Introduction to Stochastic Processes

Agissilaos G. Athanassoulis

Preface

These notes are addressed to advanced undergraduate and postgraduate students, and are meant to be reasonably self-contained: they open with a crash course in probability, including combinatorics and elementary probability first. These are worked through concrete classical problems, to develop the students' intuition. Then the probability space, random variables, and limit theorems are introduced. The measure-theoretic background is recalled and its necessity for rigorous results is explained; however, we don't always go into all the details. An intuitive, hands-on style, including relatively basic examples, is meant to motivate and frame the ideas for students without a deep knowledge of measure theory. Markov chains are then framed as a natural extension of sequences of random variables, and the later chapters – continuous-time Markov processes, Brownian motion and mean-square calculus, stationary processes and their spectra – make up the advanced part of the notes, closing with a glimpse of Itô calculus.

In these notes we think of probability and stochastic processes primarily as tools to model real systems. Thus, we pay attention to what is being assumed about a phenomenon when we call it random, and how those assumptions interact with the mathematical proofs. The wave-tank experiment of Chapter 1 is set up early precisely so that, by Chapter 14, the reader can watch it become a theorem. In the same spirit, the language often focuses on helping intuition develop, rather than the shortest or most general proof. Sections and proofs marked with a star are more technical and can be skipped on a first reading without loss of continuity.

Not all chapters are covered every year. Typically, in an 11 week semester, 10-12 chapters would be covered, depending on the background and interests of each cohort.

Most chapters end with a problem set; several of its problems are referenced at the corresponding points of the text. Outline solutions are gathered at the end of the notes. Targeted "further reading" pointers close most chapters, and the works that have influenced these notes most are collected in the References.

Contents

1 Introduction to stochastic modelling

1.1 Modelling and the scientific method

We will start with a conceptual introduction. Loosely speaking, we want to have a grasp of what a "random phenomenon" is or, more importantly, what the subject matter of "stochastic modelling" is. This discussion will also be crucial in highlighting the relations between various mathematical objects and techniques that we will encounter.

One important way to conceptualise any phenomenon is through its amenability to experimental study:

Definition 1.1 (Experiment, ω)**.** *(i). An experiment is the controlled repetition of a phenomenon, so that the main factors affecting it (the "conditions" of the experiment) are known.*

(ii). The outcomes of an experiment, $\omega \in \Omega$, must be well defined and possible to measure with sufficient accuracy.

(iii). An experiment must be reproducible.

Some remarks are in order:

- **Repetitions under same conditions:** In practice we have either

  (i). a single experimental setup (e.g. device), and run multiple repetitions on it, or

  (ii). we have several different copies of a setup which we accept as "identical".

  An example of the first kind is Hippolyte Fizeau's rotating cogwheel arrangement that was used to measure the speed of light with 5% error in 1849. Examples of phenomena that we often model as "identical repetitions" include

  - the amount of time it takes for the Earth to perform a full revolution around the Sun
  - the total number of phone calls in the UK on a business day
  - the total volume of water going through the Tay every year,
  - the highest wave to appear in the North Atlantic each week,
  - the percentage of mortgages that default within a year,
  - or the heads-or-tails outcome of identically setup coin tosses.

  It is not always an easy question whether the main factors affecting the phenomenon are the same, as these may not be known, or not be possible to measure. E.g., if the climate changes fast enough, year-by-year weather-related data perhaps should not be thought of as different realisations of the same experiment. In any kind of real-world study of noisy phenomena, this type of question is in fact very important, and usually not possible to answer with absolute certainty. The investigators should keep track of their assumptions, and test them wherever possible.

- **Outcomes:** the set Ω containing all the possible outcomes of the experiment will be called the **sample space**. The complexity of studying the experiment mathematically is determined by the structure of Ω – not merely its cardinality ($\mathbb{R}$ and many function spaces have the *same* cardinality; what changes is the available structure: topology, norms, compactness, regularity of paths): the easiest case is when Ω is finite or countable; a harder case is when $\Omega = \mathbb{R}$ (e.g. we need to be worrying about errors); and a harder yet case is when Ω is a space of functions (at which point e.g. different ways of measuring errors are not equivalent). We will see in great detail the different mathematical requirements of different kinds of sample spaces.

  An important concept here is that we can take an experiment on ω and form composite experiments for events $A \subseteq \Omega$, $\omega \in A$: for example, if the experiment ω is the number of total phone calls in the UK in a business day, we can define the composite experiment $\omega > 50{,}000{,}000$.

- **Reproducibility:** In the deterministic setting, reproducibility means that an experiment repeated under the same conditions must have the same outcomes. However this is often not the case, as can already be glimpsed from the examples listed above. This led to the concept of

  Definition 1.2 (Stochastic reproducibility)**.** *Consider the event $A \subseteq \Omega$. If for any sequence of repetitions $\omega_1, \omega_2, \ldots, \omega_n, \ldots$ under the same conditions, the number*

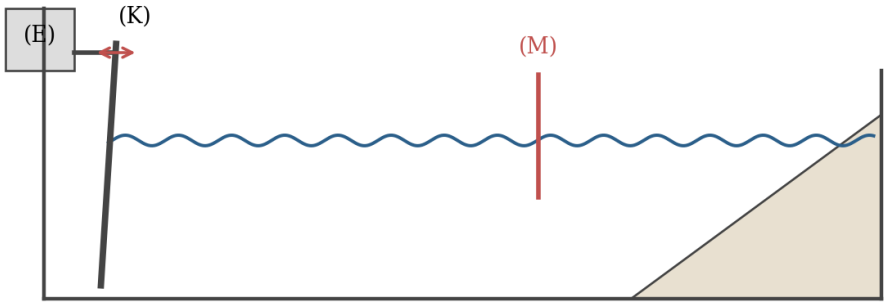
  $$\frac{\#\{n = 1, 2, \ldots, N \,|\, \omega_n \in A\}}{N}$$

  has a deterministic limit as $N \to \infty$, then the experiment $\omega \in A$ is stochastically reproducible.

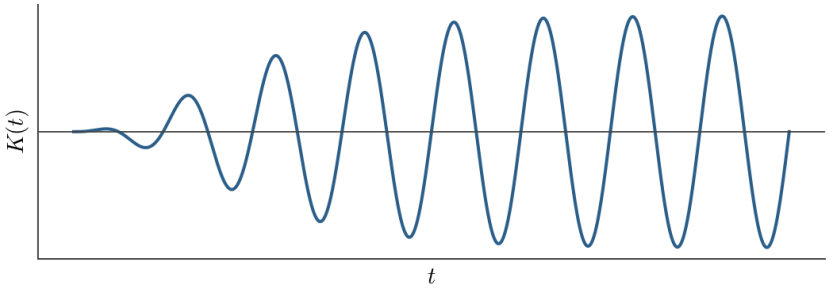
  A note of caution: taken literally, "any sequence of repetitions" is too strong. For a fair coin, an unending sequence of heads is a possible – if extraordinary – sequence of repetitions, and its relative frequency does not converge to $\frac{1}{2}$. The precise statement is the Law of Large Numbers (Chapter 6): the relative frequency converges for *almost every* sequence of repetitions, in a sense that will be made exact. At the modelling level we keep the loose phrasing, with this understanding. Stochastic reproducibility is also called statistical stability, or statistical regularity.

  The idea is that, if the experiment $\omega \in A$ is stochastically reproducible for "all meaningful events" A, then the experiment ω is stochastically reproducible. However, when Ω is uncountable (e.g. $\Omega = \mathbb{R}$,) we have to be more systematic in talking about "all events". We will see this in extended detail in Chapter 3.

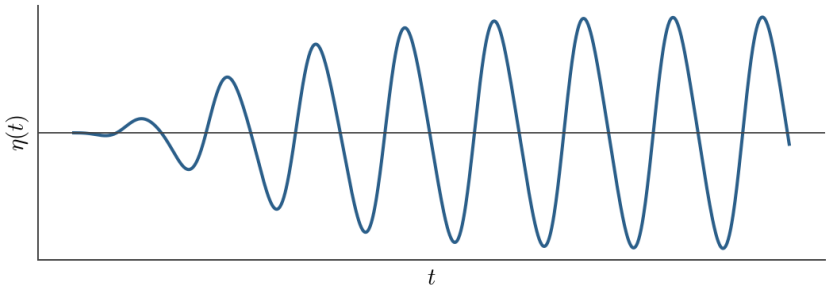
Example 1.3 (An experiment with function-valued outcomes: Mechanically induced waves)**.** *Consider a wave tank with a wavemaker fixed on its end:*



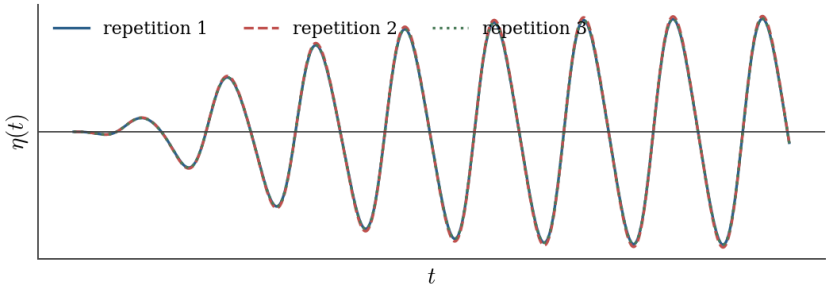
The experimentalist programs the wavemaker so that its tip performs a movement in time determined by $K(t)$:



The elevation of the water surface at point M, $\eta(t)$, *is measured as a function of time:*



Up to instrument and measurement errors, multiple controlled realisations of the same experiment yield the same result:



1.2 Beyond deterministic reproducibility

Example 1.4. *Consider the problem ω of throwing a "fair" die. In that case the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, i.e. the outcomes are the elements of Ω. This is not deterministically reproducible, since identical repetitions will lead to different outcomes.*

Moreover, we can consider the composite experiment $\omega \leqslant 3$. This is now a different experiment, with outcomes 0 (if $\omega \in \{4, 5, 6\}$) and 1 (if $\omega \in \{1, 2, 3\}$)[1]. It is still not deterministically reproducible. Essentially we created a coin toss out of a die.

However, for a large number of repetitions N, the relative frequency

$$\frac{\#\{n = 1, 2, \dots N \,|\, \omega_n \leqslant 3\}}{N}$$

is deterministically reproducible, and will be described as "the probability of the event $A = \{\omega \leqslant 3\}$", and denoted by $P(A)$. By definition, this means that the experiment $\omega \leqslant 3$ is stochastically reproducible.

Elementary considerations on relative frequencies lead to stipulating that

$$\begin{aligned} P(A) &\geqslant 0, \\ P(\Omega) &= 1, \\ A \cap B = \emptyset \quad \Rightarrow \quad P(A \cup B) &= P(A) + P(B). \end{aligned}$$

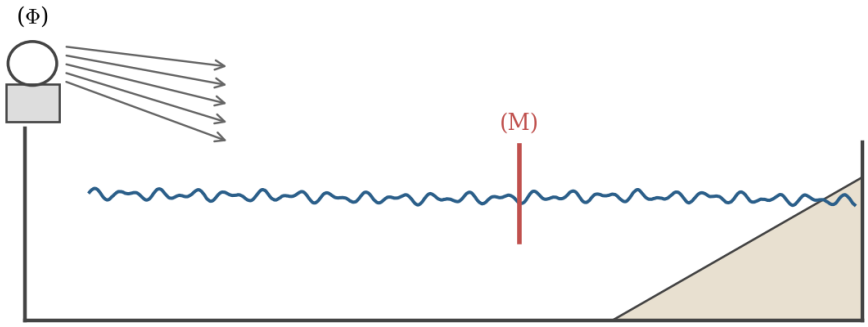
for any "meaningful events" A, B.

Moreover, in this particular example, the meaningful events are all the subsets of Ω – a finite collection, since Ω is finite. This is not the case for large (e.g. uncountable) Ω.

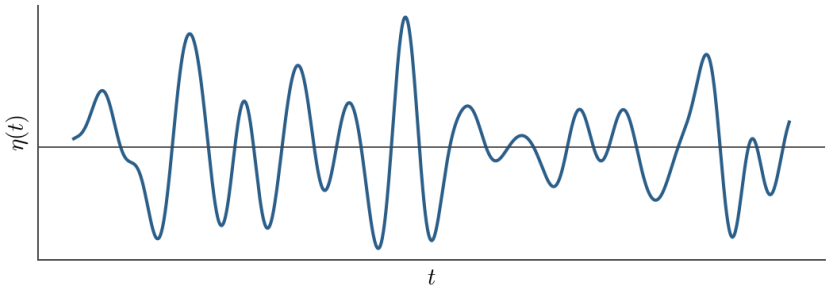
These considerations can be used to analyse experiments of the form $\omega \in A$ even when they are not deterministically reproducible, as long as they are stochastically reproducible. The fully fledged concept of a stochastic experiment is inherently measure-theoretic, and follows later on. Before we proceed to that however, let us look at one more experiment where the outcome is a *function*:

Example 1.5 (Wind waves – a stochastic experiment with function-valued outcomes)**.** *An engineer wants to study how wind generates waves in the sea. To that end, he has set up a large testing tank with a big fan (Φ) blowing over it. Some way downstream, at the point of measurement (M), the elevation of the water surface is recorded as a function of time, $\eta(t)$.*

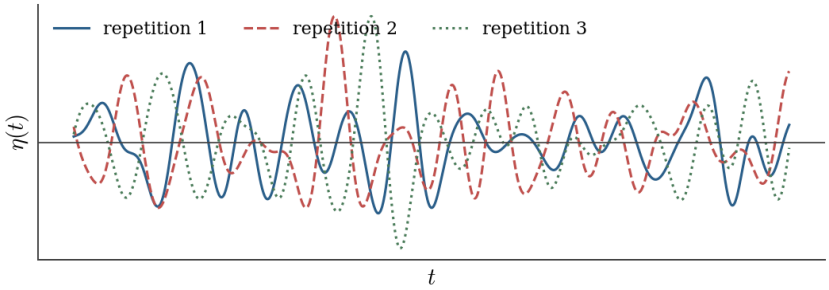
[1] Stochastically reproducible experiments with only two outcomes are called Bernoulli trials, and were one of the earliest attempts to analyse systematically stochastic (i.e. non-deterministically-reproducible) phenomena using relatively simple tools.



A realisation of this experiment can be seen here:



Now, even if the experimenter works triple overtime to make all the conditions identical, subsequent repetitions of the experiment will have different *outcomes:*



This turns out to be a stochastic experiment with function-valued outcomes; these "random functions" are also called "stochastic processes".

Now we can proceed and clearly formulate the following

Definition 1.6 (Random Experiment $(\Omega, \mathcal{E}, P)$)**.** *(i). An experiment is the controlled repetition of a phenomenon, so that the main factors affecting it (the "conditions" of the experiment) are known.*

(ii). The events $E \in \mathcal{E}$, are well defined, and it is possible to check whether they took place.

(iii). The probability of any event $E \in \mathcal{E}$ is stochastically reproducible, and is given by a probability measure $P : \mathcal{E} \to [0, 1]$

$$\lim_{N\to\infty} \frac{\#\{n = 1, 2, \dots N \,|\, \omega_n \in E\}}{N} = P(E)$$

for any sequence ω_n of repetitions of the experiment under the same conditions.

Systematically checking whether an experiment is a random experiment as specified above is a problem in statistics. From now on we will assume we are dealing with random experiments in that sense.

What does "random" mean?

Probability theory can drive surprisingly robust and specific conclusions about very noisy and messy problems, but this can only be achieved drawing on *stochastically reproducible facts*.

Thus a stochastic phenomenon is not at all the same as in the casual use of the term "random" to mean merely "an experiment whose outcome is not known in advance"[2]; not much quantitative analysis can be done if all we know about an experiment is that the outcome is not known in advance.

If an event cannot be meaningfully thought of as having independent identical repetitions, then we cannot possibly do any kind of systematic stochastic modelling. This applies in particular to singular or unique events. Questions like "what is the probability that life appears randomly in the Universe", "what is the date when humanity will become extinct", or "what is the magnitude of the largest earthquake that will ever happen on Earth" are not amenable to stochastic modelling. Any kind of investigation of such questions involves a number of assumptions about "all possible Universes" or something along those lines, and is therefore entirely non-empirical.

On the historical development of probability

Before we proceed to the next chapter, a note is warranted on the development of the broader field of probability. Historically, the first mathematical studies that crystallised as elements of modern probability theory were carried out by Pascal, Fermat and Bernoulli on games of chance in the 1600s. A significant reason was the limitation of mathematics and science of that time with respect to studying more natural problems.

Many textbooks have taken up games of chance examples simply out of continuity with existing studies. However, the modern era of probability theory, including random processes, was reached after a watershed moment: the realisation that Measure Theory is the natural language of probability. Dealing with large sample spaces (especially when outcomes are functions) is only possible in the context of Measure Theory. The following table helps appreciate how recent this development is.

[2]Unfortunately such simplistic statements can be found in otherwise nice textbooks, such as [19].

Calculus		Probability	
Mean Value Theorem: Michel Rolle	1691	Concept of statistical regularity: G. Cardano	1576
De l'Hôpital	1696		
Taylor series: James Gregory	1671	Laws of large numbers: Bernoulli	1713
Brook Taylor	1715	Laplace	1810
Colin Maclaurin	1742	Poisson	1837
Death of Leibniz	1716	Measure Theory: Lebesgue	1901
Death of Newton	1727	Modern foundation of probability	
Modern exposition: A. L. Cauchy	1823	theory: A. N. Kolmogorov	1933
		Random processes: J. L. Doob	1940's

The relatively new and not-yet-settled character of the field can be easily gleaned through some of the most popular textbooks: looking at a handful of textbooks titled "Stochastic Processes" or "Stochastic Differential Equations" one will observe many different selections of material and even more different emphases and flavours. This is very different from the well-settled areas of mathematics, such as calculus and differential equations.

Further reading: Chapters 1–2 of Gnedenko [1] develop the foundations in a similar classical spirit; Kobayashi, Mark and Turin [2] give an applications-first account. The stochastic modelling of wave systems, which inspires the running examples of this course, is developed in depth in Athanassoulis [13].

2 Combinatorics and elementary probability

In this chapter we do probability in the simplest possible setting: that of a **finite sample space**. This setting is rich enough to introduce, through concrete computations, most of the concepts that drive the rest of the course – events, probability measures, conditional probability, independence – while requiring no machinery beyond careful counting. In Chapter 3 we will rebuild the foundations so that they cover infinite sample spaces as well; the present chapter builds intuition and computational fluency and, in Section 2.6, makes clear *why* the rebuilding is necessary.

2.1 Probability on finite sample spaces

Consider a random experiment with a finite sample space,

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}.$$

Here we can take the family of events to be *all* the subsets of Ω, i.e. $\mathcal{E} = \mathcal{P}(\Omega)$, and a probability measure is completely specified by the values it assigns to the individual outcomes,

$$p_j := P[\{\omega_j\}] \geqslant 0, \qquad \sum_{j=1}^{M} p_j = 1,$$

since then, by additivity, for any event $A \subseteq \Omega$

$$P[A] = \sum_{j:\, \omega_j \in A} p_j.$$

In the light of Chapter 1, the numbers p_j are stochastically reproducible relative frequencies: they are what repetitions of the experiment give us.

Definition 2.1 (Classical probability)**.** *If all the outcomes of a finite sample space are equally likely ("equiprobable"),*

$$p_1 = p_2 = \cdots = p_M = \frac{1}{M},$$

then for any event $A \subseteq \Omega$

$$P[A] = \frac{|A|}{|\Omega|} = \frac{\#\ \textit{of favourable outcomes}}{\#\ \textit{of all possible outcomes}}.$$

Remarks:

- In the classical setting, computing probabilities *becomes counting* – hence the importance of the counting toolkit of the next section.
- Words like "fair", "well-shuffled", "at random" are modelling assumptions declaring certain outcomes equiprobable. Whether such an assumption holds for a physical experiment is exactly the question of stochastic reproducibility discussed in Chapter 1.

- Some basic rules follow immediately by counting: $P[A^c] = 1 - P[A]$; if $AB = \emptyset$ then $P[A \cup B] = P[A] + P[B]$; and in general

$$P[A \cup B] = P[A] + P[B] - P[AB],$$

since the outcomes of AB are counted twice in $|A| + |B|$.

Example 2.2. *Throw two fair dice; what is the probability that the sum is* 7*? Take*

$$\Omega = \{(i, j) \mid i, j \in \{1, \ldots, 6\}\}, \qquad |\Omega| = 36,$$

i.e. ordered *pairs, all equiprobable. The favourable outcomes are* $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$, $(6, 1)$, *so* $P = \frac{6}{36} = \frac{1}{6}$.

Observe the importance of choosing the sample space so that its outcomes really are equiprobable: the 21 unordered *pairs* $\{i, j\}$ *are* not *equiprobable (e.g.* $\{1, 6\}$ *is twice as likely as* $\{3, 3\}$*), and counting with them without care leads to wrong answers.*

2.2 The counting toolkit

The product rule. If a selection is performed in k successive stages, with n_1 options in the first stage, n_2 options in the second stage *no matter what was chosen in the first*, and so on, then the total number of possible selections is $n_1 n_2 \cdots n_k$. For example, there are $52 \cdot 51$ ways to draw two cards from a deck in order.

The sum rule. If a selection can be made either from a collection of n_1 options or from a disjoint collection of n_2 options, the total number of options is $n_1 + n_2$.

These two rules generate the four fundamental counts. Let n be the number of distinct objects available, from which we select k:

1. **Ordered selection with repetition:** each of the k slots can be filled in n ways, giving n^k.

2. **Ordered selection without repetition:** the slots can be filled in $n, n-1, \ldots, n-k+1$ ways, giving
$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$
In the special case $k = n$ these are the **permutations** of n objects, $n!$ in total.

3. **Unordered selection without repetition (combinations):** each unordered selection of k distinct objects corresponds to exactly $k!$ ordered ones, so their number is
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

4. **Unordered selection with repetition:** their number is $\binom{n+k-1}{k}$; see the "stars and bars" argument below.

	without repetition	with repetition
ordered	$\dfrac{n!}{(n-k)!}$	n^k
unordered	$\binom{n}{k}$	$\binom{n+k-1}{k}$

Stars and bars. The following three counts are equal:

(i). the number of unordered selections of k objects out of n kinds, with repetition allowed;

(ii). the number of ways k identical objects can be distributed among n distinct containers;

(iii). the number of nonnegative integer solutions of $x_1 + x_2 + \cdots + x_n = k$.

Indeed, a selection as in (i) is completely described by how many objects of each kind were selected – which is exactly (iii) – and x_j can be read as the content of the j^{th} container – which is exactly (ii). To count them, encode each solution of (iii) as a string of k stars and $n-1$ bars: e.g. for $n = 3$, $k = 7$, the solution $2 + 1 + 4$ is encoded as $\star\star \mid \star \mid \star\star\star\star$. Every arrangement of the $k + n - 1$ symbols corresponds to exactly one solution, and there are $\binom{n+k-1}{k}$ ways to choose the positions of the stars.

Permutations with repeated elements. The number of distinct words that can be written using n_1 copies of one letter, n_2 of a second, …, n_r of an r^{th} (with $n = n_1 + \cdots + n_r$) is the **multinomial coefficient**

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!\, n_2! \cdots n_r!},$$

since permuting the identical copies of each letter ($n_1!\, n_2! \cdots n_r!$ ways) does not change the word.

Binomial coefficients. The numbers $\binom{n}{k}$ satisfy

$$\binom{n}{k} = \binom{n}{n-k}, \qquad \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k},$$

the latter (Pascal's rule, which generates Pascal's triangle) by splitting the selections according to whether they include the n^{th} object or not. They owe their name to the **binomial theorem**,

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

which itself is proved by counting: expanding the product $(a+b)(a+b)\cdots(a+b)$, the term $a^k b^{n-k}$ appears once for every choice of k parentheses out of n to contribute an a.
Practice with the toolkit – Pascal's rule, stars and bars, and a counting identity – is Exercise 2.1 in the problem set.

2.3 Classic problems, worked out

Example 2.3 (Roulette)**.** *A roulette wheel has* 37 *sections, numbered* 0 *to* 36*. What is the probability that all the numbers in* 8 *spins are different?*

Solution: *Take as sample space all ordered 8-tuples* $(i_1, \ldots, i_8)$, $i_j \in \{0, 1, \ldots, 36\}$*: assuming the roulette is fair, all* 37^8 *of them are equiprobable. The favourable outcomes are the tuples with distinct entries – ordered selections without repetition – of which there are* $37 \cdot 36 \cdots 30$*. Thus*

$$P = \frac{37 \cdot 36 \cdots 30}{37^8} = \prod_{j=0}^{7} \frac{37 - j}{37} \approx 0.4432.$$

Example 2.4 (The birthday problem)**.** *A class has* n *children (*$n \leqslant 365$*, no twins); assume all birth dates equally likely and ignore February 29. How large must* n *be so that* $P[\textit{some two children share a birthday}] > \frac{1}{2}$*?*

Solution: *As in the roulette example (which is the same problem in disguise!),*

$$P[\textit{all } n \textit{ birthdays distinct}] = \prod_{j=0}^{n-1} \frac{365 - j}{365},$$

which decreases with n *and drops below* $\frac{1}{2}$ *for the first time at* $n = 23$*:*

$$P[\textit{coincidence, } n = 22] \approx 0.476, \qquad P[\textit{coincidence, } n = 23] \approx 0.507.$$

*Most people find this number surprisingly small – the intuition "*23 *children can only cover a small part of the calendar" is about* fixed *birthdays, while here all* $\binom{23}{2} = 253$ pairs *get a chance to collide.*

Example 2.5 (The airport shuttle)**.** *An airport shuttle carries* 25 *passengers and makes* 7 *stops; each passenger is equally likely to get off at any stop, independently of the others. What is the probability that at least one passenger gets off at each stop?*

Solution: *The sample space consists of the* 7^{25} *equiprobable assignments of passengers to stops. Consider the events*

$$A_i = \{\textit{no passenger gets off at the } i^{th} \textit{ stop}\}, \qquad i = 1, \ldots, 7;$$

we want $1 - P[A_1 \cup \cdots \cup A_7]$*. For any* j *distinct indices* $i_1 < \cdots < i_j$*,*

$$P[A_{i_1} A_{i_2} \cdots A_{i_j}] = \left(\frac{7 - j}{7}\right)^{25},$$

since all passengers must then choose among the remaining $7 - j$ *stops. The inclusion-exclusion formula (cf. Lemma 3.2) gives*

$$P[\textit{every stop is used}] = \sum_{j=0}^{7} (-1)^j \binom{7}{j} \left(\frac{7 - j}{7}\right)^{25} \approx 0.8562.$$

Example 2.6 (A game of dice)**.** *Two players take turns throwing a fair die; whoever rolls a* 6 *first wins. What is the probability that the player who rolls first wins?*

Solution: *The first player wins on their* $(k+1)^{st}$ *throw (*$k = 0, 1, 2, \ldots$*) iff the first* $2k$ *rolls of the game show no* 6 *and roll* $2k+1$ *shows a* 6*; by the product rule this has probability*

$\left(\frac{5}{6}\right)^{2k}\frac{1}{6}$. *These events are disjoint, so*

$$P[\textit{first player wins}] = \sum_{k=0}^{\infty}\left(\frac{25}{36}\right)^{k}\frac{1}{6} = \frac{\frac{1}{6}}{1-\frac{25}{36}} = \frac{6}{11}.$$

Going first helps: $\frac{6}{11} > \frac{1}{2}$. *Observe that here the sample space is* countably infinite*, and the computation is powered by the geometric series – a first hint that counting alone will not carry us forever.*

Two more classics – Lord Yarborough's wager and the four-piles problem – are left for the problem set (Exercises 2.2 and 2.3).

2.4 Conditional probability and independence, via examples

If we know that the event B took place, how should the probability of another event A be updated? In the classical setting the answer is transparent: the effective sample space shrinks to B, its outcomes remain equally likely, and so the updated probability of A should be

$$\frac{|AB|}{|B|} = \frac{|AB|/|\Omega|}{|B|/|\Omega|} = \frac{P(AB)}{P(B)}.$$

The last expression makes sense beyond the equiprobable case (in terms of relative frequencies: among the many repetitions in which B occurred, the fraction in which A also occurred), and we denote it

$$P(A|B) := \frac{P(AB)}{P(B)}, \qquad P(B) > 0,$$

read "the probability of A given B."

This can be easily motivated through simple examples:

Example 2.7. *Two cards are drawn randomly in succession from a deck of cards.*

(i). What is the unconditional probability that the second card is an ace?

(ii). What is the conditional probability that the second card is an ace, if the first card was also an ace?

Solution: Let A denote the event that the second card is an ace, and B the event that the first card is an ace. So question (i) seeks $P(A)$, while question (ii) seeks $P(A|B)$.

Observe that

$$A = (A \cap B) \cup (A \cap B^c) = AB + AB^c$$

in shorthand notation. Moreover, since the events AB, AB^c are mutually exclusive,

$$P(A) = P(AB) + P(AB^c).$$

Now we can count: all the possible ways to draw two cards from a $52-$card deck are $52 \cdot 51$. Of these $4 \cdot 3$ are favourable to the event AB, and $48 \cdot 4$ are favourable to the event AB^c, thus

$$P(A) = \frac{4 \cdot 3 + 48 \cdot 4}{52 \cdot 51} = \frac{204}{2652} = \frac{1}{13}.$$

If the first card is known to be an ace, then at the second draw there are 3 *aces within* 51 *cards, so*

$$P(A|B) = \frac{3}{51} = \frac{1}{17}.$$

Observe that

$$P(B) = \frac{4}{52} = \frac{1}{13}.$$

One easily checks that the formula for conditional probability $P(A|B)$ (read "the probability of A under the condition that B took place"),

$$P(A|B) = \frac{P(AB)}{P(B)}, \qquad \text{defined whenever } P(B) > 0,$$

holds in our example. However, counting cards doesn't exactly prove this formula as much as verify it. For now we adopt the formula as our working definition on finite sample spaces; the deeper reconstruction – in which conditioning is shown to define a genuine new probability space – is carried out in Chapter 4.

The multiplication rule. Rearranging the definition, $P(AB) = P(A|B)P(B) = P(B|A)P(A)$, and more generally, for any events with $P(A_1A_2\cdots A_{N-1}) > 0$,

$$P(A_1A_2\cdots A_N) = P(A_1)\,P(A_2|A_1)\,P(A_3|A_2A_1)\cdots P(A_N|A_{N-1}\cdots A_1),$$

as one sees by expanding each conditional probability and cancelling the telescoping product. This rule turns sequential experiments into products, as in the following

Example 2.8. *Lots of 100 computer chips are subject to quality control according to the following process: 5 chips are successively selected at random out of the 100, and if there is even one defective among them, the lot is rejected. If a specific lot contains 5 defective chips, what is the probability that it will be accepted?*

Solution: Denote A_n, $n \in \{1,2,3,4,5\}$, *the event that the* n^{th} *object is not defective. The probability that the lot is accepted is* $p = P(A_1A_2A_3A_4A_5)$, *which can be recast as*

$$p = P(A_1A_2A_3A_4A_5) = P(A_1)\ P(A_2|A_1)\ P(A_3|A_2A_1)\ P(A_4|A_3A_2A_1)\ P(A_5|A_4A_3A_2A_1).$$

Clearly, $P(A_1) = \frac{95}{100}$. *Moreover,*

$$P(A_2|A_1) = \tfrac{94}{99}, \qquad P(A_3|A_2A_1) = \tfrac{93}{98} \qquad P(A_4|A_3A_2A_1) = \tfrac{92}{97}, \qquad P(A_5|A_4A_3A_2A_1) = \tfrac{91}{96}.$$

So finally

$$p = \frac{95\cdot 94\cdot 93\cdot 92\cdot 91}{100\cdot 99\cdot 98\cdot 97\cdot 96} \approx 0.7696.$$

Exercise: Prove that the probability of the lot being accepted would be the same if the 5 *chips to be checked were randomly selected at once (as opposed to successively).*

Total probability and Bayes' rule. If the events $D_1, D_2, \ldots, D_r$ form a partition of Ω (they are disjoint and their union is Ω; cells with $P(D_j) = 0$ are simply omitted from the

sums below), then, splitting A over the partition and using the multiplication rule,

$$P(A) = \sum_{j=1}^{r} P(AD_j) = \sum_{j=1}^{r} P(A|D_j)P(D_j),$$

which is called the **total probability formula**. Combining it with $P(D_j|A)P(A) = P(A|D_j)P(D_j)$ we obtain **Bayes' rule**,

$$P(D_j|A) = \frac{P(A|D_j)P(D_j)}{\sum_{m=1}^{r} P(A|D_m)P(D_m)},$$

which "inverts" the direction of the conditioning: it computes the probability of a *cause* D_j given an observed *effect* A. The following classic example shows how counterintuitive this inversion can be.

Example 2.9 (False positives)**.** *A rare but potentially fatal disease has an incidence of 1 in* 10^5 *in the population at large. There is a diagnostic test available, but it is quite crude. If you have the disease, the test is positive with probability* $\frac{9}{10}$*; if you do not, the test is positive with probability* $\frac{1}{20}$*. Your test result is positive. What is the probability that you have the disease?*

Solution: *Write* D *for the event that you have the disease, and* T *for the event that the test is positive. By virtue of Bayes' rule*

$$\begin{aligned} P(D \mid T) &= \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P\left(T \mid D^{\mathrm{c}}\right) P\left(D^{\mathrm{c}}\right)} \\ &= \frac{\frac{9}{10} \cdot \frac{1}{10^5}}{\frac{9}{10} \cdot \frac{1}{10^5} + \frac{1}{20} \cdot \frac{10^5-1}{10^5}} \approx 0.00018. \end{aligned}$$

It is more likely that the result of the test is incorrect than that you have the disease. On the other hand, the positive test result makes it ≈ 18 *times more likely to have the disease compared to the default baseline risk. (This example follows Grimmett and Welsh [4].)*

Example 2.10 (Monty Hall problem)**.** *Consider the following game show: there are three doors; one has behind it a car, and the other two have behind them goats. The game is played as follows:*

1. *The contestant chooses one door as his selection.*
2. *Then the host opens one of the two unchosen doors. He has to open a door with a goat behind it. That way, there remain two unopened doors – the one chosen originally by the contestant and one more. If both unopened doors have goats behind them, he has to choose between them with probability* $\frac{1}{2}$*.*
3. *The host now gives the opportunity to the contestant to either keep his choice of door, or choose the other unopened door.*

(a). What is the probability to win if the contestant always keeps his original choice of door?

(b). What is the probability to win if the contestant always changes door in step 3?

Solution: *Originally, the contestant has the same information for each door. Thus, whichever door he chooses, he has probability* $\frac{1}{3}$ *to pick the car:*

$$A = \{\textit{contestant originally chose the car}\}, \qquad P[A] = \frac{1}{3}.$$

Whatever happens next, as long as he doesn't alter his choice, will not change this. So the answer for (a) *is* $\frac{1}{3}$.

However, observe that

$$A^c = \{\textit{contestant originally did not choose the car}\} = \{\textit{the car is in the unopened door in Step 3}\}$$

and $P(A^c) = 1 - P(A) = \frac{2}{3}$.

Thus, denoting

$$B = \{\textit{contestant changes door in step 3 and wins}\}$$

we compute

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}.$$

Remark: *This is famous as a counterintuitive problem. Listing of the sample space in detail and simulations of a large number of games have been used to accompany and validate this computation.*

It must be said that the precise setting of the problem & rules matter. For example, assuming no a priori information about the doors is not as obvious as it may sound: in a "real life" setting the doors would be identifiable (left, centre, right) and a contestant could e.g. look at the statistics from previous games to see which door was more likely to have the car behind it. If the producers did not place the car in a truly random way, this would likely be picked up. Another issue is the selection of the door to open; e.g. if the host always opened the right-most allowed door, that would also be picked up etc.

Independence. Sometimes the knowledge that B occurred does *not* change the probability of A, i.e. $P(A|B) = P(A)$. By the multiplication rule this is equivalent (when $P(A), P(B) > 0$) to the symmetric condition

$$P(AB) = P(A)P(B),$$

which we adopt as the definition: such events are called **(stochastically) independent**. For example, if a card is drawn, *replaced*, the deck reshuffled, and a second card drawn, then the events "first card is an ace" and "second card is an ace" are independent; drawing *without* replacement they are not – as computed in Example 2.7, the conditional probability changes from $\frac{1}{13}$ to $\frac{1}{17}$. The systematic treatment of independence (for more than two events the story is subtler than one might expect) is taken up in Chapter 4.

2.5 A stretch example: the Pascal distribution

The following worked problem combines everything in this chapter – the product rule, combinations, and a series summation – and is worth studying in detail. It will reappear later in the course, as the **Pascal** (or **negative binomial**) **distribution**.

Example 2.11 (The Pascal distribution)**.** *Independent trials, each succeeding with probability $p \in (0,1)$ (and failing with probability $q := 1-p$), are repeated indefinitely. Fix $r \in \mathbb{N}$. What is the probability $f(n)$ that the r^{th} success occurs exactly at the n^{th} trial?*

Solution: *The event in question happens if and only if*

(i). there are exactly $r-1$ successes among the first $n-1$ trials, and

(ii). the n^{th} trial is a success.

Any specific *arrangement of $r-1$ successes and $(n-1)-(r-1) = n-r$ failures in the first $n-1$ trials has probability $p^{r-1}q^{n-r}$ by the product rule, and there are $\binom{n-1}{r-1}$ such arrangements (choose which trials succeed). Multiplying finally by p for the n^{th} trial,*

$$f(n) = \binom{n-1}{r-1} p^r q^{n-r}, \qquad n = r, r+1, r+2, \ldots$$

Do these probabilities add up to 1? *Substituting $m = n-r$:*

$$\sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r q^{n-r} = p^r \sum_{m=0}^{\infty} \binom{m+r-1}{m} q^m = p^r (1-q)^{-r} = 1,$$

where we used the **negative binomial series**

$$\sum_{m=0}^{\infty} \binom{m+r-1}{m} x^m = \frac{1}{(1-x)^r}, \qquad |x| < 1,$$

which itself has a counting proof: writing $(1-x)^{-r} = \left(\sum_{i \geqslant 0} x^i\right)^r$, the coefficient of x^m counts the ways to write $m = m_1 + \cdots + m_r$ with nonnegative integers – and by stars and bars this is $\binom{m+r-1}{m}$. The interpretation of the normalisation is that, with probability 1, the r^{th} success eventually occurs.

Special case $r = 1$: *$f(n) = pq^{n-1}$ is the* **geometric distribution** *– the waiting time for the first success. This is precisely the structure of Example 2.6 (with success = "a 6 is rolled").*

Looking ahead: *once expectations are available (Chapter 5) one can show that the average waiting time for the r^{th} success is $\frac{r}{p}$ – e.g. on average 6 rolls of a fair die per 6 rolled – and this example will return in the Chapter 5 problem set as the Pascal distribution.*

2.6 The limits of counting

This chapter may suggest that probability theory *is* counting, dressed up. For finite – even countably infinite – sample spaces this view is workable. The following famous example shows what goes wrong beyond that.

Example 2.12 (Bertrand's paradox)**.** *A chord of the unit circle is drawn "at random". What is the probability that it is longer than* $\sqrt{3}$*, the side of the inscribed equilateral triangle? Here are three natural interpretations of "at random":*

(i). Random endpoints: *the two endpoints are chosen uniformly on the circle. Fixing the first endpoint, the chord is longer than* $\sqrt{3}$ *iff the second endpoint falls in the opposite third of the circle:*

$$P = \frac{1}{3}.$$

(ii). Random distance: *the chord is taken perpendicular to a fixed radius, its distance* d *from the centre uniform in* $[0, 1]$*. The chord is longer than* $\sqrt{3}$ *iff* $d < \frac{1}{2}$*:*

$$P = \frac{1}{2}.$$

(iii). Random midpoint: *the midpoint of the chord is chosen uniformly in the disk. The chord is longer than* $\sqrt{3}$ *iff its midpoint lies within distance* $\frac{1}{2}$ *of the centre:*

$$P = \frac{(\frac{1}{2})^2\pi}{1^2\pi} = \frac{1}{4}.$$

Three equally "natural" postulates of equiprobability – three different answers. With an uncountable sample space, "equally likely" has no intrinsic meaning: a probability measure must be specified, *and different measures are different models. There is no paradox once this is understood; but it does mean that the classical formula "favourable over possible" cannot be the foundation of the theory.*

Having just spent a whole chapter using games of chance as our driving examples, a critical note is warranted on **the casino approach**, i.e. on treating games of chance as the paradigm for all of stochastic modelling.

This approach, found in many textbooks still today, represents historical continuity with the pre-modern studies and their limitations in dealing with continuous random variables. Indeed, a discrete (countable) sample space is a common feature of all "games of chance" paradigms. However, this is not the only mark of the casino approach: conceptually, it is built upon the dogma of "pure randomness". The investigator doesn't need to worry about whether the repetitions are under "the same conditions", or what their laws are, since this is always assumed known by construction. Some basic "elementary events" are considered, and they are always taken to be equiprobable (fair coin toss, fair dice etc). Then the problem becomes essentially **combinatorial**, i.e. a counting problem. This has the advantage of simple, easy to grasp examples on key aspects (such as Bernoulli trials, relative frequency etc).

However, this approach also has some important disadvantages:

(i). In more complicated problems (larger sample space Ω) it is not always possible to determine the "elementary events" that should be equiprobable. Joseph Bertrand developed

several paradoxes in problems with uncountable sample spaces, where different "natural" choices of equiprobable events lead to different results. The events are way too complicated for the equiprobability of elementary events to make sense. So, strictly speaking, what Bertrand proved is that on an uncountable sample space the instruction "choose uniformly at random" does *not*, by itself, specify a probability measure: one must state explicitly *which* quantity is uniformly distributed. Different choices are different models – and give different answers.

(ii). More broadly, the "casino approach" trivialises the basic question of whether we are indeed looking at different realisations of the same experiment, what are the basic assumptions we make about our stochastic phenomenon, and how we test them[3].

(iii). Finally, a pedagogical over-reliance on casino examples (finite sample spaces, equiprobable events) when studying stochastic methods trivialises the measure theory aspect of understanding the families of events and their structure, both in the modelling/conceptual sense as discussed above, as well as in the technical sense – which is equally important in quantitative studies.

So the casino approach to probability theory, if taken too far, comes at the expense of conceptual, scientific and technical understanding.

Where does this leave us? We need a framework that works for an arbitrary sample space Ω: a systematic way to say which subsets of Ω are events, what a probability measure is, and how it behaves under limiting operations. This is the subject of the next chapter.

Further reading: Grimmett and Welsh [4] is a crisp modern introduction to probability, covering the material of this and the following chapters; Tijms [3] is a rich source of problems and paradoxes in the spirit of this chapter; Lerma [5] covers the counting toolkit in greater depth.

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 2.1. *1. Prove Pascal's rule both combinatorially (by splitting the selections according to whether they include the n^{th} object) and algebraically.*

[3]For example, stipulating that the price of a stock performs an unbiased random walk around its "true value" means that there never are any bubbles. *Detecting* a bubble by data analysis is much more interesting (and profitable) than simply denying that bubbles exist.

The need to go through the basics and ask ourselves whether we have a proper stochastic experiment can be summarised in the following story, from "The Black Swan" [20]:

Someone asks Dr. John and Fat Tony to "assume that a coin is fair, i.e., has an equal probability of coming up heads or tails when flipped. I flip it ninety-nine times and get heads each time. What are the odds of my getting tails on my next throw?"

Dr. John: the odds are not affected by the previous outcomes so the odds must still be 50:50.

Fat Tony: The coin gotta be loaded. This ain't a fair game.

While the Dr is correct, Fat Tony has a point too. After 99 heads of a *fair* coin toss in a row, tails is indeed not "due"; however, if we did observe 99 heads in a row (or its equivalent, an event with probability $2^{-99} \approx 10^{-30}$) in a real-life setting, it *would be a very good idea to re-examine whether the coin is indeed fair.*

2. *Write out in detail the equivalence of the three "stars and bars" formulations.*

3. *How many solutions does* $x_1 + x_2 + \cdots + x_n = k$ *have in* positive *integers* ($x_j \geqslant 1$)*?*

4. *Setting* $a = b = 1$ *in the binomial theorem yields* $\sum_k \binom{n}{k} = 2^n$*. Which familiar fact about the number of subsets of a finite set does this recover?*

Exercise 2.2 (Lord Yarborough's wager)**.** *Lord Yarborough used to offer fellow bridge players the following wager: if their bridge hand (*13 *cards randomly drawn from a well-shuffled deck) contained no card higher than a nine (figures and aces counting as higher), he would pay them £ 1000; otherwise they would pay him £ 1.*

What are the odds that Yarborough would lose?

(Hint: both an ordered sample space – drawing the cards one at a time – and an unordered one – a set of 13 cards taken at once – work, and provide two solutions. There are 32 cards no higher than a nine. Answer: $\binom{32}{13}/\binom{52}{13} \approx \frac{1}{1828}$, so the wager was profitable for his lordship.)

Exercise 2.3 (Four piles)**.** *A well-shuffled deck is divided into four piles of* 13 *cards each. What is the probability that each pile contains exactly one ace?*

(Hint: it suffices to track the 4 positions of the aces among the 52 slots – all $\binom{52}{4}$ position sets being equiprobable. Answer: $13^4/\binom{52}{4} \approx 0.1055$.)

Exercise 2.4 (Fifty-fifty?)**.** *A fair coin is tossed* 100 *times. What is the probability of getting exactly* 50 *heads?*

(Hint: define the sample space to be all sequences of length 100 of 1s and 0s. To evaluate the result, use Stirling's approximation, $n! \approx \sqrt{2\pi n}\, n^n e^{-n}$.)

Exercise 2.5 (Milk containers)**.** *In how many ways can* 7 *identical quarts of milk be distributed among* 3 *distinct containers? Solve it by stars and bars, and check your answer by relating it to a count from Section 2.2.*

Exercise 2.6 (Two urns, I)**.** *Urn I contains* 3 *white and* 4 *black balls; Urn II contains* 2 *white and* 6 *black balls. You pick a ball at random from Urn I and place it in Urn II; then you pick a ball at random from Urn II. What is the probability that it is black?*

Exercise 2.7 (Two urns, II)**.** *With the urns as above (in their original composition), you pick one of the two urns at random (probability* $\frac{1}{2}$ *each) and draw a ball at random from it. Given that the ball is black, what is the probability that you picked Urn I?*

(Hint: Bayes' rule.)

3 The probability space $(\Omega, \mathcal{E}, P)$

3.1 Sample space Ω

Consider an experiment; we have already denoted by Ω the sample space, i.e. the set of all possible outcomes $\omega \in \Omega$. ω is called the outcome, or realisation.

3.2 Event σ-algebra $\mathcal{E}$

We will also need the family $\mathcal{E}$ of "all interesting / meaningful events".

Set theory	Probabilistic interpretation
$\omega \in A$	the event A is realised
$\omega \in A^c$	the event A is not realised
$\omega \in A \cup B$	at least one of the events A, B is realised
$\omega \in A \cap B$	both of the events A, B are realised
$\omega \in A \setminus B$	A is realised but B is not
$\emptyset$	impossible event
Ω	certain event
$A \subseteq B$	A implies B

Here we need to refresh some basic facts about sets. Certainly $\mathcal{E}$ will be a subset of the powerset,

$$\mathcal{P}(\Omega) = \{A | A \subseteq \Omega\}.$$

The powerset is in natural bijection with $2^\Omega = \{f : \Omega \to \{0,1\}\}$ through the correspondence

$$\pi : \mathcal{P}(\Omega) \to 2^\Omega : A \mapsto \pi(A) = f$$

where

$$\omega \in A \qquad \Leftrightarrow \qquad f(\omega) = 1.$$

The cardinality of the powerset is $|2^\Omega| = 2^{|\Omega|}$, e.g.

$$\Omega = \{1,2,3\} \quad \Rightarrow \quad 2^\Omega = \Big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,3\}, \{1,2,3\}\Big\}$$

For infinite sets, $|2^\Omega| = 2^{|\Omega|}$ is strictly larger than $|\Omega|$.
(Recalling why $\mathbb{Q}$ is countable while $[0,1]$ is uncountable is Exercise 3.2 in the problem set.)

It turns out that for infinite sets $\mathcal{P}(\Omega)$ is too large, and we typically only need a "small" subset of the powerset. There are several natural closure requirements on the class of "all interesting events":

$$\begin{aligned} A \in \mathcal{E} \quad &\Rightarrow \quad \Omega \setminus A = A^c \in \mathcal{E}, \\ A, B \in \mathcal{E} \quad &\Rightarrow \quad A \cup B,\ A \cap B \in \mathcal{E} \\ A_n \in \mathcal{E}\ \forall n \quad &\Rightarrow \quad \bigcup_n A_n,\ \bigcap_n A_n \in \mathcal{E} \end{aligned}$$

These mean that $\mathcal{E}$ is a σ-algebra.

If closure is required only under *finite* unions (equivalently, finite intersections), the family is called just an **algebra**; the prefix σ- marks closure under countably infinite operations. Several equivalent packagings of the definition of a σ-algebra are in circulation; proving their equivalence is Exercise 3.3 in the problem set.

3.3 Probability measure P, elementary properties and continuity

Definition 3.1 (Axioms of probability)**.** *The probability $P : \mathcal{E} \to \mathbb{R}$ is a set function which assigns values to events $E \in \mathcal{E}$ and satisfies the following axioms (Kolmogorov's axioms of probability):*

(i). $\forall E \in \mathcal{E}, \qquad P(E) \geqslant 0.$

(ii). $P(\Omega) = 1.$

(iii). [**σ-additivity**] *If $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{E}$ are disjoint events, $A_i \cap A_j = \emptyset \ \forall i \neq j$, then*

$$P(\bigcup_n A_n) = \sum_n P(A_n).$$

In measure-theory terminology, P is a finite, nonnegative measure on Ω.

If property (iii) was enforced only for finite unions /sums, we would get a finitely-additive set function, which means we wouldn't be allowed to take limits!

Lemma 3.2 (Elementary properties of the probability measure)**.** *For every $A, B, A_n \in \mathcal{E}$:*

1. $P(A^c) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. $0 \leqslant P(A) \leqslant 1$
4. $A \subseteq B \quad \Rightarrow \quad P(A) \leqslant P(B)$
5. $P(A \cup B) \leqslant P(A) + P(B)$
6. *Inclusion exclusion formula:*
   $P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B) = P(A) + P(B) - P(A \cap B)$
7. *High-order Inclusion exclusion formula:*

$$P(A_1 \cup \cdots \cup A_n) = \sum_{1\leqslant i\leqslant n} P(A_i) - \sum_{1\leqslant i_1<i_2\leqslant n} P(A_{i_1}A_{i_2}) + \sum_{1\leqslant i_1<i_2<i_3\leqslant n} P(A_{i_1}A_{i_2}A_{i_3}) + \dots$$
$$\cdots + (-1)^{n+1}P(A_1A_2A_3\dots A_n) = \sum_{j=1}^{n}(-1)^{j+1} \sum_{1\leqslant i_1<\cdots<i_j\leqslant n} P(A_{i_1}A_{i_2}\dots A_{i_j})$$

Proof: Properties 1–6 follow directly from the axioms – proving them is Exercise 3.1 in the problem set (*in proving each statement it helps to use the ones before it*). Property 7 follows from Property 6 by induction – Exercise 3.12. □

The list of properties above continues with properties expressing the **continuity** of the probability measure along limits of sequences of events; the simplest instance is

$$B_n \subseteq B_{n+1}\ \forall n \qquad \Longrightarrow \qquad P(\bigcup_n B_n) = \lim_n P(B_n).$$

These properties, together with the machinery of lim inf and lim sup for sequences of sets that they require, are developed in the technical Section 3.5 at the end of this chapter. That section can be skipped on a first reading, but its results are used repeatedly in the sequel (e.g. through the Borel-Cantelli Lemma in the study of Brownian motion).

3.4 Representations of a measure

The thought of performing a random experiment brings up interesting mathematical questions, such as:

- what is the smallest class of events $\underline{\mathcal{E}}$ so that if we know P on $\underline{\mathcal{E}}$ we can extend it to all of $\mathcal{E}$? (For $\Omega = \mathbb{R}$, all the left-open, right-closed intervals $\{(a, b]\}$ suffice.)
- why do we bother studying σ-algebras of events, and not take just the powerset (all possible subsets of Ω) as in the finite case? For uncountable Ω, the natural measures typically cannot be extended to the whole powerset: there are "too many events" in the powerset, and it is not possible in general to define all their probabilities in a way that makes sense. A concrete example is the existence of non-Lebesgue-measurable subsets of $[0, 1]$ (so that "uniform length" cannot be defined for all subsets of $[0, 1]$).
- In practice the specification of $\mathcal{E}$ and P go together; e.g. given some semi-algebra $\underline{\mathcal{E}}$ on which P is known, $\mathcal{E}$ is a σ-algebra containing $\underline{\mathcal{E}}$ on which P can be consistently extended.

Let's unpack the previous questions:

Definition 3.3 (Semi-algebra)**.** *A family of sets $\underline{\mathcal{E}} \subseteq \mathcal{P}(\Omega)$ is called a semi-algebra if it has the following properties:*

1. $\emptyset, \Omega \in \underline{\mathcal{E}}$
2. $A, B \in \underline{\mathcal{E}} \quad \Rightarrow \quad A \cap B \in \underline{\mathcal{E}}$
3. $A, B \in \underline{\mathcal{E}} \quad \Rightarrow \quad \exists N \in \mathbb{N}, \{C_n\}_{n=1}^N \subseteq \underline{\mathcal{E}}$ *such that* $A \setminus B = \bigcup_{n=1}^N C_n$

(In property 3 the sets C_n can moreover be taken disjoint – Exercise 3.4.)

Theorem 3.4 (Carathéodory extension theorem – stated without proof)**.** *Let $\underline{\mathcal{E}}$ be a semi-algebra on Ω and let $P : \underline{\mathcal{E}} \to [0, 1]$, $P(\Omega) = 1$, be σ-additive on $\underline{\mathcal{E}}$ (i.e. whenever a countable disjoint union of elements of $\underline{\mathcal{E}}$ belongs to $\underline{\mathcal{E}}$, its measure is the sum of the measures). Then P extends uniquely to a probability measure on the σ-algebra generated by $\underline{\mathcal{E}}$.*

This is the workhorse that turns a formula on simple sets into a full probability measure; it is not a routine verification, and its proof belongs to a measure theory course (see [7], [8]). Verifying its hypotheses and applying it – to construct Lebesgue measure on $(0, 1]$ and the probability measure of an arbitrary cdf on $\mathbb{R}$ – is the content of Exercises 3.5, 3.6 and 3.7 in the problem set.

The smallest σ-algebra on $\mathbb{R}$ which contains all intervals is the Borel family $\mathcal{B}$. This is strictly smaller than the σ-algebra of all Lebesgue-measurable sets $\mathcal{M}$, but in our context the Borel sets are perfectly sufficient.

The abstract machinery of measure theory is very powerful, and can handle with the same proofs many different situations. However defining measures on function spaces (i.e. a situation where Ω is a function space) represents a big leap, as many key facts (that did hold as long as e.g. $\Omega \approx \mathbb{R}$) no longer hold. Historically, "Stochastic Processes" forced people to come up with "Basic Measure Theory when Ω is a Set of Functions (instead of a set of numbers)".

3.5 Technical toolkit: limits of sets and continuity of the measure *

This section collects the more technical material of the chapter: lim inf and lim sup for sequences of sets, the continuity properties of probability measures, and the Borel-Cantelli Lemma. It continues Lemma 3.2.

Recap: lim sup and lim inf of sequences of real numbers. Let us briefly recall the concepts of lim sup and lim inf in the context of real numbers: let $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of real numbers.

If there exists a subsequence a_{k_n} such that $\lim_n a_{k_n} = +\infty$ we will say that the sequence is unbounded from above. For any sequence that is not unbounded from above, there exists some upper bound $M \in \mathbb{R}$ such that $a_n \leqslant M$ for all $n \in \mathbb{N}$. The smallest number M that can play this role is called the **least upper bound**, or **supremum** of the sequence.

For example, $a_n = \frac{1}{n}$ has a supremum of 1, which is also its maximum. However, the sequence $b_n = -\frac{1}{n}$ has a supremum of 0, which is not a maximum (i.e. this value is not attained for any index $n \in \mathbb{N}$). The existence of the supremum for every bounded set of real numbers is equivalent to the completeness of the real numbers (i.e. this would fail for the rational numbers).

Similarly, the **infimum** or **greatest lower bound** exists for sets bounded from below.

Given any sequence a_n, one can always define the

$$\liminf_n a_n := \lim_n \left(\inf_{k \geqslant n} \{a_k\} \right).$$

(If the sequence is not bounded below, we will get $\liminf_n a_n = -\infty$.)

(Why does that limit always exist? – Exercise 3.8.)

Similarly,

$$\limsup_n a_n := \lim_n \left(\sup_{k \geqslant n} \{a_k\} \right).$$

It can be seen (Exercise 3.8) that

$$\lim_n a_n \text{ exists} \quad \iff \quad \liminf a_n = \limsup a_n.$$

Definition 3.5 (lim sup and lim inf of sequences of sets)**.** *Given a sequence of sets A_n, denote*

$$\limsup_n A_n := \bigcap_{n\in\mathbb{N}} \bigcup_{m\geqslant n} A_m,$$
$$\liminf_n A_n := \bigcup_{n\in\mathbb{N}} \bigcap_{m\geqslant n} A_m.$$

Lemma 3.6 (Characterisations and monotone limits)**.** *For any sequence of sets,*

$$\limsup_n A_n = \{\omega \in \Omega| \quad \omega \text{ belongs to infinitely many of the } A_n\}$$

and

$$\liminf_n A_n = \{\omega \in \Omega| \quad \omega \text{ belongs to all of the } A_n \text{ except possibly a finite number of them}\}.$$

Moreover, if A_n is increasing, i.e. $n > m \Rightarrow A_n \supseteq A_m$, then

$$\limsup_n A_n = \liminf_n A_n = \bigcup_n A_n.$$

Similarly, if A_n is decreasing, i.e. $n > m \Rightarrow A_n \subseteq A_m$, then

$$\limsup_n A_n = \liminf_n A_n = \bigcap_n A_n.$$

Proof: See Exercise 3.9 in the problem set. □

Lemma 3.7 (lim inf and lim sup within σ-algebras)**.** *Let $\mathcal{A}$ be a σ-algebra, and $A_n \in \mathcal{A}$ $\forall n$. Then*

$$\limsup_n A_n := \bigcap_{n\in\mathbb{N}} \bigcup_{m\geqslant n} A_m \in \mathcal{A},$$
$$\liminf_n A_n := \bigcup_{n\in\mathbb{N}} \bigcap_{m\geqslant n} A_m \in \mathcal{A},$$

and that

$$\liminf_n A_n \subseteq \limsup_n A_n.$$

Proof: Membership in $\mathcal{A}$ follows from the closure of a σ-algebra under countable unions,

countable intersections and complements – Exercise 3.10. For the inclusion:

$$x \in \liminf_n A_n \iff x \in \bigcup_{n\in\mathbb{N}} \bigcap_{m\geqslant n} A_m \iff \exists N \in \mathbb{N} : x \in \bigcap_{m\geqslant N} A_m \iff$$

$$\iff \exists N \in \mathbb{N} : x \in A_m \ \forall m \geqslant N \implies$$

$$\implies x \in \bigcup_{m\geqslant n} A_m \ \forall n \iff x \in \bigcap_{n\in\mathbb{N}} \bigcup_{m\geqslant n} A_m \iff$$

$$\iff x \in \limsup_n A_n.$$

□

Lemma 3.8 (Continuity properties of the probability measure)**.** *Continuing Lemma 3.2: for every $A_n, B_n \in \mathcal{E}$,*

8. *If $B_n \in \mathcal{E}$ is an increasing sequence, $B_n \subseteq B_{n+1}$, then*

$$P(\bigcup_n B_n) = \lim_n P(B_n).$$

9. *If $B_n \in \mathcal{E}$ is a decreasing sequence, $B_n \supseteq B_{n+1}$, then*

$$P(\bigcap_n B_n) = \lim_n P(B_n).$$

10. $P(\bigcap_n A_n) \leqslant \inf_n P(A_n), \qquad P(\bigcup_n A_n) \geqslant \sup_n P(A_n).$

11. $P(\liminf_n A_n) \leqslant \liminf_n P(A_n) \leqslant \limsup_n P(A_n) \leqslant P(\limsup_n A_n).$

12. $\lim_{N\to\infty} P(\bigcup_{n=1}^{N} A_n) = P(\bigcup_{n\in\mathbb{N}} A_n), \qquad \lim_{N\to\infty} P(\bigcap_{n=1}^{N} A_n) = P(\bigcap_{n\in\mathbb{N}} A_n)$

Proof: Property 8: In this case B_n is increasing; set

$$C_1 = B_1, \qquad C_n = B_n \setminus B_{n-1} \quad (n \geqslant 2).$$

The C_n are pairwise disjoint, and

$$B_n = \biguplus_{j=1}^{n} C_j \qquad \text{and} \qquad \bigcup_n B_n = \biguplus_{j=1}^{\infty} C_j$$

where $\uplus$ stands for disjoint union. Now due to property (iii) of Definition 3.1

$$P(B_n) = \sum_{j=1}^{n} P(C_j), \qquad P(\bigcup_n B_n) = \sum_{j=1}^{\infty} P(C_j).$$

Hence

$$\lim_n P(B_n) = \lim_n \sum_{j=1}^{n} P(C_j) = \sum_{j=1}^{\infty} P(C_j) = P(\bigcup_n B_n).$$

Property 9 is Exercise 3.11 (*hint: complements turn a decreasing sequence into an increasing one*).

For Property 10, observe that $\bigcap_m A_m \subseteq A_n$ for all n; thus

$$P(\bigcap_m A_m) \leqslant P(A_n) \qquad \forall n \in \mathbb{N}.$$

By taking the infimum of this inequality, the result follows. The same logic applies for the other part.

For Property 11, observe that $B_n := \bigcap_{m \geqslant n} A_m$ is increasing, thus

$$P(\bigcup_{n\in\mathbb{N}} \bigcap_{m\geqslant n} A_m) = P(\bigcup_{n\in\mathbb{N}} B_n) = \lim_n P(B_n)$$

where we used Property 8 for the last equality. Now we have

$$P(\bigcup_{n\in\mathbb{N}} \bigcap_{m\geqslant n} A_m) = \lim_n P(\bigcap_{m\geqslant n} A_m) = \liminf_n P(B_n) \leqslant \liminf_n P(A_n)$$

where in the last equation we used the fact that

$$B_n \subseteq A_n \implies P(B_n) \leqslant P(A_n) \implies \liminf_n P(B_n) \leqslant \liminf_n P(A_n).$$

Collecting all our work for Property 11 so far, we have proven the first inequality. The second inequality is real analysis ($\liminf \leqslant \limsup$ always). The third follows with the same logic as the first – Exercise 3.11.

Property 12 follows by applying Property 8 to the increasing sequence $B_N := \bigcup_{n=1}^{N} A_n$ (resp. Property 9 to the decreasing sequence $\bigcap_{n=1}^{N} A_n$). □

Property 12 can also be seen as a special case of Property 11, in light of the following

Definition 3.9 (Limit of a sequence of sets)**.** *Let $A_n \in \mathcal{E}$. We will say that the limit $\lim_n A_n$ exists if and only if*

$$\limsup_n A_n = \liminf_n A_n,$$

in which case

$$\lim_n A_n := \limsup_n A_n = \liminf_n A_n.$$

Definition 3.5 and Lemmata 3.6, 3.7 show that this definition is consistent with elementary considerations. Property 11 of Lemma 3.8 means that, if $\lim_n A_n$ exists,

$$\lim_n P(A_n) = P(\lim_n A_n).$$

Later on, in a computation related to Brownian motion, we will use the following

Theorem 3.10 (Borel-Cantelli Lemma)**.** *Let P be a probability measure on the σ-algebra $\mathcal{E}$, and consider a sequence of sets $E_n \in \mathcal{E}$. Then*

$$\sum_{n\in\mathbb{N}} P[E_n] < \infty \implies P[\limsup_n E_n] = 0.$$

Proof. By Property 4 of Lemma 3.2 and σ-subadditivity (which follows by combining finite subadditivity – Property 5 of Lemma 3.2 – with continuity from below, Property 8 above), for every n

$$P[\limsup_m E_m] \leqslant P[\bigcup_{m\geqslant n} E_m] \leqslant \sum_{m\geqslant n} P[E_m] \xrightarrow[n\to\infty]{} 0,$$

where the tail of the convergent series vanishes in the limit. □

This is just an example of the intrinsic connections of random processes and measure theory.

Further reading: An excellent, accessible yet fully rigorous introduction to set theory is Moschovakis [6]. Chapters 2–4 of Royden and Fitzpatrick [7] cover basic measure theory in a self-contained way; the timeless classic is Halmos [8].

Exercises

Several of the problems below are referenced at the corresponding points of the text; Exercises 3.8–3.11 supply the proofs deferred from the technical toolkit (Section 3.5).

Exercise 3.1. *Prove Properties 1–6 of Lemma 3.2 from Kolmogorov's axioms, using in each proof the properties already established.*

Hint: $A \uplus A^c = \Omega$; $B = A \uplus (B \setminus A)$; $A \cup B = A \uplus (B \setminus A)$.

Exercise 3.2 (A taste of set theory)**.** *Recall / find the definition of a countable set, and of an uncountable set.*

Prove that $\mathbb{Q}$ is countable. Prove that $[0,1] \subseteq \mathbb{R}$ is uncountable.

Exercise 3.3 (Equivalent Definitions of σ-algebra)**.** *One of several equivalent definitions of a σ-algebra $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a family of sets with the properties*

(i). $\emptyset \in \mathcal{A}$,

(ii). $A \in \mathcal{A} \quad \Rightarrow \quad A^c \in \mathcal{A}$,

(iii). $A_n \in \mathcal{A} \quad \forall n \quad \Rightarrow \quad \bigcup_n A_n \in \mathcal{A}$.

Prove that if $\mathcal{A}' \subseteq \mathcal{P}(\Omega)$ has the properties

(i'). $\Omega \in \mathcal{A}'$,

(ii'). $A \in \mathcal{A}' \quad \Rightarrow \quad A^c \in \mathcal{A}'$,

(iii'). $A_n \in \mathcal{A}' \quad \forall n \quad \Rightarrow \quad \bigcap_n A_n \in \mathcal{A}'$,

then $\mathcal{A}'$ is a σ-algebra.

Hint: use the identity $(A \cup B)^c = A^c \cap B^c$

Exercise 3.4. *Prove that without loss of generality property 3 of Definition 3.3 can be rewritten as $A, B \in \underline{\mathcal{E}} \quad \Rightarrow \quad \exists N \in \mathbb{N}$, disjoint sets $\{C_n\}_{n=1}^N \subseteq \underline{\mathcal{E}}$ such that $A \setminus B = \biguplus_{n=1}^{N} C_n$*

Exercise 3.5. *Prove that the set of all intervals $(a, b]$ with* extended endpoints $-\infty \leqslant a \leqslant b \leqslant +\infty$ *(conventions: $(a, +\infty] := (a, +\infty)$, $(a, a] = \emptyset$ – so $\emptyset$ and $\mathbb{R}$ are included) form a semi-algebra of $\mathbb{R}$.* Watch the endpoints: with finite endpoints only, the complement of $(a, b]$ contains the ray $(b, +\infty)$, which is not a finite union of bounded intervals – that family would not be a semi-algebra. *Similarly that all rectangles $(a, b] \times (c, d]$ form a semi-algebra of $\mathbb{R}^2$.*

Exercise 3.6 (Cumulative distribution function, I of II)**.** *Prove that the intervals $(a, b] \subseteq (0, 1]$ form a semi-algebra.*

Prove that Lebesgue measure

$$P((a, b]) = b - a$$

satisfies the hypotheses of Theorem 3.4, and hence extends to a probability measure on the σ-algebra generated by $\{(a, b] \subseteq (0, 1]\}$.

Prove that this measure assigns zero mass to every singleton, $P[\{x\}] = 0$ (use continuity of measure), and deduce, by σ-additivity, that every countable subset of $(0, 1]$ has measure zero. Note the order of the two steps: σ-additivity alone does not make countable sets null – an atomic measure violates no axiom.

Recall / find a proof that non-measurable subsets of $[0, 1]$ exist.

Exercise 3.7 (Cumulative distribution function, II of II)**.** *Prove that the σ-algebra generated by $\{(a, b] \subseteq \mathbb{R}\}$ is the Borel σ-algebra.*

Let $F : \mathbb{R} \to [0, 1]$ be a non-decreasing, right-continuous function with $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$.

Prove that the set function

$$P((a, b]) = F(b) - F(a)$$

is σ-additive on the intervals (taken again with extended endpoints, setting $F(-\infty) := 0$, $F(+\infty) := 1$), and conclude by Theorem 3.4 that it extends to a probability measure on the Borel σ-algebra.

Hint: the σ-additivity is the one genuinely nontrivial step – real measure-theoretic work, not a routine check. Reduce it to the following: if $(a, b] = \biguplus_n (a_n, b_n]$, then $F(b) - F(a) = \sum_n \big(F(b_n) - F(a_n)\big)$. For the hard inequality, use right-continuity of F to enlarge each $(a_n, b_n]$ to a slightly longer open interval at $\varepsilon 2^{-n}$ cost, and compactness (Heine-Borel) of $[a', b]$ for $a' > a$ to reduce to a finite cover.

Give a sufficient condition so that $P(\{x\}) = 0$ for all $x \in \mathbb{R}$.

Give an example of a cumulative distribution function F for which there exists an $x_* \in \mathbb{R}$ such that $P(\{x_*\}) > 0$.

Exercise 3.8. *Let a_n be a sequence of real numbers. Prove that the limits defining* $\liminf_n a_n$ *and* $\limsup_n a_n$ *always exist in* $\mathbb{R} \cup \{\pm\infty\}$*, and that* $\lim_n a_n$ *exists if and only if* $\liminf_n a_n = \limsup_n a_n$ *(in which case all three coincide).*

Hint: $\inf_{k \geqslant n} a_k$ is non-decreasing in n, and monotone sequences always have limits in $\mathbb{R} \cup \{\pm\infty\}$.

Exercise 3.9. *Prove Lemma 3.6: the characterisations of* $\limsup_n A_n$ *("ω belongs to infinitely many of the A_n") and of* $\liminf_n A_n$ *("ω belongs to all but finitely many"), and the values of both for monotone sequences of sets.*

Hint: unpack the definitions – $\omega \in \bigcap_n \bigcup_{m \geqslant n} A_m$ says: for every n there exists $m \geqslant n$ with $\omega \in A_m$. For monotone sequences, the inner union or intersection stabilises.

Exercise 3.10. *Prove the membership claim of Lemma 3.7: if $A_n \in \mathcal{A}$ for all n, then* $\limsup_n A_n$, $\liminf_n A_n \in \mathcal{A}$.

Hint: a σ-algebra is closed under countable unions by definition; combine with complements (De Morgan) to get countable intersections.

Exercise 3.11. *Complete the proof of Lemma 3.8: prove Property 9, and the third inequality of Property 11.*

Hint: for Property 9, pass to complements – B_n^c is increasing – and use Properties 1 and 8. For the third inequality of Property 11, mirror the proof of the first one, using the decreasing sets $\bigcup_{m \geqslant n} A_m$ and Property 9.

Exercise 3.12. *Prove the third order inclusion-exclusion formula,*

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

Can you prove the general case (Property 7 of Lemma 3.2) by induction?

Exercise 3.13. *For a sequence of events $A_n \in \mathcal{E}$, we know that*

$$P(\liminf_n A_n) \leqslant \liminf_n P(A_n) \leqslant \limsup_n P(A_n) \leqslant P(\limsup_n A_n).$$

Present a completely self-contained proof in your own words. Feel free to use Section 3.5 to help you.

Exercise 3.14. *Prove that, for any finite set A, $|\mathcal{P}(A)| = 2^{|A|}$.*

(Hint: one way is to show that $\mathcal{P}(A)$ is in bijection with the set of all functions $\phi : A \to \{0, 1\}$, and count the elements of that set. Another way is by induction.)

4 Conditional probability, revisited

Within the foundations of the correspondence between a random experiment and a probability space, is the concept of conditional probability: if I know that event B took place, what is the probability that event A took place as well?

In Chapter 2 we introduced the conditional probability

$$P(A|B) = \frac{P(AB)}{P(B)}$$

on finite sample spaces, where it could be motivated – and verified – by direct counting (cf. Example 2.7). We now recreate the concept within the general framework of Chapter 3. The payoff is that conditioning becomes available on arbitrary probability spaces, and is revealed to be a genuine probability measure in its own right.

A way to derive this formula is by going back to our probability space: if I know that B took place, this means that – as long as I am in the conditional setting – my new sample space is really $\Omega_B = B$, and now the conditional events are $E_B := E \cap B = EB$, i.e. the σ-algebra should become $\mathcal{E}_B = \{EB | E \in \mathcal{E}\}$. So now I have to modify P into some P_B so that $(\Omega_B, \mathcal{E}_B, P_B)$ makes sense as a probability space.

Definition 4.1 (Conditional probability space)**.** *Given a probability space $(\Omega, \mathcal{E}, P)$ and an event $B \in \mathcal{E}$ with $P[B] > 0$, the conditional probability space $(\Omega_B, \mathcal{E}_B, P_B)$ is defined as*

- $\Omega_B = B$
- $\mathcal{E}_B = \{EB \,|\, E \in \mathcal{E}\}$
- $P_B(A_B) = \frac{P(A_B)}{P(B)}$, *i.e. for any $A \in \mathcal{E}$ we set* $\quad P(A|B) := P_B(AB) = \frac{P(AB)}{P(B)}$.

One can check that this has all the required properties to be a probability space in the sense of the previous section.

The conditional probability formula

$$P(A|B) = \frac{P(AB)}{P(B)} \qquad \Rightarrow \qquad P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

has many interesting consequences.

Lemma 4.2 (Extended conditional probability formula)**.** *Let $A_n \in \mathcal{E}$ with $P\left(\bigcap_{n=1}^{N-1} A_n\right) > 0$. Then*

$$P\left(\bigcap_{n=1}^{N} A_n\right) = P(A_1)\ P(A_2|A_1)\ P(A_3|A_2A_1)\ \ldots\ P(A_N|A_{N-1}\ldots A_2A_1).$$

Proof. Expanding each factor on the right-hand side by the definition of conditional probability, the product telescopes:

$$P(A_1)\ \frac{P(A_1A_2)}{P(A_1)}\ \frac{P(A_1A_2A_3)}{P(A_1A_2)}\ \cdots\ \frac{P(A_1A_2\cdots A_N)}{P(A_1A_2\cdots A_{N-1})} = P(A_1A_2\cdots A_N).$$

All the denominators are nonzero, since by monotonicity $P(A_1 \cdots A_{N-1}) > 0$ implies $P(A_1 \cdots A_n) > 0$ for all $n \leqslant N-1$. □

We already saw this formula at work in Chapter 2, in the quality control computation of Example 2.8.

Lemma 4.3 (Total probability lemma)**.** *Let $D_n \in \mathcal{E}$ be a partition, i.e.*

$$m \neq n \quad \Rightarrow \quad D_n D_m = \emptyset, \qquad \bigcup_n D_n = \Omega.$$

Then

$$P(A) = \sum_n P(A|D_n)P(D_n).$$

Therefore,

$$P(D_n|B) = \frac{P(D_n)P(B|D_n)}{P(B)} = \frac{P(D_n)P(B|D_n)}{\sum\limits_m P(D_m)P(B|D_m)}.$$

Proof. Since the D_n are disjoint and cover Ω, we have $A = A \cap \Omega = \biguplus\limits_n AD_n$, so by σ-additivity and the conditional probability formula

$$P(A) = \sum_n P(AD_n) = \sum_n P(A|D_n)P(D_n),$$

with the convention that terms with $P(D_n) = 0$ are omitted. For the second claim, observe that $P(D_n|B) = \frac{P(D_n B)}{P(B)} = \frac{P(D_n)P(B|D_n)}{P(B)}$, and expand the denominator $P(B)$ using the first claim. □

The latter is also called Bayes' formula, and it allows one to "invert" the question, i.e. use knowledge of $P(B|D_m)$ to talk about $P(D_n|B)$, which can be very useful in real-world problems.

Corollary 4.4 (Bayes formula for two events)**.** *A simpler version of Lemma 4.3 when the partition $\{D_n\}$ consists just of two sets, $\{A, A^c\}$, is*

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

Considerations on conditional probability naturally raise the question of stochastic independence: what if the knowledge that B happened doesn't affect whether A happened? This is the case if

$$P(A|B) = P(A), \qquad P(B|A) = P(B),$$

and the conditional probability formula quickly leads to the following

Definition 4.5 (Stochastic independence)**.** *The events $A, B \in \mathcal{E}$ are called stochastically independent iff*

$$P(AB) = P(A)P(B).$$

The events $A_n \in \mathcal{E}$ are called pairwise stochastically independent if, for any $n \neq m$

$$P(A_n A_m) = P(A_n)P(A_m).$$

The events $A_n \in \mathcal{E}$ are called collectively stochastically independent if, for any finite collection of different indices $i_1, i_2, \ldots i_N$, we have

$$P(A_{i_1} A_{i_2} \ldots A_{i_N}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_N}).$$

Practice with these definitions – including the classic construction of events that are pairwise but not collectively independent – is Exercise 4.1.

Sometimes stochastic independence is stipulated (for non-mathematical reasons), and then used to recover the probability measure P from limited information.

Example 4.6. *Find the probability that a machine operating at time t_0 will not break down before time $t_0 + t$, under the assumptions that:*

1. *this probability depends only on t (and in particular it does not depend on how long the machine has been operating before t_0)*
2. *for short times Δt, the probability that the machine breaks down is proportional to Δt up to $o(\Delta t)$.*

Solution: Let $p(t)$ be the probability that machine which is running at the time of observation t_0 has not broken down by $t_0 + t$ (i.e. after time t has passed). Define the event

$$A_t := \{ \text{ the machine still runs after time } t \text{ has passed } \}.$$

By construction, $p(t) := P[A_t]$.

Now, Assumption 2 means that

$$1 - p(\Delta t) = a\Delta t + o(\Delta t)$$

for some constant a.

Moreover observe that

$$A_{t+\Delta t} = \{ \text{ the machine still runs after time } t + \Delta t \text{ has passed} \} \subseteq A_t.$$

So the conditional probability formula yields

$$P(A_{t+\Delta t}|A_t) = \frac{P(A_{t+\Delta t} \cap A_t)}{P(A_t)} = \frac{P(A_{t+\Delta t})}{P(A_t)}$$

Moreover, Assumption 1 means that

$$P(A_{t+\Delta t}|A_t) = P(A_{\Delta t}).$$

By combining the last two equations we get

$$p(t+\Delta t) = p(\Delta t)p(t)$$

hence

$$\frac{p(t+\Delta t)-p(t)}{\Delta t} = p(t)\frac{p(\Delta t)-1}{\Delta t} = -ap(t)+o(1),$$

and we are led to the differential equation

$$\frac{dp(t)}{dt} = -ap(t) \qquad \Rightarrow \qquad p(t) = Ce^{-at}.$$

The constant $C = 1$ is chosen so that $p(0) = 1$, since the machine is running at the moment of observation.

Observe that, if by X we denote the **failure time***, the time it takes for the machine to break down, then*

$$p(t) = P[\{\textit{the machine still runs after time } t \textit{ has passed}\}] =$$

$$= P[X > t] = 1 - P[X \leqslant t] = 1 - P[X \in (0,t]].$$

So this is an example of computing a **cumulative distribution function** *for X, thereby completely determining the probability measure for X.*

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 4.1 (Stochastic independence)**.** *1. Prove that if two events A, B are stochastically independent, then each of the following sets of events are stochastically independent: $\{A, B^c\}$; $\{A^c, B\}$; $\{A^c, B^c\}$.*

2. Create an example of pairwise stochastically independent events which are not collectively stochastically independent.

Hint: consider three fair coin tosses, $A = \{$tosses 1,2 agree$\}$, $B = \{$tosses 2,3 agree$\}$, $C = \{$tosses 1,3 agree$\}$.

3. Denote $\widetilde{\mathcal{E}} := \{A, B, C, A^c, B^c, C^c\}$. Prove that if

$$P(E_1E_2E_3) = P(E_1)P(E_2)P(E_3) \quad \textit{for every three different } E_j \in \widetilde{\mathcal{E}},$$

then A, B, C are also pairwise stochastically independent.

4. Create an example of events A, B, C which satisfy

$$P(ABC) = P(A)P(B)P(C)$$

but are not *pairwise stochastically independent.*

Hint: take $\Omega = [0,1]$, P *to be the Lebesgue measure and* $A = (0, \frac{1}{3})$, $B = (\frac{1}{9}, \frac{12}{27})$, $C = (\frac{8}{27}, \frac{17}{27})$.

Exercise 4.2 (Simpson's paradox)**.** *A university has two departments,* α *and* β*. In one admissions round the outcomes were as follows:*

	men		*women*	
	applied	*admitted*	*applied*	*admitted*
Dept. α	*100*	*80*	*10*	*9*
Dept. β	*10*	*2*	*90*	*27*

Consider the experiment of picking an applicant at random, and the events A *(admitted),* M*,* W *(the applicant is a man / woman),* D_α*,* D_β *(the applicant applied to department* α */* β*).*

1. *Show that* $P(A|WD_j) > P(A|MD_j)$ *for both* $j \in \{\alpha, \beta\}$*, but nevertheless* $P(A|W) < P(A|M)$*.*
2. *Explain how this is possible, by writing* $P(A|W)$ *and* $P(A|M)$ *through the total probability formula over the partition* $\{D_\alpha, D_\beta\}$ *and comparing the weights* $P(D_j|W)$*,* $P(D_j|M)$*.*

This reversal is known as **Simpson's paradox***; it is a recurring trap in the analysis of real datasets.*

Exercise 4.3 (The red aces)**.** *Consider a well-shuffled deck of cards face down. We flip cards one at a time; if a card is not a red ace we put it face up in a new pile. Find the probability that there are exactly* n *cards in the face-up pile when we find the first red ace.*

(Hint: define the events $A^k = \{\text{the } k^{th} \text{ card is not a red ace}\}$, express the event $B_n = \{\text{exactly } n \text{ cards in the pile at the first red ace}\}$ in terms of the A^k, and use the extended conditional probability formula, Lemma 4.2.)

Exercise 4.4 (Polya's urn)**.** *An urn contains* b *black and* w *white balls. A ball is drawn at random, and returned to the urn together with* c *additional balls of the same colour; this procedure is then repeated indefinitely.*

1. *Show that* $P[\text{second draw is black}] = \frac{b}{b+w}$*.*
2. *(*) Show that, for every* n*,* $P[n^{th} \text{ draw is black}] = \frac{b}{b+w}$*.*

   (Hint: total probability and induction.)

Polya's urn is a simple model of a "rich get richer" mechanism; note the striking contrast between the strong dependence between draws and the invariance of the unconditional probabilities.

Exercise 4.5 (Two children)**.** *A family has two children; assume the four birth sequences (BB, BG, GB, GG) are equally likely.*

1. *Given that at least one child is a boy, what is the probability that both are boys?*
2. *Given that the* elder *child is a boy, what is the probability that both are boys?*

The two answers differ; explain carefully what the conditioning event is in each case.

5 Introduction to Random Variables

5.1 Generalities

Definition 5.1 (Measurable space)**.** *Consider a set $\mathbb{A}$ and a σ-algebra $\mathcal{A} \subseteq \mathcal{P}(\mathbb{A})$ of subsets of $\mathbb{A}$.*

The pair $(\mathbb{A}, \mathcal{A})$ is called a measurable space.

Definition 5.2 (Measurable mapping)**.** *Let $(\mathbb{A}, \mathcal{A})$, $(\mathbb{B}, \mathcal{B})$ be two measurable spaces. A mapping*

$$f : \mathbb{A} \to \mathbb{B}$$

is called measurable if

$$\forall\ B \in \mathcal{B} \quad f^{-1}(B) \in \mathcal{A}.$$

Recall that the inverse image under a mapping $f : \mathbb{A} \to \mathbb{B}$ is well-defined for any $B \subseteq \mathbb{B}$,

$$f^{-1}(B) = \{a \in \mathbb{A} | f(a) \in B\},$$

even if f is not one-to-one. Also, the inverse image of a set can be $\emptyset$, or all of $\mathbb{A}$.

Definition 5.3 (Random variable and induced probability space)**.** *Let $(\Omega, \mathcal{E}, P)$ be a probability space, and $(\mathbb{B}, \mathcal{B})$ a measurable space. A measurable mapping*

$$X : \Omega \to \mathbb{B} : \omega \mapsto X(\omega), \qquad X^{-1}(B) \in \mathcal{E}\ \forall B \in \mathcal{B}$$

is called a random variable. Moreover, denoting

$$P_X(B) := P(X^{-1}(B)) \quad \forall B \in \mathcal{B},$$

the triplet $(\mathbb{B}, \mathcal{B}, P_X)$ is called the induced probability space.

Moreover, if $\mathbb{B}$ is countable – equipped with the σ-algebra $\mathcal{B} = 2^{\mathbb{B}}$, so that in particular every $\{X = x\}$ is an event – X is called a discrete random variable.

Why do we go to all this trouble to introduce a random variable? The following examples give some concrete ideas:

1. In example 4.6, the events $A_t \in \mathcal{E}$ are of the form

   $$A_t := \{\text{ the machine runs after time } t \text{ has passed }\}.$$

   We saw that being able to have this descriptive explanation of the events helped figure out the probability measure P. On the other hand, when dealing with the random variable X measuring the time until the machine breaks down, we have corresponding *induced* events $B_t \in \mathcal{B}$ of the form

   $$B_t = (t, +\infty).$$

It is clear that B_t *corresponds* to A_t (namely $A_t = X^{-1}(B_t)$), but it *is not* A_t. The mapping X is exactly this *correspondence*.

2. This formalism gives us a very clear way to distinguish between realisations and random variables. X is a random variable, $X(\omega)$ is a realisation. Going back e.g. to example 4.6, observe that we don't have a canonical, compact, quantitative way to refer to "individual realisations of the time until the machine breaks down", versus "the random variable of the time until the machine breaks down".

3. At a first glance, everything we need to know about a random variable is contained in the induced probability space $(\mathbb{B}, \mathcal{B}, P_X)$. However, the formalism including the mapping $X : \Omega \to \mathbb{B}$ is essential in doing operations between several different random variables. We will see this in some detail later on.

(That preimages respect disjointness – unlike forward images – and that P_X is indeed a probability measure is Exercise 5.1.)

5.2 Discrete random variables

In this subsection we will assume for simplicity that $\mathbb{B}$ is countable. In this case any random variable (R.V.) on $(\Omega, \mathcal{E}, P)$ is called a discrete R.V.

Definition 5.4 (Probability density function (pdf) for a discrete random variable)**.** *Given a probability space $(\Omega, \mathcal{E}, P)$ and a discrete random variable*

$$X : \Omega \to \mathbb{B} \qquad X^{-1}(B) \in \mathcal{E} \ \forall B \in \mathcal{B},$$

the induced probability measure $P_X : \mathcal{B} \to [0, 1]$ can be represented through the pdf

$$f_X(b) := P\big(X^{-1}(\{b\})\big) = P_X\big(\{b\}\big).$$

Then

$$P_X(B) = \sum_{b \in B} f_X(b).$$

Now we can easily construct more complicated variables out of simpler ones.

Example 5.5. *Let X, Y be two independent Bernoulli trials, i.e. they have identical sample spaces and probability measures*

$$\mathbb{B}_X = \mathbb{B}_Y = \{0, 1\}, \qquad f_X(1) = f_Y(1) = p,$$

and their joint pdf factorises, $f_{(X,Y)}(x, y) = f_X(x) f_Y(y)$ (this is the definition of independence for discrete random variables; more on independence later). Now consider the fraction of successes in two trials

$$Z := \frac{X + Y}{2}, \ \textit{i.e.} \ Z : \Omega \to \mathbb{R} : \omega \mapsto \frac{X(\omega) + Y(\omega)}{2}.$$

This has sample space

$$\mathbb{B}_Z = \{0, \frac{1}{2}, 1\}$$

and pdf

$$f_Z(0) = (1-p)^2, \qquad f_Z(\frac{1}{2}) = 2p(1-p), \qquad f_Z(1) = p^2.$$

Example 5.6. *Recall the problem of Example 2.7, and denote X the number of aces in the first draw and Y the number of aces in the second draw. As was computed in Example 2.7, each of these random variables separately works out as*

$$\begin{aligned} \mathbb{B}_X &= \{0,1\}, \qquad f_X(1) = \tfrac{1}{13}, \\ \mathbb{B}_Y &= \{0,1\}, \qquad f_Y(1) = \tfrac{1}{13}. \end{aligned}$$

However, X and Y are no longer independent; that is the joint probability density $f_{(X,Y)}(x,y)$ is not given merely by the products of the individual pdfs, but instead

$$f_{(X,Y)}(1,1) = \frac{12}{52\cdot 51}, \quad f_{(X,Y)}(0,1) = \frac{48\cdot 4}{52\cdot 51}, \quad f_{(X,Y)}(1,0) = \frac{4\cdot 48}{52\cdot 51}, \quad f_{(X,Y)}(0,0) = \frac{48\cdot 47}{52\cdot 51}.$$

So now if we consider a new random variable $Z = \frac{X+Y}{2}$ (the average number of aces per draw in two draws) we will have

$$\mathbb{B}_Z = \{0, \frac{1}{2}, 1\}, \qquad Z : \Omega \to \mathbb{R} : \omega \mapsto \frac{X(\omega) + Y(\omega)}{2}$$

and pdf

$$f_Z(0) = f_{(X,Y)}(0,0), \qquad f_Z(1) = f_{(X,Y)}(1,1), \qquad f_Z(\frac{1}{2}) = \sum_{x+y=1} f_{(X,Y)}(x,y).$$

Definition 5.7 (Mathematical expectation)**.** *Let $\mathbb{B} \subseteq \mathbb{R}$ and denote*

$$\mathrm{E}[X] := \sum_{b\in\mathbb{B}} b f_X(b) = \int_\Omega X(\omega) dP(\omega)$$

the mathematical expectation of X whenever the series is absolutely convergent.

The mean value has the following properties:

Lemma 5.8 (Properties of the mathematical expectation)**.** *1. If $\mathrm{E}[X]$ and $\mathrm{E}[Y]$ are finite then, for any $a, b \in \mathbb{R}$,*

$$\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y].$$

2. If $A \leqslant X(\omega) \leqslant B$ for all $\omega \in \Omega$, then

$$A \leqslant \mathrm{E}[X] \leqslant B.$$

3. If a random variable Y is defined through a transformation on the random variable X,

$Y = g(X)$, *then*

$$\mathrm{E}[Y] = \sum_{b \in \mathbb{B}} g(b) f_X(b).$$

Lemma 5.9 (Equivalent definitions of variance)**.** *The following two definitions of the variance are equivalent:*

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

The proofs of Lemmata 5.8 and 5.9, together with the verification that Definition 5.7 is well posed, are Exercise 5.2.

A discrete random variable of singular importance is the following:

Example 5.10 (The Poisson random variable)**.** *Let $\lambda > 0$. The random variable $X : \Omega \to \mathbb{N} \cup \{0\}$ with pdf*

$$f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

is called the **Poisson random variable with parameter $\boldsymbol{\lambda}$**. *The pdf is correctly normalised, since*

$$\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1,$$

and its first moments work out as

$$\mathrm{E}[X] = \sum_{k=1}^{\infty} k\, e^{-\lambda} \frac{\lambda^k}{k!} = \lambda \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda,$$

$$\mathrm{E}[X(X-1)] = \sum_{k=2}^{\infty} k(k-1) e^{-\lambda} \tfrac{\lambda^k}{k!} = \lambda^2 \sum_{k=2}^{\infty} e^{-\lambda} \tfrac{\lambda^{k-2}}{(k-2)!} = \lambda^2$$

$$\Longrightarrow \qquad \mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \lambda.$$

The Poisson random variable is the universal model for **counts of rare events***: misprints on a page, radioactive decays in a second, phone calls arriving in a minute (cf. the examples of Chapter 1).* Why *it plays this role is explained by the law of rare events in Section 6.3; and in Chapter 11 it will be upgraded from a static count to a full random process, the Poisson process.*

The variance and expectation are **moments** that can be used through Chebychev-type inequalities.

5.3 Chebychev Inequalities

Lemma 5.11 (Chebychev Inequality I, also called Markov's Inequality)**.** *If $X(\omega) \geqslant 0$, then, for any $t > 0$,*

$$P[X \geqslant t] \leqslant \frac{\mathrm{E}[X]}{t}$$

Proof:

$$P[\{\omega \in \Omega : X(\omega) \geqslant t\}] = \int\limits_{\{\omega : X(\omega) \geqslant t\}} dP(\omega) \leqslant \int\limits_{\{\omega : X(\omega) \geqslant t\}} \frac{X(\omega)}{t} dP(\omega) \leqslant \frac{\mathrm{E}[X]}{t},$$

where we used the fact that $X(\omega) \geqslant 0$ in the last step.

Lemma 5.12 (Chebychev inequality II)**.** *Let* $\mathrm{E}[X]$, $\mathrm{Var}[X]$ *exist. Then, for any* $t > 0$,

$$P[|X - \mathrm{E}[X]| \geqslant t] \leqslant \frac{\mathrm{Var}[X]}{t^2}$$

Proof: Apply Lemma 5.11 to the non-negative R.V. $Y = (X - \mathrm{E}[X])^2$.

Remark 5.13. *Observe that the proof does not use the fact that* $\mathbb{B}$ *is countable, i.e. the same proofs will also apply to the continuous case (uncountable* $\mathbb{B}$*).*

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 5.1. *Prove that* $A \cap B = \emptyset \implies X^{-1}(A) \cap X^{-1}(B) = \emptyset$*. Show that this is not true for the* forward image,

$$X[A] := \{t | \exists a \in A \text{ such that } t = X(a)\}.$$

Prove that the induced probability measure P_X*, as defined in Definition 5.3, satisfies the axioms of probability, cf. Definition 3.1.*

Exercise 5.2. *Prove that Definition 5.7 makes sense, i.e. that*

$$\sum_{b \in \mathbb{B}} b f_X(b) = \int_\Omega X(\omega) dP(\omega)$$

(Hint: Start with a finite $\mathbb{B}$. Then the integral becomes one of a simple function.)

Moreover, prove Lemmata 5.8, 5.9.

Exercise 5.3. *Consider independent random variables* X, Y *with* $P[X = 1] = P[X = -1] = \frac{1}{2}$, $P[Y = 1] = P[Y = -1] = \frac{1}{2}$*. Show that* $Z := XY$ *is independent of each of* X *and* Y*, but not of* $X + Y$*.*

(Hint: show that $f_{ZX}(z, x) = f_Z(z) f_X(x)$ and similarly for Z, Y; then show the counterpart for $Z, X + Y$ fails.)

Exercise 5.4 (Discrete convolution)**.** *Let* $X, Y : \Omega \to \mathbb{N} \cup \{0\}$ *be independent. Show that* $Z := X + Y$ *has pdf*

$$f_Z(n) = \sum_{j=0}^{n} f_X(j) f_Y(n - j), \qquad n = 0, 1, 2, \ldots$$

(Hint: $\{Z = n\} = \biguplus_j \{X = j \text{ AND } Y = n - j\}$.)

Exercise 5.5. *For the red aces problem (Exercise 4.3), compute the average number of cards that come before the first red ace.*

Exercise 5.6 (St Petersburg paradox)**.** *A player pays an entry fee of x dollars and a coin is tossed repeatedly until it first comes up heads. If this happens at toss number k, the prize is 2^k dollars. To decide on a reasonable entry fee, compute the average prize. What do you conclude?*

Exercise 5.7. *If X, Y are independent random variables with finite means and variances, prove that*

$$\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y], \qquad \mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Exercise 5.8 (Pascal distribution)**.** *(*) For the Pascal random variable of Section 2.5 (number of trials until the r^{th} success), show that*

$$\mathrm{E}[X] = \frac{r}{p}, \qquad \mathrm{Var}[X] = r\frac{1-p}{p^2}.$$

(Hint: write X as the sum of r i.i.d. geometric waiting times.)

6 Sequences of independent trials

6.1 Bernoulli Law of Large Numbers and Ramifications

A Bernoulli trial is a random experiment with two possible outcomes, "success" and "failure",

$$\Omega = \{ \underbrace{0}_{failure}, \underbrace{1}_{success} \}.$$

The probability of "success" is p and of "failure" is $q := 1 - p$. The importance of Bernoulli trials is that even very complicated experiments can be broken down to Bernoulli trials by selecting an event $\mathcal{E} \ni A \subseteq \Omega$ as "success" and A^c as "failure". We saw a very simple instance of that in example 1.4. More interesting examples can include things like

- "success" if the number of phone calls in the UK on a business day exceeds 50,000,000
- "success" if the daily production of an oil rig exceeds 130% of its nominal average daily value
- "success" if a specific machine keeps running without incident for 1000 hours since the last maintenance

etc.

Two distinct modelling assumptions are being made here, and it pays to separate them: (*i*) **identical distribution** – every trial has the *same* probability of success p; and (*ii*) **independence** – the outcome of each trial carries no information about the others. Neither implies the other: flip one coin once and copy the result forever ($X_n = X_1 \ \forall n$) – every trial has success probability $\frac{1}{2}$, yet the sequence is as dependent as possible. Bernoulli trials assume *both* – which can be a strong assumption in the more complicated examples above – and are therefore often called *independent* trials.

So we consider a RV $X : \Omega \to \{0, 1\}$ with probability density function (pdf) f_X as follows

$$f_X(1) := p, \qquad f_X(0) = 1 - p.$$

Moreover, out of this we create a sequence of independent and identically distributed random variables $\{X_n\}$. Independence here means that for any $N \in \mathbb{N}$ and any $A_1, \ldots, A_N \in \mathcal{B}$

$$P[X_1 \in A_1 \text{ and } X_2 \in A_2 \text{ and } \ldots \text{ and } X_N \in A_N] = \prod_{n=1}^{N} P[X_n \in A_n].$$

Now denote by $B_{N,p}$ the number of "successes" in a total of N independent trials,

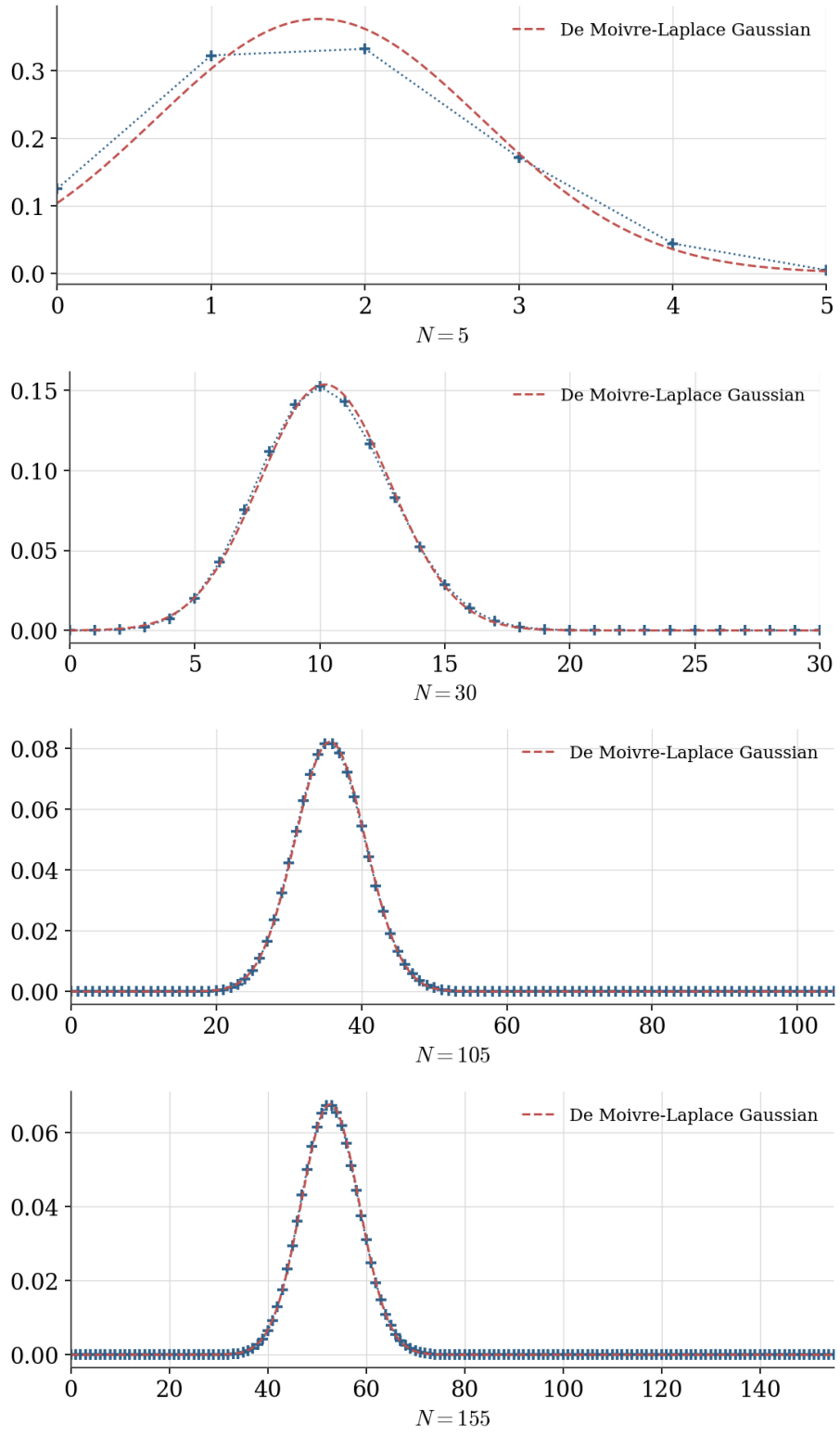
$$B_{N,p} = \sum_{n=1}^{N} X_n.$$

Clearly, the possible values for $B_{N,p}$ are

$$B_{N,p}(\omega) \in \{0, 1, 2, \ldots N\}.$$

$B_{N,p}$ is also called binomial because of the form of its pdf.
The derivation of the pdf, and of the moments $\mathrm{E}[B_{N,p}] = Np$, $\mathrm{Var}[B_{N,p}] = Np(1-p)$ – used repeatedly below – is Exercise 6.1.

Plots of the pdf $f_{B_{N,p}}(x)$ for $p = 0.34$ and various N.



Theorem 6.1 (Bernoulli Law of large numbers)**.** *Let X_n be iid Bernoulli trials with parameter*

p, as above. Then, for any $t > 0$,

$$P\left[\left|\frac{\sum_{n=1}^{N} X_n}{N} - p\right| > t\right] \leqslant \frac{p(1-p)}{Nt^2}.$$

Proof: Denote

$$\eta := \frac{\sum_{n=1}^{N} X_n}{N} - p.$$

Then $\mathrm{E}[\eta] = 0$ and

$$\mathrm{Var}[\eta] = \mathrm{E}[\eta^2] - \cancelto{0}{\mathrm{E}^2[\eta]} = \mathrm{E}[\left(\frac{\sum_{n=1}^{N} X_n}{N} - p\right)^2] = \mathrm{E}[\left(\frac{B_{N,p}}{N} - p\right)^2]$$
$$= \tfrac{1}{N^2}\,\mathrm{E}[(B_{N,p})^2] - \tfrac{2p}{N}\,\mathrm{E}[B_{N,p}] + p^2 = \tfrac{p+(N-1)p^2}{N} - 2p^2 + p^2 = \tfrac{p(1-p)}{N}.$$

The result now follows by applying Chebychev II (Lemma 5.12) to the RV η. □

The significance of the Bernoulli Law of Large Numbers is that it quantifies precisely the rate with which averaged quantities from many independent realisations converge to their expected values. That is, it guarantees that our random variables do exhibit the statistical reproducibility that we were trying to describe in the first place. For example, we have a consistency criterion for our stochastic modelling: if we observe a large number of independent trials, our findings must necessarily start converging as described above – *otherwise it means that we are not really looking at independent trials.* There are many results bearing the name "Law of Large Numbers" (LLN), and all of them, broadly speaking, quantify statistical reproducibility in various contexts.

Theorem 6.2 (Bernoulli LLN for finitely many outcomes)**.** *Let* $X : \Omega \to \{0, 1, 2, \ldots r\}$ *with* $P[X = j] = p_j$*. Consider the sequence of iid RV* X_n*,* $P[X_n = j] = p_j$ *and denote*

$$B_N^j = |\{n = 1, 2, \ldots, N | X_n = j\}|$$

the number of times the outcome j *appears in* N *independent trials. Then*

$$P\left[\left|\frac{B_N^j}{N} - p_j\right| > t\right] \leqslant \frac{1}{4Nt^2}.$$

(The proof is a direct adaptation of that of Theorem 6.1 – Exercise 6.2.)

As an aside, the Bernoulli Law of Large Numbers yields a classical constructive proof of the Weierstrass approximation theorem:

Theorem 6.3 (Weierstrass Theorem)**.** *Let* $f \in C[0, 1]$*. Then* $\forall \varepsilon > 0\ \exists\ p_N(x)$ *polynomial of degree at most* N *such that*

$$\max_{0 \leq x \leq 1} |p_N(x) - f(x)| \leq \varepsilon.$$

Proof: To prove this theorem we will use the Bernstein polynomial defined as

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

Then for $x \in [0,1]$,

$$\begin{aligned}
|b_N(x) - f(x)| &= \left| \sum_{k=0}^{N} \binom{N}{k} x^k (1-x)^{N-k} \left(f\left(\frac{k}{N}\right) - f(x) \right) \right| \\
&\leq \sum_{k:\left|\frac{k}{N}-x\right|<\delta} \binom{N}{k} x^k (1-x)^{N-k} \left| f\left(\frac{k}{N}\right) - f(x) \right| \\
&\quad + \sum_{k:\left|\frac{k}{N}-x\right|\geqslant\delta} \binom{N}{k} x^k (1-x)^{N-k} \left| f\left(\frac{k}{N}\right) - f(x) \right| \\
&:= I_1 + I_2
\end{aligned}$$

To complete the proof, it is enough to show that, for every $\varepsilon > 0$, we can make each of the sums I_1 and I_2 less or equal than $\dfrac{\varepsilon}{2}$. We start from I_1:

$$\begin{aligned}
I_1 &= \sum_{k:\left|\frac{k}{N}-x\right|<\delta} \binom{N}{k} x^k (1-x)^{N-k} \left| f\left(\frac{k}{N}\right) - f(x) \right| \\
&\leq \sum_{k:\left|\frac{k}{N}-x\right|<\delta} \binom{N}{k} x^k (1-x)^{N-k} \frac{\varepsilon}{2} \\
&\leq \frac{\varepsilon}{2},
\end{aligned}$$

where the first inequality in the estimation of I_1 holds by the uniform continuity of $f \in C[0,1]$ and choosing δ small enough. Finally, for the sum I_2 we obtain

$$\begin{aligned}
I_2 &= \sum_{k:\left|\frac{k}{N}-x\right|\geqslant\delta} \binom{N}{k} x^k (1-x)^{N-k} \left| f\left(\frac{k}{N}\right) - f(x) \right| \\
&\leq 2M \sum_{k:\left|\frac{k}{N}-x\right|\geqslant\delta} \binom{N}{k} x^k (1-x)^{N-k} \\
&= 2MP\left[\left| \frac{B_{N,x}}{N} - x \right| \geqslant \delta \right] \leq 2M \frac{x(1-x)}{N\delta^2} \leq \frac{M}{2N\delta^2} \leq \frac{\varepsilon}{2},
\end{aligned}$$

where the first inequality in the estimation of I_2 holds by the boundedness of f ($|f(x)| \leq M$ for all $x \in [0,1]$) and the last inequality holds for large enough N. □

6.2 De Moivre-Laplace Theorem and Ramifications

Theorem 6.4 (De Moivre-Laplace)**.** *Denote $B_{N,p}$ the binomial random variable as above. Then*

$$\lim_{N\to\infty} \frac{P[B_{N,p}=m]}{\frac{1}{\sqrt{2\pi Np(1-p)}}e^{-\frac{1}{2}\frac{(m-Np)^2}{Np(1-p)}}} = 1,$$

uniformly over m such that $\frac{|m-Np|}{\sqrt{Np(1-p)}}$ remains bounded. Summing the local approximation over m in the range $a \leqslant \frac{m-Np}{\sqrt{Np(1-p)}} \leqslant b$ (a Riemann sum of the Gaussian) yields the weaker, distributional consequence

$$\lim_{N\to\infty} P[a \leqslant \frac{B_{N,p}-Np}{\sqrt{Np(1-p)}} \leqslant b] = \int_{x=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

In using the De Moivre-Laplace Theorem in practice we have to use the cdf,

$$\Phi(b) = \int_{x=-\infty}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \qquad \Rightarrow \qquad \int_{x=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = \Phi(b) - \Phi(a). \tag{1}$$

One checks that

$$\lim_{b\to+\infty} \Phi(b) = 1, \qquad \lim_{b\to-\infty} \Phi(b) = 0, \qquad \Phi(0) = \frac{1}{2}, \qquad 1 - \Phi(b) = \Phi(-b).$$

Example 6.5 (Application of De Moivre-Laplace & continuity correction)**.** *If we roll a fair die 720 times, find the probability of getting exactly 113 sixes.*
Solution using the binomial distribution: "success" in an individual trial now has probability $1/6$ *and* $N = 720$*, so*

$$P[\textit{113 sixes in 720 rolls}] = \binom{720}{113} (\frac{1}{6})^{113} (\frac{5}{6})^{607}.$$

This is the exact answer, but observe that it is computationally demanding by hand (and even a bit numerically unstable on the computer, due to division of huge numbers yielding small results) to perform. In any case the result is ≈ 0.0318*.*
Solution using the De Moivre-Laplace: the first thing we need to deal with here is that the binomial is a discrete RV while gaussian is continuous. Thus we will apply the so-called **continuity correction** *of articulating the event $B_{N,p} = 113$ in a way compatible with the (continuous) Gaussian RV appearing in the De Moivre-Laplace Theorem:*

$$P[\textit{113 sixes in 720 rolls}] = P[112.5 < B_{N,p} < 113.5] = P[-7.5 < B_{N,p} - \underbrace{Np}_{120} < -6.5] =$$
$$= P[-0.75 < \frac{B_{N,p}-Np}{\underbrace{\sqrt{Np(1-p)}}_{10}} < -0.65] \approx \Phi(-0.65) - \Phi(-0.75).$$

Of course a computation of the cdf is now required. However this is a quadrature of a smooth

and rapidly decaying function, i.e. integrating it numerically can be done accurately, quickly and is very stable. Once we implement its integral, we can then look up any similar probability. The answer we find in this case is ≈ 0.0312, *which is correct to within* 2%.

The De Moivre-Laplace hints towards a convergence of the relative frequency

$$\frac{\#\text{ of successes in } N \text{ independent trials}}{N}$$

towards p, and moreover with a specific rate of convergence. A corollary of Theorem 6.4 is that

$$\lim_{N\to\infty} P\Big[\frac{|B_{N,p} - Np|}{\sqrt{Np(1-p)}} > 2\Big] = 2\big(1 - \Phi(2)\big) \approx 4.55\% \leqslant 5\%$$

and

$$\lim_{N\to\infty} P\Big[\frac{|B_{N,p} - Np|}{\sqrt{Np(1-p)}} > 3\Big] = 2\big(1 - \Phi(3)\big) \approx 0.27\% \leqslant 0.3\%.$$

Now for the first time we have a consistency criterion for our stochastic modelling: if we observe a large number of independent trials, our findings must start resembling the DeMoivre-Laplace theorem – *persistent disagreement is evidence against the model of independent identically distributed trials.* (Note what is and is not being claimed: the failure could sit in the independence, in the constancy of p, or elsewhere – the data alone does not say which assumption broke.) Moreover, we can answer a number of questions quantitatively now:

Example 6.6 (Confidence Intervals)**.** *Assume now we have a biased coin with success probability* p. *According to the Bernoulli LLN* $p = \lim_{N\to\infty} \frac{B_{N,p}}{N}$; *however, in practice we only have a finite* N, *and we can evaluate quantitatively the impact of this finite* N. *Denote* $\hat{p} = \frac{B_{N,p}}{N}$. *What is the probability that* $\hat{p}$ *is no more than* ε *away from* p*?*

$$\begin{aligned}
\alpha &= P\left[|\hat{p} - p| < \varepsilon\right] = P\left[\left|\frac{B_{N,p}}{N} - p\right| < \varepsilon\right] = P\left[-N\varepsilon < B_{N,p} - Np < N\varepsilon\right] \\
&= P\left[-\varepsilon\sqrt{\frac{N}{p(1-p)}} < \frac{B_{N,p} - Np}{\sqrt{Np(1-p)}} < \varepsilon\sqrt{\frac{N}{p(1-p)}}\right] \\
&\approx \Phi\left(\varepsilon\sqrt{\frac{N}{p(1-p)}}\right) - \Phi\left(-\varepsilon\sqrt{\frac{N}{p(1-p)}}\right) \\
&= 2\Phi\left(\varepsilon\sqrt{\frac{N}{p(1-p)}}\right) - 1.
\end{aligned}$$

Since $p \in [0,1]$ *and thus* $p(1-p) \leq \frac{1}{4}$, *we obtain*

$$\varepsilon\sqrt{\frac{N}{p(1-p)}} \geq 2\varepsilon\sqrt{N}. \tag{2}$$

Furthermore, since $\Phi(t)$ *is increasing we get*

$$P\left[|\hat{p} - p| < \varepsilon\right] \gtrsim 2\Phi\left(2\varepsilon\sqrt{N}\right) - 1,$$

the $\gtrsim$ *recording that a normal approximation was made above. Everything that follows is therefore a* normal-approximation design rule *– excellent in practice for large* N *– and not a rigorous finite-*N *guarantee. (Rigorous finite-*N *bounds follow from concentration inequalities, e.g. Chebychev as in the Bernoulli LLN, at the price of substantially larger constants.) Q: What is the* N *needed to find* p *within* 0.05 *(i.e.,* $\varepsilon = 0.05$*) with probability at least* 99%*? A:* (2) *implies that*

$$2\Phi\left(2\varepsilon\sqrt{N}\right) - 1 \geq 0.99 \;\Rightarrow\; \Phi\left(2\varepsilon\sqrt{N}\right) \geq \frac{1.99}{2} \approx 0.995$$
$$\Rightarrow 2 \times 0.05 \times \sqrt{N} \geq \Phi^{-1}(0.995) \approx 2.58 \;\Rightarrow\; N \geq 665.64.$$

Hence the N *needed is* $N = 666$.
Q: Given N*, what is the* ε *we can report at a given confidence level (e.g.* 95%*)?*
A: Using (2) *we get*

$$P\left[|\hat{p} - p| < \varepsilon\right] \geq 95\% \;\Rightarrow\; \Phi\left(2\varepsilon\sqrt{N}\right) \geq \frac{1.95}{2} \;\Rightarrow\; \Phi\left(2\varepsilon\sqrt{N}\right) \geq 0.975 \;\Rightarrow\; 2\varepsilon\sqrt{N} \geq 1.96.$$

For $N = 1000$*, the relation* $2\varepsilon\sqrt{N} \geq 1.96$ *implies* $\varepsilon \geq 0.031$ *with* 95% *confidence.*

6.3 The law of rare events: an introduction to the Poisson process

The De Moivre-Laplace Theorem describes the binomial $B_{N,p}$ for large N and *fixed* p: the count concentrates around Np, with Gaussian fluctuations of size $\sqrt{Np(1-p)}$. A different – and equally important – regime is that of **rare events**: many trials, each individually unlikely, so that the *expected* number of successes stays of order one. The binomial then has a different limit:

Theorem 6.7 (Poisson limit)**.** *Fix* $\lambda > 0$ *and* $r \in \mathbb{N} \cup \{0\}$*. Then*

$$\lim_{N\to\infty} P[B_{N,\frac{\lambda}{N}} = r] = e^{-\lambda}\frac{\lambda^r}{r!},$$

i.e. the binomial distribution with $N \to \infty$*,* $p \to 0$*,* $Np = \lambda$ *converges to the Poisson distribution with parameter* λ *(cf. Example 5.10).*

Proof: Setting $p = \frac{\lambda}{N}$ in the binomial pdf,

$$\binom{N}{r}p^r(1-p)^{N-r} = \tfrac{N!}{r!(N-r)!}\left(\tfrac{\lambda}{N}\right)^r\left(1-\tfrac{\lambda}{N}\right)^{N-r} = \tfrac{N!}{(N-r)!}\tfrac{1}{N^r} \cdot \tfrac{\lambda^r}{r!} \cdot \left(1-\tfrac{\lambda}{N}\right)^{N-r} =$$

$$= \tfrac{N}{N}\cdot\tfrac{N-1}{N}\cdots\tfrac{N-r+1}{N} \cdot \tfrac{\lambda^r}{r!} \cdot \left(1-\tfrac{\lambda}{N}\right)^{N}\cdot\left(1-\tfrac{\lambda}{N}\right)^{-r},$$

and the proof is concluded by the elementary limits

$$\lim_{N\to\infty}\frac{N}{N}\cdot\frac{N-1}{N}\cdots\frac{N-r+1}{N} = 1, \qquad \lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^N = e^{-\lambda}, \qquad \lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^{-r} = 1.$$

□

Remarks:

- The De Moivre-Laplace Theorem 6.4 and the Poisson limit are the two complementary limits of the binomial: *common* events give Gaussian fluctuations around a large mean; *rare* events give Poisson-distributed counts of order one.

- This explains the ubiquity of the Poisson distribution: whenever a count is the aggregate of very many independent opportunities, each of small probability – calls placed by many subscribers in a given second, misprints among thousands of characters, decays among astronomically many nuclei – the law of rare events applies, regardless of the details.

- **An elementary first look at the Poisson process.** Let the observation window itself become a time variable: suppose events occur along time "at rate λ", and denote $N(t)$ the number of events in $(0,t]$. Split $(0,t]$ into $N \gg 1$ slots of length $\frac{t}{N}$; if each slot contains an event with probability $\approx \frac{\lambda t}{N}$, independently of the others, then $N(t) = B_{N,\lambda t/N}$, and the Poisson limit gives

$$P[N(t)=n] = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \qquad \mathrm{E}[N(t)] = \lambda t.$$

  In particular $P[N(t)=0] = e^{-\lambda t}$: the waiting time until the first event is exponential (cf. Example 4.6). The family $\{N(t)\}_{t\geqslant 0}$ – the first genuine random process we construct from scratch – is the **Poisson process**; its systematic treatment, as the simplest continuous-time Markov chain, is given in Chapter 11.

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 6.1. *Prove that*

$$f_{B_{N,p}}(x) = \binom{N}{x} p^x (1-p)^{N-x}, \qquad x \in \{0,1,2,\dots N\}.$$

Moreover prove that

$$\sum_{x=0}^{N} f_{B_{N,p}}(x) = 1, \qquad \mathrm{E}[B_{N,p}] = Np,$$

$$\mathrm{E}[(B_{N,p})^2] = Np + N(N-1)p^2, \qquad \mathrm{Var}[B_{N,p}] = Np(1-p).$$

Exercise 6.2. *Prove Theorem 6.2 by adapting the proof of Theorem 6.1.*

Exercise 6.3 (Proof of the De Moivre-Laplace Theorem)**.** *(*) Prove Theorem 6.4 along the following lines. Denote* $P_N(m) = P[B_{N,p} = m]$ *and* $x = \frac{m-Np}{\sqrt{Np(1-p)}}$.

1. *Using Stirling's approximation* $s! \approx \sqrt{2\pi s}\, s^s e^{-s}$, *show that*

$$P_N(m) \approx \frac{1}{\sqrt{2\pi}} \sqrt{\frac{N}{m(N-m)}}\, A, \qquad A := \left(\frac{Np}{m}\right)^m \left(\frac{N(1-p)}{N-m}\right)^{N-m}.$$

2. *Show that* $\frac{Np}{m} = \left(1 + x\sqrt{\frac{1-p}{Np}}\right)^{-1}$ *and* $\frac{N(1-p)}{N-m} = \left(1 - x\sqrt{\frac{p}{N(1-p)}}\right)^{-1}$.

3. *Taylor expand* $\log A$ *(note that* $x\sqrt{\frac{1-p}{Np}}$ *and* $x\sqrt{\frac{p}{N(1-p)}}$ *are* $o(1)$ *as* $N \to \infty$, *uniformly for bounded* x*) to obtain* $\log A = -\frac{x^2}{2} + O(N^{-1/2})$.

4. *Conclude that* $P_N(m) \approx \frac{1}{\sqrt{2\pi Np(1-p)}} e^{-\frac{x^2}{2}}$.

Exercise 6.4. *How many independent tosses of a biased coin are needed to determine its success probability* p *within* ± 0.05 *with* 99% *confidence? Within* ± 0.01*?*

Exercise 6.5. *A page of* 500 *characters is typed, each character being mistyped with probability* $\frac{1}{200}$, *independently. Using the law of rare events, estimate the probability that the page contains at least two misprints.*

Exercise 6.6 (The Shannon-McMillan entropy theorem)**.** *(*) Let* X_n *be i.i.d. with values in* $\{1, \dots, r\}$, $P[X_n = j] = p_j > 0$, *and let*

$$H := -\sum_{j=1}^{r} p_j \ln p_j$$

(the **entropy** *of the source). Denote* $\vec{\Omega}$ *the set of strings* $\vec{\omega} = (\omega_1, \dots, \omega_N)$. *Show that for every* $\varepsilon > 0$ *and* N *large enough there exists a set of "typical strings"* $\vec{\Omega}_N \subseteq \vec{\Omega}$ *such that*

(i). $e^{N(H-\varepsilon)} \leqslant |\vec{\Omega}_N| \leqslant e^{N(H+\varepsilon)}$;

(ii). $P[\vec{\Omega}_N] \geqslant 1 - \varepsilon$;

(iii). every $\vec{\omega} \in \vec{\Omega}_N$ *has* $e^{-N(H+\varepsilon)} \leqslant P[\{\vec{\omega}\}] \leqslant e^{-N(H-\varepsilon)}$.

(Hint: take $\vec{\Omega}_N = \{\vec{\omega} : |\frac{B_N^j}{N} - p_j| \leqslant \delta \ \forall j\}$, where B_N^j counts the occurrences of j; then (ii) is the Bernoulli LLN. For (iii), write $P[\{\vec{\omega}\}] = p_1^{B_N^1} \cdots p_r^{B_N^r} = \exp[N \sum_j \frac{B_N^j}{N} \ln p_j]$ and compare the exponent with $-NH$. For (i), sandwich: $1 \geqslant P[\vec{\Omega}_N] \geqslant \frac{1}{2}$ for large N, and apply (iii) – with $\frac{\varepsilon}{2}$ in place of ε for the lower bound.)

Comment: *since* $H \leqslant \ln r$, *with equality only for uniform* p_j, *the theorem says that a source with non-uniform statistics effectively "lives" on* $e^{NH} \ll r^N$ *typical strings – the engine behind data compression. It is also a practical detector: the empirical entropy of a symbol stream distinguishes* text in an unknown or encoded language *(whose* H *sits well below* $\ln r$, *like every natural language) from* a random collection of symbols *(*$H \approx \ln r$*) – without deciphering anything.*

7 Continuous Random Variables: pdfs, moments, correlation and independence

7.1 Expectation and Variance

In this section we will assume $\mathbb{B} = \mathbb{R}^N$ for some $N \in \mathbb{N}$. Thus an induced measure will be a measure on $\mathbb{R}^N$. What do all possible probability measures on $\mathbb{R}^N$ look like? The answer is given by the following

Theorem 7.1 (Lebesgue-Radon-Nikodym Theorem)**.** *Let P be a probability measure on $\mathbb{R}^N$ and μ be the Lebesgue measure on $\mathbb{R}^N$. Then P can be decomposed to three parts,*

$$P = \nu + \tau + \sigma$$

where ν is absolutely continuous with respect to Lebesgue measure, τ is a pure point measure, and σ is a singular continuous measure (i.e. σ is concentrated on a set of Lebesgue measure zero, but has no atoms).

This means that there exists a measurable function $f : \mathbb{R}^N \to \mathbb{R}$ with $f \geqslant 0$, $\int\limits_{\mathbb{R}^N} f(x)d\mu(x) \leqslant 1$ so that

$$\nu(A) = \int\limits_{x \in A} f(x)dx,$$

and moreover there exist $x_j \in \mathbb{R}^N, p_j \in [0, \infty)$ (finite or countably infinite) so that

$$\tau = \sum_j p_j \delta(x - x_j).$$

Singular continuous measures exist (the classical example is the Cantor measure), but they rarely appear in applications; in everything that follows we will assume $\sigma = 0$.

Remarks:

- In this section we will consider that all the probability measures on $\mathbb{R}^N$ that we encounter are absolutely continuous. Thus, **when we refer to a continuous R.V., unless otherwise specified, we will assume that the induced probability measure is absolutely continuous**, and can therefore be completely described by its **probability density function** (pdf) f_X,

  $$P_X(A) = \int\limits_{x \in A} f_X(x)dx.$$

  This use of terminology is widespread.

- For a continuous R.V. in the sense discussed above, the **cumulative distribution function** (cdf)

  $$F_X(x) = P[X \leqslant x]$$

is continuous, and

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy, \qquad \text{with } F_X'(x) = f_X(x) \text{ for almost every } x.$$

When we say that we know how a R.V. is distributed (or, colloquially, that "we know a R.V.") we mean that we know either the cdf F_X or the pdf f_X (and of course then we can pass from one to the other).

- Theorem 7.1 is a combination of the generalised versions of two standard Theorems, namely the Lebesgue decomposition theorem, and the Radon-Nikodym theorem. The generalisation with respect to the textbook versions refers to the fact that Lebesgue measure is σ-finite, and not finite.

Definition 7.2 (Mathematical expectation)**.** *Let $X : \Omega \to \mathbb{R}^N$ be a continuous R.V. Its mathematical expectation is defined as*

$$\mathrm{E}[X] := \int_{\Omega} X(\omega)dP(\omega) = \int_{\mathbb{R}^N} x f_X(x)dx$$

if the integrals exist (i.e. if $\int_{\Omega} |X(\omega)|dP(\omega) < \infty$).

(The equivalence of the two expressions is Exercise 7.1.)
Remarks:

- The mathematical expectation is also called expectation, mean, average, mean value. We will use all these terms interchangeably.
- The random variable X with pdf

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

has no expectation since

$$\lim_{R\to\infty} \int_{-R}^{R} \frac{|x|}{\pi(1+x^2)}dx = +\infty.$$

Whenever we use the expectation, technically we should assume (or prove from other assumptions) that it is well defined and finite. Such an assumption is often omitted for brevity, however we should be mindful of its necessity.

Lemma 7.3 (Properties of the mathematical expectation)**.** *Assume X, Y are continuous R.V. with finite expectations. Then*

1. $\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]$.
2. $A \leqslant X(\omega) \leqslant B \quad \forall \omega \in \Omega \qquad \Longrightarrow \qquad A \leqslant \mathrm{E}[X] \leqslant B$

3. *If* $Y = g(X)$ *for some function* $g : \mathbb{R}^N \to \mathbb{R}$, *then*

$$\mathrm{E}[Y] = \int_{\Omega} g(X(\omega))dP(\omega) = \int_{\mathbb{R}^N} g(x) f_X(x)dx.$$

Definition 7.4 (Variance)**.** *Let* $X : \Omega \to \mathbb{R}$ *be a continuous R.V. Its variance is defined as*

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

if it exists. The variance is a central moment of the pdf, namely

$$\mathrm{Var}[X] = \int_{\mathbb{R}} (x - \mathrm{E}[X])^2 f_X(x)dx.$$

(The equivalence of the two expressions is Exercise 7.2.)

Lemma 7.5 (Properties of the Variance)**.** *Assume* $\mathrm{E}[X]$, $\mathrm{Var}[X]$ *exist, and* $a, b \in \mathbb{R}$. *Then*

1. $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$
2. $A \leqslant X(\omega) \leqslant B \quad \forall \omega \in \Omega \qquad \Longrightarrow \qquad \mathrm{Var}[X] \leqslant \big(\frac{B-A}{2}\big)^2$

Proof:

1. $\mathrm{E}[(aX + b - \mathrm{E}[aX + b])^2] = \mathrm{E}[\big(a(X - \mathrm{E}[X])\big)^2] = a^2 \mathrm{E}[(X - \mathrm{E}[X])^2] = a^2 \mathrm{Var}[X]$.
2. Using property 1 above,

$$\mathrm{Var}[X] = \mathrm{Var}[X - \tfrac{A+B}{2}] = \mathrm{E}[(X - \tfrac{A+B}{2})^2] - \mathrm{E}[X - \tfrac{A+B}{2}]^2 \leqslant \mathrm{E}[(X - \tfrac{A+B}{2})^2] =$$
$$= \int_{\Omega} \big(X - \tfrac{A+B}{2}\big)^2 dP(\omega) \leqslant (\tfrac{B-A}{2})^2 \underbrace{\int_{\Omega} dP(\omega)}_{1}$$

   where in the last step we used the bound

$$A \leqslant X \leqslant B \qquad \Longrightarrow \qquad \frac{A-B}{2} \leqslant X - \frac{A+B}{2} \leqslant \frac{B-A}{2}.$$

7.2 Excursion to Analysis: The Riemann-Stieltjes integral *

Definition 7.6 ($p-$variation)**.** *Let* $f : [a, b] \to \mathbb{R}$ *and* $\Pi = \{t_j | j = 0, 1, \ldots, n\}$ *be a partition of* $[a, b]$. *Then the* $p-$*variation with respect to the partition* Π *is defined as*

$$V_p(f, \Pi) = \sum_{j=0}^{n-1} |f(t_{j+1}) - f(t_j)|^p.$$

We will say that the $p-$*variation of* f *is well defined and equal to* v,

$$V_p(f) = v$$

iff

$$\forall \varepsilon > 0 \quad \exists \text{ partition } \Pi_\varepsilon \ : \ |V_p(f, \Pi) - v| < \varepsilon \quad \forall \Pi \geqslant \Pi_\varepsilon,$$

where $\Pi \geqslant \Pi_\varepsilon$ means that Π is a refinement of Π_ε (i.e. Π contains all the points of Π_ε). $V_1(f)$ is often called the total variation, and $V_2(f)$ is called the quadratic variation.

Remark 7.7 (A warning on conventions)**.** *The definition above takes a limit* under refinements *(a "net" limit). For $p = 1$ the sums can only increase under refinement (triangle inequality), so this agrees with the more common definition $V_1(f) = \sup_\Pi V_1(f, \Pi)$; for $p > 1$ the refinement limit and the supremum genuinely differ. Moreover, in stochastic analysis "quadratic variation" refers to yet a third notion – a limit along partitions of vanishing mesh, in mean square or in probability. For Brownian motion the distinction is essential: see the warning following Theorem 12.11.*

Definition 7.8 (The Riemann-Stieltjes integral)**.** *Let $f, g : [a, b] \to \mathbb{R}$. We will say that the Riemann-Stieltjes integral*

$$\int_{t=a}^{b} f(t)dg(t) = I \in \mathbb{R}$$

exists iff

$$\forall \varepsilon > 0 \quad \exists \text{ partition } \Pi_\varepsilon \ : \ |\sum_j f(s_j)(g(s_{j+1}) - g(s_j)) - I| < \varepsilon \quad \forall \Pi = \{s_j\} \geqslant \Pi_\varepsilon.$$

If the above condition does not hold we will say that the integral diverges. Moreover, if

$$\forall M > 0 \quad \exists \text{ partition } \Pi_M \ : \ \pm \sum_j f(s_j)(g(s_{j+1}) - g(s_j)) > M \quad \forall \Pi = \{s_j\} \geqslant \Pi_M,$$

we will say that the integral diverges to $\pm\infty$.

Note that the definition fixes the sample point at the *left* endpoint s_j. For the integrals of this section the choice of sample point ends up being immaterial; famously, for the stochastic integrals of Chapter 15 it will not be.

Theorem 7.9. *If $f \in C[a, b]$, $g \in C^1[a, b]$, then the Riemann-Stieltjes integral $\int f dg$ exists and it is equal to a Riemann integral,*

$$\int_{t=a}^{b} f(t)dg(t) = \int_{t=a}^{b} f(t)g'(t)dt.$$

Theorem 7.10. *If $f \in C[a, b]$ and $V_1(g) < +\infty$, then the Riemann-Stieltjes integral*

$$\int_{t=a}^{b} f(t)dg(t)$$

exists.

If $V_1(g) = +\infty$, then there are functions $f \in C[a, b]$ for which the integral $\int f dg$ diverges.

Using the Riemann-Stieltjes integral, we have

$$P[X \in A] = \int_A f_X(x)dx = \int_A dF_X(x)$$

for any interval A. Observe that the latter expression makes sense even when the cdf has jumps (i.e. when the RV is discrete or mixed, in addition to when it is continuous). Strictly speaking, in the presence of jumps the right framework is the **Lebesgue-Stieltjes** integral: one builds the measure $\mu_F\big((a,b]\big) := F(b) - F(a)$ by Carathéodory extension, exactly as in Chapter 3, and integrates against it – endpoint conventions then matter, e.g. $\mu_F(\{x\})$ is the jump of F at x. The elementary Riemann-Stieltjes definition above is what we will actually use, as motivation for stochastic integration.

7.3 Covariance and correlation

Definition 7.11 (Covariance and Correlation coefficient)**.** *Let X, Y be two R.V. with values in $\mathbb{R}$. Then their covariance is defined as*

$$\mathrm{Cov}[X,Y] = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

if it exists. Moreover, assuming $0 < \mathrm{Var}[X], \mathrm{Var}[Y] < \infty$, their correlation coefficient is defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\ \mathrm{Var}[Y]}}.$$

Lemma 7.12 (Properties of the Covariance)**.** *1.* $(\mathrm{Cov}[X,Y])^2 \leqslant \mathrm{Var}[X] \cdot \mathrm{Var}[Y]$.

2. If $|\rho(X,Y)| = 1$, then there exist $a, b \in \mathbb{R}$ so that

$$Y = aX + b \qquad \textit{almost surely.}$$

3. $|\rho(aX + b, cY + d)| = |\rho(X,Y)|$ for any $a, b, c, d \in \mathbb{R}$ with $ac \neq 0$.

Proof: Consider the R.V. $Z := \Big(t(X - \mathrm{E}[X]) + (Y - \mathrm{E}[Y])\Big)^2$, and observe that $Z \geqslant 0$. Thus it follows that

$$0 \leqslant \mathrm{E}[\Big(t(X - \mathrm{E}[X]) + (Y - \mathrm{E}[Y])\Big)^2] = t^2\,\mathrm{Var}[X] + 2t\,\mathrm{Cov}[X,Y] + \mathrm{Var}[Y].$$

Thus the discriminant for the quadratic equation

$$t^2\,\mathrm{Var}[X] + 2t\,\mathrm{Cov}[X,Y] + \mathrm{Var}[Y] = 0$$

is

$$\Delta \leqslant 0 \implies 4\,\mathrm{Cov}[X,Y]^2 - 4\,\mathrm{Var}[X]\,\mathrm{Var}[Y] \leqslant 0.$$

This proves property 1.

Moreover, using again this discriminant computation, observe that

$$|\rho(X,Y)| = 1 \implies \mathrm{Cov}[X,Y]^2 = \mathrm{Var}[X]\,\mathrm{Var}[Y] \implies \exists t_* \,:\, t_*^2\,\mathrm{Var}[X] + 2t_*\,\mathrm{Cov}[X,Y] + \mathrm{Var}[Y] = 0$$

and hence

$$\mathrm{E}[\Big(t_*(X - \mathrm{E}[X]) + (Y - \mathrm{E}[Y])\Big)^2] = 0.$$

But this is now a non-negative R.V. with mean 0; by virtue of the First Chebychev Inequality, Lemma 5.11, it follows that

$$P[Y = \underbrace{-t_*}_{a} X + \underbrace{t_*\,\mathrm{E}[X] + \mathrm{E}[Y]}_{b}] = 1,$$

thus property 2 follows.

For property 3, observe that

$$\rho(aX+b, cY+d) = \frac{\mathrm{E}[(aX - a\,\mathrm{E}[X])(cY - c\,\mathrm{E}[Y])]}{\sqrt{\mathrm{Var}[aX+b]\,\mathrm{Var}[cY+d]}} = \frac{ac\,\mathrm{Cov}[X,Y]}{|ac|\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} = \frac{ac}{|ac|}\rho(X,Y).$$

Example 7.13. *Recall Example 5.6. Compute $\rho(X,Y)$.*

Solution: One needs to use all the definitions in the discrete setting, but in this case it is straightforward. In Example 5.6 it was seen that

$$f_X(1) = \frac{1}{13}, \qquad f_Y(1) = \frac{1}{13}$$

and

$$f_{XY}(1,1) = \tfrac{12}{52\cdot 51}, \qquad f_{XY}(0,1) = f_{XY}(1,0) = \tfrac{48\cdot 4}{52\cdot 51}, \qquad f_{XY}(0,0) = \tfrac{48\cdot 47}{52\cdot 51}.$$

Thus we compute

$$\mathrm{E}[X] = \frac{1}{13} = \mathrm{E}[Y], \qquad \mathrm{Var}[X] = \frac{1}{13} - \frac{1}{13^2} = \frac{12}{13^2} = \mathrm{Var}[Y]$$

and

$$\mathrm{Cov}[X,Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \tfrac{12}{13}\tfrac{12}{13}\tfrac{12}{52\cdot 51} - 2\tfrac{12}{13^2}\tfrac{48\cdot 4}{52\cdot 51} + \tfrac{1}{13^2}\tfrac{48\cdot 47}{52\cdot 51} = \\ = \tfrac{12^3 - 8\cdot 12\cdot 48 + 48\cdot 47}{13^2\cdot 52\cdot 51} = \tfrac{12}{13^2}\tfrac{-52}{52\cdot 51} = \tfrac{-12}{13^2\cdot 51}$$

leading to

$$\rho(X,Y) = \frac{\frac{-12}{13^2\cdot 51}}{\frac{12}{13^2}} = -\frac{1}{51} \approx -0.0196$$

The space of variables with finite variance is in fact a *Hilbert space* – the geometry that powers Chapter 13; setting this up carefully is Exercise 7.3.

Remark: The motivation behind taking central moments, i.e. working with $X - \mathrm{E}[X]$ instead of X, is precisely to be in a setup where the variance can be tied into a Hilbert space structure as above. This allows the use of powerful, robust techniques pretty much "for free".

Thus, if X_n, X have mean zero,

$$\|X_n - X\|_{L^2(\Omega)} \to 0 \iff \mathrm{Var}[X_n - X] \to 0 \implies P[|X_n - X| > t] \to 0 \ \forall t > 0$$

by virtue of Chebychev's Inequality, Lemma 5.12. We will see concrete examples of techniques exploiting this later.

Lemma 7.14 (Variance of a sum of R.V.)**.** *Assume $\{X_n\}_{n\in\mathbb{N}}$ are continuous R.V. with finite mean and variance. Then*

$$\mathrm{Var}[\sum_{n=1}^{N} X_n] = \sum_{n=1}^{N} \mathrm{Var}[X_n] + 2 \sum_{1\leqslant n<m\leqslant N} \mathrm{Cov}[X_n, X_m].$$

Definition 7.15 (Independent R.V.)**.** *The family $\{X_n\}_{n\in\mathbb{N}}$ of R.V. is independent if, for any collection $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{B}$ the events*

$$X_n \in A_n$$

are collectively independent. It follows that when a joint pdf exists, *it factorises,*

$$f_{X_1 X_2 \ldots X_N}(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} f_{X_n}(x_n),$$

and conversely, if a joint pdf exists and factorises as above, the R.V. are independent. (A caveat: independence of individually continuous R.V. does not *by itself guarantee that a* joint *pdf exists – when we write joint pdfs, joint absolute continuity is being assumed.)*

Lemma 7.16 (Dependence and correlation)**.** *If X, Y are independent R.V. they are uncorrelated, i.e.* $\mathrm{Cov}[X, Y] = 0$.

The converse is not true.

Proof: The reader should readily check the first claim using the definition of covariance, and the fact about joint pdfs of independent R.V. outlined in definition 7.15.

For the second claim, consider the R.V. X with pdf

$$f_X(x) = \frac{1}{2}\chi_{[-1,1]}(x) = \begin{cases} \frac{1}{2}, & \text{if } x \in [-1,1], \\ 0, & \text{if } x \notin [-1,1]. \end{cases}$$

Then X, X^2 are uncorrelated by symmetry, but obviously not independent. To check the non-correlation, observe that

$$\mathrm{Cov}[X, X^2] = \int_\Omega (X(\omega) - 0)(X^2(\omega) - \mathrm{E}[X^2])dP(\omega) =$$
$$= \int_{\{\omega: X(\omega)>0\}} X(\omega)(X^2(\omega) - \mathrm{E}[X^2])dP(\omega) + \int_{\{\omega: X(\omega)<0\}} X(\omega)(X^2(\omega) - \mathrm{E}[X^2])dP(\omega) =$$
$$= \int_{\{\omega: X(\omega)>0\}} X(\omega)(X^2(\omega) - \mathrm{E}[X^2])dP(\omega) - \int_{\{\omega: X(\omega)>0\}} X(\omega)(X^2(\omega) - \mathrm{E}[X^2])dP(\omega) = 0.$$

7.4 Modes of Convergence of RV

There are several different notions of convergence of random variables. Let X_n, X be random variables on the probability space $(\Omega, \mathcal{E}, P)$.

(i) **Convergence in distribution**

$$X_n \to X \text{ in distribution} \iff \lim_{n\to\infty} P[X_n \leqslant x] = P[X \leqslant x]$$

for all x at which F_X is continuous.

(ii) **Convergence in probability**

$$X_n \to X \text{ in probability} \iff \lim_{n\to\infty} P[|X_n - X| > \epsilon] = 0 \quad \forall \epsilon > 0.$$

(iii) **Convergence in $r-$mean** Assume $r \geqslant 1$ and $\mathrm{E}[|X_n|^r]$, $\mathrm{E}[|X|^r] < \infty$. Then

$$X_n \to X \text{ in } r-\text{mean} \iff \lim_{n\to\infty} \mathrm{E}[|X_n - X|^r] = 0.$$

If $\boldsymbol{r = 1}$ this is also called **convergence in mean**.
If $\boldsymbol{r = 2}$ this is also called **convergence in mean square (m.s.)**.

(iv) **Almost sure convergence**

$$X_n \to X \text{ almost surely} \iff P[\lim_{n\to\infty} X_n = X] = 1.$$

Almost sure convergence is often shortened as **a.s.**, and also called **almost everywhere convergence (a.e.)** and **convergence with probability 1**.

(v) **Sure convergence**

$$X_n \to X \text{ surely} \iff \lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \forall \omega \in \Omega.$$

In general the mode of convergence lower in the list is stronger than the convergence earlier in the list, i.e.

$$\text{(v)} \implies \text{(iv)} \implies \text{(ii)} \implies \text{(i)}.$$

Moreover, convergence in $r-$mean implies convergence in probability (apply the Chebychev-Markov inequality to $|X_n - X|^r$), so

$$\text{(iii)} \implies \text{(ii)} \implies \text{(i)}.$$

On the other hand – and contrary to a common first impression – almost sure convergence does *not* imply convergence in $r-$mean, even when all the moments involved are finite. The standard counterexample lives on $\Omega = (0, 1)$ with the uniform probability: the random variables

$$X_n = n^{1/r}\, \chi_{(0,\frac{1}{n})}$$

converge to 0 almost surely, yet $\mathrm{E}[|X_n|^r] = 1$ for every n. To upgrade almost sure convergence to convergence in $r-$mean one needs additional control on the tails, e.g. domination: if $X_n \to X$ almost surely and $|X_n| \leqslant Y$ with $\mathrm{E}[Y^r] < \infty$, then $X_n \to X$ in $r-$mean (dominated convergence).

7.5 Law of Large Numbers

Corollary 7.17 (Weak Law of large numbers I)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be independent identically distributed (i.i.d.) continuous R.V. with finite mean and variance. Then*

$$P\left[\left|\frac{\sum_{n=1}^{N} X_n}{N} - \mathrm{E}[X_1]\right| > t\right] \leqslant \frac{\mathrm{Var}[X_1]}{Nt^2}$$

Outline of the proof: Lemma 7.14 and property 1 from Lemma 7.5 imply

$$\mathrm{Var}[\tfrac{\sum_{n=1}^{N} X_n}{N}] = \tfrac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}[X_n] + \tfrac{2}{N^2}\sum_{1\leqslant n<m\leqslant N}\mathrm{Cov}[X_n, X_m] = \tfrac{\mathrm{Var}[X_1]}{N},$$

where in the last step we used the i.i.d. assumption. The result follows from the Chebychev Inequality, Lemma 5.12.

Corollary 7.18 (Weak Law of large numbers II)**.** *Assume $\{X_n\}_{n\in\mathbb{N}}$ have finite and uniformly bounded mean and variance, i.e. there is some $M > 0$ so that*

$$|\mathrm{E}[X_n]| + \mathrm{Var}[X_n] \leqslant M \qquad \forall n \in \mathbb{N},$$

and moreover they are weakly correlated in the sense that

$$|n - m| > L \implies \mathrm{Cov}[X_n, X_m] = 0.$$

Then,

$$P\left[\left|\frac{\sum_{n=1}^{N} X_n}{N} - \sum_{n=1}^{N}\frac{\mathrm{E}[X_n]}{N}\right| > t\right] \leqslant \frac{1+2L}{N}\frac{M}{t^2}$$

The proof follows along the same lines as the previous one.

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 7.1. *Show that the two definitions of the mathematical expectation are equivalent.*

Exercise 7.2. *Show that the two definitions of the variance are equivalent.*

Exercise 7.3. *Using Property 1 of Lemma 7.12, show that the space of all variables with finite mean and variance,* after identifying variables that agree almost surely, *carries the well-defined inner product*

$$\langle X, Y \rangle_{L^2(\Omega)} = \int_\Omega X(\omega) Y(\omega) dP(\omega).$$

(Cauchy-Schwarz makes the pairing finite, and the inner-product axioms are then direct. That the space is moreover complete *– hence a Hilbert space, denoted $L^2(\Omega)$ – is a standard theorem of measure theory which we take as given, cf. [7].)*

Moreover, show that if we only consider the R.V. with mean zero, $\{X : \mathrm{E}[X] = 0,\ \mathrm{Var}[X] < \infty\}$ the square of the induced norm is

$$\|X\|^2_{L^2(\Omega)} = \langle X, X \rangle_{L^2(\Omega)} = \mathrm{Var}[X].$$

Exercise 7.4 (Gaussian integrals)**.** *Prove that $\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$.* (Hint: compute the square of the integral in polar coordinates.) *Using this, verify that the $\mathcal{N}(\mu, \sigma^2)$ pdf integrates to* 1, *and has mean μ and variance σ^2.*

Exercise 7.5. *Let X be exponential with parameter λ. Compute* $\mathrm{E}[X]$ *and* $\mathrm{Var}[X]$, *and prove the memoryless property*

$$P[X > x + y \,|\, X > x] = P[X > y].$$

Exercise 7.6 (Pareto distribution)**.** *The Pareto distribution with scale $\theta > 0$ and shape $a > 0$ has pdf $f_X(x) = \frac{a\theta^a}{x^{a+1}} \chi_{(\theta,\infty)}(x)$. Verify the normalisation and compute* $\mathrm{E}[X]$ *and* $\mathrm{Var}[X]$, *carefully noting for which a they exist.*

Exercise 7.7. *Express $P[(X, Y) \in [x_1, x_2] \times [y_1, y_2]]$ in terms of the joint cdf F_{XY}. Show that letting $x_2 \to x_1$, $y_2 \to y_1$ recovers the relation $f_{XY} = \frac{\partial^2}{\partial x \partial y} F_{XY}$.*

Exercise 7.8 (Triangular pdf)**.** *Let $a, b > 0$ and consider the triangular pdf which rises linearly from $-a$ to* 0 *and falls linearly from* 0 *to b. Normalise it, and compute its mean and variance. For $b > a$, discuss the sign of the skewness and the relative positions of mode, median and mean.*

Exercise 7.9 (Histograms)**.** *Let X have a smooth pdf f_X. Assuming that relative frequencies over many independent realisations converge to probabilities, show that the (appropriately normalised) histogram of the observations converges to f_X.*

(Hint: consider events of the form $X \in [n\Delta x, (n+1)\Delta x]$.)

Exercise 7.10. *For the uniform random variable $X \sim \mathcal{U}(a, b)$, compute the mean, the variance, the skewness $\beta = \mathrm{E}[(\frac{X-\mu}{\sigma})^3]$ and the kurtosis $\gamma = \mathrm{E}[(\frac{X-\mu}{\sigma})^4]$. (For comparison, the normal distribution has $\beta = 0$, $\gamma = 3$.)*

Exercise 7.11 (Probability integral transform)**.** *Let X be a continuous random variable with strictly increasing cdf F_X. Prove that $Y := F_X(X)$ is uniform on $[0, 1]$. (This is the basis of sampling arbitrary distributions from uniform random numbers.)*

8 Operations on Continuous Random Variables

In this section we continue to use the assumption that our continuous R.V. have no probability concentrated on single points, i.e. their cdf's are continuous.

8.1 Operations on one variable

Lemma 8.1 ($Y = \frac{1}{\lambda}X$)**.** *Let the continuous R.V. X have cdf $F_X(x)$ and pdf $f_X(x)$, and let $\lambda > 0$. Then the R.V. $Y = \frac{1}{\lambda}X$ is distributed with*

$$F_Y(y) = F_X(\lambda y), \qquad f_Y(y) = \lambda f_X(\lambda y)$$

Proof:

$$F_Y(t) = P[Y \leqslant t] = P[\frac{X}{\lambda} \leqslant t] = P[X \leqslant t\lambda] = F_X(\lambda t)$$

and therefore

$$f_Y(t) = \partial_t F_Y(t) = \partial_t F_X(\lambda t) = \lambda f_X(\lambda t).$$

□

Lemma 8.2 ($Y = X^2$)**.** *Let the continuous R.V. X have cdf $F_X(x)$. Then the R.V. $Y = X^2$ has cdf*

$$F_Y(y) = \begin{cases} 0, & y \leqslant 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}), & y > 0 \end{cases}$$

and pdf

$$f_Y(y) = \begin{cases} 0, & y \leqslant 0 \\ \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}, & y > 0 \end{cases}$$

Proof: For any $t > 0$

$$F_Y(t) = P[Y \leqslant t] = P[X^2 \leqslant t] = P[X \leqslant \sqrt{t} \ \cap \ X \geqslant -\sqrt{t}] = \\ = P[X \leqslant \sqrt{t} \ \setminus \ X \leqslant -\sqrt{t}] = F_X(\sqrt{t}) - F_X(-\sqrt{t}).$$

Obviously for $t \leqslant 0$, $F_Y(t) = 0$.

The pdf is computed by differentiation,

$$f_Y(t) = \frac{\partial}{\partial t} F_Y(t).$$

□

Lemma 8.3 ($Y = |X|$)**.** *Let the continuous R.V. X have cdf $F_X(x)$. Then the R.V. $Y = |X|$ has cdf*

$$F_Y(y) = \begin{cases} 0, & y \leqslant 0 \\ F_X(y) - F_X(-y), & y > 0 \end{cases}$$

and pdf

$$f_Y(y) = \begin{cases} 0, & y \leqslant 0 \\ f_X(y) + f_X(-y), & y > 0 \end{cases}$$

Proof: For any $t > 0$

$$F_Y(t) = P[Y \leqslant t] = P[|X| \leqslant t] = P[X \leqslant t \ \cap \ X \geqslant -t] = \\ = P[X \leqslant t \ \setminus \ X \leqslant -t] = F_X(t) - F_X(-t).$$

Obviously for $t \leqslant 0$, $F_Y(t) = 0$.

The pdf is computed by differentiation,

$$f_Y(t) = \frac{\partial}{\partial t} F_Y(t).$$

□

Lemma 8.4 ($Y = aX + b$, $a > 0$)**.** *Let the continuous R.V. X have cdf $F_X(x)$. Then the R.V. $Y = aX + b$ with $a > 0$ has cdf*

$$F_Y(y) = F_X(\frac{y-b}{a})$$

and pdf

$$f_Y(y) = \frac{1}{a} f_X(\frac{y-b}{a})$$

Proof: Observe that

$$F_Y(t) = P[Y \leqslant t] = P[aX + b \leqslant t] = P[X \leqslant \tfrac{t-b}{a}] = F_X(\tfrac{t-b}{a}).$$

The pdf is computed by differentiation. □

(The case $a < 0$ is Exercise 8.1.)

Lemma 8.5 ($Y = X^{2n+1}$, $n \in \mathbb{N}$)**.** *Let the continuous R.V. X have cdf $F_X(x)$. Then the R.V. $Y = X^{2n+1}$ with $n \in \mathbb{N}$ has cdf*

$$F_Y(y) = F_X(y^{\frac{1}{2n+1}})$$

and pdf

$$f_Y(y) = \frac{1}{2n+1} \, |y|^{\frac{-2n}{2n+1}} f_X(y^{\frac{1}{2n+1}})$$

Proof: For any $t \in \mathbb{R}$ (odd powers are increasing on all of $\mathbb{R}$)

$$F_Y(t) = P[Y \leqslant t] = P[X^{2n+1} \leqslant t] = P[X \leqslant t^{\frac{1}{2n+1}}] = F_X(t^{\frac{1}{2n+1}})$$

where we used the fact that x^{2n+1} is one-to-one, and odd-order (real) roots are uniquely defined for any $x \in \mathbb{R}$, e.g. $\sqrt[3]{-8} = -2$.

The pdf is computed by differentiation. □

Lemma 8.6 ($U = \frac{1}{X}$)**.** *Let the continuous R.V. X have cdf $F_X(x)$. Then the R.V. $U = \frac{1}{X}$ has cdf*

$$F_U(u) = \begin{cases} F_X(0) - F_X(\frac{1}{u}), & u < 0 \\ F_X(0), & u = 0 \\ 1 + F_X(0) - F_X(\frac{1}{u}), & u > 0 \end{cases}$$

and pdf

$$f_U(u) = \frac{1}{u^2} f_X(\frac{1}{u}), \qquad u \neq 0.$$

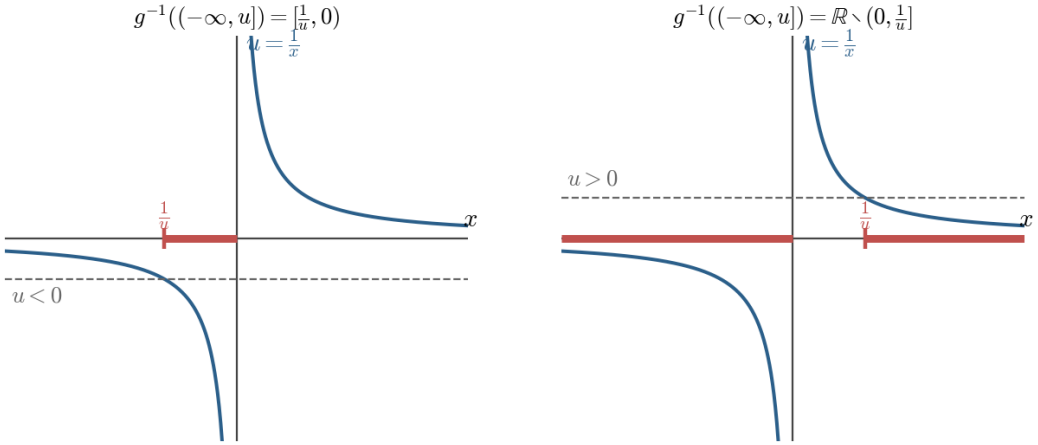
Proof: First of all, denote $g : \mathbb{R} \setminus \{0\} \to \mathbb{R} : x \mapsto \frac{1}{x}$. Since X is continuous, $P[X = 0] = 0$, so $U = g(X)$ is defined almost surely (define it arbitrarily on the null event $\{X = 0\}$). We work with exact preimages of g, and only then discard null endpoint events. Then,

$$u < 0 \implies g^{-1}\big((-\infty, u]\big) = [\tfrac{1}{u}, 0)$$

and

$$u > 0 \implies g^{-1}\big((-\infty, u]\big) = (-\infty, 0) \cup [\tfrac{1}{u}, +\infty)$$

see also the graph below.



This observation will drive the computation of the cdf: if $t < 0$, then

$$F_U(t) = P[U \leqslant t] = P[\tfrac{1}{X} \leqslant t] = P\big[g(X) \in (-\infty, t]\big] =$$
$$= P[X \in g^{-1}\big((-\infty, t]\big)] = P[\tfrac{1}{t} \leqslant X < 0] = F_X(0) - F_X(\tfrac{1}{t}),$$

Moreover, if $t > 0$

$$F_U(t) = P[U \leqslant t] = P[\tfrac{1}{X} \leqslant t] = P\big[g(X) \in (-\infty, t]\big] =$$
$$= P[X \in g^{-1}\big((-\infty, t]\big)] = P[X < 0] + P[X \geqslant \tfrac{1}{t}] = F_X(0) + 1 - F_X(\tfrac{1}{t}),$$

where in evaluating the cdf's we used $P[X = 0] = P[X = \frac{1}{t}] = 0$. For $t = 0$, $P[\frac{1}{X} \leqslant 0] = P[X < 0] = F_X(0)$ (again since $P[X = 0] = 0$).

The pdf is computed by differentiation. □

8.2 Joint pdfs and cdfs

When working with two R.V., e.g. X and Y, in general we must work with their joint distribution. **In this section we assume that (X, Y) is jointly absolutely continuous**, i.e. that there exists a **joint pdf** $f_{XY}(x, y)$ with

$$P[(X, Y) \in A \subseteq \mathbb{R}^2] = \iint\limits_{(x,y)\in A} f_{XY}(x, y)dxdy.$$

The **joint cdf** $F_{XY}(x,y)$ is defined as

$$F_{XY}(x,y) = P[X \leqslant x \ \cap Y \leqslant y]$$

and therefore it follows that

$$F_{XY}(x,y) = \int_{a=-\infty}^{x} \int_{b=-\infty}^{y} f_{XY}(a,b)dadb \iff f_{XY} = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y) \quad \text{(a.e.)}$$

The pdf and cdf for the individual R.V. can be extracted from the joint ones through

$$F_X(x) = P[X \leqslant x \ \cap \ Y \in \mathbb{R}] = \lim_{y \to \infty} F_{XY}(x,y)$$

and, accordingly,

$$f_X(x) = \tfrac{\partial}{\partial x} F_X(x) = \tfrac{\partial}{\partial x} \lim_{y\to\infty} F_{XY}(x,y) = \tfrac{\partial}{\partial x} \int_{a=-\infty}^{x} \int_{y=-\infty}^{+\infty} f_{XY}(a,y)dady = \int_{y=-\infty}^{+\infty} f_{XY}(x,y)dy.$$

This integrating-out of a variable

$$f_X(x) = \int_{y=-\infty}^{+\infty} f_{XY}(x,y)dy$$

leads to a **marginal distribution**, i.e. we would say that f_X is a marginal of f_{XY}.

In the special case that X and Y are independent, then

$$F_{XY}(x,y) = F_X(x)F_Y(y), \qquad f_{XY}(x,y) = f_X(x)f_Y(y).$$

Remark: joint absolute continuity is a genuine assumption – it is *not* implied by X and Y separately having pdfs. E.g. for $Y = X$ with X absolutely continuous, the joint law is concentrated on the diagonal of $\mathbb{R}^2$, a set of zero area: no joint pdf exists.

8.3 Operations involving two or more variables

Lemma 8.7 ($Z = X + Y$)**.** *The sum of two continuous R.V. $Z = X + Y$ has pdf*

$$f_Z(t) = \int_{s=-\infty}^{+\infty} f_{XY}(s, t-s)ds = \int_{s=-\infty}^{+\infty} f_{XY}(t-s, s)ds$$

In the special case that X, Y are independent, it follows that

$$f_Z(t) = \int_{s\in\mathbb{R}} f_X(s)f_Y(t-s)ds = f_X * f_Y(t).$$

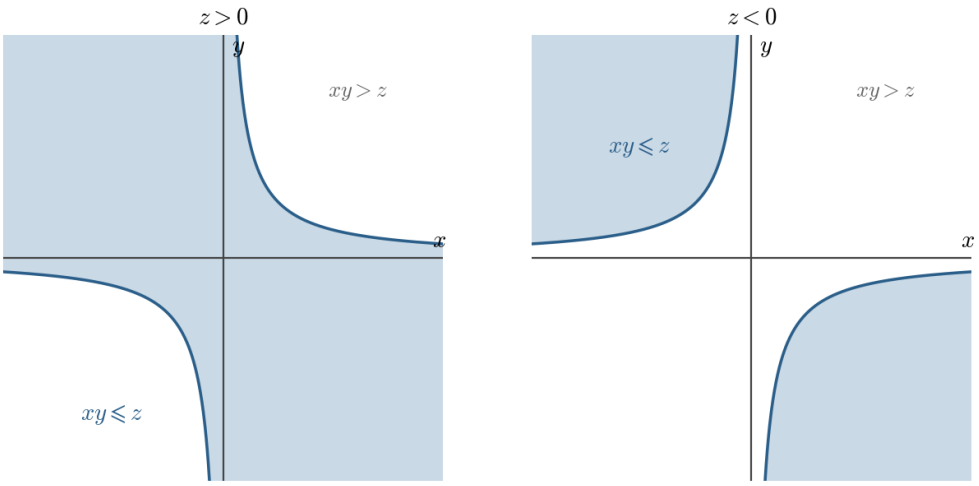


Figure 1: Level sets of $x \cdot y = z$, which can be used to help identify the sets $\{(x, y) : x \cdot y \leqslant z\}$ for $z < 0$, $z > 0$.

Proof: Observe that

$$F_Z(t) = P[X + Y \leqslant t] = \int_{x=-\infty}^{+\infty} \int_{y=-\infty}^{t-x} f_{XY}(x, y)\, dy\, dx$$

and moreover differentiation leads to

$$f_Z(t) = \tfrac{\partial}{\partial t} F_Z(t) = \int_{x=-\infty}^{+\infty} f_{XY}(x, t-x) dx = \int_{s=-\infty}^{+\infty} f_{XY}(t-s, s) ds$$

where in the last step a change of variables was made. □

Lemma 8.8 ($Z = X \cdot Y$)**.** *The product of two continuous R.V. $Z = X \cdot Y$ has cdf (displayed here for $z > 0$; the case $z < 0$ is analogous, see the proof)*

$$F_Z(z) = \int_{x=-\infty}^{0} \int_{y=\frac{z}{x}}^{+\infty} f_{XY}(x, y)\, dy\, dx + \int_{x=0}^{+\infty} \int_{y=-\infty}^{\frac{z}{x}} f_{XY}(x, y)\, dy\, dx$$

and pdf

$$f_Z(z) = \int_{x=-\infty}^{+\infty} \frac{1}{|x|} f_{XY}(x, \frac{z}{x}) dx = \int_{y=-\infty}^{+\infty} \frac{1}{|y|} f_{XY}(\frac{z}{y}, y) dy.$$

Proof: The proof essentially consists on identifying the sets $\{(x, y) : x \cdot y \leqslant z\}$ in the two separate cases of $z < 0$, and $z > 0$ (and technically in the case $z = 0$, although that also follows by continuity from either of the other two cases), see also Figure 1. For $z > 0$,

$$F_Z(z) = P[X \cdot Y \leqslant z] = \int_{x=-\infty}^{0} \int_{y=\frac{z}{x}}^{+\infty} f_{XY}(x, y)\, dy\, dx + \int_{x=0}^{+\infty} \int_{y=-\infty}^{\frac{z}{x}} f_{XY}(x, y)\, dy\, dx.$$

For $z < 0$,

$$F_Z(z) = P[X \cdot Y \leqslant z] = \int_{x=-\infty}^{0} \int_{y=-\infty}^{\frac{z}{x}} f_{XY}(x,y)\, dy\, dx + \int_{x=0}^{+\infty} \int_{y=\frac{z}{x}}^{+\infty} f_{XY}(x,y)\, dy\, dx.$$

For $z = 0$,

$$P[X \cdot Y \leqslant 0] = P[X \geqslant 0 \ \cap \ Y \leqslant 0] + P[X \leqslant 0 \ \cap \ Y \geqslant 0] =$$
$$= \int_{x=0}^{\infty} \int_{y=0}^{\infty} [f_{XY}(x,-y) + f_{XY}(-x,y)]\, dy\, dx.$$

The pdf follows by differentiating once with respect to z (Leibniz' rule; the change of variables in the inner integral is what produces the $\frac{1}{|x|}$ Jacobian factor). □

Lemma 8.9 ($Z = \min\{X, Y\}$)**.** *The minimum of two R.V. $Z = \min\{X, Y\}$ has cdf*

$$F_Z(z) = F_X(z) + F_Y(z) - F_{XY}(z,z)$$

and pdf

$$f_Z(z) = f_X(z) + f_Y(z) - \int_{s=-\infty}^{z} f_{XY}(z,s)ds - \int_{s=-\infty}^{z} f_{XY}(s,z)ds$$

Proof:

$$F_Z(z) = P[X \leqslant z \ \cup \ Y \leqslant z] = P[X \leqslant z \ \cap \ Y \leqslant z] +$$

$$+P[X \geqslant z \ \cap \ Y \leqslant z] + P[X \leqslant z \ \cap \ Y \geqslant z] =$$
$$= F_{XY}(z,z) + \big(F_Y(z) - F_{XY}(z,z)\big) + \big(F_X(z) - F_{XY}(z,z)\big) =$$

$$= F_X(z) + F_Y(z) - F_{XY}(z,z).$$

The same result can also be obtained by using the formula $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ where $A = \{X \leqslant z\}$, $B = \{Y \leqslant z\}$. (In the decomposition above the three events overlap on $\{X = z\} \cup \{Y = z\}$, which is null under our absolute continuity convention; for exact set identities use a strict inequality on one side.)

For the differentiation leading to the pdf one needs to use the auxiliary computation

$$\frac{\partial}{\partial z} F_{XY}(z,z) = \frac{\partial}{\partial x} F_{XY}(x,z)|_{x=z} + \frac{\partial}{\partial y} F_{XY}(z,y)|_{y=z} = \int_{b=-\infty}^{z} f_{XY}(z,b)db + \int_{a=-\infty}^{z} f_{XY}(a,z)da$$

□

(A consistency check – that $\int f_Z\, dz = 1$ – is Exercise 8.4.)

Lemma 8.10 ($Z = \max\{X_1, X_2, \ldots X_N\}$)**.** *The maximum of N R.V. $Z = \max\{X_1, X_2, \ldots X_N\}$ has cdf*

$$F_Z(z) = F_{X_1 X_2 \ldots X_N}(z, z, \ldots z)$$

and pdf

$$f_Z(z) = \int\limits_{\substack{x_j=-\infty \\ j\neq 1}}^{z} f_{X_1X_2\ldots X_N}(z, x_2, \ldots, x_N)dx_2 \ldots dx_N +$$

$$+ \int\limits_{\substack{x_j=-\infty \\ j\neq 2}}^{z} f_{X_1X_2\ldots X_N}(x_1, z, \ldots, x_N)dx_1 \ldots dx_N +$$
$$+ \cdots + \int\limits_{\substack{x_j=-\infty \\ j\neq N}}^{z} f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, z)dx_1dx_2 \ldots dx_{N-1}.$$

In the special case that the X_j are independent and identically distributed, we get

$$F_Z(z) = F_X^N(z), \qquad f_Z(z) = NF_X^{N-1}(z)f_X(z).$$

Proof:

$$F_Z(z) = P[Z \leqslant z] = P[X_1 \leqslant z \cap X_2 \leqslant z \cap \ldots \cap X_N \leqslant z] = F_{X_1X_2\ldots X_N}(z, z, \ldots z).$$

The pdf follows by differentiation, and with a computation similar to that of Lemma 8.9. □

Computing the n^{th} largest (or smallest) out of N random variables $X_1, X_2, \ldots X_N$ is often referred to as **"order statistics"** or **"rank statistics"**. This appears in many practical problems.

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 8.1. *Compute how $Y = aX + b$ is distributed when $a < 0$.*

Exercise 8.2. *Given the continuous R.V. X, compute how the R.V. $Y_1 = e^X$, $Y_2 = (\frac{1}{2})^X$, $Y_3 = \log_{10}(X)$ and $Y_4 = \log_{\frac{1}{2}} X$ are distributed (for Y_3, Y_4 assume $P[X > 0] = 1$).*

Exercise 8.3. *Compute how $Z_1 = \frac{X}{Y}$ and how $Z_2 = |X + Y|$ are distributed.*

Exercise 8.4. *Prove that $\int\limits_{z=-\infty}^{+\infty} f_Z(z)dz = 1$ in Lemma 8.9.*

Exercise 8.5. *Let X, Y be independent exponential random variables with parameters λ, μ. Show that $\min\{X, Y\}$ is exponential with parameter $\lambda + \mu$, and that $P[X < Y] = \frac{\lambda}{\lambda+\mu}$.*

Exercise 8.6. *Let X, Y be independent, uniform on $(0, 1)$. Compute the pdf of $Z = X + Y$.*

Exercise 8.7. *Let $X \sim \mathcal{N}(0, 1)$. Using the lemma for $Y = X^2$, show that*

$$f_{X^2}(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} \chi_{(0,\infty)}(y)$$

(the chi-squared distribution with one degree of freedom).

9 The Central Limit Theorem (CLT)

We saw in Chapter 6 that the sum of many Bernoulli trials converges in an appropriate sense to a gaussian distribution. This kind of result in fact applies in a much more general setting. However, in order to see a proof of that, we first need some basic facts about the **Fourier transform**.

9.1 The Fourier transform (FT)

Definition 9.1 (Fourier transform and Characteristic function)**.** *Let $f(x) : \mathbb{R}^n \to \mathbb{C}$ and denote*

$$\mathcal{F}_{x\to k}[f] = \widehat{f} = \int_x e^{-2\pi i k\cdot x} f(x)dx$$

whenever the integral exists. Then $\widehat{f} : \mathbb{R}^n \to \mathbb{C}$ is called the **Fourier transform (FT)** *of f.*

In probability, the term **characteristic function** *is often used. Given a continuous R.V. X, its characteristic function $\phi_X(k)$ is defined as*

$$\phi_X(k) := \mathrm{E}[e^{-2\pi i k X}] = \mathcal{F}_{x\to k}[f_X(x)] = \widehat{f}_X(k).$$

This is simply another way to talk about the Fourier transform of the pdf $f_X(x)$.

Remark 9.2. *In different communities different normalisations of the FT are used, e.g. $\int_x e^{-ix\cdot k} f(x)dx$. These lead to equivalent theories, with some book-keeping differences (how the 2π show up eventually). We use this normalisation because it leads to a "clean-cut" inverse Fourier transform, cf. Lemma 9.13 below. In general, one has to be aware of the normalisation used in a given context / textbook, but it is merely a technical issue.*

Lemma 9.3 (Elementary operational properties of the Fourier transform)**.** *Let $f : \mathbb{R} \to \mathbb{C}$. Whenever the integrals exist[4], the following properties hold:*

1. $\mathcal{F}_{x\to k}[-2\pi i x f(x)] = \partial_k \widehat{f}(k)$
2. $\mathcal{F}_{x\to k}[\partial_x f(x)] = 2\pi i k \widehat{f}(k)$
3. $\mathcal{F}_{x\to k}[f(x+L)] = e^{2\pi i k L}\widehat{f}(k)$
4. $\mathcal{F}_{x\to k}[e^{-2\pi i x L} f(x)] = \widehat{f}(k+L)$
5. $\mathcal{F}_{x\to k}[f(\lambda x)] = \frac{1}{|\lambda|}\widehat{f}(\frac{k}{\lambda}), \qquad \lambda \neq 0$

Proof: The proof follows directly from the definition. For example, 1 is proved by

$$\mathcal{F}_{x\to k}[-2\pi i x f(x)] = \int_x e^{-2\pi i k x}(-2\pi i) x f(x)dx = \partial_k \int_x e^{-2\pi i k x} f(x)dx = \partial_k \widehat{f}(k)$$

[4] in the Lebesgue sense, i.e. the integrals converge absolutely. For the differentiation rules the honest hypotheses are slightly stronger than convergence of the displayed integrals: differentiating $\widehat{f}$ under the integral sign requires $xf(x) \in L^1$, and the rule for $\widehat{f'}$ requires f absolutely continuous with $f' \in L^1$ (which makes the boundary terms vanish). All the properties hold in particular for Schwartz functions, introduced below – which is where we will actually use them.

and similarly for the other properties.

The interchange of integration and differentiation whenever the integrals converge absolutely follows by the dominated convergence theorem, although that is not the point here.

□

Thus the Fourier transform maps "moments" (expressions of the form $x^m f(x)$) to derivatives on the Fourier side ($\partial_k^m \widehat{f}(k)$) and conversely. This is further quantified through the following

Lemma 9.4 (Riemann-Lebesgue lemma)**.** *For any $f \in L^1(\mathbb{R})$, i.e.*

$$\|f\|_{L^1} := \int_x |f(x)|dx < \infty,$$

it follows that

$$|\widehat{f}(k)| \leqslant \int_x |f(x)|dx, \tag{3}$$

and

$$\lim_{|k|\to\infty} |\widehat{f}(k)| = 0. \tag{4}$$

Moreover, if f together with its derivatives up to order m is absolutely integrable,

$$\int_x |\partial_x^j f(x)|dx < \infty, \qquad j = 0, 1, \ldots, m,$$

then there exists a constant $C > 0$ such that

$$|\widehat{f}(k)| \leqslant C|k|^{-m}. \tag{5}$$

Proof: Equation (3) follows by the elementary observation

$$|\widehat{f}(k)| \leqslant |\int_x e^{-2\pi ikx} f(x)dx| \leqslant \int_x \cancel{|e^{-2\pi ikx}|}^{1} |f(x)|dx = \int_x |f(x)|dx. \tag{6}$$

To proceed, first assume $f : \mathbb{R} \to \mathbb{C}$ is compactly supported and smooth. Then

$$\begin{aligned} \widehat{f}(k) = \int_x e^{-2\pi ikx} f(x)dx = \tfrac{1}{-2\pi ik} \int_x (e^{-2\pi ikx})' f(x)dx = \tfrac{1}{2\pi ik} \int_x e^{-2\pi ikx} f(x)'dx \qquad \Rightarrow \\ \Rightarrow \qquad |\widehat{f}(k)| \leqslant \tfrac{1}{2\pi|k|} \int_x |\partial_x f(x)|dx \end{aligned} \tag{7}$$

In particular, for any smooth function of compact support equation (4) follows.

Now for general f in $L^1(\mathbb{R})$, we can always find[5] a smooth function of compact support f^ε so that

$$\|f - f^\varepsilon\|_{L^1} := \int_x |f(x) - f^\varepsilon(x)|dx \leqslant \varepsilon.$$

[5] A constructive way to do this would be through $f^\varepsilon(x) = \frac{1}{\sqrt{2\pi}\eta} \int_y e^{-\frac{(x-y)^2}{2\eta^2}} f(y)dy \cdot \phi_R(x)$ where $\phi_R(x)$ is the smooth cutoff $\phi_R(x) := \chi_{\{|x|<R\}}(x) e^{-\frac{1}{R^2-|x|^2}}$, and η is sufficiently small, R sufficiently large.

Now applying the previous ideas and the triangle inequality for the L^1 norm we have

$$\limsup_{|k|\to\infty} |\widehat{f}(k)| = \limsup_{|k|\to\infty} |\int_x e^{-2\pi ikx}\big((f(x)-f^\varepsilon(x))+f^\varepsilon(x)\big)dx| \leqslant \|f-f^\varepsilon\|_{L^1} + \lim_{|k|\to\infty} |\widehat{f^\varepsilon}(k)| \leqslant \varepsilon+0.$$

By taking $\varepsilon \to 0$, $\limsup_{|k|\to\infty} |\widehat{f}(k)| = 0$, and since $|\widehat{f}(k)| \geqslant 0$, equation (4) follows. (The lim sup is the honest object here: it always exists, whereas the limit is only known to exist at the end of the argument.)

Finally, for equation (5), the integration by parts of equation (7) can be applied m times directly under the stated hypothesis: an L^1 function whose derivative is also in L^1 is absolutely continuous and tends to 0 at $\pm\infty$, so all boundary terms vanish, and we arrive at $|\widehat{f}(k)| \leqslant (2\pi|k|)^{-m}\|\partial_x^m f\|_{L^1}$. (Note that regularisation alone would *not* prove this: closeness in L^1 gives no control of derivatives.)

□

To proceed to the more powerful properties of the Fourier transform – those that we will use to prove the CLT – we will need to talk about distributions and their Fourier transforms.

Definition 9.5 (Schwartz class of test functions, $\mathcal{S}(\mathbb{R}^n)$)**.** *Consider $f \in C^\infty(\mathbb{R}^n \to \mathbb{C})$. Then f is a Schwartz test function iff*

$$f \in \mathcal{S}(\mathbb{R}^n) \iff \sup_{x\in\mathbb{R}^n} |x^a \partial_x^b f(x)| < \infty \quad \text{for all multiindices } a, b.$$

This means that any expression of the form

$$x_1^{a_1} x_2^{a_2} \dots x_n^{a_n} \partial_{x_1}^{b_1} \partial_{x_2}^{b_2} \dots \partial_{x_n}^{b_n} f(x)$$

is bounded over $\mathbb{R}^n$.

Schwartz test functions are "nice functions". In particular, all the integrals that appeared earlier in this section exist with no problems when $f \in \mathcal{S}(\mathbb{R}^n)$. By definition, if $f \in \mathcal{S}(\mathbb{R}^n)$ then any (finite) number of derivatives and moments of f is still in $\mathcal{S}(\mathbb{R}^n)$.

Now using the properties of Lemma 9.3 one can readily show that

Lemma 9.6 (Schwartz class is closed under Fourier transforms)**.**

$$\mathcal{F}[\mathcal{S}(\mathbb{R}^n)] \subseteq \mathcal{S}(\mathbb{R}^n).$$

These "nice functions" are very useful when we integrate "not so nice" expressions against them. Then they can absorb some of the "nastiness" and allow us to complete operations that otherwise wouldn't make sense. Perhaps the best example of that is weak differentiation.

Example 9.7. *Consider the trivial R.V. $X = 1$ (i.e. it is not really random, it always takes on the value 1). For many reasons, we may want our machinery to be able to handle such cases as well without breaking down (for example we may not know in advance which variables*

really are random). If we treat this X as a random variable, we would say it has cdf

$$F_X(x) = P[X \leqslant x] = \begin{cases} 0, x < 1, \\ 1, x \geqslant 1. \end{cases}$$

Moreover, if we proceeded as we did earlier, we would say that

$$f_X(x) = F'_X(x).$$

Clearly, the classical derivative of this $F_X(x)$ doesn't make sense. However, the distributional, or weak derivative makes sense. The idea is the following: ok, I don't know what $F'_X(x)$ is. However, if it existed, when integrating against a nice function $\phi(x) \in \mathcal{S}(\mathbb{R})$ I would get

$$\int_x F'_X(x)\phi(x)dx = -\int_x F_X(x)\phi'(x)dx = -\int_{x=1}^{x=\infty} \phi'(x)dx = \phi(1) - \cancelto{0}{\phi(\infty)} = \phi(1).$$

Thus we say that

$$F'_X(x) = \delta(x-1)$$

where the "δ function" is defined through

$$\int_x \delta(x-1)\phi(x)dx = \phi(1) \qquad \forall \phi \in \mathcal{S}(\mathbb{R}).$$

We can thus give rigorous meaning to a whole number of formulas and manipulations that we have worked out for smooth objects, by interpreting them "in weak sense". This doesn't mean that "everything makes sense" in weak sense; for a proper study of exactly what works and what not one should study the Theory of Distributions. The idea is captured in the following

Definition 9.8 (Tempered distributions, $\mathcal{S}'(\mathbb{R}^n)$)**.** *The space of tempered distributions is the space of all linear functionals $T : \mathcal{S}(\mathbb{R}^n) \to \mathbb{C}$ which are continuous in the sense that if*

$$\lim_m \sup_{x \in \mathbb{R}^n} |x^a \partial_x^b \phi_m(x)| = 0 \quad \text{for all multiindices } a, b,$$

then

$$\lim_m T(\phi_m) = 0.$$

Example 9.9. *The term "functionals" above essentially is there to remind us that distributions must be integrated against test-functions, i.e. a distribution $T(x)$ is a functional through*

$$T : \mathcal{S}(\mathbb{R}^n) \to \mathbb{C} : \phi(x) \to \int_x T(x)\phi(x)dx.$$

Thus δ-functions are functionals in the sense that

$$\delta(x - x_0) : \mathcal{S}(\mathbb{R}^n) \to \mathbb{C} : \phi(x) \mapsto \phi(x_0).$$

By definition, if $T \in \mathcal{S}'(\mathbb{R}^n)$ then any (finite) number of derivatives and moments of T is still in $\mathcal{S}'(\mathbb{R}^n)$. For example,

$$(x^3 - 3x)\delta'''(x) : \phi(x) \mapsto \int_x (x^3 - 3x)\delta'''(x)\phi(x)dx = (-1)^3 \int_x \delta(x) \left[(x^3 - 3x)\phi(x)\right]'''dx.$$

Definition 9.10 (Fourier transform of distributions)**.** *Let $f \in \mathcal{S}'(\mathbb{R}^n)$. Then*

$$\mathcal{F}[f] : \mathcal{S}(\mathbb{R}^n) \to \mathbb{C} : \phi(x) \to \int_x f(x)\mathcal{F}[\phi](x)dx$$

Remark 9.11. *The definition gives sense to the computation*

$$\int_x \int_s e^{-2\pi ixs} f(s)ds\, \phi(x)dx = \int_s \int_x e^{-2\pi ixs}\phi(x)dx\, f(s)ds = \int_s f(s)\widehat{\phi}(s)ds$$

which is how we would like things to work out. Observe that the last expression makes sense because of Lemma 9.6, i.e. because $\widehat{\phi} \in \mathcal{S}(\mathbb{R}^n)$.

There is one formula that implies most of the interesting properties of the Fourier transform:

Lemma 9.12.

$$\int_k e^{-2\pi i(x-y)k}dk = \delta(x - y).$$

Proof: First of all observe that

$$\int_{k=-M}^{M} e^{-2\pi i(x-y)k}dk = \frac{e^{-2\pi i(x-y)M} - e^{2\pi i(x-y)M}}{-2\pi i(x-y)} = 2M\frac{\sin(2\pi(x-y)M)}{2\pi(x-y)M} = 2M\,\mathrm{sinc}(2\pi(x-y)M).$$

The proof is complete by checking that

$$\lim_{M\to\infty} \int_x 2M\,\mathrm{sinc}(2\pi(x-y)M)\phi(x)dx = \phi(y) \qquad \forall \phi \in \mathcal{S}(\mathbb{R}).$$

□

Lemma 9.13 (Inversion)**.** *Denote*

$$\mathcal{F}^{-1}_{k\to x}[g] = \check{g}(x) = \int_k e^{2\pi ikx}g(k)dk.$$

Then

$$\widehat{\check{f}} = \check{\widehat{f}} = f.$$

Moreover,

$$\widehat{\widehat{f}}(x) = f(-x),$$

and thus $\widehat{\widehat{\widehat{f}}}(x) = f(x)$, or $\mathcal{F}^{-1} = \mathcal{F}^3$ if we look at the operator level.

Proof: The proof rests on Lemma 9.12:

$$\widehat{\check{f}} = \mathcal{F}_{k\to x}[\mathcal{F}^{-1}_{s\to k}[f(s)]] = \int\limits_k e^{-2\pi ikx} \int\limits_s e^{2\pi iks} f(s)ds\, dk =$$

$$= \int\limits_{s,k} e^{-2\pi ik(x-s)} dk f(s)ds = \int\limits_s \delta(x-s)f(s)ds = f(x)$$

The other claims are proved similarly. □

Lemma 9.14 (Convolution)**.** *1.* $\mathcal{F}[fg] = \widehat{f} * \widehat{g} = \int\limits_\lambda \widehat{f}(k-\lambda)\widehat{g}(\lambda)d\lambda$.

2. $\mathcal{F}[f * g] = \widehat{f}(k)\widehat{g}(k)$

3. $\mathcal{F}^{-1}[fg] = \check{f} * \check{g}$

4. $\mathcal{F}^{-1}[f * g] = \check{f}(k)\check{g}(k)$

Proof:

$$\mathcal{F}_{x\to k}[fg] = \int\limits_x e^{-2\pi ikx} f(x)g(x)dx = \int\limits_x e^{-2\pi ikx} \int\limits_s e^{2\pi ixs}\widehat{f}(s)ds \int\limits_t e^{2\pi ixt}\widehat{g}(t)dtdx =$$
$$= \int\limits_{s,t}\int\limits_x e^{-2\pi ix(k-s-t)}dx\, \widehat{f}(s)ds\widehat{g}(t)dt = \int\limits_{s,t} \delta(k-t-s)\widehat{f}(s)ds\widehat{g}(t)dt = \int\limits_t \widehat{f}(k-t)\widehat{g}(t)dt.$$

The other claims can be proved similarly. □

(A closely related identity – Plancherel-Parseval – is Exercise 9.1.)

Finally, we can look at the FT of a pdf:

Lemma 9.15 (Taylor expansion of the characteristic function)**.** *Let X be a continuous R.V. with pdf $X \sim f$, i.e. $f(x) \geqslant 0$, $\|f\|_{L^1} = 1$. Then*

$$\widehat{f}(0) = 1, \qquad |\widehat{f}(k)| \leqslant 1$$

and the Taylor expansion around zero of $\widehat{f}(k)$ is given by

$$\widehat{f}(k) = 1 - 2\pi ik\,\mathrm{E}[X] - 2\pi^2k^2\,\mathrm{E}[X^2] + \mathcal{R}, \qquad |\mathcal{R}| \leqslant \frac{|2\pi k|^3}{6}\,\mathrm{E}[|X|^3].$$

Proof: By definition,

$$\widehat{f}(0) = \int\limits_x e^{-2\pi ikx} f(x)dx|_{k=0} = \int\limits_x f(x)dx = 1$$

and

$$|\widehat{f}(k)| \leqslant \int\limits_x |e^{-2\pi ikx} f(x)|dx = \int\limits_x |f(x)|dx = 1.$$

Moreover, Taylor expansion with the *integral form* of the remainder – which, unlike the mean-value (Lagrange) form, remains valid verbatim for complex-valued functions of a real

variable[6] – yields

$$\widehat{f}(k) = \widehat{f}(0) + k\partial_k \widehat{f}(0) + \frac{k^2}{2}\partial_k^2 \widehat{f}(0) + \mathcal{R}, \qquad |\mathcal{R}| \leqslant \frac{|k|^3}{6} \max_{|\xi|\leqslant|k|} \left|\partial_k^3 \widehat{f}(\xi)\right|.$$

Now using the previous computation and the properties of Lemma 9.3 we have

$$\widehat{f}(0) = 1, \qquad \partial_k \widehat{f}(0) = (-2\pi i)\,\mathrm{E}[X], \qquad \partial_k^2 \widehat{f}(0) = (-2\pi i)^2\,\mathrm{E}[X^2],$$

$$|\partial_k^3 \widehat{f}(\xi)| = |2\pi|^3 \left| \int_x e^{-2\pi i x\xi} x^3 f(x) dx \right| \leqslant |2\pi|^3\,\mathrm{E}[|X|^3].$$

The proof is complete. □

Remark 9.16. *What is usually called Bochner's Theorem applies to probability measures (i.e. not only absolutely continuous probability measures, but measures with singular parts as well), and states that their FT is a function of positive type. Since we won't use that here, it is not stated in full detail.*

9.2 The CLT and a simple proof

We first present the heuristic version, whose proof exposes the mechanism; the standard theorem, with a complete proof, follows right after.

Theorem 9.17 (CLT, heuristic version)**.** *Let X_n be an i.i.d. sequence, distributed according to $f_X(x)$, for which $\mathrm{E}[|X|^j]$ exist for $j \in \{1,2,3\}$. Then the average of N of the X_n is distributed, for large N, approximately as*

$$\frac{X_1 + X_2 + \cdots + X_N}{N} \sim f_Y(y) \quad \approx \quad \frac{\sqrt{N}}{\sqrt{2\pi\,\mathrm{Var}[X]}}\, e^{-\dfrac{N(y-\mathrm{E}[X])^2}{2\,\mathrm{Var}[X]}}$$

in the sense (made precise in the proof) that the characteristic function of Y is that of the Gaussian above, times a factor which tends to 1 uniformly on compact sets of k as $N \to \infty$.

Proof: Denote

$$S = X_1 + X_2 + \cdots + X_N, \qquad Y = \frac{1}{N} S.$$

Then by virtue of Lemmata 8.7 and 8.1, Y is distributed with

$$Y \sim f_Y(y) = N \underbrace{f_X * f_X * \cdots * f_X}_{N \text{ times}}(Ny)$$

and therefore, by virtue of Lemma 9.14 and Property 5 of Lemma 9.3, its characteristic function is

$$\phi_Y(k) = \widehat{f}_Y(k) = \left(\widehat{f}_X(\frac{k}{N})\right)^N = e^{N \log(\widehat{f}_X(\frac{k}{N}))}$$

[6] The mean-value theorem genuinely fails for complex-valued functions: there is no ξ with $e^{2\pi i} - e^0 = 2\pi i e^{i\xi} \cdot 1$. The integral remainder $\mathcal{R} = \frac{1}{2}\int_0^k (k-t)^2 \partial_k^3 \widehat{f}(t)\,dt$ needs no such ξ.

Now we can apply the Taylor expansion of Lemma 9.15 to obtain

$$\widehat{f}_X(\frac{k}{N}) = 1 - 2\pi i \frac{k}{N}\,\mathrm{E}[X] - 2\pi^2 \frac{k^2}{N^2}\,\mathrm{E}[X^2] + \mathcal{R}_1, \qquad |\mathcal{R}_1| \leqslant \frac{|2\pi k|^3}{6N^3}\,\mathrm{E}[|X|^3]$$

and thus

$$\widehat{f}_Y(k) = e^{N \log\left(1 - 2\pi i \frac{k}{N}\,\mathrm{E}[X] - 2\pi^2 \frac{k^2}{N^2}\,\mathrm{E}[X^2] + \mathcal{R}_1\right)}.$$

Moreover, for complex x with $|x| \leqslant \frac{1}{2}$ – true here for N large enough, uniformly for k in a compact set – the power series of the (principal branch) logarithm gives directly

$$\log(1+x) = x - \frac{1}{2}x^2 + \mathcal{R}_2, \qquad |\mathcal{R}_2| = \Big|\sum_{n\geqslant 3} \frac{(-1)^{n-1}x^n}{n}\Big| \leqslant \frac{|x|^3}{3}\big(1 + |x| + |x|^2 + \dots\big) \leqslant \frac{2\,|x|^3}{3}$$

(a real Lagrange remainder would not be legitimate here: x has nonzero imaginary part). So now we have $-2\pi i \frac{k}{N}\,\mathrm{E}[X] - 2\pi^2 \frac{k^2}{N^2}\,\mathrm{E}[X^2] + \mathcal{R}_1$ playing the role of x inside the logarithm, leading to

$$\widehat{f}_Y(k) = e^{N\left[-2\pi i \frac{k}{N}\,\mathrm{E}[X] - 2\pi^2 \frac{k^2}{N^2}\,\mathrm{E}[X^2] + 2\pi^2 \frac{k^2}{N^2}(\mathrm{E}[X])^2 + \mathcal{R}\right]} = e^{-2\pi i k\,\mathrm{E}[X] - 2\pi^2 \frac{k^2}{N}\,\mathrm{E}[X^2] + 2\pi^2 \frac{k^2}{N}(\mathrm{E}[X])^2 + N\mathcal{R}} =$$

$$= \underbrace{e^{-2\pi i k\,\mathrm{E}[X]}}_{\text{modulation}}\ \underbrace{e^{-2\pi^2 \frac{k^2}{N}\,\mathrm{Var}[X]}}_{\text{Gaussian}}\ \underbrace{e^{N\mathcal{R}}}_{\text{tends to } 1} = \mathcal{F}_{x\to k}\left[\frac{\sqrt{N}}{\sqrt{2\pi\,\mathrm{Var}[X]}}\, e^{-\dfrac{N(x-\mathrm{E}[X])^2}{2\,\mathrm{Var}[X]}}\right] e^{N\mathcal{R}}$$

where in $\mathcal{R}$ we collected everything of $O(\frac{k^3}{N^3})$.

The convergence $e^{N\mathcal{R}} \approx e^{C\frac{k^3}{N^2}} \to 1$ as $N \to \infty$ is uniform on every set $k \in [-M, M]$. □

To convert statements about characteristic functions into statements about distributions – which is what the calculation above is missing – one needs the following classical result, which we state without proof (see [11] or [1]):

Theorem 9.18 (Lévy's continuity theorem)**.** *Let X_n, X be random variables with characteristic functions ϕ_n, ϕ. If $\phi_n(k) \to \phi(k)$ for every k, then $X_n \to X$ in distribution. (Conversely, convergence in distribution implies pointwise convergence of the characteristic functions.)*

The statement above was the one closest to the proof. Using the notation $\mathcal{N}(\mu, \sigma^2)$ for the normal distribution with mean μ and variance σ^2, other versions of the CLT can be more convenient to use in different contexts:

Theorem 9.19 (CLT)**.** *Let X_n be an i.i.d. sequence, distributed according to $f_X(x)$, for which $E[|X|^j]$ exist for $j \in \{1,2,3\}$, with $\mu := \mathrm{E}[X]$ and $\sigma^2 := \mathrm{Var}[X] > 0$. Then*

$$Z_N := \sqrt{N}\left(\frac{X_1 + X_2 + \dots + X_N}{N} - \mu\right) \longrightarrow \mathcal{N}(0, \sigma^2) \qquad \textit{in distribution,}$$

i.e.

$$\lim_{N\to\infty} P\left[a \leqslant \frac{X_1 + X_2 + \dots + X_N - N\mu}{\sqrt{N\sigma^2}} \leqslant b\right] = \Phi(b) - \Phi(a),$$

where $\Phi(x) = \int_{z=-\infty}^{x} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz$ *is the cdf for the standardised normal RV,* $\mathcal{N}(0,1)$.

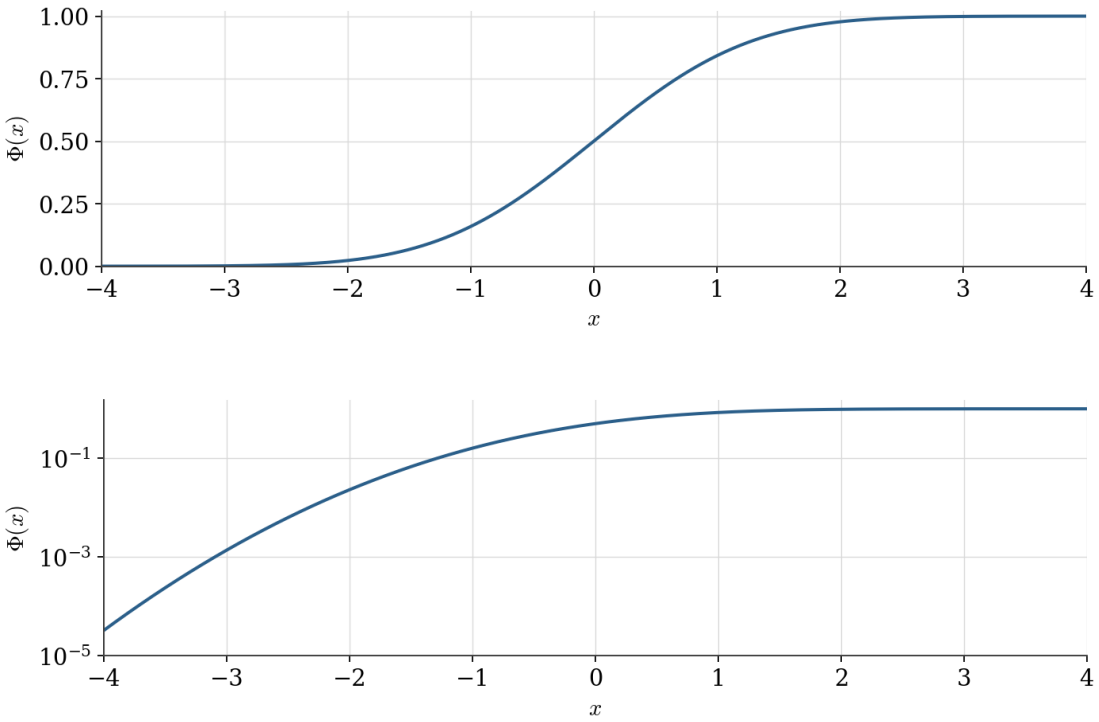
(A remark on the hypotheses: the classical CLT requires only $\mathrm{Var}[X] < \infty$; the finite third moment assumed here simply buys the short proof below, and should not be memorised as necessary.)

Proof. Repeat the computation of the previous proof for the standardised variable: by independence,

$$\phi_{Z_N}(k) = \left[\phi_{X-\mu}\left(\frac{k}{\sqrt{N}}\right)\right]^N = \left[1 - 2\pi^2\sigma^2\frac{k^2}{N} + O(N^{-3/2})\right]^N \longrightarrow e^{-2\pi^2\sigma^2k^2},$$

using the Taylor expansion of Lemma 9.15 for the centered variable $X - \mu$ (uniformly for k in compact sets). The limit is the characteristic function of $\mathcal{N}(0, \sigma^2)$, and Lévy's continuity theorem (Theorem 9.18) concludes the proof. □

Remark 9.20. *Informally we write* $Z_N \sim \mathcal{N}(0, \sigma^2)$ *"for large* N*", and even* $f_{Z_N}(x) \approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$*; note that this latter,* local *statement about densities is genuinely stronger than convergence in distribution and requires additional hypotheses – the De Moivre-Laplace Theorem 6.4 is exactly such a local limit theorem for the binomial case.*



One quickly checks that, for large N,

$$P[A \leqslant X_1 + X_2 + \cdots + X_N - N\,\mathrm{E}[X] \leqslant B] \approx \Phi\left(\frac{B}{\sqrt{N\,\mathrm{Var}[X]}}\right) - \Phi\left(\frac{A}{\sqrt{N\,\mathrm{Var}[X]}}\right)$$

or, equivalently,

$$P[A \leqslant \frac{X_1 + X_2 + \cdots + X_N}{N} - \mathrm{E}[X] \leqslant B] \approx \Phi\left(\frac{B\sqrt{N}}{\sqrt{\mathrm{Var}[X]}}\right) - \Phi\left(\frac{A\sqrt{N}}{\sqrt{\mathrm{Var}[X]}}\right).$$

Noting in addition that, due to the symmetry of Φ we have

$$\Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1,$$

this implies that, for large N,

$$P[|X_1 + X_2 + \cdots + X_N - N\,\mathrm{E}[X]| \leqslant \delta] \approx 2\Phi\left(\frac{\delta}{\sqrt{N\,\mathrm{Var}[X]}}\right) - 1$$

or, equivalently,

$$P[\left|\frac{X_1 + X_2 + \cdots + X_N}{N} - \mathrm{E}[X]\right| \leqslant \delta] \approx 2\Phi\left(\frac{\delta\sqrt{N}}{\sqrt{\mathrm{Var}[X]}}\right) - 1.$$

This can be compared with the LLN, Corollary 7.17, which can be recast as

Theorem 9.21 (Law of large numbers)**.** *Assume $\{X_n\}_{n\in\mathbb{N}}$ be independent identically distributed (i.i.d.) continuous R.V. with finite mean $\mu = \mathrm{E}[X]$ and variance $\sigma^2 = \mathrm{Var}[X]$. Then, denoting*

$$\bar{X}_N := \frac{X_1 + X_2 + \cdots + X_N}{N},$$

we have

$$P\left[\left|\bar{X}_N - \mu\right| > \delta\right] \leqslant \frac{\sigma^2}{N\delta^2}$$

Example 9.22. *A local insurance company is trying to set the yearly premium $P = 1000 + \delta$ as low as possible, provided that the following constraint is satisfied:*

$$P[\textit{total claims } > \textit{ total premium payments}] \leqslant 1\%.$$

Historically, the average claim per customer per year μ is 1000 pounds, with a variance σ^2 of 27000. Assume that this year it has $N = 3000$ customers, and this number is used as a reference figure for next year too.

If we use the CLT, we can say that we are asking that

$$P[\sum_{i=1}^{N} X_i - N \cdot 1000 \geqslant N\delta] \approx 1 - \Phi(\frac{N\delta}{\sqrt{N\,\mathrm{Var}[X]}}) = 1 - \Phi(\delta\sqrt{\frac{3000}{27000}}) \leqslant 1\%.$$

So, enforcing the 1% significance level, we get

$$1 - \Phi(\delta\sqrt{\frac{3000}{27000}}) \leqslant 0.01 \implies 0.99 \leqslant \Phi(\delta \cdot 0.33) \implies \delta \cdot 0.33 \geqslant \Phi^{-1}(0.99) = 2.33 \implies \delta \geqslant 7.$$

So, using the CLT, a premium of 1007 pounds would suffice.

In comparison, let us try to solve the problem using the LLN. A crude estimate would be

$$P[\left|\sum_{i=1}^{N} X_i - N \cdot 1000\right| \geqslant N\delta]$$

in the place of the probability of going in the red. In that case we get a failure probability of

$$P[\left|\frac{\sum_{i=1}^{N} X_i}{N} - 1000\right| \geqslant \delta] \leqslant \frac{27000}{3000 \cdot \delta^2}.$$

Enforcing the 1% *bound, we get*

$$\frac{9}{\delta^2} \leqslant 0.01 \implies \delta^2 \geqslant 900 \implies \delta \geqslant 30.$$

(Could you try to improve the latter solution in the context of the LLN? What if the histogram of claim payouts was known to be pretty symmetric around its mean?)

Other versions of the CLT are also of interest. One is the so-called Lyapunov CLT, which allows for the sum of many different - but independent - RVs.

Theorem 9.23. *Let X_i be a sequence of independent RVs, for which there exist $M, m > 0$ such that*

$$\mathrm{E}[|X_i|^n] < M \qquad \forall i \in \mathbb{N}, \quad n \in \{1,2,3\}, \qquad \textit{and} \qquad \mathrm{Var}[X_i] \geqslant m \quad \forall i \in \mathbb{N}.$$

Denote

$$B_n := \sqrt{\sum_{i=1}^{n} \mathrm{Var}[X_i]}.$$

Then the RV

$$Z_n := \frac{\sum_{i=1}^{n} (X_i - \mathrm{E}[X_i])}{B_n} \to \mathcal{N}(0,1)$$

converges in distribution to the standard normal RV.

9.3 Applications of the CLT: The Monte Carlo method

The Monte Carlo method uses repeated sampling of a random variable to approximate a quantity that may not be accessible otherwise. Assume for example that we try to compute π, which can be expressed as an integral

$$\pi = \int_{x=0}^{1} 4\sqrt{1-x^2}dx$$

(of course we have other ways to actually approximate π, this is just a simple example for the sake of illustration). This integral can be reinterpreted as $\mathrm{E}[X]$ where $X = 4\sqrt{1-U^2}$

and $U \sim \mathcal{U}(0,1)$ i.e. $f_U(u) = \chi_{[0,1]}(u)$. The CLT then guarantees that if we can generate N independent realisations of X, their mean will be very sharply concentrated around π for $N \to \infty$. But can we quantify that convergence in more detail? To that end we will need some additional ideas.

Definition 9.24. *Given R.V. $X_1, \ldots, X_N$, the* **sample mean** *is defined as*

$$\overline{X}_N := \frac{X_1 + \cdots + X_N}{N}$$

and the **sample variance** *as*

$$s_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X}_N)^2.$$

If the X_j are i.i.d. copies of a R.V. X with $\mathrm{Var}[X] < \infty$ – the setting of interest here, and the hypothesis under which the following statements hold – it can be easily seen that $\mathrm{E}[\overline{X}_N] = \mathrm{E}[X]$, $\mathrm{E}[s_N^2] = \mathrm{Var}[X]$ and (assuming in addition $\mathrm{E}[X^4] < \infty$ for the statement about s_N^2)

$$\lim_{N\to\infty} \left(\mathrm{Var}[\overline{X}_N] + \mathrm{Var}[s_N^2]\right) = 0.$$

The point is that when we have a dataset that is reasonable to see as independent realisations of the same – *possibly unknown* – RV, we can use the sample mean and sample variance in the place of the mean and variance. In that case the following variant of the CLT is useful:

Theorem 9.25. *Let X_j, $j \in \{1, 2, \ldots, N\}$ be a sequence of iid RV with finite moments up to third order. Then*

$$\lim_{N\to\infty} P\left[a \leqslant \frac{\overline{X}_N - \mathrm{E}[X]}{s_N}\sqrt{N} \leqslant b\right] = \Phi(b) - \Phi(a).$$

So now we can go back to our example of approximating an integral, and generate a **confidence interval**: assume we have a large number of independent realisations of $X = 4\sqrt{1-U^2}$ as discussed above. What is the probability that the observed mean of these realisations is within ε of π?

$$P[\overline{X}_N - \varepsilon \leqslant \pi \leqslant \overline{X}_N + \varepsilon] = P[\pi - \varepsilon \leqslant \overline{X}_N \leqslant \pi + \varepsilon] = P\left[-\tfrac{\varepsilon\sqrt{N}}{s_N} \leqslant \tfrac{\overline{X}_N - \pi}{s_N}\sqrt{N} \leqslant \tfrac{\varepsilon\sqrt{N}}{s_N}\right] \approx$$
$$\approx \Phi(\tfrac{\varepsilon\sqrt{N}}{s_N}) - \Phi(-\tfrac{\varepsilon\sqrt{N}}{s_N}) = 2\Phi(\tfrac{\varepsilon\sqrt{N}}{s_N}) - 1.$$

Thus if we decide on a confidence level, e.g. 95%, and have at hand a given set of realisations (i.e. N, s_N^2 are known) we can determine the size ε of the confidence interval. In this case

$$0.95 \approx 2\Phi(\frac{\varepsilon\sqrt{N}}{s_N}) - 1 \implies \Phi(\frac{\varepsilon\sqrt{N}}{s_N}) \approx \frac{1.95}{2} \implies \varepsilon = \frac{s_N}{\sqrt{N}}\Phi^{-1}(\frac{1.95}{2}) = \frac{s_N}{\sqrt{N}}1.96.$$

This computation shows how the confidence level plays out through Φ^{-1}; more to the point, it shows the "order of convergence in N": if, for a fixed confidence level, we want to halve the size of the confidence interval, we will need 4 times as many points. This amounts to an

order of convergence of $1/2$, since $\varepsilon = O((\frac{1}{N})^{1/2})$, which is very poor e.g. for approximating one-dimensional integrals of smooth functions. The decisive advantage of Monte Carlo is elsewhere: the $O(N^{-1/2})$ statistical rate is essentially *independent of the dimension* of the integral, while deterministic quadrature deteriorates rapidly with dimension – so Monte Carlo becomes competitive, and eventually indispensable, for high-dimensional integration. Secondary advantages include robustness under very poor smoothness of the integrand (e.g. functions no better than Hölder continuous with exponents $a < 1$), and the existence of variance-reduction and quasi-Monte Carlo techniques that improve the practical picture further.

Further reading: Reed and Simon [9] contains a no-nonsense, rigorous yet accessible introduction to the Fourier transform of about twenty pages; Stein and Shakarchi [10] is a full course on Fourier analysis. Gnedenko [1] devotes several chapters to the Central Limit Theorem and its generalisations.

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 9.1 (Plancherel-Parseval)**.** *Prove that, for any $f, g \in \mathcal{S}(\mathbb{R}^n)$,*

$$\int_x f(x)\overline{g(x)}dx = \int_k \widehat{f}(k)\overline{\widehat{g}(k)}dk.$$

Exercise 9.2 (The Gaussian is self-reproducing under FT)**.** *Show that the pdf of $\mathcal{N}(0, \sigma^2)$ has Fourier transform $e^{-2\pi^2\sigma^2k^2}$.*

(Hint: differentiate the FT with respect to k and integrate by parts to derive a first-order ODE for it.)

Exercise 9.3. *Using Fourier transforms and the convolution rule, show that the sum of independent $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ random variables is $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Exercise 9.4. *In a Monte Carlo computation the sample standard deviation stabilises at $s_N \approx 0.9$. Roughly how many samples are needed to obtain a 95% confidence interval of half-width 10^{-3}? What does this say about Monte Carlo as a numerical integration method?*

10 Markov Chains

Here we will start looking at the simplest examples of Markov processes, i.e. stochastic processes that have the Markov property. For now, we can describe the **Markov property** in a limited context as follows: imagine we have a chronological sequence of random variables so that, for any time t_0, knowing the present random variable $t = t_0$ gives all possible information about the future $t > t_0$, that is knowing what happened before t_0 does not add any information. We will see more precise formulations below.

10.1 Markov Chains with a Finite Number of States

Before we talk about Markov chains themselves, we need to talk about the objects that will be used to define them more precisely.

Definition 10.1 (Stochastic matrix)**.** *A matrix $A \in \mathbb{R}^{r\times r}$ is called a row-stochastic matrix iff*

$$a_{ij} \geqslant 0 \quad \text{and} \quad \sum_{j=1}^{r} a_{ij} = 1 \quad \forall i \in \{1, 2, \ldots, r\}.$$

For brevity we will call row-stochastic matrices just stochastic matrices from now on.

Definition 10.2 (Positive cone and pdf's)**.** *The positive cone in $\mathbb{R}^r$ is the set*

$$\Pi^+ := \{(x_1, \ldots, x_r) \in \mathbb{R}^r \mid x_j > 0 \ \forall j\}$$

We will write $\mathbb{R}^r \ni \vec{x} > 0$ iff $\vec{x} \in \Pi^+$; similarly, $\vec{x} \geqslant 0$ iff $x_j \geqslant 0 \ \forall j$.

Moreover, we will say $\mu \in \mathbb{R}^r$ is a pdf if $\mu_j \geqslant 0$, $\sum_{j=1}^{r} \mu_j = 1$. That is, we could think of it as a discrete pdf on the outcomes $\Omega = \{1, 2, \ldots, r\}$.

Lemma 10.3 (Equivalent definitions of stochastic matrices)**.** *The following are equivalent:*

1. *A is a stochastic matrix.*
2. *For any vector $\mathbb{R}^r \ni \vec{x} \geqslant 0$, $\vec{x}A \geqslant 0$ and moreover for $\vec{1} := (1, 1, \ldots, 1)$, $A\vec{1} = \vec{1}$*
3. *If $\mu \in \mathbb{R}^r$ is a pdf then μA is also a pdf.*

Lemma 10.4 (Products of stochastic matrices)**.** *Let A, B be stochastic matrices. Then*

$$C := AB$$

is also a stochastic matrix.

Definition 10.5 (Markov Chain)**.** *The Markov Chain generated by the state space $X = \{1, 2, \ldots, r\}$, the initial pdf $\mu_0 \in \mathbb{R}^r$ and the stochastic matrices $A(1) := [a_{ij}(1)]$, $A(2) := [a_{ij}(2)]$, $\ldots$ $A(N) := [a_{ij}(N)]$ is the probability measure on* **sequences of states**

$$\Omega := X^{N+1}$$

given by

$$P[(\omega_0, \omega_1, \ldots, \omega_N) = (i_0, i_1, \ldots, i_N)] = (\mu_0)_{i_0}\, a_{i_0 i_1}(1) a_{i_1 i_2}(2) \ldots a_{i_{N-1} i_N}(N).$$

Furthermore, a Markov Chain is called **homogeneous** *if* $A(1) = A(2) = \cdots = A(N) = A$.

Remark 10.6. *1. The intuition behind this definition is that we have a "system" with r possible "states". The states are randomly initialised by the initial pdf μ_0, and evolve in time according to the given stochastic matrices. In that context, the random variable ω_n is the state of the system at the n^{th} time step.*

2. The stochastic matrices $A(n)$ are also called **transition matrices**, *since*

$$a_{ij}(n) = P[\omega_n = j \,|\, \omega_{n-1} = i], \qquad n \geqslant 1.$$

In words, $A(n)$ collects the probabilities of the transitions into *the n^{th} time step: the state becomes j at step n if at step $n-1$ it was i. (Note that this is the indexing consistent with the path-probability formula of the definition, where $A(n)$ multiplies the transition $\omega_{n-1} \to \omega_n$.)*

3. The **Markov property** *is built into this construction, in the sense that by definition*

$$P[\omega_{n+1} = j \,|\, \omega_n = i_n, \omega_{n-1} = i_{n-1}, \ldots \omega_0 = i_0] = P[\omega_{n+1} = j \,|\, \omega_n = i_n].$$

4. A Markov Chain can be thought of as a generalisation of a sequence of independent Bernoulli *trials. Indeed, now we have a finite number of possible states (as opposed to only two), and a special kind of dependence is allowed (cf. above), which is still much weaker than general dependence of all times to all times.*

Observe that a sequence of independent Bernoulli trials corresponds to the trivial homogeneous Markov Chain with

$$X = \{0, 1\}, \qquad \mu_0 = (1-p, p), \qquad A = \begin{bmatrix} 1-p & p \\ 1-p & p \end{bmatrix}.$$

5. A homogeneous Markov chain can be represented by a **directed graph**, *where the vertices are the states and the (weighted) directed edges represent the transition probability from one vertex to another (there should be no directed edge when that transition probability is zero). Then the Markov Chain is a probability distribution on the paths of length n on the graph.*

6. A direct consequence of the definition is that, if

$$p_j(n) = P[\omega_n = j],$$

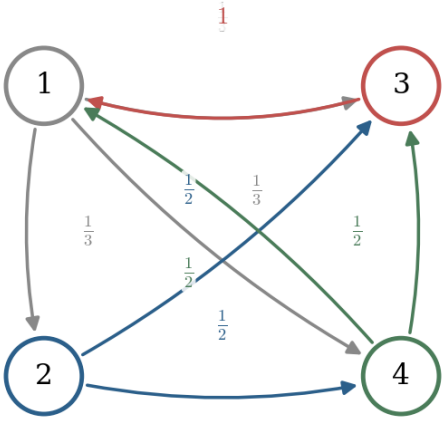
then, for a homogeneous Markov process with transition matrix $A = [a_{ij}]$

$$p_j(n+1) = \sum_i a_{ij} p_i(n).$$

This is often called the **Markov equation** *and it shows how evolving in time probabilities amounts to taking matrix products. (Observe that a straightforward modification is possible when the chain is not homogeneous as well.)*

The proof is just the total probability formula (cf. Lemma 4.3),

$$\begin{aligned} p_j(n+1) = P[\omega_{n+1} = j] = P[\{\omega_{n+1} = j\} \cap (\cup_i \{\omega_n = i\})] = \\ = \sum_i P[\omega_{n+1} = j \,|\, \omega_n = i] P[\omega_n = i] = \sum_i a_{ij} p_i(n). \end{aligned}$$



10.2 Random walks

Definition 10.7 (Random walk)**.** *A random walk is a homogeneous Markov chain with state space* $X = \mathbb{Z}^d$ *(the definition of a Markov chain extends verbatim to countably infinite state spaces) which is moreover* spatially homogeneous*: the transition probabilities depend only on the displacement,*

$$P[\omega_{n+1} = j | \omega_n = i] = P[\omega_{n+1} = j - i | \omega_n = 0] = p_{j-i} \qquad \forall\, n, i, j.$$

Definition 10.8 (Simple symmetric random walk)**.** *A spatially homogeneous random walk is called simple symmetric if*

$$p_z = \begin{cases} \frac{1}{2d}, & z \in \{\pm\vec{e}_1, \pm\vec{e}_2, \ldots, \pm\vec{e}_d\} \\ 0, & \textit{otherwise} \end{cases}$$

Simple symmetric random walks are very interesting and their properties hold clues for more complicated problems. A very important qualitative feature is whether "they stay close to their starting point" or "drift away" with high probability. A first way to quantify this is

as follows:

Definition 10.9 (Recurrent and transient random walks)**.** *Consider a random walk, and denote A_n the set of all trajectories starting from $\vec{0}$ and returning for the first time to $\vec{0}$ in exactly n steps. Denote $f_n = \sum_{\omega \in A_n} P[\omega]$. The random walk is called* **recurrent** *if*

$$\sum_{n=1}^{\infty} f_n = 1$$

and **transient** *if*

$$\sum_{n=1}^{\infty} f_n < 1.$$

The point of the definition is that, at any moment in time, with probability 1 a recurrent random walk will come back to its starting point in a finite number of steps.

Theorem 10.10 (Polya)**.** *The simple symmetric random walk is recurrent for $d = 1$ and $d = 2$, and transient for $d \geqslant 3$.*

For proofs see Norris [15]; a particularly readable account is given in Privault [16].

Let us now analyse more systematically a $1-$dimensional random walk:

Example 10.11 (Gambler's ruin)**.** *Consider the random walk on $\mathbb{Z}$ with transition probabilities*

$$p_1 = p, \qquad p_{-1} = q = 1 - p, \qquad p_z = 0 \;\; z \notin \{-1, 1\},$$

and starting point $i > 0$.

We can think of this as follows: a gambler bets \$1 in each round, and wins \$1 with probability p, or loses his bet with probability q. Moreover, the gambler announces that if he gets to have $N > i$ dollars, he will consider that he has "won", and will stop betting.

In this setting, the set of outcomes Ω contains sequences of states $(\omega_0, \omega_1, \dots, \omega_n, \dots)$ between 0 and N; moreover only the last states can be 0 and N. This includes sequences of different lengths, e.g. if $i = 2$, $N = 5$

$$(2, 1, 0), \qquad (2, 3, 2, 3, 2, 3, 4, 5)$$

and also infinite sequences, e.g.

$$(2, 3, 2, 3, 2, 3, 2, 3, 2, 3, \dots).$$

Probabilities of finite *path events ("cylinders": the first n steps take specified values) are computed as products of the transition probabilities p, q; these determine the probability measure on the space of full trajectories. (Note that an individual infinite trajectory typically carries probability zero – it is the cylinder events that carry the measure.) The initialisation at i dollars means that the initial pdf is $\mu = \vec{e}_i$.*

Question: What is the probability that the gambler will win (i.e. reach $N > i$ before reaching 0)?

Solution: Denote $P(i) = P(i;N)$ the probability that the gambler wins (reaches N dollars) having started with i dollars. (For now we will suppress the explicit dependence on N.) This can happen in two ways: either winning the first round, and proceeding to win from $i+1$ dollars, or losing the first round and proceeding to win from $i-1$ dollars, i.e.

$$P(i) = pP(i+1) + qP(i-1) \iff (p+q)P(i) = pP(i+1) + qP(i-1) \iff$$
$$\iff P(i+1) - P(i) = \tfrac{q}{p}(P(i) - P(i-1)) \iff P(i+1) - P(i) = (\tfrac{q}{p})^i(P(1) - P(0))$$

Of course $P(0) = 0$, $P(N) = 1$ will serve as our boundary conditions, leading to

$$P(i+1) - P(i) = (\frac{q}{p})^i P(1)$$

and moreover by telescopic summation

$$P(i+1) = \big(P(i+1) - P(i)\big) \; + \; \big(P(i) - P(i-1)\big) \; + \; \ldots \; + \; \big(P(2) - P(1)\big) \; + \; P(1) =$$
$$= P(1) \sum_{j=0}^{i} \tfrac{q^j}{p^j}$$

which adds up to

$$P(i+1) = P(1)\frac{1 - (\frac{q}{p})^{i+1}}{1 - \frac{q}{p}}, \qquad \text{if } q \neq p,$$

and

$$P(i+1) = P(1) \cdot (i+1), \qquad \text{if } q = p = \frac{1}{2}.$$

Moreover $P(N) = 1$ normalises the value of $P(1)$ in each case,

$$P(N) = 1 \implies NP(1) = 1 \implies P(1) = \frac{1}{N}, \qquad \text{if } q = p = \frac{1}{2},$$

and

$$P(N) = 1 \implies P(1)\frac{1 - (\frac{q}{p})^N}{1 - \frac{q}{p}} = 1 \implies P(1) = \frac{1 - \frac{q}{p}}{1 - (\frac{q}{p})^N}, \qquad \text{if } q \neq p,$$

leading finally to the answer

$$P(i) = \begin{cases} \dfrac{1 - (\frac{q}{p})^i}{1 - (\frac{q}{p})^N}, & \text{if } q \neq p, \\ \\ \dfrac{i}{N}, & \text{if } q = p = \frac{1}{2}. \end{cases}$$

Question: What is the probability that the gambler will not go bankrupt if he plays "for ever"?

Solution: We can answer this by taking the limit $N \to \infty$ in our previous computation. If $p > q$, then we find that the probability of winning becomes

$$\lim_{N \to \infty} P(i;N) = 1 - (\frac{q}{p})^i > 0.$$

If $p \leqslant q$,

$$\lim_{N\to\infty} P(i;N) = 0.$$

Question: What is the expected length of the game (i.e. the number of rounds until the gambler either wins or goes bankrupt) if initially the gambler has \$$i$, for $p = q = \frac{1}{2}$?

Solution: We are called to compute the expectation of the random variable L_i*, which is the length of the game given that we started at* $\omega_0 = i$*. To do this we will have to compute the pdf for* L_i*; to that end we will first need to introduce an auxiliary object, namely*

$$R(i,n) = R(i,n;N) = P[0 < \omega_m < N \ \forall 0 \leqslant m \leqslant n \mid \omega_0 = i],$$

the probability that at timestep n *the gambler is still in the game having started from* i*. Obviously* $R(i,n)$ *for extreme values should satisfy*

$$R(i,0) = 1 \ \forall i \in \{1,2,\dots,N-1\}, \qquad R(0,n) = R(N,n) = 0 \ \forall n.$$

Moreover, for any $0 < i < N$

$$R(i,n) = P[0 < \omega_m < N \ \forall 0 \leqslant m \leqslant n \mid \omega_0 = i] = P[0 < \omega_m < N \ \forall 1 \leqslant m \leqslant n \mid \omega_0 = i] =$$
(*the constraint at* $m = 0$ *is automatic given* $\omega_0 = i$*; the event now involves only times* $m \geqslant 1$)

$$= P[\underbrace{\omega_1 = i+1}_{A_1} \ \ AND \ \ \underbrace{0 < \omega_m < N \ \forall 1 \leqslant m \leqslant n}_{S} \mid \underbrace{\omega_0 = i}_{A_0}] +$$
$$+ P[\underbrace{\omega_1 = i-1}_{A_1'} \ \ AND \ \ \underbrace{0 < \omega_m < N \ \forall 1 \leqslant m \leqslant n}_{S} \mid \underbrace{\omega_0 = i}_{A_0}] =$$

$$= \frac{P[SA_1A_0]}{P[A_0]} + \frac{P[SA_1'A_0]}{P[A_0]} = \frac{\cancel{P[A_0]}P[A_1|A_0]P[S|A_1A_0]}{\cancel{P[A_0]}} + \frac{\cancel{P[A_0]}P[A_1'|A_0]P[S|A_1'A_0]}{\cancel{P[A_0]}}$$

where we used the extended conditional probability formula, Lemma 4.2. Moreover, since S *involves only the times* $m \geqslant 1$ *– the future relative to the conditioning on* ω_1 *– the Markov property applies cleanly (cf. Remark 10.6.3):* $P[S|A_1'A_0] = P[S|A_1']$, $P[S|A_1A_0] = P[S|A_1]$*, so that we proceed to*

$$R(i,n) = P[A_1|A_0]P[S|A_1] + P[A_1'|A_0]P[S|A_1'] = pR(i+1,n-1) + qR(i-1,n-1).$$

So altogether we have the initial-boundary value difference problem

$$\begin{gathered} R(i,n) = pR(i+1,n-1) + qR(i-1,n-1), \\ R(i,0) = 1 \ \forall i \in \{1,2,\dots,N-1\}, \qquad R(0,n) = R(N,n) = 0 \ \forall n \end{gathered} \tag{8}$$

How is this related to the length of the game? By definition, for any $n > 0$

$$R(i,n) = P[\text{the game is still going on at timestep } n \mid \omega_0 = i] =$$
$$= P[\text{the length of the game } L \text{ is greater than } n \mid \omega_0 = i] \implies$$

$$\implies L_i(n) := P[L = n \mid \omega_0 = i] = P[n-1 < L \leqslant n \mid \omega_0 = i] =$$
$$= P[n-1 < L \mid \omega_0 = i] - P[n < L \mid \omega_0 = i] = R(i,n-1) - R(i,n).$$

This has to be augmented with $L_i(0) = 0$, since for every $0 < i < N$ there is at least one round in the game.

Thus the expected length of the game (conditional that we started with $\omega_0 = i$) is

$$\overline{L}(i) := \mathrm{E}[L_i] = \sum_{n=1}^{\infty} nL_i(n) = \sum_{n=1}^{\infty} n\big(R(i,n-1) - R(i,n)\big) = \sum_{n=0}^{\infty} R(i,n),$$

assuming the series converge.

Claim: *The series converges and moreover $\sum\limits_n n^\delta \max\limits_i R(i,n) < \infty$ for all $\delta \in \mathbb{N}$. (Can you prove it on your own?) For now we take it for granted and proceed with the rest of the proof.*

Thus we have to sum up eq. (8) in n :

$$\sum_{n=1}^{\infty} R(i,n) = p\sum_{n=1}^{\infty} R(i+1,n-1) + q\sum_{n=1}^{\infty} R(i-1,n-1)$$

$$1 + \sum_{n=1}^{\infty} R(i,n) = 1 + p\sum_{n'=0}^{\infty} R(i+1,n') + q\sum_{n'=0}^{\infty} R(i-1,n')$$

$$\overline{L}(i) = 1 + p\overline{L}(i+1) + q\overline{L}(i-1)$$

which must be augmented by the boundary conditions $\overline{L}(0) = \overline{L}(N) = 0$. Now we can solve this in a way similar to what we did before. From here on we proceed for $p = q = \frac{1}{2}$:

$$\overline{L}(i+1) - \overline{L}(i) = \tfrac{q}{p}\big(\overline{L}(i) - \overline{L}(i-1)\big) - \tfrac{1}{p} \implies \overline{L}(i+1) - \overline{L}(i) = \big(\overline{L}(i) - \overline{L}(i-1)\big) - 2$$
$$\implies \overline{L}(i+1) - \overline{L}(i) = \overline{L}(1) - \cancelto{0}{L(0)} - 2i$$

and thus, by telescopic summation

$$\overline{L}(i) = \overline{L}(i) - \overline{L}(i-1) \ + \ \overline{L}(i-1) - \overline{L}(i-2) \ + \ \cdots + \ \overline{L}(1) - \overline{L}(0) =$$
$$= \sum_{j=0}^{i-1} (\overline{L}(1) - 2j) = i\overline{L}(1) - i(i-1) = i(\overline{L}(1) - (i-1)).$$

The condition $\overline{L}(0) = 0$ is satisfied automatically, while

$$\overline{L}(N) = 0 \implies \overline{L}(1) = N - 1,$$

so finally

$$\overline{L}(i) = i(N-i).$$

If we assume N is even and start with $i = \frac{N}{2}$, this result tells us that, on average, the time the symmetric random walk will take to cover a region of radius $\rho = \frac{N}{2}$ is ρ^2. In other words, the "typical displacement" during time Δt scales with the square root of Δt.

Question: What is the variance of the duration of the game, if it starts from $\$i$? Assume $p = q = \frac{1}{2}$.

Solution: Continuing on our previous analysis,

$$\mathrm{Var}[L(i)] = \mathrm{E}[L_i^2] - \overline{L}_i^2, \qquad \overline{L}_i = \sum_{n=0}^{\infty} R(i,n),$$

and

$$\mathrm{E}[L_i^2] = \sum_{n=1}^{\infty} n^2\big(R(i,n-1) - R(i,n)\big) = 2\sum_{n=1}^{\infty} nR(i,n) + \overline{L}_i$$

Denote

$$B(i) := \sum_{n=1}^{\infty} nR(i,n);$$

by summing up in n eq. (8) *and rearranging we derive*

$$p\big(B(i+1) - B(i)\big) = q\big(B(i) - B(i-1)\big) - \overline{L}(i) + 1 \implies$$
$$B(i+1) - B(i) = B(i) - B(i-1) - 2\overline{L}(i) + 2 \implies$$

$$B(i+1) - B(i) = B(1) - \cancelto{0}{B(0)} - 2\sum_{j=1}^{i} \overline{L}_j + 2i = B(1) + \tfrac{2}{3}i^3 + (1-N)i^2 + (\tfrac{7}{3} - N)i$$

where we used the lemmata

$$\sum_{i=0}^{n} i = \frac{n(n+1)}{2}, \qquad \sum_{i=0}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

Requiring $B(N) = 0$ leads to

$$0 = B(N) = B(N) - B(0) = \sum_{i=0}^{N-1}(B(i+1) - B(i)) =$$
$$= NB(1) + \tfrac{2}{3}\sum_{i=0}^{N-1} i^3 + (1-N)\sum_{i=0}^{N-1} i^2 + (\tfrac{7}{3} - N)\sum_{i=0}^{N-1} i =$$
$$= NB(1) + \tfrac{N^2(N-1)^2}{6} + (1-N)\tfrac{(N-1)N(2N-1)}{6} + (\tfrac{7}{3} - N)\tfrac{(N-1)N}{2} \implies$$
$$B(1) = (N-1)\big(\tfrac{(N-1)^2}{6} + \tfrac{N}{2} - \tfrac{7}{6}\big).$$

where we also used the fact that

$$\sum_{i=0}^{n} i^3 = \frac{n^2(n+1)^2}{4}.$$

Moreover,

$$B(i) = \sum_{j=0}^{i-1}(B(j+1) - B(j)) =$$
$$= iB(1) + \tfrac{2}{3}\sum_{j=0}^{i-1} j^3 + (1-N)\sum_{j=0}^{i-1} j^2 + (\tfrac{7}{3} - N)\sum_{j=0}^{i-1} j =$$
$$= iB(1) + \tfrac{i^2(i-1)^2}{6} + (1-N)\tfrac{(i-1)i(2i-1)}{6} + (\tfrac{7}{3} - N)\tfrac{i(i-1)}{2} =$$
$$= iB(1) + i(i-1)\tfrac{i(i-1)-(N-1)(2i-1)+7-3N}{6}$$

Thus finally

$$\mathrm{Var}[L(i)] =$$
$$= 2\left[i(N-1)\big(\tfrac{(N-1)^2}{6} + \tfrac{N}{2} - \tfrac{7}{6}\big) + i(i-1)\tfrac{i(i-1)-(N-1)(2i-1)+7-3N}{6}\right] + \overline{L}(i) - \overline{L}^2(i)$$

10.3 Potentials

The considerations that came up in the study of "gambler's ruin" can be formalised and extended as follows:

Definition 10.12 (Absorbing state)**.** *A state $j \in X$ of a Markov chain is called absorbing, if*

$$P[\omega_{n+1} = l \mid \omega_n = j] = 0 \qquad \forall l \neq j.$$

So the "gambler's ruin" problem is really a finite-state Markov chain, where the states 0 and N are absorbing, i.e. the game "ends" when one of these is reached. Absorbing states are very convenient for the modelling of such situations.

Definition 10.13 (Generalised first passage time)**.** *Given a Markov chain with state space X and $A \subseteq X$, the random variable $T_A^{(n_0)}$ is defined as*

$$T_A^{(n_0)} := \inf\{n \geqslant n_0 \mid \omega_n \in A\}$$

If $A = \{j\}$, we will write for brevity $T_j^{(n_0)} := T_{\{j\}}^{(n_0)}$.

Definition 10.14 (Potential)**.** *Given a Markov chain with state space X and* disjoint *$A, B \subseteq X$ ($A \cap B = \emptyset$ – otherwise the boundary conditions below conflict), the potential*

$$\Phi_{A,B} : X \to \mathbb{R} : i \mapsto \Phi_{A,B}(i)$$

is defined as

$$\Phi_{A,B}(i) = P[T_A^{(0)} < T_B^{(0)} \mid \omega_0 = i].$$

Theorem 10.15. *Consider a homogeneous Markov chain with state space $X = \{1, \ldots, r\}$ and transition matrix $P = [p_{ij}]$. Then the potential $\Phi(i) := \Phi_{A,B}(i)$ satisfies the boundary value problem*

$$\begin{cases} \Phi(i) = \sum_j p_{ij}\Phi(j), & i \notin A \cup B \\ \Phi(i) = 1, & i \in A \\ \Phi(i) = 0, & i \in B. \end{cases}$$

Proof: The potential $\Phi_{A,B}(i)$ obviously satisfies the boundary conditions $\Phi_{A,B}(i) = 1$, $i \in A$, $\Phi_{A,B}(i) = 0$, $i \in B$.

For the equation itself, we will condition on the first step: for any $i \notin A \cup B$,

$$\Phi_{A,B}(i) = P[T_A^{(0)} < T_B^{(0)} \mid \omega_0 = i] = \sum_{j=1}^{r} P[\omega_1 = j \mid \omega_0 = i] P[T_A^{(0)} < T_B^{(0)} \mid \omega_1 = j, \omega_0 = i] =$$
$$= \sum_{j=1}^{r} p_{ij} P[T_A^{(1)} < T_B^{(1)} \mid \omega_1 = j, \omega_0 = i] = \sum_{j=1}^{r} p_{ij} P[T_A^{(0)} < T_B^{(0)} \mid \omega_0 = j] = \sum_{j=1}^{r} p_{ij} \Phi_{A,B}(j)$$

□

Mean first-passage times. The same first-step conditioning yields equations for expected hitting times. For $A \subseteq X$ define the **first-passage time** $T_A := \min\{n \geqslant 0 : \omega_n \in A\}$ and

$$k_i := \mathrm{E}[T_A \,|\, \omega_0 = i].$$

Theorem 10.16. *For a homogeneous Markov chain with transition matrix $P = [p_{ij}]$, and assuming $\mathrm{E}[T_A \,|\, \omega_0 = i] < \infty$ for the states under consideration (for a finite chain this holds whenever A is reached almost surely from i), the mean first-passage times satisfy*

$$\begin{cases} k_i = 1 + \sum_j p_{ij} k_j, & i \notin A, \\ k_i = 0, & i \in A. \end{cases}$$

Proof: For $i \notin A$ at least one step is taken; conditioning on the first step and using the Markov property and homogeneity, exactly as in the potentials theorem, $k_i = \sum_j p_{ij}(1 + k_j)$. □

Remark 10.17 (The gambler's ruin, systematised)**.** *The hands-on computations of the gambler's ruin example are precisely these boundary value problems on $X = \{0, 1, \ldots, N\}$: the winning probability $P(i)$ is the potential $\Phi_{\{N\},\{0\}}(i)$, and the expected duration $\overline{L}(i)$ solves the mean first-passage system for $A = \{0, N\}$, which is exactly the difference equation $\overline{L}(i) = 1 + p\overline{L}(i+1) + q\overline{L}(i-1)$ derived there by hand. This is the general pattern: behind every "one-line formula" of Markov chain theory sits a first-step-conditioning computation of the kind we did explicitly. A systematic, computation-forward development of hitting probabilities, first-passage times and their many applications can be found in Privault [16] and Norris [15].*

10.4 The ergodic theorem and the law of large numbers for Markov chains

Having built intuition on a hands-on example, we now turn to the systematic, "industrial-scale" theory: what happens to a Markov chain in the long run?

Definition 10.18 (Ergodic matrices)**.** *A stochastic matrix P is called ergodic if there exists $s \in \mathbb{N}$ so that $P^s = [p_{ij}^{(s)}]$ has strictly positive entries, $p_{ij}^{(s)} > 0$ for all $i, j \in \{1, 2, \ldots, r\}$.*

A homogeneous Markov chain with ergodic transition matrix is often called itself an ergodic Markov chain. In such a Markov chain, it is possible with non-zero probability to go from

any state i to any state j in *exactly* s steps – that is what $[P^s]_{ij} > 0$ says ("within s steps" is a weaker statement).

Definition 10.19 (Stationary or invariant distribution)**.** *Given a homogeneous Markov chain with transition matrix P, a pdf π is called stationary or invariant if*

$$\pi = \pi P$$

Theorem 10.20 (Ergodic Theorem for Markov Chains)**.** *For every homogeneous Markov chain with ergodic transition matrix P, there exists a unique stationary distribution $\pi > 0$ so that*

$$\lim_{n\to\infty} \mu P^n = \pi \qquad \forall \text{ pdf } \mu.$$

Proof: Step 1 (Introduction and representation of the metric)

We will use the metric

$$d(\mu', \mu'') := \frac{1}{2}\sum_{i=1}^{r} |\mu'_i - \mu''_i|$$

This is a rescaled l^1 distance, and it admits the representation

$$d(\mu', \mu'') = \frac{1}{2}\left[\sum_{i\in I}(\mu'_i - \mu''_i) \; - \; \sum_{i\in I^C}(\mu'_i - \mu''_i)\right], \qquad I = \{i \,|\, \mu'_i > \mu''_i\}.$$

The benefit of this representation is that it is a usual sum, i.e. linear operations (such as taking common factors) are now permitted.

Moreover, if μ', μ'' are pdfs it satisfies the obvious bound

$$d(\mu', \mu'') \leqslant \frac{1}{2}\sum_{i=1}^{r} \big(|\mu'_i| + |\mu''_i|\big) = 1.$$

Step 2 (The contraction estimate $d(\mu'P^s, \mu''P^s) \leqslant (1-a)d(\mu', \mu'')$)

Let $s \in \mathbb{N}$ be such that $p_{ij}^{(s)} > 0 \; \forall i, j$, and μ', μ'' two pdfs. Observe that

$$d(\mu'P^s, \mu''P^s) = \frac{1}{2}\left[\sum_{j\in J}\big((\mu'P^s)_j - (\mu''P^s)_j\big) - \sum_{j\in J^C}\big((\mu'P^s)_j - (\mu''P^s)_j\big)\right]$$

where

$$J := \{j \,|\, (\mu'P^s)_j > (\mu''P^s)_j\}$$

Thus we now have

$$d(\mu'P^s, \mu''P^s) = \tfrac{1}{2}\Big[\sum_{j\in J}\sum_{i=1}^{r}(\mu'_i - \mu''_i)p_{ij}^{(s)} -$$

$$- \sum_{j\in J^C}\sum_{i=1}^{r}(\mu'_i - \mu''_i)p_{ij}^{(s)}\Big] = \tfrac{1}{2}\sum_{i=1}^{r}(\mu'_i - \mu''_i)\Big[\sum_{j\in J} p_{ij}^{(s)} - \sum_{j\in J^C} p_{ij}^{(s)}\Big]$$

At this point observe that

$$J^C = \emptyset \implies (\mu' P^s)_j > (\mu'' P^s)_j \ \forall j \implies \sum_j (\mu' P^s)_j > \sum_j (\mu'' P^s)_j,$$

which is absurd since both $\mu' P^s$, $\mu'' P^s$ are pdfs and have sum equal to one. So J^C has at least one element and the smallest possible value of $\sum_{j \in J^C} p_{ij}^{(s)}$ is $\min_{i,j} p_{ij}^{(s)} > 0$, which we will denote from now on with $a := \min_{i,j} p_{ij}^{(s)} > 0$.

Continuing our computation, observe that by the same argument $J \neq \emptyset$ as well, so that for every i

$$-\sum_{j \in J^C} p_{ij}^{(s)} \leqslant -a, \qquad \sum_{j \in J} p_{ij}^{(s)} \leqslant 1 - a,$$

hence the bracket is bounded in absolute value by $1 - a$, and we obtain

$$d(\mu' P^s, \mu'' P^s) \leqslant (1-a)\frac{1}{2}\sum_{i=1}^{r} |\mu'_i - \mu''_i| = (1-a) d(\mu', \mu'') \tag{9}$$

Step 3 (Existence of the limit)

Now we have all we need to show that, given a pdf μ, the sequence $\mu_n := \mu P^n$ is a Cauchy sequence in the metric $d(\cdot, \cdot)$. Indeed observe that

$$d(\mu P^{N+n}, \mu P^{N+m}) = d(\mu P^{sN'+n'}, \mu P^{sN'+m'}) \leqslant (1-a)^{N'} d(\mu P^{n'}, \mu P^{m'}) \leqslant (1-a)^{N'},$$

where of course N' is the quotient $N' = \lfloor \frac{N}{s} \rfloor$ and $m' = m + rem(N, s) = m + N - \lfloor \frac{N}{s} \rfloor s$, $n' = n + rem(N, s)$.

Thus, given an $\epsilon > 0$, if $N > s + s\frac{\log \epsilon}{\log(1-a)}$ then

$$d(\mu_n, \mu_m) < \epsilon \qquad \forall\, n, m > N,$$

i.e. μ_n is a Cauchy sequence.

By virtue of the completeness of $\mathbb{R}^r$ in the l^1 norm, the limit exists. We will denote it by

$$\mu_\infty := \lim_n \mu_n.$$

Step 4 (Stationarity, uniqueness and positivity of the limit)

One readily checks that, since each of the μ_n is a pdf, μ_∞ is a pdf as well.

Moreover,

$$\mu_\infty = \lim_{n\to\infty} \mu_n = \lim_{n\to\infty} \mu_{n+1} = \lim_{n\to\infty} \mu P^{n+1} = (\lim_{n\to\infty} \mu P^n) P = \mu_\infty P,$$

i.e. μ_∞ is an invariant measure. Now, if π', π'' are two invariant measures, then

$$d(\pi', \pi'') = d(\pi' P^s, \pi'' P^s) \leqslant (1-a) d(\pi', \pi''),$$

which is impossible unless $d(\pi', \pi'') = 0 \implies \pi' = \pi''$. Thus, there can only be one invariant

distribution, which from now on we will denote by π to emphasise the independence from the initial pdf μ.

To show that every coordinate of π is strictly positive, let us assume that for some k, $\pi_k = 0$. Then

$$0 = \pi_k = \sum_{i=1}^{r} \pi_i p_{ik}^{(s)}.$$

Since all of π_i, $p_{ij}^{(s)}$ are non-negative, the only way the sum is zero is if all of the terms $\pi_i p_{ik}^{(s)}$, $i = 1, \ldots, r$ are zero. However, since π has sum 1 (as a pdf), not all π_i can be zero; so assume $\pi_l > 0$. But then

$$\pi_l p_{lk}^{(s)} = 0 \implies p_{lk}^{(s)} = 0,$$

which is impossible since $p_{ij}^{(s)} \geqslant a > 0$.

The proof is complete. □

Alongside the ergodic theorem one has the following law of large numbers. Note that it does *not* follow directly from the convergence $\mu P^n \to \pi$ of the one-time distributions: it concerns *time averages along a single trajectory* of the (dependent!) sequence $\omega_1, \omega_2, \ldots$, and its proof requires a separate argument (see [15] or [11]).

Theorem 10.21 (Law of Large Numbers for Markov Chains)**.** *Let π be the stationary distribution of an ergodic Markov chain with transition matrix $P = [p_{ij}]$. Then for any $\epsilon > 0$*

$$\lim_{n\to\infty} P[|\frac{\nu_i^n}{n} - \pi_i| > \epsilon] = 0$$

where $\nu_i^n(\omega)$ is the number of occurrences of i in n timesteps,

$$\nu_i^n(\omega) = \#\{l \in \{1, 2, \ldots n\} \mid \omega_l = i\}$$

and

$$\lim_{n\to\infty} P[|\frac{\nu_{ij}^n}{n} - \pi_i p_{ij}| > \epsilon] = 0$$

where $\nu_{ij}^n(\omega)$ is the number of transitions from state i to state j in n timesteps,

$$\nu_{ij}^n(\omega) = \#\{l \in \{0, 1, \ldots n-1\} \mid \omega_l = i \ AND \ \omega_{l+1} = j\}.$$

Further reading: Koralov and Sinai [11] discuss Markov chains and random walks along lines similar to this chapter; Knill [14] has a nice treatment of random walks in Chapter 3; for potential theory and much more, see Norris [15]; Privault [16] is a systematic, example-driven treatment of exactly the first-passage machinery of this chapter.

Exercises

Exercise 10.1 (Two-state chains)**.** *Consider the two-state chain with transition matrix* $P = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$, $a, b \in (0, 1)$. *Find the stationary distribution, show that the chain*

is ergodic, and compute P^n explicitly (hint: eigenvalues 1 and $1-a-b$). At which rate does μP^n converge to the stationary distribution?

Exercise 10.2. *Decide which of the following are ergodic:*

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

Exercise 10.3. *(*) For the gambler's ruin with $p \neq q$, derive the expected duration of the game,*

$$\overline{L}(i) = \frac{i}{q-p} - \frac{N}{q-p} \cdot \frac{1-(\frac{q}{p})^i}{1-(\frac{q}{p})^N}.$$

(Hint: follow the method used for $p = q = \frac{1}{2}$, keeping the inhomogeneous term.)

Exercise 10.4 (Success runs)**.** *A fair coin is tossed repeatedly. Using mean first-passage equations on a suitable small state space (e.g. "no progress", "one useful toss", "done"), compute the expected number of tosses until the first appearance of the pattern HH; then of the pattern HT. The two answers differ – explain why.*

Exercise 10.5. *For the four-state chain drawn in Section 10.1 ($p_{12} = p_{13} = p_{14} = \frac{1}{3}$; $p_{23} = p_{24} = \frac{1}{2}$; $p_{31} = 1$; $p_{41} = p_{43} = \frac{1}{2}$), starting from state 1:*

1. *compute the probability of visiting state 3 before state 2;*
2. *compute the mean first-passage time to state 3.*

11 Markov processes in continuous time

The Markov chains of Chapter 10 evolve at discrete time steps $n = 0, 1, 2, \dots$ In this chapter time becomes continuous, and we develop Markov processes in two stages, climbing the ladder of state spaces.

In the first stage the states remain *discrete*: phone calls arrive, customers join and leave a queue, individuals are born and die – at random moments in continuous time. Here the entire analysis reduces to ordinary differential equations: we construct the Poisson process, then general continuous-time chains and their master equations, then birth-death processes. In the second stage (Section 11.4) the states become *continuous* as well, and the same strategy – Markov property, Chapman-Kolmogorov, conditioning on a short time step – yields the Kolmogorov forward and backward *partial* differential equations. The starred Section 11.5 connects the two stages: diffusions arise as scaling limits of frequent, small jumps. The most important Markov process of all, Brownian motion, then receives its own Chapter 12.

11.1 The Poisson process

Definition 11.1 (Counting process)**.** *A random process $\{N(t)\}_{t\geqslant 0}$ is called a counting process if its realisations are non-decreasing, right-continuous, take values in $\mathbb{N}\cup\{0\}$, and $N(0) = 0$. We interpret $N(t) - N(s)$ as the number of "events" in the time interval $(s, t]$.*

Definition 11.2 (Poisson process)**.** *A counting process is a Poisson process with rate $\lambda > 0$ if*

(i). it has independent increments: for any $t_0 < t_1 < \cdots < t_r$, the random variables $N(t_{j+1}) - N(t_j)$ are independent;

(ii). it has stationary increments: the distribution of $N(t + s) - N(t)$ depends only on s;

(iii). events are simple and occur at rate λ:

$$P[N(t + \Delta t) - N(t) = 1] = \lambda\Delta t + o(\Delta t), \qquad P[N(t + \Delta t) - N(t) \geqslant 2] = o(\Delta t).$$

Observe that postulate (iii) is the infinitesimal version of the law of rare events (Section 6.3): splitting $(0, t]$ into $N \gg 1$ slots of length $\frac{t}{N}$, each slot contains an event with probability $\approx \frac{\lambda t}{N}$, independently of the others – and $B_{N,\lambda t/N}$ converges to a Poisson count. The following theorem confirms this expectation, by the method we will use throughout this chapter.

Theorem 11.3. *For a Poisson process with rate λ,*

$$P[N(t) = n] = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \qquad n = 0, 1, 2, \dots$$

i.e. $N(t)$ is a Poisson random variable with parameter λt.

Proof: Denote $p_n(t) := P[N(t) = n]$. Conditioning on what happens in $(t, t + \Delta t]$ and using (i), (iii),

$$p_0(t + \Delta t) = p_0(t)\big(1 - \lambda\Delta t\big) + o(\Delta t) \quad \Longrightarrow \quad p_0'(t) = -\lambda p_0(t), \quad p_0(0) = 1,$$

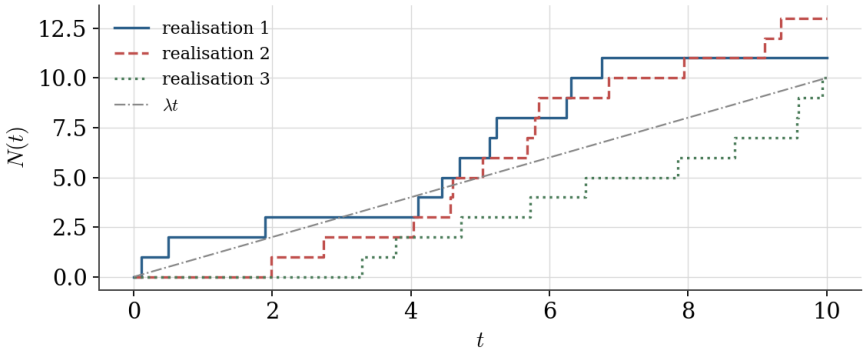
so that $p_0(t) = e^{-\lambda t}$ – exactly as in Example 4.6. For $n \geqslant 1$, the state $N(t+\Delta t) = n$ can be reached from n (no event) or from $n-1$ (one event), all other transitions having probability $o(\Delta t)$:

$$p_n(t+\Delta t) = p_n(t)(1-\lambda\Delta t)+p_{n-1}(t)\lambda\Delta t+o(\Delta t) \quad \Longrightarrow \quad p_n'(t) = -\lambda p_n(t)+\lambda p_{n-1}(t), \quad p_n(0) = 0.$$

Multiplying by the integrating factor $e^{\lambda t}$, $\left(e^{\lambda t}p_n(t)\right)' = \lambda e^{\lambda t}p_{n-1}(t)$, and by induction, if $p_{n-1}(t) = e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}$ then

$$e^{\lambda t}p_n(t) = \int_0^t \lambda\frac{(\lambda s)^{n-1}}{(n-1)!}ds = \frac{(\lambda t)^n}{n!}.$$

□



Lemma 11.4 (Interarrival times)**.** *Let T_1 be the time of the first event, and T_k the time between the $(k-1)^{st}$ and the k^{th} event. Then the T_k are i.i.d. exponential random variables with parameter λ.*

Proof sketch: $P[T_1 > t] = P[N(t) = 0] = e^{-\lambda t}$, i.e. T_1 is exponential. Given the time of the $(k-1)^{st}$ event, the process "restarts" by the independence and stationarity of increments, making the subsequent interarrival time an independent copy of T_1. (Making the restart argument fully rigorous requires conditioning on the past trajectory; we do not spell out the details.) □

Thus the Poisson process and the exponential distribution are two views of the same object: Poisson-distributed *counts* $\Longleftrightarrow$ exponential, memoryless *gaps*. The machine of Example 4.6 fails at the first event time of a Poisson process.

A toolbox of standard manipulations of the Poisson process – superposition, thinning, uniform conditional order statistics, competing clocks – is collected in Exercise 11.1, and will be used freely below.

11.2 Continuous-time Markov chains and the master equations

The Poisson process moves only "upwards". The general object is the following:

Definition 11.5 (Continuous-time Markov chain)**.** *Let X be a finite or countable state space. A process $\{\omega(t)\}_{t\geqslant 0}$ with values in X is a continuous-time Markov chain with jump rates*

$q_{ij} \geqslant 0$ $(i \neq j)$, $q_i := \sum_{j \neq i} q_{ij} < \infty$, *if it has the Markov property (the future depends on the past only through the present state) and*

$$P[\omega(t+\Delta t) = j \,|\, \omega(t) = i] = \begin{cases} q_{ij}\Delta t + o(\Delta t), & j \neq i, \\ 1 - q_i \Delta t + o(\Delta t), & j = i. \end{cases}$$

The matrix Q with $Q_{ij} = q_{ij}$ $(i \neq j)$ and $Q_{ii} = -q_i$ is called the generator of the chain; its rows sum to zero.

Denote $p_{ij}(t) := P[\omega(s+t) = j | \omega(s) = i]$ (independent of s) and $P(t) = [p_{ij}(t)]$. By the total probability formula and the Markov property, exactly as for discrete chains,

$$P(t+s) = P(t)P(s), \qquad P(0) = I,$$

the **Chapman-Kolmogorov equation** in this setting (cf. its continuous-state counterpart in Section 11.4).

Remark 11.6 (The transition-probability formulation, following [12])**.** *One can equally well take conditional probabilities as the primitive object. Introduce transition probabilities $P\{j, s \,|\, i, s_0\}$, $s_0 < s$; the Markov property is the requirement*

$$P\{\omega_s = j \,|\, \omega_{s_0} = i,\, \omega_{s_1} = i_1, \ldots, \omega_{s_k} = i_k\} = P\{\omega_s = j \,|\, \omega_{s_0} = i\} \qquad \forall\, s > s_0 > s_1 > \cdots > s_k,$$

and the Chapman-Kolmogorov equation takes the form

$$P\{j, s \,|\, i_1, s_1\} = \sum_{i_0} P\{i_0, s_0 \,|\, i_1, s_1\}\, P\{j, s \,|\, i_0, s_0\}, \qquad s_1 < s_0 < s,$$

valid with no *homogeneity assumption. Our matrix $P(t)$ is the homogeneous special case $P\{j, s|i, s_0\} = p_{ij}(s - s_0)$.*

Theorem 11.7 (Kolmogorov equations / master equation)**.** *For a continuous-time Markov chain with bounded rates,*

$$P'(t) = P(t)\, Q \quad \textit{(forward)}, \qquad P'(t) = Q\, P(t) \quad \textit{(backward)}.$$

Moreover, the distribution $p_j(t) := P[\omega(t) = j]$ satisfies the **master equation**

$$p_j'(t) = \underbrace{\sum_{i \neq j} p_i(t)\, q_{ij}}_{\textit{inflow}} - \underbrace{p_j(t)\, q_j}_{\textit{outflow}}.$$

Proof sketch: Condition on the last short interval: by Chapman-Kolmogorov and the rate postulates,

$$p_{ij}(t+\Delta t) = \sum_k p_{ik}(t) p_{kj}(\Delta t) = p_{ij}(t)(1 - q_j \Delta t) + \sum_{k \neq j} p_{ik}(t) q_{kj} \Delta t + o(\Delta t),$$

subtract $p_{ij}(t)$, divide by Δt and let $\Delta t \to 0$. For a finite state space this is immediate; for countable X one needs the $o(\Delta t)$ of Definition 11.5 to be uniform over the states – a standing assumption here (the general, "regular chain", construction is in [15]). Conditioning on the *first* short interval instead yields the backward equation. The master equation follows by multiplying the forward equation by the initial distribution. □

Remarks:

- For a finite state space, the solution is the matrix exponential $P(t) = e^{tQ}$.
- This is the same derivation strategy as for the Kolmogorov equations of Section 11.4 – Markov property → Chapman-Kolmogorov → condition on a short first/last step – with the generator matrix Q playing the role that the differential operator $a\,\partial_x + \frac{b}{2}\partial_x^2$ plays there. Note, however, that the theorems of Section 11.4 do *not* apply directly here: their assumption (i) explicitly forbids jumps, while our chains move *only* by jumps. Section 11.5 explains how the two settings connect.
- A stationary distribution is a pdf π on X with $\pi Q = 0$; then $p(0) = \pi \implies p(t) = \pi$ $\forall t$.

11.3 Birth-death processes

Definition 11.8 (Birth-death process)**.** *A birth-death process is a continuous-time Markov chain on $X = \mathbb{N} \cup \{0\}$ which only jumps to nearest neighbours:*

$$q_{n,n+1} = \lambda_n \quad \textit{(birth rates)}, \qquad q_{n,n-1} = \mu_n \quad \textit{(death rates)}, \qquad \mu_0 = 0,$$

all other rates being zero. The master equation reads, for $n \geqslant 1$,

$$p_n'(t) = \lambda_{n-1} p_{n-1}(t) + \mu_{n+1} p_{n+1}(t) - (\lambda_n + \mu_n) p_n(t),$$

with the separate boundary equation $p_0'(t) = \mu_1 p_1(t) - \lambda_0 p_0(t)$ at $n = 0$ (equivalently: the displayed equation holds for all $n \geqslant 0$ with the conventions $\lambda_{-1} = \mu_0 = 0$, $p_{-1} \equiv 0$).

Standard examples:

model	birth rates	death rates
Poisson process	$\lambda_n = \lambda$	$\mu_n = 0$
M/M/1 queue	$\lambda_n = \lambda$	$\mu_n = \mu$ for $n \geqslant 1$
M/M/∞ queue	$\lambda_n = \lambda$	$\mu_n = n\mu$
linear birth-death (population)	$\lambda_n = n\lambda$	$\mu_n = n\mu$

In the queueing models, n is the number of customers in the system: arrivals form a Poisson process with rate λ, and each service completes at rate μ (one server working in M/M/1; every customer served simultaneously in M/M/∞). In the population model each individual independently gives birth at rate λ and dies at rate μ – whence the factors n.

Theorem 11.9 (Stationary distributions of birth-death processes)**.** *Assume $\mu_k > 0$ for all $k \geqslant 1$. A pdf π on $\mathbb{N} \cup \{0\}$ is stationary if and only if it satisfies the* **detailed balance** *equations*

$$\lambda_n \pi_n = \mu_{n+1} \pi_{n+1} \qquad \forall n \geqslant 0,$$

in which case

$$\pi_n = \pi_0 \prod_{k=1}^{n} \frac{\lambda_{k-1}}{\mu_k},$$

and such a (normalisable) π exists if and only if $\sum_n \prod_{k=1}^{n} \frac{\lambda_{k-1}}{\mu_k} < \infty$.

Proof: Define the probability flux across the "edge" between $n-1$ and n,

$$F_n := \lambda_{n-1} \pi_{n-1} - \mu_n \pi_n, \qquad n \geqslant 1.$$

The stationary master equation $0 = \lambda_{n-1}\pi_{n-1} + \mu_{n+1}\pi_{n+1} - (\lambda_n + \mu_n)\pi_n$ says precisely that $F_{n+1} = F_n$ for all $n \geqslant 1$, while at the boundary ($n = 0$, $\mu_0 = 0$) it says $F_1 = 0$. Hence all fluxes vanish, which is detailed balance; the product formula follows by iterating $\pi_{n+1} = \frac{\lambda_n}{\mu_{n+1}} \pi_n$, and π_0 is fixed by normalisation. □

Example 11.10 (The M/M/1 queue)**.** *For $\lambda_n = \lambda$, $\mu_n = \mu$, set $\rho := \frac{\lambda}{\mu}$ (the "traffic intensity"). The product formula gives $\pi_n = \pi_0 \rho^n$, which is normalisable iff $\rho < 1$:*

$$\pi_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, \ldots$$

a geometric distribution. The stationary mean number of customers in the system *is*

$$\sum_{n=0}^{\infty} n \pi_n = \frac{\rho}{1-\rho},$$

which blows up as $\rho \uparrow 1$: a server loaded at 95% of capacity carries on average 19 customers in the system (of which $\rho^2/(1-\rho) \approx 18.05$ are waiting in the queue proper) – long queues appear well before the system is "overloaded". For $\rho \geqslant 1$ there is no stationary distribution, but the two regimes differ qualitatively: for $\rho > 1$ the chain is transient *– the queue grows without bound; at the critical value $\rho = 1$ it is* null recurrent *– it returns to the empty state infinitely often, yet its excursions are so long that no stationary distribution exists.*

(The M/M/∞ analogue – with a Poisson stationary law – is Exercise 11.2.)

Theorem 11.11 (Extinction)**.** *Consider the linear birth-death process ($\lambda_n = n\lambda$, $\mu_n = n\mu$), for which the state 0 is absorbing ("extinction"). Starting from i individuals,*

$$P[\textit{extinction} \,|\, \omega(0) = i] = \begin{cases} 1, & \lambda \leqslant \mu, \\ \left(\frac{\mu}{\lambda}\right)^i, & \lambda > \mu. \end{cases}$$

Proof idea: Watch the process only at its jump times (the *embedded chain*). From any state $n \geqslant 1$, the next jump is a birth with probability $\frac{\lambda_n}{\lambda_n + \mu_n} = \frac{\lambda}{\lambda+\mu}$ (cf. Exercise 11.1.4, with the two competing exponential clocks) – independent of n, because the factors n cancel.

So the embedded chain is precisely the random walk of the gambler's ruin, Example 10.11, with $p = \frac{\lambda}{\lambda+\mu}$, $q = \frac{\mu}{\lambda+\mu}$, and extinction is ruin against an infinitely rich adversary: taking $N \to \infty$ in the gambler's ruin formulas gives probability of ruin 1 when $p \leqslant q$ ($\lambda \leqslant \mu$) and $\left(\frac{q}{p}\right)^i = \left(\frac{\mu}{\lambda}\right)^i$ when $p > q$. □

Note the qualitative conclusion: for $\lambda > \mu$ the population is *not* safe – it still dies out with positive probability $(\mu/\lambda)^i$, small only if the initial population is large.

11.4 Continuous state space: the Kolmogorov equations

One can think of a continuous-time, continuous-space Markov process as a continuous limit of a Markov chain. To that end, we will need some definitions:

Definition 11.12 (Generalised conditional probability)**.** *Let $(\Omega, \mathcal{E}, P)$ be a probability space, X a continuous R.V. on it, and consider the events*

$$B \in \mathcal{E}, \qquad A_\epsilon = \{x - \frac{\epsilon}{2} < X < x + \frac{\epsilon}{2}\} \in \mathcal{E}.$$

Observe that

$$P[B|A_\epsilon] = \frac{P[BA_\epsilon]}{P[A_\epsilon]} = \frac{P[BA_\epsilon]}{F_X(x + \frac{\epsilon}{2}) - F_X(x - \frac{\epsilon}{2})}$$

Then

$$P[B|X = x] := \lim_{\epsilon \to 0} \frac{P[BA_\epsilon]}{F_X(x + \frac{\epsilon}{2}) - F_X(x - \frac{\epsilon}{2})}$$

if the limit exists.

The term "generalised" is used because, for a continuous random variable, $P[X = x] = 0$, so this new kind of conditional probability we end up with is not included in the original discussion of conditional probability. Two caveats are in order: the limit may fail to exist; and any assignment of conditional probabilities is canonical only up to P_X-null sets – where the pointwise limit exists it is a definite number, the null-set ambiguity belongs to the abstract object; the fully general object – the *regular conditional probability* – is beyond our scope (see [11]). In the case we actually use, where the variables have a joint pdf, everything reduces to the explicit formula $f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$ for $f_X(x) > 0$, derived below.

Lemma 11.13 (Continuous total probability formula)**.** *In the context of Definition 11.12, we have*

$$P[B] = \int_{x \in \mathbb{R}} P[B|X = x] f_X(x) dx$$

assuming all the objects appearing in the equation above are well defined.

Remark: as indicated in the title of the section, this is really a formal result. We don't have a good enough way to check in detail all the assumptions involved in the proof yet. So the characterisation "lemma" is somewhat loose. (One concrete instance: the sketch below uses a single representative $\overline{x}_i$ per cell to stand in for two different ratios at once; no mean-value theorem guarantees a common point. Under the smoothness assumed, each ratio equals its

value at $\overline{x}_i$ up to errors vanishing with ϵ, uniformly – that is what actually justifies the step.)

Sketch of the proof: Denote $I_i^\epsilon = [x_i, x_{i+1})$ a uniform partition of $\mathbb{R}$ with $x_{i+1} = x_i + \epsilon \ \forall i$, $\lim_{i\to\pm\infty} x_i = \pm\infty$, and

$$A_i^\epsilon = \{X \in I_i^\epsilon\}.$$

By the classical total probability formula we have

$$P[B] = P[B \cap \Omega] = P[B \cap (\cup_i A_i^\epsilon)] = \sum_{i=-\infty}^{+\infty} P[B|A_i^\epsilon]P[A_i^\epsilon] =$$
$$= \sum_{i=-\infty}^{+\infty} \underbrace{\frac{P[BA_i^\epsilon]}{F_X(x_{i+1}) - F_X(x_i)}}_{\to P[B|X=\overline{x}_i]} \underbrace{\frac{F_X(x_{i+1}) - F_X(x_i)}{x_{i+1} - x_i}}_{\to f_X(\overline{x}_i)} \underbrace{(x_{i+1} - x_i)}_{\to dx|_{x\approx\overline{x_i}}}$$

for some $x_i \leqslant \overline{x}_i \leqslant x_{i+1}$.

Now, as the above schematics indicate, if the conditional probability $P[B|X = x]$ exists for all x, and F_X is smooth enough, taking the limit $\epsilon \to 0$ above (and assuming that the resulting integral is finite) completes the proof. □

In the same way that we have generalised conditional probabilities of the form $P[B|X = x]$ we can have generalised conditional pdf's and cdf's.

Definition 11.14 (Generalised conditional cdf's and pdf's)**.** *Assume the continuous random variables X, Y have a smooth enough joint cdf $F_{XY}(x, y)$. Then*

$$P[Y \leqslant y|X = x] = F_Y(y|X = x) = \lim_{\substack{\epsilon\to 0\\ s\to\infty}} \frac{F_{XY}(x + \frac{\epsilon}{2}, y) - F_{XY}(x - \frac{\epsilon}{2}, y)}{F_{XY}(x + \frac{\epsilon}{2}, s) - F_{XY}(x - \frac{\epsilon}{2}, s)} = \lim_{s\to\infty} \frac{\frac{\partial}{\partial x}F_{XY}(x, y)}{\frac{\partial}{\partial x}F_{XY}(x, s)}$$

if the limit exists. Moreover the conditional pdf is

$$f_Y(y|X = x) = \frac{\partial}{\partial y}F_Y(y|X = x)$$

if it exists, in the sense that

$$P[Y \in B|X = x] = \int_B f_Y(y|X = x)dy.$$

Definition 11.15 (Conditional cdf and pdf for a random process)**.** *Consider a random process, i.e. a collection*

$$\{X(t)\}_{t\in[a,b]}, \qquad [a, b] \subseteq \mathbb{R} \tag{10}$$

of random variables, which in general are dependent with each other. One possible way to describe such a random process is through the conditional cdf:

$$F(t, x; \tau, y) := P[X(\tau) \leqslant y|X(t) = x], \qquad t \leqslant \tau, \quad x, y \in \mathbb{R},$$

and the corresponding conditional pdf

$$f(t,x;\tau,y) := \frac{\partial}{\partial y}F(t,x;\tau,y), \qquad t \leqslant \tau, \quad x,y \in \mathbb{R}.$$

Remarks:

- The only difference from Definition 11.14 is that the different random variables are now different time instances of the same random process. That is why we need to introduce explicitly as arguments the times t, τ.
- If our random process has (an appropriate version of) the **Markov property**, cf. Remark 10.6.3, then we would expect that the conditional cdf or pdf provides a *complete* description for the underlying probability measure. If that is the case, we should be able to answer all questions in terms only of $f(t,x;\tau,y)$ or $F(t,x;\tau,y)$. We will see some examples of that later on.

  A random process with the Markov property is called **a Markov process**.

Remark 11.16 (Transition functions, following [12])**.** *The clean axiomatic object underneath this discussion is the* **transition function** $P(C,t;x,s)$, $s<t$ *– to be read as "the probability that* $\omega_t \in C$*, given* $\omega_s = x$*" – required to satisfy:*

(i). for fixed s,t,x*, it is a probability measure in* C *(over the Borel subsets of* $\mathbb{R}^d$*);*

(ii). for every bounded measurable f*, the function* $x \mapsto \int f(y)\, d_y P(y,t;x,s)$ *is measurable;*

(iii). $P(C,s;x,s) = 1$ *if* $x \in C$ *and* 0 *otherwise.*

A Markov process is then specified by a transition function together with an initial distribution μ_0*, its finite-dimensional distributions being built up by iterated integration. Our conditional cdf is recovered as* $F(s,x;t,y) = P\big((-\infty,y],t;x,s\big)$*. This formulation keeps track of exactly what is needed, and postpones the regularity questions (existence of densities, of the limits defining conditional probabilities, etc.) to where they belong.*

Lemma 11.17 (Chapman-Kolmogorov-Markov equation)**.** *Consider a random process* $X(t)$ *as in eq.* (10) *with the Markov property, and assume its conditional cdf and pdf are well defined. Then, for all* $t<s<\tau$ *we have*

$$F(t,x;\tau,y) = \int_z F(s,z;\tau,y)\frac{\partial}{\partial z}F(t,x;s,z)dz, \tag{11}$$

$$f(t,x;\tau,y) = \int_z f(s,z;\tau,y)f(t,x;s,z)dz, \tag{12}$$

Sketch of the Proof: The proof for eq. (11) mimicks closely the proof for the continuous total probability formula. Let us define the events

$$A = \{X(t) = x\}, \qquad B_i = \{X(s) \in I_i^\epsilon\}, \qquad C = \{X(\tau) \leqslant y\}.$$

where the intervals $I_i^\epsilon = [z_i, z_{i+1})$ form a fine partition of $\mathbb{R}$ as in the proof of Lemma 11.13.

Then

$$F(t,x;\tau,y) = P[X(\tau) \leqslant y | X(t) = x] = P[C|A] = P[C \cap (\bigcup_i B_i)|A] =$$
$$= \sum_i \frac{P[AB_iC]}{P[A]} = \sum_i \frac{\cancel{P[A]}P[B_i|A]P[C|B_iA]}{\cancel{P[A]}}$$

Here two remarks are in order:

- first of, technically $P[A] = 0$; this discussion makes sense through the regularisation $A^\epsilon = \{X(t) \in (x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2})\}$ and by assuming that the "$\frac{0}{0}$" limit exists. This assumption is already implicit in working with the conditional cdf. We will not talk about this further.

- To proceed from here, we use

$$P[C|B_iA] \approx P[C|B_i].$$

  A warning is in order: this is *not* literally the Markov property. The Markov property makes the future independent of the past given the *exact* present state $X(s) = z$; the event B_i only locates $X(s)$ inside a bin, and conditioning additionally on the past can, in principle, change the distribution of $X(s)$ *within* the bin – and hence the law of the future. The discrepancy disappears as the bins shrink, under our standing smoothness assumptions on the conditional densities; this is why the final formula is correct. The honest formulation of the result is the exact identity between transition densities, eq. (12), and a fully rigorous treatment proves it at that level (see [11]).

So now we have

$$F(t,x;\tau,y) \approx \sum_i P[B_i|A]P[C|B_i] =$$

$$= \sum_i P[X(s) \in [z_i, z_{i+1})|X(t) = x]P[X(\tau) \leqslant y | X(s) \in [z_i, z_{i+1})] =$$
$$= \sum_i \big(F(t,x;s,z_{i+1}) - F(t,x;s,z_i)\big) \cdot P[X(\tau) \leqslant y | X(s) \in [z_i, z_{i+1})] \approx$$
$$\approx \sum_i \tfrac{\partial}{\partial z} F(t,x;s,z)|_{\overline{z}_i} \underbrace{(z_{i+1} - z_i)}_{\to dz|_{\overline{z}_i}} \cdot$$

$$\cdot \underbrace{P[X(\tau) \leqslant y | X(s) \in [z_i, z_{i+1})]}_{\to F(s,\overline{z}_i;\tau,y)}$$

for $\overline{z}_i \in [z_i, z_{i+1})$ as $\epsilon \to 0$.

To prove equation (12), we simply observe that $\frac{\partial}{\partial z} F(t,x;s,z) = f(t,x;s,z)$ and take the partial derivative of eq. (11) in y. □

The Chapman-Kolmogorov-Markov equations can be used to derive an equation for the evolution in time of the conditional cdf's and pdf's, under additional assumptions:

Theorem 11.18 (Backward Kolmogorov equations)**.** *Consider a Markov process $X(t)$, and assume that $F(t,x;\tau,y)$ is smooth in all variables for $t<\tau$ and moreover*

(i) $\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int\limits_{|x-z|\geqslant\delta} f(t-\Delta t,x;t,z)dz=0\ \forall\delta>0$

(ii) $\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int\limits_{|x-z|<\delta} (z-x)f(t-\Delta t,x;t,z)dz=a(t,x)\ \forall\delta>0$

(iii) $\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int\limits_{|x-z|<\delta} (z-x)^2f(t-\Delta t,x;t,z)dz=b(t,x)\ \forall\delta>0.$

Then its conditional cdf satisfies the backward Kolmogorov equation

$$\frac{\partial}{\partial t}F(t,x;\tau,y)=-a(t,x)\frac{\partial}{\partial x}F(t,x;\tau,y)-\frac{b(t,x)}{2}\frac{\partial^2}{\partial x^2}F(t,x;\tau,y). \tag{13}$$

The conditional pdf also satisfies the same equation.

Proof: Observe that, by virtue of the Chapman-Kolmogorov-Markov equation for the times $t-\Delta t<t<\tau$

$$F(t-\Delta t,x;\tau,y)=\int\limits_z F(t,z;\tau,y)f(t-\Delta t,x;t,z)dz$$

and trivially

$$F(t,x;\tau,y)=F(t,x;\tau,y)\int\limits_z f(t-\Delta t,x;t,z)dz=\int\limits_z F(t,x;\tau,y)f(t-\Delta t,x;t,z)dz.$$

Now combining these two equations we get

$$\begin{aligned}\tfrac{F(t-\Delta t,x;\tau,y)-F(t,x;\tau,y)}{\Delta t}&=\tfrac{1}{\Delta t}\int\limits_z\big(F(t,z;\tau,y)-F(t,x;\tau,y)\big)f(t-\Delta t,x;t,z)dz=\\&=\underbrace{\frac{1}{\Delta t}\int\limits_{z:|x-z|\geqslant\delta}\big(F(t,z;\tau,y)-F(t,x;\tau,y)\big)f(t-\Delta t,x;t,z)dz}_{I_1}+\\&\quad+\underbrace{\frac{1}{\Delta t}\int\limits_{z:|x-z|<\delta}\big(F(t,z;\tau,y)-F(t,x;\tau,y)\big)f(t-\Delta t,x;t,z)dz}_{I_2}\end{aligned}$$

for some $\delta<1$ (to be determined later).

Now observe that, since $0\leqslant F(t_1,x_1;t_2,x_2)\leqslant 1$ for any t_1,t_2,x_1,x_2, we have

$$|I_1|\leqslant\frac{2}{\Delta t}\int\limits_{z:|x-z|\geqslant\delta} f(t-\Delta t,x;t,z)dz\to 0\qquad\text{as }\Delta t\to 0$$

by Assumption (i). Moreover, to proceed with I_2, we will Taylor expand $F(t,z;\tau,y)$ in z around $z=x$, i.e. use the expansion

$$F(t,z;\tau,y)=F(t,x;\tau,y)+(z-x)\frac{\partial}{\partial x}F(t,x;\tau,y)+\frac{1}{2}(z-x)^2\frac{\partial^2}{\partial x^2}F(t,x;\tau,y)+R,$$

where of course the Taylor remainder is of the form

$$|R| \leqslant \frac{1}{6}|z-x|^3 \left| \frac{\partial^3}{\partial x^3} F(t,x;\tau,y)|_{x=\xi} \right|$$

for some ξ between x and z. In particular δ can be chosen to be no larger than

$$\delta \leqslant \left(\frac{6}{\sup\limits_{|\xi-x|\leqslant 1} \left| \frac{\partial^3}{\partial x^3} F(t,x;\tau,y)|_{x=\xi} \right|} \right)^2 \implies |R| \leqslant |z-x|^2\sqrt{\delta}$$

So now combining everything we get

$$\frac{F(t-\Delta t,x;\tau,y)-F(t,x;\tau,y)}{\Delta t} = I_1 + \frac{1}{\Delta t}\frac{\partial}{\partial x}F(t,x;\tau,y) \int\limits_{z:|x-z|<\delta} (z-x) f(t-\Delta t,x;t,z)dz +$$
$$+\frac{1}{2\Delta t}\frac{\partial^2}{\partial x^2}F(t,x;\tau,y) \int\limits_{z:|x-z|<\delta} (z-x)^2 f(t-\Delta t,x;t,z)dz + \underbrace{\frac{1}{\Delta t}\int\limits_{z:|x-z|<\delta} R\cdot f(t-\Delta t,x;t,z)dz}_{I_3}.$$

Since $f(t-\Delta t,x;t,z)$ is a conditional pdf, its integral in z is bounded by 1, so we have

$$|I_3| \leqslant \frac{\sqrt{\delta}}{\Delta t} \int\limits_{z:|x-z|<\delta} (z-x)^2 f(t-\Delta t,x;t,z)dz \to \sqrt{\delta}\, b(t,x) \quad \text{as } \Delta t \to 0.$$

Now taking the limit $\Delta t \to 0$ we get

$$-\frac{\partial}{\partial t}F(t,x;\tau,y) = a(t,x)\frac{\partial}{\partial x}F(t,x;\tau,y) + \frac{b(t,x)}{2}\frac{\partial^2}{\partial x^2}F(t,x;\tau,y) + O\big(\sqrt{\delta}\, b(t,x)\big).$$

The left-hand side does not depend on δ, so letting $\delta \to 0$ – after $\Delta t \to 0$, in this order – completes the proof. (The supremum in the choice of δ above is over the compact set $|\xi - x| \leqslant 1$ – recall $\delta < 1$ – so smoothness of F suffices; no global bound on derivatives is needed.)

To observe that $f(t,x;\tau,y)$ satisfies the same equation, simply take the partial derivative with respect to y. □

Theorem 11.19 (Forward Kolmogorov equations)**.** *Under the assumptions of Theorem 11.18, the conditional pdf satisfies the equation*

$$\frac{\partial}{\partial \tau} f(t,x;\tau,y) + \frac{\partial}{\partial y}\left[a(\tau,y) f(t,x;\tau,y)\right] - \frac{1}{2}\frac{\partial^2}{\partial y^2}\left[b(\tau,y) f(t,x;\tau,y)\right] = 0.$$

Remarks:

- This equation is also called the Fokker-Planck equation in mathematical physics. For $a = 0$, $b = const.$ it is the heat equation (probability diffuses with no preferred direction), while for $a = const.$, $b = 0$ it describes simply a constant motion (constant deterministic rate of change, and no diffusive part).

- The proof has some more technical steps, but is based on the same idea as for Theorem 11.18; for a complete treatment see [11].
- This program, of deriving a PDE for the evolution of the conditional pdf, and solving it to understand the structure of a Markov process, can in fact be carried out, and is widely used. However, to build up such a theory to a satisfactory level, we would have to understand better the continuous limits we took (cf. Definition 11.12, Lemma 11.13 etc), and when these exist or not. The next subsection gives a formal bridge between jump processes and diffusions; a systematic treatment of these analytical questions lies beyond the scope of these notes.

11.5 From jumps to diffusions *

Finally, let us connect the two halves of this chapter. Place a birth-death process on the fine grid $X = h\mathbb{Z}$ (sites $x = nh$, $h \ll 1$), with rates chosen to be

$$\lambda(x) = \frac{b(x)}{2h^2} + \frac{a(x)}{2h}, \qquad \mu(x) = \frac{b(x)}{2h^2} - \frac{a(x)}{2h}$$

for given smooth functions $a(x)$ and $b(x) > 0$ (positive rates for h small enough – pointwise in x; a single h for all x requires e.g. $\inf b > 0$ and $|a|/b$ bounded, or simply restricting to a compact region, which is all this formal computation needs). Writing the master equation for $p(x,t)$ and setting $F(x) := \lambda(x)p(x)$, $G(x) := \mu(x)p(x)$,

$$\partial_t p(x,t) = F(x-h) + G(x+h) - F(x) - G(x) = h\big(G-F\big)'(x) + \frac{h^2}{2}\big(F+G\big)''(x) + O(h^3 \cdot h^{-2}),$$

by Taylor expansion. Since $F - G = \frac{ap}{h}$ and $F + G = \frac{bp}{h^2}$, this becomes

$$\partial_t p = -\partial_x\big(a(x)\,p\big) + \frac{1}{2}\partial_x^2\big(b(x)\,p\big) + O(h),$$

i.e. in the limit $h \to 0$ the master equation becomes the **forward Kolmogorov (Fokker-Planck) equation** of Theorem 11.19. Frequent, small jumps aggregate into a drift a and a diffusion b – and, conversely, this computation explains the meaning of the assumptions (i)-(iii) of the Kolmogorov equations in Section 11.4: they say exactly that, at vanishing scales, the process has a well-defined drift and diffusion coefficient and no macroscopic jumps.

Further reading: Norris [15] is the standard reference for continuous-time Markov chains, at a level continuous with this chapter; queues and birth-death models are developed at length in [2].

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 11.1 (Poisson process toolbox)**.** *1.* **(Superposition)** *If $N_1(t)$, $N_2(t)$ are independent Poisson processes with rates λ_1, λ_2, then $N_1(t) + N_2(t)$ is a Poisson process with rate $\lambda_1 + \lambda_2$.*

2. **(Thinning)** *Each event of a Poisson process with rate λ is kept with probability q and discarded with probability $1 - q$, independently of everything else. Show that the kept events form a Poisson process with rate $q\lambda$.*

3. Show that for $0 < s < t$,

$$P[N(s) = k \,|\, N(t) = n] = \binom{n}{k} \left(\frac{s}{t}\right)^k \left(1 - \frac{s}{t}\right)^{n-k},$$

i.e. given the total count, the events are distributed over $(0, t]$ like independent uniform points.

4. Two independent Poisson processes with rates λ, μ are running. Show that the probability that the next event (after any given time) comes from the first process is $\frac{\lambda}{\lambda+\mu}$.

Exercise 11.2. *Show that the M/M/∞ queue has stationary distribution Poisson with parameter $\frac{\lambda}{\mu}$.*

Exercise 11.3 (Pure death process)**.** *N_0 individuals are alive at $t = 0$; each dies independently at rate μ (i.e. $\mu_n = n\mu$, $\lambda_n = 0$). Show that*

$$p_n(t) = \binom{N_0}{n} e^{-n\mu t} \left(1 - e^{-\mu t}\right)^{N_0 - n},$$

either by solving the master equation or by a direct probabilistic argument (each individual is still alive at time t with probability $e^{-\mu t}$, independently).

Exercise 11.4 (M/M/2 queue)**.** *Customers arrive at rate λ and are served by* two *servers, each working at rate μ (so $\mu_1 = \mu$, $\mu_n = 2\mu$ for $n \geqslant 2$). Using detailed balance, find the stationary distribution and the condition for its existence.*

12 Brownian motion

12.1 Continuous limit of a random walk

The random process known as Brownian motion takes its name from the observation through a microscope by Robert Brown in 1827 of the movement of a grain of pollen suspended in water. The grain followed an erratic trajectory, and at the time explaining this physically was not possible.

As molecular physics matured, it was understood that the grain was colliding with the surrounding molecules of water, which were performing thermal oscillations. However, a mathematical theory systematically characterising this movement took much longer to develop. Here we will roughly follow the historical development of such a theory.

Consider a simple, symmetric random walk in 1 dimension (cf. Definition 10.8),

$$\omega^{(n)} = (\omega^{(n)}(0), \omega^{(n)}(1), \dots, \omega^{(n)}(k), \dots, \omega^{(n)}(n)) = \Big(\sum_{j=1}^{k} X_j \Big)_{k=0}^{n},$$

$$\text{where } P[X_j = 1] = P[X_j = -1] = \tfrac{1}{2} \quad \forall j,$$

and its rescaled version

$$w^{(n)}(t) := \sqrt{\frac{T}{n}}\, \omega^{(n)}(\frac{tn}{T}), \qquad \frac{nt}{T} = k.$$

By an abuse of notation we identify the vector $\big(w^{(n)}(\frac{kT}{n})\big)_{k=0}^{n}$ to its linearly-interpolated version

$$w^{(n)}(t) = \begin{cases} \sqrt{\frac{T}{n}}\, \omega^{(n)}(\frac{tn}{T}), & \text{if } t = \frac{kT}{n} \text{ for some } k \in \{0, 1, \dots, n\}, \\ (k + 1 - \frac{nt}{T})\, w^{(n)}(\frac{kT}{n}) + (\frac{nt}{T} - k)\, w^{(n)}(\frac{(k+1)T}{n}), & \\ \qquad \text{if } t \in (\frac{kT}{n}, \frac{(k+1)T}{n}) \text{ for some } k \in \{0, 1, \dots, n-1\}. & \end{cases}$$

By construction, all of the 2^n realisations $w^{(n)}(t)$ are equiprobable, i.e. each realisation appears with probability $\frac{1}{2^n}$.

Thus the $w^{(n)}(t)$ define a probability measure on $C[0, T]$: for every E Borel subset[7] of $C[0, T]$,

$$P^{(n)}[E] = \frac{\#\{\text{realisations of } w^{(n)} \text{ that belong to } E\}}{2^n}.$$

Theorem 12.1. *Let $0 < t_1 < t_2 < T$. Then, as $n \to \infty$,*

$$w^{(n)}(t_2) - w^{(n)}(t_1) \;\longrightarrow\; \mathcal{N}(0, t_2 - t_1) \qquad \textit{in distribution}$$

[7] When we talk about the Borel subsets of $C[0, T]$ we mean the smallest σ-algebra generated by open sets in the natural topology, which is generated by the sup norm. This coincides with the smallest σ-algebra containing the family of all open balls,

$$\Big\{ \{g \in C[0,T] \mid \sup_{t \in [0,T]} |f(t) - g(t)| < r\} \;:\; f \in C[0,T],\ r > 0 \Big\}.$$

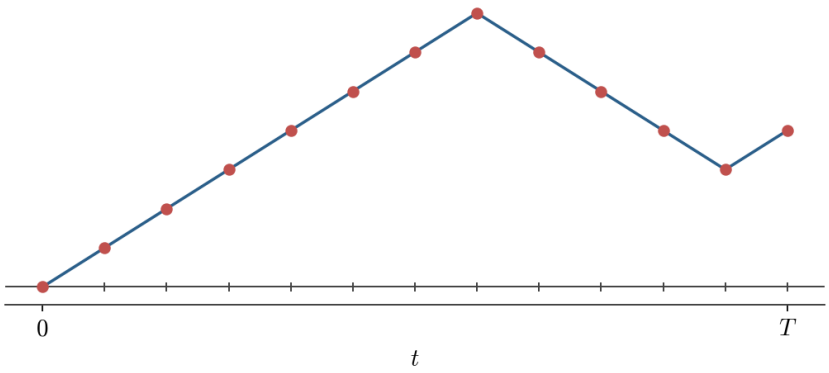


Figure 2: A random walk and the continuous function in $C[0,T]$ it defines by linear interpolation.

(note that for each finite n the increment has a discrete distribution; the Gaussian law appears only in the limit).

Proof: For simplicity we will assume that t_1, t_2 are of the form $t_j = \frac{k_j T}{n}$ for some $k_1 < k_2$. Then, by definition,

$$w^{(n)}(t_2) - w^{(n)}(t_1) = \sqrt{\frac{T}{n}} \left(\omega^{(n)}(k_2) - \omega^{(n)}(k_1) \right) = \sqrt{\frac{T}{n}} \sum_{j=k_1+1}^{k_2} X_j .$$

We will use heavily the fact that

$$k_2 - k_1 = \frac{n}{T}(t_2 - t_1) \gg 1$$

since $n \gg 1$ i.e. $n \to \infty$. Thus we have a sum of a very large number of i.i.d. random variables. This can be recast in the form

$$\begin{aligned} w^{(n)}(t_2) - w^{(n)}(t_1) &= \sqrt{\tfrac{T}{n}} \sum_{j=k_1+1}^{k_2} X_j = \sqrt{\tfrac{t_2-t_1}{k_2-k_1}} \sum_{j=1}^{k_2-k_1} X_j = \sqrt{t_2-t_1} \sqrt{\tfrac{T}{n(t_2-t_1)}} \sum_{j=1}^{n\frac{t_2-t_1}{T}} X_j = \\ &= \sqrt{t_2-t_1} \sqrt{\tfrac{n(t_2-t_1)}{T}} \tfrac{T}{n(t_2-t_1)} \sum_{j=1}^{n\frac{t_2-t_1}{T}} X_j = \sqrt{t_2-t_1} \sqrt{\tfrac{n(t_2-t_1)}{T}} \sum_{j=1}^{n\frac{t_2-t_1}{T}} \tfrac{X_j}{\frac{n(t_2-t_1)}{T}} = \\ &= \sqrt{t_2-t_1} \underbrace{\sqrt{M} \sum_{j=1}^{M} \left(\frac{X_j}{M} - \mathrm{E}[X] \right)}_{Z} \end{aligned}$$

for $M = n\frac{t_2-t_1}{T} \gg 1$ (since $n \to \infty$). By virtue of the CLT (Theorem 9.19) the random variable Z converges in distribution, as $M \to \infty$, to $\mathcal{N}(0, \mathrm{Var}[X]) = \mathcal{N}(0,1)$, since $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \frac{1}{2} \cdot 1^2 + \frac{1}{2} \cdot (-1)^2 - 0 = 1$. Note that Z itself is still a *discrete* random variable for each finite M – it does not have a density; it is the limiting variable $Z_\infty \sim \mathcal{N}(0,1)$ that has the pdf

$$f_{Z_\infty}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} .$$

Since multiplication by a constant preserves convergence in distribution, and by Lemma 8.1 $\sqrt{t_2 - t_1}\, Z_\infty \sim \mathcal{N}(0, t_2 - t_1)$ (with pdf $\frac{1}{\sqrt{2\pi(t_2-t_1)}} e^{-\frac{x^2}{2(t_2-t_1)}}$), the increment $w^{(n)}(t_2) - w^{(n)}(t_1)$ converges in distribution to $\mathcal{N}(0, t_2 - t_1)$, as claimed. □

For general (non-grid) times, take $k_j = \lfloor nt_j/T \rfloor$: then $|t_j - \frac{k_j T}{n}| \leqslant \frac{T}{n}$, and the correction $w^{(n)}(t_j) - w^{(n)}(\frac{k_j T}{n})$ – at most one increment plus a linear interpolation – has variance at most $\frac{T}{n} \to 0$, so it does not affect the limit in distribution.

Observe moreover that, by construction, if $0 \leqslant k_1 < k_2 < k_3 \leqslant n$, the random variables

$$\omega^{(n)}(k_1), \qquad \omega^{(n)}(k_2) - \omega^{(n)}(k_1), \qquad \omega^{(n)}(k_3) - \omega^{(n)}(k_2)$$

are independent. This is reflected in the independence of

$$w^{(n)}(t_1), \qquad w^{(n)}(t_2) - w^{(n)}(t_1), \qquad w^{(n)}(t_3) - w^{(n)}(t_2)$$

as long as the t_j are of the form $t_j = \frac{k_j T}{n}$ (general times are handled by the same grid approximation as above).

The random process $\{B(t)\}_{t\in[0,T]}$, known as **Brownian motion**, is obtained as the limit of $w^{(n)}(t)$ as $n \to \infty$. This is made precise in the following

What should "convergence of the measures $P^{(n)}$" mean? The standard notion is **weak convergence**: $P^{(n)} \rightharpoonup W$ iff

$$\int F\, dP^{(n)} \longrightarrow \int F\, dW$$

for every bounded functional $F : C[0,T] \to \mathbb{R}$ continuous in the sup norm. Convergence of all the finite-dimensional distributions – which is essentially what we checked above – is necessary but *not* sufficient for this; the missing ingredient, which we will not develop, is *tightness*, a compactness property ruling out probability mass escaping onto ever rougher paths.

Theorem 12.2 (Existence of the Wiener measure and Brownian motion)**.** *The probability measures $P^{(n)}$ have a weak limit*

$$P^{(n)} \rightharpoonup W$$

which is also a probability measure on $C[0,T]$. This will be called the Wiener measure.

We will not go into the proof; a complete construction of the Wiener measure can be found in Koralov and Sinai [11].

12.2 Fundamental properties of Brownian motion

It follows now that for $0 \leqslant t_1 < t_2 \leqslant T$

$$B(0) = 0, \qquad B(t_1) \sim \mathcal{N}(0, t_1), \qquad B(t_2) - B(t_1) \sim \mathcal{N}(0, t_2 - t_1).$$

Moreover, for any $0 \leqslant t_0 < t_1 < \cdots < t_r \leqslant T$ the random variables

$$B(t_{j+1}) - B(t_j), \qquad j = 0, 1, \ldots, r-1$$

are independent. This is called the **property of independent increments**, and it is crucial.

But what is the joint pdf for the evaluation of $B(t)$ at several time instants?

Lemma 12.3. *Let $0 < t_0 < t_1 < \cdots < t_r \leqslant T$. Then the vector-valued R.V.*

$$\vec{X} = (B(t_0), B(t_1), \ldots, B(t_r))$$

has pdf

$$f_{\vec{X}}(x_0, x_1, \ldots, x_r) = \prod_{j=0}^{r} \frac{1}{\sqrt{2\pi(t_j - t_{j-1})}} e^{-\frac{(x_j - x_{j-1})^2}{2(t_j - t_{j-1})}}$$

where $x_{-1} = t_{-1} = 0$.

Proof: Consider the linear mapping $H : \mathbb{R}^{r+1} \to \mathbb{R}^{r+1}$

$$H_j(\vec{x}) := x_j - x_{j-1}$$

where of course x_{-1} is again understood as 0. This H maps the values of our random process $B(t_j)$ to the **increments** $B(t_j) - B(t_{j-1})$ (where once again t_{-1} is set to zero), i.e.

$$\begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_r \end{pmatrix} := \begin{pmatrix} B(t_0) \\ B(t_1) - B(t_0) \\ \vdots \\ B(t_r) - B(t_{r-1}) \end{pmatrix} = H \begin{pmatrix} B(t_0) \\ B(t_1) \\ \vdots \\ B(t_r) \end{pmatrix} = \vec{H}(\vec{X}).$$

We know that the increments η_j are independent and Gaussian, i.e. the pdf for $\vec{\eta}$ is

$$f_{\vec{\eta}}(\eta_0, \eta_1, \ldots, \eta_r) = \prod_{j=0}^{r} \frac{1}{\sqrt{2\pi(t_j - t_{j-1})}} e^{-\frac{\eta_j^2}{2(t_j - t_{j-1})}}.$$

So for any $C \subseteq \mathbb{R}^{r+1}$,

$$P[\vec{X} \in C] = P[\vec{\eta} \in H(C)] = \int\limits_{\vec{\eta} \in \vec{H}(C)} \prod_{j=0}^{r} \frac{1}{\sqrt{2\pi(t_j - t_{j-1})}} e^{-\frac{\eta_j^2}{2(t_j - t_{j-1})}} d\eta_0 d\eta_1 \ldots d\eta_r =$$

$$= \int\limits_{\vec{X} \in C} \prod_{j=0}^{r} \frac{1}{\sqrt{2\pi(t_j - t_{j-1})}} e^{-\frac{H_j(\vec{X})^2}{2(t_j - t_{j-1})}} \left| \frac{\partial(H_0, H_1, \ldots, H_r)}{\partial(x_0, x_1, \ldots, x_r)} \right| dx_0 dx_1 \ldots dx_r =$$

$$= \int\limits_{C} \prod_{j=0}^{r} \frac{1}{\sqrt{2\pi(t_j - t_{j-1})}} e^{-\frac{(x_j - x_{j-1})^2}{2(t_j - t_{j-1})}} dx_0 dx_1 \ldots dx_r$$

since the Jacobian

$$\left|\frac{\partial(H_0, H_1, \ldots, H_r)}{\partial(x_0, x_1, \ldots, x_r)}\right| = \det \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ -1 & 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & & 0 \\ & & & & \\ 0 & 0 & \ldots & -1 & 1 \end{bmatrix} = 1.$$

□

More generally, most questions about Brownian motion can be answered by reformulating them to a question about increments.

Lemma 12.4. *Let $0 < t_1 < t_2 \leqslant T$. Then*

$$\mathrm{Cov}[B(t_1), B(t_2)] = t_1.$$

Proof: Since $\mathrm{E}[B(t)] = 0 \ \forall t \in [0, T]$, and using the property of independent increments,

$$\mathrm{Cov}[B(t_1), B(t_2)] = \mathrm{E}[B(t_1)B(t_2)] = \mathrm{E}[(B(t_2) - B(t_1) + B(t_1))B(t_1)] =$$
$$= \mathrm{E}[(B(t_2) - B(t_1))B(t_1)] + \mathrm{E}[B(t_1)^2] = \cancelto{0}{\mathrm{E}[B(t_2) - B(t_1)]}\cancelto{0}{\mathrm{E}[B(t_1)]} + \mathrm{Var}[B(t_1)] = t_1$$

□

12.3 Further properties of Brownian motion *

Theorem 12.5.

$$P[B(t) \text{ is differentiable at even one point } t_0 \in [0, T]] = 0$$

We will in fact prove a stronger statement, namely the set

$$E := \{w \in C[0, T] \,|\, \exists t_0 \in [0, T], A > 0, \gamma > \frac{1}{2} : |w(t) - w(t_0)| \leqslant A|t - t_0|^\gamma\}$$

has Wiener measure

$$W[E] = 0.$$

Remark: Recall that if the continuous function $w(t)$ is differentiable at a single point t_0, it would automatically satisfy a **local Lipschitz condition**

$$|w(t) - w(t_0)| \leqslant \big(|w'(t_0)| + 1\big)|t - t_0| \qquad \forall t : |t - t_0| < \epsilon.$$

Since t is contained in $[0, T]$, w has a maximum, and thus a Lipschitz condition is trivially true for $|t - t_0| > \epsilon$,

$$|w(t) - w(t_0)| \leqslant 2 \max_t |w(t)| \leqslant \frac{2 \max_t |w(t)|}{\epsilon}|t - t_0| \qquad \forall t : |t - t_0| \geqslant \epsilon.$$

Thus by selecting $L = \max\{|w'(t_0)| + 1, \frac{2\max_t |w(t)|}{\epsilon}\}$, w would satisfy a **global Lipschitz condition**

$$|w(t) - w(t_0)| \leqslant L|t - t_0| \qquad \forall t \in [0, T].$$

Any function satisfying a global Lipschitz condition on $[0, T]$ automatically satisfies a **global Hölder condition with exponent $\beta < 1$**, i.e.

$$|w(t) - w(t_0)| \leqslant B|t - t_0|^\beta \qquad \forall t \in [0, T].$$

The main idea is that for $|t - t_0| < 1$ and $\beta < 1$, $|t - t_0|^\beta > |t - t_0|$. Thus the local Hölder condition follows automatically, and with the same constant. Extending to a global condition is trivial. Thus

$$\{w \in C[0,T] \,|\, \exists t_0 \in [0,T] : w \text{ is differentiable at } t_0\} \subseteq$$
$$\subseteq \{w \in C[0,T] \,|\, \exists t_0 \in [0,T], A > 0, \gamma > \tfrac{1}{2} : |w(t) - w(t_0)| \leqslant A|t - t_0|^\gamma\}$$

An example of a function satisfying a Hölder condition with exponent $\beta < 1$ around a point, but not a Lipschitz condition around the same point, is $f(x) = |x|^\beta$ around $x = 0$.

The following proof is the most technical argument of these notes; it may be skipped on a first reading – but it is worth returning to, as a showcase of the measure-theoretic machinery in action.

Proof: Select

$$A > 0, \gamma > \frac{1}{2}, \text{ and } a \in (0,1) \text{ such that } \gamma(1-a) > \frac{1}{2}.$$

Moreover choose some "large" $m \in \mathbb{N}$.

Now for each $k_0 \in \{0, 1, 2, \ldots, 2^m\}$ define

$$D_m(A, k_0) := \{w \in C[0,T] \,|\, \left| w(\frac{k+1}{2^m}T) - w(\frac{k}{2^m}T) \right| \leqslant \frac{2AT^\gamma}{2^{m\gamma(1-a)}} \quad \forall k : |k - k_0| + 2 \leqslant 2^{ma}\}.$$

Claim: If for some $t_0 \in [\frac{k_0}{2^m}T, \frac{k_0+1}{2^m}T]$

$$|w(t) - w(t_0)| \leqslant A|t - t_0|^\gamma, \tag{14}$$

then $w \in D_m(A, k_0)$.

Proof of the Claim: Assume w satisfies eq. (14). Then

$$|w(\tfrac{k+1}{2^m}T) - w(\tfrac{k}{2^m}T)| \leqslant |w(\tfrac{k+1}{2^m}T) - w(t_0)| + |w(\tfrac{k}{2^m}T) - w(t_0)| \leqslant 2A\left(\tfrac{|k-k_0|+2}{2^m}T\right)^\gamma$$

where we used the fact that

$$\frac{k_0}{2^m}T \leqslant t_0 \leqslant \frac{k_0+1}{2^m}T \implies |\frac{k+1}{2^m}T - t_0|, |\frac{k}{2^m}T - t_0| \leqslant \frac{|k-k_0|+2}{2^m}T.$$

Moreover, if we restrict k so that $|k - k_0| + 2 \leqslant 2^{ma}$ as in the definition of $D_m(A, k_0)$, it

follows that

$$|w(\frac{k+1}{2^m}T) - w(\frac{k}{2^m}T)| \leqslant 2A\left(\frac{|k-k_0|+2}{2^m}T\right)^\gamma = 2A\left(\frac{|k-k_0|+2}{2^{ma}}\frac{2^{ma}}{2^m}T\right)^\gamma \leqslant \frac{2AT^\gamma}{2^{m\gamma(1-a)}}.$$

The proof of the claim is complete.

Now let us bound the probability of the event $D_m(A,k_0)$. From what we know about increments of the Brownian motion, $w(\frac{k+1}{2^m}T) - w(\frac{k}{2^m}T)$ is a Gaussian R.V. with mean 0 and variance $\frac{T}{2^m}$. So (using the argument of Figure 3)

$$P[|w(\frac{k+1}{2^m}T) - w(\frac{k}{2^m}T)| \leqslant \frac{2AT^\gamma}{2^{m\gamma(1-a)}}] \leqslant \frac{2^{\frac{m}{2}}}{\sqrt{2\pi T}}\frac{4AT^\gamma}{2^{m\gamma(1-a)}} = \underbrace{\frac{4AT^\gamma}{\sqrt{2\pi T}}}_{C_1}\frac{1}{2^{m[\gamma(1-a)-\frac{1}{2}]}}$$

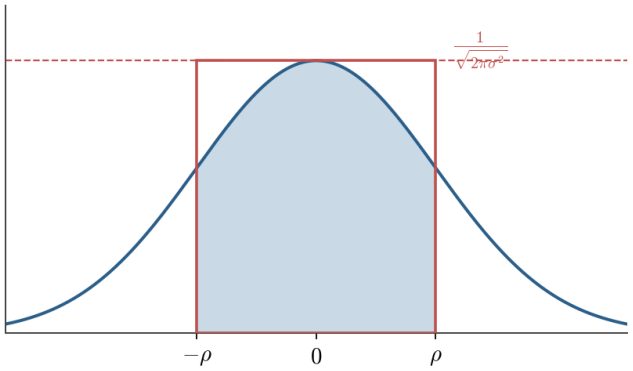


Figure 3: $\int_{x=-\rho}^{\rho} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx \leqslant 2\rho\frac{1}{\sqrt{2\pi\sigma^2}}$

However, we require that the increment "is small" for *all* admissible k, i.e. all k with $|k-k_0|+2 \leqslant 2^{ma}$ and $0 \leqslant k \leqslant 2^m - 1$. Even in the worst boundary cases (k_0 near 0 or near 2^m) these number at least $2^{ma} - 2 \geqslant 2^{ma-1}$ (for $ma \geqslant 2$, i.e. all m large enough), and the corresponding increments live on pairwise disjoint dyadic intervals, hence are *independent*. Keeping 2^{ma-1} of them,

$$P\left[|w(\tfrac{k+1}{2^m}T) - w(\tfrac{k}{2^m}T)| \leqslant \tfrac{2AT^\gamma}{2^{m\gamma(1-a)}} \quad \forall k : |k-k_0|+2 \leqslant 2^{ma}\right] \leqslant \left(C_1\frac{1}{2^{m[\gamma(1-a)-\frac{1}{2}]}}\right)^{2^{ma-1}} =$$
$$= C_1^{2^{ma-1}} e^{-m\underbrace{[\log(2)(\gamma(1-a)-\frac{1}{2})]}_{C_2>0}2^{ma-1}}$$

Now if we define

$$D_m(A) = \bigcup_{k_0} D_m(A,k_0),$$

(the events $D_m(A,k_0)$ of course depend on the exponent γ as well; we suppress this in the notation) by virtue of the claim this event includes the event E of the statement; moreover,

by the additivity of the measure,

$$P[D_m(A)] \leqslant \sum_{k_0=0}^{2^m} P[D_m(A,k_0)] \leqslant (2^m+1)\, C_1^{2^{ma-1}} e^{-C_2 m 2^{ma-1}}$$

Moreover, this is true for any m, i.e.

$$E \subseteq \bigcap_{m\in\mathbb{N}} D_m(A).$$

By virtue of the Borel-Cantelli lemma (cf. Theorem 3.10) it is sufficient to prove that $\sum_m P[D_m(A)] < \infty$; then it follows automatically that $P[\bigcap_{m\in\mathbb{N}} D_m(A)] \leqslant P[\limsup_m D_m(A)] = 0$.

Claim: $\sum_m P[D_m(A)] < \infty$.

Proof of the claim: Taking logarithms in the bound above,

$$\log P[D_m(A)] \leqslant (m+1)\log 2 \,+\, 2^{ma-1}\big(\log C_1 - C_2\, m\big).$$

Since $\log C_1 - C_2 m \to -\infty$, for all m large enough $\log C_1 - C_2 m \leqslant -1$, and then

$$\log P[D_m(A)] \leqslant (m+1)\log 2 - 2^{ma-1} \longrightarrow -\infty$$

superexponentially fast – the power 2^{ma-1} crushes the linear term. In particular $P[D_m(A)] \leqslant 2^{-m}$ for all m large enough, and the series converges by comparison with $\sum_m 2^{-m}$. The proof of the claim is complete.

Finally, one last observation is needed: the event E involves *all* $A > 0$, $\gamma > \frac{1}{2}$ – an uncountable family – while we have treated one pair (A,γ) at a time, and an uncountable union of null events need not be null. The gap is closed by reducing to a countable family: if a path satisfies the Hölder estimate for *some* $A > 0$, $\gamma > \frac{1}{2}$, then it also satisfies it with the integer constant $\lceil A\, T^{\gamma-q}\rceil$ and any rational exponent $q \in (\frac{1}{2},\gamma)$ (since $|t-t_0|^\gamma \leqslant T^{\gamma-q}|t-t_0|^q$ on $[0,T]$). Hence E is contained in the *countable* union, over $A \in \mathbb{N}$ and rational $q > \frac{1}{2}$, of the null events treated above, and $P[E] = 0$. The proof of the Theorem is complete. □

We will also list with no proof the following

Theorem 12.6. *For every* $0 < \gamma < \frac{1}{2}$

$$E := \{w \in C[0,T] \,|\, \exists A > 0 : |w(t_1) - w(t_2)| \leqslant A|t_1 - t_2|^\gamma \quad \forall t_1, t_2 \in [0,T]\}$$

has Wiener measure

$$W[E] = 1.$$

These last two results (cf. also Theorems 12.11 and 12.13 below) can be interpreted as saying that $B(t)$ has, at each point $t \in [0,T]$, more-or-less the smoothness that the function $f(t) = \sqrt{|t|}$ has at $t = 0$.

This "bad" behaviour of "typical" continuous functions ("typical" according to the Wiener measure) causes substantial difficulties, some of which we will see in detail subsequently.

Before we proceed however, we will see a more compact definition of Brownian motion that will also be helpful:

Definition 12.7 (Gaussian Random Variable)**.** *A vector valued R.V.* $\vec{X} : \Omega \to \mathbb{R}^d$ *is called a Gaussian R.V. if every linear combination* $\sum_{i=1}^d a_i X_i$, $\vec{a} \in \mathbb{R}^d$, *is a (possibly degenerate, i.e. constant) one-dimensional Gaussian random variable. Its law is completely determined by the mean* $\vec{m} = \mathrm{E}[\vec{X}]$ *and the covariance matrix* $V_{ij} = \mathrm{Cov}[X_i, X_j]$, *which is symmetric and positive* semi*definite. In the nondegenerate case, when* V *is positive definite,* $\vec{X}$ *has the pdf*

$$f_{\vec{X}}(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(V)}} e^{-\frac{(\vec{x}-\vec{m})V^{-1}(\vec{x}-\vec{m})^T}{2}}.$$

Degenerate Gaussian vectors (singular V*, law supported on a lower-dimensional subspace) are genuinely needed: e.g. any vector of Brownian values which includes* $B(0) = 0$ *has singular covariance matrix.*

Definition 12.8 (Gaussian Random Process)**.** *A random process* $\{X(t)\}_{t\in[a,b]}$ *is called Gaussian if, for any finite collection* $\{t_0, t_1, \ldots, t_r\} \subseteq [a, b]$ *the random variable*

$$\vec{Y} = (X(t_0), X(t_1), \ldots, X(t_r))$$

is Gaussian as in Definition 12.7.

Definition 12.9 (Standard and Generalised Brownian motion)**.** *A Gaussian random process* $\{B(t)\}_{t\in[0,T]}$ *with mean value*

$$m(t) = \mathrm{E}[B(t)]$$

and covariance

$$V(s,t) = \mathrm{Cov}[B(s), B(t)]$$

is called Brownian motion if it has continuous realisations (as guaranteed by the construction of the previous subsection – continuity of paths is part of the definition), and, for any ordered *finite collection of times* $0 \leqslant t_0 < t_1 < \cdots < t_r \leqslant T$, *the random variables*

$$B(t_0), \quad B(t_1) - B(t_0), \quad \ldots \quad B(t_r) - B(t_{r-1}),$$

are independent and $V(s,t) = V(\min(s,t), \min(s,t))$.

If in addition $m(t) = 0$, $V(s,t) = \min(s,t)$, $B(t)$ *is the standard Brownian motion.*

Definition 12.10 (Multi-dimensional Brownian motion)**.** *A random process*

$$\{\vec{B}(t)\}_{t\in[0,T]} = \left\{ \begin{pmatrix} B_1(t) \\ B_2(t) \\ \vdots \\ B_d(t) \end{pmatrix} \right\}_{t\in[0,T]}$$

is d*-dimensional Brownian motion if the coordinates* $B_j(t)$ *are* independent 1*-dimensional Brownian motions as in Definition 12.9; in the standard case, equivalently,* $\mathrm{E}[B_i(s)B_j(t)] =$

$\delta_{ij} \min(s,t)$. *The independence is essential: taking* $B_1 = B_2 = B$ *produces a process whose coordinates are each Brownian, but which moves only along the diagonal of the plane – not a* 2*-dimensional Brownian motion.*

Now let us look at some more properties of Brownian motion related to the p-variation. To do that, we first recall some notions from Calculus:

Recall from Section 7.2 the notions of a partition $P = \{a = t_0 < t_1 < \cdots < t_n = b\}$ of $[a,b]$, of refinement ($P_1 \leqslant P_2$ iff P_2 contains all the points of P_1), and of the p-variation $V_p(f)$ – the limit, over refinements, of $V_p(f,P) = \sum_j |f(t_{j+1}) - f(t_j)|^p$, with V_1 the total variation and V_2 the quadratic variation. In addition we will use the **step size** of a partition,

$$|P| := \max_j (t_{j+1} - t_j).$$

(That every non-decreasing function has finite total variation is Exercise 12.1.)

Theorem 12.11 (Variation of Brownian motion)**.** *Let* $\{B(t)\}_{t\in[0,T]}$ *be the standard Brownian motion, restricted on some interval* $[a,b] \subseteq [0,T]$*, and let* P_n *be any sequence of partitions of* $[a,b]$ *with* $|P_n| \to 0$*. Then the* **quadratic variation** *sums converge to* $b-a$ *in mean square,*

$$\lim_{n\to\infty} \mathrm{E}\left[\Big(\sum_j \big(B(t^n_{j+1}) - B(t^n_j)\big)^2 \;-\; (b-a)\Big)^2\right] = 0,$$

while the total variation is infinite almost surely,

$$P[V_1(B) = +\infty] = 1.$$

Proof of the first claim. The increments are independent, with $\mathrm{E}[(\Delta B_j)^2] = \Delta t_j$; hence $\mathrm{E}[\sum_j (\Delta B_j)^2] = b - a$. Moreover $\mathrm{Var}[(\Delta B_j)^2] = 2(\Delta t_j)^2$ (fourth moment of a Gaussian), so by independence

$$\mathrm{Var}\Big[\sum_j (\Delta B_j)^2\Big] = 2\sum_j (\Delta t_j)^2 \leqslant 2\,|P_n|\,(b-a) \to 0.$$

For the second claim see e.g. [11]. □

Remark 12.12 (A warning on terminology)**.** *The quadratic variation of Brownian motion is a limit* along partitions of vanishing step size*, in the mean-square sense (or in probability) – it is* not *the supremum over all partitions, nor a pathwise limit under arbitrary refinements. In fact the "total* 2*-variation"* $\sup_P \sum_j |\Delta B_j|^2$ *is infinite almost surely, exactly like* V_1*. Whenever we write* $V_2(B)$ *below, the partition-limit (quadratic variation) sense is meant.*

By taking $b-a$ arbitrarily small, this result seems to suggest that

$$\text{“}(dB(t))^2 = dt\text{”}.$$

Such an interpretation is in fact further supported by the following

Theorem 12.13. *Let $\{B(t)\}_{t\in[0,T]}$ be the standard Brownian motion, $[a,b]\subseteq[0,T]$, $P=\{t_j\}$ a partition of $[a,b]$ and $f\in C[a,b]$. Then*

$$\lim_{|P|\to 0}\sum_j f(B(t_j))(B(t_{j+1})-B(t_j))^2=\int_{t=a}^{b} f(B(t))dt \quad \text{in probability.}$$

But why do we care so much about integrals involving the Brownian motion, anyway? Two examples help motivate this question.

12.4 Introduction to Stochastic Calculus

Example 12.14. *Let us go back to the pollen grain suspended in a liquid. Robert Brown's observation is usually interpreted as saying that the displacement is Brownian motion*

$$x(t)-x(0)=\sigma B(t).$$

One can ask: how would this change in a more viscous liquid? How would viscosity (i.e. friction) effects show up? In the case of a larger particle, when the collisions with the liquid molecules are negligible, if a near-spherical object is moving with velocity $v(t)$ at time t, viscosity (friction) would be decelerating it according to

$$\frac{dv(t)}{dt}=-Cv(t) \quad \Leftrightarrow \quad dv=-Cvdt$$

where the coefficient C depends on the size of the grain and the properties of the materials.

Ornstein and Uhlenbeck observed that, if we formally interpret the aggregate effect of the collisions as acting on the velocity*,*

$$v(t)-v(0)=\sigma B(t) \quad \Rightarrow \quad \int_{s=0}^{t}\frac{dv(s)}{ds}ds=\sigma B(t)=\sigma\int_{s=0}^{t}\frac{dB(s)}{ds}ds \quad \forall t$$

then we formally get

$$\int\frac{dv}{dt}dt=\sigma\int\frac{dB}{dt}dt \quad \rightsquigarrow \quad \frac{dv}{dt}=\sigma\frac{dB}{dt} \quad \rightsquigarrow \quad dv=\sigma dB.$$

So now combining both effects (viscosity and collisions) one can propose the equation

$$dv=-Cv\,dt+\sigma dB.$$

This is a **Stochastic Differential Equation***, and we will need to give sense to it in a precise way.*

Example 12.15. *Let us assume that the price of a stock is given by*

$$\underbrace{m_0}_{\text{stable average value}}+\underbrace{B(t)}_{\text{standard Brownian motion}}.$$

Brownian motion is a model widely used to describe the fluctuations in price due to a myriad of unpredictable "random" reasons adding up. This is a reasonable model for fluctuations which are due to some people buying or selling some stock for routine reasons, i.e. the aggregate effect of several small fluctuations none of which fundamentally changes the inherent worth of the company.

If I make transactions with my stocks at times t_j, so that during $[t_j, t_{j+1})$ I hold $f(t_j)$ stocks, the profit that I make from these transactions due to the random fluctuations of the price is

$$\sum_{j=0}^{n-1} f(t_j)(B(t_{j+1}) - B(t_j)).$$

It is clear that if I would try to come up with a strategy of "as rapid transactions as needed in order to maximise profits", I would be looking at **Stochastic Riemann-Stieltjes integrals** *of the form*

$$\int_0^t f(s)dB(s).$$

To make sense of such objects, we will need to develop a calculus for random processes.

It may seem that the latter example is much less speculative than the former. After all, sums of the form

$$\sum_{j=0}^{n-1} f(t_j)(B(t_{j+1}) - B(t_j))$$

are perfectly well-defined, unlike some of the objects appearing in Example 12.14. However, in fact the integral

$$\int_0^t f(s)dB(s)$$

is no less problematic. Let us briefly look at why that is.

12.5 Why the Riemann-Stieltjes integral fails for Brownian motion

Recall from Section 7.2 the definition of the Riemann-Stieltjes integral $\int_a^b f(t)\,dg(t)$ and the two basic facts about it: for $f \in C[a,b]$ the integral exists whenever $V_1(g) < +\infty$ (and reduces to the Riemann integral $\int fg'\,dt$ when $g \in C^1$); while if $V_1(g) = +\infty$, there exist continuous integrands f for which it diverges.

The problem with the stochastic integral $\int f(t)dB(t)$ is that, for almost all ω, $V_1(B(t,\omega)) = +\infty$, cf. Theorem 12.11. So for almost all the realisations of $B(t,\omega)$, it is possible that $\int f(t)dB(t,\omega)$ diverges. Thus making sense of the random variable

$$I : \Omega \to \mathbb{R} : \omega \mapsto \int f(t)dB(t,\omega)$$

is not straightforward.

Further reading: Koralov and Sinai [11] construct the Wiener measure and develop the properties of Brownian motion in full rigour; Sinai's Princeton lecture notes [12] cover much

of this course's material in a compact, self-contained way.

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 12.1. *Prove that every non-decreasing function $f : [a, b] \to \mathbb{R}$ has finite total variation.*

Exercise 12.2 (Scaling invariance)**.** *Let $B(t)$ be standard Brownian motion and $c > 0$. Show that $Y(t) := cB(t/c^2)$ is again a standard Brownian motion.*

Exercise 12.3 (Brownian bridge)**.** *Let $X(t) := B(t) - tB(1)$ for $t \in [0, 1]$. Compute $\mathrm{E}[X(t)]$ and $\mathrm{Cov}[X(s), X(t)]$. Is X a Brownian motion? What is special about $t = 0$ and $t = 1$?*

Exercise 12.4. *(*) Compute $P[B(2) > 0 \,|\, B(1) > 0]$.*

(Hint: write $B(2) = B(1) + (B(2) - B(1))$, a sum of two independent $\mathcal{N}(0, 1)$; the answer is $\frac{3}{4}$.)

13 Mean-square Calculus for Random Processes

13.1 Introduction and Modes of Convergence

We will now establish a framework in which we can systematically discuss integrals (and derivatives) of random processes. However, before we do that, we need a systematic way to talk about convergence of random variables. To explain this, observe that

$$\int_{t=a}^{b} \underbrace{X(t,\omega)}_{\text{random process}} dt = \underbrace{I(\omega)}_{\text{random variable}}.$$

Thus, if we define a sequence $I_n(\omega)$ of Riemann sums (random variables), to discuss whether it converges to the R.V. $I(\omega)$, we need a satisfactory framework for the convergence $I_n \to I$.

We recall from Section 7.4 the modes of convergence of random variables – in distribution, in probability, in $r-$mean, almost surely, surely – and the implications between them. In this chapter the central mode is **convergence in mean square**:

$$X_n \to X \text{ in mean square (m.s.)} \iff \lim_{n\to\infty} \mathrm{E}[|X_n - X|^2] = 0,$$

which (for random variables with finite second moments) implies convergence in probability and in distribution.

13.2 Hilbert spaces of Random Variables and Random Processes

Let $(\Omega, \mathcal{E}, P)$ be a probability space. We will define the Hilbert space of square-integrable random variables as

$$\mathfrak{h} := \{X : \Omega \to \mathbb{R} \mid \mathrm{E}[X^2] < \infty\}, \tag{15}$$

equipped with the inner product

$$\langle X, Y\rangle := \mathrm{E}[XY] = \int_{\omega\in\Omega} X(\omega)Y(\omega)dP(\omega).$$

Remarks:

1. The inner product $\mathrm{E}[XY]$ is well-defined because

   $$|\mathrm{E}[XY]| \leqslant \mathrm{E}[|XY|] \leqslant \sqrt{\mathrm{E}[X^2]\,\mathrm{E}[Y^2]}.$$

   This follows from the **Cauchy-Schwarz inequality** on Ω,

   $$\left|\int_{\omega\in\Omega} X(\omega)Y(\omega)dP(\omega)\right| \leqslant \int_{\omega\in\Omega} |X(\omega)Y(\omega)|dP(\omega) \leqslant \sqrt{\int_{\omega\in\Omega} X^2(\omega)dP(\omega) \int_{\omega\in\Omega} Y^2(\omega)dP(\omega)}.$$

   In fact $\mathfrak{h}$ is just an abstract L^2 space on Ω equipped with the measure P, $\mathfrak{h} = L^2_P(\Omega)$

2. Just like for the function space $L^2[a,b]$, the Hilbert space $\mathfrak{h}$ is not exactly a space of random variables, but rather of classes of equivalence of random variables. Indeed, if

$$\|X\|_{\mathfrak{h}} := \sqrt{\mathrm{E}[X^2]}$$

is going to be a norm, it means that every random variable X with $\mathrm{E}[X^2] = 0$ is identified with the "always zero" random variable. More generally, in the context of $\mathfrak{h}$ we will identify any two R.V. with

$$\mathrm{E}[(X-Y)^2] = \int\limits_{\omega\in\Omega} (X(\omega)-Y(\omega))^2 dP(\omega) = 0$$

This implies that $P[X=Y]=1$; in particular there could be a set $E \subseteq \Omega$ with $P[E]=0$ so that $X(\omega) \neq Y(\omega)$ for $\omega \in E$.

Similarly, we can introduce the Hilbert space of square-integrable random processes $\{X(t)\}_{t\in[a,b]}$,

$$\mathfrak{H}_{[a,b]} := \left\{ \{X(t)\}_{t\in[a,b]} \mid \mathrm{E}\left[\int\limits_{t=a}^{b} X^2(t)dt\right] < \infty \right\}$$

equipped with the inner product

$$\langle X(t), Y(t)\rangle := \mathrm{E}\left[\int\limits_{t=a}^{b} X(t)Y(t)dt\right] = \int\limits_{[a,b]\times\Omega} X(t,\omega)Y(t,\omega)dtdP(\omega).$$

Remarks:

1. Later on we will see an easy way to check the square-integrability condition $\mathrm{E}\left[\int\limits_{t=a}^{b} X^2(t)dt\right] < \infty$ in practice, cf. Corollary 13.12.

2. Here we also have the Cauchy-Schwarz inequality, in the sense that

$$|\langle X(t), Y(t)\rangle| \leqslant \left|\int\limits_{[a,b]\times\Omega} X(t,\omega)Y(t,\omega)dtdP(\omega)\right| \leqslant \int\limits_{[a,b]\times\Omega} |X(t,\omega)Y(t,\omega)|\, dtdP(\omega) \leqslant$$

$$\leqslant \sqrt{\int\limits_{[a,b]\times\Omega} X^2(t,\omega)dtdP(\omega) \int\limits_{[a,b]\times\Omega} Y^2(t,\omega)dtdP(\omega)}$$

3. Similarly, we identify random processes for which

$$\mathrm{E}\left[\int\limits_{t=a}^{b} (X(t)-Y(t))^2 dt\right] = 0.$$

Now we can set up limits for random variables and calculus for random processes in **mean square sense (m.s.s.)**, i.e. taking all limits in $\mathfrak{h}$ and / or $\mathfrak{H}$.

Convention (the right mental picture): although $\mathfrak{H}_{[a,b]}$ is a convenient ambient space, the basic object of this chapter *is* the map $t \mapsto X(t) \in \mathfrak{h}$, defined for *every* $t \in [a,b]$ – and this is henceforth our standing convention for what a random process is in this chapter (with joint measurability of $(t,\omega) \mapsto X(t,\omega)$ tacitly assumed whenever products of integrals appear). Writing $X \in \mathfrak{H}_{[a,b]}$ adds the integrability requirement $\int_a^b \mathrm{E}[X(t)^2]\,dt < \infty$. M.s. continuity, m.s. differentiation and the correlation functions below are statements about the $\mathfrak{h}$-valued function of time. The distinction matters because point evaluations $X(t_0)$ are *not* well defined on the equivalence classes of $\mathfrak{H}_{[a,b]}$ alone – an issue explored in the exercise below.

Definition 13.1 (Convergence in m.s.s. and continuity in m.s.s.)**.** *Let $X_n, Y \in \mathfrak{h}$. We will say that*

$$\lim_{n\to\infty} X_n = Y \text{ in m.s.s.} \iff \lim_{n\to\infty} \mathrm{E}[(X_n - Y)^2] = 0.$$

Moreover, for a random process $\{X(t)\}_{t\in[a,b]}$ (i.e., per our convention, $X(t) \in \mathfrak{h}$ for every t), $t_0 \in [a,b]$, we will say that

$$\underbrace{\lim_{t\to t_0} X(t)}_{\text{limit of R.V.}} = Y \text{ in m.s.s.} \iff \lim_{t\to t_0} \mathrm{E}[(X(t) - Y)^2] = 0.$$

Finally, we will say that the random process $X(t)$ is **continuous in m.s.s.** *at the point $t_0 \in [a,b]$ if*

$$\lim_{t\to t_0} X(t) = X(t_0) \text{ in m.s.s.} \iff \lim_{t\to t_0} \mathrm{E}[(X(t) - X(t_0))^2] = 0.$$

These definitions are straightforward, but how do we check them in practice? The key point is that we only need to check **second moments**.

Definition 13.2 (Autocorrelation and Cross-correlation)**.** *Let $X(t), Y(t) \in \mathfrak{H}_{[a,b]}$. We will denote R_{XX} the autocorrelation*

$$R_{XX}(t_1,t_2) = \mathrm{E}[X(t_1)X(t_2)]$$

and R_{XY} the cross-correlation

$$R_{XY}(t_1,t_2) = \mathrm{E}[X(t_1)Y(t_2)].$$

Most properties of $\mathfrak{H}$ random processes can be checked / computed in terms of the auto/cross-correlation functions. This is just a function of two variables, and it is easy to measure directly in many cases – unlike the full probability measure for a random process.

Centered second moments also appear sometimes; the covariance (also called autocovariance sometimes)

$$\begin{aligned} C_{XX}(t_1,t_2) := \mathrm{Cov}[X(t_1),X(t_2)] &= \mathrm{E}[(X(t_1) - \mathrm{E}[X(t_1)])(X(t_2) - \mathrm{E}[X(t_2)])] = \\ &= R_{XX}(t_1,t_2) - \mathrm{E}[X(t_1)]\,\mathrm{E}[X(t_2)] \end{aligned}$$

and the cross-covariance

$$C_{XY}(t_1,t_2) := \operatorname{Cov}[X(t_1),Y(t_2)] = \operatorname{E}[(X(t_1)-\operatorname{E}[X(t_1)])(Y(t_2)-\operatorname{E}[Y(t_2)])] = \\ = R_{XY}(t_1,t_2) - \operatorname{E}[X(t_1)]\operatorname{E}[Y(t_2)].$$

Theorem 13.3. *A random process $X(t) \in \mathfrak{H}_{[a,b]}$ is continuous in m.s.s. at the point $t_0 \in [a,b]$ if and only if the autocorrelation $R_{XX}(t_1,t_2)$ is continuous at the point $(t_1,t_2)=(t_0,t_0)$.*

Proof: Part I: $R_{XX}(t_1,t_2)$ continuous at $(t_0,t_0) \Rightarrow X(t)$ m.s.s. continuous.

Observe that

$$\operatorname{E}[(X(t_0+h)-X(t_0))^2] = \operatorname{E}[X(t_0+h)^2] + \operatorname{E}[X(t_0)^2] - 2\operatorname{E}[X(t_0)X(t_0+h)] = \\ = R_{XX}(t_0+h,t_0+h) + R_{XX}(t_0,t_0) - 2R_{XX}(t_0+h,t_0).$$

Thus

$$\lim_{h_1,h_2\to 0} R_{XX}(t_0+h_1,t_0+h_2) = R_{XX}(t_0,t_0) \implies \lim_{h\to 0} \operatorname{E}[(X(t_0+h)-X(t_0))^2] = 0.$$

Part II: $X(t)$ m.s.s. continuous $\Rightarrow R_{XX}(t_1,t_2)$ continuous at (t_0,t_0).

Observe that

$$R_{XX}(t_0+h_1,t_0+h_2) - R_{XX}(t_0,t_0) = \operatorname{E}[X(t_0+h_1)X(t_0+h_2)] - \operatorname{E}[X(t_0)X(t_0)] = \\ = \operatorname{E}[\big(X(t_0+h_1)-X(t_0)+X(t_0)\big)X(t_0+h_2)] - \operatorname{E}[X(t_0)X(t_0)] = \\ = \operatorname{E}[\big(X(t_0+h_1)-X(t_0)\big)X(t_0+h_2)] + \operatorname{E}[X(t_0)\big(X(t_0+h_2)-X(t_0)\big)].$$

Thus, by virtue of the triangle inequality and the Cauchy-Schwarz inequality,

$$|R_{XX}(t_0+h_1,t_0+h_2) - R_{XX}(t_0,t_0)| \leqslant \\ \leqslant |\operatorname{E}[\big(X(t_0+h_1)-X(t_0)\big)X(t_0+h_2)]| + |\operatorname{E}[X(t_0)\big(X(t_0+h_2)-X(t_0)\big)]| \leqslant \\ \leqslant \sqrt{\operatorname{E}[\big(X(t_0+h_1)-X(t_0)\big)^2]\operatorname{E}[X(t_0+h_2)^2]} + \sqrt{\operatorname{E}[\big(X(t_0+h_2)-X(t_0)\big)^2]\operatorname{E}[X(t_0)^2]}$$

To proceed we will use the following

Claim: If $X(t)$ is m.s.s. continuous at t_0, then $\operatorname{E}[X(t_0+h_2)^2]$ is bounded for $|h_2|$ small enough.

Using this claim, and the fact that $\lim_{h\to 0} \operatorname{E}[\big(X(t_0+h_2)-X(t_0)\big)^2] = 0$ by assumption, it follows that

$$\lim_{h_1,h_2\to 0} |R_{XX}(t_0+h_1,t_0+h_2) - R_{XX}(t_0,t_0)| = 0.$$

Proof of the claim:

$$\operatorname{E}[X(t_0+h_2)^2] = \operatorname{E}[\big(X(t_0+h_2)-X(t_0)+X(t_0)\big)^2] = \\ = \operatorname{E}[\big(X(t_0+h_2)-X(t_0)\big)^2] + \operatorname{E}[X(t_0)^2] + 2\operatorname{E}[\big(X(t_0+h_2)-X(t_0)\big)X(t_0)]$$

and thus, by virtue of the Cauchy-Schwarz inequality,

$$|\,\mathrm{E}[X(t_0+h_2)^2]| \leqslant \underbrace{\mathrm{E}[\big(X(t_0+h_2)-X(t_0)\big)^2]}_{\to 0} + \underbrace{\mathrm{E}[X(t_0)^2]}_{\text{independent of } h} \; + \\ +2\sqrt{\underbrace{\mathrm{E}[\big(X(t_0+h_2)-X(t_0)\big)^2]}_{\to 0}\;\underbrace{\mathrm{E}[X(t_0)^2]}_{\text{independent of } h}}\,.$$

□

Remark 13.4. *Under our standing convention every value $X(t_0)$ is an element of $\mathfrak{h}$ from the outset – there is nothing to prove there. The subtle point is different: changing a process at a* single *time instant does not change it as an element of $\mathfrak{H}_{[a,b]}$ (the two processes differ on a set of product measure zero), so "the value at t_0" is meaningful for the $\mathfrak{h}$-valued map but* not *for the $\mathfrak{H}_{[a,b]}$-equivalence class. M.s. continuity is what ties the two views together: it pins down $X(t_0)$ as the $\mathfrak{h}$-limit of the neighbouring values, i.e. continuous representatives are unique where they exist. Exercise 13.1 in the problem set explores exactly this.*

Example 13.5 (M.s.s. continuity is not "continuity of paths".)**.** *Consider the random process*

$$X(t,\omega) = \begin{cases} 0, & t < Y(\omega) \\ 1, & t \geqslant Y(\omega) \end{cases}$$

where the continuous R.V. Y has exponential pdf $f_Y(y) = e^{-y}\chi_{[0,+\infty)}(y)$.

Let us first compute the pdf for the discrete random variable $X(t)$ for some fixed t. Clearly, $X(t)$ can only take the values 0 and 1, and

$$P[X(t)=0] = P[t<Y] = \int_t^{+\infty} e^{-x}dx = e^{-t},$$

and thus

$$P[X(t)=1] = 1-e^{-t},$$

leading to a mean value

$$\mathrm{E}[X(t)] = 0\cdot P[X(t)=0] + 1\cdot P[X(t)=1] = 1-e^{-t}.$$

Moreover, the autocorrelation function for $X(t)$ is

$$R_{XX}(t_1,t_2) = \mathrm{E}[X(t_1)X(t_2)] = 0\cdot P[X(t_1)=X(t_2)=0] + 0\cdot P[X(t_1)=1, X(t_2)=0] + \\ +0\cdot P[X(t_1)=0, X(t_2)=1] + 1\cdot P[X(t_1)=X(t_2)=1] = P[\{t_1 \geqslant Y\}\cap\{t_2\geqslant Y\}] = \\ = P[\min(t_1,t_2)\geqslant Y] = \int_0^{\min(t_1,t_2)} e^{-x}dx = 1-e^{-\min(t_1,t_2)}.$$

Thus, while all realisations of this R.P. are **not** *continuous functions, the autocorrelation is continuous and hence the R.P. is nevertheless m.s.s. continuous.*

One can moreover prove the following

Theorem 13.6. *If $f : \mathbb{R} \to \mathbb{R}$ is a continuous function of at most linear growth, $|f(x)| \leqslant C(1 + |x|)$, and $X(t) \in \mathfrak{H}_{[a,b]}$ is a m.s.s. continuous R.P., then $f(X(t)) \in \mathfrak{H}_{[a,b]}$ is a m.s.s. continuous R.P.*

Definition 13.7 (M.s. derivative)**.** *A random process $X(t)$ is m.s. differentiable at t if the limit*

$$X'(t) := \lim_{h\to 0} \frac{X(t+h) - X(t)}{h} \qquad \text{exists in the m.s. sense.}$$

Theorem 13.8. *If the mixed derivative $\frac{\partial^2}{\partial t_1 \partial t_2} R_{XX}(t_1, t_2)$ exists and is continuous at (t, t), then X is m.s. differentiable at t, with $\mathrm{E}[(X'(t))^2] = \frac{\partial^2}{\partial t_1 \partial t_2} R_{XX}(t, t)$. If, moreover, the mixed derivative exists and is continuous on all of $[a, b]^2$, then $X'(t)$ exists for every t and the full two-time identity holds:*

$$R_{X'X'}(t_1, t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} R_{XX}(t_1, t_2).$$

(Mind the geography of the hypotheses: the identity at (t_1, t_2) requires m.s. differentiability at both *times, so the pointwise hypothesis buys only the diagonal statement.)*

The proof uses the Cauchy criterion of Lemma 13.9 below, in exactly the same way as for the stochastic integrals of the next section: everything is checked on second moments. We will use m.s. differentiation in Chapter 14, where for stationary processes the criterion takes a particularly clean spectral form.

Before we move on to integrals, we need to elaborate somewhat the convergence in our Hilbert spaces:

Lemma 13.9. *Let X_n, $n \in \mathbb{N}$ be a sequence of random variables in $\mathfrak{h}$. The sequence X_n has a limit in $\mathfrak{h}$ if and only if*

$$\lim_{m,n\to\infty} \mathrm{E}[\big(X_n - X_m\big)^2] = 0.$$

This is a direct consequence of the completeness of $\mathfrak{h}$.

13.3 M.s.s. stochastic Riemann integrals

Let $X(t) \in \mathfrak{H}_{[a,b]}$, and $P = \{t_j\}$ a partition of $[a, b]$. Define

$$I(X, P) := \sum_j X(t_j)(t_{j+1} - t_j).$$

For each P, this is now a random variable in $\mathfrak{h}$ (check!!). Let us choose a sequence of partitions $P_n = \{t_j^n\}$ with step sizes that go to zero, $\lim_{n\to\infty} |P_n| = 0$. When does the sequence

$$I_n := I(X, P_n) = \sum_j X(t_j^n)(t_{j+1}^n - t_j^n)$$

have a limit in n? By virtue of Lemma 13.9 we only need to check

$$\mathrm{E}[(I_n - I_m)^2] = \mathrm{E}[I_n^2] + \mathrm{E}[I_m^2] - 2\,\mathrm{E}[I_n I_m].$$

Clearly, this limit is zero if and only if $\lim_{n,m\to\infty} \mathrm{E}[I_n I_m]$, $\lim_{n\to\infty} \mathrm{E}[I_n^2]$ exist and are equal. But

$$\mathrm{E}[I_n^2] = \mathrm{E}[\big(\sum_j X(t_j^n)(t_{j+1}^n - t_j^n)\big)^2] =$$
$$= \mathrm{E}[\sum_{i,j} X(t_i^n)X(t_j^n)(t_{j+1}^n - t_j^n)(t_{i+1}^n - t_i^n)] = \sum_{i,j} R_{XX}(t_i^n, t_j^n)(t_{j+1}^n - t_j^n)(t_{i+1}^n - t_i^n).$$

But this is the Riemann sum for the 2−dimensional integral

$$\int_{t_1=a}^{b}\int_{t_2=a}^{b} R_{XX}(t_1,t_2)dt_1dt_2$$

generated by the partition P_n for t_1 and t_2. One can similarly work out the $\mathrm{E}[I_n I_m]$ term as the Riemann sum on the partition P_n for t_1 and P_m for t_2.

Thus the limit exists (and furthermore is independent of the choice of partitions) as long as the Riemann integral $\int_{t_1=a}^{b}\int_{t_2=a}^{b} R_{XX}(t_1,t_2)dt_1dt_2$ exists.

This approach leads to the following

Definition 13.10. *Given R.P. $X(t) \in \mathfrak{H}$ the stochastic Riemann integral $\int_{t=a}^{b} X(t)dt = I \in \mathfrak{h}$ exists if*

$$\forall \epsilon > 0 \quad \exists \text{ partition } P_\epsilon \ : \ \|I - \sum_j X(t_j)(t_{j+1} - t_j)\|_{\mathfrak{h}} < \epsilon \quad \forall \text{ partitions } P = \{t_j\} \supseteq P_\epsilon.$$

Theorem 13.11. *Given R.P. $X(t) \in \mathfrak{H}$ with autocorrelation R_{XX} continuous on $[a,b]^2$ (more generally: Riemann integrable), the stochastic Riemann integral $\int_{t=a}^{b} X(t)dt = I \in \mathfrak{h}$ exists as in Definition 13.10, and*

$$\mathrm{E}[I^2] = \int_{t_1=a}^{b}\int_{t_2=a}^{b} R_{XX}(t_1,t_2)\, dt_1dt_2.$$

Corollary 13.12. *The criterion of square integrability*

$$\mathrm{E}\left[\int_{t=a}^{b} X^2(t)dt\right] = \int_{[a,b]\times\Omega} X^2(t,\omega)dtdP(\omega) < \infty$$

can be reformulated as

$$\int_{t=a}^{b} R_{XX}(t,t)\, dt < \infty.$$

Observe how much simpler it is to check the new formulation (essentially only a second moment is needed) versus the original formulation (which involves the whole probability measure).

Exercises

Several of the problems below are referenced at the corresponding points of the text.

Exercise 13.1. *Consider the following R.P. for* $t \in [0,1]$ *:*

$$X_1(t,\omega) = \begin{cases} tY, & t > 0, \\ Y, & t = 0 \end{cases}$$

and

$$X_2(t,\omega) = \begin{cases} tY, & t > 0, \\ -Y, & t = 0 \end{cases}$$

where Y *is an exponential R.V.,* $f_Y(y) = e^{-y}\chi_{[0,+\infty)}(y)$.

1. *Prove that* $X_1(t), X_2(t) \in \mathfrak{H}_{[0,1]}$.
2. *Prove that* $X_1(t) = X_2(t)$ *in* $\mathfrak{H}_{[0,1]}$, *i.e.* $\|X_1(t) - X_2(t)\|_{\mathfrak{H}_{[0,1]}} = 0$. *Is the random variable* $X_1(0,\omega)$ *the same as* $X_2(0,\omega)$ *in* $\mathfrak{h}$*?*
3. *Compute* $\mathrm{E}[X_j(t)]$, $R_{X_jX_j}(t_1,t_2) = \mathrm{E}[X_j(t_1)X_j(t_2)]$ *for* $j \in \{1,2\}$. *Is* $R_{X_jX_j}(t_1,t_2)$ *continuous on every point on its diagonal* $\{t_1 = t_2\}$*? Use the answer to the previous question to conclude for which* $t \in [0,1]$ *is* $X_j(t)$ *continuous.*

The point of the problem is that, since in $\mathfrak{H}_{[0,1]}$ *we lump together different random processes into equivalence classes, for* $X(t_0)$ *to be a well-defined R.V. it has to be the same*[8] *no matter which of the elements of the equivalence class we choose.*

Exercise 13.2. *Let* A, B *be uncorrelated random variables with mean zero and variances* σ_A^2, σ_B^2, *and consider the random process* $X(t) = A + Bt$. *Compute* $R_{XX}(t_1,t_2)$. *Show that* X *is m.s. continuous and m.s. differentiable, and identify its m.s. derivative.*

Exercise 13.3. *For the Poisson process* $N(t)$ *with rate* λ, *show that*

$$R_{NN}(s,t) = \lambda \min(s,t) + \lambda^2 st.$$

Conclude that $N(t)$ *is m.s. continuous – even though every realisation has jumps (compare with the example on m.s. continuity vs. continuity of paths).*

[8] The same in $\mathfrak{h}$, i.e. there is room for differences on sets of measure zero.

14 Stationary processes and spectral representation

Many signals encountered in practice – the wind-generated waves of Chapter 1, electronic noise, turbulent velocities – look statistically "the same at all times": there is no drift, no beginning and no end, only sustained random fluctuation. This chapter develops the mathematical form of that statement, and its single most important consequence: such processes can be decomposed over frequencies, with the Fourier transform of Chapter 9 once again doing the work. Our treatment follows Sinai's Princeton lectures [12] and the book that grew out of them, Koralov and Sinai [11].

14.1 Stationarity

Definition 14.1 (Stationarity)**.** *A random process $\{X(t)\}_{t\in\mathbb{R}}$ is strictly stationary if all its finite-dimensional distributions are invariant under time shifts: for every r, times $t_1,\dots,t_r$ and shift h,*

$$(X(t_1+h),\dots,X(t_r+h)) \sim (X(t_1),\dots,X(t_r)).$$

It is wide-sense stationary (WSS) if it has finite second moments,

$$\mathrm{E}[X(t)] = m \quad \text{(constant)}, \qquad \text{and} \qquad R_{XX}(t_1,t_2) \text{ depends only on } \tau := t_2 - t_1,$$

in which case we write $R(\tau) := R_{XX}(t, t+\tau)$.

Strict stationarity (plus finite second moments) implies wide-sense stationarity; the converse fails in general, *but holds for Gaussian processes*, since their finite-dimensional distributions are completely determined by the mean and the covariance (cf. Definition 12.7). For a systematic treatment of both notions see [11, Chapters 15–16].

Example 14.2 (The random-phase oscillator)**.** *Let $A, \nu_0 > 0$ be constants and $\Theta \sim \mathcal{U}(0, 2\pi)$. Then*

$$X(t) = A\cos(2\pi\nu_0 t + \Theta)$$

is WSS (in fact strictly stationary): $\mathrm{E}[X(t)] = \frac{A}{2\pi}\int_0^{2\pi}\cos(2\pi\nu_0 t+\theta)\,d\theta = 0$, and by the product-to-sum formula

$$\begin{aligned} R(t,t+\tau) &= A^2\,\mathrm{E}[\cos(2\pi\nu_0 t+\Theta)\cos(2\pi\nu_0(t+\tau)+\Theta)] = \\ &= \tfrac{A^2}{2}\,\mathrm{E}[\cos(2\pi\nu_0\tau) + \cos(2\pi\nu_0(2t+\tau)+2\Theta)] = \tfrac{A^2}{2}\cos(2\pi\nu_0\tau), \end{aligned}$$

the double-frequency term averaging to zero over the uniform phase. Note that the randomness is minimal (one random variable), yet the process is genuinely stationary – and it will be the building block of Section 14.6.

Example 14.3. *Brownian motion is* not *stationary: $R_{BB}(s,t) = \min(s,t)$ is not a function of $t-s$ (and $\mathrm{Var}[B(t)] = t$ grows). Its* increments*, however, are stationary. The Ornstein–Uhlenbeck process – Brownian motion with restoring friction, cf. Example 12.14 and Chapter*

15 – is *stationary when started appropriately, with*

$$m = 0, \qquad R(\tau) = \frac{\sigma^2}{2C} e^{-C|\tau|};$$

we will take this as given here and verify it in Chapter 15.

14.2 Properties of the autocorrelation function

Lemma 14.4. *Let $X(t)$ be WSS (real-valued) with autocorrelation $R(\tau)$. Then*

1. $R(0) = \mathrm{E}[X(t)^2] \geqslant 0$;
2. $R(-\tau) = R(\tau)$;
3. $|R(\tau)| \leqslant R(0)$;
4. *R is positive semidefinite: for any $t_1, \dots, t_r$ and $c_1, \dots, c_r \in \mathbb{R}$, $\sum_{i,j} c_i c_j R(t_i - t_j) \geqslant 0$;*
5. *X is m.s. continuous (at every t) if and only if R is continuous at $\tau = 0$.*

Proof: (1) and (2) are immediate from the definition. (3) is Cauchy-Schwarz: $|\mathrm{E}[X(t)X(t+\tau)]| \leqslant \sqrt{\mathrm{E}[X(t)^2]\,\mathrm{E}[X(t+\tau)^2]} = R(0)$. For (4),

$$0 \leqslant \mathrm{E}\Big[\Big(\sum_i c_i X(t_i)\Big)^2\Big] = \sum_{i,j} c_i c_j R(t_i - t_j).$$

(5) is the m.s. continuity criterion of Chapter 13 specialised to the stationary case: $\mathrm{E}[(X(t+h) - X(t))^2] = 2\big(R(0) - R(h)\big)$. □

14.3 Spectral density and the Wiener-Khinchin theorem

Definition 14.5 (Spectral density)**.** *Let $X(t)$ be WSS with $m = 0$ and $R \in L^1(\mathbb{R})$. The spectral density of the process is the Fourier transform of the autocorrelation,*

$$S(\nu) := \int_\tau e^{-2\pi i \nu \tau} R(\tau)\, d\tau = \widehat{R}(\nu),$$

so that, by Fourier inversion (Lemma 9.13; legitimate e.g. when $S \in L^1$ and R is continuous – in full generality the spectral representation of a WSS autocorrelation is Bochner's theorem, of which the Wiener-Khinchin framework below is the absolutely continuous case),

$$R(\tau) = \int_\nu e^{2\pi i \nu \tau} S(\nu)\, d\nu, \qquad \text{and in particular} \qquad \mathrm{E}[X^2] = R(0) = \int_\nu S(\nu) d\nu.$$

The last identity is the point: the total power of the process is distributed over frequencies with density $S(\nu)$. For this interpretation to be consistent, S had better be nonnegative:

Theorem 14.6 (Wiener-Khinchin)**.** *In the setting of Definition 14.5, S is real, even, and*

$$S(\nu) \geqslant 0 \qquad \forall \nu.$$

Proof sketch: Real and even follow from R real and even. For the positivity, compute the power of the "windowed Fourier coefficient" of the process: using Fubini (formally – this is where a full proof needs care) and stationarity,

$$0 \leqslant \tfrac{1}{T}\,\mathrm{E}\left[\left|\int_{t=0}^{T} e^{-2\pi i\nu t}X(t)dt\right|^2\right] = \tfrac{1}{T}\int_{t=0}^{T}\int_{s=0}^{T} e^{-2\pi i\nu(t-s)}R(t-s)\,dt\,ds =$$
$$= \int_{\tau=-T}^{T}\Big(1-\tfrac{|\tau|}{T}\Big)e^{-2\pi i\nu\tau}R(\tau)d\tau \underset{T\to\infty}{\longrightarrow} S(\nu),$$

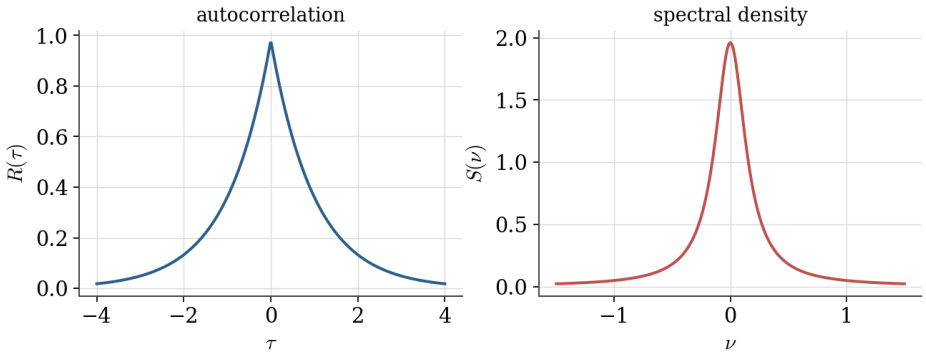
by dominated convergence. □

A complete proof – including the general case, where the spectral measure is not absolutely continuous and positivity is expressed through Bochner's theorem – can be found in [12] and [11].

Example 14.7 (The Ornstein-Uhlenbeck pair)**.** *For* $R(\tau) = \frac{\sigma^2}{2C}e^{-C|\tau|}$*:*

$$S(\nu) = \frac{\sigma^2}{2C}\int_{\tau} e^{-2\pi i\nu\tau}e^{-C|\tau|}d\tau = \frac{\sigma^2}{2C}\cdot\frac{2C}{C^2+4\pi^2\nu^2} = \frac{\sigma^2}{C^2+4\pi^2\nu^2},$$

a **Lorentzian** *spectrum: flat for* $|\nu| \ll C$*, decaying like* ν^{-2} *beyond the cutoff frequency* $\sim C$*.*



Remarks:

- For the random-phase oscillator, $R(\tau) = \frac{A^2}{2}\cos(2\pi\nu_0\tau)$ is not in L^1, but its FT makes sense distributionally (Chapter 9 again):

$$S(\nu) = \frac{A^2}{4}\big(\delta(\nu-\nu_0)+\delta(\nu+\nu_0)\big)$$

  – **spectral lines**: all the power sits at the frequency ν_0.

- **White noise** is the idealisation $S(\nu) \equiv \sigma^2$ (all frequencies equally represented, like white light), i.e. $R(\tau) = \sigma^2\delta(\tau)$: values at different times are uncorrelated, and the total power $\int S\,d\nu$ is infinite. This is not an honest process with finite second moments, but a very useful idealisation; morally it is "the derivative of Brownian motion" (recall $(dB)^2 = dt$, Theorem 12.13), a statement made precise by the Itô calculus of Chapter 15.

14.4 Linear filters

Let $h \in L^1(\mathbb{R})$ and let $X(t)$ be WSS ($m = 0$, R_{XX} bounded and continuous). Define the filtered process

$$Y(t) = \int_s h(s)\, X(t-s)\, ds,$$

the integral being the m.s. stochastic Riemann integral of Chapter 13. (It exists here: on bounded windows by continuity of R_{XX}; and the truncations $\int_{|s|\leqslant T}$ are Cauchy in $\mathfrak{h}$ as $T \to \infty$, because $|R_{XX}| \leqslant R_{XX}(0)$ and $h \in L^1$.) Filters of this form model any linear, time-invariant transformation: electrical circuits, moving averages, differentiation, the response of a linear structure to random forcing.

Theorem 14.8 (Filtered spectra)**.** *Y is WSS, and*

$$R_{YY}(\tau) = \int_s \int_u h(s)h(u)R_{XX}(\tau + s - u)\, ds\, du, \qquad S_{YY}(\nu) = |\widehat{h}(\nu)|^2\, S_{XX}(\nu).$$

Proof sketch: The formula for R_{YY} is a direct computation (expectations commute with the m.s. integrals). Taking Fourier transforms and using the convolution rules (Lemma 9.14) together with $\widehat{h(-\cdot)}(\nu) = \overline{\widehat{h}(\nu)}$ for real h:

$$S_{YY} = \widehat{h} \cdot \overline{\widehat{h}} \cdot S_{XX} = |\widehat{h}|^2 S_{XX}.$$

□

Example 14.9. *1.* **Differentiation.** *Formally $h = \delta'$, $\widehat{h}(\nu) = 2\pi i\nu$ (Lemma 9.3), so*

$$S_{X'X'}(\nu) = 4\pi^2\nu^2\, S_{XX}(\nu).$$

In particular the m.s. derivative of a WSS process exists iff $\int \nu^2 S(\nu)d\nu < \infty$: differentiation amplifies high frequencies, and a process with a heavy-tailed spectrum (e.g. the Lorentzian, with $\int \nu^2 S = \infty$) is m.s. continuous but not *m.s. differentiable.*

2. **Exponential smoothing (RC filter).** *For $h(s) = e^{-Cs}\chi_{[0,\infty)}(s)$, $|\widehat{h}(\nu)|^2 = \frac{1}{C^2+4\pi^2\nu^2}$. Feeding in white noise, $S_{XX} \equiv \sigma^2$, produces exactly the Lorentzian spectrum of Example 14.7:* the Ornstein-Uhlenbeck process is low-pass-filtered white noise.

Remark (the spectral representation). What we have used here is the spectral description of the *second moments*. There is a deeper statement lying underneath: the process itself can be represented as a stochastic integral over frequencies,

$$X(t) = \int_\nu e^{2\pi i\nu t}\, dZ(\nu),$$

where Z is a process with orthogonal increments satisfying $\mathrm{E}[|dZ(\nu)|^2] = S(\nu)d\nu$ – the rigorous, continuum version of the random-amplitude superpositions of Section 14.6. This *spectral representation theorem* is developed in [12] and [11, Chapter 15].

14.5 The Hilbert space of a stationary process; ergodicity *

This section presents, following Sinai [12], the geometric picture underneath everything in this chapter.

Let $X(t)$ be WSS with $m = 0$, and let $\mathfrak{H}_X$ be the closure in $\mathfrak{h}$ of the (complexified) linear span of the values of the process – on the complexification the inner product is $\langle X, Y\rangle := \mathrm{E}[X\overline{Y}]$ –,

$$\mathfrak{H}_X := \overline{\Big\{ \textstyle\sum_j c_j X(t_j) \Big\}} \subseteq \mathfrak{h}.$$

The shift is a unitary operator. For $h \in \mathbb{R}$ define U_h on the span by $U_h X(t) := X(t+h)$. By stationarity,

$$\langle U_h X(t), U_h X(s)\rangle = R(t-s) = \langle X(t), X(s)\rangle,$$

so U_h preserves inner products and extends to a **unitary operator** on $\mathfrak{H}_X$ – wide-sense stationarity *is* the statement that time evolution acts unitarily on the Hilbert space of the process. (For strictly stationary processes there is a parallel statement one level deeper: the time shift preserves the probability measure P itself; this is the starting point of ergodic theory [12].)

The fundamental isometry. Consider the map

$$X(t) \;\longmapsto\; e^{2\pi i \nu t} \;\in\; L^2\big(\mathbb{R}, S(\nu)d\nu\big).$$

It preserves inner products, since

$$\langle X(t), X(s)\rangle = R(t-s) = \int_\nu e^{2\pi i\nu(t-s)} S(\nu)d\nu = \big\langle e^{2\pi i\nu t}, e^{2\pi i\nu s}\big\rangle_{L^2(S\,d\nu)},$$

and therefore extends to an isometry of $\mathfrak{H}_X$ onto $L^2(S\,d\nu)$ (onto: the exponentials are dense). Under this isometry the shift U_h becomes *multiplication by* $e^{2\pi i\nu h}$: the spectral density is precisely the spectral decomposition, in the sense of operator theory, of the unitary time evolution. All the operations of this chapter become transparent on the other side of the isometry: linear filters act as multiplication by $\widehat{h}(\nu)$ (Theorem 14.8), the m.s. derivative as multiplication by $2\pi i\nu$, and the spectral representation $X(t) = \int e^{2\pi i\nu t} dZ(\nu)$ is the isometry read from right to left.

Ergodicity. For strictly stationary processes, **Birkhoff's ergodic theorem** [12] guarantees that time averages

$$\frac{1}{T}\int_0^T f\big(X(t+\cdot)\big)\,dt$$

converge almost surely as $T \to \infty$; the process is called **ergodic** if the limit is deterministic and equal to the ensemble average $\mathrm{E}[f]$ – equivalently, if every shift-invariant event has probability 0 or 1. For *Gaussian* stationary processes there is a clean criterion, due to the interplay with the isometry above:

Theorem 14.10 (cf. [12])**.** *A stationary Gaussian process is ergodic if and only if its spectral measure has no atoms (in particular: whenever it has a spectral density).*

The mechanism behind the "only if" is instructive: an atom of the spectral measure at ν_0 is a periodic component that never decorrelates. Concretely, in a superposition $\sum_j A_j \cos(2\pi\nu_j t + \Theta_j)$ with *random* independent amplitudes, the time average of X^2 converges to $\sum_j \frac{A_j^2}{2}$ – a *random* quantity, not $\mathrm{E}[X^2]$: observing one realisation forever still does not reveal the ensemble. Spectral lines break ergodicity; a spectrum without atoms restores it. Note the precise hierarchy here: for stationary Gaussian processes, absence of atoms is exactly ergodicity, and is equivalent to decorrelation *in the averaged sense* $\frac{1}{T}\int_0^T R(\tau)^2 d\tau \to 0$; the stronger pointwise decay $R(\tau) \to 0$ – automatic when a spectral *density* exists, by the Riemann-Lebesgue lemma – corresponds to the stronger *mixing* property. Ergodicity and asymptotic decorrelation are related, but they are not the same thing.

14.6 Application: synthesis of random seas; ergodicity

Superpose random-phase oscillators (Example 14.2) with independent phases Θ_j:

$$X(t) = \sum_{j=1}^{J} a_j \cos(2\pi\nu_j t + \Theta_j) \quad \Longrightarrow \quad R(\tau) = \sum_{j=1}^{J} \frac{a_j^2}{2} \cos(2\pi\nu_j \tau),$$

a WSS process whose spectrum consists of lines of weight $\frac{a_j^2}{4}$ at $\pm\nu_j$. Now let the frequencies fill a grid of *positive* frequencies $\nu_j > 0$ with spacing $\Delta\nu$ and choose the amplitudes from a *target spectrum* $S(\nu)$,

$$a_j^2 = 4\, S(\nu_j)\, \Delta\nu :$$

(the factor 4 matches the *two-sided* spectrum used throughout this chapter, where each cosine carries mass $\frac{a_j^2}{4}$ at $+\nu_j$ *and* at $-\nu_j$; the engineering literature usually works with the one-sided spectrum $S^+ := 2S$ over $\nu > 0$, for which the familiar formula $a_j^2 = 2\, S^+(\nu_j)\Delta\nu$ holds) the result is a stationary process whose spectrum approximates S. This is precisely how the wind-wave records shown in Chapter 1 were synthesised, and how "sea states" are specified in ocean engineering: empirical spectra (e.g. of JONSWAP type) are fitted to wave measurements, and time series are generated from them for simulation and design. For large J, provided no single component dominates, the CLT moreover makes such superpositions approximately Gaussian processes. (For *finite* J the random-phase superposition is not Gaussian, so the ergodicity criterion of Section 14.5 – a statement about Gaussian processes – does not apply to it verbatim.)

Remark (ergodicity): in practice R and S are estimated from a *single* long record, replacing the ensemble average $\mathrm{E}[X(t)X(t+\tau)]$ by the time average $\frac{1}{T}\int_0^T X(t)X(t+\tau)dt$. For this to converge to $R(\tau)$, time averages must reproduce ensemble averages – an **ergodicity** property – the continuous-time analogue of the law of large numbers for Markov chains (Chapter 10), discussed in Section 14.5. In the light of the theorem stated there, the minimal requirement is a spectrum without atoms; in practice one asks for the stronger mixing property $R(\tau) \to 0$ (automatic when a spectral density exists), so that distant stretches of the record behave like nearly independent samples.

Further reading: The primary mathematical references are Sinai's lectures [12] and Koralov

and Sinai [11, Chapters 15–16], which develop the spectral theory of stationary processes in full, including the spectral representation theorem and ergodic theorems. Kobayashi, Mark and Turin [2] develop spectral analysis with signal-processing and communications applications; for the spectral description of ocean waves and the stochastic modelling of wave systems, see Athanassoulis [13].

Exercises

Exercise 14.1. *1. Complete the computations of Example 14.2, and show that the superposition of Section 14.6 (with independent uniform phases) is WSS with the stated autocorrelation.*

2. Verify the Fourier transform pair of Example 14.7.

3. Let X be WSS and $Y(t) = X(t+a) - X(t)$ for fixed $a > 0$. Show that $S_{YY}(\nu) = 4\sin^2(\pi\nu a)\, S_{XX}(\nu)$. Which frequencies does this filter suppress?

4. Show that if $\int \nu^4 S(\nu) d\nu < \infty$ then X has two m.s. derivatives, and $R_{X'X'}(\tau) = -R''_{XX}(\tau)$.

Exercise 14.2. *Let X be the (stationary) Ornstein-Uhlenbeck process of Example 14.7. Using the filtered-spectra theorem, write down the spectral density that the m.s. derivative X' would have, and show that its total power $\int S_{X'X'}(\nu) d\nu$ is infinite. Conclude that the OU process is m.s. continuous but* not *m.s. differentiable.*

Exercise 14.3 (Borel's normal numbers)**.** *(*) Let $\omega \in [0,1]$ with decimal digits $\omega_1\omega_2\omega_3\ldots$; under the uniform (Lebesgue) measure, the digits form an i.i.d. sequence, uniform on $\{0,1,\ldots,9\}$ – in particular a strictly stationary, ergodic process. Call ω* normal *if, for every finite word $a = (a_1,\ldots,a_m)$, the frequency of occurrences of a along the digit sequence tends to 10^{-m}. Use Birkhoff's ergodic theorem (Section 14.5) to show that almost every $\omega \in [0,1]$ is normal. (Following [12].)*

15 A glimpse of Itô calculus *

This closing chapter is an outline, not a full development: definitions and results are stated without proofs, as a preview of where the machinery of this course leads. It is meant as material for a final-lecture seminar; references for serious study are given at the end.

1. The problem. We saw that $V_1(B) = \infty$ almost surely (Theorem 12.11), so integrals $\int f\,dB$ against Brownian motion cannot be defined pathwise as Riemann-Stieltjes integrals – yet the motivating applications (the Ornstein-Uhlenbeck equation of Example 12.14, trading strategies) demand exactly such integrals. The way out uses the two structures we have built: the finite *quadratic* variation of Brownian motion, "$(dB)^2 = dt$" (Theorem 12.13), and the Hilbert space framework of Chapter 13.

2. The Itô integral. For a *simple, non-anticipating* integrand – piecewise constant, $f(t) = f_{t_j}$ on $[t_j, t_{j+1})$, with f_{t_j} depending only on the history up to time t_j – define

$$\int_0^T f\,dB := \sum_j f_{t_j}\big(B(t_{j+1}) - B(t_j)\big).$$

Two remarks. First, the evaluation is at the *left* endpoint: the gambler places the bet before seeing the increment. (The precise formulation of "depends only on the history" uses a **filtration**: an increasing family $\mathcal{F}_t$ of σ-algebras recording the information available up to time t, with each f_{t_j} required to be $\mathcal{F}_{t_j}$-measurable – *adapted*. The cross terms in the isometry below then vanish because the future increment has zero conditional expectation given $\mathcal{F}_{t_j}$.) Unlike the Riemann-Stieltjes case, here the choice of evaluation point changes the answer – other choices give different calculi (e.g. Stratonovich). Second, one computes the **Itô isometry** – the cross terms vanish by the conditional-expectation argument of the previous parenthesis: for $j < k$, $\mathrm{E}[f_{t_j} f_{t_k} \Delta B_j \Delta B_k] = \mathrm{E}\left[f_{t_j} f_{t_k} \Delta B_j\ \mathrm{E}(\Delta B_k | \mathcal{F}_{t_k})\right] = 0$. (Independence of the increments from each other would *not* suffice: f_{t_k} may well depend on the earlier increments.) The result:

$$\mathrm{E}\left[\Big(\int_0^T f\,dB\Big)^2\right] = \mathrm{E}\left[\int_0^T f^2\,dt\right],$$

i.e. $f \mapsto \int f dB$ is an isometry, defined on the closure in $\mathfrak{H}_{[0,T]}$ of the simple non-anticipating integrands, into $\mathfrak{h}$ – so it extends to general non-anticipating square-integrable integrands by taking m.s. limits, exactly as in Chapter 13.

3. Itô's formula. For $f(t,x)$ with continuous $\partial_t f, \partial_x f, \partial_x^2 f$ (i.e. $f \in C^{1,2}$):

$$df(t,B(t)) = \left[\partial_t f + \frac{1}{2}\partial_x^2 f\right](t,B(t))\,dt + \partial_x f(t,B(t))\,dB(t),$$

the extra second-order term being the price of $(dB)^2 = dt$. For example, with $f = \frac{x^2}{2}$:

$$\int_0^t B\,dB = \frac{B(t)^2}{2} - \frac{t}{2},$$

where classical calculus would give no $-\frac{t}{2}$.

4. Stochastic differential equations. The expression $dX = a(t,X)dt + b(t,X)dB$ is shorthand for the integral equation $X(t) = X(0) + \int_0^t a(s,X(s))ds + \int_0^t b(s,X(s))dB(s)$, with the last term an Itô integral. Two fundamental examples:

- **Ornstein-Uhlenbeck** (resolving Example 12.14): $dv = -Cv\,dt + \sigma dB$ has the explicit solution
$$v(t) = e^{-Ct}v(0) + \sigma \int_0^t e^{-C(t-s)}dB(s),$$
a Gaussian process with $\mathrm{Var}[v(t)] \to \frac{\sigma^2}{2C}$. As $t \to \infty$ it approaches the stationary process with autocorrelation $\frac{\sigma^2}{2C}e^{-C|\tau|}$ and Lorentzian spectrum met in Chapter 14 – three chapters of this course meeting in one formula.

- **Geometric Brownian motion:** $dS = \mu S\,dt + \sigma S\,dB$ has solution
$$S(t) = S(0)\,e^{(\mu - \frac{\sigma^2}{2})t + \sigma B(t)},$$
the $-\frac{\sigma^2}{2}$ being an Itô correction (apply Itô's formula to $\log S$). This is the standard model of asset prices underlying the Black-Scholes theory of option pricing.

5. The dictionary with PDEs. Under the usual existence/uniqueness assumptions on the coefficients, a solution of $dX = a(t,X)dt + \sqrt{b(t,X)}\,dB$ is a Markov process; when a sufficiently regular transition density exists, it satisfies the forward Kolmogorov (Fokker-Planck) equation of Theorem 11.19 with drift a and diffusion b – and conditional expectations of functionals satisfy the backward equation. This two-way dictionary between SDEs and PDEs is where stochastic and deterministic analysis merge, and it is the engine behind applications from statistical physics to mathematical finance.

Further reading: Øksendal [17] is the standard accessible entry point to Itô calculus; Part II of Koralov and Sinai [11] continues at a level consistent with these notes; Evans [18] is a compact and readable set of lectures; Sinai's Princeton notes [12] include lectures on stochastic integrals in the same spirit as this chapter.

Appendix: outline solutions and answers

Outline solutions and answers for the end-of-chapter problem sets. They are deliberately terse – checkpoints, not model solutions. Problems referenced at specific points of the text are direct verifications of the adjacent development; the surrounding text is their solution sketch.

Chapter 2

2.4. $\binom{100}{50}2^{-100} \approx \frac{1}{\sqrt{50\pi}} \approx 0.0796$ (Stirling). Note how *small* "the most likely outcome" is.
2.5. Nonnegative integer solutions of $x_1 + x_2 + x_3 = 7$: $\binom{9}{2} = 36$.
2.6. Condition on the colour of the transferred ball: $\frac{3}{7} \cdot \frac{6}{9} + \frac{4}{7} \cdot \frac{7}{9} = \frac{46}{63}$.
2.7. Bayes: $\frac{\frac{1}{2}\cdot\frac{4}{7}}{\frac{1}{2}\cdot\frac{4}{7}+\frac{1}{2}\cdot\frac{6}{8}} = \frac{16}{37} \approx 0.43$.

Chapter 3

3.1. In order: 1. $A \uplus A^c = \Omega$ and additivity; 2. take $A = \Omega$ in 1; 3. from 1 and axiom (i); 4. $B = A \uplus (B \setminus A)$; 5. from 6; 6. $A \cup B = A \uplus (B \setminus A)$ and $P(B) = P(AB) + P(B \setminus A)$.
3.8. $\inf_{k\geqslant n} a_k$ is non-decreasing in n; for the equivalence, squeeze $\inf_{k\geqslant n} a_k \leqslant a_n \leqslant \sup_{k\geqslant n} a_k$.
3.9. Direct unpacking of $\cap\cup$ and $\cup\cap$; in the monotone cases the inner operation stabilises.
3.10. Closure under countable unions plus De Morgan.
3.11. Property 9: complements and Property 8 (note $P(A^c) = 1 - P(A)$). Property 11, third inequality: the sets $\bigcup_{m\geqslant n} A_m$ decrease, apply Property 9 and $B_n \supseteq A_n$.
3.12. Apply the two-event formula to $(A \cup B) \cup C$ and collect terms; for the general case, induction with the same splitting.
3.13. Set $B_n = \bigcap_{m\geqslant n} A_m$ (increasing) and $C_n = \bigcup_{m\geqslant n} A_m$ (decreasing); use continuity along monotone sequences and $B_n \subseteq A_n \subseteq C_n$.
3.14. Each subset corresponds to its indicator function $A \to \{0,1\}$; there are $2^{|A|}$ such functions. Or: induction, splitting subsets by whether they contain the last element.

Chapter 4

4.2. $P(A|WD_\alpha) = 0.9 > 0.8$, $P(A|WD_\beta) = 0.3 > 0.2$, yet $P(A|W) = \frac{36}{100} < \frac{82}{110} = P(A|M)$. The point: $P(A|W) = \sum_j P(A|WD_j)P(D_j|W)$, and the weights differ drastically – women mostly applied to the hard department.
4.3. $P(B_n) = \prod_{k=0}^{n-1} \frac{50-k}{52-k} \cdot \frac{2}{52-n} = \frac{2(51-n)}{52\cdot 51}, \quad n = 0, \ldots, 50$ (check: $\sum_n P(B_n) = 1$).
4.4. First step conditioning: $\frac{b}{b+w} \cdot \frac{b+c}{b+w+c} + \frac{w}{b+w} \cdot \frac{b}{b+w+c} = \frac{b}{b+w}$. For (2), induct: conditioning on the first draw, the urn after one step is again a Polya urn.
4.5. (a) $\frac{1}{3}$ (conditioning event {BB,BG,GB}); (b) $\frac{1}{2}$ (conditioning event {BB,BG}).

Chapter 5

5.3. Z is ± 1 with probability $\frac{1}{2}$; check $f_{ZX} = f_Z f_X$ on the four points. But $P[Z = -1, X + Y = 2] = 0 \neq P[Z = -1]P[X + Y = 2]$.
5.4. Disjoint decomposition of $\{Z = n\}$ plus independence.

5.5. $\mathrm{E} = \sum_n n\frac{2(51-n)}{52\cdot 51} = \frac{50}{3} \approx 16.7$.
5.6. $\mathrm{E}[\text{prize}] = \sum_k 2^k \cdot 2^{-k} = +\infty$: a random variable with infinite expectation; no finite entry fee is "fair".
5.7. Use the factorisation of the joint pdf; for the variance, expand and use the first identity on the cross term.
5.8. $X = T_1 + \cdots + T_r$ with T_j i.i.d. geometric ($\mathrm{E} = \frac{1}{p}$, $\mathrm{Var} = \frac{1-p}{p^2}$); use linearity and independence.

Chapter 6

6.3. Follow the given steps; all the ingredients (Stirling, the substitution for x, the logarithmic expansion) are stated in the exercise.
6.4. $2\varepsilon\sqrt{N} \geqslant 2.58$: $N \geqslant 666$ for ± 0.05; $N \geqslant 16\,641$ for ± 0.01. (Normal-approximation design values, cf. the discussion there.)
6.5. $\lambda = \frac{500}{200} = 2.5$: $1 - e^{-2.5}(1 + 2.5) \approx 0.71$.
6.6. Follow the hint: (ii) is the LLN; (iii) follows from the exponential form of the string probabilities, since the exponent differs from $-NH$ by $N\sum_j(\frac{B_N^j}{N} - p_j)\ln p_j$, controlled by δ; (i) follows by sandwiching $|\vec{\Omega}_N|$ between the bounds that (iii) puts on total probability.

Chapter 7

7.4. Polar coordinates for the square of the integral; then substitute and use symmetry / integration by parts.
7.5. $\mathrm{E} = \frac{1}{\lambda}$, $\mathrm{Var} = \frac{1}{\lambda^2}$; memorylessness from $P[X > x] = e^{-\lambda x}$.
7.6. $\mathrm{E}[X] = \frac{a\theta}{a-1}$ for $a > 1$; $\mathrm{Var}[X] = \frac{a\theta^2}{(a-1)^2(a-2)}$ for $a > 2$; both diverge otherwise – heavy tails.
7.7. $p = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$; divide by $dx\,dy$ and let them vanish.
7.8. $\mathrm{E}[X] = \frac{b-a}{3}$, $\mathrm{Var}[X] = \frac{a^2+ab+b^2}{18}$. For $b > a$: positive skewness; mode $0 <$ median $<$ mean.
7.9. The bin frequency estimates $P[X \in [n\Delta x, (n+1)\Delta x]] \approx f_X(x)\Delta x$.
7.10. $\mu = \frac{a+b}{2}$, $\sigma^2 = \frac{(b-a)^2}{12}$, $\beta = 0$, $\gamma = \frac{9}{5}$.
7.11. $P[Y \leqslant y] = P[X \leqslant F_X^{-1}(y)] = y$ for $y \in [0, 1]$.

Chapter 8

8.2. Monotone transformations: for increasing g, $F_Y(y) = F_X(g^{-1}(y))$; for decreasing g, $F_Y(y) = 1 - F_X(g^{-1}(y))$ (up to null endpoint events). E.g. $f_{e^X}(y) = \frac{1}{y}f_X(\ln y)$ for $y > 0$.
8.3. $Z_1 = \frac{X}{Y}$: split over the sign of Y; $f_{Z_1}(z) = \int |y| f_{XY}(zy, y)\,dy$. $Z_2 = |X + Y|$: $F_{Z_2}(z) = F_{X+Y}(z) - F_{X+Y}(-z)$ for $z \geqslant 0$, so $f_{Z_2}(z) = f_{X+Y}(z) + f_{X+Y}(-z)$.
8.5. $P[\min > t] = e^{-\lambda t}e^{-\mu t}$; for the second part condition on X (or cf. Exercise 11.1).
8.6. Triangular: $f_Z(z) = z$ on $(0, 1)$, $2 - z$ on $(1, 2)$.
8.7. Substitute into $f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$.

Chapter 9

9.2. The ODE is $\partial_k \widehat{f} = -4\pi^2\sigma^2 k\widehat{f}$ with $\widehat{f}(0) = 1$.
9.3. The pdf of the sum is the convolution; its FT is the product $e^{-2\pi ik(\mu_1+\mu_2)}e^{-2\pi^2(\sigma_1^2+\sigma_2^2)k^2}$.
9.4. $N \approx \left(\frac{1.96\cdot 0.9}{10^{-3}}\right)^2 \approx 3\times 10^6$: the $O(N^{-1/2})$ convergence is the price of Monte Carlo.

Chapter 10

10.1. $\pi = \left(\frac{b}{a+b}, \frac{a}{a+b}\right)$; eigen-decomposition gives $P^n = \Pi + (1-a-b)^n(I-\Pi)$ where Π has rows π; geometric convergence at rate $|1-a-b|$.
10.2. Not ergodic (periodic); not ergodic (reducible); ergodic ($P^2 > 0$); ergodic (check $P^s > 0$ for suitable s).
10.3. The difference equation $\overline{L}(i) = 1 + p\overline{L}(i+1) + q\overline{L}(i-1)$ has particular solution $\frac{i}{q-p}$; fit the homogeneous solution $c_1 + c_2(\frac{q}{p})^i$ to the boundary conditions.
10.4. HH: 6; HT: 4. (For HH, progress is destroyed by a tail; for HT, a head is never wasted.)
10.5. Hitting probability $\frac{3}{5}$; mean first-passage time $\frac{22}{9} \approx 2.4$ (solve the two 3×3 linear systems from the potentials / mean-hitting theorems).

Chapter 11

11.3. Each individual survives to time t with probability $e^{-\mu t}$, independently: binomial. Direct verification in the master equation also works.
11.4. $\pi_n = \pi_0 \frac{(\lambda/\mu)^n}{2^{n-1}}$ for $n \geqslant 1$; normalisable iff $\lambda < 2\mu$; with $\rho := \frac{\lambda}{\mu}$, normalisation gives $\pi_0 = \frac{2-\rho}{2+\rho}$, hence $\pi_n = \frac{2-\rho}{2+\rho}\cdot\frac{\rho^n}{2^{n-1}}$, $n \geqslant 1$.

Chapter 12

12.2. Y is Gaussian with mean 0 and $\mathrm{Cov}[Y(s), Y(t)] = c^2\min(s/c^2, t/c^2) = \min(s,t)$; continuity of paths is inherited from B (continuous rescaling), completing the verification that Y is standard BM.
12.3. $\mathrm{E}[X(t)] = 0$, $\mathrm{Cov}[X(s), X(t)] = \min(s,t) - st$. Not a BM (variance $t(1-t)$ is not t); it is pinned to 0 at both endpoints – the Brownian bridge.
12.4. With $U = B(1)$, $V = B(2) - B(1)$ independent standard normals, the pair (U, V) is rotationally symmetric, and $\{U > 0, U + V > 0\}$ is a wedge of angle $\frac{3\pi}{4}$: hence $P[U > 0, U + V > 0] = \frac{3\pi/4}{2\pi} = \frac{3}{8}$, and dividing by $P[U > 0] = \frac{1}{2}$ gives $\frac{3}{4}$.

Chapter 13

13.2. $R(t_1, t_2) = \sigma_A^2 + \sigma_B^2 t_1 t_2$, continuous with continuous mixed derivatives: m.s. derivative exists and equals B.
13.3. For $s < t$: $\mathrm{E}[N(s)N(t)] = \mathrm{E}[N(s)^2] + \mathrm{E}[N(s)]\,\mathrm{E}[N(t) - N(s)] = (\lambda s + \lambda^2 s^2) + \lambda s\cdot\lambda(t-s)$. Continuity of R on the diagonal gives m.s. continuity.

Chapter 14

14.1. (1)–(2) direct computations; (3) $\widehat{h}(\nu) = e^{2\pi i\nu a} - 1$; the filter kills frequencies with $\nu a \in \mathbb{Z}$; (4) differentiate under the integral sign twice.

14.2. $S_{X'X'}(\nu) = \frac{4\pi^2\nu^2\sigma^2}{C^2+4\pi^2\nu^2} \to \sigma^2 \neq 0$ as $|\nu| \to \infty$: not integrable, so X' would have infinite power – no m.s. derivative.

14.3. Apply Birkhoff's theorem to the shift on digit sequences with $f(\omega) = \chi\{\omega_1 \dots \omega_m = a\}$: the time average converges a.s. to $\mathrm{E}[f] = 10^{-m}$ by ergodicity of i.i.d. sequences. Intersect over the countably many words a.

References

[1] B. V. Gnedenko, *The Theory of Probability*, 6th edition, Gordon and Breach (1997).

[2] H. Kobayashi, B. L. Mark and W. Turin, *Probability, Random Processes, and Statistical Analysis*, Cambridge University Press (2012).

[3] H. Tijms, *Probability: A Lively Introduction*, Cambridge University Press (2017).

[4] G. Grimmett and D. Welsh, *Probability: An Introduction*, 2nd edition, Oxford University Press (2014).

[5] M. A. Lerma, *Notes on Discrete Mathematics*, lecture notes, Northwestern University.

[6] Y. Moschovakis, *Notes on Set Theory*, Springer (1994).

[7] H. Royden and P. Fitzpatrick, *Real Analysis*, 4th edition, Pearson (2010).

[8] P. R. Halmos, *Measure Theory*, Springer (1950).

[9] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis*, revised edition, Academic Press (1980).

[10] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, Princeton University Press (2003).

[11] L. B. Koralov and Y. G. Sinai, *Theory of Probability and Random Processes*, 2nd edition, Springer (2007).

[12] Ya. G. Sinai, *Notes on Probability and Stochastic Processes*, Lecture Notes for the Stochastic Processes course, Princeton University (2003).

[13] G. A. Athanassoulis, *Stochastic Modelling of Wave Systems*, Lecture Notes (in Greek), National Technical University of Athens.

[14] O. Knill, *Probability Theory and Stochastic Processes with Applications*, Overseas Press (2009).

[15] J. R. Norris, *Markov Chains*, Cambridge University Press (1997).

[16] N. Privault, *Understanding Markov Chains: Examples and Applications*, 2nd edition, Springer (2018).

[17] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th edition, Springer (2003).

[18] L. C. Evans, *An Introduction to Stochastic Differential Equations*, American Mathematical Society (2013).

[19] E. Çinlar, *Introduction to Stochastic Processes*, Dover (2013).

[20] N. N. Taleb, *The Black Swan*, Random House (2007).