tag{2.6}$$

We shall always use these shorthand notations.

Before looking at some examples, let us record the following rather banal remark, which is sometimes good to keep in mind. For a given probability measure μ on a measurable space $(E, \mathcal{E})$, can we construct a random variable X with law $\mathbb{P}_X = \mu$? Obviously yes, just by setting $(\Omega, \mathcal{A}, \mathbb{P}) = (E, \mathcal{E}, \mu)$ and $X(\omega) = \omega$.

2.3.1 Elementary special cases

Let us now review some special cases of random variables, some of which you may already have seen in school.

Let X be a random variable with values in $(E, \mathcal{E})$.

- *Discrete random variables.* Here E is finite or countable, and $\mathcal{E} = \mathcal{P}(E)$. In that case,

$$\mathbb{P}_X = \sum_{x \in E} p_x\, \delta_x\,, \tag{2.7}$$

  where $p_x := \mathbb{P}(X = x)$ and δ_x is the delta measure from Theorem 1.5. To verify (2.7), we write, for any $B \in \mathcal{E}$,

$$\begin{aligned}\mathbb{P}_X(B) = \mathbb{P}(X \in B) &= \mathbb{P}\left(\bigcup_{x \in B} \{X = x\}\right)\\ &= \sum_{x \in B} \mathbb{P}(X = x) = \sum_{x \in E} p_x \delta_x(B)\,,\end{aligned}$$

  where in the third step we used crucially the σ-additivity of measures, since E is at most countable by assumption.

Example 2.23 (Theorem 2.9 continued) For $n \in \mathbb{N}^*$ let us compute the probablility that we first obtain a 6 on the nth throw,

$$\begin{aligned}
\mathbb{P}(X = n) &= \mathbb{P}(\omega_1 \neq 6, \dots, \omega_{n-1} \neq 6, \omega_n = 6) \\
&= \mathbb{P}\left(\bigcup_{i_1,\dots,i_{n-1}=1}^{5} \{\omega_1 = i_1, \dots, \omega_{n-1} = i_{n-1}, \omega_n = 6\} \right) \\
&= \sum_{i_1,\dots,i_{n-1}=1}^{5} \mathbb{P}\big(\omega_1 = i_1, \dots, \omega_{n-1} = i_{n-1}, \omega_n = 6\big) \\
&= 5^{n-1} \left(\frac{1}{6}\right)^n \\
&= \frac{1}{6}\left(\frac{5}{6}\right)^{n-1} .
\end{aligned}$$

This computation shows the power of a clear and rigorous formulation in solving very concrete problems. In particular, we find that the probability that we never throw a 6 is

$$\mathbb{P}(X = \infty) = 1 - \mathbb{P}(X < \infty) = 1 - \sum_{n \in \mathbb{N}^*} P(X = n) = 1 - 1 = 0 .$$

Nevertheless, the event $\{X = \infty\} = \{\omega \in \Omega : \omega_k < 6 \text{ for all } k \in \mathbb{N}^*\}$ is enourmous, in particular uncountable.

- *Continuous random variables.* Let $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and suppose that $\mathbb{P}_X$ is *absolutely continuous*[1] with respect to Lebesgue measure. This means that there exists a measurable function $p \colon \mathbb{R}^d \to [0, \infty)$ such that

$$\mathbb{P}_X(B) = \int_B p(x)\, \mathrm{d}x ,$$

where $\mathrm{d}x$ denotes Lebesgue measure on $\mathbb{R}^d$. The function p is called the *density of the law of* X, sometimes just *density of* X.

[1] As you may know from measure theory, the notion of absolute continuity that we use here is usually a consequence of the general definition, by the so-called Radon-Nikodym theorem. For our purposes, however, the above definition is sufficient.

2.4 Expectation

WEEK 2

Let X be a random variable with values in $\mathbb{R}$. We would like to determine a number that represents the typical, or mean, value of X. For example, consider a random variable X that is equal to a with probability p and to b with probability $1-p$. In other words,

$$\mathbb{P}_X = p\delta_a + (1-p)\delta_b\,.$$

In school we learn that the mean value of X is

$$pa + (1-p)b = \int_R x\,\mathbb{P}_X(\mathrm{d}x) = \int_\Omega X(\omega)\,\mathbb{P}(\mathrm{d}\omega)\,,$$

where the second identity follows by definition of $\mathbb{P}_X$ (Theorem 2.22). Note that which probability space X is defined on does not matter.

Definition 2.24 Let X be a random variable with values in $\mathbb{R}$. The *expectation* of X is

$$\mathbb{E}[X] := \int X(\omega)\,\mathbb{P}(\mathrm{d}\omega)\,,$$

which is well-defined if $X \geqslant 0$ (in which case $\mathbb{E}[X] \in [0,\infty]$) or if $\mathbb{E}[|X|] < \infty$ (in which case X is called *integrable*).
If $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ has values in $\mathbb{R}^d$, then we define

$$\mathbb{E}[X] := (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d])\,.$$

The following result states how to compute the expectation of a function of a random variable.

Proposition 2.25 *Let X be a random variable with values in $(E,\mathcal{E})$, and let $f : E \to \mathbb{R}\cup\{\infty\}$ be measurable. Then*

$$\mathbb{E}[f(X)] = \int f\,\mathrm{d}\mathbb{P}_X$$

provided that $f \geqslant 0$ or $\mathbb{E}[|f(X)|] < \infty$.

Proof The proof is an archetypal argument from measure theory, which we recall here. See also the discussion after Theorem 1.9. Consider first the case where f is an indicator function, $f = \mathbf{1}_B$ for $B \in \mathcal{E}$. Then

$$\mathbb{E}[f(X)] = \mathbb{E}[\mathbf{1}_B(X)] = \mathbb{P}(X \in B) = \mathbb{P}_X(B) = \int_B \mathrm{d}\mathbb{P}_X = \int \mathbf{1}_B\,\mathrm{d}\mathbb{P}_X = \int f\,\mathrm{d}\mathbb{P}_X\,,$$

as desired.

Moreover, by linearity of the integrals over $\mathbb{P}$ in $\mathbb{E}[\cdot]$ as well as over $\mathbb{P}_X$, we deduce that the claim holds for any simple function f.

Next, let f be an arbitrary nonnegative measurable function. Choose a sequence of simple functions f_n that converge to f monotonically from below. Then by using the claim for simple functions, as well as the monotone convergence theorem twice (since (f_n) and $(f_n(X))$ are pointwise nondecreasing sequences on E and Ω, respectively), we obtain

$$\mathbb{E}[f(X)] = \lim_{n\to\infty} \mathbb{E}[f_n(X)] = \lim_{n\to\infty} \int f_n\,\mathrm{d}\mathbb{P}_X = \int f\,\mathrm{d}\mathbb{P}_X\,.$$

This concludes the proof for $f \geqslant 0$. If f is a general integrable function, we split $f = f_+ - f_-$ into its positive and negative parts, and apply the result for positive f to f_+ and f_- separately. □

Remark 2.26 The proof used a trick that is so trivial that it may go unnoticed: the probability of an event A can be expressed as the expectation of its indicator function,

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

This trick, as simple as it seems, is extremely useful and consequenty ubiquitous in probability. We shall use it repeatedly in this class.

Example 2.27 (Theorem 2.8 continued) Let us compute the expectation of the sum of two throws of a die:

$$\mathbb{E}[X] = \frac{1}{36} \sum_{i,j=1}^{6} (i+j) = \frac{1}{36} \left(6 \sum_{i=1}^{6} i + 6 \sum_{j=1}^{6} j \right) = 7 .$$

Example 2.28 (Theorem 2.9 continued) Let us compute the expected number of throws until we obtain a 6:

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} n \, \mathbb{P}(X = n) = \frac{1}{6} \sum_{n=1}^{\infty} n \left(\frac{5}{6} \right)^{n-1} = \frac{1}{6} \frac{1}{(1/6)^2} = 6 ,$$

where we used the geometric series identity $\sum_{n=1}^{\infty} n x^{n-1} = \frac{1}{(1-x)^2}$, which is proved by differentiating in x.

We conclude this section by introducing the notion of *conditional expectation*.

Definition 2.29 Let B and event satisfying $\mathbb{P}(B) > 0$. Let X be a random variable. The *conditional expectation of X given B* is

$$\mathbb{E}[X \mid B] := \frac{\mathbb{E}[X \mathbf{1}_B]}{\mathbb{P}(B)} ,$$

i.e. the expectation of X with respect to the probability measure $\mathbb{P}(\cdot \mid B)$.

In particular, for any events A and B we clearly have

$$\mathbb{E}[\mathbf{1}_A \mid B] = \mathbb{P}(A \mid B) .$$

Example 2.30 (Theorem 2.8 continued) Let us compute the expectation of the sum of two throws of a die, given that each number is even. The event B on which we condition is

$$B = \{2, 4, 6\}^2 .$$

We find $\mathbb{P}(B) = \frac{1}{4}$ and

$$\mathbb{E}[X \mathbf{1}_B] = \frac{1}{36} \sum_{i,j \in \{2,4,6\}} (i+j) = \frac{1}{36} \left(3 \sum_{i \in \{2,4,6\}} i + 3 \sum_{j \in \{2,4,6\}} j \right) = 2 .$$

Hence,

$$\mathbb{E}[X \mid B] = \frac{2}{1/4} = 8 .$$

This is larger than the unconditional expectation from Theorem 2.27, which is hardly surprising: conditioning on a number from 1 to 6 being even makes it on average larger.

2.5 The classical laws

In this section we go to the zoo. We encounter the most common and useful laws and learn their names, along with any parameters they depend on. We leave it as a simple exercise to check that each of them is indeed a probability measure (i.e. that the total measure equals one).

Discrete laws

The following laws are defined on $(E, \mathcal{E})$ with E finite or countable and $\mathcal{E} = \mathcal{P}(E)$.

- *Uniform.* Let E be a finite set and define $\mathbb{P}(X = x) = \frac{1}{\#E}$ for all $x \in E$.
- *Bernoulli ($p \in [0,1]$).* Let $E = \{0, 1\}$ and set $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. (This law models a coin toss that is biased for $p \neq \frac{1}{2}$.)
- *Binomial ($n \in \mathbb{N}^*$, $p \in [0,1]$).* Let $E = \{0, 1, \dots, n\}$ and
$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$
for $k \in E$. Here $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. (This law models the number of heads when tossing a biased coin n times; see the exercises.)
- *Geometric ($p \in (0,1)$).* Let $E = \mathbb{N}$ and
$$\mathbb{P}(X = k) = (1-p)p^k$$
for all $k \in \mathbb{N}$. (This law models the number of heads before the first tails when tossing a biased coin. See also Theorem 2.9 and its continuation in Section 2.3.1.) Note that the different convention $\mathbb{P}(X = k) = (1-p)^{k-1}p$ for $k \in E = \mathbb{N}^*$ is also often used in the literature.
- *Poisson ($\lambda > 0$).* Let $E = \mathbb{N}$ and
$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}$$
for all $k \in \mathbb{N}$. (This law models the number of rare events observed in a long time interval. More precisely, if X_n has binomial law with parameters n and p_n satisfying $np_n \to \lambda$ as $n \to \infty$, then $\mathbb{P}(X_n = k) \to \mathbb{P}(X = k)$ for all k. See the exercises.)

Continuous laws

The following laws are defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. They are continuous, and therefore determined by their densities $p(x)$.

- *Uniform on $[a, b]$.* Let
$$p(x) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(x)\,.$$
- *Exponential ($\lambda > 0$).* Let
$$p(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{[0,\infty)}(x)\,.$$

- *Gaussian or normal ($m \in \mathbb{R}$, $\sigma > 0$).* Let
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} .$$
(m is called the mean and σ the standard deviation.)

2.6 Cumulative distribution function

The law of a real-valued random variable can be fully and equivalently characterised by a function on $\mathbb{R}$.

Definition 2.31 Let X be a random variable with values in $\mathbb{R}$. We define its *cumulative distribution function* $F_X : \mathbb{R} \to [0,1]$ by
$$F_X(t) := \mathbb{P}(X \leqslant t) .$$
For brevity, sometimes we simply speak of the *distribution function* of X.

Proposition 2.32 *If $F = F_X$ is the distribution function of a random variable X, then*
(i) F is nondecreasing;
(ii) $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$;
(iii) F is right-continuous, i.e. $\lim_{s\downarrow t} F(s) = F(t)$ for all $t \in \mathbb{R}$.

Proof The proof of (i) is obvious.

Let us prove (ii). It uses some basic facts from measure theory, which are reviewed in Exercise 1.1. Since the events $\{X \leqslant n\}$ are increasing, we find
$$\lim_{t\to\infty} F(t) = \lim_{n\to\infty} \mathbb{P}(X \leqslant n) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \{X \leqslant n\}\right) = \mathbb{P}(\Omega) = 1 ,$$
where the second step follows by σ-additivity of $\mathbb{P}$ and the last step from $\bigcup_{n\in\mathbb{N}}\{X \leqslant n\} = \Omega$ since $X \in \mathbb{R}$. Analogously, since the events $\{X \leqslant -n\}$ are decreasing, we find
$$\lim_{t\to-\infty} F(t) = \lim_{n\to\infty} \mathbb{P}(X \leqslant -n) = \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \{X \leqslant -n\}\right) = \mathbb{P}(\emptyset) = 0 ,$$
where the second step follows by σ-additivity of $\mathbb{P}$ and the last step from $\bigcap_{n\in\mathbb{N}}\{X \leqslant -n\} = \emptyset$ since $X \in \mathbb{R}$.

To prove (iii), let (a_n) be a strictly decreasing sequence tending to 0. Since the events $\{X \leqslant t + a_n\}$ are decreasing, we find
$$\lim_{n\to\infty} F(t + a_n) = \lim_{n\to\infty} \mathbb{P}(X \leqslant t + a_n) = \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \{X \leqslant t + a_n\}\right) = \mathbb{P}(X \leqslant t) = F(t) ,$$
where the second step follows by σ-additivity of $\mathbb{P}$ and the last step from $\bigcap_{n\in\mathbb{N}}\{X \leqslant t + a_n\} = \{X \leqslant t\}$. □

As a helpful check to see if you understood the proof of (iii), try to see what goes wrong if you try to prove that F is left-continuous, which is in general wrong.

Note that the F is discontinuous whenever $\mathbb{P}_X$ has an atom. For instance, if X is equal to a constant a, then $\mathbb{P}_X = \delta_a$ and hence $F_X(t) = \mathbf{1}_{t\geqslant a}$.

Conversely, it is natural to ask whether any function F satisfying the three properties (i), (ii), (iii) is the distribution function of some random variable. As the following proposition shows, the answer is yes!

This very important result is not just a theoretical curiosity; it is extremely useful. Indeed, the proof relies on an explicit construction that is of great use in both theoretical probability and applications. It is the standard algorithm for generating random variables with any given distribution. See Theorem 2.34 below.

Proposition 2.33 *If $F : \mathbb{R} \to [0,1]$ satisfies (i), (ii), (iii) from Theorem 2.32 then there exists a random variable X with values in $\mathbb{R}$ such that $F = F_X$.*

Proof To construct X, we have to start by constructing the probability space. We take $\Omega = (0,1)$, $\mathcal{A} = \mathcal{B}((0,1))$, and $\mathbb{P}$ to be Lebesgue measure. Then we define

$$X(\omega) := \sup\{s \in \mathbb{R} : F(s) < \omega\}\,.$$

We note that $X(\omega) \in \mathbb{R}$ for any $\omega \in (0,1)$ by the assumption (ii). The rest of the proof consists in showing that this explicit choice satisfies $F = F_X$.

To that end, we shall show that for all $t \in \mathbb{R}$

$$\{\omega \in \Omega : X(\omega) \leqslant t\} = \{\omega \in \Omega : \omega \leqslant F(t)\}\,. \tag{2.8}$$

Supposing that (2.8) is true, we obtain

$$F_X(t) = \mathbb{P}(X \leqslant t) = \mathbb{P}\big(\{\omega \in \Omega : \omega \leqslant F(t)\}\big) = F(t)$$

by definition of Lebesgue measure, as desired.

Hence, what remains is to show (2.8), which follows from the two following observations.

- Suppose that $\omega \leqslant F(t)$. Then, by definition of X and of sup, we immediately deduce that $t \geqslant X(\omega)$.
- Suppose that $\omega > F(t)$. Since F is right-continuous, there exists $\varepsilon > 0$ such that $F(t+\varepsilon) < \omega$. Thus, again by definition of X, we deduce that $X(\omega) \geqslant t + \varepsilon$. Hence, $X(\omega) > t$. □

As a helpful check to see that you understood the proof, you can try to spot exactly where we used each of the assumptions (i), (ii), (iii) on F.

Remark 2.34 The function $X : (0,1) \to \mathbb{R}$ constructed in the above proof is often called *inverse* of F, and denoted by F^{-1}, even if F is not bijective. (The function F can be locally constant.) If F is bijective, then F^{-1} coincides with the usual inverse. The construction of the proof is remarkable. It provides an algorithm for generating a random variable X with distribution function F starting from a random variable Y which is uniformly distributed on $(0,1)$: simply set $X := F^{-1}(Y)$. Thus, the problem is reduced to the generation of the special and simple random variable Y.
To be more concrete, suppose that we have a computer that generates a random floating point number ω that is uniformly distributed in $(0,1)$. (We shall see later that such a generator can be easily constructed by taking the binary digits of ω to be independent Bernoulli random variables with parameter $p = 1/2$. Thus, all that we need is a random number generator that generates such Bernoulli random variables.)
We wish to generate a standard Gaussian random variable, with parameters $m = 0$

and $\sigma = 1$. To that end, we define the function

$$F(t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \mathrm{e}^{-\frac{x^2}{2}} \, \mathrm{d}x \, .$$

The function F is famously not an elementary function, but it can be easily computed numerically, and most software packages have an implementation of (a version of) it, often called Erf. Then the desired Gaussian random variable is $X := F^{-1}(\omega)$, where the right-hand side is again evaluated numerically.

2.7 The σ-algebra generated by a random variable

Every random variable X gives naturally rise to a σ-algebra, which is the smallest (i.e. coarsest) σ-algebra on Ω with respect to which X is measurable. To build intuition, consider the case where X is a random variable with values in $\{1, 2, 3\}$, and define the events $A_i := X^{-1}(\{i\})$ for $i = 1, 2, 3$. Then X is measurable with respect to the σ-algebra

$$\begin{aligned} \mathcal{B} &:= \{\emptyset, A_1, A_2, A_3, A_1 \cup A_2, A_1 \cup A_3, A_2 \cup A_3, \Omega\} \\ &= \{X^{-1}(\emptyset), X^{-1}(\{1\}), X^{-1}(\{2\}), X^{-1}(\{3\}), \\ &\qquad X^{-1}(\{1,2\}), X^{-1}(\{1,3\}), X^{-1}(\{2,3\}), X^{-1}(\{1,2,3\})\} \, . \end{aligned}$$

You can convince yourself that this is the smallest σ-algebra with respect to which X is measurable.

In some sense, $\mathcal{B}$ captures the *resolving power of* X, but it does not contain the full information about X. For instance, the random variable $Y = 2X$ generates the same σ-algebra as X, but it is clearly different from X. However, both X and Y have the same ability to resolve the probability space Ω. This is the basic intuition behind the following definition.

Definition 2.35 Let X be a random variable with values in a measurable space $(E, \mathcal{E})$. Then the *σ-algebra generated by X* is

$$\sigma(X) := \{X^{-1}(B) : B \in \mathcal{E}\} \, .$$

Note that, as advertised above, this is the smallest σ-algebra with respect to which X is measurable. Indeed, clearly any such σ-algebra will have to contain all sets of the form $X^{-1}(B)$ for $B \in \mathcal{E}$; moreover, the set $\sigma(X)$ is a σ-algebra.

2.8 Moments and inequalities

Definition 2.36 Let X be a random variable with values in $\mathbb{R}$ and $p \geqslant 1$. The *p-th moment of X* is $\mathbb{E}[X^p]$, which is well-defined under either of the following conditions:

- $p \in \mathbb{N}^*$ and $\mathbb{E}[|X|^p] < \infty$;
- $X \geqslant 0$.

In probability, we say that some property $P(\omega)$ depending on the realisation ω holds *almost surely* instead of almost everywhere (as in measure theory) if $\mathbb{P}(P \text{ true}) = 1$. We often abbreviate a.s. for almost surely.

We use the following definition from measure theory (see the course *Calculus II*).

Definition 2.37 For $p \in [1, \infty]$, we denote by $L^p \equiv L^p(\Omega, \mathcal{A}, \mathbb{P})$ the usual L^p-space with norm denoted by $\|X\|_p$.

It might be helpful to do a quick review of measure theory to recall how these spaces are defined. As in measure theory, there is a technical annoyance, which arises from the need to identify random variables that are almost surely equal.

- For $p \in [1, \infty)$ we denote by $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ the set of real-valued random variables X satisfying $\mathbb{E}[|X|^p] < \infty$.
- We denote by $\mathcal{L}^\infty(\Omega, \mathcal{A}, \mathbb{P})$ the set of real-valued random variables X such that there exists a constant C satisfying $|X| \leqslant C$ almost surely.
- For $p \in [1, \infty]$ we define the equivalence relation $\sim$ on $\mathcal{L}^p$ by setting $X \sim Y$ if and only if $X = Y$ almost surely.
- For $p \in [1, \infty]$ we define the quotient space

$$L^p(\Omega, \mathcal{A}, \mathbb{P}) := \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P}) / \sim \ .$$

Thus an element of L^p is an *equivalence class of random variables*. Throughout the following, and in accordance with the literature, we usually skirt around this issue by abusing notation and identifying an element of L^p with a representative of its class. This convention is consistent provided that all operations performed on such representatives do not depend on the choice of the representative within its class. It is always good to keep the precise definition in mind, as this subtlety is sometimes important.

- For $p \in [1, \infty)$ and $X \in L^p$ we write

$$\|X\|_p := \left(\mathbb{E}[|X|^p]\right)^{1/p}.$$

  Note that this definition makes sense, since it is independent of the representative of X.
- We write

$$\|X\|_\infty := \inf\{C \geqslant 0 : |X| \leqslant C \text{ a.s.}\}.$$

  This number is sometimes called the essential supremum of $|X|$. It is independent of the representative of X (unlike $\sup|X|$).

The following result[2] was proved in the course *Calculus II*.

Proposition 2.38 *For each $p \in [1, \infty]$, the space $L^p(\Omega, \mathcal{A}, \mathbb{P})$ is a Banach space.*

The following inequality is the most important inequality in all of analysis.

Proposition 2.39 (Hölder's inequality) *Let $p, q \in [1, \infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$ (with the convention $\frac{1}{\infty} = 0$). Then for any random variables X, Y we have*

$$\|XY\|_1 \leqslant \|X\|_p \|Y\|_q .$$

[2] Note that this result is one place where taking the quotient in the definition of L^p is essential; it is wrong for the space $\mathcal{L}^p$.

Note that Propositions 2.38 and 2.39 are true for any measure space, not just for a probability space.

Let us list some obvious but important special cases of Hölder's inequality:

(i) $\|X\|_p \leqslant \|X\|_q$ if $1 \leqslant p \leqslant q$.
(ii) $\mathbb{E}[|XY|] \leqslant \|X\|_2 \|Y\|_2$ (Cauchy-Schwarz inequality).
(iii) $\mathbb{E}[|X|]^2 \leqslant \mathbb{E}[X^2]$.

Note that (ii) is true for any measure space, while (i) and (iii) are only true for a probability space.

Definition 2.40 Let $X \in L^2$. The *variance of* X is

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$$

and its *standard deviation* is $\sigma_X := \sqrt{\mathrm{Var}(X)}$.

Just as the expectation measures the typical mean value of X, the variance measures the typical spread of X around its mean value. It is important to realise that the variance is not the only quantity to quantify this spread, it is merely the most convenient and the most popular one. For example, another quantity that measures the spread is $\mathbb{E}[|X - \mathbb{E}[X]|]$; as we shall see in the exercises, this quantity has advantages and disadvantages as compared to the variance, and it is sometimes used in statistics where it is closely related to the *median* of X (see the exercises).

Remark 2.41 The following observations follow immediately from the definition of the variance.

(i) $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
(ii) For all $a \in \mathbb{R}$ we have $\mathbb{E}[(X-a)^2] = \mathrm{Var}(X) + (\mathbb{E}[X] - a)^2$, and hence
$$\mathrm{Var}(X) = \inf_{a \in \mathbb{R}} \mathbb{E}[(X-a)^2]$$
This gives another, so-called variational, interpretation of the variance.
(iii) $\mathrm{Var}(X) = 0$ if and only if X is almost surely constant.

Next we state the most important inequality in probability, which is traditionally associated with at least the names of Bienaymé, Chebyshev, and Markov. We shall call it Chebyshev's inequality, as it is also commonly known, for historical reasons that we do not go into here.

Proposition 2.42 (Chebyshev) *Let $f : \mathbb{R} \to [0, \infty)$ be nondecreasing and X a random variable. Then for all $a \in \mathbb{R}$ we have*

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[f(X)]}{f(a)}.$$

Proof Since f is nondecreasing, on the event $X \geqslant a$ we have $f(X) \geqslant f(a)$. Thus,

$$\mathbb{P}(X \geqslant a) = \mathbb{E}[\mathbf{1}_{X \geqslant a}] \leqslant \mathbb{E}\left[\mathbf{1}_{X \geqslant a} \frac{f(X)}{f(a)}\right] \leqslant \mathbb{E}\left[\frac{f(X)}{f(a)}\right],$$

as claimed. □

Here are some important and famous special cases of Chebyshev's inequality:

(i) If $X \geqslant 0$ and $a > 0$ then $\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}$ (often called Markov's inequality).

(ii) If $X \in L^2$ and $a > 0$ then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant a) \leqslant \frac{\mathrm{Var}(X)}{a^2}$$

(often simply called Chebyshev's inequality)

(iii) $\mathbb{P}(X \geqslant a) \leqslant \mathrm{e}^{-ta}\mathbb{E}[\mathrm{e}^{tX}]$ for any $t > 0$ (often called Chernov's inequality). Since this inequality holds for any $t > 0$, one can even take the infimum over t to deduce that $\mathbb{P}(X \geqslant a) \leqslant \mathrm{e}^{-I(a)}$, where

$$I(a) := \sup_{t>0}\{ta - \log \mathbb{E}[\mathrm{e}^{tX}]\}\,.$$

This estimate is often very sharp, and it plays a fundamental role in the so-called theory of large deviations and statistical mechanics, which however goes beyond the scope of this course.

Finally, the notion of variance can be generalised to the *covariance* of several random variables, which roughly measures how strongly they tend to fluctuate jointly.

Definition 2.43 For $X, Y \in L^2$ define the *covariance of X and Y* as

$$\mathrm{Cov}(X, Y) := \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\,.$$

For a random vector $X = (X_1, \dots, X_d)$ with values in $\mathbb{R}^d$ with $X_i \in L^2$ for all $i = 1, \dots, d$, we define the *$d \times d$ covariance matrix*

$$\mathrm{Cov}(X) := (\mathrm{Cov}(X_i, X_j))_{i,j=1}^{d}\,.$$

The covariance matrix of a random vector is one of the most fundamental objects of study in high-dimensional statistics and machine learning. We shall discuss some of its properties in the exercises.

CHAPTER 3

Independence

WEEK 3

This chapter is devoted to independence, which is a fundamentally probabilistic notion. Although the basic idea behind independence is very simple, a precise statement general enough for later applications requires some care.

3.1 Independent events

Independence is classically stated on the level of events. In the real world, two events are typically independent if they describe events that are causally unrelated. For instance, if I flip a coin twice, whether I geat heads the first and the second time are independent events. The mathematical definition of course goes beyond any causal or mechanical interpretations in the real world. Informally, A and B being independent means that knowing that B happened gives no information about the probability of A happening. More formally, the *conditional probability* (see Theorem 2.12) $\mathbb{P}(A \mid B)$ is equal to $\mathbb{P}(A)$. This leads to the following definition.

Definition 3.1 Two events $A, B \in \mathcal{A}$ are *independent* if

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

Example 3.2

(i) (Theorem 2.8 continued.) When throwing a die twice, obtaining a six on the first throw and obtaining a six on the second throw are independent events. More precisely, setting

$$A = \{6\} \times \{1, \dots, 6\} , \qquad B := \{1, \dots, 6\} \times \{6\} ,$$

we find $\mathbb{P}(A \cap B) = \frac{1}{36} = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

(ii) When throwing a single die, the events

$$A = \{1, 2\} , \qquad B = \{1, 3, 5\}$$

are independent: $\mathbb{P}(A \cap B) = \frac{1}{6} = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

The notion of independence extends from two events to any, possibly infinite, collection of events.

Definition 3.3 A collection of events $\{A_i\}_{i \in I}$ is *independent* if for any finite $J \subset I$ we have

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i) .$$

Remark 3.4 Even if I is finite, for the collection $\{A_i\}_{i\in I}$ to be independent, it is not sufficient that $\mathbb{P}(\bigcap_{i\in I} A_i) = \prod_{i\in I} \mathbb{P}(A_i)$.
Moreover, for the collection $\{A_i\}_{i\in I}$ to be independent, it is not sufficient that all pairs A_i and A_j be independent (pairwise independence).
To see this, consider flipping an unbiased coin twice, so that $\Omega = \{0,1\}^2$ with the uniform probability measure. Define the events

$$A = \{1\} \times \{0,1\}\,, \qquad B = \{0,1\} \times \{1\}\,, \qquad C = \{0\} \times \{0\} \cup \{1\} \times \{1\}\,.$$

(What is their interpretation?) Then we have

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}\,,$$
$$\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \frac{1}{4}\,,$$
$$\mathbb{P}(A \cap B \cap C) = \frac{1}{4}\,.$$

We conclude that they are not independent, although they are pairwise independent.

3.2 Intermezzo: monotone class lemma*

In order to extend the notion of independence to random variables, a notion that plays a fundamental role in probability, we shall need a powerful tool from measure theory: the monotone class lemma. It is usually not covered in a course in measure theory. Thus, in this section we perform a measure-theoretic excursion. The section is marked with an asterisk, which means that it does not belong to the core material of the course; in particular, if you wish you can skip over the proofs, which will also not be asked in the exam. All that you have to know from this section is Theorem 3.7 and Theorem 3.9.

Let E be a set.

Definition 3.5 A collection $\mathcal{M} \subset \mathcal{P}(E)$ is a *monotone class* if

(i) $E \in \mathcal{M}$;
(ii) If $A, B \in \mathcal{M}$ satisfy $A \subset B$ then $B \setminus A \in \mathcal{M}$;
(iii) If $A_n \in \mathcal{M}$ and $A_n \subset A_{n+1}$ for all $n \in \mathbb{N}$ then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{M}$.

The term *monotone class* comes from the last property, which distinguishes it from a σ-algebra, and states that the union of an increasing family of elements of $\mathcal{M}$ is again in $\mathcal{M}$. There is a priori no very clear intuitive interpretation of this definition. Its usefulness will become apparent a posteriori, through its applications; see for instance Theorem 3.9 and Theorem 3.10 below.

Similarly to Theorem 1.2, any collection of subsets of E generates a monotone class.

Definition 3.6 The monotone class generated by a collection $\mathcal{C} \subset \mathcal{P}(E)$ is

$$\mathcal{M}(\mathcal{C}) := \bigcap_{\substack{\mathcal{M} \text{ is a monotone class} \\ \mathcal{C}\subset\mathcal{M}}} \mathcal{M}\,.$$

It is left as an exercise to check that the intersection of monotone classes is a monotone class, so that in particular $\mathcal{M}(\mathcal{C})$ is always a monotone class.

The following result is the main tool proved in this section. To state it, we need the following definition.

Definition 3.7 A collection $\mathcal{C} \subset \mathcal{P}(E)$ is *stable under finite intersections* if for any $A, B \in \mathcal{C}$ we have $A \cap B \in \mathcal{C}$.

Proposition 3.8 (Monotone class lemma) *If $\mathcal{C} \subset \mathcal{P}(E)$ is stable under finite intersections, then $\mathcal{M}(\mathcal{C}) = \sigma(\mathcal{C})$.*

Proof Note first that a σ-algebra is a monotone class (this is left as an easy exercise). Hence, we trivially have the inclusion $\mathcal{M}(\mathcal{C}) \subset \sigma(\mathcal{C})$.

To prove the reverse inclusion, $\sigma(\mathcal{C}) \subset \mathcal{M}(\mathcal{C})$, it suffices to show that $\mathcal{M}(\mathcal{C})$ is a σ-algebra[1].

We shall proceed in several steps.

Claim. A monotone class $\mathcal{M}$ is a σ-algebra if and only if it is stable under finite intersections.

Clearly, a σ-algebra is a monotone class that is stable under finite intersections. For the reverse implication, suppose that $\mathcal{M}$ is a monotone class that is stable under finite intersections. Then $\mathcal{M}$ is also stable under finite unions, since

$$A, B \in \mathcal{M} \;\Rightarrow\; A^c, B^c \in \mathcal{M} \;\Rightarrow\; A^c \cap B^c \in \mathcal{M} \;\Rightarrow\; A \cup B \in \mathcal{M}\,.$$

Let now $A_1, A_2, \cdots \in \mathcal{M}$ and set $B_n := A_1 \cup \cdots \cup A_n$. Then, by the property we just proved, $B_n \in \mathcal{M}$. Moreover, since $B_n \subset B_{n+1}$, by Theorem 3.5 we conclude that $\bigcup_n A_n = \bigcup_n B_n \in \mathcal{M}$. We have therefore verified Theorem 1.1, and hence proved the Claim.

By the Claim, it suffices to show that $\mathcal{M}(\mathcal{C})$ is stable under finite intersections, i.e.

$$A, B \in \mathcal{M}(\mathcal{C}) \quad\Longrightarrow\quad A \cap B \in \mathcal{M}(\mathcal{C})\,. \tag{3.1}$$

To that end, we first fix $A \in \mathcal{C}$, and define

$$\mathcal{M}_1 := \{B \in \mathcal{M}(\mathcal{C}) : A \cap B \in \mathcal{M}(\mathcal{C})\}\,.$$

Since by assumption $\mathcal{C}$ is stable under finite intersections, we have

$$\mathcal{C} \subset \mathcal{M}_1\,. \tag{3.2}$$

Moreover, we claim that

$$\mathcal{M}_1 \text{ is a monotone class.} \tag{3.3}$$

To verify (3.3), let us verify the three properties (i)–(iii) of Theorem 3.5. Property (i) is trivial. To verify (ii), we take $B, B' \in \mathcal{M}_1$ satisfying $B \subset B'$, and note that

$$A \cap (B' \setminus B) = (A \cap B') \setminus (A \cap B) \in \mathcal{M}(\mathcal{C})\,,$$

where the last step follows from the facts that $\mathcal{M}(\mathcal{C})$ is a monotone class and that $A \cap B'$ and $A \cap B$ are in $\mathcal{M}(\mathcal{C})$ by definition of $\mathcal{M}_1$. This shows that $B' \setminus B \in \mathcal{M}_1$, and hence yields (ii). Finally, to prove (iii), let us take $B_n \in \mathcal{M}_1$ such that $B_n \subset B_{n+1}$. Then

$$A \cap \Big(\bigcup_n B_n\Big) = \bigcup_n (A \cap B_n) \in \mathcal{M}(\mathcal{C})\,,$$

[1] Since in that case $\mathcal{M}(\mathcal{C})$ is a σ-algebra containing $\mathcal{C}$, and hence it contains $\sigma(\mathcal{C})$ by Theorem 1.2

since $A \cap B_n \in \mathcal{M}(\mathcal{C})$ by definition of $\mathcal{M}_1$, and $\mathcal{M}(\mathcal{C})$ is a monotone class. We conclude that $\bigcup_n B_n \in \mathcal{M}_1$. This concludes the proof of property (iii), and hence of (3.3).

Next, from (3.2) and (3.3), we deduce that $\mathcal{M}(\mathcal{C}) \subset \mathcal{M}_1$. This means that

$$\forall A \in \mathcal{C}, \forall B \in \mathcal{M}(\mathcal{C}), A \cap B \in \mathcal{M}(\mathcal{C}) . \tag{3.4}$$

Next, we repeat exactly the same argument with fixed $B \in \mathcal{M}(\mathcal{C})$ and

$$\mathcal{M}_2 := \{A \in \mathcal{M}(\mathcal{C}) : A \cap B \in \mathcal{M}(\mathcal{C})\} .$$

From (3.4) we know that $\mathcal{C} \subset \mathcal{M}_2$.

We may repeat the proof of (3.3) almost to the letter to show that $\mathcal{M}_2$ is a monotone class. Since $\mathcal{C} \subset \mathcal{M}_2$, we conclude that $\mathcal{M}(\mathcal{C}) \subset \mathcal{M}_2$. This immediately implies (3.1), and hence concludes the proof. □

The monotone class lemma may seem rather abstract, but it is very useful in probability. It allows to verify equality of two probability measures μ and ν on a much smaller set $\mathcal{C}$ of events than the full σ-algebra. Typically, verifying the equality $\mu(A) = \nu(A)$ for all $A \in \mathcal{A}$ is practically impossible. However, it is often very easy to construct a simple collection of events $\mathcal{C}$ (for instance intervals, rectangles, or cylinder sets) on which the equality is trivial. The monotone class lemma then allows to deduce equality on all sets $A \in \mathcal{A}$. That is its great power. The following result is a typical application of this idea.

Corollary 3.9 *Let μ and ν be two probability measures on $(\Omega, \mathcal{A})$. If there exists a collection $\mathcal{C} \subset \mathcal{A}$ that is stable under finite intersections such that $\sigma(\mathcal{C}) = \mathcal{A}$ and $\mu(A) = \nu(A)$ for all $A \in \mathcal{C}$, then $\mu = \nu$.*

Proof Let $\mathcal{G} := \{A \in \mathcal{A} : \mu(A) = \nu(A)\}$. Then $\mathcal{C} \subset \mathcal{G}$ and it is easy to check that $\mathcal{G}$ is a monotone class. Moreover, by Theorem 3.8,

$$\mathcal{M}(\mathcal{C}) = \sigma(\mathcal{C}) = \mathcal{A} ,$$

and the claim follows since $\mathcal{M}(\mathcal{C}) \subset \mathcal{G}$. □

We shall use Theorem 3.9 throughout this class. Theorem 3.12 below is a typical application. Here is an immediate application that shows its power in proving a famous and nontrivial result.

Example 3.10 (Uniqueness of Lebesgue measure) There exists at most one probability measure λ on $([0, 1], \mathcal{B}([0, 1]))$ such that $\lambda((a, b]) = b - a$ for all $0 < a < b \leqslant 1$. For the proof, simply invoke Theorem 3.9 with $\mathcal{C} = \{(a, b] : 0 < a < b \leqslant 1\}$, the set of half-open intervals (which is obviously stable under finite intersections).

3.3 Independent σ-algebras and random variables

On the most fundamental, and general, level, independence is formulated for σ-algebras. This notion then naturally extends to random variables through their generated σ-algebras (Theorem 2.35).

Definition 3.11

(i) The σ-algebras $\mathcal{B}_1, \dots, \mathcal{B}_n \subset \mathcal{A}$ are *independent* if for all $A_1 \in \mathcal{B}_1, \dots, A_n \in \mathcal{B}_n$ we have $\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$.

(ii) The random variables $X_1, \dots, X_n$ are *independent* if $\sigma(X_1), \dots, \sigma(X_n)$ are independent.

Explicitly, recalling Theorem 2.35, we see that (ii) means that for all $F_1 \in \mathcal{E}_1, \dots, F_n \in \mathcal{E}_n$ we have[2]

$$\mathbb{P}(X_1 \in F_1, \dots, X_n \in F_n) = \mathbb{P}(X_1 \in F_1) \cdots \mathbb{P}(X_n \in F_n)\,, \tag{3.5}$$

where X_i takes values in the measurable space $(E_i, \mathcal{E}_i)$.

The following result is very convenient: it gives a concrete characterisation of independence of random variables that is very useful when working with independent random variables.

Proposition 3.12 *The random variables $X_1, \dots, X_n$ are independent if and only if the law of $(X_1, \dots, X_n)$ is the product of the laws of $X_1, \dots, X_n$, i.e.*

$$\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}\,. \tag{3.6}$$

In this case we have

$$\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] = \mathbb{E}[f_1(X_1)] \cdots \mathbb{E}[f_n(X_n)]$$

for any measurable and nonnegative functions f_i.

Proof Let $(E_i, \mathcal{E}_i)$ be the target space of X_i. Let $F_i \in \mathcal{E}_i$ for all i. Then we have

$$\mathbb{P}_{(X_1,\dots,X_n)}(F_1 \times \cdots \times F_n) = \mathbb{P}(X_1 \in F_1, \dots, X_n \in F_n)\,,$$
$$\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}(F_1 \times \cdots \times F_n) = \mathbb{P}(X_1 \in F_1) \cdots \mathbb{P}(X_n \in F_n)\,.$$

Using (3.5), we conclude that $X_1, \dots, X_n$ are independent if and only if $\mathbb{P}_{(X_1,\dots,X_n)}$ and $\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ coincide on all rectangles of the form $F_1 \times \cdots \times F_n$. The family of such rectangles,

$$\mathcal{C} = \{F_1 \times \cdots \times F_n : F_i \in \mathcal{E}_i \,\forall i\}$$

is stable under finite intersections (Theorem 3.7) and it satisfies $\sigma(\mathcal{C}) = \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n$ (recall Theorem 1.3 (ii)). By Theorem 3.9 we therefore conclude that $X_1, \dots, X_n$ are independent if and only if $\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$.

For the last statement, we use the Fubini-Tonelli theorem (Theorem 1.14) to conclude

$$\mathbb{E}\left[\prod_i f_i(X_i)\right] = \int \prod_i f_i(x_i)\, \mathbb{P}_{X_1}(\mathrm{d}x_1) \cdots \mathbb{P}_{X_n}(\mathrm{d}x_n)$$
$$= \prod_i \int f_i(x_i)\, \mathbb{P}_{X_i}(\mathrm{d}x_i) = \prod_i \mathbb{E}[f_i(X_i)]\,. \quad \square$$

Theorem 3.12 shows how to *construct* independent random variables $X_1, \dots, X_n$ with given laws $\mu_1, \dots, \mu_n$ on the spaces $(E_1, \mathcal{E}_1), \dots (E_n, \mathcal{E}_n)$, respectively. Indeed, simply choose $\Omega = E_1 \times \cdots \times E_n, \mathcal{A} = \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n, \mathbb{P} = \mu_1 \otimes \cdots \otimes \mu_n$, and set $X_i(\omega_1, \dots, \omega_n) := \omega_i$. Clearly, (3.6) holds.

Let us record some obvious but important properties of independent random variables.

[2] Recalling the convention (2.6).

Remark 3.13

(i) If $X_1, \dots, X_n$ are independent random variables with values in $\mathbb{R}$, then $\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n]$ provided that $\mathbb{E}[|X_i|] < \infty$ for all i. In particular, if $X_1, \dots, X_n \in L^1$ then $X_1 \cdots X_n \in L^1$. Without independence, this is in general false. For instance, for $X_1 = X_2 = X \in L^1$ in general we have $X \notin L^2$ (i.e. $X^2 \notin L^1$).

(ii) If $X_1, X_2 \in L^2$ are independent then $\mathrm{Cov}(X_1, X_2) = 0$. In words: independent random variables are uncorrelated. The reverse implication (uncorrelated random variables are independent) is in general wrong.

Example 3.14 To illustrate (ii), consider a random variable $X_1 \in L^2$ on $\mathbb{R}$ with a symmetric density p, i.e. $p(x) = p(-x)$. Let $\chi \in \{\pm 1\}$ be a random variable with law $\mathbb{P}(\chi = +1) = \mathbb{P}(\chi = -1) = \frac{1}{2}$. Let X_1 and χ be independent. Define $X_2 := \chi \cdot X_1$. Then we have

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] = \mathbb{E}[\chi X_1^2] = \mathbb{E}[\chi]\mathbb{E}[X_1^2] = 0\,,$$

so that X_1 and X_2 are uncorrelated. Nevertheless, X_1 and X_2 are not independent. Indeed, if they were independent, then $|X_1|$ and $|X_2| = |X_1|$ would also be independent, but a random variable that is independent of itself is necessarily constant[a]. Clearly, $|X_1|$ cannot be constant, since it has density $2p(x)\mathbf{1}_{x \geqslant 0}$.
Remarkably, if the law of (X_1, X_2) is Gaussian, then independence of X_1 and X_2 is equivalent to them being uncorrelated. This is a consequence of Wick's theorem (Exercise 3.2).

[a] If X is independent of itself then $\mathrm{Var}(X) = 0$. Hence, by Chebyshev, $\mathbb{P}(|X - \mathbb{E}[X]| > t) = 0$ for all $t > 0$, which implies that $\mathbb{P}(X \neq \mathbb{E}[X]) = 0$.

Remark 3.15 Let X, Y, Z be independent random variables. Then X and $f(Y, Z)$ are independent for any measurable function f. Indeed, by Theorem 3.12,

$$\begin{aligned}\mathbb{P}(X \in A, f(Y, Z) \in B) &= (\mathbb{P}_X \otimes \mathbb{P}_Y \otimes \mathbb{P}_Z)(A \times f^{-1}(B)) \\ &= \mathbb{P}_X(A) \cdot \mathbb{P}_Y \otimes \mathbb{P}_Z(f^{-1}(B)) = \mathbb{P}(X \in A) \cdot \mathbb{P}(f(Y, Z) \in B)\,.\end{aligned}$$

This principle of regrouping random variables can easily be generalised in the obvious way to more random variables.

3.4 The Borel-Cantelli lemma

WEEK 4

In this section we encounter the first deep result that has a genuinely probabilistic flavour. It is the standard tool to prove that some asymptotic property holds almost surely, and it lies at the heart of many important results in probability. A typical application is given in Theorem 3.20 below.

Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events in $\mathcal{A}$. We define the new events

$$\limsup_n A_n := \bigcap_{n\geqslant 0}\bigcup_{k\geqslant n} A_k$$
$$\liminf_n A_n := \bigcup_{n\geqslant 0}\bigcap_{k\geqslant n} A_k\,.$$

The following interpretation is crucial, and it is often the better way to understand these events:

$$\limsup_n A_n = \{\omega : \omega \in A_k \text{ infinitely often}\}\,,$$
$$\liminf_n A_n = \{\omega : \omega \in A_k \text{ eventually}\}\,.$$

In other words:

$\limsup_n A_n$ is the set of realisations that appear in infinitely many A_k.

$\liminf_n A_n$ is the set of realisations that appear in all A_k beyond a certain index.

It is very important that you play with these different formulations until you are comfortable with them.

We always have

$$\liminf_n A_n \subset \limsup_n A_n\,. \tag{3.7}$$

If you're not sure why this is true, pause here until you see why.

The reason behind the terminology lim sup and lim inf stems from the fact that

$$\mathbf{1}_{\limsup_n A_n} = \limsup_n \mathbf{1}_{A_n}\,, \qquad \mathbf{1}_{\liminf_n A_n} = \liminf_n \mathbf{1}_{A_n}\,,$$

where the right-hand sides are the usual lim sup and lim inf on $\mathbb{R}$; see the exercises. (This remark also provides another way of seeing (3.7).)

The event $\limsup_n A_n$ is more useful and more commonly used than $\liminf_n A_n$, which appears quite rarely in probability.

Proposition 3.16 (Borel-Cantelli lemma)

(i) If $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$ then $\mathbb{P}(\limsup_n A_n) = 0$. In other words, almost surely A_n happens only finitely often.

(ii) If $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty$ and $(A_n)_{n\in\mathbb{N}}$ are independent, then $\mathbb{P}(\limsup_n A_n) = 1$. In other words, almost surely A_n happens infinitely often.

Remark 3.17 In (ii), the independence is important. The conclusion is clearly wrong if we take $A_n = A$ for all $n \in \mathbb{N}$ with $0 < \mathbb{P}(A) < 1$.

Proof of Theorem 3.16 For (i), we write (using Exercise 1.1)

$$\mathbb{P}(\limsup_n A_n) = \lim_n \mathbb{P}\left(\bigcup_{k\geqslant n} A_k\right) \leqslant \lim_n \sum_{k\geqslant n} \mathbb{P}(A_k) = 0\,,$$

since the sum $\sum_k \mathbb{P}(A_k)$ is finite.

For (ii), we choose $N \in \mathbb{N}$ and write for any $n \geqslant N$, by independence of the events A_k,

$$\mathbb{P}\left(\bigcap_{k=N}^{n} A_k^c\right) = \prod_{k=N}^{n} \mathbb{P}(A_k^c) = \prod_{k=N}^{n} (1-\mathbb{P}(A_k)) \leqslant \prod_{k=N}^{n} \mathrm{e}^{-\mathbb{P}(A_k)} = \mathrm{e}^{-\sum_{k=N}^{n}\mathbb{P}(A_k)}\,,$$

which tends to zero as $n \to \infty$ because of the assumption $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty$. Here we used the basic inequality $1 - x \leqslant \mathrm{e}^{-x}$ (which follows for instance by convexity of the function e^{-x}). We conclude (by Exercise 1.1) that

$$\mathbb{P}\left(\bigcap_{k\geqslant N} A_k^c\right) = 0\,,$$

and hence also

$$\mathbb{P}\left(\bigcup_{N\geqslant 0}\bigcap_{k\geqslant N} A_k^c\right) = 0\,,$$

which is equivalent to

$$\mathbb{P}\left(\bigcap_{N\geqslant 0}\bigcup_{k\geqslant N} A_k\right) = 1\,. \qquad \square$$

Remark 3.18 A somewhat different proof of (i) follows from the observation that

$$\mathbb{E}\left[\sum_n \mathbf{1}_{A_n}\right] = \sum_n \mathbb{P}(A_n) < \infty\,,$$

so that the random variable $\sum_n \mathbf{1}_{A_n}$ is almost surely finite, which implies that A_n happens only finitely often.

The Borel-Cantelli lemma states that whether A_n happens infinitely often depends on *how fast* the sequence $\mathbb{P}(A_n)$ tends to zero. This principle is well illustrated by the following example, which provides a good intuition for Theorem 3.16 (i).

Example 3.19 Let $b_1, b_2, \dots \in [0,1)$. We partition $[0,\infty)$ into intervals $[a_n, a_{n+1})$ of length b_n, by setting $a_0 = 0$ and $a_n = a_{n-1} + b_n$ for $n \geqslant 1$. Now we "fold" these intervals into the unit interval $[0,1)$ using the function $f(x) := x - \lfloor x \rfloor$, the fractional part of x. Thus, we define $A_n := f([a_n, a_{n+1}))$. You may find it helpful to draw these sets.

- If $\sum_n b_n = \infty$ then $\bigcup_n [a_n, a_{n+1}) = [0,\infty)$. This means that the folded sets A_n keep on "passing through" the interval $[0,1)$, and every $\omega \in [0,1)$ is contained in infinitely many sets A_n.
- If $\sum_n b_n < \infty$ then $\bigcup_n [a_n, a_{n+1})$ is a finite interval. This means that the folded sets A_n do not cover enough ground to keep on passing through $[0,1)$, and every $\omega \in [0,1)$ is contained in only finitely many sets A_n.

These observations can be interpreted in light of the Borel-Cantelli lemma on the probability space $([0,1), \mathcal{B}([0,1)), \mathbb{P})$, where $\mathbb{P}$ is Lebesgue measure. Then $\mathbb{P}(A_n) = b_n$ (why?), and Theorem 3.16 (i) is applicable. In particular, we see that the condition $\sum_n \mathbb{P}(A_n) < \infty$ is not only sufficient but in general also necessary. Note

that Theorem 3.16 (ii) does not apply to this example, because the events (A_n) are not independent (why?).

Example 3.20 In this example we investigate the statistics of digits of random numbers. For simplicity, we work with binary digits, although the same argument can be easily adapted to any basis (such as 10). Take the probability space $([0,1), \mathcal{B}([0,1)), \mathbb{P})$, where $\mathbb{P}$ is Lebesgue measure. Use the notation $\omega = 0.\omega_1\omega_2\omega_3\cdots$ for the binary digits $\omega_n \in \{0,1\}$ of $\omega \in [0,1)$, i.e. $\omega = \sum_{n=1}^{\infty} \omega_n 2^{-n}$ (with the usual convention that we cannot have $\omega_n = 1$ for all n above a certain index; this ensures the uniqueness of the binary representation). This definition is not very direct, and it turns out to be more convenient to define the digits through the events

$$B_n := \bigcup_{k=1}^{2^{n-1}} \left[\frac{k-1/2}{2^{n-1}}, \frac{k}{2^{n-1}}\right), \qquad n \geqslant 1\,.$$

(Draw them!) Then we define the random variable $X_n := \mathbf{1}_{B_n}$. One can then show by induction that $X_n(\omega) = \omega_n$; the details are left as an exercise (drawing the events B_n should make this clear).

Next, we find that $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = \frac{1}{2}$ for all n, and we claim that $(X_n)_{n\geqslant 1}$ are independent random variables. Indeed, for any $p \in \mathbb{N}^*$ and $x_1, \dots, x_p \in \{0,1\}$ we find

$$\mathbb{P}(X_1 = x_1, \dots, X_p = x_p) = \mathbb{P}\left(\left[\sum_{k=1}^{p} x_k 2^{-k}, \sum_{k=1}^{p} x_k 2^{-k} + 2^{-p}\right)\right)$$
$$= \frac{1}{2^p} = \prod_{k=1}^{p} \mathbb{P}(X_k = x_k)\,,$$

as desired.

Moreover, we claim that for any $p \in \mathbb{N}^*$ and $x_1, \dots, x_p \in \{0,1\}$ we have, almost surely,

$$\#\{k \in \mathbb{N} : (X_{k+1}, \dots, X_{k+p}) = (x_1, \dots, x_p)\} = \infty\,. \tag{3.8}$$

This is a simple consequence of Theorem 3.16 (ii). The only issue is that events

$$\big\{(X_{k+1}, \dots, X_{k+p}) = (x_1, \dots, x_p)\big\}\,, \qquad k \in \mathbb{N}\,,$$

are not independent (since the collections of random variables on which they depend overlap). There's a very easy solution to this, however: we can just pick every p-th of them, in which case they are independent. Thus, defining $Y_n := (X_{np+1}, \dots, X_{np+p})$, from Theorem 3.15 we conclude that $(Y_n)_{n\in\mathbb{N}}$ are independent random variables. Setting $A_n := \{Y_n = (x_1, \dots, x_p)\}$, the family $(A_n)_{n\in\mathbb{N}}$ is independent and satisfies $\mathbb{P}(A_n) = 2^{-p}$. Hence Theorem 3.16 (ii) yields (3.8).

We can even upgrade[a] (3.8) by taking a countable union over $p \in \mathbb{N}^*$ and $x_1, \dots, x_p \in \{0,1\}$ to show that almost surely

$$\#\{k \in \mathbb{N} : (X_{k+1}, \dots, X_{k+p}) = (x_1, \dots, x_p)\} = \infty \qquad \forall p \geqslant 1\,,\ \forall x_1, \dots, x_p \in \{0,1\}\,.$$

In other words: Almost every real number exhibits every finite sequence of binary digits infinitely often.

[a] This is an important general observation. Let $P(\omega, i)$ be a statement depending on the realisation ω and some index $i \in I$ in an index set I. Then the statement

$$\text{for all } i \in I, \text{ almost surely } P(\omega, i)$$

is weaker than

$$\text{almost surely, for all } i \in I,\ P(\omega, i).$$

However, if the set I is countable, these statements are equivalent by a union bound.

3.5 Sums of independent random variables

The following definition gives the law of the sum of independent random variables in terms of their laws.

Definition 3.21 Let μ and ν be probability measures on $\mathbb{R}$. The *convolution* of μ and ν is the probability measure

$$\mu * \nu := p_*(\mu \otimes \nu)\,,$$

where $p(x, y) := x + y$.

Remark 3.22 This definition is a generalisation of the usual convolution of functions that you have learned in analysis. Indeed, suppose that μ and ν both have densities, i.e. $\mu(\mathrm{d}x) = f(x)\,\mathrm{d}x$ and $\nu(\mathrm{d}x) = g(x)\,\mathrm{d}x$. Then we have $(\mu * \nu)(\mathrm{d}x) = h(x)\,\mathrm{d}x$, where

$$h(x) = \int f(x-y)\,g(y)\,\mathrm{d}y\,.$$

The right-hand side is also usually denoted by $f * g$ and it is called the convolution of f and g. To verify this assertion, take a nonnegative measurable function ϕ and calculate, using Theorem 3.21,

$$\int \phi(x)\,(\mu * \nu)(\mathrm{d}x) = \int \phi(x+y)\,f(x)\,g(y)\,\mathrm{d}x\,\mathrm{d}y = \int \phi(x)\,f(x-y)\,g(y)\,\mathrm{d}x\,\mathrm{d}y$$

as desired, where in the last step we used the change of variables $x \mapsto x - y$.

Remark 3.23 If X and Y are independent random variables then clearly $\mathbb{P}_{X+Y} = \mathbb{P}_X * \mathbb{P}_Y$. Moreover, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Next, we state a weak version of the most famous result in probability theory: the *law of large numbers*. It states that the average of a large number of independent copies of a random variable is close to its expectation. The following result is known as the *weak* law of large numbers because it falls short of the best possible statement, for two reasons: the assumption $X_n \in L^2$ can be weakened to $X_n \in L^1$, and the convergence in fact holds almost surely instead of just in L^2. Later on, we shall see how to remedy both issues. The weak law has the advantage that it is very easy to prove.

Proposition 3.24 (Weak law of large numbers) *Let $(X_n)_{n\geqslant 1}$ be independent random variables in L^2 with the same law. Then*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{L^2} \mathbb{E}[X_1]\,.$$

Proof The proof is a simple computation using Theorem 3.23:

$$\mathbb{E}\left[\left(\frac{1}{n}(X_1+\cdots+X_n)-\mathbb{E}[X_1]\right)^2\right] = \operatorname{Var}\left(\frac{1}{n}(X_1+\cdots+X_n)\right)$$
$$= \frac{1}{n^2}\operatorname{Var}(X_1+\cdots+X_n) = \frac{1}{n^2}(\operatorname{Var}(X_1)+\cdots+\operatorname{Var}(X_n)) = \frac{1}{n}\operatorname{Var}(X_1)\,. \quad \square$$

Note that, for this proof to work, it in fact suffices that the random variables (X_n) be uncorrelated (i.e. $\operatorname{Cov}(X_n, X_m) = 0$ for $n \neq m$) instead of requiring independence (recall Theorem 3.13 (ii)).

The following is a nice application of the law of large numbers to polynomial approximation in numerical analysis.

Example 3.25 (Polynomial approximation) The problem of polynomial approximation is one of the classical problems in numerical analysis: Suppose that we are given the values of an unknown function f at $n+1$ points. How do we approximate f with a polynomial of degree n?
To be more precise, let f be continuous on $[0,1]$. Define the *Bernstein polynomial*

$$f_n(x) := \sum_{k=0}^{n}\binom{n}{k}x^k(1-x)^{n-k}f\left(\frac{k}{n}\right).$$

Then we claim that f_n converges to f uniformly on $[0,1]$.
To see why, take independent random variables $X_1, \dots, X_n$, each having a Bernoulli law with parameter p. Then $S_n = X_1 + \cdots + X_n$ has binomial law (recall Exercise 2.2)

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

and hence

$$\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right] = f_n(p)\,. \tag{3.9}$$

Then by the law of large numbers we have $\frac{S_n}{n} \to p$ in L^2, so that we expect $f_n(p) \to f(p)$. Let us carry out this argument carefully. Let $\varepsilon > 0$ and choose $\delta > 0$ such that $|x-y| < \delta$ implies $|f(x)-f(y)| < \varepsilon$ (since f is continuous, and hence uniformly continuous, on the compact interval $[0,1]$.) Then we partition

$$1 = \mathbf{1}_{|S_n/n-p|<\delta} + \mathbf{1}_{|S_n/n-p|\geqslant\delta}$$

and use this to estimate

$$\left|\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right]-f(p)\right| \leqslant \varepsilon + 2\|f\|_{L^\infty}\mathbb{P}\left(\left|\frac{S_n}{n}-p\right|\geqslant\delta\right)$$
$$\leqslant \varepsilon + 2\|f\|_{L^\infty}\frac{\operatorname{Var}(S_n/n)}{\delta^2} = \varepsilon + 2\|f\|_{L^\infty}\frac{p(1-p)}{n\delta^2} \leqslant \varepsilon + \|f\|_{L^\infty}\frac{1}{2n\delta^2}\,,$$

where in the second step we used Chebyshev's inequality. Recalling (3.9), we conclude the proof.

As advertised, the convergence in L^2 in Theorem 3.24 is often not strong enough, and we would need almost sure convergence. This issue is clarified in the following remark.

Remark 3.26 If $Y_n \to Y$ in L^p for $p \geqslant 1$, it could well be that $Y_n(\omega)$ fails to converge for *every* $\omega \in \Omega$. In other words, in no realisation of the randomness does Y_n converge. This is a major weakness of convergence in L^p.

For an example, we continue with Theorem 3.19. Suppose that $\sum_n b_n = \infty$. Take the random variable $Y_n := \mathbf{1}_{A_n}$. Then $\|Y_n\|_p = b_n^{1/p} \to 0$, so that Y_n converges to $Y = 0$ in L^p. However, for all $\omega \in \Omega$ we have $Y_n(\omega) = 1$ infinitely often. In particular, although Y_n converges to 0 in L^p for any $p \geqslant 1$, almost surely Y_n does not converge. Note, however, that if $\sum_n b_n < \infty$ then $Y_n \to 0$ almost surely.

Using the Borel-Cantelli lemma, we can upgrade the convergence in L^2 to almost sure convergence. The result is the following proposition. It represents an archetypal use of the Borel-Cantelli lemma.

Proposition 3.27 (Strong law of large numbers in L^4) *Let $(X_n)_{n\geqslant 1}$ be independent random variables in L^4 with the same law. Then*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{a.s.} \mathbb{E}[X_1]\,.$$

Proof The proof is very constructive. It consists of two main ideas. The first is that if we repeat the proof of Theorem 3.24 with a higher L^p norm, we get stronger decay in n of the error probabilities. The second is that if these error probabilities are summable over n, we can use the Borel-Cantelli lemma to conclude almost sure convergence.

To begin with, without loss of generality we can replace X_n with $X_n - \mathbb{E}[X_n]$, and hence suppose that $\mathbb{E}[X_n] = 0$. Then we simply calculate, using the independence of the random variables (X_n), to get

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^n X_k\right)^4\right] &= \frac{1}{n^4}\sum_{k_1,\dots,k_4=1}^n \mathbb{E}[X_{k_1}X_{k_2}X_{k_3}X_{k_4}] \\ &= \frac{1}{n^4}\left(n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2\right) = O\left(\frac{1}{n^2}\right),\end{aligned}$$

where in the second step we classified the indices k_1, k_2, k_3, k_4 according to their coincidences: to obtain a nonzero contribution we need either all four indices to coincide, or two and two indices to coincide, which gives rise to the two terms in the third step.

We conclude that

$$\sum_{n\geqslant 1}\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^n X_k\right)^4\right] = \mathbb{E}\left[\sum_{n\geqslant 1}\left(\frac{1}{n}\sum_{k=1}^n X_k\right)^4\right] < \infty\,,$$

and hence

$$\sum_{n\geqslant 1}\left(\frac{1}{n}\sum_{k=1}^n X_k\right)^4 < \infty$$

almost surely, from which the claim follows. This is an application of the (proof of the) Borel-Cantelli lemma (see also Theorem 3.18). □

Remark 3.28 The condition that $X_n \in L^4$ is not optimal. In fact, Theorem 3.27 remains true under the weaker assumption $X_n \in L^1$ (see Theorem A.5 in Chapter A). This assumption is known to be optimal (in the sense that the conclusion must be wrong if $X_n \notin L^1$). This optimal result is usually known as the *strong law of large numbers* or just the *law of large numbers*. It turns out that, unlike the relatively straightforward proof of Theorem 3.27, the proof of the law of Theorem A.5 is somewhat involved, and was a major achievement in 20th century probability. Its proof relies on some far-reaching and deep ideas of probability theory. If you are interested, you can read all about it in Chapter A.
However, the L^4 assumption from Theorem 3.27 is good enough for most[a] applications, and in this course we shall restrict ourselves to it.

[a] Although not all; see e.g. the proof of Theorem 5.45 below.

Let us discuss some application of the strong law of large numbers.

Corollary 3.29 *Let $(A_n)_{n\geqslant 1}$ be independent of events of constant probability. Then*

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{A_k} \xrightarrow{a.s.} \mathbb{P}(A_1)\,.$$

We can use this result to bring Theorem 3.20 to a striking conclusion.

Example 3.30 (Theorem 3.20 continued) Let $p \in \mathbb{N}^*$ and $l \in \{1, \dots, p\}$. Define

$$Y_n^l := (X_{(n-1)p+l}, \dots, X_{(n-1)p+l+p-1})\,.$$

Explicitly, the sequence $Y_1^l, Y_2^l, \dots$ is

$$(X_l, \dots, X_{l+p-1}), (X_{p+l}, \dots, X_{p+l+p-1}), \dots$$

By Theorem 3.20 and Theorem 3.15, $(Y_n^l)_{n\in\mathbb{N}^*}$ is an independent family of random variables, for each $l \in \{1, \dots, p\}$. Hence, for any $x = (x_1, \dots, x_p) \in \{0,1\}^p$, applying Theorem 3.29 to the events $A_n = \{Y_n^l = x\}$, we find that

$$\frac{1}{n}\#\left\{i \in \{1, \dots, n\} : Y_i^l = x\right\} \xrightarrow{a.s.} \frac{1}{2^p}\,.$$

Since this holds for all $l \in \{1, \dots, p\}$ we deduce that

$$\frac{1}{n}\#\left\{i \in \{1, \dots, n\} : (X_i, \dots, X_{i+p-1}) = x\right\} \xrightarrow{a.s.} \frac{1}{2^p}\,.$$

Taking the countable union over $p \in \mathbb{N}^*$ and $x \in \{0,1\}^p$, we conclude: almost surely, for all $p \in \mathbb{N}^*$, $x \in \{0,1\}^p$,

$$\frac{1}{n}\#\left\{i \in \{1, \dots, n\} : (X_i, \dots, X_{i+p-1}) = x\right\} \longrightarrow \frac{1}{2^p}\,.$$

In words: for almost every real number $\omega \in [0,1)$, any finite sequence of length p appears with frequency 2^{-p} in the binary digits of ω.
Note that it is difficult to construct a number ω with this rather striking property. The easiest way to show that such a number exists is the preceding probabilistic argument. Not only does it show that such a number exists, but it shows that almost all numbers have this property.

CHAPTER 4

Convergence of random variables

WEEK 5

In this chapter we study the convergence of random variables in detail. We shall study the most important notions of convergence: almost surely, in probability, in L^p, and in law.

4.1 Notions of convergence

Let $(X_n)_{n\in\mathbb{N}^*}$ and X be random variables with values in $\mathbb{R}$. In this section we wish to understand different notions of the convergence $X_n \to X$ and any logical implications between them.

Recall that we have already seen two notions of convergence:

- $X_n \xrightarrow{\text{a.s.}} X$ if $\mathbb{P}(\lim_n X_n = X) = 1$.
- $X_n \xrightarrow{L^p} X$ if $\lim_n \mathbb{E}[|X_n - X|^p] = 0$.

The following definition is very useful, and specific to probability.

Definition 4.1 The random variables X_n converge *in probability* to X, denoted $X_n \xrightarrow{\mathbb{P}} X$, if for all $\varepsilon > 0$ we have

$$\lim_n \mathbb{P}(|X_n - X| > \varepsilon) = 0\,.$$

It is often useful to observe that this notion of convergence is *metrisable*, i.e. it arises from a metric on the space of all random variables.

Definition 4.2 Let $\mathcal{L}^0$ be the space of random variables on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{R}$, and let $L^0 := \mathcal{L}^0/\sim$, where $\sim$ is the equivalence relation defined by $X \sim Y$ if and only if $X = Y$ almost surely. For $X, Y \in L^0$ we define

$$d(X, Y) := \mathbb{E}[|X - Y| \wedge 1]\,.$$

Proposition 4.3 *The space (L^0, d) is a complete metric space, and $X_n \xrightarrow{\mathbb{P}} X$ if and only if $d(X_n, X) \to 0$.*

Proof It is easy to check that d is a metric.

Let us now verify that $X_n \xrightarrow{\mathbb{P}} X$ implies $d(X_n, X) \to 0$. Suppose that $X_n \xrightarrow{\mathbb{P}} X$ and choose an arbitrary $\varepsilon \in (0, 1]$. Then

$$d(X_n, X) = \mathbb{E}[|X_n - X| \wedge 1] = \mathbb{E}[|X_n - X| \mathbf{1}_{|X_n - X| \leqslant \varepsilon}] + \mathbb{E}[(|X_n - X| \wedge 1) \mathbf{1}_{|X_n - X| > \varepsilon}] \\ \leqslant \varepsilon + \mathbb{P}(|X_n - X| > \varepsilon) \longrightarrow \varepsilon ,$$

by assumption. Since $\varepsilon > 0$ was arbitrary, we conclude that $d(X_n, X) \to 0$.

Conversely, suppose that $d(X_n, X) \to 0$. Then for all $\varepsilon \in (0, 1]$ we have, by Chebyshev's inequality,

$$\mathbb{P}(|X_n - X| > \varepsilon) \leqslant \frac{1}{\varepsilon} \mathbb{E}[|X_n - X| \wedge 1] \to 0 ,$$

i.e. $X_n \xrightarrow{\mathbb{P}} X$.

All that remains, therefore, is to show that the metric space (L^0, d) is complete. To that end, let (X_n) be a Cauchy sequence for $d(\cdot, \cdot)$. Choose a subsequence $Y_k = X_{n_k}$ such that $d(Y_k, Y_{k+1}) \leqslant 2^{-k}$. We then use the Borel-Cantelli lemma (see also Theorem 3.18) with

$$\mathbb{E}\left[\sum_{k=1}^{\infty} (|Y_{k+1} - Y_k| \wedge 1)\right] \leqslant \sum_{k=1}^{\infty} 2^{-k} < \infty ,$$

so that

$$\sum_{k=1}^{\infty} (|Y_{k+1} - Y_k| \wedge 1) < \infty \quad \text{a.s.} ,$$

which implies

$$\sum_{k=1}^{\infty} |Y_{k+1} - Y_k| < \infty \quad \text{a.s.} .$$

Defining

$$X := Y_1 + \sum_{k=1}^{\infty} (Y_{k+1} - Y_k) ,$$

we therefore have $Y_k \xrightarrow{\text{a.s.}} X$ as $k \to \infty$. Hence,

$$d(Y_k, X) = \mathbb{E}[|Y_k - X| \wedge 1] \longrightarrow 0$$

as $k \to \infty$, by dominated convergence. We conclude that $d(X_n, X) \to 0$ as $n \to \infty$. □

The argument in the preceding proof of completeness is a general and important fact from probability and measure theory: convergence in probability does not in general imply almost sure convergence, but it does so provided that we restrict ourselves to a suitable *subsequence*. This is made precise in the following proposition.

Proposition 4.4 *Let X_n, X be random variables with values in $\mathbb{R}$.*

(i) If $X_n \xrightarrow{a.s.} X$ or $X_n \xrightarrow{L^p} X$ then $X_n \xrightarrow{\mathbb{P}} X$.

(ii) If $X_n \xrightarrow{\mathbb{P}} X$ then there exists a subsequence (X_{n_k}) such that $X_{n_k} \xrightarrow{a.s.} X$.

Proof Part (ii) was already established in the proof of Theorem 4.3. For part (i), if $X_n \xrightarrow{\text{a.s.}} X$ then $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{E}[\mathbf{1}_{|X_n - X| > \varepsilon}] \to 0$ by dominated convergence, and if $X_n \xrightarrow{L^p} X$ then $\mathbb{P}(|X_n - X| > \varepsilon) \leqslant \frac{1}{\varepsilon^p} \mathbb{E}[|X_n - X|^p] \to 0$ for any $\varepsilon > 0$. □

Remark 4.5 In Theorem 4.4 (ii), it is in general necessary to take a subsequence; see Theorem 3.26. (In this example, after taking a subsequence we can ensure that $\sum_n b_n < \infty$.)

4.2 Convergence in law

In this section we introduce the final notion of convergence of random variables in this course. We fix the dimension $d \in \mathbb{N}^*$ throughout. We denote by $C_b \equiv C_b(\mathbb{R}^d)$ the space of bounded continuous real-valued functions on $\mathbb{R}^d$.

Definition 4.6

(i) Let μ_n, $n \in \mathbb{N}^*$, and μ be probability measures on $\mathbb{R}^d$. We say that μ_n *converges weakly* to μ, denoted by $\mu_n \xrightarrow{w} \mu$, if

$$\int \varphi \, \mathrm{d}\mu_n \longrightarrow \int \varphi \, \mathrm{d}\mu \,, \qquad \forall \varphi \in C_b \,.$$

(ii) Let X_n, $n \in \mathbb{N}^*$, and X be random variables with values in $\mathbb{R}^d$. We say that X_n *converges in law*, or *in distribution*, to X, denoted by $X_n \xrightarrow{d} X$, if

$$\mathbb{P}_{X_n} \xrightarrow{w} \mathbb{P}_X \,.$$

Explicitly, this means that

$$\mathbb{E}[\varphi(X_n)] \longrightarrow \mathbb{E}[\varphi(X)] \,, \qquad \forall \varphi \in C_b \,.$$

Remark 4.7 The convergence in law $\xrightarrow{d}$ is very different in nature from the other modes of convergence $\xrightarrow{a.s.}$, $\xrightarrow{\mathbb{P}}$, $\xrightarrow{L^2}$ that we have seen up to now: it only pertains to the laws of the random variables. In particular, the random variables X_n and X can all be defined on different probability spaces. Moreover, the limit is (trivially) not unique: if X and Y are different random variables with the same law and $X_n \xrightarrow{d} X$ then clearly also $X_n \xrightarrow{d} Y$. (In contrast, the limit of weak convergence of probability measures is unique.)

Example 4.8

(i) If $a_n \to a$ then $\delta_{a_n} \xrightarrow{w} \delta_a$ (by definition of continuity).

(ii) If the law of X_n is uniform on $\{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\}$ and the law of X is Lebesgue measure on $[0,1]$, then $X_n \xrightarrow{d} X$ (by the Riemann sum approximation of integrals of continuous functions).

(iii) Let μ be a probability measure on $\mathbb{R}$ and define the scaling function $s^\eta(x) := \eta x$ for $\eta > 0$. Then $s^\eta_* \mu \xrightarrow{w} \delta_0$ as $\eta \to 0$. To show this, take a function $\varphi \in C_b$ and write, using the change of variables $x = s^\eta(y)$,

$$\int \varphi(x) \, s^\eta_* \mu(\mathrm{d}x) = \int \varphi(s^\eta(y)) \, \mu(\mathrm{d}y) = \int \varphi(\eta y) \, \mu(\mathrm{d}y) \to \varphi(0)$$

as $\eta \to 0$, by dominated convergence.

In the important special case where $\mu(\mathrm{d}x) = p(x)\,\mathrm{d}x$ has a density p, we have

$$s^\eta_* \mu(\mathrm{d}x) = \frac{1}{\eta} p\left(\frac{x}{\eta}\right) \mathrm{d}x \,.$$

The right-hand side is usually known as an *approximate delta function*. Such functions play a very important role in analysis. One such application is given in the Fourier inversion formula in Section 4.3.

Proposition 4.9 *If $X_n \xrightarrow{\mathbb{P}} X$ then $X_n \xrightarrow{\mathrm{d}} X$.*

Proof We proceed by contradiction and suppose that $X_n \xrightarrow{\mathbb{P}} X$ but X_n does not converge to X in law. The latter means that there exists $\varphi \in C_b$ such that $\mathbb{E}[\varphi(X_n)] \not\to \mathbb{E}[\varphi(X)]$. Hence, there exists a subsequence (n_k) and $\varepsilon > 0$ such that

$$\left|\mathbb{E}[\varphi(X_{n_k})] - \mathbb{E}[\varphi(X)]\right| \geqslant \varepsilon \tag{4.1}$$

for all k. Moreover, by Theorem 4.4 (ii), there exists a further subsequence (n_{k_l}) such that $X_{n_{k_l}} \to X$ a.s. as $l \to \infty$. But by dominated convergence, we have

$$|\mathbb{E}[\varphi(X_{n_{k_l}})] - \mathbb{E}[\varphi(X)]| \to 0$$

as $l \to \infty$, in contradiction to (4.1). □

Remark 4.10 The reverse implication of Theorem 4.9 is false. Worse: if $X_n \xrightarrow{\mathrm{d}} X$ then the very statement $X_n \xrightarrow{\mathbb{P}} X$ is in general meaningless! This is because $X_n \xrightarrow{\mathrm{d}} X$ does not imply that X_n and X are all defined on the same probability space, while $X_n \xrightarrow{\mathbb{P}} X$ requires all random variables to be defined on the same probability space (see Theorem 4.7). Even when all random variables are defined on the same probability space, it is easy to think of counterexamples. For example, let X have a Bernoulli law with parameter $p = 1/2$ and set $X_n := 1 - X$ for all n. Then clearly $\mathbb{P}_{X_n} = \mathbb{P}_X$, so that $X_n \xrightarrow{\mathrm{d}} X$, but because $|X - X_n| = 1$ a.s., clearly X_n does not converge to X is probability.
However, if $X_n \xrightarrow{\mathrm{d}} a$ for some constant a then the implication $X_n \xrightarrow{\mathbb{P}} a$ does hold. To show this, let $\varepsilon > 0$ and define the continuous bounded function

$$\varphi(x) := \frac{|x - a|}{\varepsilon} \wedge 1 .$$

(Plot this function!) Then

$$\mathbb{P}(|X_n - a| > \varepsilon) = \mathbb{E}[\mathbf{1}_{|X_n - a| > \varepsilon}] \leqslant \mathbb{E}[\varphi(X_n)] \to \varphi(a) = 0$$

as $n \to \infty$, by assumption $X_n \xrightarrow{\mathrm{d}} a$.

It turns out that weak convergence is a remarkably polyvalent concept, and there are many, very useful, equivalent criteria to characterise it. The following proposition is the first step in this direction. To state it, we use the notation $C_c \equiv C_c(\mathbb{R}^d)$ to denote the space of continuous functions of compact support[1]. We recall the supremum norm

$$\|\varphi\|_\infty := \sup_{x \in \mathbb{R}^d} |\varphi(x)|$$

for any $\varphi \in C_b$.

Proposition 4.11 *Let $H \subset C_c$ be such that the closure of H under $\|\cdot\|_\infty$ contains C_c. Let μ_n and μ be probability measures on $\mathbb{R}^d$. Then the following are equivalent.*

(i) $\mu_n \xrightarrow{\mathrm{w}} \mu$ *(i.e.* $\forall \varphi \in C_b$, $\int \varphi \,\mathrm{d}\mu_n \to \int \varphi \,\mathrm{d}\mu$*).*
(ii) $\forall \varphi \in C_c$, $\int \varphi \,\mathrm{d}\mu_n \to \int \varphi \,\mathrm{d}\mu$.
(iii) $\forall \varphi \in H$, $\int \varphi \,\mathrm{d}\mu_n \to \int \varphi \,\mathrm{d}\mu$.

[1] We recall that the *support* of a function $f : \mathbb{R}^d \to \mathbb{R}$ is the set $\operatorname{supp} f := \overline{\{x \in \mathbb{R}^d : f(x) \neq 0\}}$. Hence, the condition that $\operatorname{supp} f$ be compact simply means that it is bounded.

Proof The implications (i)⇒(ii) and (i)⇒(iii) are obvious. We shall show (ii)⇒(i) and (iii)⇒(ii).

To show (ii)⇒(i), suppose (ii). Let $\varphi \in C_b$. Choose a sequence $f_k \in C_c$ such that $0 \leqslant f_k \leqslant 1$ and $f_k \uparrow 1$ as $k \to \infty$ (you can take for instance $f_k(x) = (1 - |x/k|)_+$). Then we telescope

$$\begin{aligned}\int \varphi \, \mathrm{d}\mu_n - \int \varphi \, \mathrm{d}\mu = &\int \varphi \, \mathrm{d}\mu_n - \int \varphi f_k \, \mathrm{d}\mu_n \\ &+ \int \varphi f_k \, \mathrm{d}\mu_n - \int \varphi f_k \, \mathrm{d}\mu \\ &+ \int \varphi f_k \, \mathrm{d}\mu - \int \varphi \, \mathrm{d}\mu \,,\end{aligned}$$

and estimate each line on the right-hand side separately.

- For any $k \in \mathbb{N}^*$, the second line tends to zero as $n \to \infty$, by assumption (ii) since $\varphi f_k \in C_c$.
- For any $k \in \mathbb{N}^*$, the first line is estimated in absolute value by
$$\|\varphi\|_\infty \left(1 - \int f_k \, \mathrm{d}\mu_n\right) \underset{n\to\infty}{\longrightarrow} \|\varphi\|_\infty \left(1 - \int f_k \, \mathrm{d}\mu\right),$$
where we again used (ii) since $f_k \in C_c$.
- The third line is estimated in absolute value by
$$\|\varphi\|_\infty \left(1 - \int f_k \, \mathrm{d}\mu\right).$$

Putting everything together, we conclude that for any $k \in \mathbb{N}^*$ we have

$$\limsup_{n\to\infty} \left| \int \varphi \, \mathrm{d}\mu_n - \int \varphi \, \mathrm{d}\mu \right| \leqslant 2\|\varphi\|_\infty \left(1 - \int f_k \, \mathrm{d}\mu\right).$$

Since k was arbitrary, we can take $k \to \infty$, under which the right-hand side tends to zero by dominated convergence. This concludes the proof of (ii)⇒(i).

To show (iii)⇒(ii), suppose (iii). Let $\varphi \in C_c$. Choose a sequence $\varphi_k \in H$ such that $\|\varphi_k - \varphi\|_\infty \to 0$ as $k \to \infty$. Then for any $k \in \mathbb{N}^*$ we have, again by telescoping,

$$\begin{aligned}&\limsup_{n\to\infty} \left| \int \varphi \, \mathrm{d}\mu_n - \int \varphi \, \mathrm{d}\mu \right| \\ &\quad\leqslant \limsup_{n\to\infty} \left(\left| \int \varphi \, \mathrm{d}\mu_n - \int \varphi_k \, \mathrm{d}\mu_n \right| + \left| \int \varphi_k \, \mathrm{d}\mu_n - \int \varphi_k \, \mathrm{d}\mu \right| + \left| \int \varphi_k \, \mathrm{d}\mu - \int \varphi \, \mathrm{d}\mu \right| \right) \\ &\quad\leqslant 2\|\varphi - \varphi_k\|_\infty \underset{k\to\infty}{\longrightarrow} 0 \,,\end{aligned}$$

where we used that for any $k \in \mathbb{N}^*$, the middle term on the second line tends to zero as $n \to \infty$ by (iii). This concludes the proof of (iii)⇒(ii). □

Next, we state the most useful set of equivalent criteria characterising weak convergence of probability measures. It is usually called, somewhat cryptically, the *portmanteau theorem*[2].

Before stating the portmanteau theorem, we remark that $\mu_n \xrightarrow{\text{w}} \mu$ does in general not imply that $\mu_n(B) \to \mu(B)$ for all $B \in \mathcal{B}(\mathbb{R}^d)$. A counterexample is provided by Theorem 4.8 (i): $\delta_{1/n} \xrightarrow{\text{w}} \delta_0$, but $\delta_{1/n}(\{0\}) = 0$, which does not converge to $\delta_0(\{0\}) = 1$. We shall see that weak convergence is equivalent to convergence on a subset of sets $B \in \mathcal{B}(\mathbb{R}^d)$, namely those whose boundary $\partial B = \bar{B} \setminus \mathring{B}$ has limiting measure zero.

Proposition 4.12 (Portmanteau theorem) *Let μ_n and μ be probability measures on $\mathbb{R}^d$. Then the following are equivalent.*

(i) $\mu_n \xrightarrow{\text{w}} \mu$.
(ii) *For any open* $G \subset \mathbb{R}^d$, $\liminf_n \mu_n(G) \geqslant \mu(G)$.
(iii) *For any closed* $F \subset \mathbb{R}^d$, $\limsup_n \mu_n(F) \leqslant \mu(F)$.
(iv) *For any* $B \in \mathcal{B}(\mathbb{R}^d)$ *such that* $\mu(\partial B) = 0$, $\lim_n \mu_n(B) = \mu(B)$.

Proof We prove the following implications.

(i)⇒(ii). Let G be open. Then there exists a sequence $\varphi_k \in C_b$ such that $0 \leqslant \varphi_k \leqslant \mathbf{1}_G$ and $\varphi_k \uparrow \mathbf{1}_G$. For instance, we can take

$$\varphi_k(x) := (k \operatorname{dist}(x, G^c)) \wedge 1 .$$

(Note that the property $\varphi_k \uparrow \mathbf{1}_G$ holds because G is open.) Since $\varphi_k \leqslant \mathbf{1}_G$, we find

$$\liminf_n \mu_n(G) \geqslant \sup_k \left(\liminf_n \int \varphi_k \, \mathrm{d}\mu_n \right) = \sup_k \int \varphi_k \, \mathrm{d}\mu = \mu(G) ,$$

where the last step follows by monotone convergence.

(ii)⇔(iii). This is obvious by taking $F = G^c$.

(ii), (iii)⇒(iv). Let $B \in \mathcal{B}(\mathbb{R}^d)$. Then by (iii) we have

$$\limsup_n \mu_n(B) \leqslant \limsup_n \mu_n(\bar{B}) \leqslant \mu(\bar{B})$$

and by (ii) we have

$$\liminf_n \mu_n(B) \geqslant \liminf_n \mu_n(\mathring{B}) \geqslant \mu(\mathring{B}) .$$

If $\mu(\partial B) = 0$ then $\mu(\bar{B}) = \mu(\mathring{B}) = \mu(B)$, and we conclude (iv).

(iv)⇒(i). This is the last remaining implication. Let $\varphi \in C_b$ and suppose without loss of generality that $\varphi \geqslant 0$ (otherwise split $\varphi = \varphi_+ - \varphi_-$ with $\varphi_+, \varphi_- \geqslant 0$). With $K := \sup_x \varphi(x)$ we have (recall Exercise 2.4)

$$\int \varphi(x) \, \mu(\mathrm{d}x) = \int \int_0^K \mathbf{1}_{t \leqslant \varphi(x)} \, \mathrm{d}t \, \mu(\mathrm{d}x) = \int_0^K \mu(E_t^\varphi) \, \mathrm{d}t ,$$

[2]The origin of this term is somewhat unclear. In English, a *portmanteau* is traditionally a large suitcase made of leather, that opens into two equal parts, usually used to transport coats. Similarly, the portmanteau theorem bundles together multiple conditions that are each equivalent to weak convergence. (Confusingly, although the word is obviously of French origin, in French the word *portemanteau* means something altogether different: a standing piece of furniture on which one can hang coats.) In his 1871 work *Through the Looking Glass*, the sequel to *Alice in Wonderland*, Lewis Carroll coined the term *portmanteau* to denote a word that has been obtained by gluing pieces of other words together (such as “motel” from “motor” and “hotel”).

where in the last step we used Fubini's theorem and defined

$$E_t^\varphi := \{x \in \mathbb{R}^d : \varphi(x) \geqslant t\}\,.$$

By the same argument,

$$\int \varphi(x)\,\mu_n(\mathrm{d}x) = \int_0^K \mu_n(E_t^\varphi)\,\mathrm{d}t\,.$$

To conclude the argument, we make two claims.

- First, $\partial E_t^\varphi \subset \{x \in \mathbb{R}^d : \varphi(x) = t\}$. To see this, we note that since φ is continuous, E_t^φ is closed as the preimage of a closed set. Moreover,

  $$\overset{\circ}{E}{}_t^\varphi \supset \{x \in \mathbb{R}^d : \varphi(x) > t\}\,,$$

  since the right-hand side is open (as the preimage of an open set) and contained in E_t^φ. Hence,

  $$\partial E_t^\varphi = E_t^\varphi \setminus \overset{\circ}{E}{}_t^\varphi \subset \{x \in \mathbb{R}^d : \varphi(x) = t\}\,,$$

  as claimed.
- Second, the set

  $$\big\{t \in [0,K] : \mu(\{x : \varphi(x) = t\}) > 0\big\}$$

  is at most countable. This follows from the observation that this set can be written as

  $$\bigcup_{k\geqslant 1}\left\{t \in [0,K] : \mu(\{x : \varphi(x) = t\}) \geqslant \frac{1}{k}\right\},$$

  and for each $k \geqslant 1$ the set on the right-hand side is a set of cardinality at most k (recall that μ has total measure 1), in particular finite.

Putting both claims together, we use (iv) to conclude that $\mu_n(E_t^\varphi) \to \mu(E_t^\varphi)$ as $n \to \infty$ for almost all t. Hence, by dominated convergence we have

$$\int \varphi(x)\,\mu_n(\mathrm{d}x) = \int_0^K \mu_n(E_t^\varphi)\,\mathrm{d}t \longrightarrow \int_0^K \mu(E_t^\varphi)\,\mathrm{d}t = \int \varphi(x)\,\mu(\mathrm{d}x)\,,$$

as desired. □

As a corollary of the portmanteau theorem, we deduce yet another criterion for convergence in law on $\mathbb{R}$: pointwise convergence of the distribution function at its points of continuity (think again of Example 4.8 (i) for why the last condition is needed).

Proposition 4.13 *Let X_n, X be real-valued random variables. Then $X_n \xrightarrow{\mathrm{d}} X$ if and only if $F_{X_n}(x) \to F_X(x)$ for all x where F_X is continuous.*

Proof The "only if" implication is immediate from Theorem 4.12 (iv). Indeed, by Theorem 4.12 (iv), convergence in law implies that

$$F_{X_n}(x) = \mathbb{P}(X_n \leqslant x) = \mathbb{P}_{X_n}((-\infty, x]) \longrightarrow \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X \leqslant x) = F_X(x)$$

for all $x \in \mathbb{R}$ such that $\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x) = 0$ (since $\partial(-\infty, x] = \{x\}$). Moreover, if F is continuous at x it means that $\lim_{n\to\infty} F(x - 1/n) = F(x)$, which implies that $\mathbb{P}(X = x) = \mathbb{P}(X \leqslant x) - \mathbb{P}(X < x) = 0$.

For the "if" implication, we abbreviate $\mu := \mathbb{P}_X$ and $F := F_X$ as well as $\mu_n := \mathbb{P}_{X_n}$ and $F_n := F_{X_n}$. First we claim that the set D of points of discontinuity of F is at most countable. This is an exercise in real analysis that we recall here. Since F is right-continuous and nondecreasing, at any $x \in D$ we have $F(x-) := \lim_{y\uparrow x} F(y) < F(x)$. Hence, there exists

$q(x) \in \mathbb{Q} \cap (F(x-), F(x))$. By monotonicity of F, the map $q : D \to \mathbb{Q}$ is injective, which proves the claim. In particular, the set $\mathbb{R} \setminus D$ of points of continuity of F is dense in $\mathbb{R}$.

Next, by right-continuity and by definition of $F(x-)$, for any $x \in \mathbb{R}$ and any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$F(x+\delta) \leqslant F(x) + \varepsilon\,, \qquad F(x-\delta) \geqslant F(x-) - \varepsilon\,.$$

Choosing $a, b \in \mathbb{R} \setminus D$ satisfying $x - \delta \leqslant a \leqslant x \leqslant b \leqslant x + \delta$ (which is possible by density of $\mathbb{R} \setminus D$), we have, by assumption and by monotonicity of F,

$$\lim_n F_n(a) = F(a) \geqslant F(x-\delta) \geqslant F(x-) - \varepsilon\,,$$

which implies

$$\liminf_n F_n(x-) \geqslant \liminf_n F_n(a) \geqslant F(x-) - \varepsilon\,.$$

Since $\varepsilon > 0$ was arbitrary, we conclude that

$$\liminf_n F_n(x-) \geqslant F(x-)\,. \tag{4.2}$$

Let us repeat the same argument for the lim sup:

$$\lim_n F_n(b) = F(b) \leqslant F(x+\delta) \leqslant F(x) + \varepsilon\,,$$

which implies

$$\limsup_n F_n(x) \leqslant \limsup_n F_n(b) \leqslant F(x) + \varepsilon\,.$$

Since $\varepsilon > 0$ was arbitrary, we conclude that

$$\limsup_n F_n(x) \leqslant F(x)\,. \tag{4.3}$$

Let[3] $a < b$, and notice that $\mu((a,b)) = F(b-) - F(a)$. From (4.2) and (4.3), we therefore conclude

$$\liminf_n \mu_n((a,b)) = \liminf_n (F_n(b-) - F_n(a)) \geqslant F(b-) - F(a) = \mu((a,b))\,. \tag{4.4}$$

Hence, we have verified Theorem 4.12 (ii) for the special case that G is an interval.

To obtain the general case, we recall from analysis that any open set G can be written as a countable disjoint union of open intervals I_k, i.e. $G = \bigcup_{k\geqslant 1} I_k$. (For the proof, we simply decompose G into its connected components, which are intervals, and note that, since each such interval contains a point in $\mathbb{Q}$ unique to that interval, there are at most countably many intervals.) Hence,

$$\begin{aligned}\liminf_n \mu_n(G) = \liminf_n \mu_n\Big(\bigcup_{k\geqslant 1} I_k\Big) &= \liminf_n \sum_{k\geqslant 1} \mu_n(I_k) \\ &\geqslant \sum_{k\geqslant 1} \liminf_n \mu_n(I_k) \geqslant \sum_{k\geqslant 1} \mu(I_k) = \mu(G)\,,\end{aligned}$$

where in the third step we used Fatou's lemma (for the counting measure $\sum_{k\geqslant 1}$), and in the fourth step we used (4.4). We have therefore proved Theorem 4.12 (ii) for a general open set G. □

[3] Note that these a, b are different from the ones above used to prove (4.2) and (4.3).

4.3 Characteristic function

WEEK 7

The term *characteristic function* is used in probability theory to denote the *Fourier transform* of a law. As we shall see, it is a beautiful and incredibly powerful tool. Before giving the precise definition, for your general mathematical culture it is helpful to review the key ideas and definitions of Fourier analysis. For any $\xi \in \mathbb{R}^d$, we define the *plane wave* to be the function $\mathbb{R}^d \to \mathbb{C}$ defined by

$$x \mapsto \mathrm{e}^{-\mathrm{i}\xi\cdot x}\,.$$

To understand the term of plane wave, you can simply decompose $\mathrm{e}^{-\mathrm{i}\xi\cdot x}$ into its real and imaginary parts[4] and plot these as a function of x (for instance for $d = 2$): you will see a series of parallel waves, like ones at open sea far from the shore.

The main idea behind Fourier analysis is that *any function can be represented as a superposition of plane waves and the corresponding coefficients are explicitly computable*. This is very plainly illustrated in the following finite-dimensional setting. For each $N \in \mathbb{N}^*$, define the *discrete cube*

$$\Lambda := \{0, 1, \dots, N-1\}^d$$

and the *dual cube*

$$\Lambda^* := \frac{2\pi}{N}\Lambda\,.$$

Consider the finite-dimensional complex Hilbert spaces $V := \mathbb{C}^\Lambda$ and $V^* := \mathbb{C}^{\Lambda^*}$. We use the notations $f = (f(x))_{x\in\Lambda} \in V$ and $f = (f(\xi))_{\xi\in\Lambda^*} \in V^*$ for vectors in these spaces. They carry the complex inner products

$$\langle f\,,g\rangle_V := \sum_{x\in\Lambda} \overline{f(x)}g(x)\,, \qquad \langle f\,,g\rangle_{V^*} := \sum_{\xi\in\Lambda^*} \overline{f(\xi)}g(\xi)\,.$$

For any $\xi \in \Lambda^*$ we define the vector $e_\xi \in V$ as the normalized plane wave

$$e_\xi(x) := \frac{1}{N^{d/2}}\,\mathrm{e}^{-\mathrm{i}\xi\cdot x}\,.$$

Now the truly wonderful fact is that the family $(e_\xi)_{\xi\in\Lambda^*}$ is an orthonormal basis of V! I strongly recommend that you check this carefully; it is a simple exercise using finite geometric series.

The *Fourier transform* of a vector $f \in V$ is the vector $\widehat{f} \in V^*$ defined by

$$\widehat{f}(\xi) := \langle e_\xi\,, f\rangle\,. \tag{4.5}$$

In other words, Fourier transformation is nothing but a change of basis from one orthonormal basis (the standard basis of $\mathbb{C}^\Lambda$) to another orthonormal basis (the basis (e_ξ)). Hence, we can write f as a superposition of plane waves,

$$f = \sum_{\xi\in\Lambda^*} \widehat{f}(\xi)\,e_\xi\,. \tag{4.6}$$

The relations (4.5) and (4.6) can be explicitely written as

$$\widehat{f}(\xi) = \frac{1}{N^{d/2}}\sum_{x\in\Lambda} \mathrm{e}^{\mathrm{i}\xi\cdot x}\,f(x)\,, \qquad f(x) = \frac{1}{N^{d/2}}\sum_{\xi\in\Lambda^*} \mathrm{e}^{-\mathrm{i}\xi\cdot x}\,\widehat{f}(\xi)\,, \tag{4.7}$$

respectively. The former is usually called the Fourier transform and the latter the inverse Fourier transform. Remarkably, they have almost exactly the same form (up to the sign of the argument).

[4] Fourier analysis is indeed sometimes performed for real functions only, which requires dealing with the real and imaginary parts of $\mathrm{e}^{-\mathrm{i}\xi\cdot x}$ separately, resulting in complicated formulas involving sines and cosines. This approach leads to a hot complicated mess, which makes everything harder without any advantages to make up for it.

Summarising, Fourier transformation can be viewed as simply a change of orthonormal basis. This is somewhat complicated by the fact that, as in this class, it is often applied in infinite dimensions, which leads to analytic complications (see e.g. the precise statement of Theorem 4.16 below, as well as Theorem 4.17 for a simplified formulation under stronger analytic assumptions). It is a tremendously useful tool for many reasons. One such reason is that it diagonalises all differential operators (to see why, you can immediately check that differentiating a plane wave $e^{-i\xi\cdot x}$ gives $-i\xi$ times the same plane wave, so that a plane wave is an eigenfunction of the derivative operator). As a consequence, it is the most important and celebrated tool in all of analysis, upon which basically the entire modern theory of partial differential equations is founded. In this section we shall see other remarkable properties that make it particularly useful in probability theory. For another application, see Theorem 5.21 below.

Let us now bring this introductory digression to a close and return to probability theory. We begin with the following definition.

Definition 4.14

(i) Let μ be a finite complex measure[a] on $\mathbb{R}^d$. Define the *Fourier transform of* μ, denoted by $\widehat{\mu} : \mathbb{R}^d \to \mathbb{C}$, through

$$\widehat{\mu}(\xi) := \int e^{i\xi\cdot x}\, \mu(dx)\,.$$

(ii) Let X be a real-valued random variable. Define the *characteristic function of* X, denoted by $\Phi_X : \mathbb{R}^d \to \mathbb{C}$, as the Fourier transform of its law $\mathbb{P}_X$. That is,

$$\Phi_X(\xi) = \widehat{\mathbb{P}}_X(\xi) = \int e^{i\xi\cdot x}\, \mathbb{P}_X(dx) = \mathbb{E}[e^{i\xi\cdot X}]\,.$$

[a] This means $\mu = \mu_1 + i\mu_2$, where μ_1 and μ_2 are signed measures of finite total variation.

By dominated convergence, $\Phi_X \in C_b(\mathbb{R}^d)$.

The most important observation in all of Fourier analysis is the following computation for a Gaussian. For $\sigma > 0$, define

$$g_\sigma(x) := \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}\,, \tag{4.8}$$

the density of the Gaussian law with mean zero and variance σ^2.

Proposition 4.15 *Let $X \in \mathbb{R}$ be a Gaussian random variable with law $g_\sigma(x)\,dx$. Then*

$$\Phi_X(\xi) = e^{-\frac{\sigma^2}{2}\xi^2}\,.$$

Proof By definition,

$$\Phi_X(\xi) = \int \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}\, e^{i\xi x}\, dx\,.$$

By the change of variables $x \mapsto \sigma x$, we may suppose that $\sigma = 1$ and compute

$$f(\xi) := \int \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, e^{i\xi x}\, dx\,.$$

Differentiating under the integral and then integrating by parts, we find

$$\begin{aligned} f'(\xi) &= \int \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}\, \mathrm{i}x\, \mathrm{e}^{\mathrm{i}\xi x}\, \mathrm{d}x \\ &= \int \frac{1}{\sqrt{2\pi}} (-\mathrm{i})\partial_x \left(\mathrm{e}^{-\frac{x^2}{2}}\right) \mathrm{e}^{\mathrm{i}\xi x}\, \mathrm{d}x \\ &= \int \frac{1}{\sqrt{2\pi}} (-1) \left(\mathrm{e}^{-\frac{x^2}{2}}\right) \xi\, \mathrm{e}^{\mathrm{i}\xi x}\, \mathrm{d}x \\ &= -\xi f(\xi)\,. \end{aligned}$$

Thus, f satisfies the ordinary differential equation

$$\begin{cases} f(0) = 1 \\ f'(\xi) = -\xi f(\xi)\,. \end{cases}$$

As seen in analysis (since f' is a Lipschitz continuous function of f), this equation has a unique solution, $f(\xi) = \mathrm{e}^{-\frac{\xi^2}{2}}$. □

Thanks to the preceding computation, we can *invert* the Fourier transform in the following sense. For simplicity, set $d = 1$; the case $d > 1$ is done in exactly the same way.

Since the measure μ can be quite rough (it need not have a density), it is very helpful to *mollify*[5] it by convolving (recall Theorem 3.21 and Theorem 3.22) it with the smooth function (4.8). This convolution has density

$$f_\sigma(x) := \int g_\sigma(x - y)\, \mu(\mathrm{d}y)\,. \tag{4.9}$$

Lemma 4.16 (Fourier inversion formula for measures) *For any finite complex measure μ on $\mathbb{R}$, we have*

$$f_\sigma(x) = \frac{1}{2\pi} \int \mathrm{e}^{-\mathrm{i}\xi x}\, \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2}\, \widehat{\mu}(\xi)\, \mathrm{d}\xi\,. \tag{4.10}$$

Proof By Theorem 4.15 with σ replaced by $1/\sigma$, we have

$$\sigma\sqrt{2\pi} g_\sigma(x) = \mathrm{e}^{-\frac{x^2}{2\sigma^2}} = \int \mathrm{e}^{\mathrm{i}\xi x}\, g_{1/\sigma}(\xi)\, \mathrm{d}\xi\,.$$

Hence,

$$\begin{aligned} f_\sigma(x) &= \int g_\sigma(x - y)\, \mu(\mathrm{d}y) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int \int \mathrm{e}^{\mathrm{i}\xi(x-y)}\, g_{1/\sigma}(\xi)\, \mathrm{d}\xi\, \mu(\mathrm{d}y) \\ &= \frac{1}{2\pi} \int \int \mathrm{e}^{\mathrm{i}\xi(x-y)}\, \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2}\, \mathrm{d}\xi\, \mu(\mathrm{d}y) \\ &= \frac{1}{2\pi} \int \mathrm{e}^{\mathrm{i}\xi x}\, \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2} \int \mathrm{e}^{-\mathrm{i}\xi y} \mu(\mathrm{d}y)\, \mathrm{d}\xi \\ &= \frac{1}{2\pi} \int \mathrm{e}^{\mathrm{i}\xi x}\, \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2}\, \widehat{\mu}(-\xi)\, \mathrm{d}\xi\,, \end{aligned}$$

where in the fourth step we used Fubini's theorem. The claim follows by the change of variables $\xi \mapsto -\xi$. □

[5] Note that the function g_σ is (the density of) an approximate delta function (recall Theorem 4.8 (iii)).

Remark 4.17 If the measure μ is sufficiently regular, then the Fourier inversion formula takes on a simpler form because one can take the limit $\sigma \to 0$ and hence get rid of the mollifiers g_σ. Suppose that $\mu(\mathrm{d}x) = f(x)\,\mathrm{d}x$ has a continuous density f that also satisfies $\widehat{f} := \widehat{\mu} \in L^1$. (The latter condition is true provided that f is smooth enough.) Then by taking $\sigma \to 0$ in (4.10), using Theorem 4.8 (iii) on the left-hand side and dominated convergence on the right-hand side, we find the *Fourier inversion formula for regular functions*

$$f(x) = \frac{1}{2\pi} \int \mathrm{e}^{-\mathrm{i}\xi x}\, \widehat{f}(\xi)\,\mathrm{d}\xi\,,$$

where we recall that the Fourier transformation is given by

$$\widehat{f}(\xi) = \int \mathrm{e}^{\mathrm{i}\xi x}\, f(x)\,\mathrm{d}x\,.$$

Therefore, inverse Fourier transformation is, up to a sign in the argument, simply Fourier transformation itself! Compare these expressions to the finite-dimensional ones from (4.7).

The characteristic function provides yet another, extremely useful, equivalent criterion for convergence in law of random variables (to complement Propositions 4.11 and 4.12) – pointwise convergence of the characteristic function.

Proposition 4.18 *Let μ_n and μ be probability measures on $\mathbb{R}^d$. Then $\mu_n \xrightarrow{\mathrm{w}} \mu$ if and only if $\widehat{\mu}_n(\xi) \to \widehat{\mu}(\xi)$ for all $\xi \in \mathbb{R}^d$.*

Proof The "only if" implication is obvious by definition of weak convergence, since the real and imaginary parts of the function $x \mapsto \mathrm{e}^{\mathrm{i}\xi\cdot x}$ are continuous and bounded for all $x \in \mathbb{R}^d$.

To prove the "if" implication, we again suppose for simplicity that $d = 1$ (the case $d > 1$ is very similar). Suppose therefore that $\widehat{\mu}_n(\xi) \to \widehat{\mu}(\xi)$ for all $\xi \in \mathbb{R}^d$. For $\varphi \in C_c(\mathbb{R})$ we have, by Fubini's theorem,

$$\int g_\sigma * \varphi\,\mathrm{d}\mu = \int \varphi(x)\,(g_\sigma * \mu)(x)\,\mathrm{d}x\,.$$

The function $g_\sigma * \mu$ is simply (4.9), so that Theorem 4.16 yields

$$\int g_\sigma * \varphi\,\mathrm{d}\mu = \int \varphi(x)\,\frac{1}{2\pi} \int \mathrm{e}^{-\mathrm{i}\xi x}\, \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2}\, \widehat{\mu}(\xi)\,\mathrm{d}\xi\,\mathrm{d}x\,.$$

An analogous formula holds for μ_n. By dominated convergence, for any $\sigma > 0$ we have

$$\int \mathrm{e}^{-\mathrm{i}\xi x}\, \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2}\, \widehat{\mu}_n(\xi)\,\mathrm{d}\xi \longrightarrow \int \mathrm{e}^{-\mathrm{i}\xi x}\, \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2}\, \widehat{\mu}(\xi)\,\mathrm{d}\xi$$

as $n \to \infty$ for all x, so that another application of dominated convergence (to the integral over x) yields, for all $\varphi \in C_c$,

$$\int g_\sigma * \varphi\,\mathrm{d}\mu_n \longrightarrow \int g_\sigma * \varphi\,\mathrm{d}\mu \tag{4.11}$$

as $n \to \infty$.

To conclude the argument, we define the space of functions

$$H := \{g_\sigma * \varphi : \sigma > 0, \varphi \in C_c\}\,.$$

If we can prove that the closure of H under $\|\cdot\|_\infty$ contains C_c, then the proof will be complete by applying Theorem 4.11 to (4.11).

What remains, therefore, is to prove that the closure of H under $\|\cdot\|_\infty$ contains C_c. To that end, choose $\varphi \in C_c$ and estimate

$$\begin{aligned}\|g_\sigma * \varphi - \varphi\|_\infty &= \sup_x \left| \int \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2\sigma^2}} \big(\varphi(x-y) - \varphi(x)\big)\, \mathrm{d}y \right| \\ &= \sup_x \left| \int \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \big(\varphi(x-\sigma y) - \varphi(x)\big)\, \mathrm{d}y \right| .\end{aligned}$$

Now let $\varepsilon > 0$ and choose $K > 0$ such that

$$\int_{|y|>K} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}}\, \mathrm{d}y \leqslant \frac{\varepsilon}{\|\varphi\|_\infty} .$$

Splitting the y-integration into $|y| \leqslant K$ and $|y| > K$, we conclude that

$$\|g_\sigma * \varphi - \varphi\|_\infty \leqslant \sup_x \left| \int_{|y|\leqslant K} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \big(\varphi(x-\sigma y) - \varphi(x)\big)\, \mathrm{d}y \right| + 2\varepsilon .$$

On the support of the integral, the vector σy has norm bounded by σK, so that by uniform continuity of φ we deduce that the right-hand side converges to 2ε as $\sigma \to 0$. This concludes the proof. □

4.4 The central limit theorem

The *central limit theorem* is, together with the law of large numbers, the second most fundamental result in probability. It states that the sum of a large number of independent identically distributed random variables has approximately a Gaussian distribution, no matter what the distribution of these variables is. This provides at least a partial theoretical justification[6] for the ubiquity of the Gaussian distribution in probability and statistics. This represents the first instance of a remarkable phenomenon in probability and statistical physics called *universality*: if you take a complicated system made up of many small parts, the behaviour of the system on large scales is *universal* in the sense that it does not depend on the details of the individual parts[7]. In this instance, the universal behaviour is the Gaussian distribution of the sum, no matter the distribution of the individual random variables.

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed real-valued random variables in L^1. The strong law of large numbers states that

$$\frac{1}{n}(X_1 + \cdots + X_n) \longrightarrow \mathbb{E}[X_1]$$

almost surely as $n \to \infty$. It is natural to ask how *fast* this convergence takes place, i.e. what is the typical size, or scale, of $\frac{1}{n}(X_1 + \cdots + X_n) - \mathbb{E}[X_1]$, as a function of n.

For $X_1 \in L^2$, the answer is easy. Indeed, since

$$\mathbb{E}\big[(X_1 + \cdots + X_n - n\mathbb{E}[X_1])^2\big] = \mathrm{Var}(X_1 + \cdots + X_n) = n\,\mathrm{Var}(X_1) ,$$

we find that

$$\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n - n\mathbb{E}[X_1]) \tag{4.12}$$

[6] Another, perhaps more pragmatic, justification is that, if one does not know the distribution of a random variable one is considering, we have no choice but to guess, and the Gaussian is a particularly convenient guess. Even if this is not correct, in many applications the Gaussian is a good enough approximation.

[7] As a consequence, some very complicated systems admit a remarkably simple emergent effective description on large scales, although the full analysis of their individual components is hopelessly complicated. An example is the derivation of the emergent laws of hydrodynamics from a microscopic theory of matter. This idea is also famously at the core of Isaac Asimov's Foundation trilogy.

is typically of order one (since the expectation of its square is equal to $\mathrm{Var}(X_1)$, which does not depend on n).

The central limit theorem is a more precise version of this observation, as it even identifies the limiting law of (4.12).

Proposition 4.19 (Central limit theorem) *Let $X_1, X_2, \dots$ be independent identically distributed random variables in L^2, with variance σ^2. Then, as $n \to \infty$, the quantity (4.12) converges in law to a Gaussian random variable with mean zero and variance σ^2.*

Proof Using the technology of characteristic functions developed in the previous section, the proof is remarkably straightforward. First, without loss of generality we may suppose that $\mathbb{E}[X_1] = 0$ (otherwise just replace X_n with $X_n - \mathbb{E}[X_n]$).

We shall use that for any random variable $X \in L^2$ we have[8]

$$\Phi_X(\xi) = 1 + \mathrm{i}\xi\mathbb{E}[X] - \frac{1}{2}\xi^2\mathbb{E}[X^2] + o(\xi^2) \tag{4.13}$$

as $\xi \to 0$. To show (4.13), we differentiate under the expectation, using that $X \in L^2$, to obtain

$$\Phi'_X(\xi) = \mathrm{i}\,\mathbb{E}[X\,\mathrm{e}^{\mathrm{i}\xi X}]\,,$$

and differentiating again yields

$$\Phi''_X(\xi) = -\mathbb{E}[X^2\,\mathrm{e}^{\mathrm{i}\xi X}]\,.$$

Note that differentiating inside the expectation is allowed since $X \in L^2$. By Taylor's theorem, we therefore have

$$\begin{aligned}\Phi_X(\xi) &= 1 + \mathrm{i}\,\mathbb{E}[X]\,\xi - \int_0^\xi \mathbb{E}[X^2\,\mathrm{e}^{\mathrm{i}tX}]\,(\xi - t)\,\mathrm{d}t \\ &= 1 + \mathrm{i}\,\mathbb{E}[X]\,\xi - \frac{1}{2}\xi^2\mathbb{E}[X^2] - \int_0^\xi \mathbb{E}[X^2\,(\mathrm{e}^{\mathrm{i}tX} - 1)]\,(\xi - t)\,\mathrm{d}t\,.\end{aligned}$$

The expectation under the last integral tends to zero as $t \to 0$, by the dominated convergence theorem. Hence, the whole integral is $o(\xi^2)$, and we obtain (4.13).

With $Z_n := \frac{X_1 + \cdots + X_n}{\sqrt{n}}$ we have, by independence of the variables $X_1, \dots, X_n$,

$$\Phi_{Z_n}(\xi) = \mathbb{E}\left[\exp\left(\mathrm{i}\xi\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)\right] = \mathbb{E}[\exp(\mathrm{i}\xi X_1/\sqrt{n})]^n = \Phi_{X_1}(\xi/\sqrt{n})^n\,.$$

By (4.13), we therefore get, for any $\xi \in \mathbb{R}$,

$$\Phi_{Z_n}(\xi) = \left(1 - \frac{\sigma^2\xi^2}{2n} + o\left(\frac{\xi^2}{n}\right)\right)^n \longrightarrow \mathrm{e}^{-\frac{\sigma^2}{2}\xi^2}$$

as $n \to \infty$. The claim now follows from Propositions 4.15 and 4.18. □

[8] Here we recall the "little-o" notation for some complex-valued function f and nonnegative function g: "$f(\xi) = o(g(\xi))$ as $\xi \to 0$" means that $\lim_{\xi\to 0} \frac{f(\xi)}{g(\xi)} = 0$; informally: "$f$ is much smaller than g". Contrast this to the "big-O" notation: "$f(\xi) = O(g(\xi))$" means that $\frac{|f(\xi)|}{g(\xi)} \leqslant C$ for some constant C independent of ξ; informally: "f is not much larger than g".

CHAPTER 5

Markov chains and random walks

WEEK 8

Markov chains are some of the most useful and fundamental objects in probability theory. They constitute the paradigm of a random dynamical system, and as such have innumerable applications from physics to biology and economics[1]. Informally, a Markov chain is a *stochastic process*, i.e. a random variable depending on time, whereby the law of the future of the process depends only on the present and not the whole past. In other words, the process has no memory: knowing its entire history gives me no useful information for predicting the future as compared to knowing just its value today.

Before moving on, let us briefly unwrap the notion of a stochastic process. In this course, time is always *discrete*, i.e. an integer $n \in \mathbb{N}$. (Think n labelling days starting from some arbitrary starting point.) A stochastic process is a family of random variables $(X_n)_{n\in\mathbb{N}}$ taking values in some set S (which can be, for instance, $\mathbb{R}^d$, $\mathbb{Z}^d$, or some finite set). In other words, a stochastic process on the probability space Ω is a function

$$X : \mathbb{N} \times \Omega \to S\,,$$

such that $X_n(\cdot)$ is measurable for each $n \in \mathbb{N}$. It is helpful to note that a stochastic process can be though of in two different ways.

- As a collection of random variables. Here, one chooses a time $n \in \mathbb{N}$ and regards $\omega \mapsto X_n(\omega)$ as a function of $\omega \in \Omega$.
- As a random collection of trajectories. Here, one chooses a realization $\omega \in \Omega$ and regards $n \mapsto X_n(\omega)$ as a fixed S-valued sequence.

5.1 Definition and basic properties

Throughout this chapter, we fix a set S, which we always assume to be *discrete*, i.e. finite or countable. A prominent example of the latter is $S = \mathbb{Z}^d$.

Definition 5.1

(i) A stochastic process $(X_n)_{n\in\mathbb{N}}$ is a *Markov chain*, or a *Markov process*, if for any $n \in \mathbb{N}$ and $x_0, \dots, x_n, y \in S$ we have

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n, \dots, X_0 = x_0) = \mathbb{P}(X_{n+1} = y \mid X_n = x_n)\,. \tag{5.1}$$

(ii) The Markov chain is *homogeneous* if the function

$$(x, y) \mapsto \mathbb{P}(X_{n+1} = y \mid X_n = x)$$

does not depend on $n \in \mathbb{N}$.

[1] Applications of Markov chains represents a good proprtion of all of science, and as such we do not even attempt an overview here. Some famous examples are Markov chain Monte Carlo (see Section 5.9 below) and Google's PageRank algorithm.

Throughout this chapter, we shall always and without further mention suppose that (X_n) is a *homogeneous Markov chain*.

Theorem 5.1 says informally that, conditioned on the past $X_0, \dots, X_{n-1}, X_n$, the law of the future X_{n+1} depends only on the present X_n. The homogeneous property says that this law does not depend on the time n.

Definition 5.2 We define the *transition matrix* $Q : S \times S \to [0,1]$ of the chain (X_n) through

$$Q(x,y) := \mathbb{P}(X_1 = y \mid X_0 = x)\,.$$

By homogeneity, we have $Q(x,y) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$ for all $n \in \mathbb{N}$. The number $Q(x,y)$ is therefore the probability of going from x to y in one time step of the chain.

The following remark follows immediately from Theorem 5.2.

Remark 5.3 The transition matrix Q satisfies the follows properties.

(i) For all $x, y \in S$ we have $Q(x,y) \in [0,1]$.
(ii) For all $x \in S$ we have $\sum_{y \in S} Q(x,y) = 1$.

Definition 5.4 A matrix Q satisfying the properties (i) and (ii) from Theorem 5.3 is called *stochastic*.

Proposition 5.5 *Let Q be the transition matrix of the chain (X_n). Then Q and the law of X_0 fully determine the law of the entire process (X_n), through the formula*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_0 = x_0)Q(x_0,x_1)\cdots Q(x_{n-1},x_n)\,. \tag{5.2}$$

Proof By definition of conditional expectation and by Theorem 5.1, we get

$$\begin{aligned}
&\mathbb{P}(X_n = x_n, \dots, X_0 = x_0) \\
&\quad = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0)\,\mathbb{P}(X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\
&\quad = Q(x_{n-1},x_n)\,\mathbb{P}(X_{n-1} = x_{n-1}, \dots, X_0 = x_0)\,.
\end{aligned}$$

Repeatedly applying the same argument to the second term on the right-hand side yields the claim. □

Definition 5.6 The law of X_0 is called the *initial distribution* of the chain (X_n).

Remark 5.7 Conversely, given a probability measure μ and a stochastic matrix Q, can we construct a Markov chain with initial distribution μ and transition matrix Q? The answer is yes.
For any finite time $N \in \mathbb{N}$, we can define the law of the vector $(X_0, \dots, X_N)$ as in (5.2):

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_N = x_N) := \mu(x_0)Q(x_0,x_1)\cdots Q(x_{N-1},x_N)\,.$$

Because Q is stochastic, the right-hand side is indeed a probability measure on S^n, and moreover for any $n \leqslant N$ the relation (5.2) holds. Hence, for any $n \leqslant N-1$, we have

$$\begin{aligned}\mathbb{P}(X_{n+1} = y \mid X_0 = x_0, \dots, X_n = x_n) &= \frac{\mathbb{P}(X_{n+1} = x_{n+1}, \dots, X_0 = x_0)}{\mathbb{P}(X_n = x_n, \dots, X_0 = x_0)} \\ &= \frac{\mu(x_0)Q(x_0, x_1)\cdots Q(x_{n-1}, x_n)Q(x_n, y)}{\mu(x_0)Q(x_0, x_1)\cdots Q(x_{n-1}, x_n)} \\ &= Q(x_n, y)\,.\end{aligned}$$

Therefore, the process $X_0, X_1, \dots, X_N$ satisfies (5.1) up to time N. This construction can be extended to infinite times with a bit more work (e.g. using Kolmogorov's extension theorem), which we shall not go into here.

Summarizing, one can equivalently consider

(i) a *Markov chain* (as in Theorem 5.1), or
(ii) its *initial distribution* and its *transition matrix*.

The latter two in (ii) are often easier to work with.

Example 5.8 Consider the following very simple weather model. Let $p, q \in [0,1]$. Suppose that a day is either dry (D) or rainy (R). If today is rainy, then with probability p tomorrow is rainy. If today is dry, then with probability q tomorrow is dry. This specifies the transition matrix Q on the state space $\{\mathrm{R}, \mathrm{D}\}$ through

$$Q(\mathrm{R},\mathrm{R}) = p\,, \qquad Q(\mathrm{R},\mathrm{D}) = 1-p\,, \qquad Q(\mathrm{D},\mathrm{D}) = q\,, \qquad Q(\mathrm{D},\mathrm{R}) = 1-q\,.$$

It is very convenient to interpret $Q \in [0,1]^{S\times S}$ literally as a matrix and use the usual notations for matrix multiplication, according to the following definition.

Definition 5.9 Let Q be a stochastic matrix on S.

(i) For $n \in \mathbb{N}$ we define the matrix $Q^n \in [0,1]^{S\times S}$ through $Q^0(x,y) = \delta_{xy}$, $Q^1(x,y) := Q(x,y)$, and

$$Q^{n+1}(x,y) := \sum_{z\in S} Q^n(x,z)Q(z,y)$$

by recurrence.

(ii) For a bounded function $f : S \to \mathbb{R}$ we define $Qf(x) := \sum_{y\in S} Q(x,y)f(y)$.

(iii) For a probability measure $\mu : S \to [0,1]$ we define $\mu Q(x) := \sum_{y\in S} \mu(y)Q(y,x)$.

Hence, according to the conventions of linear algebra, we interpret Q as an $S \times S$ matrix, f as an S-dimensional column vector, and μ as an S-dimensional row vector.

Note that if μ is a probability measure then so is μQ^n for any $n \in \mathbb{N}$. Moreover, by (5.2) we have

$$\mathbb{P}(X_n = x_n \mid X_0 = x_0) = Q^n(x_0, x_n)\,.$$

Thus, the matrix Q^n has the interpretation of the n-step transition matrix of the Markov chain.

For a Markov chain (X_n) with initial distribution μ and transition matrix Q, in matrix notation we can for instance compute

$$\begin{aligned}\mathbb{E}[f(X_n)] = \sum_{x_n \in S} f(x_n)\mathbb{P}(X_n = x_n) &= \sum_{x_n \in S}\sum_{x_0 \in S} f(x_n)\mathbb{P}(X_n = x_n \mid X_0 = x_0)\mathbb{P}(X_0 = x_0) \\ &= \sum_{x_n \in S}\sum_{x_0 \in S} f(x_n)Q^n(x_0, x_n)\mu(x_0) = \mu Q^n f\,.\end{aligned}$$

Example 5.10 (Theorem 5.8 continued) For the Markov chain from Theorem 5.8 we can use the matrix notation to write

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

as an $\{\mathrm{R}, \mathrm{D}\} \times \{\mathrm{R}, \mathrm{D}\}$ matrix.

We conclude this section with a few more examples.

Example 5.11 (Random walk on $\mathbb{Z}^d$) Let $d \in \mathbb{N}^*$ and ν a probability measure on $\mathbb{Z}^d$. Define the stochastic matrix $Q(x, y) := \nu(y - x)$. A Markov chain with transition matrix Q is called a *random walk* on $\mathbb{Z}^d$. The interpretation is that at each step of the walk, the chain takes a random step from its current location, with the law of the step being given by ν. The random walk can also be explicitly written as a sum

$$X_n = X_0 + \sum_{i=1}^{n} Z_i\,, \tag{5.3}$$

where $Z_1, Z_2, \dots$ are independent random variables with law ν, independent of X_0. Indeed, to verify Theorem 5.1, we simply note that (5.3) satisfies

$$\begin{aligned}\mathbb{P}(X_{n+1} = y \mid X_n = x_n, \dots, X_0 = x_0) &= \mathbb{P}(X_n + Z_{n+1} = y \mid X_n = x_n, \dots, X_0 = x_0) \\ &= \mathbb{P}(Z_{n+1} = y - x) \\ &= Q(x, y)\,,\end{aligned}$$

by independence of Z_{n+1} and $(X_0, Z_1, \dots, Z_n)$.
If

$$\nu(x) = \frac{1}{2d}\mathbf{1}_{|x|=1}\,, \tag{5.4}$$

then the walk is called *simple* (the walk jumps to the nearest neighbours with equal probability).

Example 5.12 (Knight's random walk) Random walks can take place in a more general setting than the lattice $\mathbb{Z}^d$. As an example, let $S = \{1, \dots, 8\}^2$ be a chessboard. For each square $x \in S$ denote by $K(x) \subset S$ the set of squares of the board that can be reached by a single move of a knight from x. Then we can define the "knight's random walk" as the Markov chain with transition matrix

$$Q(x, y) = \frac{1}{|K(x)|}\mathbf{1}_{y \in K(x)}\,.$$

Thus, at each step, the knight moves uniformly at random to any square it can reach.

Example 5.13 (Simple random walk on general graph) The previous example can be greatly generalised as follows. Let (S, E) be a connected graph on the vertex set S such that the degree $D_x := |\{e \in E : x \in e\}|$ of each vertex $x \in S$ is finite (such a graph is called *locally finite*). Define the stochastic matrix Q on $S \times S$ through

$$Q(x, y) := \frac{1}{D_x} \mathbf{1}_{\{x,y\} \in E} .$$

Here, at each step the walk moves uniformly at random along an incident edge.

Example 5.14 (Ehrenfest model of diffusion) Here is a primitive model of diffusion names after its inventors, Tatiana and Paul Ehrenfest. Suppose that we have a container containing N gas molecules. The container is split in two halves separated by a small hole. Sometimes molecules will travel from one half to the other. We are interested in the number x of molecules in the left half. We take the following, rather naive, model of diffusion: at each time step we choose one molecule uniformly at random and bring it to the other side of the container.
Thus, we set $S = \{0, 1, \dots, N\}$ and define the transition matrix

$$Q(x, y) = \begin{cases} \frac{N-x}{N} & \text{if } y = x + 1 \\ \frac{x}{N} & \text{if } y = x - 1 \\ 0 & \text{otherwise} . \end{cases}$$

5.2 The Markov property

In this section we introduce a result, often simply referred to the *Markov property*, which will be our main tool in studying Markov chains. It is a generalisation of the condition (5.1) from Theorem 5.1.

Proposition 5.15 (Markov property) *Let $n, m \in \mathbb{N}$. Let f be a nonnegative function on S^{m+1} and $A \subset S^{n+1}$. Then for any $x \in S$ we have*

$$\mathbb{E}[f(X_n, \dots, N_{n+m}) \mid X_n = x, (X_0, \dots, X_n) \in A] = \mathbb{E}[f(X_0, \dots, X_m) \mid X_0 = x] .$$

In words, the Markov property says that the law of the entire future of the process, conditioned on the entire past and being at x today, is the same as the law of a process simply starting at x at time zero. In other words, what happened before the present does not matter. This is a strong manifestation of the memoryless and homogeneity properties of Markov chains.

Proof of Theorem 5.15 For $x_0, \dots, x_{n-1} \in S$, we get from Theorem 5.1

$$\begin{aligned} &\mathbb{E}[f(X_n, \dots, X_{n+m}) \mid X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_n = x] \\ &= \sum_{y_1, \dots, y_m \in S} f(x, y_1, \dots, y_m) \\ &\quad \times \mathbb{P}(X_{n+m} = y_m, \dots, X_{n+1} = y_1 \mid X_0 = x_0, \dots, X_{n-1} = x_{n-1}, X_n = x) \\ &= \sum_{y_1, \dots, y_m \in S} f(x, y_1, \dots, y_m) \, \mathbb{P}(X_{n+m} = y_m, \dots, X_{n+1} = y_1 \mid X_n = x) \\ &= \mathbb{E}[f(X_0, \dots, X_m) \mid X_0 = x] . \end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[f(X_n, \dots, N_{n+m})\mathbf{1}_{X_n=x,(X_0,\dots,X_n)\in A}] \\
&= \sum_{x_0,\dots,x_{n-1}\in S} \mathbf{1}_{(x_0,\dots,x_{n-1},x)\in A}\, \mathbb{E}[f(X_n, \dots, N_{n+m})\mathbf{1}_{X_0=x_1,\dots,X_{n-1}=x_{n-1},X_n=x}] \\
&= \sum_{x_0,\dots,x_{n-1}\in S} \mathbf{1}_{(x_0,\dots,x_{n-1},x)\in A}\, \mathbb{E}[f(X_n, \dots, N_{n+m}) \mid X_0 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = x] \\
&\quad \times \mathbb{P}(X_0 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = x) \\
&= \sum_{x_0,\dots,x_{n-1}\in S} \mathbf{1}_{(x_0,\dots,x_{n-1},x)\in A}\, \mathbb{E}[f(X_0, \dots, N_m) \mid X_0 = x] \\
&\quad \times \mathbb{P}(X_0 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = x) \\
&= \mathbb{E}[f(X_0, \dots, N_m) \mid X_0 = x]\, \mathbb{P}(X_n = x, (X_0, \dots, X_n) \in A)\,,
\end{aligned}$$

and the claim follows after dividing by $\mathbb{P}(X_n = x, (X_0, \dots, X_n) \in A)$. □

The Markov property can be rewritten in the following form, which is sometimes useful. At a first reading, I suggest that you skip over it and return to it when you read the proof of Theorem 5.45 where it is used.

Corollary 5.16 *Let $n, m \in \mathbb{N}$. Let f be a nonnegative function on S^{m+1} and g a nonnegative function on S^{n+1}. Then for any $x \in S$ we have*

$$\begin{aligned}
&\mathbb{E}[f(X_n, \dots, N_{n+m})\, g(X_0, \dots, X_n) \mid X_n = x] \\
&\qquad = \mathbb{E}[f(X_0, \dots, X_m) \mid X_0 = x]\, \mathbb{E}[g(X_0, \dots, X_n) \mid X_n = x_n]\,.
\end{aligned}$$

Proof Abbreviating $x_n := x$, we obtain from Theorem 5.15

$$\begin{aligned}
&\sum_{x_0,\dots,x_{n-1}\in S} g(x_0, \dots, x_n)\, \mathbb{E}[f(X_n, \dots, N_{n+m})\, \mathbf{1}_{X_0=x_0} \cdots \mathbf{1}_{X_{n-1}=x_{n-1}} \mid X_n = x_n] \\
&= \sum_{x_0,\dots,x_{n-1}\in S} g(x_0, \dots, x_n)\, \mathbb{E}[f(X_n, \dots, N_{n+m}) \mid X_0 = x_0, \dots, X_n = x_n] \\
&\qquad \times \mathbb{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1} \mid X_n = x_n) \\
&= \sum_{x_0,\dots,x_{n-1}\in S} g(x_0, \dots, x_n)\, \mathbb{E}[f(X_0, \dots, N_m) \mid X_0 = x_n] \\
&\qquad \times \mathbb{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1} \mid X_n = x_n) \\
&= \mathbb{E}[f(X_0, \dots, N_m) \mid X_0 = x_n]\, \mathbb{E}[g(X_0, \dots, X_n) \mid X_n = x_n]\,.
\end{aligned}$$

□

5.3 Recurrence and transience

WEEK 9

Suppose that a Markov chain starts from state $x \in S$. How long does it take for it to return to x? How often will it return? These questions are two of the most central ones in the study of Markov chains, and lead to the notions of recurrence and transience of Markov chains.

To study these questions, we introduce two fundamental variables. In their definition, we always use the convention that $\inf \emptyset = \infty$.

Definition 5.17 Let $x \in S$.

- The time of first visit at x, or first return to x, is
$$H_x := \inf\{n \geqslant 1 : X_n = x\}\,.$$
- The number of visits at x is
$$N_x := \sum_{n \geqslant 0} \mathbf{1}_{X_n = x}\,.$$

We emphasize that H_x cannot be 0, i.e. if the chain starts at x then H_x counts the time of the first return to x. On the other hand, if the chain starts at x then this initial visit is counted in N_x.

Remark 5.18 The random variable H_x is an example of a *stopping time*, i.e. a random time $T \in \mathbb{N}$ such that, for each $n \in \mathbb{N}$, the event $\{T = n\}$ is in $\sigma(X_0, \dots, X_n)$ (i.e. it is determined only by the values of the process up to time n). The intuition is that the decision to stop at a certain time n, i.e. the event $\{T = n\}$, can only be made based on the information $X_0, \dots, X_n$ available up to time n: we cannot see into the future. Indeed, for $T = H_x$ we have

$$\{H_x = n\} = \{X_1 \neq x, \dots, X_{n-1} \neq x, X_n = x\} \in \sigma(X_0, \dots, X_n)\,, \tag{5.5}$$

Definition 5.19 We often use the abbreviations
$$\mathbb{P}_x(\cdot) := \mathbb{P}(\cdot \mid X_0 = x)\,, \qquad \mathbb{P}_\mu(\cdot) := \sum_{x \in S} \mu(x)\, \mathbb{P}_x(\cdot)\,,$$
where μ is a probability measure on S. We denote by $\mathbb{E}_x$ and $\mathbb{E}_\mu$ the corresponding expectations.

The next proposition establishes the crucial dichotomy for the number of visits at x.

Proposition 5.20 *Let $x \in S$.*

(i) If $\mathbb{P}_x(H_x < \infty) = 1$ then $N_x = \infty$ $\mathbb{P}_x$-a.s. In this case x is called recurrent.

(ii) If $\mathbb{P}_x(H_x < \infty) < 1$ then $\mathbb{E}_x[N_x] < \infty$. In this case x is called transient.

Proof The proof is a typical application of the Markov property, Theorem 5.15. To avoid technical issues, we shall first work with the truncated number of visits,

$$N_x^m := \sum_{n=0}^{m} \mathbf{1}_{X_n = x} ,$$

which only depends on a finite number of random variables.

The main idea of the proof is to *condition* on the random time H_x, i.e. to sum over all of its possible values, to compute the probability of at least $k+1$ visits at x. For $k \geqslant 1$ we get

$$\begin{aligned}
\mathbb{P}_x(N_x^m \geqslant k+1) &= \sum_{n \geqslant 1} \mathbb{P}_x(N_x^m \geqslant k+1, H_x = n) \\
&= \sum_{n \geqslant 1} \mathbb{P}_x\Big(\sum_{i=0}^{m} \mathbf{1}_{X_i = x} \geqslant k+1, H_x = n \Big) \\
&= \sum_{n \geqslant 1} \mathbb{P}_x\Big(\sum_{i=n}^{m} \mathbf{1}_{X_i = x} \geqslant k, H_x = n \Big) ,
\end{aligned}$$

where the last step follows from the definition of the time of first return H_x. Using (5.5) we can use Theorem 5.15 to get

$$\begin{aligned}
\mathbb{P}_x(N_x^m \geqslant k+1) &= \sum_{n \geqslant 1} \mathbb{P}_x\Big(\sum_{i=n}^{m} \mathbf{1}_{X_i = x} \geqslant k \,\Big|\, X_1 \neq x, \dots, X_{n-1} \neq x, X_n = x \Big) \mathbb{P}(H_x = n) \\
&= \sum_{n \geqslant 1} \mathbb{P}\Big(\sum_{i=0}^{m-n} \mathbf{1}_{X_i = x} \geqslant k \,\Big|\, X_0 = x \Big) \mathbb{P}(H_x = n) \\
&= \sum_{n \geqslant 1} \mathbb{P}_x(N_x^{m-n} \geqslant k) \mathbb{P}(H_x = n) .
\end{aligned}$$

Next, we let $m \to \infty$. To that end, we note that for any $k \geqslant 1$ the sequence of random variables $\mathbf{1}_{N_x^m \geqslant k}$ is pointwise nondecreasing in m, with limit $\mathbf{1}_{N_x \geqslant k}$. Hence, by monotone convergence (recall Theorem 2.26), we conclude that

$$\mathbb{P}_x(N_x \geqslant k+1) = \sum_{n \geqslant 1} \mathbb{P}_x(N_x \geqslant k) \mathbb{P}(H_x = n) = \mathbb{P}_x(N_x \geqslant k) \mathbb{P}(H_x < \infty) .$$

(Notice that on the right-hand side, we used monotone convergence twice; once for the sum over n and then for the expectation $\mathbb{E}_x$.)

Since $\mathbb{P}_x(N_x \geqslant 1) = 1$ trivially, we conclude by induction that

$$\mathbb{P}_x(N_x \geqslant k) = \mathbb{P}_x(H_x < \infty)^{k-1} . \tag{5.6}$$

Using (5.6), it is now easy to conclude the proof.

- If $\mathbb{P}_x(H_x < \infty) = 1$ then $\mathbb{P}_x(N_x \geqslant k) = 1$ for all k, and hence

$$\mathbb{P}_x(N_x = \infty) = \mathbb{P}_x\Big(\bigcap_{k \geqslant 1} \{N_x \geqslant k\} \Big) = 1 .$$

- If $\mathbb{P}_x(H_x < \infty) < 1$ then

$$\begin{aligned}
\mathbb{E}_x[N_x] &= \sum_{k \geqslant 1} \mathbb{P}_x(N_x \geqslant k) = \sum_{k \geqslant 1} \mathbb{P}_x(H_x < \infty)^{k-1} \\
&= \frac{1}{1 - \mathbb{P}_x(H_x < \infty)} = \frac{1}{\mathbb{P}_x(H_x = \infty)} < \infty .
\end{aligned}$$

□

Informally, we have a dichotomy for a Markov chain starting from x: either the walk returns to x almost surely (recurrent) or there is a positive probability that it never returns to x (transient). Theorem 5.20 gives a simple criterion for analysing the recurrence or transience: check whether the sum

$$\mathbb{E}_x[N_x] = \sum_{n\in\mathbb{N}} \mathbb{P}_x(X_n = x) = \sum_{n\in\mathbb{N}} Q^n(x,x) \tag{5.7}$$

is finite. Thus, the problem reduces to a question about the analysis of the matrix Q. We illustrate this for the simple random walk on $\mathbb{Z}^d$.

Example 5.21 (Theorem 5.11 continued) Let us consider the simple random walk on $\mathbb{Z}^d$ from Theorem 5.11. In principle, one can find explicit combinatorial formulas for $Q^n(x,x)$ and perform an asymptotic analysis to determine the convergence of (5.7). In practice, this can be rather tedious and the analysis depends strongly on the precise transition matrix we are considering.
A far more powerful and versatile approach is to use Fourier analysis (cf. Section 4.3). We observe first that for $x \in \mathbb{Z}^d$ we have

$$\int_{-[\pi,\pi]^d} \frac{\mathrm{d}\xi}{(2\pi)^d}\, \mathrm{e}^{\mathrm{i}\xi\cdot x} = \delta_{x0}\,,$$

as follows by a simple application of Fubini's theorem to evaluate the integral. Hence,

$$\mathbb{P}_0(X_n = 0) = \int_{-[\pi,\pi]^d} \frac{\mathrm{d}\xi}{(2\pi)^d}\, \mathbb{E}_0[\mathrm{e}^{\mathrm{i}\xi\cdot X_n}]$$

By the representation (5.3), we find for the characteristic function of X_n

$$\mathbb{E}_0[\mathrm{e}^{\mathrm{i}\xi\cdot X_n}] = \Phi(\xi)^n\,,$$

where we abbreviated

$$\Phi(\xi) := \mathbb{E}_0[\mathrm{e}^{\mathrm{i}\xi\cdot Z_1}] = \frac{1}{d}\sum_{i=1}^{d} \cos(\xi_i)\,,$$

where the last step follows from by an explicit calculation using the law (5.4) of Z_1. By Fubini's theorem, we conclude that for any $0 < \lambda < 1$ we have

$$\sum_{n\in\mathbb{N}} \lambda^n \mathbb{P}_0(X_n = 0) = \int_{-[\pi,\pi]^d} \frac{\mathrm{d}\xi}{(2\pi)^d} \sum_{n\in\mathbb{N}} \lambda^n \Phi(\xi)^n = \int_{-[\pi,\pi]^d} \frac{\mathrm{d}\xi}{(2\pi)^d} \frac{1}{1-\lambda\Phi(\xi)}\,.$$

We shall now take the limit $\lambda \uparrow 1$. By monotone convergence, the left-hand side converges to (5.7). As for the right-hand side, we note that there exists a constant $c > 0$ such that $\phi(x) \leqslant 1 - c$ if $\xi \in [-\pi.\pi]^d \setminus [-1,1]^d$ and that $\Phi(\xi) \geqslant 0$ for $\xi \in [-1,1]^d$. We now split the integral over $[-\pi.\pi]^d$ into two pieces: $[-\pi.\pi]^d \setminus [-1,1]^d$ and $[-1,1]^d$. We can now take the limit $\lambda \uparrow 1$ by applying dominated convergence to the former piece and monotone convergence to the latter piece. This yields

$$\sum_{n\in\mathbb{N}} \mathbb{P}_0(X_n = 0) = \int_{-[\pi,\pi]^d} \frac{\mathrm{d}\xi}{(2\pi)^d} \frac{1}{1-\Phi(\xi)}\,. \tag{5.8}$$

The denominator of the integral has a singularity at $\xi = 0$, whose integrability we need to analyse. From Taylor's theorem we obtain for $t \in [-1,1]$

$$\cos(t) = 1 - \frac{1}{2}t^2 + \frac{1}{24}s^4$$

for some $0 \leqslant s \leqslant t$, from which we deduce that

$$1 - \frac{1}{2}t^2 \leqslant \cos(t) \leqslant 1 - \frac{11}{24}t^2$$

for $t \in [-1,1]$. Hence,

$$\frac{1}{2d}|\xi|^2 \geqslant 1 - \Phi(\xi) \geqslant \frac{11}{24d}|\xi|^2$$

in a neighbourhood of 0. We conclude that the integral in (5.8) is finite if and only if $d \geqslant 3$.
Summarising: the simple random walk on $\mathbb{Z}^2$ is recurrent if $d \leqslant 2$ and transient for $d \geqslant 3$. Or, as Shizuo Kakutani put it, "A drunk man will find his way home, but a drunk bird may get lost forever[a]."

[a] This is a memorable quote, but you may already spot a flaw in the analogy: the earth is a finite planet and its atmosphere has a finite thickness. Even in an infinite flat earth model, the drunk bird would still be doing an effectively two-dimensional random walk, since it cannot fly into space: the vertical direction is bounded. A more accurate, but admittedly less catchy, version of this quote would therefore be: "A drunk man on an infinite flat earth will find his way home, but a drunk alien in a spacecraft in an infinite universe may get lost forever."

5.4 Stationary and reversible measures

Throughout this section, μ always denotes a measure on S satisfying $\mu(S) > 0$ and $0 \leqslant \mu(x) < \infty$ for all x. In particular, $\mu(S)$ may be infinite.

Definition 5.22 (Stationary measure) Let Q be a stochastic matrix on S. A measure μ is *stationary* with respect to μ if $\mu = \mu Q$. Explicitly, this means that

$$\mu(y) = \sum_{x \in S} \mu(x) Q(x, y) , \qquad \forall y \in S .$$

To explain in more detail the adjective *stationary*, suppose that (X_n) is a Markov chain with initial distribution μ and transition matrix Q, such that μ is stationary with respect to Q. Then the law of X_n is

$$\mathbb{P}_\mu(X_n = x) = \mu Q^n \delta_x = \mu Q^{n-1} \delta_x = \cdots \mu \delta_x = \mu(x)$$

for all $n \in \mathbb{N}$. Hence, the law of X_n is μ for all n. In this sense, the process (X_n) is in the stationary state μ: it is an example of a dynamic process in equilibrium.

Example 5.23 (Theorem 5.11 continued) Consider a random walk on $\mathbb{Z}^d$ with transition matrix $Q(x, y) = \nu(y - x)$, as in Theorem 5.11. The counting measure μ on $\mathbb{Z}^d$ is stationary with respect to Q. (Because $\sum_x Q(x, y) = \sum_x \nu(y - x) = 1$.)

Definition 5.24 (Reversible measure) The measure μ is *reversible* with respect to Q if

$$\mu(x) Q(x, y) = \mu(y) Q(y, x) , \qquad \forall x, y \in S . \tag{5.9}$$

The condition (5.9) is often called *detailed balance*. Interpreting $Q(x, y)$ as the flow of probability from x to y per unit probability in x, the detailed balance condition states that the flow from x to y is the same as the flow from y to x. It is a remarkable and extremely useful condition in both pure mathematics and innumerable applications, for instance in so-called Markov chain Monte Carlo methods (see Section 5.9 below). Its usefulness relies on the following observation.

Remark 5.25 Reversibility is a stronger condition than stationarity (i.e. Theorem 5.24 implies Theorem 5.22). Indeed, if μ is reversible then

$$\sum_x \mu(x)Q(x,y) = \sum_x \mu(y)Q(y,x) = \mu(y)\,.$$

Thus, a simple way to verify that a measure is stationary is to verify that it is reversible. In practice, reversibility is often much easier to check, since it presents a simple local equation that can often be explicitly solved, whereas it can be very difficult to find stationary measures. However, it is worth keeping in mind that there are measures that are stationary but not reversible; see Theorem 5.26.

Remark 5.26 The converse implication is wrong: in general stationarity does not imply reversibility. For instance, in the random walk on $\mathbb{Z}^d$ from Theorem 5.11, if ν is not symmetric, $\nu(-x) \neq \nu(x)$ for some $x \in \mathbb{Z}^d$, then the counting measure μ is stationary (see Theorem 5.23) but not reversible.

Example 5.27 Consider the simple asymmetric random walk on $\mathbb{Z}$, with transition matrix

$$Q(x,x+1) = p\,, \qquad Q(x,x-1) = 1-p\,, \qquad \forall x \in \mathbb{Z}\,,$$

where $0 < p < 1$.
We know from Theorem 5.23 and Theorem 5.26 that the counting measure on $\mathbb{Z}$ is stationary but, if $p \neq 1/2$, not reversible.
Can we find a reversible measure μ? The equation we have solve is

$$\mu(x)Q(x,x+1) = \mu(x+1)Q(x+1,x)\,,$$

i.e.

$$\mu(x)\,p = \mu(x+1)\,(1-p)\,,$$

whose can be solved by induction to yield

$$\mu(x) = \mu(0)\left(\frac{p}{1-p}\right)^x\,. \tag{5.10}$$

This measure coincides with the counting measure if and only if $p = 1/2$, in which case μ is stationary and reversible.
If $p \neq 1/2$, we conclude that Q has two different stationary measures: the counting measure, which is not reversible, and μ, which is reversible.

Example 5.28 (Theorem 5.13 continued) Consider the simple random walk on a connected locally finite graph (S,E). Can we find a stationary measure? It is much easier to look for a reversible measure by solving the detailed balance equation, which can be written as

$$\mu(x)\frac{1}{D_x} = \mu(y)\frac{1}{D_y}$$

for any adjacent x, y. Since the graph is connected, we easily find that μ must be of the form

$$\mu(x) = CD_x$$

for some positive constant C.

Example 5.29 (Theorem 5.14 continued) Let us look for a stationary measure of the Ehrefest model from Theorem 5.14. As before, the best approach is to find a reversible measure by solving the detailed balance equations, which read

$$\mu(x)\frac{N-x}{N} = \mu(x+1)\frac{x+1}{N}\,, \qquad 0 \leqslant x \leqslant N-1\,.$$

This can be solved by induction on x to yield

$$\mu(x) = C\binom{N}{x}\,, \qquad 0 \leqslant x \leqslant N\,.$$

For a probability measure, we take $C = 2^{-N}$. This stationary measure corresponds to the equilibrium measure of the diffusion model. It is strongly concentrated around $x = N/2$, as one would expect based on the physical interpretation of the model.

5.5 The Green function

In this Section we introduce a tool of great power, which bears many names in the literature: the *Green function*, the *potential kernel*, the *fundamental matrix*, ... Green functions appear throughout mathematics; informally, Green functions are always inverses of some basic matrix or operator underlying the problem one is considering (see also Theorem 5.31 (i) below).

Definition 5.30 The *Green function* of a Markov chain is the function $U : S^2 \to [0,\infty]$ defined by

$$U(x,y) := \mathbb{E}_x[N_y]\,,$$

i.e. the expected number of visits in y starting from x.

Remark 5.31

(i) Clearly,

$$U(x,y) = \mathbb{E}_x\left[\sum_{n\in\mathbb{N}} \mathbf{1}_{X_n=y}\right] = \sum_{n\in\mathbb{N}} \mathbb{P}_x(X_n = y) = \sum_{n\in\mathbb{N}} Q^n(x,y)\,.$$

Hence, $U(x,y) > 0$ if and only if there is an $n \in \mathbb{N}$ such that $Q^n(x,y) > 0$. In matrix notation, the Green function is therefore formally given by the Neumann series

$$U = \sum_{n\in\mathbb{N}} Q^n = (I-Q)^{-1}\,.$$

(ii) By Theorem 5.20, x is recurrent if and only if $U(x,x) = \infty$.

Proposition 5.32 *If $x \neq y$ then*

$$U(x,y) = \mathbb{P}_x(H_y < \infty)U(y,y)\,.$$

This is rather intuitive: to count the number of visits at y starting from x, one has to first find the probability of going from x to y and then count the number of visits at y starting from y; in fact, this is an informal summary of the proof given below.

Proof We compute

$$\begin{aligned} U(x,y) &= \mathbb{E}_x[N_y] \\ &= \mathbb{E}_x[N_y \mathbf{1}_{H_y<\infty}] \\ &= \sum_{n,k\in\mathbb{N}} \mathbb{E}_x[\mathbf{1}_{X_k=y}\mathbf{1}_{H_y=n}] \\ &= \sum_{n,\ell\in\mathbb{N}} \mathbb{E}_x[\mathbf{1}_{X_{n+\ell}=y}\mathbf{1}_{H_y=n}] \\ &= \sum_{n,\ell\in\mathbb{N}} \mathbb{P}(X_{n+\ell}=y \mid H_y=n, X_n=y)\,\mathbb{P}_x(H_y=n) \\ &= \sum_{n,\ell\in\mathbb{N}} \mathbb{P}(X_\ell=y \mid X_0=y)\,\mathbb{P}_x(H_y=n) \\ &= U(y,y)\mathbb{P}_x(H_y<\infty)\,, \end{aligned}$$

where in the fourth step we set $k = n + \ell$ and used the definition of H_y, and in the sixth step we used Theorem 5.15. □

By Theorem 5.31 (i), $U(x,y) > 0$ is equivalent to $Q^n(x,y) > 0$ for some $n \in \mathbb{N}$. This means that it is possible to go from x to y with positive probability in finite time. Does this imply that $U(y,x) > 0$? In general, the answer is clearly no (think of an example!). However, if x is recurrent, then the answer is yes.

Proposition 5.33 *Let that $x, y \in S$ such that x is recurrent. If $U(x,y) > 0$ then y is also recurrent and*

$$\mathbb{P}_y(H_x < \infty) = 1\,. \tag{5.11}$$

In particular, by Theorem 5.32, $U(y,x) > 0$.

Proof The main work of the proof is to show (5.11). We start by computing the probability of going from x to y in time n and not returning to x between time n and $n+m$. Using Theorem 5.15, it is

$$\begin{aligned} &\mathbb{P}_x(H_y=n, X_{n+1}\neq x, \dots, X_{n+m}\neq x) \\ &\quad= \mathbb{P}(X_{n+1}\neq x, \dots, X_{n+m}\neq x \mid X_n=y, H_y=n)\,\mathbb{P}_x(H_y=n) \\ &\quad= \mathbb{P}_y(X_1\neq x, \dots, X_m\neq x)\,\mathbb{P}_x(H_y=n)\,. \end{aligned}$$

Taking $m \to \infty$ yields

$$\begin{aligned} \mathbb{P}_y(H_x=\infty)\,\mathbb{P}_x(H_y=n) &= \mathbb{P}_x\left(\{H_y=n\} \cap \bigcap_{k\in\mathbb{N}^*} \{X_{n+k}\neq x\}\right) \\ &\leqslant \mathbb{P}_x(H_y=n, N_x<\infty)\,. \end{aligned}$$

Summing over $n \in \mathbb{N}$ yields

$$\mathbb{P}_y(H_x=\infty)\,\mathbb{P}_x(H_y<\infty) \leqslant \mathbb{P}_x(H_y<\infty, N_x<\infty) \leqslant \mathbb{P}_x(N_x<\infty) = 0\,,$$

by Theorem 5.20. Since $\mathbb{P}_x(H_y < \infty) > 0$ by $U(x,y) > 0$ and Theorem 5.32, we conclude $\mathbb{P}_y(H_x = \infty) = 0$, which is (5.11). In particular, by Theorem 5.32 we also have $U(y,x) > 0$.

What remains is to show that y is recurrent. The idea how to show this is to use $U(x,y) > 0$ and $U(y,x) > 0$ to transport the question of recurrence of y to the recurrence of x. More precisely, because $U(x,y) > 0$ there exists $n \in \mathbb{N}$ such that $Q^n(x,y) > 0$, and because

$U(y,x) > 0$ there exists $m \in \mathbb{N}$ such that $Q^m(y,x) > 0$. Now for any $k \in \mathbb{N}$ we have, by definition of matrix multiplication,

$$Q^{n+k+m}(y,y) \geqslant Q^m(y,x)Q^k(x,x)Q^n(x,y)\,,$$

which implies

$$U(y,y) \geqslant \sum_{k\in\mathbb{N}} Q^{n+k+m}(y,y) \geqslant Q^m(y,x)\left(\sum_{k\in\mathbb{N}} Q^k(x,x)\right)Q^n(x,y)\,.$$

The first and third terms on the right-hand side are strictly positive, while the second term is simply $U(x,x) = \infty$, by recurrence of y. We conclude that $U(y,y) = \infty$ and hence y is recurrent. □

We conclude this section a notion that describes the connectedness of states in a Markov chain. Consider the following somewhat silly example. Take two Markov chains on the disjoint state spaces S_1 and S_2, with transition matrices Q_1 and Q_2 respectively. Combine these two chains into one on the combined the space $S = S_1 \cup S_2$, where the transition matrix is given by

$$Q(x,y) = \begin{cases} Q_1(x,y) & \text{if } x,y \in S_1 \\ Q_2(x,y) & \text{if } x,y \in S_2 \\ 0 & \text{otherwise}\,. \end{cases}$$

These two sub-chains evolve without any knowledge of each other, and one can never go from S_1 to S_2. More precisely, $U(x,y) = 0$ if x and y belong to differents sets S_1 and S_2. Such a chain is *reducible*, in the sense that we can break it apart into two chains without changing anything in its behaviour. The following definition precludes precisely this kind of behaviour.

Definition 5.34 A Markov chain is *irreducible* if $U(x,y) > 0$ for all $x, y \in S$.

Equivalently, the chain is irreducible if for any $x, y \in S$ there exists $n \in \mathbb{N}$ such that $Q^n(x,y) > 0$. In other words, irreducibility means that one can go with positive probability from any state to any other state in finite time.

Remark 5.35 If a chain is irreducible and has a recurrent state, then, by Theorem 5.33, all states are recurrent. In this case we call the whole chain *recurrent*.

Definition 5.36 Let x and y be recurrent states. We say that they *communicate* if $U(x,y) > 0$.

Note that this definition is unambiguous in the sense that the conditions $U(x,y) > 0$ and $U(y,x) > 0$ are equivalent by Theorem 5.33. Thus, x and y communicate if and only if there exists $n \in \mathbb{N}$ such that $Q^n(x,y) > 0$.

5.6 Existence and uniqueness of stationary measures

WEEK 10

In this section we give an explicit general formula for a stationary measure of a recurrent Markov chain, provided that it has a recurrent state. Moreover, under an obviously necessary irreducibility condition, we show that this measure is the unique stationary measure. The construction of this measure is very natural: its value at y is simply the expected number of visits at y during the first excursion from some fixed reference state x back to x.

Proposition 5.37 *Suppose that $x \in S$ is recurrent. Then the measure*

$$\nu_x(y) := \mathbb{E}_x\left[\sum_{k=0}^{H_x-1} \mathbf{1}_{X_k=y}\right]$$

is stationary. Moreover, $\nu_x(y) > 0$ if and only if x and y communicate.

Proof The main idea of the proof is to condition on the value z just before the chain visist y. Since $X_0 = X_{H_x} = x$, we find

$$\begin{aligned}
\nu_x(y) &= \mathbb{E}_x\left[\sum_{k=1}^{H_x} \mathbf{1}_{X_k=y}\right] \\
&= \sum_{z\in S} \mathbb{E}_x\left[\sum_{k=1}^{H_x} \mathbf{1}_{X_{k-1}=z}\mathbf{1}_{X_k=y}\right] \\
&= \sum_{z\in S}\sum_{k\in\mathbb{N}^*} \mathbb{E}_x\left[\mathbf{1}_{k\leqslant H_x}\mathbf{1}_{X_{k-1}=z}\mathbf{1}_{X_k=y}\right] \\
&= \sum_{z\in S}\sum_{k\in\mathbb{N}^*} \mathbb{P}_x\left(k \leqslant H_x, X_{k-1}=z, X_k=y\right) \\
&= \sum_{z\in S}\sum_{k\in\mathbb{N}^*} \mathbb{P}_x\left(X_k=y \mid k\leqslant H_x, X_{k-1}=z\right)\mathbb{P}_x(k\leqslant H_x, X_{k-1}=z) \\
&= \sum_{z\in S}\sum_{k\in\mathbb{N}^*} Q(z,y)\,\mathbb{E}_x\left[\mathbf{1}_{k\leqslant H_x}\mathbf{1}_{X_{k-1}=z}\right] \\
&= \sum_{z\in S} Q(z,y)\,\mathbb{E}_x\left[\sum_{k=1}^{H_x}\mathbf{1}_{X_{k-1}=z}\right] \\
&= \sum_{z\in S} Q(z,y)\,\nu_x(z)\,,
\end{aligned}$$

where in the sixth step we used the Markov property from Theorem 5.15 combined with the remark that the event

$$\{k \leqslant H_x\} = \{X_1 \neq x, \dots, X_{k-1} \neq x\}$$

depends only on the values of X up to time $k-1$. We have shown that $\nu_x = Q\nu_x$, i.e. that ν_x is stationary.

To show the final assertion, note first that if x and y do not communicate then $\mathbb{E}_x[N_y] = U(x,y) = 0$, and hence $\nu_x(y) = 0$.

Next, suppose that x and y communicate, and let us show that $0 < \nu_x(y) < \infty$. First, since there exists $n \in \mathbb{N}$ such that $Q^n(x,y) > 0$, we conclude that

$$1 = \nu_x(x) = \sum_{z\in S} \nu_x(z)Q^n(z,x) \geqslant \nu_x(y)Q^n(y,x)\,,$$

from which we deduce that $\nu_x(y) < \infty$. Second, since there exists $m \in \mathbb{N}$ such that $Q^m(y,x) > 0$, we conclude that

$$\nu_x(y) = \sum_{z \in S} \nu_x(z) Q^m(z,x) \geqslant \nu_x(x) Q^m(y,x) = Q^m(y,x)\,,$$

from which we deduce that $\nu_x(y) > 0$. □

Is the stationary measure ν_x unique, i.e. is it independent of the choice of x? In general, the answer is clearly no. There are two reasons for this, both of which turn out to be "silly". First, if the Markov chain is not irreducible, as in the example preceding Theorem 5.34, choosing x in S_1 or in S_2 will clearly yields two measures that are supported on different disjoint sets. Let us therefore assume that our chain is irreducible. Even then, because of the obvious constraint $\nu_x(x) = 1$ for all x, for each x we get in general a different measure. However, it turns out that this dependence on x is only via a multiplicative positive constant. Up to this constant, the measure ν_x is indeed unique for any irreducible recurrent chain.

Proposition 5.38 *If the Markov chain is irreducible and recurrent, then it has (up to a multiplicative constant in $(0,\infty)$) a unique stationary measure.*

Proof Suppose that μ is a stationary measure. We shall show, by induction on $p \in \mathbb{N}$, that for all $x, y \in S$ we have

$$\mu(y) \geqslant \mu(x)\, \mathbb{E}_x\left[\sum_{k=0}^{p \wedge (H_x - 1)} \mathbf{1}_{X_k = y}\right]. \tag{5.12}$$

Note first that if $x = y$ then (5.12) is trivially true (with an equality) for all p. Suppose therefore that $x \neq y$ and let us show (5.12) by induction on p. By stationarity of μ and

the induction assumption (5.12) for p, we get

$$\begin{aligned}
\mu(y) &= \sum_{z\in S} \mu(z)Q(z,y) \\
&\geqslant \mu(x) \sum_{z\in S} \mathbb{E}_x\left[\sum_{k=0}^{p\wedge(H_x-1)} \mathbf{1}_{X_k=z}\right] Q(z,y) \\
&= \mu(x) \sum_{z\in S} \sum_{k=0}^{p} \mathbb{E}_x\left[\mathbf{1}_{k\leqslant H_x-1}\mathbf{1}_{X_k=z}\right] Q(z,y) \\
&= \mu(x) \sum_{z\in S} \sum_{k=0}^{p} \mathbb{P}_x\left(k \leqslant H_x-1, X_k=z\right) Q(z,y) \\
&= \mu(x) \sum_{z\in S} \sum_{k=0}^{p} \mathbb{P}_x\left(k \leqslant H_x-1, X_k=z\right) \mathbb{P}_x\left(X_{k+1}=y \mid k \leqslant H_x-1, X_k=z\right) \\
&= \mu(x) \sum_{z\in S} \sum_{k=0}^{p} \mathbb{P}_x\left(X_{k+1}=y, k \leqslant H_x-1, X_k=z\right) \\
&= \mu(x) \sum_{z\in S} \sum_{k=0}^{p} \mathbb{E}_x\left[\mathbf{1}_{X_{k+1}=y}\mathbf{1}_{k\leqslant H_x-1}\mathbf{1}_{X_k=z}\right] \\
&= \mu(x)\, \mathbb{E}_x\left[\sum_{k=0}^{p\wedge(H_x-1)} \mathbf{1}_{X_{k+1}=y}\right] \\
&= \mu(x)\, \mathbb{E}_x\left[\sum_{k=1}^{(p+1)\wedge H_x} \mathbf{1}_{X_k=y}\right] \\
&= \mu(x)\, \mathbb{E}_x\left[\sum_{k=0}^{(p+1)\wedge(H_x-1)} \mathbf{1}_{X_k=y}\right],
\end{aligned}$$

where in the fifth step we used the Markov property from Theorem 5.15 combined with

$$\{k \leqslant H_x - 1\} = \{X_1 \neq x, \dots, X_k \neq x\},$$

and in the last step we used that $x \neq y$. We have therefore shown (5.12) for $p+1$, and the proof of (5.12) is hence complete.

Taking $p \to \infty$ in (5.12), we get by monotone convergence

$$\mu(y) \geqslant \mu(x)\, \mathbb{E}_x\left[\sum_{k=0}^{H_x-1} \mathbf{1}_{X_k=y}\right] = \mu(x)\nu_x(y).$$

By stationarity of μ and of ν_x (see Theorem 5.37), we therefore find for any $n \in \mathbb{N}^*$

$$\mu(x) = \sum_{z\in S} \mu(z)Q^n(z,x) \geqslant \sum_{z\in S} \mu(x)\nu_x(z)Q^n(z,x) = \mu(x)\nu_x(x) = \mu(x).$$

The inequality is therefore an equality, and we have

$$\sum_{z\in S} \mu(z)Q^n(z,x) = \sum_{z\in S} \mu(x)\nu_x(z)Q^n(z,x).$$

Since $\mu(z) \geqslant \mu(x)\nu_x(z)$ we conclude that $\mu(z) = \mu(x)\nu_x(z)$ whenever $Q^n(z,x) > 0$. By irreducibility, for any x and z there exists $n \in \mathbb{N}^*$ such that $Q^n(z,x) > 0$. We conclude that

$$\mu = \mu(x)\, \nu_x$$

for any x. □

5.7 Positive and null recurrence

It is now easy to show the following remarkable result about recurrent Markov chains. Recall that recurrent means that $H_x < \infty$ a.s. under $\mathbb{P}_x$. In this case, the expectation $\mathbb{E}_x[H_x]$ may be finite or infinite, which leads to an imporant refinement of the notion of recurrence.

Proposition 5.39 *Consider an irreducible and recurrent Markov chain.*

(i) Either there exists a stationary probability measure μ, in which case $\mathbb{E}_x[H_x] = \frac{1}{\mu(x)}$ for all x;

(ii) or any stationary measure has infinite total mass, in which case we have $\mathbb{E}[H_x] = \infty$ for all x.

Definition 5.40 In case (i) we say that the chain *positive recurrent* and in case (ii) we say that it is *null recurrent*.

Proof of Theorem 5.39 By Theorem 5.38, the stationary measure μ is unique up to a constant, and we can choose it to be either a probability measure (case (i)) or a measure of infinite total mass (case (ii)). Either way, for any $x \in S$ we can write $\mu = C\nu_x$ for some constant C depending on x.

In case (i), we have

$$1 = \mu(S) = C\nu_x(S)\,,$$

which implies $C = \frac{1}{\nu_x(S)}$ and hence

$$\mu(x) = \frac{\nu_x(x)}{\nu_x(S)} = \frac{1}{\nu_x(S)}\,.$$

On the other hand, by Fubini's theorem,

$$\nu_x(S) = \sum_{y\in S} \mathbb{E}_x\left[\sum_{k=0}^{H_x-1} \mathbf{1}_{X_k=y}\right] = \mathbb{E}_x\left[\sum_{k=0}^{H_x-1} 1\right] = \mathbb{E}_x[H_x]\,, \tag{5.13}$$

as claimed.

In case (ii), $\nu_x(S) = \infty$, so that (5.13) implies $\mathbb{E}_x[H_x] = \infty$. □

Clearly, if S is finite then any recurrent chain is always positive recurrent. Thus, null recurrent chains can only occur on an infinite state space.

Example 5.41 (Examples 5.11 and 5.21 continued) In Theorem 5.21, we saw that the simple random walk on $\mathbb{Z}^d$ is recurrent for $d \leqslant 2$. Moreover, in Theorem 5.23, we saw that that counting measure on $\mathbb{Z}^d$ is a stationary measure. Since the total mass of the counting measure on $\mathbb{Z}^d$ is infinite, we conclude that the simple random walk on $\mathbb{Z}^d$ is null recurrent for $d \leqslant 2$.

Let us suppose that an irreducible Markov chain has a stationary probability measure. Theorem 5.39 tells us that if the chain is recurrent then it is in fact positive recurrent. What if we do not a priori know that it is recurrent? It turns out that this assumption is in fact not needed.

Proposition 5.42 *If an irreducible Markov chain has a stationary probability measure, then it is recurrent (and hence positive recurrent).*

Proof Let μ be a stationary probability measure and let $y \in S$ satisfy $\mu(y) > 0$. Then, by Theorem 5.32,

$$\begin{aligned}\mu(S)U(y,y) &= \sum_{x\in S} \mu(x)U(y,y)\\ &\geqslant \sum_{x\in S} \mu(x)U(x,y)\\ &= \sum_{n\in\mathbb{N}}\sum_{x\in S} \mu(x)Q^n(x,y)\\ &= \sum_{n\in\mathbb{N}} \mu(y)\\ &= \infty\,.\end{aligned}$$

Since $\mu(S) = 1$, we conclude that $U(y,y) = \infty$, so that y is recurrent. The claim now follows from Theorem 5.35. □

Note that the existence of a stationary measure with infinite mass does not imply anything about the recurrence of the chain. For instance, in Theorem 5.23 we saw that the simple random walk on $\mathbb{Z}^d$ has the counting measure as a stationary measure, but it is recurrent for $d \leqslant 2$ (see Theorem 5.21) and transient for $d \geqslant 3$.

Now that we have worked hard in deriving the theory behind stationary measures, let us see some applications.

Propositions 5.39 and 5.42 gives a very powerful tool for computing $\mathbb{E}_x[H_x]$ whenever it is finite. Indeed, it suffices to find a stationary probability measure μ, in which case we know that $\mathbb{E}_x[H_x] = \frac{1}{\mu(x)}$.

Example 5.43 (Random chess) A rook moves randomly on a chessboard: at each step, it makes uniformly at random any legal move (motion along rows or columns). How many moves on average does it take to return to its inital square?
This problem is a random walk on a finite graph in disguise (Examples 5.13 and 5.28). The vertex set is the set of squares on the chessboard, $S = \{1,\dots,8\}^2$. There is an edge between $x = (x_1, x_2)$ and $y = (y_1, y_2)$ if and only if $x_1 = y_1$ or $x_2 = y_2$, under the additional constraint $x \neq y$. Clearly, the chain is irreducible, and we already worked out the stationary measure in Theorem 5.28: $\mu(x) = CD_x$, where $C > 0$ is a normalization constant that ensures that μ is a probability measure. Since a rook can move from any square to any of 14 squares, we find that $D_x = 14$ for all $x \in S$. We have the condition

$$1 = \sum_{x\in S} \mu(x) = C \cdot 14 \cdot 64\,,$$

from which we deduce that $C = \frac{1}{14\cdot 64}$, and therefore

$$\mathbb{E}_x[H_x] = \frac{1}{\mu(x)} = \frac{1}{CD_x} = \frac{14\cdot 64}{14} = 64\,.$$

Suppose now that instead of a rook we play with a king, which can move to any of the eight squares sharing an edge or a corner with the original square. In this case, D_x depends on x. We consider three types of squares:

(c) corner: $D_x = 3$

(e) edge but no corner: $D_x = 5$
(b) neither edge nor corner: $D_x = 8$.

There are 4 squares of kind (c), 24 squares of kind (e), and 36 squares of kind (d). Thus we find

$$1 = \sum_{x \in S} \mu(x) = C(4 \cdot 3 + 24 \cdot 5 + 36 \cdot 8) = C \cdot 420.$$

We conclude that $\mathbb{E}_x[H_x]$ is $\frac{420}{3} = 140$ for x of kind (c), $\frac{420}{5} = 84$ for x of kind (e), and $\frac{420}{8} = 52.5$ for x of kind (b).

Example 5.44 (Asymmetric random walk on $\mathbb{N}$) Let us consider a random walk on $\mathbb{N}$. It is asymmetric in the sense that the probability p of taking a step to the right may be different from the probability $1 - p$ of taking a step to the left. Unlike the random walk on $\mathbb{Z}$ studied in Examples 5.11 and 5.21, this walk has a reflecting barrier at 0. The precise definition is as follows. Let $0 < p < 1$. For $x \in \mathbb{N}^*$ set

$$Q(x, x+1) := p\,, \qquad Q(x, x-1) := 1 - p\,,$$

and moreover $Q(0, 1) = 1$. (All other entries of Q vanish.) It is clear that Q is a stochastic matrix. It describes a p-asymmetric random walk on $\mathbb{N}$, which bounces off 0 back to the right whenever it hits it.

This chain is clearly irreducible. Let us look for a stationary measure. As usual, it is much easier to look for a reversible measure. The detailed balance equations from Theorem 5.24 read

$$\begin{aligned} \mu(x)p &= \mu(x+1)(1-p) \qquad \text{for } x \geqslant 1 \\ \mu(0) &= \mu(1)(1-p)\,, \end{aligned}$$

which cen be easily solved by induction to yield

$$\mu(x) = C \begin{cases} 1 - p & \text{if } x = 0 \\ \left(\frac{p}{1-p}\right)^{x-1} & \text{if } x \geqslant 1\,, \end{cases} \tag{5.14}$$

where C is a normalization constant. For $p < \frac{1}{2}$ the measure μ is finite (and hence can be chosen to be a probability measure). By Theorem 5.42, we conclude that for $p < \frac{1}{2}$ the chain is positive recurrent.

What about $p \geqslant \frac{1}{2}$? In that case the stationary measure is infinite, and we cannot conclude anything about recurrence or transience from it (all that we can say is that the chain is not positive recurrent).

Instead, we shall use a *coupling argument* that relates, or *couples*, the chain X_n to a suitable random walk on $\mathbb{Z}$, which can be more easily analysed. We consider the cases $p = \frac{1}{2}$ and $p > \frac{1}{2}$ separately.

$p = \frac{1}{2}$ Let (Y_n) be the simple random walk on $\mathbb{Z}$ (see Theorem 5.11). Then we claim that $X_n := |Y_n|$ is a simple random walk on $\mathbb{N}$ with transition matrix Q. To show this, let $y, x_0, \dots, x_n \in S$ and abbreviate

$$B := \{|Y_0| = x_0, \dots, |Y_{n-1}| = x_{n-1}\}\,.$$

Consider first the case $x_n = 0$, so that, by the Markov property from Theorem 5.15,

$$\begin{aligned} &\mathbb{P}(|Y_{n+1}| = y \mid |Y_n| = x_n, \dots, |Y_0| = x_0) \\ &= P(|Y_{n+1}| = y \mid Y_n = 0, B) \\ &= \mathbf{1}_{y=1} \\ &= Q(0, y)\,, \end{aligned}$$

as desired.

Next, if $x_n \neq 0$ we get

$$\begin{aligned}
&\mathbb{P}(|Y_{n+1}| = y \mid |Y_n| = x_n, \dots, |Y_0| = x_0) \\
&\quad = \mathbb{P}(|Y_{n+1}| = y \mid |Y_n| = x_n, B) \\
&\quad = \mathbb{P}(|Y_{n+1}| = y \mid Y_n = x_n, B)\frac{\mathbb{P}(Y_n = x_n, B)}{\mathbb{P}(|Y_n| = x_n, B)} \\
&\qquad + \mathbb{P}(|Y_{n+1}| = y \mid Y_n = -x_n, B)\frac{\mathbb{P}(Y_n = -x_n, B)}{\mathbb{P}(|Y_n| = x_n, B)}\,.
\end{aligned}$$

The first factor of each term on the right-hand side is $\frac{1}{2}\mathbf{1}_{|y-x_n|=1} = Q(x_n, y)$ by Theorem 5.15 and the definition of Y_n (note that this is only correct because $p = \frac{1}{2}$). Thus we conclude that

$$\begin{aligned}
&P(|Y_{n+1}| = y \mid |Y_n| = x_n, \dots, |Y_0| = x_0) \\
&\qquad = Q(x_n, y)\left(\frac{\mathbb{P}(Y_n = x_n, B)}{\mathbb{P}(|Y_n| = x_n, B)} + \frac{\mathbb{P}(Y_n = -x_n, B)}{\mathbb{P}(|Y_n| = x_n, B)}\right) = Q(x_n, y)\,,
\end{aligned}$$

as desired.

We conclude that

$$\mathbb{E}_0\left[\sum_{n\in\mathbb{N}} \mathbf{1}_{X_n=0}\right] = \mathbb{E}_0\left[\sum_{n\in\mathbb{N}} \mathbf{1}_{|Y_n|=0}\right] = \mathbb{E}_0\left[\sum_{n\in\mathbb{N}} \mathbf{1}_{Y_n=0}\right] = \infty\,,$$

where the last step follows from the recurrence of the simple random walk on $\mathbb{Z}$ (Theorem 5.21). Hence, for $p = \frac{1}{2}$ the chain (X_n) is recurrent. Since it has a stationary measure (5.14) of infinite total mass, from Theorem 5.39 we conclude that it is null recurrent.

$p > \frac{1}{2}$ For $p > \frac{1}{2}$, the previous coupling argument does not work because of the lack of symmetry. However, a somewhat different coupling to the asymmetric random walk on $\mathbb{Z}$ does work. Let (Y_n) be the asymmetric random walk on $\mathbb{Z}$ starting from 0 from Theorem 5.27. It has the transition matrix

$$\mathbb{P}(Y_{n+1} = y \mid Y_n = x) = p\mathbf{1}_{y=x+1} + (1-p)\mathbf{1}_{y=x-1}\,.$$

The idea of the argument is to define a walk X_n on $\mathbb{N}$ in terms of Y_n by imposing that X_n takes a step to the right whenever Y_n does, and X_n takes a step to the left whenever X_n does, unless $X_n = 0$, in which case X_n takes a step to the right even if Y_n takes a step to the left.

More formally, $X_0 := 0$ and

$$X_{n+1} := X_n + \begin{cases} Y_{n+1} - Y_n & \text{if } X_n > 0 \\ 1 & \text{if } X_n = 0\,. \end{cases} \tag{5.15}$$

From the definition and a simple induction argument, we get that

$$X_n \geqslant Y_n\,, \qquad \forall n \in \mathbb{N}\,. \tag{5.16}$$

Moreover, we claim that (X_n) thus defined is a Markov chain with transition matrix Q. To show this, let $y, x_0, \dots, x_n \in S$ and abbreviate

$$B := \{X_0 = x_0, \dots, X_n = x_n\}\,.$$

By the definition (5.15), the vector $(X_0, \dots, X_n)$ is a deterministic function of the vector $(Y_0, \dots, Y_n)$, and hence we can also write $B = \{(Y_0, \dots, Y_n) \in A\}$ for some set $A \subset S^{n+1}$. Now if $x_n = 0$ we have

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n, \dots, X_0 = x_0) = \mathbf{1}_{y=1} = Q(0, y)\,,$$

as desired. On the other hand, if $x_n > 0$, we get

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = y \mid X_n = x_n, \dots, X_0 = x_0) \\
&\quad = \mathbb{P}(X_{n+1} = y \mid B) \\
&\quad = \mathbb{P}(Y_{n+1} = Y_n + y - x_n \mid B) \\
&\quad = \sum_{z \in S} \mathbb{P}(Y_{n+1} = z + y - x_n \mid Y_n = z, B) \, \frac{\mathbb{P}(Y_n = z, B)}{\mathbb{P}(B)} \\
&\quad = \sum_{z \in S} Q(x_n, y) \, \frac{\mathbb{P}(Y_n = z, B)}{\mathbb{P}(B)} \\
&\quad = Q(x_n, y) \,,
\end{aligned}$$

where in the fourth step we used the Markov property Theorem 5.15 for the Markov chain (Y_n). We conclude that (X_n) is indeed a Markov chain with transition matrix Q.

To conclude the analysis, we note that the strong law of large numbers from Theorem 3.27 implies that

$$\lim_{n \to \infty} \frac{Y_n}{n} = 2p - 1$$

almost surely, because each step of the random walk (Y_n) has expectation $\mathbb{E}[Y_1 - Y_0] = 2p - 1$. Since $2p - 1 > 0$, we conclude that, almost surely, $Y_n \to +\infty$ as $n \to \infty$. From (5.16) we decude that almost surely $X_n \to +\infty$ as $n \to \infty$, and therefore (X_n) is transient.

We summarise: the random walk on $\mathbb{N}$ is

- positive recurrent for $p < \frac{1}{2}$,
- null recurrent for $p = \frac{1}{2}$,
- transient for $p > \frac{1}{2}$.

5.8 Asymptotic behaviour

WEEK 11

In this section we study the following important question. Suppose that we take an irreducible and recurrent Markov chain. In Section 5.6 we saw that the chain has a unique stationary measure μ. Let f be a nonnegative function on S. How are the *time average* $\sum_{i=0}^{n} f(X_i)$ and the *space average* $\int f \, \mathrm{d}\mu$ related? The following result says that for large n they coincide almost surely (up to a rescaling), no matter where the chain starts from. Such results are usually known as *ergodic theorems*.

Proposition 5.45 *Consider an irreducible and recurrent Markov chain with stationary measure μ. Let $f, g : S \to [0, \infty)$ such that $0 < \int g \, \mathrm{d}\mu < \infty$. Then, for all $x \in S$, we have $\mathbb{P}_x$-a.s.*

$$\frac{\sum_{i=0}^{n-1} f(X_i)}{\sum_{i=0}^{n-1} g(X_i)} \longrightarrow \frac{\int f \, \mathrm{d}\mu}{\int g \, \mathrm{d}\mu}$$

as $n \to \infty$.

As a consequence (taking $g = 1$), if the chain is positive recurrent, i.e. it has a stationary probability measure, the space average can be computed as the almost surely limit of the time average.

Corollary 5.46 *If the Markov chain is irreducible and positive recurrent with stationary probability measure μ, then, for all $x \in S$, we have $\mathbb{P}_x$-a.s.*

(5.17)
$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \longrightarrow \int f \, \mathrm{d}\mu$$

as $n \to \infty$.

Proof of Theorem 5.45 First, by monotone approximation, we may suppose that $\int f \, \mathrm{d}\mu < \infty$. (Otherwise, consider a sequence of functions f_m supported on finite subsets of S, that converges from below to f, and use the monotone convergence theorem.)

The main tool of the proof is the sequence of random times T_k, corresponding to the kth return of the chain to x. That is,

$$T_0 := 0\,, \qquad T_1 := H_x = \inf\{n > 0 : X_n = x\}\,, \qquad T_{k+1} := \inf\{n > T_k : X_n = x\}\,.$$

Since (X_n) is recurrent, $T_k < \infty$ a.s. for all $k \in \mathbb{N}$.

For $k \in \mathbb{N}$ we define the sum over the kth excursion,

$$Z_k := \sum_{n=T_k}^{T_{k+1}-1} f(X_n)\,.$$

The main observation is that $(Z_k)_{k\in\mathbb{N}}$ are independent identically distributed random variables. Intuitively, this is clear from the Markov property, since each excursion from x back to x is a fresh start, the chain forgetting everything about its past except that it starts at x again.

More formally, by Theorem 3.12, it suffices to show that for any bounded and measureable functions $g_0, \dots, g_k$ we have

$$\mathbb{E}_x\left[\prod_{i=0}^{k} g_i(Z_i)\right] = \prod_{i=0}^{k} \mathbb{E}_x[g_i(Z_0)]\,. \tag{5.18}$$

We show (5.18) by induction on k. The case $k = 0$ is obvious. For the induction step, we condition on T_k and T_{k+1} to get

$$\begin{aligned}
&\mathbb{E}_x\left[\prod_{i=0}^{k} g_i(Z_i)\right] \\
&= \sum_{m,n\in\mathbb{N}^*} \mathbb{E}_x\left[g_k(Z_k)\mathbf{1}_{T_{k+1}=n+m}\mathbf{1}_{T_k=n}\prod_{i=0}^{k-1} g_i(Z_i)\right] \\
&= \sum_{m,n\in\mathbb{N}^*} \mathbb{E}_x\left[g_k\Big(\sum_{l=n}^{n+m-1} f(X_l)\Big)\mathbf{1}_{T_{k+1}=n+m}\mathbf{1}_{T_k=n}\prod_{i=0}^{k-1} g_i(Z_i)\,\middle|\, X_n = x\right]\mathbb{P}_x(X_n = x) \\
&= \sum_{m,n\in\mathbb{N}^*} \mathbb{E}_x\left[g_k\Big(\sum_{l=0}^{m-1} f(X_l)\Big)\mathbf{1}_{H_x=m}\right]\mathbb{E}_x\left[\mathbf{1}_{T_k=n}\prod_{i=0}^{k-1} g_i(Z_i)\,\middle|\, X_n = x\right]\mathbb{P}_x(X_n = x) \\
&= \sum_{m,n\in\mathbb{N}^*} \mathbb{E}_x\left[g_k(Z_0)\mathbf{1}_{H_x=m}\right]\mathbb{E}_x\left[\mathbf{1}_{T_k=n}\prod_{i=0}^{k-1} g_i(Z_i)\right] \\
&= \mathbb{E}_x[g_k(Z_0)]\,\mathbb{E}_x\left[\prod_{i=0}^{k-1} g_i(Z_i)\right],
\end{aligned}$$

where in the third step we used Theorem 5.16 combined with the observation that on the event $\{T_k = n\}$ we have

$$\{T_{k+1} = n+m\} = \{X_{n+1} \neq x, \dots, X_{n+m-1} \neq x, X_{n+m} = x\}\,,$$

and in the fourth step we used that $X_n = x$ on the event $\{T_k = n\}$. Now (5.18) follows by induction.

Next, by Propositions 5.37 and 5.38, we have $\mu = \mu(x)\nu_x$ and $\nu_x(x) = 1$. Hence,

$$\mathbb{E}_x[Z_0] = \mathbb{E}_x\left[\sum_{n=0}^{H_x-1}\sum_{y\in S} f(y)\mathbf{1}_{X_n=y}\right] = \sum_{y\in S} f(y)\nu_x(y) = \frac{1}{\mu(x)}\int f\,\mathrm{d}\mu\,.$$

By independence of the family $(Z_k)_{k\in\mathbb{N}}$ and the law of large numbers[2] we therefore conclude that

$$\frac{1}{n}\sum_{k=0}^{n-1} Z_k \longrightarrow \frac{1}{\mu(x)}\int f\,\mathrm{d}\mu \tag{5.19}$$

$\mathbb{P}_x$-a.s.

What remains is to go from such an average over an integer multiple of excursions to an average over a given time. To that end, for $n \in \mathbb{N}$ we denote by $N_x(n)$ the number of returns to x before time n, i.e.

$$T_{N_x(n)} \leqslant n < T_{N_x(n)+1}\,.$$

Then we clearly have

$$\sum_{i=0}^{T_{N_x(n)}-1} f(X_i) \leqslant \sum_{i=0}^{n-1} f(X_i) \leqslant \sum_{i=0}^{T_{N_x(n)+1}-1} f(X_i)\,,$$

[2] In general we need the optimal version from Theorem A.5, since, as just shown, we only know that Z_0 has a finite expectation, i.e. is in L^1.

i.e.

$$\sum_{k=0}^{N_x(n)-1} Z_k \leqslant \sum_{i=0}^{n-1} f(X_i) \leqslant \sum_{k=0}^{N_x(n)} Z_k \,.$$

Dividing by $N_x(n)$ and using (5.19) yields

$$\frac{1}{N_x(n)} \sum_{i=0}^{n-1} f(X_i) \longrightarrow \frac{1}{\mu(x)} \int f \,\mathrm{d}\mu$$

$\mathbb{P}_x$-a.s.

The claim now follows by replacing f with g and dividing the two results. □

Remark 5.47 From Theorem 5.45 and Theorem 5.46, we deduce by choosing $f(y) = \mathbf{1}_{x=y}$ that for an irreducible positive recurrent chain with arbitrary initial distribution and stationary probability distribution μ we have a.s. for all $x \in S$

$$\frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}_{X_k=x} \longrightarrow \mu(x)$$

as $n \to \infty$. (A similar argument shows that if the chain is null recurrent then the limit is 0.) Taking the expectation and using dominated convergence yields

$$\frac{1}{n} \sum_{i=0}^{n-1} \mathbb{P}(X_k = x) \longrightarrow \mu(x) \,.$$

In words, the time averages of the laws converge to μ. It is natural to ask whether this holds without averaging over time: does the law of X_n converge to μ for any irreducible positive recurrent chain? In other words, does the measure $\mu_n(x) := \mathbb{P}(X_n = x)$ converge to μ in some sense?
It is easy to see that in general the answer is no. Consider the very simple example $S = \{1, 2\}$ and $Q = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. This chain is clearly irreducible and positive recurrent. Taking for instance an initial distribution δ_2, we clearly have

$$\mathbb{P}(X_n = x) = \mathbf{1}_{x-n \text{ is even}} \,.$$

In other words, the chain jumps between 1 and 2.
The problem with the above example is *periodicity*: the chain has a period of two, meaning that one can only return to the initial state in an even number of steps. It turns out that if in addition we impose that the chain is *aperiodic*, i.e. has no nontrivial period, then μ_n indeed converges to μ in the total variation distance. We shall not go into further details in this course.

5.9 Markov chain Monte Carlo and the Metropolis–Hastings algorithm

We conclude this chapter with a remarkable application of the theory of Markov chains. It is the original and most important algorithm in so-called *Markov chain Monte Carlo* (often abbreviated as MCMC). MCMC methods are some of the most important numerical algorithms ever devised, and they are used in countless applications, from fundamental sciences to economics and weather forecasts.

Suppose that we are given a large finite set S and a probability measure μ on S. We would like to compute the average of some function f, i.e. find $\int f \,\mathrm{d}\mu$. In most applications

of interest the set S is huge and the measure μ might also be hard to evaluate, thus rendering any simple-minded numerical evaluation of the integral hopeless.

The idea behind Monte Carlo methods is to draw a sample of random variables $X_0, X_1, \dots$ that on average reproduce the law μ, and hence we can hope that

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \longrightarrow \int f \, \mathrm{d}\mu \tag{5.20}$$

as $n \to \infty$. A simple way to guarantee convergence in (5.20) is to take the variables X_n to be independent with law μ, in which case (5.20) clearly holds a.s. by the law of large numbers. However, in practice, even this is often either extremely difficult or impossible. Let us consider a typical and celebrated example.

Example 5.48 (Ising model) The *Ising model* is the simplest and most famous model of *ferromagnetism*, the remarkable property of certain materials to exhibit spontaneous magnetisation (as used e.g. in fridge magnets). Despite being over a hundred years old, it is still actively studied and many important questions about it remain unsolved. Although seemingly simple, it is known to exhibit a remarkably intricate behaviour, in particular a host of phase transitions – abrupt transitions from one phase of matter to another – which can be studied theoretically (both analytically and numerically).

Let (V, E) be a finite connected graph on the vertex set V. Typically, one should think of V being a subset of the lattice $\mathbb{Z}^d$ such as $V = \{0, \dots, L-1\}^d$, and two elements $u, v \in V$ are adjacent if and only if they are nearest neighbours. At every vertex v sits a so-called spin, which can point up $+1$ or down -1. More formally, we consider a spin configuration $x = (x_v)_{v \in V} \in S := \{\pm 1\}^V$. With each spin configuration $x \in S$ we associate the *energy*

$$H(x) := -J \sum_{\{u,v\} \in E} x_u x_v - \sum_{v \in V} h_v x_v \,,$$

where $h \in \mathbb{R}^V$ is called an external magnetic field. (The physical intuition behind this definition is that, if $J > 0$, neighbouring spins like to align: each pair of neighbouring spins that are aligned lowers the energy by 1, while each pair of neighbouring spins that point in opposite directions increases the energy by 1. The material is called *ferromagnetic*. If $J < 0$ the the opposite behaviour is true, and neighbouring spins favour opposite orientations. The material is called *antiferromagnetic*. An external magnetic field introduces a bias in the orientation favoured by the spins.)

The probability of the spin configuration σ is given by the *Boltzmann-Gibbs* distribution from statistical mechanics:

$$\mu(x) := \frac{1}{Z} \mathrm{e}^{-\beta H(x)} \,, \qquad Z := \sum_{x \in S} \mathrm{e}^{-\beta H(x)} \,,$$

where $\beta > 0$ is a fixed parameter that has the physical interpretation of the inverse temperature. (The physical intuition behind this definition is that configurations of low energy occur with higher probability than configurations of high energy. The stronger this imbalance between high and low energies is, the lower the temperature of the system. In fact, this parameter β can be regarded as one possible mathematically rigorous definition of the rather mysterious physical notion of temperature in statistical mechanics.)

One is typically interested in quantities such as the magnetisation at a given vertex $v \in V$,

$$\int x_v \, \mu(\mathrm{d}x) \,, \tag{5.21}$$

as well as the two-point correlation function at two vertices u, v,

$$\int x_u \, x_v \, \mu(\mathrm{d}x) \,, \tag{5.22}$$

The size of the set S is $2^{|V|}$, which means that any direct numerical evaluation of such an integral is utterly hopeless for large V, no matter the computing power at one's disposal. Hence, some kind of Monte Carlo is required for the numerical evaluation of (5.21) and (5.22). However, there is no known effective way of generating random samples X with law μ.

Instead of trying to generating independent samples with law μ, which is typically impossible, the remarkable idea is the following.

Definition 5.49 (MCMC) Let μ be a probability measure on a finite set S. In *Markov chain Monte Carlo*, one constructs an irreducible Markov chain with stationary measure μ.

If we have constructed such a chain, then we can use it to sample the measure μ through (5.17), choosing a large enough n to obtain a good enough aproximation. (Indeed, by Theorem 5.42, the chain is positive recurrent and we can apply Theorem 5.46.)

How to construct such a chain? The original and most famous, still widely used, algorithm is the following.

Definition 5.50 (Metropolis-Hastings algorithm) Let μ be a probability measure on a finite set S. Suppose that $\mu(x) = C\rho(x)$ for some constant $C > 0$ (which does not need to be determined) and a known function $\rho > 0$.

(i) Choose the initial state X_0 in some arbitrary fashion.

(ii) Choose a stochastic matrix G on S satisfying the following conditions: $G(x,x) = 0$ for all x; the associated Markov chain is irreducible; $G(x,y) > 0$ if and only if $G(y,x) > 0$. This matrix is called the *proposal function*.

(iii) Construct the chain by defining inductively X_{n+1} as a function of $x = X_n$ as follows.

- Choose a candidate y randomly according to the law $G(x,\cdot)$.
- Calculate the *acceptance ratio*

$$A(x,y) := 1 \wedge \frac{\rho(y)\, G(y,x)}{\rho(x)\, G(x,y)} \,. \tag{5.23}$$

- Set $X_{n+1} := y$ with probability $A(x,y)$ (acceptance of proposal) or $X_{n+1} := x$ with probability $1 - A(x,y)$ (rejection of proposal).

In practice, the latter step is performed by drawing a random variable U with uniform law on $[0,1]$, and then setting

$$X_{n+1} := \begin{cases} y & \text{if } U \leqslant A(x,y) \\ x & \text{otherwise} \,. \end{cases}$$

Explicitly, the transition matrix Q of the Metropolis-Hastings chain is

$$Q(x,y) = \begin{cases} G(x,y)A(x,y) & \text{if } x \neq y \\ 1 - \sum_{z \in S} G(x,z)A(x,z) & \text{if } x = y \,. \end{cases}$$

For many applications, it is crucial that the Metropolis-Hastings algorithm only requires the knowledge of $\rho(x)$ and not of the constant C (and hence of the actual measure μ); the latter can be prohibitively difficult to determine, while ρ is often known and simple. See for instance Theorem 5.52 below.

Proposition 5.51 *The Metropolis-Hastings algorithm is an instance of MCMC: it defines an irreducible Markov chain with reversible measure μ.*

Proof Let us first show that the chain is irreducible. By the irreducibility assumption on G, for any two $x, y \in S$ there exists n such that $G^n(x,y) > 0$. Since $A(x,y) > 0$ whenever $G(x,y) > 0$, by assumption on ρ and G, we hence conclude that we also have $Q^n(x,y) > 0$, and the chain is irreducible.

What remains is to verify that μ is reversible. The detailed balance equations (5.9) for the matrix Q read

$$\mu(x)Q(x,y) = \mu(y)Q(y,x)$$

for all $x \neq y$, which are equivalently writtten as

$$\rho(x)G(x,y)A(x,y) = \rho(y)G(y,x)A(y,x)$$

for all $x \neq y$. This is equivalent to the condition

$$\frac{A(x,y)}{A(y,x)} = \frac{\rho(y)G(y,x)}{\rho(x)G(x,y)} \tag{5.24}$$

for all x, y satisfying $G(x,y) > 0$. For the choice (5.23), we always have $A(x,y) = 1$ or $A(y,x) = 1$, and hence it satisfies (5.24). □

It is common to choose the proposal function to be symmetric, $G(x,y) = G(y,x)$, in which case the acceptance ratio (5.23) simplifies to

$$A(x,y) = 1 \wedge \frac{\rho(y)}{\rho(x)}\,. \tag{5.25}$$

Example 5.52 (Theorem 5.48 continued) Let us apply the Metropolis-Hastings algorithm to the Ising model. In this case, we have

$$\rho(x) = \mathrm{e}^{-\beta H(x)}\,.$$

Note that computing the constant $C = \frac{1}{Z}$ (and hence determining the measure μ) is practially impossible (it is a sum over $2^{|V|}$ terms), but the function ρ itself is simple and easy to compute numerically.

We have a lot of freedom in choosing the proposal function, and different choices lead to different versions of the algorithm. We shall consider the simplest choice: choose a vertex v uniformly at random and flip the corresponding spin. That is, we map $x \mapsto x(v)$, where $x(v)$ is the configuration of spins obtained from x by flipping the spin at v: $x(v)_u = (-1)^{\delta_{uv}} x_u$. More formally,

$$G(x,y) = \begin{cases} \frac{1}{|V|} & \text{if } y = x(v) \text{ for some } v \in V \\ 0 & \text{otherwise}\,. \end{cases}$$

It is easy to check that G is irreducible and symmetric and satisfies $G(x,x) = 0$ for all x. The acceptance ratio (5.25) is

$$A(x,x(v)) = 1 \wedge \mathrm{e}^{\beta H(x) - \beta H(x(v))} = 1 \wedge \mathrm{e}^{-2\beta J \sum_{u\in V} \mathbf{1}_{\{u,v\}\in E} x_u x_v - 2\beta h_v x_v}\,.$$

Note that the right-hand side is trivial to evaluate numerically, as it involves just a sum over over the neighbours of v.

CHAPTER 6

Introduction to statistics

WEEK 12

In this final chapter we give an introduction to statistics. The goal of statistics is of an entirely different kind from that of probability, and indeed of any area of mathematics. In contrast to mathematics, which can be regarded as reasoning based on axioms and logic, entirely unconcerned with any physical reality, statistics is an empirical and pragmatic science whose goal is to understand the real world by analysing data from empirical observations. Very briefly, this difference can be summarised as follows.

- In probability, one is given a probability measure and random variables, and one studies their behaviour.
- In statistics, one is given a collection of observations obtained by repeating a random experiment, and one wishes to determine the law of the underlying random variable.

In practice, as in this course, the study of statistics is often combined with that of probability because it relies heavily on the tools and language of probability.

6.1 Estimators

We suppose that the observations are obtained from repeated experiments that are performed independently under the same conditions. This leads to the following notion.

Definition 6.1 An *n-sample* drawn from $\mathbb{P}$ is a family $X_1, \dots, X_n$ of independent random variables with law $\mathbb{P}$.

In *parametric statistics*, we suppose that the probability measure $\mathbb{P} \equiv \mathbb{P}_\theta$ depends on a parameter θ in some parameter set Θ. The goal is to *estimate* the parameter $\theta \in \Theta$ from an n-sample drawn from $\mathbb{P}_\theta$. Put differently, the rules of the game are the following.

- Known: an n-sample (the observation).
- Unknown: the parameter θ of $\mathbb{P}_\theta$.

The choice of the parametrisation $\theta \mapsto \mathbb{P}_\theta$ is a subjective decision to be made by the statistician, depending on various constraints and her insight into the nature of the questions she is investigating. As such, for a given question and n-sample of observations, there is usually no right or wrong parametrisation, although there are certainly reasonable and less reasonable choices. One has to appeal to common sense. Here are some examples along with reasonable choices of $\mathbb{P}_\theta$.

Example 6.2

- A yes/no opinion poll among n individuals. We choose $\mathbb{P}_\theta$ to the Bernoulli distribution on $\{0,1\}$ with parameter $\theta \in [0,1] = \Theta$.
- The lifetime of an appliance. We choose $\mathbb{P}_\theta$ to be the exponential distribution on $[0,\infty)$ with parameter $\theta \in (0,\infty) = \Theta$.
- The size of an individual in a homogeneous population. We choose $\mathbb{P}_\theta$ to be the normal distribution with mean m and variance σ^2, i.e. $\theta = (m,\sigma^2) \in \mathbb{R}\times[0,\infty) = \Theta$.

We seek to determine the parameter θ from the observed n-sample, or, more generally, any function $f(\theta)$ of the parameter.

Definition 6.3

(i) A *statistic* is a measurable function of an n-sample.

(ii) Let f be a function on Θ. An *estimator* of $f(\theta)$ is a statistic with values in $f(\Theta)$.

Note that an estimator can only depend on the sample (which is observable) and not on the parameter θ (which is unknown). It is customary in statistics to denote estimators with decorated symbols, such as $\hat{f}$, $\tilde{f}$, $\bar{f}$, to distinguish them from (unknown) deterministic functions of θ.

A good estimator $\hat{f} = F(X_1,\dots,X_n)$ of $f(\theta)$ should with high probability be close to $f(\theta)$ provided that the underlying n-sample was drawn from $\mathbb{P}_\theta$. This leads to the following condition.

Definition 6.4 For each $n \in \mathbb{N}^*$, let $\hat{f}_n = F_n(X_1,\dots,X_n)$ be an estimator of $f(\theta)$ depending on an n-sample $X_1,\dots,X_n$ drawn from $\mathbb{P}_\theta$. The family of estimators $\hat{f}_n$ is called *consistent* if for all $\theta \in \Theta$ we have

$$\hat{f}_n \xrightarrow{\mathbb{P}} f(\theta)$$

as $n \to \infty$.

Example 6.5 (Empirical mean) The *empirical mean*

$$\bar{X}_n := \frac{1}{n}(X_1 + \cdots + X_n)$$

is a consistent estimator of the mean $f(\theta) := \mathbb{E}_\theta[X_1]$, by the law of large numbers.

A classical characteristic of an estimator is its bias.

Definition 6.6 The *bias* of an estimator $\hat{f}$ of $f(\theta)$ is $\mathbb{E}_\theta[\hat{f}] - f(\theta)$. If the bias is zero for all $\theta \in \Theta$, the estimator is called *unbiased*. Otherwise, it is called *biased*.

Being unbiased can be a desirable property for an estimator. However, in practice it is not always a good idea to insist on a lack bias: it can indeed happen that a biased estimator performs better than an unbiased one. We shall examples of this later on. What one almost always will require, however, is that for large values of n the bias tends to zero.

Definition 6.7 A family of estimators $\hat{f}_n$ of $t(\theta)$ is *asymptotically unbiased* if for all $\theta \in \Theta$ we have

$$\lim_{n\to\infty} \mathbb{E}_\theta[\hat{f}_n] = f(\theta)\,.$$

Example 6.8 (Theorem 6.5 continued) The empirical mean $\bar{X}_n$ from Theorem 6.5 is an unbiased estimator of the mean $f(\theta) := \mathbb{E}_\theta[X_1]$.

Example 6.9 (Empirical variance) We would like to estimate the variance

$$\sigma^2 = f(\theta) := \mathrm{Var}_\theta(X_1) = \mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2$$

of the sample. A natural way to come with an estimator is to replace $\mathbb{E}_\theta$ with an empirical average over the n-sample:

$$\tilde{\sigma}^2 := \frac{1}{n}(X_1^2 + \cdots + X_n^2) - \left(\frac{1}{n}(X_1 + \cdots + X_n)\right)^2 .$$

By the law of large numbers, $\tilde{\sigma}^2$ is a consistent estimator of σ^2.
Is it biased? Let us find out:

$$\mathbb{E}[\tilde{\sigma}^2] = \mathbb{E}_\theta[X_1^2] - \frac{1}{n}\mathbb{E}_\theta[X_1^2] - \frac{n-1}{n}\mathbb{E}_\theta[X_1]^2 = \frac{n-1}{n}\sigma^2\,,$$

so that the bias is $-\sigma^2/n$. Hence, $\tilde{\sigma}^2$ is biased bu asymptotically unbiased. The bias can be removed by considering the slightly modified estimator

$$S_n^2 := \frac{n}{n-1}\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\,,$$

which is usually called the empirical variance. It is a consistent and unbiased estimator of the variance.

We saw a few examples of estimators that we defined essentially by guessing. In general, how does one find estimators? There is no general formula or algorithm, but there are some general recipies that are a good place to start. We discuss the two most common methods: the *method of moments* and the *maximum likelihood estimator*.

We have already used the *method of moments* in the Examples 6.5 and 6.9. In general, for $f(\theta) = \mathbb{E}_\theta[g(X_1)]$ for some function g, we can estimate $f(\theta)$ using the estimator

$$\hat{f} := \frac{1}{n}(g(X_1) + \cdots + g(X_n))\,,$$

which is unbiased and consistent (by the law of large numbers). A classical choice, giving the method its name, is $g(x) := x^r$ for some exponent $r \in \mathbb{R}$, which allows to estimate $\theta = h(\mathbb{E}_\theta[X_1^r])$ using the estimator

$$\hat{\theta} := h\left(\frac{1}{n}(X_1^r + \cdots + X_n^r)\right).$$

This estimator is consistent by the law of large numbers, but in general biased.

Example 6.10 If $\mathbb{P}_\theta$ is the exponential law with parameter θ, then we have $\mathbb{E}_\theta[X_1] = 1/\theta$, and therefore

$$\hat{\theta} = 1/\bar{X}_n$$

is a consistent estimator of θ.

The *maximum likelihood estimator* is somewhat more subtle and based on a simple but powerful idea: given a realisation $x_1, \dots, x_n$ of an n-sample, we look for $\theta \in \Theta$ for which the probability of observing $x_1, \dots, x_n$ is the highest. Informally, we look for the θ that best explains the observed sample. We shall construct an estimator based on this idea. For its definition, it is necessary to distinguish the discrete and continuous cases.

Definition 6.11 (Likelihood)

(i) Let $\mathbb{P}_\theta$ discrete. The *likelihood* at $x_1, \dots, x_n$ is the function

$$L(\theta; x_1, \dots, x_n) := \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i) .$$

(ii) Let $\mathbb{P}_\theta(\mathrm{d}x) = f_\theta(x)\,\mathrm{d}x$ be continuous with density f_θ. The *likelihood* at $x_1, \dots, x_n$ is the function

$$L(\theta; x_1, \dots, x_n) := \prod_{i=1}^{n} f_\theta(x_i) .$$

In each case, L should be regarded as a function of θ, with $x_1, \dots, x_n$ acting as fixed parameters.

Definition 6.12 (Maximum likelihood estimator) The *maximum likelihood estimator* of θ is the estimator

$$\hat{\theta} := \operatorname{argmax}_\theta L(\theta; X_1, \dots, X_n) ,$$

where argmax denotes the[a] value of $\theta \in \Theta$ at which the function L attains its maximum.

[a] Note that such a maximum might not exist or, if it exists, it might not be unique. Hence, the operator argmax is rarely used in mathematics, but it is very useful in statistics, where such questions of existence and uniqueness are always understood to be solved by common sense (e.g. choosing one of the maxima in some prescribed way in case the maximum is not unique.)

Example 6.13 If $\mathbb{P}_\theta$ is the exponential law with parameter θ and $x_1, \dots, x_n \geqslant 0$ is a realisation of an n-sample, then

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} \theta \mathrm{e}^{-\theta x_i} .$$

The maximising argument is easily determined by differentiation[a] in θ:

$$\hat{\theta} = \frac{n}{X_1 + \cdots + X_n} ,$$

which coincides with the estimator from Theorem 6.10 found using the method of moments.

[a] Here, and in many other situations, the following remark is helpful: because the function log is strictly increasing, maximising L is equivalent to maximising $\log L$. The derivative of the latter is often easier to compute. For this reason, one often consider the *log-likelihood* instead of the likelihood.

Example 6.14 If $\mathbb{P}_\theta$ is the uniform law on $[0, \theta]$ and $x_1, \dots, x_n \geqslant 0$ is a realisation of an n-sample, then

$$L(\theta; x_1, \dots, x_n) = \frac{1}{\theta^n} \mathbf{1}_{x_1 \leqslant \theta} \cdots \mathbf{1}_{x_n \leqslant \theta} = \frac{1}{\theta^n} \mathbf{1}_{\max_i x_i \leqslant \theta} .$$

The maximum likelihood estimator is therefore

$$\hat{\theta} = \max_i X_i \,.$$

We conclude this section by quantifying the quality of an estimator. Since one can often find several estimators for the same quantity, it is important to be able to quantify their accuracy and compare them.

Definition 6.15 Let $\hat{f}$ be an estimator of $f(\theta)$. The *quadratic risk* of $\hat{f}$ is

$$R_{\hat{f}}(\theta) := \mathbb{E}_\theta[(\hat{f} - f(\theta))^2] \,.$$

Remark 6.16 By writing $\hat{f} - f(\theta) = \hat{f} - \mathbb{E}_\theta[\hat{f}] + \mathbb{E}_\theta[\hat{f}] - f(\theta)$ and expanding in Theorem 6.15, we obtain

$$R_{\hat{f}}(\theta) = \mathrm{Var}_\theta(\hat{f}) + (\mathbb{E}_\theta[\hat{f}] - f(\theta))^2 \,. \tag{6.1}$$

In particular, if $\hat{f}$ is unbiased then

$$R_{\hat{f}}(\theta) = \mathrm{Var}_\theta(\hat{f}) \,.$$

Definition 6.17 If $\hat{f}$ and $\tilde{f}$ are estimators of $f(\theta)$, we say that $\hat{f}$ is *better* than $\tilde{f}$ if, for all $\theta \in \Theta$,

$$R_{\hat{f}}(\theta) \leqslant R_{\tilde{f}}(\theta) \,.$$

The relation (6.1) shows that the quadratic risk can be viewed as a sum of the variance and the square of the bias. In general, in order to minimise the quadratic risk, it can sometimes be advantageous to introduce a bias provided this sufficiently reduces the variance.

Example 6.18 (Theorem 6.14 continued) Let $\mathbb{P}_\theta$ be the uniform law on $[0, \theta]$. The estimator

$$\bar{\theta} = \frac{2}{n}(X_1 + \cdots + X_n)$$

is an unbiased estimator of θ. Its quadratic risk is

$$R_{\bar{\theta}}(\theta) = \frac{4}{n}\,\mathrm{Var}_\theta(X_1) = \frac{\theta^2}{3n} \,. \tag{6.2}$$

Let us compare $\bar{\theta}$ to the maximum likelihood estimator $\hat{\theta} = \max_i X_i$ from Theorem 6.14. Clearly, $\hat{\theta}$ is biased because $\mathbb{E}_\theta[\hat{\theta}] < \theta$. To compute the quadratic risk of $\hat{\theta}$, let us first compute its cumulative distribution function

$$\mathbb{P}_\theta(\hat{\theta} \leqslant x) = \mathbb{P}_\theta(X_1 \leqslant x, \dots, X_n \leqslant x) = \mathbb{P}(X_1 \leqslant x)^n = \left(\frac{x}{\theta}\right)^n$$

for $x \leqslant \theta$. Differentiating in x, we deduce that the density of the law of $\hat{\theta}$ is

$$f_{\hat{\theta}}(x) = \frac{n}{\theta^n} x^{n-1} \mathbf{1}_{0 \leqslant x \leqslant \theta} \,.$$

We deduce that

$$\mathbb{E}_\theta[\hat{\theta}] = \int x\, f_{\hat{\theta}}(x)\, \mathrm{d}x = \frac{n}{n+1}\theta \,,$$

so that $\hat{\theta}$ is asymptotically unbiased. For the quadratic risk, we obtain

$$R_{\hat{\theta}}(\theta) = \int (x-\theta)^2 f_{\hat{\theta}}(x)\,\mathrm{d}x = \frac{2\theta}{(n+1)(n+2)}\,. \tag{6.3}$$

Comparing (6.2) and (6.3), we conclude: for $n \geqslant 3$ the estimator $\hat{\theta}$ is better than $\bar{\theta}$, despite being biased. (Note that the bias of $\hat{\theta}$ can be removed by considering the estimator $\frac{n+1}{n}\hat{\theta}$.)

6.2 Confidence intervals

Suppose that we have a sample drawn from $\mathbb{P}_\theta$ and we wish to estimate $f(\theta) \in \mathbb{R}$. Often one not only wishes to estimate $f(\theta)$ but one would also like to have a notion of how likely it is that this estimate is close to the true value $f(\theta)$. Thus, we seek an interval I, depending only on the sample (and hence random), such that we know that $f(\theta) \in I$ with a certain confidence.

Definition 6.19 (Confidence interval) Let $0 < \gamma < 1$. Let $X_1, \dots, X_n$ be an n-sample drawn from $\mathbb{P}_\theta$. Let $f : \Theta \to \mathbb{R}$. An interval $I = I(X_1, \dots, X_n)$ is a *confidence interval for $f(\theta)$ with confidence level γ* if

$$\mathbb{P}_\theta(f(\theta) \in I) \geqslant \gamma\,, \qquad \forall \theta \in \Theta\,.$$

If there is equality, we call the I a strict[a] confidence interval.

[a] We note that there is no universal convention in the literature regarding this terminology, and both $\geqslant$ and $=$ are commonly used in the definition of confidence intervals.

Example 6.20 Suppose that $\mathbb{P}_\theta$ is the normal distribution with mean θ and variance one. Let

$$I := [\bar{X}_n - a, \bar{X}_n + a]\,.$$

We know that $\sqrt{n}(\bar{X}_n - a) =: Z$ is normal with mean zero and variance one. The condition of a strict confidence interval for θ reads

$$\gamma = \mathbb{P}_\theta(\theta \in I) = \mathbb{P}_\theta(|\bar{X}_n - a| \leqslant a) = \mathbb{P}(|Z| \leqslant a\sqrt{n})\,.$$

The right-hand side is an explicit function of the standard normal distribution can be easily computed numerically. For instance, for the confidence level $\gamma = 90\%$ we have

$$I = \left[\bar{X}_n - \frac{1.64}{\sqrt{n}}, \bar{X}_n + \frac{1.64}{\sqrt{n}}\right].$$

Example 6.21 Suppose that $\mathbb{P}_\theta$ is the uniform distribution on $[0, \theta]$. We use the estimator $\hat{\theta} = \max_i X_i$ from Theorem 6.14. Clearly, $\hat{\theta} \leqslant \theta$. Consider the interval

$$I = [\hat{\theta}, C\hat{\theta}]$$

for $C > 1$. The condition of a strict confidence interval for θ reads

$$\gamma = \mathbb{P}_\theta(\theta \in I) = \mathbb{P}_\theta(\theta \leqslant C\hat{\theta}) = 1 - \mathbb{P}_\theta\left(\hat{\theta} < \frac{\theta}{C}\right) = 1 - \frac{1}{C^n}\,,$$

where in the last step we used the law of $\hat{\theta}$ computed in Theorem 6.18. We conclude that the strict confidence interval with confidence level γ is

$$I = \left[\hat{\theta}, \left(\frac{1}{1-\gamma}\right)^{1/n} \hat{\theta}\right].$$

In the previous examples, thanks to a rather special form of the law $\mathbb{P}_\theta$, we could compute the probability $\mathbb{P}_\theta(f(\theta) \in I)$ exactly, and hence obtain stric confidence intervals. In general, such an exact computation is not possible and one is reduced to finding non-strict confidence intervals.

Example 6.22 Suppose that $\mathbb{P}_\theta$ has variance $\mathbb{E}_\theta[X_1^2] = \sigma^2$ and expectation $\mu = \mathbb{E}_\theta[X_1^2]$. We would like to estimate μ. As an estimator, we use the empirical mean $\bar{X}_n$. From Chebyshev's inequality we get

$$\mathbb{P}_\theta(|\bar{X}_n - \mu| < \delta) \geqslant 1 - \frac{\sigma^2}{n\delta^2},$$

from which we conclude that

$$I := \left[\bar{X}_n - \frac{\sigma}{\sqrt{n(1-\gamma)}}, \bar{X}_n + \frac{\sigma}{\sqrt{n(1-\gamma)}}\right] \tag{6.4}$$

is a confidence interval for μ with confidence level γ.

If n is large, by the Central Limit Theorem the estimate from the previous example is highly wasteful, as $\bar{X}_n$ is asymptotically Gaussian. The following definition is sometimes used to capture this phenomenon.

Definition 6.23 (Asymptotic confidence interval) Let $0 < \gamma < 1$. Let $X_1, \dots, X_n$ be an n-sample drawn from $\mathbb{P}_\theta$. Let $f : \Theta \to \mathbb{R}$. An interval $I_n = I_n(X_1, \dots, X_n)$ is an *asymptotic confidence interval for $f(\theta)$ with confidence level γ* if

$$\lim_{n\to\infty} \mathbb{P}_\theta(f(\theta) \in I_n) \geqslant \gamma, \qquad \forall \theta \in \Theta.$$

If there is equality, we call the I_n a strict asymptotic confidence interval.

Example 6.24 (Theorem 6.22 continued) Let us return to Theorem 6.22. By the Central Limit Theorem,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(\bar{X}_n \in \left[\mu - \frac{a\sigma}{\sqrt{n}}, \mu + \frac{a\sigma}{\sqrt{n}}\right]\right) = \mathbb{P}(|Z| \leqslant a),$$

where Z is a standard Gaussian random variable. Given γ, choose a such that $\mathbb{P}(|Z| \leqslant a) = \gamma$. In that case the interval

$$I_n := \left[\bar{X}_n - \frac{a\sigma}{\sqrt{n}}, \bar{X}_n + \frac{a\sigma}{\sqrt{n}}\right] \tag{6.5}$$

is an asymptotic confidence interval for μ with confidence level γ. Note that, if one aims for high confidence levels, where γ is close to 1, the interval (6.5) is much smaller than (6.4), because of the strong Gaussian decay of Z.
For instance, suppose that I am measuring the mean size μ of a population. The average uncertainty is $\sigma = 0.73$. How many samples do I need to determine μ with an accuracy of 0.1? To answer the question, I first have to choose a confidence level;

this is a choice to be made by the statistician based on personal preferences and needs. Let us say that I would like $\gamma = 99\%$. This yields $a \approx 2.58$ and hence

$$I_n = \left[\bar{X}_n - \frac{1.88}{\sqrt{n}}, \bar{X}_n + \frac{1.88}{\sqrt{n}} \right].$$

I need to choose n such that $\frac{1.88}{\sqrt{n}} > 0.1$, i.e. $n \geqslant 355$.

6.3 Hypothesis testing

WEEK 13

Example 6.25 Someone gives you a coin and you would like to determine whether it is fair. You do so by flipping it many times and recording the results. From the observed results you would like to determine which of the following hypotheses is true.

(H_0) The coin is fair (heads or tails occur with equal probability).
(H_1) The coin is biased (heads or tails occur with different probabilities).

The hypothesis (H_0) is called the *null hypothesis*, while the hypothesis (H_1) is called the *alternative hypothesis*.

How to choose the null hypothesis is an art and requires insight and experience on the part of the statistician. In general, the null hypothesis describes the default or standard scenario, where the statistical phenomenon or effect one is looking for is absent (in the above example, a bias in the coin). The alternative hypothesis describes the scenario where, on the contrary, the phenomenon or effect is present.

Example 6.26 A defendent is accused of a crime. In most modern systems of justice, the null and alternative hypotheses are:

(H_0) The defendent is innocent.
(H_1) The defendent is guilty.

In this example, the effect one is looking for is the guilt of the defendent.

The goal of statistical tests is to determine whether the observed sample provides sufficient evidence to reject the null hypothesis (and hence conclude that the phenomenon or effect one is investigating is present).

More formally, we consider an n-sample $X_1, \dots, X_n$ drawn from $\mathbb{P}_\theta$ with unknown $\theta \in \Theta$. We partition the parameter space in two:

$$\Theta = \Theta_0 \cup \Theta_1 , \qquad \Theta_0 \cap \Theta_1 = \emptyset .$$

We then define the hypotheses

$$(H_0) : \theta \in \Theta_0 \qquad \text{(null hypothesis)}$$
$$(H_1) : \theta \in \Theta_1 \qquad \text{(alternative hypothesis)} ,$$

and aim to determine, using the given n-sample, which of these two hypotheses is correct.

To that end, we follow the following *test procedure* for testing H_0 versus H_1:

(i) Define a *rejection region*, which is an event

$$D = D(X_1, \dots, X_n) .$$

(ii) Reject H_0 if and only if D holds.

There are two different kinds of errors that one can make:

- *Error of first kind:* H_0 is true but we reject it.
- *Error of second kind:* H_0 is false but we do not reject it.

In Theorem 6.26, the error of first kind is to convict an innocent defendant, while the error of second kind is to clear a guilty defendant.

The probabilities of commiting these two errors are quantified by the following definition.

Definition 6.27 Consider a statistical test determined by its rejection region D.

(i) The *confidence* of the test is

$$1-\alpha := \inf_{\theta\in\Theta_0} \mathbb{P}_\theta(D^c)\,.$$

We also call α the *risk* of the test.

(ii) The *power* of the test is

$$1-\beta := \inf_{\theta\in\Theta_1} \mathbb{P}_\theta(D)\,.$$

Thus, α is the probability of making a mistake of first kind, while β is the probability of making a mistake of second kind.

Example 6.28 Let I be a confidence interval for θ with confidence level $1-\alpha$. Suppose that $\theta_0 \in \Theta$. We want to test the null hypothesis $\theta = \theta_0$ versus the alternative hypothesis $\theta \neq \theta_0$. Then the rejection region $D = \{\theta_0 \notin I\}$ yields a test of H_0 versus H_1 with confidence at least $1-\alpha$. Indeed, by Theorem 6.19, we have

$$\mathbb{P}_{\theta_0}(D^c) = \mathbb{P}_{\theta_0}(\theta_0 \in I) \geqslant 1-\alpha\,.$$

Next, we consider a few concrete Gaussian examples.

Example 6.29 Let $X_1, \dots, X_n$ be an n-sample drawn from the Gaussian law with mean μ and variance σ^2. Let $\mu_0 \in \mathbb{R}$. We want to test the null hypothesis $\mu = \mu_0$ versus the alternative hypothesis $\mu \neq \mu_0$. To construct the rejection region, we use the empirical mean $\bar{X}_n$ (see Theorem 6.5).

(i) Suppose first that the variance σ^2 is known and only the mean μ is unknown. For $C > 0$ define

$$D := \left\{|\bar{X}_n - \mu_0| \geqslant C\right\}.$$

We require a test with confidence at least 95%, which means

$$\mathbb{P}_{\mu_0}(|\bar{X}_n - \mu_0| \geqslant C) = 0.05\,,$$

which gives the condition $C \approx \frac{1.96\cdot\sigma}{\sqrt{n}}$ (where we used that $\bar{X}_n$ is Gaussian with mean μ and variance σ^2/n).

(ii) Suppose now that we know neither the mean μ nor the variance σ^2. We use the empirical mean $\bar{X}_n$ and the empirical variance S_n^2 from Theorem 6.9, and set

$$D := \left\{\frac{|\bar{X}_n - \mu_0|}{S_n} \geqslant C\right\}.$$

The law of the random variable

$$T_{n-1} := \frac{\sqrt{n}}{S_n}(\bar{X}_n - \mu)$$

is called *Student's t distribution with n degrees of freedom*; it has an explicit form that can be computed or found in the literature, which we shall not go into here. We require a test with confidence at least 95%, which means

$$\mathbb{P}_{\mu_0}(D) = \mathbb{P}(|T_{n-1}| \geqslant C/\sqrt{n}) = 0.05\,,$$

from which one can solve $C \approx \frac{2.093}{\sqrt{n}}$. Thus, the rejection region for a test with confidence 95% is

$$D = \left\{ \frac{|\bar{X}_n - \mu_0|}{S_n} \geqslant \frac{2.093}{\sqrt{n}} \right\}.$$

Next, we consider testing of so-called simple, or binary, hypotheses, where $\Theta = \{\theta_0, \theta_1\}$ consists of just two elements. The null hypothesis is $\theta = \theta_0$ and the alternative hypothesis is $\theta = \theta_1$. A powerful (in fact, the most powerful, see Theorem 6.31 below) test in this situation is the *Neyman-Pearson test*, defined as follows. Recall the likelihood L from Theorem 6.11. Define the *likelihood ratio*

$$R(\theta_0, \theta_1; x_1, \dots, x_n) := \frac{L(\theta_1; x_1, \dots, x_n)}{L(\theta_0; x_1, \dots, x_n)}.$$

The Neyman-Pearson test is defined by the rejection region

$$D = \left\{ R(\theta_0, \theta_1; X_1, \dots, X_n) > C \right\}.$$

Intuitively, a larger valure of R indicates that θ_1 is more likely than θ_0, and hence a rejection of the null hypothesis $\theta = \theta_0$ should be more likely.

Example 6.30 A person has two coins. One is fair. For the other, the probability of obtaining heads is twice that of obtaining tails. She chooses one of the coins, tosses it 100 times and obtains 60 heads and 40 tails. Which coin did she pick?
We model this with an n-sample $X_1, \dots, X_n$ drawn from a Bernoulli distribution with parameter p. The null hypothesis (fair coin) is $p = 1/2$ while the alternative hypothesis (biased coin) is $p = 2/3$. The number of heads is $K := X_1 + \cdots + X_n$. The likelihood for a sample with $K = k$ is

$$L(p, k) = p^k (1-p)^{n-k} = (1-p)^n \left(\frac{p}{1-p} \right)^k.$$

The likelihood ratio $R = R(1/2, 2/3)$ is

$$R(k) = \left(\frac{1 - 2/3}{1 - 1/2} \right)^n \left(\frac{2/3}{1/3} \right)^k = \left(\frac{2}{3} \right)^n 2^k.$$

We use the Neyman-Pearson with rejection region

$$D = \{R(K) > C\} = \{K > C'\}$$

where we used that $k \mapsto R(k)$ is monotone increasing and we set $C' = R^{-1}(C)$. Suppose that we want a test with a confidence of 90%. This results in the condition

$$\mathbb{P}_{1/2}(K > C') = 0.1.$$

The law of K for $n = 100$ can be evaluated numerically or approximated by the Central Limit Theorem. This yields $C' \approx 54.6$. Since we observed $K = 60 > C'$ we are in rejection region, and hence we reject the null hypothesis. Thus we can say, with 90% confidence, that the person chose the biased coin.

We conclude this chapter with a remarkable theoretical result, known as the Neyman-Pearson lemma, which states that, within the context of simple hypothesis testing, the Neyman-Pearson test is the most powerful test at any given confidence level.

Proposition 6.31 (Neyman-Pearson lemma) *The Neyman-Pearson test is the most powerful test at any given confidence level. More precisely, let $\Theta = \{\theta_0, \theta_1\}$ and*

suppose that $\mathbb{P}_\theta$ has a density for each $\theta \in \Theta$. Let $0 < \alpha < 1$. Suppose that the constant C in the Neyman-Pearson rejection region

$$D = \{R(\theta_0, \theta_1) > C\}$$

is chosen to that the risk $\mathbb{P}_{\theta_0}(D) = \alpha$. Then for any rejection region B with risk $\mathbb{P}_{\theta_0}(B) = \alpha$ we have

$$\mathbb{P}_{\theta_1}(B) \leqslant \mathbb{P}_{\theta_1}(D)$$

with a strict inequality if $\mathbb{P}_{\theta_1}(D \setminus B) > 0$.

Proof We use the notation $x = (x_1, \dots, x_n)$ and $\mathbb{P}_\theta(\mathrm{d}x_1) \cdots \mathbb{P}_\theta(\mathrm{d}x_n) = f_\theta(x)\,\mathrm{d}x$. From $\mathbb{P}_{\theta_0}(D) = \mathbb{P}_{\theta_0}(B) = \alpha$ we get

$$\int_{D\setminus B} f_{\theta_0}(x)\,\mathrm{d}x = \alpha - \int_{D\cap B} f_{\theta_0}(x)\,\mathrm{d}x = \int_{B\setminus D} f_{\theta_0}(x)\,\mathrm{d}x\,.$$

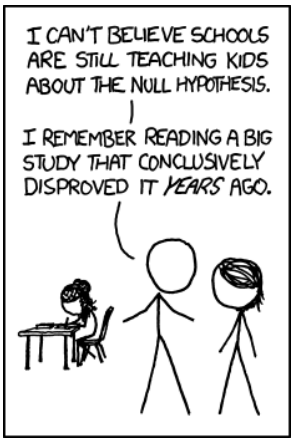
Since $D \setminus B \subset D$ and $B \setminus D \subset D^c$, we therefore get, by definition of the event D.

$$\int_{D\setminus B} f_{\theta_1}(x)\,\mathrm{d}x \geqslant C \int_{D\setminus B} f_{\theta_0}(x)\,\mathrm{d}x = C \int_{B\setminus D} f_{\theta_0}(x)\,\mathrm{d}x \geqslant \int_{B\setminus D} f_{\theta_1}(x)\,\mathrm{d}x\,,$$

where the first inequality is strict if $\mathbb{P}_{\theta_1}(D \setminus B) > 0$. Thus we conclude that

$$\mathbb{P}_{\theta_1}(D) = \mathbb{P}_{\theta_1}(D \setminus B) + \mathbb{P}_{\theta_1}(D \cap B) \geqslant \mathbb{P}_{\theta_1}(B \setminus D) + \mathbb{P}_{\theta_1}(D \cap B) = \mathbb{P}_{\theta_1}(B)\,,$$

as claimed. □



This is the end of this course. I hope you enjoyed it!

Now you know all of the fundamentals of probability. If you liked what you learned (as I hope!), you are fully equipped to go on and learn about more advanced topics such as martingales and Brownian motion.

APPENDIX A

The strong law of large numbers

In this appendix we prove the strong law of large numbers under the optimal condition on the random variables. Recall that we already proved the weak law of large numbers in Theorem 3.24, which had the deficiency of establishing convergence in L^2 instead of almost surely, which led to issues explained in Theorem 3.26. This issue was remedied in the strong law of large numbers in L^4 in Theorem 3.27. But the latter result still required the random variables X_n to lie in L^4 instead of in the optimal space, L^1. (This space is optimal since we clearly want $\mathbb{E}[X_1]$ to be well-defined and finite.)

We shall need the following tool from measure theory, which is a consequence of the monotone class lemma.

Lemma A.1 *For each $i = 1, \dots, n$, let $\mathcal{C}_i \subset \mathcal{A}$ be a collection of events stable under intersections containing Ω. Define $\mathcal{B}_i := \sigma(\mathcal{C}_i)$. If for all $C_1 \in \mathcal{C}_1, \dots, C_n \in \mathcal{C}_n$ we have*

$$\mathbb{P}(C_1 \cap \cdots \cap C_n) = \mathbb{P}(C_1) \cdots \mathbb{P}(C_n) ,$$

then $\mathcal{B}_1, \dots, \mathcal{B}_n$ are independent.

Proof We use the monotone class lemma from Section 3.2, whose notations we also take over. Fix $C_2 \in \mathcal{C}_2, \dots, C_n \in \mathcal{C}_n$, and define

$$\mathcal{M}_1 := \left\{ B_1 \in \mathcal{B}_1 : \mathbb{P}(B_1 \cap C_2 \cap \cdots \cap C_n) = \mathbb{P}(B_1)\, \mathbb{P}(C_2) \cdots \mathbb{P}(C_n) \right\} .$$

By assumption, $\mathcal{C}_1 \subset \mathcal{M}_1$. Moreover, it is easy to verify that $\mathcal{M}_1$ is a monotone class. Hence,

$$\mathcal{M}_1 \supset \mathcal{M}(\mathcal{C}_1) = \sigma(\mathcal{C}_1) = \mathcal{B}_1 ,$$

where the second step follows from the monotone class lemma (Theorem 3.8). We conclude: for all $B_1 \in \mathcal{B}_1$, $C_2 \in \mathcal{C}_2, \dots, C_n \in \mathcal{C}_n$, we have

$$\mathbb{P}(B_1 \cap C_2 \cap \cdots \cap C_n) = \mathbb{P}(B_1)\, \mathbb{P}(C_2) \cdots \mathbb{P}(C_n) .$$

We now continue in this fashion, moving on to the second argument. More precisely, fix $B_1 \in \mathcal{B}_1$, $C_3 \in \mathcal{C}_3, \dots, C_n \in \mathcal{C}_n$ and define

$$\mathcal{M}_2 := \left\{ B_2 \in \mathcal{B}_2 : \mathbb{P}(B_1 \cap B_2 \cap C_3 \cap \cdots \cap C_n) = \mathbb{P}(B_1)\, \mathbb{P}(B_2)\, \mathbb{P}(C_3) \cdots \mathbb{P}(C_n) \right\} .$$

As above, it is easy to see that $\mathcal{M}_2$ is a monotone class, and by the previous step we know that $\mathcal{C}_2 \subset \mathcal{M}_2$. By the monotone class lemma, we find that $\mathcal{M}_2 \supset \mathcal{B}_2$. By repeating this procedure n times we arrive at the claim. □

Our proof of the strong law of large numbers rests on the following fundamental result. To state it, let $(X_n)_{n \geqslant 1}$ be a family of random variables. For $n \geqslant 1$ we define the σ-algebra

$$\mathcal{B}_n := \sigma(X_n, X_{n+1}, \dots) = \sigma\left(\bigcup_{k \geqslant n} \sigma(X_k) \right)$$

as well as the *tail σ-algebra*

$$\mathcal{B}_\infty := \bigcap_{n\geqslant 1} \mathcal{B}_n \,.$$

Proposition A.2 (Kolmogorov's zero-one law) *Let $(X_n)_{n\geqslant 1}$ be independent random variables. Then $\mathcal{B}_\infty$ satisfies a* zero-one law *in the sense that any* tail event $B \in \mathcal{B}_\infty$ *satisfies $\mathbb{P}(B) = 0$ or $\mathbb{P}(B) = 1$.*

Remark A.3 It is important to understand the meaning of the objects in Theorem A.2. The σ-algebra $\mathcal{B}_n$ contains all the information from time n onwards, i.e. it discards all information up to time $n-1$. The tail events in $\mathcal{B}_\infty$ are precisely those whose occurrence can be determined if an arbitrarily large but finite initial segment of the variables X_k is discarded. For example $\{\sup_n X_n \leqslant 1\}$ is not in $\mathcal{B}_\infty$, since it clearly depends on all random variables X_n. But $\{\limsup_n X_n \leqslant 1\}$ is in $\mathcal{B}_\infty$, since it depends only on the "distant future", i.e. changing any finite number of variables X_n does not change its occurrence.
Kolmogorov's zero-one law is remarkable: it states that any tail event occurs almost surely or its complement occurs almost surely. As we shall see, the tail σ-algebra is rich (i.e. large) enough to make this statement very useful.

Proof of Theorem A.2 Define $\mathcal{D}_n := \sigma(X_1, \dots, X_n)$ (the σ-algebra containing the information up to time n). Then we claim that $\mathcal{D}_n$ and $\mathcal{B}_{n+1}$ are independent. This sounds intuitively obvious, as $\mathcal{D}_n$ contains information up to time n, and $\mathcal{B}_{n+1}$ information starting from time $n+1$. For a rigorous proof, we proceed in two steps.

- For any $k \geqslant n+1$ we define $\mathcal{B}_{n+1,k} := \sigma(X_{n+1}, \dots, X_k)$. Define the collections

$$\mathcal{C}_1 := \left\{ B_1 \cap \cdots \cap B_n : B_i \in \sigma(X_i)\, \forall i \right\},$$
$$\mathcal{C}_2 := \left\{ B_{n+1} \cap \cdots \cap B_k : B_i \in \sigma(X_i)\, \forall i \right\}.$$

  Clearly, these collections are stable under finite intersections and $\mathcal{D}_n = \sigma(\mathcal{C}_n)$ and $\mathcal{B}_{n+1,k} = \sigma(\mathcal{C}_2)$. By Theorem A.1 and independence of the random variables (X_n), we therefore conclude that $\mathcal{D}_n$ and $\mathcal{B}_{n+1,k}$ are independent.
- Define the collections $\mathcal{C}_1 := \mathcal{D}_n$ and $\mathcal{C}_2 := \bigcup_{k\geqslant n+1} \mathcal{B}_{n+1,k}$, which are clearly stable under finite intersections. Thus, $\sigma(\mathcal{C}_1) = \mathcal{D}_n$ and $\sigma(\mathcal{C}_2) = \mathcal{B}_{n+1}$. By the previous step and Theorem A.1, we conclude that $\mathcal{D}_n$ and $\mathcal{B}_{n+1}$ are independent, as desired.

Next, choose $\mathcal{C}_1 := \bigcup_{n\geqslant 1} \mathcal{D}_n$ and $\mathcal{C}_2 := \mathcal{B}_\infty$. Since $\mathcal{D}_n$ and $\mathcal{B}_{n+1}$ are independent for all n, we conclude that $\mathbb{P}(C_1 \cap C_2) = \mathbb{P}(C_1)\,\mathbb{P}(C_2)$ for all $C_1 \in \mathcal{C}_1$ and $C_2 \in \mathcal{C}_2$. By Theorem A.1, we deduce that $\sigma(\mathcal{C}_1) = \sigma(X_1, X_2, \dots) = \mathcal{B}_1$ and $\mathcal{B}_\infty$ are independent. Since $\mathcal{B}_\infty \subset \mathcal{B}_1$, we conclude that $\mathcal{B}_\infty$ is independent of itself! This means that any tail event $B \in \mathcal{B}_\infty$ satisfies $\mathbb{P}(B) = \mathbb{P}(B \cap B) = \mathbb{P}(B)^2$, from which the zero-one law follows. □

At first sight, this proof seems quite strange. It is in fact nothing but a careful justification of a simple fact: $\mathcal{B}_\infty$ is independent of itself. Since we are working with rather abstract σ-algebras, it is important to proceed slowly and carefully, as we tried to do above. The zero-one law has deep implications in probability. The strong law of large numbers, which we are about to state and prove, is one. The following remark is another one.

Remark A.4 Let $(X_n)_{n\geqslant 1}$ be independent random variables. Clearly,

$$X_+ := \limsup_{k\to\infty} \frac{1}{k}(X_1 + \cdots + X_k) = \limsup_{k\to\infty} \frac{1}{k}(X_n + \cdots + X_k)$$

for any $n \in \mathbb{N}^*$. Hence, X_+ is $\mathcal{B}_n$-measurable for all $n \in \mathbb{N}^*$, which implies that X_+ is $\mathcal{B}_\infty$-measurable. The same holds for X_- where lim sup is replaced with lim inf. In particular, the event

$$\left\{ \frac{1}{k}(X_1 + \cdots + X_k) \text{ converges} \right\} = \{X_- = X_+\}$$

is $\mathcal{B}_\infty$-measurable, and hence has either probability 1 or 0. In the former case, the limiting random variable $X_- = X_+$ is $\mathcal{B}_\infty$-measurable, and it is therefore almost surely constant (exercise). In summary: averages of independent random variables either diverge almost surely or converge almost surely to a constant.

We can now state and prove the strong law of large numbers.

Proposition A.5 (Strong law of large numbers) *Let $(X_n)_{n\geqslant 1}$ be independent random variables in L^1 with the same law. Then*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{a.s.} \mathbb{E}[X_1]\,.$$

Proof Let $S_n := X_1 + \cdots + X_n$ and $S_0 := 0$. Let $a > \mathbb{E}[X_1]$ and define $M := \sup_{n\in\mathbb{N}}(S_n - na)$. Thus, M is a random variable with values in $[0, \infty]$. The core of the proof is to show that

$$M < \infty \quad \text{a.s.} \tag{A.1}$$

Let us suppose first that (A.1) has been proved and use it to conclude the proof of the strong law of large numbers. By definition of M we have $S_n \leqslant na + M$ for all n and hence (A.1) implies, for all $a > \mathbb{E}[X_1]$,

$$\limsup_n \frac{S_n}{n} \leqslant a \quad \text{a.s.}$$

This implies that

$$\limsup_n \frac{S_n}{n} \leqslant \mathbb{E}[X_1] \quad \text{a.s.}\,, \tag{A.2}$$

since

$$\mathbb{P}\left(\limsup_n \frac{S_n}{n} \leqslant \mathbb{E}[X_1]\right) = \mathbb{P}\left(\bigcap_{k\in\mathbb{N}^*} \left\{\limsup_n \frac{S_n}{n} \leqslant \mathbb{E}[X_1] + \frac{1}{k}\right\}\right) = 1\,,$$

where we used that a countable intersection of events of probability one has probability one.

Replacing X_n with $-X_n$ we obtain

$$\liminf_n \frac{S_n}{n} \geqslant \mathbb{E}[X_1] \quad \text{a.s.} \tag{A.3}$$

From (A.2) and (A.3) we conclude the strong law of large numbers.

What remains, therefore, is to prove (A.1). First, we claim that $\{M < \infty\} \in \mathcal{B}_\infty$. Indeed, for all $k \geqslant 0$ we have

$$\{M < \infty\} = \left\{\sup_{n\geqslant 0}(S_n - na) < \infty\right\} = \left\{\sup_{n\geqslant k}\big((S_n - S_k) - na\big) < \infty\right\} \in \sigma(X_{k+1}, X_{k+2}, \dots)\,,$$

since $S_n - S_k = X_{k+1} + X_{k+2} + \cdots + X_n$. By the zero-one law, Theorem A.2, to prove (A.1), it therefore suffices to prove that $\mathbb{P}(M = \infty) < 1$.

We proceed by contradiction and suppose that $\mathbb{P}(M = \infty) = 1$. For all $k \in \mathbb{N}$ we define

$$M_k := \sup_{0 \leqslant n \leqslant k} (S_n - na)\,, \qquad M'_k := \sup_{0 \leqslant n \leqslant k} (S_{n+1} - S_1 - na)\,.$$

Since $S_n = X_1 + \cdots + X_n$ and $S_{n+1} - S_1 = X_2 + \cdots + X_{n+1}$, we conclude that $M_k \stackrel{\mathrm{d}}{=} M'_k$ (recall (2.5)). Moreover, M_k and M'_k are increasing sequences that converge from below to their limits M and M'. By $M_k \stackrel{\mathrm{d}}{=} M'_k$, we conclude that $M \stackrel{\mathrm{d}}{=} M'$, since

$$\mathbb{P}(M' \leqslant x) = \lim_{k\to\infty} \mathbb{P}(M'_k \leqslant x) = \lim_{k\to\infty} \mathbb{P}(M_k \leqslant x) = \mathbb{P}(M \leqslant x)\,.$$

Moreover,

$$\begin{aligned} M_{k+1} &= \sup\left\{0, \sup_{1 \leqslant n \leqslant k+1} (S_n - na)\right\} \\ &= \sup\left\{0, \sup_{0 \leqslant n \leqslant k} (S_{n+1} - (n+1)a)\right\} \\ &= \sup\{0, M'_k + X_1 - a\} \\ &= M'_k - \inf\{a - X_1, M'_k\}\,. \end{aligned}$$

Hence,

$$\mathbb{E}[\inf\{a - X_1, M'_k\}] = \mathbb{E}[M'_k] - \mathbb{E}[M_{k+1}] = \mathbb{E}[M_k] - \mathbb{E}[M_{k+1}] \leqslant 0\,, \tag{A.4}$$

since the sequence (M_k) is nondecreasing. Moreover, since $M'_k \geqslant 0$ we have

$$|\inf\{a - X_1, M'_k\}| \leqslant |a - X_1|$$

for all k, so that we may apply dominated convergence to (A.4) to get

$$\mathbb{E}[\inf\{a - X_1, M'\}] \leqslant 0\,.$$

Now if $\mathbb{P}(M = \infty) = 1$ then also $\mathbb{P}(M' = \infty) = 1$ and hence $\inf\{a - X_1, M'\} = a - X_1$. But

$$\mathbb{E}[a - X_1] > 0$$

by assumption. This is the desired contradiction. □