Probability Models

2013

by Omar Lakkis

Copyright and copyleft notice

https://plus.google.com/103353428445025203078

About the cover picture. Rare flower, by Banksy (public domain).

Contents

What’s this?

This is a short introduction to probability theory and modelling. It used primarily for teaching purposes.

Course synopsis

Aims. To build on previous exposure to probability, construct and analyse mathematical models of chance processes, mainly in discrete time.

Learning outcomes. At the end of this course students are expected to be able to:

(1) set up plausible models of probability spaces for a variety of chance experiments;
(2) recognise where and when standard probability distributions are applicable;
(3) find probabilities and other properties associated with random variables and vectors;
(4) use indicator variables, and the idea of conditional expectation, appropriately;
(5) apply the Laws of Large Numbers and the Central Limit Theorem;
(6) apply extinction criteria, and describe the long-term behaviour of branching processes;
(7) understand the main properties of one-dimensional simple random walks;
(8) set up transition matrices of Markov Chains, and describe their long-term behaviour;
(9) apply the ideas to gambling, optimal growth strategy, insurance models and similar.

Syllabus. Probability spaces as models of chance experiments. Axioms, conditional probability. Random variables, distributions, densities, mass functions. Random vectors, joint and marginal distributions, conditioning. Expectation, indicator variables, laws of large numbers, moment generating functions, central limit theorem. Ideas of convergence of random variables. Branching processes, random walks, Markov chains.

Recommended books. Professor John Haigh’s book, “Probability Models” was published in February 2013 in the Springer SUMS series in its third reprinting including corrections and some reviewed parts. This book will also be useful for the Spring Term module, “Random Processes”. It contains solutions to all its exercises. Other suitable books include “Introduction to Probability”, by C Grinstead and J L Snell; (it is on the world wide web in its entirety), and “A First Course in Probability” by S Ross. More advanced, and suitable for anyone intending to get really involved is G R Grimmett and D Stirzaker “Probability” (Third edition 2001) and .

Omar Lakkis
lakkis.o.maths@gmail.com

CHAPTER 1

Probability spaces

The context is an *experiment* (such as the roll of a die) with various outcomes possible, governed by *chance*. In this section we introduce some basic notation and the Axioms of Probablity in the form of definitions.

1.1. A mathematical foundation for Chance

1.1.1. Cardinality, countable and uncountable sets. Sets can be classified into finite and infinite sets. Finite sets are rigorously defined as those that satisfy the pigeonhole principle (i.e., those were an injective map into themselves is necessarily surjective) and a basic result in set theory says that if Ω is a finite set then there exists N such that Ω and $\{1\dots N\}$ (the set of all integers between and including 1 and N) are in a one-to-one correspondence. More practically, this means that the elements of Ω can be *labelled* or *indexed*, without repetition, by using the indices in $\{1\dots N\}$

$$\Omega=\{\omega_1,\omega_2,\dots,\omega_N\} \text{ and } \big(\omega_i=\omega_j \Leftrightarrow i=j\big). \tag{1.1.1}$$

We say that this number N is the *cardinality* or *count* of Ω and we write $N=\#\Omega$.
A set $\mathscr{S}$ is called *countable* if and only if $\mathscr{S}$ is in a one-to-one correspondence (i.e., bijection) with a subset of the natural numbers $\mathbb{N}$. This makes all finite sets countable, but also some infinite sets, like $\mathbb{N}$ itself.
Following this definition any *finite set* is countable, but also some infinite sets, such as $\mathbb{N}$ itself, $\mathbb{N}_0$, $\mathbb{Z}$ and $\mathbb{Q}$ are also countable. It is well-known that the interval $[0,1]$ of the real line, or the real line, $\mathbb{R}$, itself are *uncountable sets*.
Often countable sets are *indexed* over $\mathbb{N}$ or a finite subset thereof. Sometimes $\mathbb{N}_0$ or $\mathbb{Z}$ are used as index sets. Note the values of any *sequence* form a countable set.

1.1.2. Definition of probability space and probability measure. A *probability space* consists of three elements, $(\Omega,\mathscr{F},\mathcal{P})$:

(1) Ω is the set of possible *outcomes* of the experiment, it is also referred to as the *sample space*.
(2) $\mathscr{F}$ is a collection of subsets of Ω; that is $\mathscr{F}$ is a subset of Ω's power set 2^Ω. The elements of the collection $\mathscr{F}$ are called *events* (or, more rigorously, *measurable events*) and are usually (but not always) indicated with some uppercase letter such as A, B, etc.
(3) $\mathcal{P}$ is a function defined for each event, satisfying the three fundamental Laws:
    (a) *positivity* and *maximum likelihood*: for any event $A\in\mathscr{F}$, we have

$$0\le\mathcal{P}(A)\le 1. \tag{1.1.2}$$

    (b) *impossibility* and *certainty* $\mathcal{P}(\varnothing)=0$, $\mathcal{P}(\Omega)=1$.

(c) *(disjoint) additivity* If $A \cap B = \varnothing$, then $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B)$. In fact additivity holds to a countable family of events, $(A_i)_{i \in \mathscr{I}}$ if $A_i \cap A_j = \varnothing$ when $i \neq j$, then $\mathcal{P}\big(\bigcup_{i \in \mathscr{I}} A_i\big) = \sum_{i \in \mathscr{I}} \mathcal{P}(A_i)$.

1.1.3. Example (rolling a die). Consider the game (or experiment) of rolling a six-sided die with faces numbered $1, 2, \dots, 6$. The sample space Ω is the set of all outcomes, hence we have $\Omega = \{1, 2, 3, 4, 5, 6\} = \{1 \dots 6\}$.

In this case, the set of events, $\mathscr{F}$ has a quite simple structure, it is simply the *power set* of Ω, i.e., the set of all possible subsets of Ω. Explicitly, we may enumerate this as

$$\mathscr{F} = \{\varnothing, \{1\}, \{2\}, \dots, \{6\}, \{1,2\}, \{1,3\}, \dots, \{1,2,3,4,5,6\}\}. \tag{1.1.3}$$

Like sets events are usually more easily identified with some characteristic property rather than enumeration, for example, the event "get an odd-numbered face" is equal to $\{1, 3, 5\}$.

Now the probability measure can vary, depending on the die at hand. If we are rolling a "loaded" die, we might have the probabilities, say,

$$\mathcal{P}[1] = \mathcal{P}[2] = \frac{1}{12}, \mathcal{P}[3] = \mathcal{P}[4] = \frac{1}{6}, \mathcal{P}[5] = \mathcal{P}[6] = \frac{1}{4}. \tag{1.1.4}$$

If the die is "fair", instead, then

$$\mathcal{P}[n] = 1/6 \text{ for all } n \in \Omega. \tag{1.1.5}$$

In both cases, we do not need to list the values of $\mathcal{P}$ on all the events in $\mathscr{F}$, as these can be obtained from the elementary probabilities in (1.1.4), or (1.1.5), and the definition of probability measure 1.1.2. For example, in the case of the loaded die we have

$$\mathcal{P}\big[\text{get an odd-numbered face}\big] = \mathcal{P}[1,3,5] = \frac{1}{12} + \frac{1}{6} + \frac{1}{4} = \frac{1}{2}. \tag{1.1.6}$$

While

$$\mathcal{P}\big[\text{get a less or equal than 3 face}\big] = \frac{2}{12} + \frac{1}{6} = \frac{1}{3}. \tag{1.1.7}$$

In the fair-die case we also have

$$\mathcal{P}\big[\text{get an odd-numbered face}\big] = \frac{1}{2}, \tag{1.1.8}$$

but

$$\mathcal{P}\big[\text{get a less or equal than 3 face}\big] = \frac{1}{2}. \tag{1.1.9}$$

1.1.4. Remark (Technical point). $\mathscr{F}$ consists of those subsets of Ω whose probabilities we are interested in. Its technical name is *set field*, or *set algebra*, (also known as *σ-field, sigma-field, sigma-algebra*, or *σ-algebra*)[1]; it has the properties

(i) $\varnothing$ and Ω both belong to $\mathscr{F}$;

(ii) Whenever A belongs to $\mathscr{F}$, then the complement of A, $A^c := \Omega \setminus A$, also belongs to $\mathscr{F}$

(iii) For any countable family $(A_i)_{i \in \mathscr{I}}$ of events in $\mathscr{F}$ we have $\bigcup_n A_n \in \mathscr{F}$.

[1] The symbol σ plays no mathematical role it is a relict from past notation where σ was used to indicate $\cup$ and is short for "summe", or "summation". The terminology σ-algebra is one of the many inconsistencies in mathematical conventions. It's use is so widespread, there is no risk of confusion.

These details are vital for mathematical completeness, but we will not refer to them unless strictly necessary in some of the proofs. In all the applications we shall see in this course, we can safely assume[2] that *any subset of Ω is an element of $\mathscr{F}$*.

1.1.5. Example (darts launching). As we shall see, probability spaces come in two main flavours: *discrete (or countable)* and *continuous*. The die-roll example we used in §1.1.2 is modelled by a discrete (in fact finite) space, we will see discrete spaces that are infinite. Some infinite spaces are *continuous*: take for example the point on a dart-board where the dart hits (assuming that the dart will necessarily hit one point). If we model the dart-board, say, as the open disk D of radius 1 and center $(0,0)$ in the Cartesian plane then we can talk about the probability of the dart hitting one "spot" on the board by using a probability density function $f:\mathbb{R}^2\to\mathbb{R}$ such that

$$f(\boldsymbol{x})>0 \quad \forall\, x\in D \text{ and } \int_{\mathbb{R}^2} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=1. \tag{1.1.10}$$

For example, assume the density function is given by

$$f(\boldsymbol{x}):=\frac{2/\pi^2}{1+|\boldsymbol{x}|^4}. \tag{1.1.11}$$

As an exercise check that f statisfies conditions (1.1.10).[∗] Using this density function, the corresponding *probability measure* of a region S being hit is given by

[∗]: Check!

$$\mathcal{P}(S):=\int_S f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{1.1.12}$$

With the function f given by (1.1.11) let us calculate the probability of *hitting the board*:

$$\begin{aligned}\mathcal{P}(D)&=\int_D f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\frac{2}{\pi^2}\int_D\frac{\mathrm{d}\boldsymbol{x}}{1+|\boldsymbol{x}|^4}\\&=\frac{2}{\pi}\int_0^1\frac{2\rho\,\mathrm{d}\rho}{1+\rho^4}=\frac{2}{\pi}\int_0^1\frac{\mathrm{d}z}{1+z^2}=\frac{2}{\pi}[\arctan z]_{z=0}^{z=1}=\frac{1}{2}.\end{aligned} \tag{1.1.13}$$

In this case, a possible choice for $\mathscr{F}$ is the collection of all *Borel sets* in $\mathbb{R}^2$.[3] We recall that all open and all closed subsets of $\mathbb{R}^2$ and countable unions or intersections thereof are Borel sets.

1.1.6. Exercise (hitting the bulls eye). *In the context of Example 1.1.5, assuming that the bulls eye B has radius $1/10$ and center $(0,0)$, calculate the probability to hit it. What is the probability of hitting exactly the origin $(0,0)$? Discuss your result with a friend.*

[2] But everytime you allow yourself this 'assumption" you must also make yourself feel guilty, because this is logically *wrong*, in the case of continuous probability spaces. This assumption might lead to disaster in some deeper corners of Probability Theory; but this is way beyond the scope of this course. If you are interested, a playful version of the Banach–Tarski paradox can be found on Muller and various authors (2007) with particular attention to Terence Tao's comments.

[3] If you don't know what a Borel set is, don't panic; there is no need to know this for this course, but we put the remark here for awareness.

1.1.7. Theorem (basic properties of probability). *Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space, then $\mathcal{P}$ statisfies the following properties:*

(i) complementarity:

$$\mathcal{P}(A^{\mathrm{c}}) = 1 - \mathcal{P}(A), \text{ where } A^{\mathrm{c}} := \Omega \smallsetminus A := \{\omega \in \Omega : \omega \notin A\}. \tag{1.1.14}$$

(ii) monotonicity:

$$A \subseteq B \Rightarrow \mathcal{P}(A) \leq \mathcal{P}(B). \tag{1.1.15}$$

(iii) inclusion–exclusion property:

$$\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B). \tag{1.1.16}$$

(iv) subadditivity: if $(A_i)_{i\in\mathbb{N}}$ is a sequence in $\mathscr{F}$ then

$$\mathcal{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) \leq \sum_{i=1}^{\infty} \mathcal{P}(A_i) \tag{1.1.17}$$

Proof

(i) Note that $A \cup A^{\mathrm{c}} = \Omega$ and $A \cap A^{\mathrm{c}} = \varnothing$ so by disjoint additivity we have

$$\mathcal{P}(A) + \mathcal{P}(A^{\mathrm{c}}) = 1. \tag{1.1.18}$$

(ii) Note that if $A \subseteq B$ then $B = A \cup (B \smallsetminus A)$, hence by disjoint additivity we have

$$\mathcal{P}(B) = \mathcal{P}(A) + \mathcal{P}(B \smallsetminus A). \tag{1.1.19}$$

Since $\mathcal{P}(B \smallsetminus A) \geq 0$ by positivity of $\mathcal{P}$, the result follows.

(iii) Decomposing $A = (A \smallsetminus B) \cup (A \cap B)$ and $B = (B \smallsetminus A) \cup (B \cap A)$ we have

$$\begin{aligned} \mathcal{P}(A) &= \mathcal{P}(A \smallsetminus B) + \mathcal{P}(A \cap B) \\ \mathcal{P}(B) &= \mathcal{P}(B \smallsetminus A) + \mathcal{P}(A \cap B). \end{aligned} \tag{1.1.20}$$

Thus

$$\mathcal{P}(A) + \mathcal{P}(B) = \mathcal{P}(A \smallsetminus B) + \mathcal{P}(A \cap B) + \mathcal{P}(B \smallsetminus A) + \mathcal{P}(A \cap B). \tag{1.1.21}$$

On the other hand

$$\mathcal{P}(A \cup B) = \mathcal{P}(A \smallsetminus B) + \mathcal{P}(A \cap B) + \mathcal{P}(B \smallsetminus A), \tag{1.1.22}$$

hence

$$\mathcal{P}(A) + \mathcal{P}(B) = \mathcal{P}(A \cup B) + \mathcal{P}(A \cap B). \tag{1.1.23}$$

(iv) For each $i \in \mathbb{N}$ let

$$B_i := A_i \smallsetminus \bigcup_{\substack{j\in\mathbb{N} \\ j\neq i}} A_i \tag{1.1.24}$$

□

1.1.8. Theorem (monotone convergence of events implies convergence of probabilities). *Let* $(A_i)_{i\in\mathbb{N}}$ *be a sequence in* $\mathscr{F}$*, then*

$$A_i \subseteq A_{i+1} \quad \forall\, i \in \mathbb{N} \Rightarrow \mathcal{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i\to\infty} \mathcal{P}(A_i) = \sup_{i\in\mathbb{N}} \mathcal{P}(A_i), \tag{1.1.25}$$

and

$$A_i \supseteq A_{i+1} \quad \forall\, i \in \mathbb{N} \Rightarrow \mathcal{P}\left(\bigcap_{i\in\mathbb{N}} A_i\right) = \lim_{i\to\infty} \mathcal{P}(A_i) = \inf_{i\in\mathbb{N}} \mathcal{P}(A_i). \tag{1.1.26}$$

Proof Let us write write $A_\infty := \bigcup_{i\in\mathbb{N}} A_i$, $p := \mathcal{P}(A_\infty)$ and $p_i := \mathcal{P}(A_i)$. By monotonicity we have that

$$p_i = \mathcal{P}(A_i) \le \mathcal{P}(A_{i+1}) = p_{i+1} \quad \forall\, i \in \mathbb{N}, \tag{1.1.27}$$

hence $\left(p_i\right)_{i\in\mathbb{N}}$ is an increasing sequence. On the other hand $p_i \le 1$ for all i, so the sequence has a supremum (which happens to be limit as well):

$$\sup_{i\in\mathbb{N}} p_i = \lim_{i\to\infty} p_i. \tag{1.1.28}$$

By additivity, we know that for each $i \in \mathbb{N}$ we have

$$p = \mathcal{P}(A_i \cup (A_\infty \setminus A_i)) = p_i + \mathcal{P}(A_\infty \setminus A_i). \tag{1.1.29}$$

Provided the last term in (1.1.29) has a limit for $i \to \infty$, it follows that

$$p = \lim_{i\to\infty} p_i + \lim_{i\to\infty} \mathcal{P}(A_\infty \setminus A_i). \tag{1.1.30}$$

All we need to prove then is that

$$\mathcal{P}(A_\infty \setminus A_i) \to 0 \text{ as } i \to \infty. \tag{1.1.31}$$

Using additivitiy again, we obtain

$$\begin{gathered} \sum_{i=1}^{\infty} \mathcal{P}(A_{i+1} \setminus A_i) = p, \\ \mathcal{P}(A_\infty \setminus A_i) = \sum_{j=i}^{\infty} \mathcal{P}\big(A_{j+1} \setminus A_j\big) \quad \forall\, i \in \mathbb{N} \end{gathered} \tag{1.1.32}$$

Since the first series in (1.1.32) converges it follows that the sequence of its tails converges to 0, which means that (1.1.31) is satisfied, which proves (1.1.25).
To prove (1.1.26), assume now that $(A_i)_{i\in\mathbb{N}}$ is a nested-decreasing sequence of events, then write $B_i := {A_i}^{\mathrm{c}}$ and upon noting that $(B_i)_{i\in\mathbb{N}}$ is nested-increasing, apply complementarity, De Morgan's laws, and (1.1.25) to find

$$1 - \mathcal{P}\left(\bigcap_{i\in\mathbb{N}} A_i\right) = \mathcal{P}\left(\bigcup_{i\in\mathbb{N}} {A_i}^{\mathrm{c}}\right) = \lim_{i\to\infty} \mathcal{P}({A_i}^{\mathrm{c}}) = 1 - \lim_{i\to\infty} \mathcal{P}(A_i), \tag{1.1.33}$$

which implies the first equality of (1.1.26). The second equality (and the existence of the limit) are immediate consequences of $\mathcal{P}$'s monotonicity. □

1.1.9. Theorem (first Borel–Cantelli lemma). *Suppose $(A_i)_{i\in\mathbb{N}}$ is a sequence of events, whose probabilities are the terms of a convergent series, i.e., such that*

$$\sum_{i\in\mathbb{N}} \mathcal{P}(A_i) < \infty \tag{1.1.34}$$

then the probability that $\omega \in \Omega$ belongs to infinitely many A_i's is null, i.e.,

$$\mathcal{P}\left(\bigcap_{j\in\mathbb{N}}\bigcup_{i\geq j} A_i\right) = 0. \tag{1.1.35}$$

The event $\bigcap_{j\in\mathbb{N}}\bigcup_{i\geq j} A_i$ is called $(A_i)_{i\in\mathbb{N}}$ *infinitely often* and is abbreviated as "A_i i.o." in by many authors. Sometimes it is also (more rigorously but less suggestively) written as $\limsup_{i\to\infty} A_i$.

Proof Consider the sequence of events $\left(B_j\right)_{j\in\mathbb{N}}$ given by

$$B_j := \bigcup_{i\geq j} A_i. \tag{1.1.36}$$

Then $\bigcap_{j\in\mathbb{N}} B_j = \limsup_{i\to\infty} A_i$ and $B_{j+1} \subseteq B_j$ for all $j \geq 1$. By the monotone convergence of events theorem 1.1.8 we have that

$$\mathcal{P}\left(\limsup_{i\to\infty} A_i\right) = \lim_{j\to\infty} \mathcal{P}\left(B_j\right). \tag{1.1.37}$$

On the other hand, from (1.1.34) we have

$$\mathcal{P}\left(B_j\right) \leq \sum_{i=j}^{\infty} \mathcal{P}(A_i) \to 0, \text{ as } j \to \infty. \tag{1.1.38}$$

The conclusion (1.1.35) follows from (1.1.37) and (1.1.38). □

1.2. Experiments and "Randomness"

Chance is indicated in jargon by *randomness*. It is not easy to define randomness, and in fact, the very definition of random depends on the context. In many probability models, the hardest task is sometimes that of estabilishing the right concept of randomness for the problem at hand. To appreciate this difficulty you could think about the following question.

1.2.1. Problem. *Pick a random person, what is their height?*

1.2.2. Problem. *Pick a random triangle, what are the odds it is acute (i.e., all of its angles are acute)?*

Solution. An answer to this problem is given by Strang, 2010.

1.2.3. Finite probability spaces. Consider the following experiment: given a set of N ojects, select one of N objects "at random".
Here Ω is just the finite list $\{\omega_1, \omega_2, \cdots, \omega_N\}$, for some N, and we assume there are no repetitions, i.e., $\omega_i = \omega_j \iff i = j$.
In daily language "at random" often indicates that all outcomes are equally likely, so, since nothing states otherwise, we may assume that for some fixed (or uniform) $p > 0$, $\mathcal{P}[\omega_i] = p$ for each $i = 1, \ldots, N$. Because $\mathcal{P}(\Omega) = 1$ and $\#\Omega = N$ it follows that the p in question must equal $1/N$. Unless otherwise stated, we may also assume $\mathscr{F} = 2^{\Omega}$ (the

power set of Ω, i.e., the collection of all subsets of Ω). Then, if an event $A \in \mathscr{F}$ consists of K single outcomes, disjoint additivity of $\mathcal{P}$ yields

$$\mathcal{P}(A) = \sum_{i:\omega_i \in A} \mathcal{P}[\omega_i] = \sum_{i=1}^{K} \frac{1}{N} = \frac{K}{N}. \tag{1.2.1}$$

Of course, the distribution of $\mathcal{P}$ may not be uniform. For example, in the loaded-die example in 1.1.2 $\mathcal{P}$ is not uniformly distributed.
Likewise, there are cases (although very rarely so) of finite probability spaces where the summation-algebra $\mathscr{F}$ over which $\mathcal{P}$ is defined may not be all of 2^{Ω}.

1.2.4. Example (lottery). [4]

PROBLEM. *In the British National Lottery, punters select 6 numbers from a list of 49; 6 of these 49 numbers are chosen at random to be Winning Numbers, the other 43 are thus Losing Numbers. You* randomly *select 6 numbers; what is the chance you will select k winning numbers?*

Solution. A good model for Ω here consists of all possible collections of exactly 6 objects out of 49 available choices:

$$\Omega := \left\{\omega = \left\{d^1, \ldots, d^6\right\} : \ 1 \le d^1, \ldots, d^6 \le 49 \text{ and } \left(d^i = d^j \iff i = j\right)\right\} \tag{1.2.2}$$

By the definition of *binomial coefficients* there are $\binom{49}{6}$ of these. So, assuming all draws are equally likely, i.e., that $\mathcal{P}$ is uniformly distributed, we have

$$\mathcal{P}[\omega] = p = \frac{1}{\binom{49}{6}} \quad \forall\, \omega \in \Omega. \tag{1.2.3}$$

The "winning 6 numbers", say $d_*^1, d_*^2, \ldots, d_*^6$, can be now thought to be one (given) single element, say ω^*, in the set Ω. Denote by A_k the event "k of the guesses are winners"; if $k = 6$ (i.e., we guess exactly all numbers) then the event A_6 consists of only one element, ω_*, so $\# A_6 = 1$ and

$$\mathcal{P}\left[k \text{ of the guesses are winning}\right] = \mathcal{P}(A_6) = p \# A_6 = \frac{1}{\binom{49}{6}}. \tag{1.2.4}$$

If $k = 5$ (i.e., we guess exactly 5 winners and 1 loser), then A_5 consists of those ω which have 5 elements in ω_* and 1 element in $L := \{1 \ldots 49\} \setminus \omega_*$: hence

$$\# A_5 = \binom{6}{5}\binom{49-6}{1} \text{ and } \mathcal{P}(A_5) = \frac{\binom{6}{5}\binom{49-6}{1}}{\binom{49}{6}} \tag{1.2.5}$$

Similarly for any $k \in \{0 \ldots 6\}$, we have

$$\mathcal{P}(A_k) = \frac{\binom{6}{k}\binom{43}{6-k}}{\binom{49}{6}}. \tag{1.2.6}$$

[4] In this example, and in many to come we use some basic *combinatrics*. Since this material is not directly covered in this course, it will be good to become acquainted with it from specialised sources: e.g., Lakkis, 2011, Ch. 5, for a basic treatment, and Grinstead and Snell, 1997, Ch. 3 for a more thorough discussion.

1.2.5. Problem (a random integer). *Explain why the intuitively "obvious" answer of* 50% *doesn't (always) make sense to the following question:*

> *Pick a random positive integer $n \in \mathbb{N}$, what is the likelihood that n is even?*

Hint. Try and assign a "uniform" probability $p > 0$ to all numbers. What happens to $\mathcal{P}(\mathbb{N})$?

Solution. If by "natural" probability, we mean one which is democratically treating all the outcomes equally likely, then it impossible to have such a one for $\mathbb{N}$. In other words, there can be no uniform probability measure on $\mathbb{N}$ (or any countably infinite set for that matter). Suppose there were one, say with $\mathcal{P}[n] = p$ for all $n \in \mathbb{N}$, then by the Monotone Convergence and Countable Additivity we would have

$$1 = \mathcal{P}(\mathbb{N}) = \mathcal{P}\left(\bigcup_{n\in\mathbb{N}} \{1 \dots n\}\right) = \lim_{n\to\infty} \mathcal{P}\{1\dots n\} = \lim_{n\to\infty} np = \infty, \tag{1.2.7}$$

which is absurd. So any prbability measure on $\mathbb{N}$ would have to be non-uniform, for example, we could assign

$$\mathcal{P}[n] = \frac{1}{2^n}. \tag{1.2.8}$$

Clearly $\mathcal{P}(\mathbb{N}) = \sum_{n=1}^{\infty} 1/2^n = 1$ (so $\mathcal{P}$ satisfies the probability axioms) while

$$\mathcal{P}[n \in \mathbb{N} : n \text{is odd}] = \sum_{n \text{ odd}} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{8} + \dots > \frac{1}{2}. \tag{1.2.9}$$

This means that

$$\mathcal{P}[n \in \mathbb{N} : n \text{is even}] < 1 - \frac{1}{2} = \frac{1}{2}, \tag{1.2.10}$$

which implies

$$\mathcal{P}[n \in \mathbb{N} : n \text{is even}] < \mathcal{P}[n \in \mathbb{N} : n \text{is odd}] \tag{1.2.11}$$

1.2.6. Countably infinite probability spaces. If the set Ω underlying the probability space $(\Omega, \mathscr{F}, \mathcal{P})$ is infinite and countable, we talk about a *countably infinite*, e.g., $\Omega = \mathbb{N}$, or $\Omega = \mathbb{N}_0$, or $\Omega = \mathbb{Z}$, or $\Omega = \mathbb{Q}$, but other choices are possible.
In this case, being Ω countable, we can list the outcomes (i.e., Ω's elements) as a sequence (i.e., with an index on $\mathbb{N}$)

$$\Omega := \{w_1, \omega_2, \dots\} = \{w_i\}_{i \in i \in \mathbb{N}} \text{ and } \big(\omega_i = \omega_j \iff i = j\big). \tag{1.2.12}$$

If we know the probability of each single ω_i in this list, i.e., we are given $\mathcal{P}[\omega_i] = p_i$, with $p_i \geq 0$, for each i and $\sum_i p_i = 1$, then for any given $A \subseteq \Omega$ we may use countable additivity to calculate $\mathcal{P}(A)$ as follows:

$$\mathcal{P}(A) = \mathcal{P}\left(\bigcup_{\omega_i \in A} \{\omega_i\}\right) = \sum_{i=1}^{\infty} \mathcal{P}[\omega_i] = \sum_{\omega_i \in A} p_i. \tag{1.2.13}$$

Moral: if the probability space is countable then $\mathscr{F} = 2^{\Omega}$ (the power set of Ω) it is enough to list Ω and state the value of $\mathcal{P}[\omega]$ for every ω in order to have full knowledge of $\mathcal{P}$.

1.2.7. Remark (uniform distributions on countably infinite spaces are not possible). An interesting and somewhat surprising observation is that a *countably infinite set Ω does not admit a uniform probability distribution.*

Indeed, suppose that there was a uniform probability measure on Ω, i.e., for some $p > 0$ we have

$$\mathcal{P}[\omega_i] = p > 0 \quad \forall\, i \in \mathbb{N}, \tag{1.2.14}$$

where we indexed Ω as $\{\omega_i : i = 1, \dots, \infty\}$ then by additivity we have

$$\mathcal{P}(\Omega) = \mathcal{P}\left(\bigcup_{i=1}^{\infty}\{\omega\}\right) = \sum_{i=1}^{\infty}\mathcal{P}(\omega) = \sum_{i=1}^{\infty} p = \infty, \tag{1.2.15}$$

(i.e., the series diverges) on the other hand

$$\mathcal{P}(\Omega) = 1, \tag{1.2.16}$$

which implies that the same series converges (to 1). The contradiction stems from having assumed a uniform distribution on Ω.

1.2.8. Example (a probability measure on $\mathbb{N}$). Consider the probability measure $\mathcal{P}$ on $\mathbb{N}$ such that

$$\mathcal{P}(n) = 1/2^n. \tag{1.2.17}$$

Take $\mathscr{F} = 2^{\mathbb{N}}$ and heck that $\mathcal{P}$ satisfies $\mathcal{P}(\Omega) = 1$.[$*$] Then introduce the event

[$*$]: Check!

$$A := \{\text{even number}\} = \{2n : n \in \mathbb{N}\}, \tag{1.2.18}$$

and using the additivity of $\mathcal{P}$ we calculate

$$\begin{aligned}\mathcal{P}(A) &= \sum_{n\in\mathbb{N}}\mathcal{P}[2n] = \sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^{2n} = \sum_{n=1}^{\infty} 1/4^n \\ &= \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \dots = \frac{1}{4}(1 - 1/4)^{-1} = 1/3.\end{aligned} \tag{1.2.19}$$

You should work out $\mathcal{P}$[odd number] in the same way; it should sum to 2/3, of course.

1.2.9. Exercise. *Introduce a probability measure $\mathcal{P}$ on $\mathbb{N}$ which satisfies*

$$\mathcal{P}[\textit{even number}] = \mathcal{P}[\textit{odd number}] = 50\%. \tag{1.2.20}$$

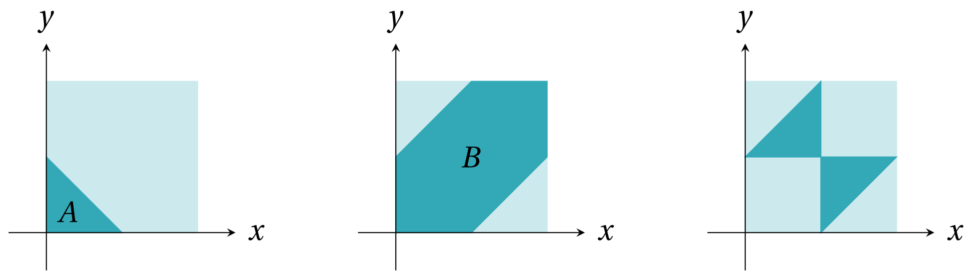
1.2.10. Uncountable probability spaces. An infinite probability space, to be strict the underlying set Ω thereof, may be uncountable. Since being uncountable implies being infinite, there is no need to stress the latter.

A seemigly paradoxical fact of life is that if Ω is uncountable it may very well happen that

$$\mathcal{P}[\omega] = 0 \quad \forall\, \omega \in \Omega. \tag{1.2.21}$$

Indeed, in the Darts Lauching Example 1.1.5, you might have noticed this. This is confusing because it seems to be implying the "sum of all probabilities" is zero (instead of being one). In fact, there is no paradox: infinite sums (series) are defined over *countable* sets, but they are not defined on *uncountable* sets, hence the right-hand side in the following relation:

$$1 = \mathcal{P}(\Omega) = \sum_{\omega\in\Omega}\mathcal{P}[\omega] = 0, \tag{1.2.22}$$



(a) Planar region $A := \{(x, y)\colon x + y < 1/2\}$

(b) Planar region $B := \{(x, y)\colon |x - y| < 1/2\}$

(c) Region of points (x, y) for which triangle possible.

FIGURE 1. A visual aid to the solution of Problem 1.2.12.

does not make sense in this case.

To add to the confusion (for the newbies, that is), although uncountable sets are of higher cardinality than countable ones, some of them, such as bounded intervals of the real line, do admit *uniform distributions*.

1.2.11. Example (an uncountable probability space). Consider the experiment: "Choose a point at random, all points with equal chance, in the unit interval."

Take $\Omega = [0, 1]$; and if $0 \le a \le b \le 1$, define $\mathcal{P}([a, b]) = b - a$. (The probability of an interval depends on its length only, not on its position.) Also define $\mathcal{P}(\varnothing) = 0$.

Note that thanks to this definition we have $\mathcal{P}(\Omega) = 1$. Moreover, knowing the probability of all closed intervals lets us, thanks to additivity and De Morgan's laws[5], use countable unions and intersections to find the probability of many reasonable subsets of the unit interval. Technically speaking, the subsets of $[0, 1]$ that can be attained by taking countable unions and complements (and thus intersections) of closed intervals are called Borel subsets. The probability of individual points will be zero, of open intervals the same as the corresponding closed interval, etc. This idea can obviously be extended for use in any finite interval, not just the unit interval.

1.2.12. Problem (the stick–triangle question). *A stick has unit length. Two points are selected, each chosen at random without reference to the other, and the stick broken in those points. What is the chance the three pieces can form a triangle?*

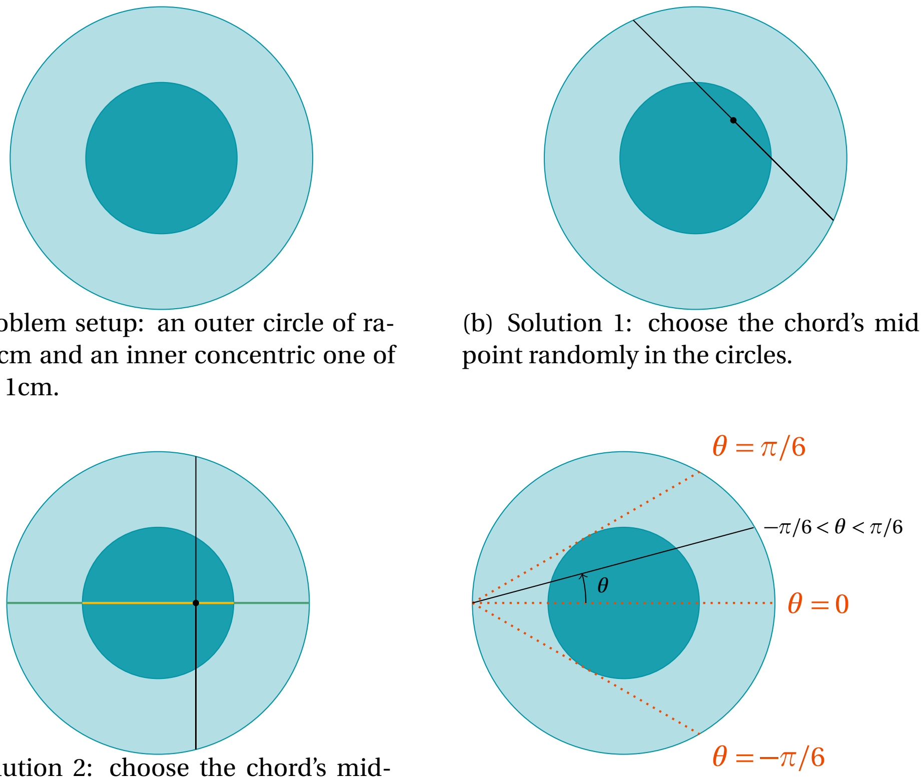
Solution. This is a two-dimensional version of the previous example. Choosing two points, x and y, on the unit line is like choosing one point at random in the unit square. (Dwell on this. It is crucial.) So Ω will consist of the unit square, and for any subset, its probability is just its area. That sets up the probability space.

So (ignoring the specific question for a moment), let A be the region where $x + y < 1/2$, and B be where $|x - y| < 1/2$. Thus: Plainly, $P(A) = 1/8$ and $P(B) = 3/4$. For the

[5] De Morgan's laws for countable collections are stated as follows:

$$\left(\bigcup_{i\in\mathbb{N}} A_i\right)^{\mathrm{c}} = \bigcap_{i\in\mathbb{N}} A_i^{\mathrm{c}} \text{ and } \left(\bigcap_{i\in\mathbb{N}} A_i\right)^{\mathrm{c}} = \bigcup_{i\in\mathbb{N}} A_i^{\mathrm{c}} \qquad (1.2.23)$$

FIGURE 2. Bertrand's Pardox: find the likelihood of a chord intersecting the concentric circle.



(a) Problem setup: an outer circle of radius 2cm and an inner concentric one of radius 1cm.

(b) Solution 1: choose the chord's midpoint randomly in the circles.

(c) Solution 2: choose the chord's midpoint randomly on the horizontal ("x") axis.

(d) Solution 3: choose the chord's angle with the horizontal ("x") axis at (-2,0).

problem set, suppose the point selected is (x, y) with $x < y$. Then (think about it), the three pieces have lengths x, $y-x$ and $1-y$, and the condition they form a triangle is simple: all must be less than 0.5.
So when $x < y$, then $x < 0.5$, $y - x < 0.5$ and $1 - y < 0.5$ define a region in the square; and if $y < x$, we similarly have $y < 0.5$, $x - y < 0.5$ and $1 - x < 0.5$, which defines another region. The diagram is Simple geometry tells us that the shaded region has one quarter of the area of the whole square, so the chance we form a triangle is $1/4$. Later, we will change the rules about how the stick is broken: this may change the probability calculation, but the *diagram* will be the same.

1.2.13. Problem (Bertrand's Paradox; Evans, 2009). [6] *Take a circle of radius 2cm in the plane and choose a chord of this circle at random. What is the probability p this chord intersects the concentric circle of radius 1cm?*

Solution. There are (at least) 3 different solutions with 3 different answers to this problem!

[6] Joseph Bertrand, not to be confused with Bertrand Russell who is famous for a "paradox" of his own as well, has provided *two* paradoxes: one in Probability Theory that we describe here, and another one in Game Theoretical Economics.

Alternative 1. Note that there is a one-to-one correspondence between all possible chords except the ones through the center (but these are "negligible" in probability), and the points in the bigger circle: by considering this point to be the midpoint of the chord.

The chord can be thus randomly chosen by randomly chosing its midpoint rule. Assuming a uniform distribution of the probability of choosing the points in the circle, we obtain

$$p = \frac{\text{midpoint in } C1}{\text{midpoint in } C2} = \frac{\pi}{\pi 4} = \frac{1}{4}. \tag{1.2.24}$$

Alternative 2. Another way of looking at the problem is the following: by symmetry we can look at the chords that are "vertical" (i.e., parallel to the y-axis). Again by considering the midpoint–chord correpondence, we see that if the midpoint lies between -1 and 1 (on the x-axis) it is a favourable case, among all the cases whose points are between -2 and 2. Hence

$$p = \frac{P\{\text{midpoint in } (-1,1)\}}{P\{\text{midpoint in } (-2,2)\}} = \frac{2}{4} = \frac{1}{2}. \tag{1.2.25}$$

Alternative 3. Yet another way, is to consider, by symmetry, that one endpoint of the chord is a the left-most point of the biggest circle, i.e., $(-2,0)$. Then the angle, say θ, the chord forms with the horizontal (x-axis) gives a one-to-one correspondence with the chords. Picking θ randomly gives us

$$p = \frac{P\{\theta : -\pi/6 < \theta < \pi/6\}}{P\{\theta : -\pi/2 < \theta < \pi/2\}} = \frac{\pi/3}{\pi} = \frac{1}{3}. \tag{1.2.26}$$

Of course, there something fishy going on here. Can you guess what?
Hint. Ignoring symmetries, identify the probability space in all three cases, and find, for each pair of spaces, a (differentiable) one-to-one correspondence between them. Look at the determinant of the Jacobian of these correspondences. Make sure you understand why looking at the Jacobian is important. If you require more help, you may have a peek at §5.4, where the solution to this conundrum will be revealed.

1.3. Product probability spaces

In §1.2.12 we have used the Cartesian two "copies" of the probability space from § 1.2.11. We now formalise a bit that construction.

1.3.1. Product of two probability spaces. Let $(\Omega_1, \mathscr{F}_1, \mathcal{P}_1)$ and $(\Omega_2, \mathscr{F}_2, \mathcal{P}_2)$ be two probability spaces. Consider the Cartesian product of sample spaces

$$\Omega := \Omega_1 \times \Omega_2, \tag{1.3.1}$$

and the following collection of subsets thereof

$$\mathscr{F}_0 := \{A_1 \times A_2 : A_1 \in \mathscr{F}_1 \text{ and } A_2 \in \mathscr{F}_2\}. \tag{1.3.2}$$

Geometrically the elements of $\mathscr{F}_0$ are interpreted as "rectangles". It is a useful exercise to convince oneself that $\mathscr{F}_0$ may *not* be a sigma-algebgra. However, a useful result in measure theory tells us that there is a "most economical" sigma-algebra, say $\mathscr{F}$, that contains $\mathscr{F}_0$ (so the rectangles of $\mathscr{F}_0$ are literally the building blocks of $\mathscr{F}$). By most economical, we mean that if $\mathscr{G}$ is another sigma-algebra with $\mathscr{F}_0 \subseteq \mathscr{G}$ then $\mathscr{F} \subseteq \mathscr{G}$. Furthermore, it is possible to *introduce a probability measure* $\mathcal{P}$ *on* $\mathscr{F}$ such that

$$\mathcal{P}(A_1 \times A_2) = \mathcal{P}_1(A_1)\mathcal{P}_2(A_2). \tag{1.3.3}$$

Furthermore, a measure such as $\mathcal{P}$ is unique, i.e., if $\mathcal{Q}$ satisfies

$$\mathcal{Q}(A_1 \times A_2) = \mathcal{P}_1(A_1)\mathcal{P}_2(A_2), \tag{1.3.4}$$

then

$$\mathcal{P}(A) = \mathcal{Q}(A) \quad \forall A \in \mathscr{F}. \tag{1.3.5}$$

The unique probability measure (space) thus defined is called the *product probability measure (space) of* $\mathcal{P}_1$ *by* $\mathcal{P}_2$ and is sometimes denoted as $\mathcal{P}_1 \otimes \mathcal{P}_2$.

1.3.2. Example (rolling 2 dies). As an example, consider the experiment of rolling two dice. Say one is fair and the other one is loaded as follows

$$\mathcal{P}_2(i) = \begin{cases} 1/4, \text{ if } i = 1,2,3, \\ 1/12, \text{ of } i = 4,5,6. \end{cases} \tag{1.3.6}$$

Then the product space is given by the sample space

$$\Omega = \{1\ldots 6\}^2 = \{(i,j) : i,j = 1,\ldots,6\} = \{(1,1),(1,2),(1,3),\ldots,(6,5),(6,6)\} \tag{1.3.7}$$

with the product probability measure $\mathcal{P} = \mathcal{P}_1 \otimes \mathcal{P}_2$ given as a table

	1	2	3	4	5	6	i
1	1/24	1/24	1/24	1/24	1/24	1/24	1/4
2	1/24	1/24	1/24	1/24	1/24	1/24	1/4
3	1/24	1/24	1/24	1/24	1/24	1/24	1/4
4	1/72	1/72	1/72	1/72	1/72	1/72	1/12
5	1/72	1/72	1/72	1/72	1/72	1/72	1/12
6	1/72	1/72	1/72	1/72	1/72	1/72	1/12
j	1/6	1/6	1/6	1/6	1/6	1/6	1

(1.3.8)

1.3.3. Product of a countable sequence of probability spaces. The procedure provided in §1.3.1 can be generalised to any number $n \in \mathbb{N}$ of spaces, by using induction. In fact, with a bit more effort a countable generalisation can be given. We do not enter the details of this construction here but we state the main result. For more details about product spaces, see Billingsley, 1995 or Jacod and Protter, 2003.

THEOREM (countable product space).

Exercises and problems on Probability spaces

Exercise 1.1 (examples of probability spaces)**.** Set up probability spaces that represent

(a) three tosses of a fair coin;

(b) one throw of a die with N sides, in which the chance of the side labelled k is proportional to k, for each of $1 \le k \le N$.

Exercise 1.2 (a discrete probability space). Suppose $\Omega = \{1,2,3,..\}$ with $p(n) = 1/(n(n+1))$ for all n. Verify that this is indeed a probability space, and find the chances the outcome is
(a) at most 3 (b) at least 10 (c) an even number

Exercise 1.3. Find an expression, in terms of $P(A)$, $P(B)$ and $P(A \cap B)$ for the probability that exactly one of A, B occurs.

Exercise 1.4. Use induction to prove Boole's inequality:

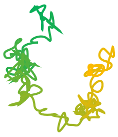
$$\mathcal{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathcal{P}(A_i). \tag{P1.4.1}$$

Exercise 1.5. Assume that the last two digits on a car number plate are equally likely to be any one of the 100 outcomes $\{00, 01, \ldots, 99\}$. Peter bets Paul, at even money, that at least two of the next n cars seen have the same last two digits. What value of n makes this bet pretty fair?

Exercise 1.6. A sphere of radius 1 has 90% of its surface randomly coloured in red, the rest being white. A cube whose main diagonal has length 1 is fitted inside the cube and can move freely with its corners always on the sphere's surface. Show that no matter how the red surface is given, it is possible to find a position of the cube such that all 8 corners are on the red part of the surface.

CHAPTER 2

Conditional probability and independence



A famous probability riddle, known as the "Tuesday Son" question goes like this: a woman says "I have two children, one of them is a girl born on a Friday." Assuming that all days are equally likely for a girl to be born and that the probability of giving birth to a girl or boy is 50%-50%, calculate the probability that both children are girls.
Spoiler: the answer is neither 1/2 nor 1/3 and the solution does require a bit of thinking (or drawing). A quick way to get the solution is to use the concept of conditional probabilty, which is the object of the current chapter.

2.1. Conditioning

2.1.1. Definition of conditional probability. Let A and B be two events in a (sigma-algebra $\mathscr{F}$ on a) probability space Ω, if $\mathcal{P}(B) > 0$, then define the *probability of A conditional to B*

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)}. \tag{2.1.1}$$

2.1.2. Remark. From (2.1.1) it follows that

$$\mathcal{P}(A \cap B) = \mathcal{P}(A|B)\mathcal{P}(B), \tag{2.1.2}$$

a true statement even if $\mathcal{P}(B) = 0$, as then both sides must be zero, hence are equal (no matter what, if any, value is assigned to $\mathcal{P}(A|B)$). This extends to the *product rule*, exemplified by

$$\mathcal{P}(A \cap B \cap C) = \mathcal{P}(A)\mathcal{P}(A|B)\mathcal{P}(C|A \cap B) \tag{2.1.3}$$

and, more generally by

$$\mathcal{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathcal{P}\left(A_i | \bigcap_{j=1}^{i-1} A_j\right) \tag{2.1.4}$$

with the convention that the empty interstection is the whole space, i.e.,

$$\bigcap_{j=1}^{0} A_j = \Omega. \tag{2.1.5}$$

2.1.3. Example.

PROBLEM. *In a game of bridge, find the chance each player has an Ace.*

Solution. As is often the case, good notation virtually solves the problem.
Define

$$A_i := \{\text{player } i \text{ gets an Ace}\}; \tag{2.1.6}$$

we want $\mathcal{P}(A_1 \cap A_2 \cap A_3 \cap A_4)$, and will use the multiplication rule extended to four events.
For $\mathcal{P}(A_1)$, there are $\binom{52}{13}$ possible hands for the player, all equally likely. The number that have one Ace is $\binom{4}{1}\binom{48}{12}$, hence

$$\mathcal{P}(A_1) = \binom{4}{1}\binom{48}{12}\Big/\binom{52}{13}. \tag{2.1.7}$$

To find $\mathcal{P}(A_2|A_1)$, notice that, *given* A_1 the second player will be dealt 13 cards from a reduced pack that has 3 Aces and 36 non-Aces. So by the same argument,

$$\mathcal{P}(A_2|A_1) = \binom{3}{1}\binom{36}{12}\Big/\binom{39}{13}. \tag{2.1.8}$$

And, by the same logic,

$$\mathcal{P}(A_3|A_1 \cap A_2) = \binom{2}{1}\binom{24}{12}\Big/\binom{26}{13}. \tag{2.1.9}$$

You could write out the formula for $\mathcal{P}(A_4|A_1 \cap A_2 \cap A_3)$ if you wanted to, but you'll find it collapses to unity (of course!). Thus the probability we want is

$$\frac{\binom{48}{12}\binom{4}{1}}{\binom{52}{13}} \times \frac{\binom{36}{12}\binom{3}{1}}{\binom{39}{13}} \times \frac{\binom{24}{12}\binom{2}{1}}{\binom{26}{13}} \tag{2.1.10}$$

which works out at around 0.1055 or 10.55%.

2.2. Bayes's Theorem

2.2.1. Definition of partition. A collection of events $B_i \in \mathscr{F}$, for $i \in \mathscr{I}$, forms a *partition* of the probability space Ω, if and only if the following are true:
(i) $\bigcup_{i\in\mathscr{I}} B_i = \Omega$,
(ii) $B_i \cap B_j = \varnothing$ when $i \neq j$.

2.2.2. Theorem (partition of total probability). *Let* $\{B_i\}_{i\in i\in\mathscr{I}}$ *be a countable partition of* Ω*. For any event* $A \in \mathscr{F}$*, we have*

$$\mathcal{P}(A) = \sum_{i\in\mathscr{I}} \mathcal{P}(A|B_i)\mathcal{P}(B_i). \tag{2.2.1}$$

Proof Given that $\{B_i\}_{i\in i\in\mathscr{I}}$ is a partition we have that

$$A = \bigcup_{i\in\mathscr{I}} A \cap B_i \text{ and } \big(i \neq j \Rightarrow (A \cap B_i) \cap \big(A \cap B_j\big) = \varnothing\big). \tag{2.2.2}$$

Thus, by countable disjoint additivity of $\mathcal{P}$ we have

$$\begin{aligned}\mathcal{P}(A) &= \sum_{i\in\mathscr{I}} \mathcal{P}(A\cap B_i)\\ \text{(definition of } \mathcal{P}(\cdot|\cdot)) \quad &= \sum_{i\in\mathscr{I}} \mathcal{P}(A|B_i)\mathcal{P}(B_i).\end{aligned} \tag{2.2.3}$$

□

2.2.3. Theorem (Bayes). *Let $(\Omega,\mathscr{F},\mathcal{P})$ be a probability space and $A, B\in\mathscr{F}$ two events, then*

$$\mathcal{P}(A)\mathcal{P}(B|A) = \mathcal{P}(A|B)\mathcal{P}(B) \tag{2.2.4}$$

Proof This an interesting consequence of the commutativity of $\cap$. Indeed by (2.1.1) we have

$$\mathcal{P}(A)\mathcal{P}(B|A) = \mathcal{P}(A\cap B) = \mathcal{P}(B\cap A) = \mathcal{P}(B)\mathcal{P}(A|B). \tag{2.2.5}$$

□

2.2.4. Remark. If $\mathcal{P}(A)\neq 0$ relation

$$\mathcal{P}(B|A) = \frac{\mathcal{P}(A|B)\mathcal{P}(B)}{\mathcal{P}(A)} \tag{2.2.6}$$

2.2.5. Corollary (Bayes and partitions). *Let $\{B_i\}_{i\in i\in\mathscr{I}}$ be a countable partition of Ω, and A an event, then for each $j\in\mathscr{I}$ we have*

$$\mathcal{P}\left(B_j|A\right) = \frac{\mathcal{P}\left(A|B_j\right)\mathcal{P}\left(B_j\right)}{\sum_{i\in\mathscr{I}}\mathcal{P}(A|B_i)\mathcal{P}(B_i)} \tag{2.2.7}$$

2.2.6. Applications.

- ⋆ Medical diagnosis: A = Symptoms, B_i = possible cause.
- ⋆ The Law: e.g., $A = \{\text{evidence is true}\}$, $B_1 = \{\text{is guilty}\}$, $B_2 = \{\text{is innocent}\}$.

2.2.7. Example.

PROBLEM. 85% *of the taxis in a town are Green,* 15% *are Blue. A witness to an accident involving a taxi states that the taxi was Blue. Tests show that, whatever the true colour of the taxi, she gets it right* 80% *of the time. What is the chance the taxi was indeed Blue?*

Solution. The events $B := \{\text{It was Blue}\}$ and $G := \{\text{It was Green}\}$ form a partition. Let $S := \{\text{She says it was Blue}\}$. We seek $\mathcal{P}(B|S)$. We are given $\mathcal{P}(S|B) = 0.8$ and $\mathcal{P}(S|G) = 0.2$, so

$$\mathcal{P}(S) = \mathcal{P}(S|B)\mathcal{P}(B) + \mathcal{P}(S|G)\mathcal{P}(G) = 0.8\times 0.15 + 0.2\times 0.85 = 0.29, \tag{2.2.8}$$

and so

$$\mathcal{P}(B|S) = \frac{\mathcal{P}(S|B)\mathcal{P}(B)}{\mathcal{P}(S)} = \frac{0.8\times 0.15}{0.29} = \frac{12}{29} \approx 41\%. \tag{2.2.9}$$

Despite the evidence of a witness who is 80% correct saying it is Blue, it is more likely to be Green.

2.3. Independence

The idea of independence is central: the occurrence of one event has no bearing on the probabilities of the other events. Precise definitions are:

2.3.1. Definition of two independent events. Events A, B are *independent* if $\mathcal{P}(A\cap B)=\mathcal{P}(A)\mathcal{P}(B)$.

2.3.2. Definition of many independent events. Events $\{A_i\}_{i\in i\in\mathcal{I}}$ are independent if, for all finite n, and all subcollections of size n,

$$\mathcal{P}\big(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_n}\big)=\mathcal{P}\big(A_{i_1}\big)\mathcal{P}\big(A_{i_2}\big)\ldots\mathcal{P}\big(A_{i_n}\big). \tag{2.3.1}$$

Exercises and problems on Conditional probability and independence

Exercise 2.1. Show that, if

$$\mathcal{P}(A|C)>\mathcal{P}(B|C) \text{ and } \mathcal{P}(A|C^{\mathrm{c}})>\mathcal{P}(B|C^{\mathrm{c}}), \tag{P2.1.1}$$

then

$$\mathcal{P}(A)>\mathcal{P}(B). \tag{P2.1.2}$$

Translate this result into an English sentence, beginning “Given that, whenever C occurs, A is more likely than B, and also that …”.

Exercise 2.2. A and B are mutually exclusive events with respective chances $x>0$ and $y>0$. Show that, in a series of independent repetitions of the experiment, the chance that A occurs before B is $x/(x+y)$.

Exercise 2.3. Three cards of the same size and shape are placed in a hat. One is red on both sides, one black on both, one is red on one side and black on the other. One card is selected at random, and one side exposed: it is seen to be red. What is the chance the other side is red?

Exercise 2.4. 85% of the taxis in a town are Green, 15% are Blue. A witness to an accident involving a taxi states that the taxi was Blue. Tests show that, whatever the true colour of the taxi, she gets it right 80% of the time. What is the chance the taxi was indeed Blue?

Exercise 2.5. The till has 30 £20 notes, and 20 £10 notes. Barmaid Gina, who identifies notes correctly 95% of the time, claims Derek paid his bill with a £10 note. Assuming all 50 notes in the till are equally likely, show that the chance Derek used a £10 note is $38/41$.

Derek, who claims he used a £20 note, correctly identifies £20 notes 80% of the time, and £10 notes 90% of the time. Use the result from the first part to calculate the new chance it was a £10 note.

Exercise 2.6. One coin among n in a bag is double-headed, the rest are fair. Janet selects one of them at random, tosses it k times, gets H (head) every time. What is the chance she chose the double-headed coin?

Exercise 2.7. A bridge hand of 13 cards is dealt from a well-shuffled deck of 52 cards. Find the probabilities it has:

(a) no Aces,
(b) exactly one Ace,
(c) the probability it has at least two Aces, given it has at least one Ace.
(d) Consider the probability it has at least two Aces, given it has at least the Ace of Spades. Will that be the same as, greater or less than c? Justify your answer.

Problem 2.8. Assume that the probability for someone to give birth to a child of either sex is 50%-50%, and that this is independent amongs siblings. (I.e., the sex of the first child doesn't affect the probabilities for the second child.)

(a) A man says: "I have two children." What is the likelihood that both his children are boys.
(b) A man says: "I have two children, one of them is a boy." What is the likelihood that both his children are boys?
(c) A man says: "I have two children, one of them is a boy born on tuesday." What is the likelihood that both his children are boys? You may assume that the probability of having a child (of either sex) on a tuesday is 1/7.

Problem 2.9. Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and let $A, B, C \in \mathscr{F}$. Denote by X^{c} the complement of X in Ω, or $\Omega \setminus X$. For each of the following statements say (i) whether it is true or false, (ii) prove if true, or find countrexample if false:

(a) If $\{A, B\}$ is $\mathcal{P}$-independent, then it is $\mathcal{P}(\cdot|C)$-independent.
(b) If $\{A, B\}$ is an independent pair and $\{B, C\}$ as well then $\{A \cap C, B\}$ is independent.
(c) If $\{A, B\}$ is independent and $\{A, C\}$ is independent $\mathcal{P}(\cdot|C)$ and $\mathcal{P}(\cdot|C^{\mathrm{c}})$-independent, then $\{A, B\}$ is $\mathcal{P}$-independent.

CHAPTER 3

Discrete random variables

The concept of random variable is central to modern probability. In this chapter we start by looking at a very special class of random variables, that of *discrete random variables*, which can be understood without too much technical background. Most basic concepts in probabilistic modelling can be understood in the discrete setting anyway.

3.1. Discrete random variables

In this section we describe the most commonly used discrete probability spaces and random variables.

3.1.1. Discrete random variables. We have already defined a discrete probability space to be one where the set of all possible outcomes is *discrete* (also known as *countable*), i.e., finite or countably infinite (i.e., in a one-to-one correspondence with $\mathbb{N}$). Let $(\Omega, \mathscr{F}, \mathcal{P})$ (not necessarily discrete) probability space, a *discrete random variable* X is a function defined on Ω if

(i) the set of values that X can attain (also known as the image of Ω through X), denoted as $X(\Omega)$ is countable, i.e., for some $K \subseteq \mathbb{N}$,

$$X(\Omega) = \{x_k\}_{k \in k \in K}, \tag{3.1.1}$$

(ii) and for each $k \in K$ the counterimage of x_k is an event, i.e.,

$$X^{-1}(\{x_k\}) = \{X = x_k\} \in \mathscr{F}. \tag{3.1.2}$$

If the set K is finite, then X is called a *finitely valued*, *finite*, or *simple random variable*.

3.1.2. Remark (to index or not?) Note that we could have done without the index k in the definition of discrete random variable in §3.1.1. In fact, we could use the *image* (i.e., the set of values) of X, $X(\Omega)$, as an index set. Since $X(\Omega)$ is countable this means that expressions like

$$\sum_{x \in X(\Omega)} f(x) \text{ and } \bigcup_{x \in X(\Omega)} A_x \tag{3.1.3}$$

where $f(x)$, resp. A_x, is a vector-valued, resp. event-valued, function of $x \in X(\Omega)$. Furthermore, defining $f(x) = 0$ and $A_x = \emptyset$, if $x \in \mathbb{R} \smallsetminus X(\Omega)$, these expressions can be even written as

$$\sum_{x \in \mathbb{R}} f(x) \text{ and } \bigcup_{x \in \mathbb{R}} A_x, \tag{3.1.4}$$

harmlessly as the sum and the union effectively occur on a countable family. In the rest of these notes we will be occasionally using these conventions.

3.1.3. Indicator randoms variables. The most basic example of simple (and hence discrete) random variable, is that of indicator random variable. Given an event A, its *indicator random variable* is defined as[1]

$$\mathbb{1}_A := \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{if } A \text{ does not happen.} \end{cases} \tag{3.1.5}$$

3.1.4. Iverson's brackets. This is the right spot to introduce the following *Iverson's bracket notation* popularised by Knuth (1992): if P is a proposition (i.e., a mathematical statement that has to be either true or false) then

$$[\![P]\!] := \begin{cases} 1, & \text{if } P \text{ is true,} \\ 0, & \text{if } P \text{ is false.} \end{cases} \tag{3.1.6}$$

Using Iverson's brackets the indicator variable of the set A can be equivalently written as

$$\mathbb{1}_A(\omega) = [\![\omega \in A]\!]. \tag{3.1.7}$$

3.1.5. A quotient set reduction. Although a discrete random variable is not necessarily defined on a discrete probability space, it can be "reduced" to one. This can be done in two equivalent (but different) ways which we describe next.
Let $X : \Omega \to \mathbb{R}$ be a random variable, let us consider on Ω the equivalence relation $\equiv$ given by

$$\omega \equiv \omega' \Leftrightarrow X(\omega) = X(\omega'). \tag{3.1.8}$$

The quotient set $\tilde{\Omega} = \Omega / X$ induced by $\equiv$, is defined as the set of all equivalence classes y such that

$$\omega, \omega' \in y \Leftrightarrow \omega \equiv \omega'. \tag{3.1.9}$$

The resulting set $\tilde{\Omega}$ is a discrete one (although Ω may not be so). A new probability measure $\tilde{\mathcal{P}}_X$ can be now be introduced on $\tilde{\Omega}$ as follows

$$\tilde{\mathcal{P}}_X(A) = \mathcal{P}\left\{\omega \in \Omega : \omega \in y, \text{ for some } y \in A\right\}. \tag{3.1.10}$$

It is a tedious exercise to check that this is indeed a probability measure.
Furthermore the resulting function, with domain the "new" probability space,

$$\begin{aligned} \tilde{X} : \quad \tilde{\Omega} &\to \mathbb{R} \\ y &\mapsto \tilde{X}(y) = X(\omega) \text{ for } \omega \in y \end{aligned} \tag{3.1.11}$$

is then a well-defined function (the value of $X(y)$ does not depend on the particular y-class representative ω) on the new probability space $\tilde{\Omega}$ and it can be shown that it is measurable (hence calling it a random variable is legitimate).
Noting that although Ω is not discrete, the quotient (or "X-reduced") space Ω / X is, it follows that discrete random variables can be studied to all effects in the framework of discrete probability distributions and (although their domain may not be discrete in general) are usually thought as such.

[1] In Mathematical Analysis, "indicator random variables" take the name of "(measurable) characteristic functions". Unfortunately, in Probability, the name "characteristic function" is given to what analysts call "Fourier transform" and we stick to the probabilist's conventions here.

3.1.6. The image set reduction and the induced probability distribution. An equivalent way of building $\tilde{\mathcal{P}}_X$ is to define it on the set of values of X, $X(\Omega)$, known in set theory as the *image* of X (or the image of Ω under X). Since X is a discrete random variable, $X(\Omega)$ is a discrete set, by definition and we may label its elements over a set $S \subseteq \mathbb{N}$, i.e.,

$$X(\Omega) = \{x_1, x_2, \ldots\} = \{x_i\}_{i \in i \in S}, \text{ and } x_i \neq x_j \Leftrightarrow i \neq j. \tag{3.1.12}$$

Consider then the X-induced probability measure, defined by

$$\mathcal{P}_X\{x_i\} := \mathcal{P}\big(X^{-1}\{x_i\}\big), \text{ for } i \in S, \tag{3.1.13}$$

and extended to all subsets of $X(\Omega)$ by additivity. Noting that $X(\Omega)$ and Ω/S are in a one-to-one correspondence, call it $\phi : \Omega/X \leftrightarrows X(\Omega)$, it turns out that the $\mathcal{P}_X$ introduced in §3.1.5 (call it $\tilde{\mathcal{P}}_X$ now) is closely related to the $\mathcal{P}_X$ introduced here, in that

$$\mathcal{P}_X(A) = \tilde{\mathcal{P}}_X(X^{-1}A). \tag{3.1.14}$$

This construction is seen as an action that X takes upon $\mathcal{P}$ and goes by the name of *push forward* in Measure Theory and Optimal Transport Villani, 2003 and is denoted as $X\mathcal{P}\triangleright$.

Note that $\mathcal{P}_X$ is a probability measure on $X(\Omega)$. It is called the X*-induced probability measure.*

3.1.7. Definition of discrete probability distribution. A function $f : \mathbb{R} \to \mathbb{R}$ is called a *discrete probability distribution* if and only if

(i) $f(x) \geq 0$ for all $x \in \mathbb{R}$,

(ii) the set of points x for which $f(x) \neq 0$, called the *support* of f and indicated with $\operatorname{spt} f$, is countable (finite or infinite),

(iii) f has unit sum, i.e., $\sum_{x \in \operatorname{spt} f} f(x) = 1$.

3.1.8. Remark (all distributions are induced). The induced probability measure, $\mathcal{P}_X$, of a discrete random variable X, as defined in 3.1.6, is characterised by its values on singletons, i.e., by the values of the function f defined as

$$f(x) := \mathcal{P}_X[x] = \mathcal{P}[X = x]. \tag{3.1.15}$$

It is a good exercise to show that such an f is a discrete probability distribution.
Conversely, given a discrete probability distribution f, it is always possible to find a discrete random variable X on a probability space $(\Omega, \mathscr{F}, \mathcal{P})$ such that $f(x) = \mathcal{P}_X[x]$ for $x \in \mathbb{R}$. Indeed, define the sample space $\Omega := \operatorname{spt} f$ (this is a discrete sample space), and thereon consider the probability measure $\mathcal{P}$ such that $\mathcal{P}[\omega] = f(\omega)$ for each $\omega \in \Omega$. Finally, let X be the identity map on Ω, i.e., $X(\omega) = \omega$ for $\omega \in \Omega$, then X is a random variable with respect to $\mathcal{P}$ and its distribution $\mathcal{P}_X$ satisfies $f(x) = \mathcal{P}_X[x]$, as claimed.
This observation justifies the terminology in Definition 3.1.7 and reduces the study of the distribution of discrete random variables to that of discrete probability distributions on $\mathbb{R}$.

3.2. Expectation of a discrete random variable

Random variables were born to be manipulated algebraically: extracting proper numbers from random numbers is the basis of probability theory (if those random numbers have a theoretical nature) and statistics (if those numbers come from real-life measurments). The first quantity we are usually interested in, when we do statistics, is the *average*, which goes also by the names of *mean*, *expected value* or *expectation*.

3.2.1. Motivation. Given a finite number of measurements of the same quantity (also known as *sample*) $x_1, \ldots, x_N$, their arithmetic *mean* or *average* is

$$\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{3.2.1}$$

For example, x_i could be the number of accidents on a road on day i and $N = 365$, where we conduct daily statistics over a year's time. Although not necessary, but to simplify the formulas below, let us temporarily assume that for some $n \in \mathbb{N}$,

$$x_i \in \mathbb{N} \text{ and } 0 \leq x_i \leq n \quad \forall\, i = 1, \ldots, N; \tag{3.2.2}$$

this assumption conforms with the "accidents" example, where the count can be only a nonnegative integer.

There is another way of calculating $\bar{x}$. It goes like this: for each fixed $k = 0, \ldots, n$, we may count out those x_i's such that $x_i = k$ into one set, $\mathscr{A}_k$; in symbols,

$$\mathscr{A}_k := \{ i \in \{1 \ldots N\} : \ x_i = k \}. \tag{3.2.3}$$

We can then calculate the likelihood of having k accidents on a "generic" day as

$$p_k := \frac{\# \mathscr{A}_k}{N} \tag{3.2.4}$$

You can now observe that the average can be calculated using these likelihoods as

$$\bar{x} = \sum_{k=0}^{n} k p_k. \tag{3.2.5}$$

From a practical point of view the second way of calculating the average is more economical in the accident's case. Indeed, assuming that the maximum number of daily accidents n is much smaller than 365, say $n = 20$ for very bad days, the first one requires a data-set $\{x_i : \ i = 1, \ldots, 365\}$ of size $N = 365$, whereas the second data-set $\left\{p_k : \ k = 0, \ldots, 20\right\}$ has size $n + 1 = 21$.

That was statistics where the average computed is known as the *sample average*, or *sample mean*. Now in probability theory, the second procedure turns out to be also very important. For example, assume we are given the number of accident likelihoods (or probabilities as we call them) by some canny statistician and we would like to forecast tomorrow's traffic. The accident experiment can then be modelled with a probability space Ω, whose outcomes ω model all the possible patterns on a given day (tomorrow). Next, we introduce the events

$$A_k := \{ \omega \in \Omega : \ \omega \text{ has } k \text{ accidents tomorrow} \}. \tag{3.2.6}$$

We assume the knowledge of a probability measure on Ω. Letting X be the number of accidents occuring tomorrow, then, X is a random number (or variable) and we have

$$\{X = k\} = A_k. \tag{3.2.7}$$

In practice, a reasonable approximation for $\mathcal{P} A_k$ is the p_k calculated by the Statistician above, but we can simply assume that $\mathcal{P} A_k$ is *known* for each $k = 0, \ldots, n$. Now we ask: given this probability distribution, what is the *expected number* of accidents that will occur tomorrow?

If Ω is a countable set then the natural answer is the approximation of $\bar{x}$ given by

$$\sum_{\omega\in\Omega} X(\omega)\mathcal{P}[\omega] \text{ which is also } = \sum_{k=0}^{n} k\mathcal{P}[X=k] = \sum_{k=0}^{n} k\mathcal{P}(A_k). \tag{3.2.8}$$

This motivates, and in fact is, the definition of *expected value* of a discrete random variable X on a countable probability set Ω:

$$\mathrm{E}X := \mathrm{E}[X] := \sum_{\omega\in\Omega} X(\omega)\mathcal{P}[\omega]. \tag{3.2.9}$$

If Ω is not countable, but X is discrete, then the expected value must be defined a bit more carefully as follows

$$\mathrm{E}X := \mathrm{E}[X] := \sum_{x\in X(\Omega)} x\mathcal{P}_X[x], \tag{3.2.10}$$

whenever the sum on the right hand side converges absolutely (which is an issue only when X has infinitely many values).
If X has infinitely many values and $X \geq 0$ we say that it has *finite expectation* if the series appearing in (3.2.10) converges (which is the same as converging absolutely in this case), while we say that it has *infinite expectation* if the series diverges.
If X has inifinitely many values, but changes sign, we way that it has finite expectation if the series converges absolutely. Otherwise we have to split X as follows

$$X = [X]_+ - [X]_- \tag{3.2.11}$$

where the *positive part* and *negative part* are defined by

$$[X]_+ := 0 \vee X \text{ and } [X]_- := -0 \wedge X. \tag{3.2.12}$$

Since both $[X]_+$ and $[X]_-$ are nonnegative, their expectations are either finite or infinite. Both of them are finite if and only if X has a finite expectation as defined above. If one is finite and the other is infinite, we define $\mathrm{E}X$ to be $\pm\infty$, the sign depending on which part is bigger. If both are infinite, then we declare X not to have an expectation.

3.2.2. Example.

PROBLEM. *On the same spin of a roulette wheel:*
Xavier bets one unit on outcome 15,
Yolande bets two units on the "street" $\{13, 14, 15\}$*,*
Wilf bets ten units on the column $\{1, 4, 7, \cdots, 34\}$*,*
with respective payout odds $35:1$, $11:1$ *and* $2:1$.

(a) *Write down the payoff,* X, Y *and* W*, respectively, of each of the players as a random variable.*
(b) *Calculate the expectation and the variance of each of* X, Y, W.

Solution. The outcome ω is one of the integers $0, 1, 2, \ldots, 36$; we have three random variables, X, Y, W that represent their profits. $X(\omega) = 35$ if $\omega = 15$, $X(\omega) = -1$ otherwise. $Y(\omega) = 22$ if $13 \leq \omega \leq 15$, $Y(\omega) = -2$ otherwise. $W(\omega) = 20$ if $\omega = 1$ modulo 3, $W(\omega) = -10$ otherwise.

3.2.3. Proposition (well-posedness of E**).** *Let* X *be a discrete random variable with finite expectation and suppose that*

$$X = \sum_{k \in K} x_k \mathbb{1}_{A_k} \tag{3.2.13}$$

for some pairwise disjoint family of events $\{A_k\}_{k \in k \in K}$*,* $K \subseteq \mathbb{N}$*, then*[2]

$$\mathrm{E} X = \sum_{k \in K} x_k \mathcal{P}(A_k). \tag{3.2.14}$$

Proof Since $\{A_k\}_{k \in k \in K}$ is pairwise disjoint, for each $k \in K$ there is at most one $\omega \in \Omega$ for which $\omega \in A_k$, and thus $X(\omega) = x_k \mathbb{1}_{A_k}(\omega) = x_k$. That is for each $k \in K$ there is a unique $x \in X(\Omega)$ such that $x_k = x$, giving us thus a mapping $K \ni k \mapsto x_k \in X(\Omega)$. Furthermore, for each $x \in X(\Omega)$ the set $K_x := \{k \in K : x = x_k\}$ is nonempty and thus

$$\{X = x\} = \bigcup_{k \in K_x} A_k. \tag{3.2.15}$$

Hence

$$\begin{aligned} \mathrm{E} X = \sum_{x \in X(\Omega)} x \mathcal{P}[X = x] = \sum_{x \in X(\Omega)} x \mathcal{P} \bigcup_{k \in K_x} A_k \\ = \sum_{x \in X(\Omega)} x \sum_{k \in K_x} \mathcal{P} A_k = \sum_{k \in K} x_k \mathcal{P} A_k. \end{aligned} \tag{3.2.16}$$

□

3.2.4. Proposition (monotonicity and linearity of (discrete) expectation). *Let* $X, Y : \Omega \to \mathbb{R}$ *be two discrete random variables on a probability space* $(\Omega, \mathcal{P}, \mathscr{F})$ *then we have*

(i) monotonicity of E

$$X \le Y \Rightarrow \mathrm{E}[X] \le \mathrm{E}[Y], \tag{3.2.17}$$

and

$$X < Y \Rightarrow \mathrm{E}[X] < \mathrm{E}[Y]. \tag{3.2.18}$$

($X \le Y$ *means* $X(\omega) \le Y(\omega)$ *for all* $\omega \in \Omega$*, and similarly for* $X < Y$*.)*

(ii) for any $\alpha, \beta \in \mathbb{R}$ *we have*

$$\mathrm{E}\big[\alpha X + \beta Y\big] = \alpha \mathrm{E}[X] + \beta \mathrm{E}[Y]. \tag{3.2.19}$$

Proof We leave the proof as an exercise, so you can familiarise with E. □

3.2.5. Theorem (triangle inequality for expectations). *Let* X *be a discrete random variable. Then* $|X|$ *is also a random variable and*

$$|\mathrm{E} X| \le \mathrm{E}|X|. \tag{3.2.20}$$

Proof If X takes finitely many values, say $x_1 < \cdots < x_n$, the result follows by induction from the basic triangle inequality for numbers:

$$|a + b| \le |a| + |b| \quad \forall\, a, b \in \mathbb{R}. \tag{3.2.21}$$

[2] Despite its simplicity this result is quite useful. It is valid also when the disjoint family assumption is removed as long as the series (if any) converges absolutely $\mathcal{P}$-almost surely, but we will not need this extension.

Indeed

$$|\mathrm{E}\,X| = \left|\sum_{i=1}^{n} x_i \mathcal{P}_X[x_i]\right| \leq \sum_{i=1}^{n} |x_i| \mathcal{P}_X[x_i] = \mathrm{E}\,|X|. \tag{3.2.22}$$

If X takes infinitely many values, say $\{x_i\}_{i \in i \in \mathbb{N}}$, with $x_i \neq x_j$ for all $i \neq j$, then the inequality is still true for the partial sums, and hence for the limit, leading to

$$|\mathrm{E}\,X| = \lim_{n\to\infty} \left|\sum_{i=1}^{n} x_i \mathcal{P}_X[x_i]\right| \leq \lim_{n\to\infty} \sum_{i=1}^{n} |x_i| \mathcal{P}_X[x_i] = \mathrm{E}\,|X|. \tag{3.2.23}$$

□

3.2.6. Remark (dealing with ∞). Whenever inequalities are involved we use the convention that

$$a < \infty \text{ and } -\infty < a \;\forall\; a \in \mathbb{R}. \tag{3.2.24}$$

This allows to use results such as (3.2.20) in the case where one (or both the terms are inifinite). For example, if $\mathrm{E}\,|X|$ is finite then $|\mathrm{E}\,X|$ and $\mathrm{E}\,X$ are finite. Also if $|\mathrm{E}\,X|$ (or $\mathrm{E}\,X$) is infinite then $\mathrm{E}\,|X|$ is also infinite. When we say that "$\mathrm{E}\,X$ is infinite", we mean that the series $\sum_{x \in X(\Omega)} x \mathcal{P}[X = x]$ diverges to $\pm\infty$.[3]

3.3. Independent discrete random variables

3.3.1. Definition of independent random variables. A pair of discrete random variables (X, Y) is called independent[4] if and only if for any $x, y \in \mathbb{R}$ we have

$$\mathcal{P}_{(X,Y)}\big[(x, y)\big] := \mathcal{P}\big[X = x \text{ and } Y = y\big] = \mathcal{P}_X[x]\mathcal{P}_Y\big[y\big]. \tag{3.3.1}$$

A similar definition applies for n discrete random variables X_i, $i = 1, \dots, n$. The discrete random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is called *independent*, or by abusing the languange, the discrete random variables $X_1, \dots, X_n$ are called independent, if and only if for each $(x_i)_{i \in \{i \dots 1\}n} \in \mathbb{R}^n$, we have

$$\mathcal{P}_{(X_i)_{i\in\{i\dots1\}n}}\left[(x_i)_{i\in\{i\dots1\}n}\right] = \prod_{i=1}^{n} \mathcal{P}_{X_i}[x_i]. \tag{3.3.2}$$

3.3.2. Exercise. *Prove that unless the probability measure $\mathcal{P}$ is trivial, i.e., $\mathcal{P}A = 0$ or 1 for all $A \in \mathscr{F}$, the random vector (X, X) cannot be independent for any nonconstant random variable X.*

[3] Note that this is not a license to say nonsense, such as trying to assign a value to $0 \times \infty$.

[4] Many authors, including the writing one, often talk about "independent random variables" instead of "independent vector", or "independent set". This is an abuse of language: as saying that "X and Y are independent" is not the same thing as saying "X is independent and Y is independent".

3.3.3. Theorem (independence and expectation of products). *Let X and Y be discrete random variables on $(\Omega,\mathscr{F},\mathcal{P})$. If X and Y are independent then*

$$\mathrm{E}[XY]=\mathrm{E}[X]\mathrm{E}[Y]. \tag{3.3.3}$$

Proof

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{z\in[XY]\Omega} z\mathcal{P}[XY=z] \\
&= \sum_{z\in[XY]\Omega} z\mathcal{P}\Big(\bigcup_{x\in X(\Omega)}\{X=x \text{ and } xY=z\}\Big) \\
\text{(additivity of } \mathcal{P} \text{ and independence of } X \text{ and } Y\text{)} \quad &= \sum_{z\in[XY]\Omega} z \sum_{x\in X(\Omega)} \mathcal{P}[X=x]\mathcal{P}[xY=z] \\
\text{(changing summation variable } z=xy\text{)} \quad &= \sum_{y\in Y(\Omega)} xy \sum_{x\in X(\Omega)} \mathcal{P}[X=x]\mathcal{P}\big[xY=xy\big] \\
\text{(ignoring the summands with } x=0 \text{ and rearranging terms)} \quad &= \sum_{x\in X(\Omega)} x\mathcal{P}[X=x] \sum_{y\in Y(\Omega)} y\mathcal{P}\big[Y=y\big] \\
&= \mathrm{E}[X]\mathrm{E}[Y].
\end{aligned} \tag{3.3.4}$$

□

3.3.4. Exercise (multiplicativity of expectations and independence not equivalent). *Let X and Y be two random variables.*

(a) Condition

$$\mathrm{E}[XY]=\mathrm{E}[X]\mathrm{E}[Y] \tag{3.3.5}$$

is only necessary, but not sufficient for the random variables X and Y to be independent. Convince yourself of this fact by finding two random variables X and Y that are not independent yet satisfying (P3.6.1).

(b) Although (P3.6.1) is only necessary, it often becomes sufficient if coupled with some more assumptions on X and Y. For example, if X and Y are two Bernoulli trials taking each the value 0, and satisfying (P3.6.1), then X and Y are independent. Prove it.

3.3.5. Theorem (independence and convolution). *Let X and Y be two independent discrete random variables, and let $\mathcal{P}_X$, $\mathcal{P}_Y$ and $\mathcal{P}_{X+Y}$ denote the probabilities induced by the subscript, then*

$$\mathcal{P}_{X+Y}[z]=\mathcal{P}_X*\mathcal{P}_Y\{z\}:=\sum_{x\in X(\Omega)}\mathcal{P}_X[x]\mathcal{P}_Y[z-x]. \tag{3.3.6}$$

Proof The operation defined in the second part of (3.3.6) is called a *discrete convolution.* We will encounter its continuous counterpart in Chapter 5. To prove the result, fix a $z \in \mathbb{R}$, then

$$\begin{aligned}\mathcal{P}_{X+Y}[z] &= \mathcal{P}[X+Y=z] \\ &= \mathcal{P} \bigcup_{x \in X(\Omega)} \{X = x \text{ and } Y = z-x\} \\ &= \sum_{x \in X(\Omega)} \mathcal{P}[X=x \text{ and } Y=z-x] \\ &= \sum_{x \in X(\Omega)} \mathcal{P}[X=x]\mathcal{P}[Y=z-x].\end{aligned} \tag{3.3.7}$$

□

3.4. Variance

3.4.1. Example (A motivation for variance by Graham, Knuth and Patashnik, 1994).

PROBLEM. *A lottery consist of a fair draw of* 1 *number out of* 100 *possible ones and* 100 *tickets numbered* 1 *through* 100 *are sold. Each day a number is drawn at random and a holder of that ticket number wins* £100M*, where* M *is short for one million. Someone makes us an offer we can't refuse: 2 free tickets. But we are given the choice between*

A. getting 2 *(differently numbered) tickets for today's draw,*
B. getting 1 *ticket for today's draw and* 1 *ticket for tomorrow's draw.*

Which choice should we take?

The first thing we look at is the expectations, in strategy A we obtain

$$\text{chance that one of the two tickets wins} \times \text{prize} = \frac{1}{50} 100M = 2M. \tag{3.4.1}$$

While strategy B yields

$$\begin{pmatrix} \text{chance ticket 1 wins today} \times \text{prize} \\ + \\ \text{chance ticket 2 wins tomorrow} \times \text{prize} \end{pmatrix} = \frac{1}{100}100M + \frac{1}{100}100M = 2M. \tag{3.4.2}$$

So seeminlgy there is no difference between the strategy A and stragety B. To understand this better, let us phrase the problem in a more mathematical notation. Let

$$\Omega_0 := \{1 \dots 100\}, \text{ with } \mathcal{P}_0[\omega] = \frac{1}{100} \quad \forall\, \omega \in \Omega_0, \tag{3.4.3}$$

and consider the product space

$$\Omega = \Omega_0 \times \Omega_0 \text{ with } \mathcal{P}[(\omega_1, \omega_2)] = \mathcal{P}_0[\omega_1]\mathcal{P}_0[\omega_2] = 10^{-4} \quad \forall\, \boldsymbol{\omega} = (\omega_1, \omega_2) \in \Omega. \tag{3.4.4}$$

Choose then the following random variables:

- ⋆ X_1 to be the gain from the first ticket today (used for both strategies),
- ⋆ $\tilde{X}_1$ to be the gain from the second ticket today (used for strategy A) and
- ⋆ X_2 the gain from second ticket tomorrow (used for strategy B).

Indicating by t_1, $\tilde{t}_1$ and t_2 the chosen tickets (with $t_1 \neq \tilde{t}_1$), we have

$$
\begin{gathered}
X_1(\omega_1,\omega_2) = \begin{cases} 100M & \text{if } \omega_1 = t_1 \\ 0 & \omega_1 \neq t_1, \end{cases} \quad \tilde{X}_1(\omega_1,\omega_2) = \begin{cases} 100M & \text{if } \omega_1 = \tilde{t}_1 \\ 0 & \omega_1 \neq \tilde{t}_1, \end{cases} \\
\text{and } X_2(\omega_1,\omega_2) = \begin{cases} 100M & \text{if } \omega_1 = t_2 \\ 0 & \omega_1 \neq t_2. \end{cases}
\end{gathered}
\tag{3.4.5}
$$

Using these "elementary" random variables, we can write the profits for strategy A and strategy A, respectively as

$$
Y = X_1 + \tilde{X}_1 \text{ and } Z = X_1 + X_2. \tag{3.4.6}
$$

Then we recover the previously calculated expectations

$$
\mathrm{E}[Y] = \mathrm{E}[X_1] + \mathrm{E}[\tilde{X}_1] = 100M \times \mathcal{P}[\omega_1 = t_1] + 100M \times \mathcal{P}\left[\omega_1 = \tilde{t}_1\right] = 2M, \tag{3.4.7}
$$

having used the linearity of E, and

$$
\mathrm{E}[Z] = \mathrm{E}[X_1] + \mathrm{E}[X_2] = 100M \times \mathcal{P}[\omega_1 = t_1] + 100M \times \mathcal{P}[\omega_2 = t_2] = 2M. \tag{3.4.8}
$$

Note that although Y and Z have the same expectation, they have different distributions; in particular while X_2 and X_1 are independent, X_1 and $\tilde{X}_1$ are not, because $t_1 \neq \tilde{t}_1$. Indeed,

$$
Y(\omega_1,\omega_2) = \begin{cases} 100M & \text{if } \omega_1 = t_1 \\ 100M & \text{if } \omega_1 = \tilde{t}_1 \\ 0 & \text{if } \omega_1 \in \Omega_0 \setminus \left\{t_1, \tilde{t}_1\right\} \end{cases} \tag{3.4.9}
$$

whereas

$$
Z(\omega_1,\omega_2) = \begin{cases} 200M & \text{if } (\omega_1,\omega_2) = (t_1,t_2) \\ 100M & \text{if } (\omega_1 = t_1 \text{ and } \omega_2 \neq t_2) \text{ or } (\omega_1 \neq t_1 \text{ and } \omega_2 = t_2) \\ 0 & \text{if } (\omega_1 \neq t_1 \text{ and } \omega_2 \neq t_2). \end{cases} \tag{3.4.10}
$$

It is now apparent that Z has more oscillation around its expectation than Y does. Intuitively, if we are extremely lucky we may win twice with strategy B, while a win in strategy A excludes the other win. To do some algebra let us denote

$$
\mu := \mathrm{E}[Y] = \mathrm{E}[Z], \tag{3.4.11}
$$

and consider the "squared oscillations about the average"

$$
V := \left(Y - \mu\right)^2 \text{ and } W := \left(Z - \mu\right)^2. \tag{3.4.12}
$$

(Note that we could have taken absolute values instead of squares but, for a reason that becomes apparent later in the course, the squares turn out to be more suited to the purpose of measuring the oscillations.) Noting that V and W are random variables, we are tempted to look at their expectations

$$
\begin{gathered}
\mathrm{E}[V] = (100-2)^2 M^2 \frac{2}{100} + (0-2)^2 M^2 \frac{98}{100} = 196M^2, \\
\mathrm{E}[W] = \left((200-2)^2 + 2(100-2)^2\, 99 + (0-2)^2\, 99^2\right) M^2 10^{-4} = 198M^2.
\end{gathered}
\tag{3.4.13}
$$

Now we see that *there is a difference* between the distribution of the payoff of strategy A, Y, and that of strategy B, Z. This means that Z is more "spread out" than Y around their common expectation. The numbers computed in (3.4.13) are squares (in terms of units), in order to relate them to the mean, we take the square roots, which leads to the following quantities

$$\sigma_Y := \sqrt{196M^2} = 14M \text{ and } \sigma_Z := \sqrt{198M^2} = 14.071M, \tag{3.4.14}$$

which are called the *standard deviations* of Y and Z respectively. Looking back at Y and Z, the fact that $\sigma_Z > \sigma_Y$ is reflected in the fact that strategy B has the potential of yielding a maximal gain of $200M$ instead of just $100M$ which the maximum for strategy A. On the other hand strategy B is riskier, in that, the probability of actually winning something (1.99%) is lower than that for strategy A (2%). To see this more clearly, think of the extreme situatino where you can get 100 tickets. By using strategy A in this case, you will guarantee a win, whereas using strategy A, by taking 50 tickets for today and 50 for tomorrow you will foresake the guarantee, but that's the "price to pay" in order to maximise the gain. Financial mathematics is a branch of applied probability which deals with such issues.

3.4.2. Definition of variance and standard deviation. Let X be a discrete random variable, with finite expectation $\mathrm{E}X$, its variance is defined as

$$\operatorname{var}[X] := \mathrm{E}\left[|X - \mathrm{E}X|^2\right]. \tag{3.4.15}$$

If $\operatorname{var}[X]$ is finite, the *standard deviation* of X, often denoted as σ_X, is the square root of $\operatorname{var}[X]$:

$$\sigma_X := \sqrt{\operatorname{var}[X]}. \tag{3.4.16}$$

3.4.3. Variance, statistics and Chebyshev's inequality. The main motivation behind this definition will be appreciated later in the course, when we talk about the (Weak and Strong) Law of Large Numbers and the Central Limit Theorem, where the variance plays a crucial role. As a prelude to that study, we mention a simple yet important result, relating the variance of a random variable and the probability of its oscillations around its mean.

THEOREM (discrete Chebyshev's inequality). *Let X be a discrete random variable. If X has finite mean and variance, then for any $\alpha \in \mathbb{R}^+$ we have*

$$\mathcal{P}[|X - \mathrm{E}[X]| > \alpha] \leq \frac{\operatorname{var}[X]}{\alpha^2}. \tag{3.4.17}$$

Proof We leave the proof as a recommended exercise. □

3.5. Covariance and independence

Given two random variables X and Y, both with finite expectation and finite variance. We would like to look at the way X influences Y. A useful tool for this is the so called *covariance*.

3.5.1. Definition of covariance. If X and Y are two random variables with finite mean, we define their *covariance* as

$$\operatorname{cov}[X,Y] := \mathrm{E}[(X-\mathrm{E}X)(Y-\mathrm{E}Y)]. \tag{3.5.1}$$

The *correlation* of X and Y, when both have non-zero variance, is defined as their covariance relative to their respective variances, i.e.,

$$\operatorname{cor}[X,Y] = \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{var}X \operatorname{var}Y}}. \tag{3.5.2}$$

Two random variables are called *uncorrelated* if they have zero correlation, or if one of them has zero variance (or, equivalently, that they have zero covariance).

3.5.2. Proposition (covariance the quick way). *If X and Y are two random variables with finite mean and finite covariance then*

$$\operatorname{cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \tag{3.5.3}$$

3.5.3. Theorem (independence implies zero correlation). *If X and Y are independent then they are uncorrelated, i.e.,*

$$\operatorname{cov}[X,Y] = 0. \tag{3.5.4}$$

Proof The definition of covariance, independence of X and Y and Theorem 3.3.3 imply that

$$\operatorname{cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[X]\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[Y] = 0. \tag{3.5.5}$$

□

3.5.4. Remark. Note that the converse of Theorem 3.5.3 does not hold. Namely, there are examples of random variables that are uncorrelated yet not independent.

3.5.5. Theorem (Cauchy–Bunyakovskii–Schwarz inequality). *For two random variables X and Y we have*[5]

$$\mathrm{E}[XY] \leq \sqrt{\mathrm{E}[X^2]\mathrm{E}[Y^2]}, \tag{3.5.6}$$

where the inequality is understood in the "extended reals" sense, i.e., still valid if some of the terms are infinite with the proviso that $x < \infty$ for any $x \in \mathbb{R}$ and no "fishy" operations such as $0 \times \infty$ occur.

Proof First, let us deal with the case where $\mathrm{E}\left[X^2\right]$ and $\mathrm{E}\left[Y^2\right]$ are finite (i.e., in $\mathbb{R}$). In this case, by elementary algebra we have

$$|XY| \leq \frac{1}{2}\left(X^2 + Y^2\right) \tag{3.5.7}$$

which, implies, by the Monotonicity and Linearity of Expectation Theorem that

$$\mathrm{E}|XY| \leq \frac{1}{2}\mathrm{E}\left[X^2\right] + \frac{1}{2}\mathrm{E}\left[Y^2\right] < \infty. \tag{3.5.8}$$

[5] In most textbooks this is known as the Cauchy–Schwarz inequality, and sometimes just Cauchy's or Schwarz's inequality. Cauchy first published it for finite sums and series, in Cauchy, 1821, e.g. while Bunyakovskii published the first known version for integrals in Bouniakowsky, 1859 and Schwarz published similar results, indendently of Bunyakovskii, in 1888.

This and the Triangle Inequality, $|\mathrm{E}[XY]| \leq \mathrm{E}|XY|$, shows that XY has a finite mean; that is, $\mathrm{E}[XY] \in \mathbb{R}$.
Next, X and Y being fixed, consider the following function of $t \in \mathbb{R}$,

$$f(t) := \mathrm{E}\left[(tX+Y)^2\right]. \tag{3.5.9}$$

By the Monotonicity of Expectation Theorem 3.2.4 and the fact that $(tX(\omega)+Y(\omega))^2 \geq 0$ for all $\omega \in \Omega$ and $t \in \mathbb{R}$ we get that

$$0 \leq f(t) \quad \forall\, t \in \mathbb{R}. \tag{3.5.10}$$

On the other hand, for each $t \in \mathbb{R}$ we have

$$f(t) = \mathrm{E}\left[X^2\right] t^2 + 2\,\mathrm{E}[XY]\, t + \mathrm{E}\left[Y^2\right]. \tag{3.5.11}$$

This means that $f(t)$ is a quadratic polynomial in t that is always of the same sign. The discriminant of f is thus non-positive, i.e.,

$$\mathrm{E}[XY]^2 - \mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right] \leq 0, \tag{3.5.12}$$

which after simple manipulation implies (3.5.6).
If one of the two random variables, say X, has infinite second moment, $\mathrm{E}\left[X^2\right]$, then either Y is almost surely 0, in which case also $XY = 0$ almost surely, and the inequality is still satisfied whatever value is attributed to $0 \times \infty$, or $\mathrm{E}\left[Y^2\right] > 0$ and the inequality is trivially satisfied with the rule that $a \times \infty = \infty$ for any $a > 0$. □

3.5.6. Corollary. *If a random variable X has finite second moment $\mathrm{E}\left[X^2\right]$ then it has a finite expectation and variance.*
Proof By the Cauchy–Bunyakovskii–Schwarz inequality, we have that

$$\mathrm{E}[X] = \mathrm{E}[X1] = \sqrt{\mathrm{E}[X^2]\mathrm{E}[1^2]} = \sqrt{\mathrm{E}[X^2]} < \infty. \tag{3.5.13}$$

□

3.5.7. Theorem (algebra of discrete expectation and variance). *Let $(X_1, \ldots, X_n)$ be a discrete random vector and $(c_1, \ldots, c_n) \in \mathbb{R}^n$, then*

(i) the expectation and the linear combination commute

$$\mathrm{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i\, \mathrm{E}[X_i]; \tag{3.5.14}$$

(ii) the variance is quadratic

$$\operatorname{var}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2 \operatorname{var}[X_i] + 2 \sum_{1 \leq i < j \leq n} c_i c_j \operatorname{cov}\left[X_i, X_j\right]; \tag{3.5.15}$$

(iii) and, if the random vector $(X_i)_{i \in \{i \ldots 1\} n}$ is independent then

$$\operatorname{var}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2 \operatorname{var}[X_i]. \tag{3.5.16}$$

.

Proof See Exercise 3.2. □

3.6. Common discrete distributions and random variables

3.6.1. Proposition (indicator variable's statistics). *If A is an event, then its indicator variable satisfies*

$$\mathrm{E}[\mathbb{1}_A] = \mathcal{P}(A) \text{ and } \operatorname{var}[\mathbb{1}_A] = \mathcal{P}(A)(1-\mathcal{P}(A)). \tag{3.6.1}$$

Proof The proof is left as an exercise. □

3.6.2. Discrete uniform distribution. The *uniform distribution* over a finite set $\Omega = \{\omega_1, \ldots, \omega_n\}$ has $\mathcal{P}(\omega_i) = 1/n$ for all i, we denote it as $\mathrm{U}\,S$. In particular, we use $\mathrm{U}\{1\ldots n\}$ to mean choosing one of the integers $1, \cdots, n$, all equally likely.
More generally a random variable X defined on a general probabilty space $(\Omega, \mathscr{F}, \mathcal{P})$ is said to be *discretely and uniformly distributed* if and only if X is discrete and its induced probability measure is uniform, i.e., for a finite set of values for X, say $\{x_k : k = 1, \ldots, n\}$, we have

$$\mathcal{P}[X = x_k] = \frac{1}{N} \quad \forall\, k = 1, \ldots, n. \tag{3.6.2}$$

We write that $X \in \mathrm{U}\{x_k : k = 1, \ldots, n\}$.
It is not too hard to check that if $\Omega = \{\omega_1, \ldots, \omega_n\}$ has a uniform probability measure $\mathcal{P}$, then the random variable X defined by $X(\omega_k) = k$, for $k = 1, \ldots, n$, it is in $\mathrm{U}\{1\ldots n\}$.

3.6.3. Bernoulli trials. A sequence of *Bernoulli trials* is a sequence of independent repetitions of an experiment with two possible outcomes, Success and Failure, say (or H and T, or 1 and 0), with $\mathcal{P}[\text{Success}] = p$ the same for each trial. Write $q = 1 - p = \mathcal{P}[\text{Failure}]$. This is a very simple, yet very important, example of probablity where the space has cardinality 2 (i.e., 2 elements), denoted by $\{\mathrm{F}, \mathrm{S}\}$ or $\{\mathrm{T}, \mathrm{H}\}$ or $\{0, 1\} = \mathbb{Z}_2$.
A random variable, say X, defined on any probability space Ω, that takes only 2 possible values, i.e., for some event A and $a, b \in \mathbb{R}$

$$X(\omega) = \begin{cases} a, \text{ for } \omega \in A, \\ b, \text{ for } \omega \in \Omega \smallsetminus A, \end{cases} \tag{3.6.3}$$

has a distribution similar to that of a single Bernoulli trial. So much so that such a random variable is actually called a *Bernoulli trial* by definition. Note that the probability space Ω can be quite large and needs not coincide with $\mathbb{Z}_2$. However, Ω can be "reduced" to $\mathbb{Z}_2$ for all the purposes where only X is involved, by considering the quotient space Ω/X defined as the set of equivalence classes of $\equiv$ where, for $\omega, \omega' \in \Omega$ we define

$$\omega \equiv \omega' \Leftrightarrow X(\omega) = X(\omega'). \tag{3.6.4}$$

Another way of constructing $\mathcal{P}_X$ is to simply introduce it on the image of X, $X(\Omega)$, as follows

$$\mathcal{P}_X[k] := \mathcal{P}[X = k], \text{ for } k \in X(\Omega) \tag{3.6.5}$$

and noting that $X(\Omega)$ is isomorphic (via the canonic bijection) to $\tilde{X}$.

3.6.4. Exercise (mean and variance of a single Bernoulli trial). *Let a random variable X be a Bernoulli trial on a space $\varOmega$ with possible values a in the event A, with $\mathcal{P}(A)=p$, and b otherwise. Show that*

$$\mathrm{E}[X]=ap+b(1-p) \text{ and } \operatorname{var}X=(a-b)^2p(1-p). \tag{3.6.6}$$

Denote $q:=1-p$ and consider $a=1$, $b=0$, and rewrite these formulas for that case.

3.6.5. Binomial distribution. A random variable S has the *binomial distribution*, and we write $S\in\mathrm{B}(n,p)$ if and only if

$$\mathcal{P}_S[r]=\binom{n}{r}p^r q^{n-r} \text{ for } r=0,\dots,n. \tag{3.6.7}$$

Such a random variable arises as the number of successes in a sequence of n Bernoulli trials, where each trial is independent of the others, each with success rate p and denote by $\mathcal{P}_1$ the probability measure thus induced on $\mathbb{Z}_2$. This can be modelled with the probability space

$$\varOmega:=\mathbb{Z}_2{}^n=\underbrace{\{0,1\}\times\ldots\times\{0,1\}}_{n \text{ times}} \tag{3.6.8}$$

where each outcome is given by

$$\boldsymbol{\omega}=(\omega_1,\dots,\omega_n), \text{ with } \omega_i=\begin{cases}0 & \text{with probability } 1-p,\\ 1 & \text{with probability } p,\end{cases} \text{ for } i=1,\dots,n. \tag{3.6.9}$$

It is handy to think of "1" modelling *success* and "0" *failure*, where $\boldsymbol{\omega}\varOmega$ models a sequence of independent identical experiments (i.e., with the same likelihood of success), which we call *trials*. To model *independence* of the trials we define the probability of $\{\boldsymbol{\omega}\}$ as the product of the elementary "single-trial" probability $\mathcal{P}_1$:

$$\mathcal{P}[\boldsymbol{\omega}]=\mathcal{P}_1[\omega_1]\cdots\mathcal{P}_1[\omega_n]. \tag{3.6.10}$$

Introducing the random variable the counts the number of successes in a sequence of n trials we get

$$S(\boldsymbol{\omega}):=\#\{i=1,\dots,n:\ \omega_i=1\}, \tag{3.6.11}$$

we hence have

$$\mathcal{P}[\boldsymbol{\omega}]=p^{S(\boldsymbol{\omega})}(1-p)^{S(\boldsymbol{\omega})}. \tag{3.6.12}$$

Suppose $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ are two sequences with the same number of success, i.e., $S(\boldsymbol{\omega})=S(\boldsymbol{\omega}')=s$, say, we have

$$\mathcal{P}[\boldsymbol{\omega}]=p^s(1-p)^{n-s}=\mathcal{P}\left[\boldsymbol{\omega}'\right], \tag{3.6.13}$$

which can be interpreted as "it does not matter when the successes occur, all that matters is their number". This means that we can lump all outcomes with this same probability into the event

$$\varOmega_s:=\left\{\text{there are exactly } s \text{ successes over } n \text{ trials}\right\}, \tag{3.6.14}$$

which has $\binom{n}{s}$ elements[∗] and hence probability [∗]: Check!

$$\mathcal{P}(\varOmega_s)=\binom{n}{s}p^s q^{n-s}, \text{ with } q:=1-p. \tag{3.6.15}$$

Since s is arbitrary with $s = 0, \dots, n$, it follows that the distribution function of the random variable S, $\mathcal{P}_S$, coincides with the $\mathrm{B}(n,p)$ distribution.

Note that the family $\{\Omega_0, \dots, \Omega_n\}$ constitutes a partition of Ω. Furthermore this family is the set of equivalence classes generated by the random variable S. So proceeding as we did with the single Bernoulli trial we can "reduce" the probability space Ω to the set $\{0 \dots n\}$ (sometimes denoted $\mathbb{Z}_{n+1}$) which is isomorphic to the quotient Ω/S. So to all effects of a sequence of n-trials, all that matters is the reduced probability space $\{0 \dots n\}$, rather than the much larger Ω. For this reason the Binomial distribution of n (independent) Bernoulli trials with success rate p is usually (in most textbooks) identified with the discrete *probability measure* defined by

$$\mathcal{P}\{k\} = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.6.16}$$

3.6.6. Proposition (mean and variance of the binomial). *Let Ω be the probability space of all sequences of n successive Bernoulli trials each with success chance p. Consider the random variable*

$$X(\boldsymbol{\omega}) = \textit{ number of successes in } \boldsymbol{\omega}. \tag{3.6.17}$$

Then

$$\mathrm{E}[X] = np \textit{ and } \operatorname{var} X = np(1-p). \tag{3.6.18}$$

Proof We give two alternative proofs (for didactic reasons). The first one uses calculus and is lengthy, while another one is more "probabilistic" and goes faster.

(a) In this proof it is important to recall the Binomial Formula

$$(a+b)^m = \sum_{k=0}^{m} \binom{m}{k} a^k b^{m-k}. \tag{3.6.19}$$

Computing the average is directly based on this:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{\boldsymbol{\omega}\in\Omega} X(\boldsymbol{\omega}) P[\boldsymbol{\omega}] \\
\text{(X-partitioning Ω and using $\mathcal{P}[\Omega_s] = \tbinom{n}{s} p^s q^{n-s}$)} \quad &= \sum_{s=0}^{n} \sum_{\boldsymbol{\omega}\in\{X(n,\cdot)=s\}} s p^s q^{n-s} \\
\text{(by counting $\#\{\boldsymbol{\omega} : X(n,\boldsymbol{\omega}) = s\} = \tbinom{n}{s}$)} \quad &= \sum_{s=0}^{n} s \binom{n}{s} p^s q^{n-s} \\
\text{(using $\tbinom{n}{k} k = \tbinom{n-1}{k-1} n$)} \quad &= np \sum_{s=1}^{n} \binom{n-1}{s-1} p^{s-1} q^{n-1-(s-1)} \\
\text{(substituting $k = s-1$ and using $\sum_{k=0}^{m} \tbinom{m}{k} p^k q^{m-k} = 1$)} \quad &= np.
\end{aligned} \tag{3.6.20}$$

To compute the variance we start similarly

$$\begin{aligned}
& \mathrm{E}\left[X^2\right] = \sum_{\boldsymbol{\omega}\in\Omega} X(\boldsymbol{\omega})^2 P\left[\boldsymbol{\omega}\right] \\
\text{(X-partitioning Ω and using $\mathcal{P}[\Omega_s] = \binom{n}{s} p^s q^{n-s}$)} \quad & = \sum_{s=0}^{n} \sum_{\boldsymbol{\omega}\in\{X(n,\cdot)=s\}} s^2 p^s q^{n-s} \\
\text{(by counting $\#\{\boldsymbol{\omega} : X(n,\boldsymbol{\omega}) = s\} = \binom{n}{s}$)} \quad & = \sum_{s=0}^{n} s^2 \binom{n}{s} p^s q^{n-s} \\
\text{(using $\binom{n}{k} k = \binom{n-1}{k-1} n$)} \quad & = np \sum_{s=1}^{n} \binom{n-1}{s-1} s p^{s-1} q^{n-s} \\
\text{(calculus)} \quad & = np \sum_{s=1}^{n} \binom{n-1}{s-1} \partial_p \left[p^s q^{n-s}\right] \\
& = np\partial_p \left[\sum_{s=1}^{n} \binom{n-1}{s-1} p^s q^{n-1-(s-1)}\right] \\
\text{(substituting $k = s-1$)} \quad & = np\partial_p \left[\sum_{k=0}^{n-1} \binom{n-1}{k} p^{k+1} q^{n-1+k}\right] \\
& = np\partial_p \left[p\left(p+q\right)^{n-1}\right] \\
\text{(assuming $n \geq 2$ and differentiating)} \quad & = np\left(\left(p+q\right)^{n-1} + p(n-1)\left(p+q\right)^{n-2}\right) \\
& = np\left(1 + pn - p\right) \\
& = npq + n^2 p^2.
\end{aligned} \tag{3.6.21}$$

Working out the variance this gives

$$\operatorname{var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = npq + n^2p^2 - \left(np\right)^2 = npq. \tag{3.6.22}$$

(b) Now the "short proof".[6] Denote by $X_1, X_2, \ldots, X_n$ a sequence of independent and identically distributed Bernoulli trials. Then $X = \sum_{i=1}^{n} X_i$. Hence, by additivity of expectations

$$\mathrm{E}[X] = \mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}[X_i] = np. \tag{3.6.23}$$

To compute the variances, we use the fact that X_i and X_j are independent whenever $i \neq j$, which implies that

$$\operatorname{cov}\left[X_i, X_j\right] = \mathrm{E}[X_i]\mathrm{E}\left[X_j\right] - \mathrm{E}\left[X_i X_j\right] = 0. \tag{3.6.24}$$

It follows that

$$\operatorname{var}[X] = \sum_{i=1}^{n} \operatorname{var} X_i - 2 \sum_{1\leq i<j<\leq n} \operatorname{cov}\left[X_i, X_j\right] = npq. \tag{3.6.25}$$

□

[6] The long proof is clearly tedious, but it was given for two didactic reasons: (1) the techniques here introduced are useful elsewhere, (2) it shows how probabilistic thinking can avoid lengthy analysis.

3.6.7. Remark (Binomial distribution in practice). Although we developped the binomial distribution starting from independent random variables on $\Omega = \mathbb{Z}_2{}^n$ we could have used any other single Bernoulli trial on a more general probability space Υ instead of $\mathbb{Z}_2$, as the following problem shows

PROBLEM. *The university has* $N = 8\,192$ *students and runs* 4 *quality-control surveys per year, where* 128 *students are chosen randomly out of the* N *students.*

(a) *What is the chance that Annie will never be surveyed.*

(b) *What is the chance that Annie will be picked once.*

(c) *What is the expected number of times that Annie will be picked for the survey during her* 3 *years of studies.*

3.6.8. Definition of geometric distribution. A discrete random variable X has the *geometric distribution* $\mathrm{G}_1\left(p\right)$, and we write $X \in \mathrm{G}_1\left(p\right)$, if and only if

$$\mathcal{P}[X=r] = pq^{r-1} \text{ for } r = 1, 2, \ldots. \tag{3.6.26}$$

It arises as the number of Bernoulli trials to record the first Success.
Similarly, the discrete random variable X is said to follow a 0-*geometric distribcution* distribution, $X \in \mathrm{G}_0\left(p\right)$, where

$$\mathcal{P}[X=r] = pq^r \text{ for } r = 0, 1, 2, \ldots. \tag{3.6.27}$$

such a variable arises when counting the number of Failures before the first Success.

3.6.9. Example (1.13 in Grinstead and Snell, 1997).

PROBLEM. *A fair coin is tossed until the first time a head turns up, let* X *denote the toss at which this occurs. Show that* $X \in \mathrm{G}_1(1/2)$.

Solution. A possible probability space Ω to model the experiment is

$$\Omega := \left\{(\omega_i)_{i\in\{i\ldots 1\}N} :\ \omega_i = \mathrm{T} \text{ for all } i = 1, \ldots, N-1 \text{ and } \omega_N = \mathrm{H}\right\}, \tag{3.6.28}$$

[∗]: Check! which is in a one-to-one correspondence to $\mathbb{N}$.[∗] The probability measure may be defined to be

$$\mathcal{P}[N] = \frac{1}{2^N}. \tag{3.6.29}$$

Then the random variable X is quite simply the identity, i.e., $X(N) = N$, for all $N \in \mathbb{N}$, with $X \in \mathrm{G}_1(1/2)$.
An alternative solution would be to use for Ω the set of all the infinite sequences of H's and T's, i.e., $\Omega := \{\mathrm{H,T}\}^{\mathbb{N}} = \left\{(\omega_i)_{i\in\mathbb{N}} :\ \omega_i = \mathrm{H} \text{ or } \omega_i = \mathrm{T} \quad \forall\, i \in \mathbb{N}\right\}$, with the probability given by the atomic probability of having H at the N-th toss (without any knowledge on all the other tosses)

$$\mathcal{P}[\omega_N = \mathrm{H}] = \frac{1}{2}, \text{ for each } N \in \mathbb{N}. \tag{3.6.30}$$

Then we define X, for each $\omega = (\omega_i)_{i\in\mathbb{N}} \in \Omega$ as follows

$$X(\omega) := \min\{m \in \mathbb{N} :\ \omega_m = \mathrm{H}\}. \tag{3.6.31}$$

Then we have

$$\mathcal{P}[X=N] = \mathcal{P}[\omega_1 = \cdots = \omega_{N-1} = \mathrm{T} \text{ and } \omega_N = \mathrm{H}] = \underbrace{1/2 \times \cdots \times 1/2}_{N-1 \text{ times}} \times \frac{1}{2} = \frac{1}{2^N}, \tag{3.6.32}$$

and $X \in \mathrm{G}_1(1/2)$.

3.6.10. Example.

PROBLEM. *Put one White ball in a bag. Throw a fair die repeatedly, each time inserting a Black ball and stopping after the first Six. Then select one ball at random from the bag. Find the chance it is White.*

Solution. Although not necessary to solve the problem, it is a useful exercise to set-up an appropriate probability space for this experiment. We use outcomes with two components (or coordinates) the first indicates the number of black balls and the second one the extracted ball.

$$\mathit{\Omega} = \{\boldsymbol{\omega} = (\omega_1, \omega_2) : \ \omega_1 \in \mathbb{N} \text{ and } \omega_2 \in \{\bigcirc, \bullet\}\} \tag{3.6.33}$$

(Note that there must be at least one Black ball in any outcome.) Now introduce the following events

$$A := \left\{\text{get a white ball}\right\} = \{\boldsymbol{\omega} \in \mathit{\Omega} : \ \omega_2 = \bullet\}, \tag{3.6.34}$$

and

$$B_k := \{\text{There are } k \text{ Black balls when we stop}\} = \{\boldsymbol{\omega} \in \mathit{\Omega} : \ \omega_1 = k\}. \tag{3.6.35}$$

If $k \neq l$ then $B_k \cap B_l = \varnothing$ and $\bigcup_{k=0}^{\infty} B_k = \mathit{\Omega}$, so by the Partition of Total Probability Theorem 2.2.2 we have

$$\mathcal{P}(A) = \sum_{k=1}^{\infty} \mathcal{P}(A|B_k)\mathcal{P}(B_k). \tag{3.6.36}$$

But $\mathcal{P}(A|B_k) = 1/(k+1)$ while $\mathcal{P}(B_k) = 1/6(5/6)^{k-1}$, hence

$$\mathcal{P}(A) = \sum_{k=1}^{\infty} \frac{1}{(k+1)} \frac{1}{6} \left(\frac{5}{6}\right)^{k-1} = \frac{1}{6} \left(\frac{5}{6}\right)^{-2} \sum_{k=1}^{\infty} \frac{1}{(k+1)} \left(\frac{5}{6}\right)^{k+1}. \tag{3.6.37}$$

Recall now from basic calculus of power series that, under absolute convergence conditions, we have

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{1}{(k+1)} x^{k+1} &= \sum_{k=1}^{\infty} \int_0^x \xi^k \,\mathrm{d}\xi = \int_0^x \sum_{k=1}^{\infty} \xi^k \,\mathrm{d}\xi \\ &= \int_0^x \frac{\xi}{1-\xi} \,\mathrm{d}\xi = \int_{1-x}^{1} \frac{1-\zeta}{\zeta} \,\mathrm{d}\zeta = \left[\log\zeta - \zeta\right]_{1-x}^{1} = \log\frac{1}{1-x} - x. \end{aligned} \tag{3.6.38}$$

Thus, taking $x = 5/6$, we conclude that

$$\mathcal{P}(A) = \frac{6}{25} \left(\log 6 - \frac{5}{6}\right) \approx 0.23002. \tag{3.6.39}$$

3.6.11. Definition of negative binomial distribution. The random variable X has the *negative binomial distribution*, $X \in \mathrm{NB}_0(r, p)$, if and only if X is discrete, with values x_k, $k \in \mathbb{N}_0$, $x_k \neq x_j \iff k \neq j$, and

$$\mathcal{P}[X = x_k] = \binom{k+r-1}{r-1} p^r (1-p)^k \quad \forall\, k \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}. \tag{3.6.40}$$

We say that a probability measure on an infinitely countable space $\mathit{\Omega} = \{\omega_k\}_{k \in k \in \mathbb{N}_0}$ is distributed following the negative binomial distribution if the random variable $X(\omega_k) := k$ is so.

3.6.12. Example.

PROBLEM. *Consider the sample space of sequences of coin tosses*

$$\Omega = \left\{ ! = (\omega_i)_{i\in\mathbb{N}} : \ \omega_i \in \{\mathrm{H}, \mathrm{T}\} \quad \forall\, i \in \mathbb{N} \right\}, \tag{3.6.41}$$

with the usual product probability measure $\mathcal{P}$ *built upon the Bernoulli trial probability*

$$\mathcal{P}_0(\mathrm{H}) = p \ \textit{and}\ \mathcal{P}_0(\mathrm{T}) = 1 - p =: q. \tag{3.6.42}$$

(a) Define the random variable X_2 *which counts the number of* T*'s before the* second *time* H *appears in the sequence, in symbols:*

$$X_2(!) = \max\{n \in \mathbb{N} : \#\{i < n : \ \omega_i = \mathrm{H}\} = 1\} - 2 \ \textit{for each}\ \omega = (\omega_i)_{i\in\mathbb{N}} \in \Omega. \tag{3.6.43}$$

Calculate the distribution of X_2.

(b) Next define the random variable X_r *which counts the time we get the* r*-th* H.

Solution. (a) The event that $X_2(!) = k$ means that $\omega_{k+2} = \mathrm{H}$ while $\omega_i = \mathrm{T}$ for all $i \leq k+1$ but one, say $i = j$. In set-theoretical notation we may write this as the disjoint union

$$\{X_2 = k+2\} = \bigcup_{j=1}^{k+1} A_j \tag{3.6.44}$$

where for each fixed $j = 1, \dots, k+2$ we defined

$$A_j := \left\{ ! \in \Omega : \ \omega_i = \begin{cases} \mathrm{H} \text{ for } i = j \text{ or } k+2 \\ \mathrm{T} \text{ for } i \in \{1 \dots k+1\} \setminus \{j\} \end{cases} \right\}. \tag{3.6.45}$$

It follows that

$$\mathcal{P}(A_j) = p^2 q^k, \tag{3.6.46}$$

which is independent of j. Hence, by additivity, we get

$$\mathcal{P}[X_2 = k] = \sum_{j=1}^{k+1} \mathcal{P}(A_j) = (k+1) p^2 q^k = \binom{k+2-1}{2-1} p^2 q^k, \tag{3.6.47}$$

as there are $k-1$ ways of choosing j amongst the integers less than k.

(b) For the random variable X_r we proceed similarly:

$$\{X_r = k + r\} = \bigcup_{J \in \mathscr{J}} A_J \tag{3.6.48}$$

where $\mathscr{J}$ is the collection of all possible subsets of $\{1 \dots k + r - 1\}$ with $r-1$ elements, $J = \{ j_\alpha : \ \alpha = 1, \dots, r-1 \}$, $j_\alpha < j_\beta$ and $J \subset \{1 \dots k + r - 1\}$, and

$$A_J := \left\{ ! \in \Omega : \ \omega_i = \begin{cases} \mathrm{H} \text{ for } i \in J \\ \mathrm{T} \text{ for } i \in \{1 \dots k + r - 1\} \setminus J \end{cases} \right\}. \tag{3.6.49}$$

Since the set J can be chosen in $\binom{k+r-1}{r-1}$ ways and each single outcome therein has exactly (and independently) r times H's and k times T's, and thus probability $p^r q^k$, then

$$\mathcal{P}(A_J) = \binom{k+r-1}{r-1} p^r q^k. \tag{3.6.50}$$

> *Moral: a negative binomial random variable* $X \in \mathrm{NB}_0(r,p)$ *that evaluated at* k*, gives the probability of* k *fails before the* r*-th success in a sequence of Bernoulli trials.*

3.6.13. Remark. Note that $\mathrm{NB}_0(1,p)$ is the same as $\mathrm{G}_1\left(p\right)$.

3.6.14. Definition of Poisson distribution. A random variable X has the *Poisson distribution,* if it is integer valued and satisfies

$$\mathcal{P}[X=k]=\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} \text{ for } k=0,1,2,\dots. \tag{3.6.51}$$

We write $X \in \mathrm{Poisson}(\lambda)$.

3.6.15. Remark. Poisson's distribution is a good model for a random variable that counts the number of occurences of a certain phenomenon: e.g., the number of accidents on a road in a day, the number of emails received by a user on a certain week, etc.

3.6.16. Poisson's distribution is a limit of the Binomial distribution. To build an example of the Poisson distribution, consider a Bernoulli trial, $X_0 \in \mathrm{B}(1,p)$.

3.6.17. Example.

PROBLEM. *Each packet of cornflakes contains just one of a set of N toys, all equally likely; toys in different packets are independent. Find the mean and variance of the number of cornflakes packets to buy to obtain a complete set.*

Solution. Let $X_i =$ Number of packets to buy to get the next toy, when you now have $i-1$ different toys, for $i=1,2,\cdots,N$.
Plainly, $X_1 = 1$, and $S = X_1 + X_2 + \ldots + X_N$ is the number to buy. Also, these X_i are independent. For the distribution of X_i with $i > 1$, observe that to have $X_i = k$, we first buy $k-1$ packets with toys we already have, then buy one with a new toy—a sequence of Bernoulli trials leading to the Geometric distribution $\mathrm{G}_1(N-i+1/N)$ (yes?).
We have formulae for the mean and variance of such variables, so use parts (i) and (iii) of Theorem 4.6.1 with each $c_i = 1$. Cunning manipulation leads to

$$\mathrm{E}[S]=N\sum_{i=1}^{N}\frac{1}{i} \tag{3.6.52}$$

and

$$\mathrm{var}[S]=N^2\sum_{i=1}^{N}\frac{1}{i^2}-\mathrm{E}[S], \tag{3.6.53}$$

from which useful approximations can be found.

3.6.18. Indicator random variables revisited. Indicator random variables, that were introduced earlier, can be quite useful to solve problems, that would otherwise be harder to crack. A very useful property of indicator variables is the following, suppose A and B are two events then $\mathbb{1}_A\mathbb{1}_B$ is the indicator variable of the event $A\cap B$.

3.6.19. Example (indicator variables in action).

PROBLEM. *Peter and Paul each have an ordinary well-shuffled pack of n cards each, $n > 1$. They form pair of cards by exposing the cards one at a time, simultaneously. How many of these n pairs would you expect to match?*

Solution. Let $I_r := \mathbb{1}_{\{r\text{-th pair is a match}\}}$, i.e.,

$$I_r = \begin{cases} 1 & \text{if the } r\text{-th pair of cards match} \\ 0 & \text{otherwise.} \end{cases} \tag{3.6.54}$$

The total number of matches is thus given by the random variable

$$S := \sum_{r=1}^{n} I_r. \tag{3.6.55}$$

Since $\mathcal{P}[I_r = 1] = 1/n$, the mean and variance of I_r are easily calculated using the Indicator Variable's Statistics

$$\mathrm{E}[I_r] = \frac{1}{n} \text{ and } \mathrm{var}[I_r] = \frac{1}{n}\left(1 - \frac{1}{n}\right) = \frac{n-1}{n^2} \tag{3.6.56}$$

But what about $\mathrm{cov}[I_r, I_s]$ when $r \neq s$? By definition, writing $p = 1/n$, we have

$$\mathrm{cov}[I_r, I_s] = \mathrm{E}\big[(I_r - p)(I_s - p)\big] = \mathrm{E}[I_r I_s] - p\,(\mathrm{E}[I_r] + \mathrm{E}[I_s]) + p^2 = \mathrm{E}[I_r I_s] - p^2. \tag{3.6.57}$$

Now, given that I_r and I_s are indicator variables, then also $I_r I_s$ is and

$$I_r I_s = [\![r\text{-th and } s\text{-th pairs match}]\!] = \mathbb{1}_{\{I_r=1\}\cap\{I_s=1\}} \tag{3.6.58}$$

it follows that

$$\mathrm{E}[I_r I_s] = \mathcal{P}(\{I_r = 1\} \cap \{I_s = 1\}) \tag{3.6.59}$$

Recall

$$\mathcal{P}(\{I_r = 1\} \cap \{I_s = 1\}) = \mathcal{P}[I_r = 1 | I_s = 1]\,\mathcal{P}[I_s = 1]. \tag{3.6.60}$$

We already know $\mathcal{P}(I_s = 1) = 1/n$, and then realise that

$$\mathcal{P}[I_r = 1 | I_s = 1] = \frac{1}{n-1}. \tag{3.6.61}$$

Thus $P(I_r = 1 \cap I_s = 1) = 1/(n(n-1))$.
Hence $\mathrm{E}[I_r I_s] = \mathcal{P}(I_r = 1 \cap I_s = 1) = 1/(n(n-1))$, from which $\mathrm{cov}[I_r, I_s] = 1/(n^2(n-1))$ and manipulation leads to $\mathrm{E}[S] = 1$, $\mathrm{var}[S] = 1$.

Exercises and problems on Discrete random variables

Exercise 3.1. Assuming that X and Y are discrete random variables, prove the following statement.
Let $X, Y : \Omega \to \mathbb{R}$ be two discrete random variables on a probability space $(\Omega, \mathcal{P}, \mathscr{F})$ then we have

(i) *monotonicity* of E

$$X \leq Y \Rightarrow \mathrm{E}[X] \leq \mathrm{E}[Y], \tag{P3.1.1}$$

and

$$X < Y \Rightarrow \mathrm{E}[X] < \mathrm{E}[Y]. \tag{P3.1.2}$$

($X \leq Y$ means $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, and similarly for $X < Y$.)

(ii) for any $\alpha, \beta \in \mathbb{R}$ we have

$$\mathrm{E}\big[\alpha X + \beta Y\big] = \alpha\,\mathrm{E}[X] + \beta\,\mathrm{E}[Y]. \tag{P3.1.3}$$

Exercise 3.2 (algebra of expectation and variance)**.** Let $(X_1, \ldots, X_n)$ be a random vector and $(c_1, \ldots, c_n) \in \mathbb{R}^n$. Show that

(i) the expectation and the linear combination commute

$$\mathrm{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i\,\mathrm{E}[X_i]; \tag{P3.2.1}$$

(ii) the variance is quadratic

$$\mathrm{var}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2\,\mathrm{var}[X_i] + 2 \sum_{1 \le i < j \le n} c_i c_j\,\mathrm{cov}\big[X_i, X_j\big]; \tag{P3.2.2}$$

(iii) and, if the random vector $(X_i)_{i \in \{i \ldots 1\} n}$ is independent then

$$\mathrm{var}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2\,\mathrm{var}[X_i]. \tag{P3.2.3}$$

.

Problem 3.3. Prove the following result.

THEOREM (discrete Chebyshev inequality). *Let X be a discrete random variable. If X has finite mean and variance, then for any $\alpha \in \mathbb{R}^+$ we have*

$$\mathcal{P}[|X - \mathrm{E}[X]| > \alpha] \le \frac{\mathrm{var}[X]}{\alpha^2}. \tag{P3.3.1}$$

Hint. Start by proving that for any $\alpha \ge 0$

$$\alpha^2 \mathcal{P}[\alpha \le |X|] \le \mathrm{E}\big[X^2\big]. \tag{P3.3.2}$$

Exercise 3.4. If each of 12 jurors, independently, has probability 0.9 of deciding the accused is guilty, find the chance of a majority Guilty verdict of $10-2$ or higher.

Exercise 3.5. In both parts, $X \in \mathrm{Poisson}(\lambda)$.

(a) A red marble is placed in an urn, and X blue marbles are added. One marble is selected at random from those in the urn, all equally likely. Find, in terms of λ, the probability that it is the red one.

(b) Find $\mathrm{E}[1/(1+X)]$.

Exercise 3.6 (multiplicativity of expectations and independence not equivalent)**.** Let X and Y be two random variables.

(a) Condition

$$\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y] \tag{P3.6.1}$$

is only necessary, but not sufficient for the random variables X and Y to be independent. Convince yourself of this fact by finding two random variables X and Y that are *not independent* yet satisfying (P3.6.1).

(b) Although (P3.6.1) is only necessary, it often becomes sufficient if coupled with some more assumptions on X and Y. For example, if X and Y are two Bernoulli trials taking each the value 0, and satisfying (P3.6.1), then X and Y are independent. Prove it.

Exercise 3.7. An urn initially has one White ball and one Black ball. Peter selects one ball at random, and replaces it in the urn, together with another ball of the same colour. He repeats this process, all selections being at random from the balls then in the urn, until he selects the White ball for the first time. Let X denote the number of Black balls in the urn at this moment: show that $\mathcal{P}[X=k] = 1/(k(k+1))$ for $k = 1, 2, 3, \ldots$.
Peter replaces the White ball, along with another White ball as usual, and chooses one more ball at random. Show that the overall chance that this ball is Black is $1/2$.

CHAPTER 4

General random variables and expectation

In Chapter 3 we have introduced discrete random variables, but many phenomenons in real life are better described with a continuum-valued (i.e., real-valued of $\mathbb{R}$-valued) random variables, which we shall call simply *random variables*. In fact, discrete random variables are but a special case of (continuum-valued) random variables.

Just as discrete random variables can be summed up to define expectation and variance. Summation of discrete is also useful to calculate other *moments*. In this chapter we extend the concept of summation from discrete to general random variables by using general *measure* and *integration* theory. Because a full development of measure and integration requires a whole course, we just outline this procedure and keep the technicalities at the intuitive level. In this spirit, we note that the random variable is not discrete, summation usually goes by the name of integration, but many authors interchange integration and *summation* as this is still a legitimate (and perhaps more intuitive) terminology. The same goes for *integral* (with resepect to a probability measure or density), *average* and *expectation*: these are all interchangeable terms.

4.1. What are random variables?

4.1.1. Motivating example. Many phenomenons that can be modelled with probability spaces may be very complex. For example, think of the following experiment: watching the traffic on a stretch of a road for a given day. A reasonable probability space Ω here is the space of all possible histories (of movies) of all traffic for the given day from 00:00 until 24:00. That is all the possible scenarios that traffic could have during that day. This is a huge amount of information, that could be very complicated to track or process (even for a computer with a very performant software it would take a tremendous amount of memory to try and list all the possible scenarios). We are clearly wasting memory. To save us some time and effort, we should focus only on the "important" bits of traffic. For example, we could be look at the "average speed". Or we could count the amount of accidents that happen during that day. Each outcome (i.e., a possible traffic scenario) $\omega \in \Omega$ will determine these quantities exactly. Mathematically speaking the average speed is a function of the outcome (i.e., the history of the traffic on that day) so let's call it $S(\omega)$ for $\omega \in \Omega$. Similarly, the number of accidents for a given outcome ω could be denoted by $X(\omega)$. The advantages of using these quantities, over looking at the whole history:

- ⋆ we are using numbers (a real one for $S(\omega)$ and an integer for $X(\omega)$) whereas to describe ω we need something much more complicated,

- ⋆ many ω's in Ω will lead to the same value of S or X (e.g., you could have very different histories, yet end up with 3 accidents, this means $X(\omega)=X(\omega')=3$ but with $\omega\neq\omega'$).
- ⋆ S and X are easier to "observe" in practice: for S you need a speed camera and for X a human that identifies collision.

4.1.2. Definition of random variable (also known as random number). Given a probability space $(\Omega,\mathscr{F},\mathcal{P})$, a function $X:\Omega\to\mathbb{R}$ is called $\mathcal{P}$*-random variable* (or $\mathcal{P}$*-random number*) if and only if

$$\mathcal{P}[X\leq x] \text{ where } \{X\leq x\} \text{ is short for } \{\omega\in\Omega:\ X(\omega)\leq x\} \tag{4.1.1}$$

is and event in $\mathscr{F}$ for all real x. In symbols

$$X^{-1}(-\infty,x]\in\mathscr{F}\quad\forall\, x\in\mathbb{R}. \tag{4.1.2}$$

Often the probability measure $\mathcal{P}$ is understood from context and we say simply random variable. Random variables are sometimes also called *random numbers*, to emphasise their numerical nature.[1]

4.1.3. Definition of random vector. A *random vector* $\boldsymbol{X}=(X_1,\dots,X_n)$ similarly is a function with values in $\mathbb{R}^n$ such that $\mathcal{P}[\boldsymbol{X}\leq\boldsymbol{x}]$ is defined for all vectors $\boldsymbol{x}\in\mathbb{R}^n$. This is equivalent to saying that for each $i=1,\dots,n$, X_i is a random variable.

4.1.4. Remark (why do we bother?) Definitions 4.1.2 and 4.1.3 are technical, needed for mathematical rigour when developping the theory of probability on general probability spaces. In practice, "any reasonable" function X will be OK, and we usually suppress the probability space when working. But it can be useful to invoke it, especially when several random variables are related.

4.2. Distribution functions

4.2.1. Definition of distribution function. For any random variable X, its *(cumulative) distribution function* is $\operatorname{cdf}_X(x):=F_X(x):=\mathcal{P}[X\leq x]$. We usually prefer the second notation (the first is used when some other variable named F happens to be around, in order to avoid confusion). We also suppress the subscript X where convenient, writing $F(x)=\mathcal{P}[X\leq x]$.

4.2.2. Theorem (basic properties of distribution functions). *Let X be a random variable on the probability space $(\Omega,\mathscr{F},\mathcal{P})$ and let $F=\operatorname{cdf}_X$ denote its distribution function. The following hold:*

(i) F is monotone increasing, i.e.,

$$x\leq y\ \Rightarrow\ F(x)\leq F(y).$$

(ii) $\lim_{x\to-\infty}F(x)=0$ *and* $\lim_{x\to\infty}F(x)=1$*;*

(iii) F is ladlag, i.e., it has a left and right limits at each $x_*\in\mathbb{R}$, i.e., both $\lim_{x\to x_*^-}F(x)$ *and* $\lim_{x\to x_*^+}F(x)$ *exist (but need not coincide);*

[1] The concept of $\mathcal{P}$-random variable can be generalised to other "$\mathcal{P}$-random things", where "things" can be a vector, a function, a measure, a set or any other mathematically meaningful element of a measure or topological space.

(iv) And F is cad, or right continuous, i.e.,

$$\lim_{x\to x_*^+} F(x) = F(x_*) \tag{4.2.1}$$

or, equivalently, as x_n converges to x_* from above, so $F(x_n)$ converges to $F(x_*)$,

(v) For $x_1 < x_2$ we have

$$\mathcal{P}[x_1 < X \le x_2] = F(x_2) - F(x_1). \tag{4.2.2}$$

4.2.3. Remark (one-sided continuity of distribution functions, ladlag and cadlag). In (iii), the limits need not be equal; take for example, a probablity space Ω and a *constant (or certain) random variable* $X(\omega) = 0$, for all $\omega \in \Omega$. Then

$$F(x) := \mathrm{pdf}_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \ge 0. \end{cases} \tag{4.2.3}$$

This function F satisfies all properties in Theorem 4.2.2, but

$$0 = \lim_{x\to 0^-} F(x) \ne \lim_{x\to 0^+} F(x) = 1. \tag{4.2.4}$$

A function that has a right and left limit at each point is called *ladlag*, or more properly *làdlàg*, from the French "limitée à droite et limitée à gauche". A function that is furthermore right-continuous is called *càd* for "continue à droite", hence Theorem 4.2.2 says that a distribution function is *càdlàg*, or (having pity on non-French typists) simply *cadlag*.

4.3. Random vectors, joint distributions and independence

4.3.1. Definition of indepedent random vector components. The idea of random variables being *independent* is that knowledge about one of them has no bearing on the values of the others. More precisely, we write:
The components of the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ are *independent* if, for all vectors $\boldsymbol{x} = (x_1, \dots, x_n)$, we have

$$\begin{aligned} &\mathcal{P}[X_1 \le x_1 \text{ and } X_2 \le x_2 \text{ and } \dots \text{ and } X_n \le x_n] \\ &\qquad = \mathcal{P}[X_1 \le x_1]\mathcal{P}[X_2 \le x_2] \cdots \mathcal{P}[X_n \le x_n]. \end{aligned} \tag{4.3.1}$$

4.3.2. Proposition (characterisation of independent random variables). *A random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is independent if and only if*

$$\mathcal{P}[\boldsymbol{X} \in A_1 \times \cdots \times A_n] = \prod_{i=1}^n \mathcal{P}[X_i \in A_i], \quad \forall A_1, \dots, A_n \in \mathscr{B}^1. \tag{4.3.2}$$

4.3.3. Vector manipulation. This is a good point to introduce some vector manipulation operations. Let $d \in \mathbb{N}$. If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we define

- ⋆ $\boldsymbol{x} < \boldsymbol{y}$ if and only if $x_i < y_i$ for all $i = 1, \dots, d$,
- ⋆ $\boldsymbol{x} \le \boldsymbol{y}$ if and only if $x_i \le y_i$.

As for numbers $\boldsymbol{x} \le \boldsymbol{y} \Rightarrow \boldsymbol{x} < \boldsymbol{y}$, but, unlike real numbers, *not all vectors can be compared.* For example if $\boldsymbol{x} = (1,2)$ and $\boldsymbol{y} = (2,1)$, then neither $\boldsymbol{x} \le \boldsymbol{y}$ nor $\boldsymbol{y} \le \boldsymbol{x}$ holds true.
For each $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $x \in \mathbb{R}$ we define

$$\begin{aligned} \tilde{\boldsymbol{x}}_i(x) &:= \left[y_j\right]_{j=1,\dots,d} \in \mathbb{R}^d : y_j := x_j \llbracket i \ne j \rrbracket + x \llbracket i = j \rrbracket \\ \boldsymbol{x}'_i &= \left[z_j\right]_{j=1,\dots,d-1} \in \mathbb{R}^{d-1} : z_j := x_j \llbracket j < i \rrbracket + x_{j+1} \llbracket j \ge i \rrbracket. \end{aligned} \tag{4.3.3}$$

In words, $\tilde{\boldsymbol{x}}_i(x)$ equals $\boldsymbol{x}$ with its i-th component replaced by x, while $\boldsymbol{x}'_i$ equals $\boldsymbol{x}$ with its i-th component removed (also known as "i-pop" in computer science).
Also for a vector $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$, we write $\boldsymbol{x} \to +\boldsymbol{\infty}$ (in limits) as shorthand for $x_i \to +\infty$ for all i. Similar meaning for $\boldsymbol{x} \to -\boldsymbol{\infty}$.

4.3.4. Definition of joint distribution. Consider the random vector $\boldsymbol{X} = (X_1, \dots, X_d)$, we define its *(cumulative) distribution function,* (also known as *joint distribution function of* $X_1, \dots, X_d$), as the function F given by

$$F(x_1, \dots, x_d) := \mathcal{P}[X_i \le x_i \quad \forall\, i = 1, \dots, n]. \tag{4.3.4}$$

We will denote the function F as $\operatorname{cdf}_{\boldsymbol{X}}$.

4.3.5. Theorem (marginal (cumulative) distribution). *Let* $\boldsymbol{X} = (X_i)_{i \in \{i \dots 1\}d}$ *be a random vector. Then, for each* $i = 1, \dots, n$, *and each* $x \in \mathbb{R}$, *the following limit exists:*

$$\lim_{\boldsymbol{x}'_i \to \infty} \operatorname{cdf}_{\boldsymbol{X}}(\tilde{\boldsymbol{x}}_i(x)), \tag{4.3.5}$$

and we have

$$\operatorname{cdf}_{X_i}(x) = \lim_{\boldsymbol{x}'_i \to \infty} \operatorname{cdf}_{\boldsymbol{X}}(\tilde{\boldsymbol{x}}_i(x)). \tag{4.3.6}$$

The limit defined by (4.3.5) is called a marginal distribution and is often denoted F_{X_i}, *or* F_i, *for satisfying property (4.3.6). Interpreting the notation in (4.3.5) is essential: it means that the* i*-th entry of* $\boldsymbol{x}$ *is "frozen" while all the other ones are sent to infinity.*
Proof The proof is left as an exercise. □

4.3.6. Example (marginal distribution). Suppose the random vector (X, Y) has

$$\operatorname{cdf}_{(X,Y)}(x, y) = \left(1 - \mathrm{e}^{-(x+y)}\right) \mathbb{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(x, y). \tag{4.3.7}$$

What is cdf_X?
Fixing $x \in \mathbb{R}$ and taking $y \to +\infty$, by the Marginal Distribution Theorem 4.3.5, we have

$$\begin{aligned} \operatorname{cdf}_X(x) &= \lim_{y \to \infty} \operatorname{cdf}_{(X,Y)}(x, y) = \lim_{y \to \infty} \left(1 - \mathrm{e}^{-(x+y)}\right) \mathbb{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(x, y) \\ &= \lim_{y \to \infty} \left(1 - \mathrm{e}^{-x}\, \mathrm{e}^{-y}\right) \mathbb{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(x, y) = \left(1 - \mathrm{e}^{-x}\right) \mathbb{1}_{\mathbb{R}^+}(x). \end{aligned} \tag{4.3.8}$$

4.4. Simple random variables and their expectation

4.4.1. Definition of simple random variable. A random variable X is called *simple* if for a finite collection of events $A_1, \dots, A_n$ and a corresponding collection of real numbers $a_i \in \mathbb{R}$, such that

$$X(\omega) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\omega) \text{ for } \omega \in \Omega. \tag{4.4.1}$$

Equivalently X is a simple random variable if and only if $X \in \mathrm{DRV}(\mathcal{P})$ and it takes *only finitely many values.*

4.4.2. Remark. The events A_i appearing can be taken to be disjoint, and the values a_i all different and nonzero. Indeed, suppose this is not the case, we can build such a "choice" A'_j and a'_j, $j = 1, \dots, m$, as follows. Noting that the variable X is finite (i.e., it takes finitely many values) these are at most all the elements of the set

$$C := \left\{ \sum_{i=1}^{n} \sigma_i a_i : \sigma \in \{0,1\}^n \right\}, \tag{4.4.2}$$

which is clearly finite of cardinality less or equal to 2^n. Since $X(\Omega) \subseteq C$, it is finite and we may write it as

$$X(\Omega) := \left\{ a'_j : j = 1, \dots, m \right\} \text{ where } a'_j < a'_{j+1} \quad \forall\, j = 1, \dots, m-1. \tag{4.4.3}$$

Now define, for each $j = 1, \dots, m$

$$A'_j := X^{-1}\left(\left\{a'_j\right\}\right), \tag{4.4.4}$$

who must all be in $\mathscr{F}$ given X is a random variable. It now follows that

$$X = \sum_{j=1}^{m} a'_j \mathbb{1}_{A'_j}. \tag{4.4.5}$$

Moreover, the representation of the simple random variable X via (4.4.5) with disjoint events A'_j, $j = 1, \dots, n$, and strictly increasing a'_j's is unique, i.e., if (4.4.5) holds and also

$$X = \sum_{k=1}^{l} a''_k \mathbb{1}_{A''_k}, \tag{4.4.6}$$

with the a''_k's all different and increasing and A''_k's disjoint, then we have

$$m = l \text{ and } a'_j = a''_j, A'_j = A''_j \quad \forall\, j = 1, \dots, m. \tag{4.4.7}$$

We will refer to this unique representation of a simple random variable as its *simplest representation.*

4.4.3. Remark (why a new name for the same old thing?) Another name for simple random variable is *finite (discrete) random variable*, which is a special case of discrete random variables that we covered extensively in Chapter 4. The reason we employ another name comes from the theory of integration, where the terminology *simple functions* is almost universally employed.

Given that we have covered simple random variables, we already defined the expectation of a simple random variable $X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$, as

$$\mathrm{E}\, X := \mathrm{E}[X] := \sum_{x \in X(\Omega)} x \mathcal{P}(X = x), \tag{4.4.8}$$

which implies that

$$\mathrm{E}\, X = \sum_{i=1}^{n} a_i \mathcal{P}(A_i). \tag{4.4.9}$$

4.5. Expectation

4.5.1. Integration by approximation. In order to extend the definition of expectation for simple random variables to all random variables we use *approximation from below.* Let X be a general random variable and let us assume that X is almost surely nonnegative, i.e., $\mathcal{P}[X < 0] = 0$. Then it is possible to find a sequence of simple random variables X_n, $n \in \mathbb{N}$, such that $X_n \le X_{n+1}$ almost surely, for all n and $\sup X_n = X$. By the monotonicity of expectation for simple random variables, 3.2.4, it follows that $\mathrm{E}\, X_n \le \mathrm{E}\, X_{n+1}$, so it would be natural to define

$$\mathrm{E}\, X := \mathrm{E}[X] := \sup_n \mathrm{E}\, X_n. \tag{4.5.1}$$

The problem with this definition is that the value of $\mathrm{E}\, X$ might depend on the particular choice of approximating sequence X_n. In fact, it is possible to show that there is no such risk, but it is laborious. An easier way to define $\mathrm{E}\, X$ is to consider the set of *all* simple random variables Y such that $Y \le X$ (this will include any sequence approaching X from below) and then take the least upper bound of the set of their expectations to be the expectation of X. In symbols:

$$\mathrm{E}\, X := \mathrm{E}[X] := \sup \{\mathrm{E}\, Y : Y \text{ simple random variable and } 0 \le Y \le X\}. \tag{4.5.2}$$

The least upper bound in (4.5.2) always exists: it is either a real number, in which case we say that X has *finite expectation*; or, it is ∞ in which case X is said to have *infinite expectation.*

4.5.2. Exercise. *Let X be a nonnegative discrete random variable with infinitely (countably) many different values. Using the definition of $\mathrm{E}\, X$ as the supermum of the expectation of all the underlying simple random variables, show that X has finite expectation if and only if*

$$\sum_{x \in X(\Omega)} x \mathcal{P}[X = x] \tag{4.5.3}$$

converges and coincides with $\mathrm{E}\, X$.

4.5.3. Definition of positive and negative parts. For any $x \in \mathbb{R}$ we define the *positive part* and *negative part* functions, respetively, as follows:

$$[x]_+ := 0 \vee x \text{ and } [x]_- := 0 \vee -x. \tag{4.5.4}$$

Immediate properties of these functions are

$$[x]_- = -0 \wedge x = [-x]_+ \text{ and } [x]_+ [x]_- = 0. \tag{4.5.5}$$

Furthermore, using the absolute value

$$|x| = [x]_+ + [x]_- , \, [x]_+ = \frac{|x| + x}{2} \text{ and } [x]_- = \frac{|x| - x}{2}. \tag{4.5.6}$$

If X is a random variable then it can be decomposed into the following *positive-negative parts decomposition*

$$X = [X]_+ - [X]_- . \tag{4.5.7}$$

4.5.4. Definition of finite expectation, integrable, summable random variables, L_1. A random variable X is said to have *finite expectation*, or to be *integrable* or *summable*, if and only if $\mathrm{E}[[X]_+]$ and $\mathrm{E}[[X]_-]$, as supremums defined by (4.5.2), are both finite. This is equivalent to saying that $\mathrm{E}|X|$ is finite.

We indicate with $\mathscr{L}^1(\Omega, \mathscr{F}, \mathcal{P})$ or simply $\mathscr{L}^1(\mathcal{P})$ or $\mathscr{L}^1(\Omega)$ the set of all random variables on $(\Omega, \mathscr{F})$ that have finite expectation with respect to $\mathcal{P}$.

We say that two random variables X, Y are *$\mathcal{P}$-almost surely equal* if and only if

$$\mathcal{P} X = Y = 1. \tag{4.5.8}$$

It is an exercise to check that $\mathcal{P}$-almost sure equality is an equivalence relation in $\mathscr{L}^1(\mathcal{P})$ and hence we define the quotient set

$$\mathrm{L}_1(\mathcal{P}) := \mathscr{L}^1(\mathcal{P}) / \equiv_{\mathcal{P}\text{-a.s.}} . \tag{4.5.9}$$

When $\mathcal{P}$ is clear from the context, we write $\mathrm{L}_1(\Omega)$.

4.5.5. Theorem (monotonicity and linearity of expectation). *For any $X, Y \in \mathscr{L}^1(\mathcal{P})$, $a, b \in \mathbb{R}$ we have*

$$X \leq Y \Rightarrow \mathrm{E} X \leq \mathrm{E} Y, \tag{4.5.10}$$

and

$$\mathrm{E}[a X + b Y] = a \, \mathrm{E}[X] + b \, \mathrm{E}[Y]. \tag{4.5.11}$$

Proof For each fixed $\varepsilon > 0$ there is a simple random variable S_ε such that $S_\varepsilon \leq X$ and $\mathrm{E} S_\varepsilon \geq \mathrm{E} X + \varepsilon$. Since $S_\varepsilon \leq Y$ it follows that $\mathrm{E} S_\varepsilon \leq \mathrm{E} Y$ and thus

$$\mathrm{E} X + \varepsilon \leq \mathrm{E} Y. \tag{4.5.12}$$

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\mathrm{E} X \leq \mathrm{E} Y. \tag{4.5.13}$$

Linearity follows from the linearity for simple functions and approximation. Suppose first that $X, Y \geq 0$ and $\alpha, \beta \geq 0$ then, for each $\varepsilon > 0$, there are simple functions X_ε and Y_ε such that

$$\mathrm{E} X - \varepsilon < \mathrm{E} X_\varepsilon \leq \mathrm{E} X \text{ and } \mathrm{E} Y - \varepsilon < \mathrm{E} Y_\varepsilon \leq \mathrm{E} Y. \tag{4.5.14}$$

Write $Z := \alpha X + \beta Y$ and $Z_\varepsilon := \alpha X_\varepsilon + \beta Y_\varepsilon$, it follows that

$$\mathrm{E} Z \geq \mathrm{E} Z_\varepsilon = \alpha \mathrm{E} X_\varepsilon + \beta \, \mathrm{E} Y_\varepsilon \geq \alpha \mathrm{E} X + \beta \, \mathrm{E} Y - (\alpha + \beta)\varepsilon, \tag{4.5.15}$$

hence

$$\alpha \mathrm{E} X + \beta \, \mathrm{E} Y \leq \mathrm{E} Z. \tag{4.5.16}$$

□

4.5.6. Proposition. $\mathrm{L}_1(\Omega)$ *is a vector space.*
Proof It is an exercise to check this. □

4.5.7. Theorem (Beppo Levi's monotone convergence). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative random variables on $(\Omega,\mathscr{F},\mathcal{P})$ such that $X_n \leq X_{n+1}$, for all $n \in \mathbb{N}$, then*

$$\lim_{n\to\infty} X_n = \sup_{n\in\mathbb{N}} X_n =: X \tag{4.5.17}$$

is a random variable and

$$\mathrm{E}\,X = \lim_{n\to\infty} \mathrm{E}\,X_n = \sup_{n\in\mathbb{N}} \mathrm{E}\,X_n. \tag{4.5.18}$$

Proof Omitted. □

4.5.8. Theorem (Lebesgue's dominated convergence). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables such that*

$$\lim_{n\to\infty} X_n \text{ exists } =: X \tag{4.5.19}$$

and

$$X_n \leq Y \quad \forall\, n \in \mathbb{N}, \tag{4.5.20}$$

for a given $Y \in \mathscr{L}^1(\mathcal{P})$, then $X \in \mathscr{L}^1(\mathcal{P})$ and

$$\mathrm{E}\,X = \lim_{n\to\infty} \mathrm{E}\,X_n. \tag{4.5.21}$$

Proof Omitted. □

4.5.9. Corollary.

$$\|X\|_{\mathrm{L}_1(\Omega)} := \mathrm{E}\,|X|, \tag{4.5.22}$$

then $(\mathrm{L}_1(\Omega), \|\cdot\|_{\mathrm{L}_1(\Omega)})$ is a complete normed vector space (also known as Banach space).

4.6. Properties of expectation and variance

Recall the definition of expectation of a random variable X is

$$\mathrm{E}[X] := \int_\Omega X(w)\,\mathrm{d}\mathcal{P}(x). \tag{4.6.1}$$

Based on this definition, basic properties of expectation—as well as variance, which is nothing but the expectation of the random variable $(X - \mathrm{E}\,X)^2$—follow from those of integrals with respect to the measure $\mathcal{P}$.

4.6.1. Theorem (algebra of expectation and variance). *Let $(X_1,\dots,X_n)$ be a random vector and $(c_1,\dots,c_n) \in \mathbb{R}^n$, then*

(i) the expectation and the linear combination commute

$$\mathrm{E}\left[\sum_{i=1}^n c_i X_i\right] = \sum_{i=1}^n c_i\,\mathrm{E}[X_i]; \tag{4.6.2}$$

(ii) the variance is quadratic

$$\mathrm{var}\left[\sum_{i=1}^n c_i X_i\right] = \sum_{i=1}^n c_i^2\,\mathrm{var}[X_i] + 2 \sum_{1\leq i<j\leq n} c_i c_j \,\mathrm{cov}\big[X_i, X_j\big]; \tag{4.6.3}$$

(iii) and, if the random vector $(X_i)_{i\in\{i\ldots1\}n}$ *is independent then*

$$\operatorname{var}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2 \operatorname{var}[X_i]. \tag{4.6.4}$$

Proof The proof is identical to that of Theorem 3.5.7. □

4.6.2. Proposition (Cauchy–Bunyakovskii–Schwarz inequality). *Let* X, Y *be two random variables, then*

$$\mathrm{E}[XY]^2 \le \mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right]. \tag{4.6.5}$$

4.6.3. Exercise. *Prove the Cauchy–Bunyakovskii–Schwarz inequality, i.e., that*

$$\mathrm{E}[XY]^2 \le \mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right]$$

by considering the expression $\mathrm{E}\left[(X+\theta Y)^2\right]$. *(Hint: the mean of this square quantity can never be negative. So as a quadratic in* θ*, it cannot have two distinct real roots.)*

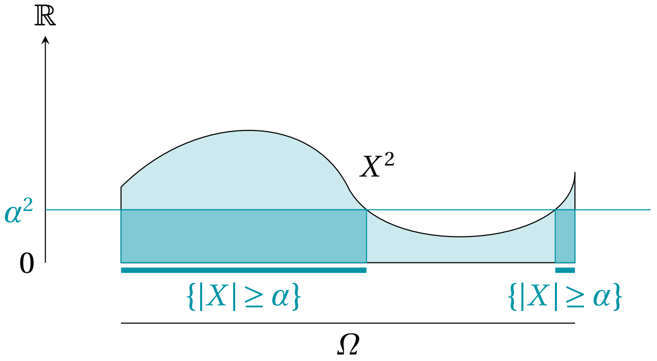
4.6.4. Theorem (Chebyshev's inequality). *Let* X *be a random variable. If* $\alpha \ge 0$ *then*

$$\alpha^2 \mathcal{P}[\alpha \le |X|] \le \mathrm{E}\left[X^2\right]. \tag{4.6.6}$$

Proof If $\mathrm{E}\left[X^2\right] = \infty$ the inequality is trivially satisfied. Supposing $\mathrm{E}\left[X^2\right]$ and recalling the definition of expectation and some elementary properties of the integral we see that

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \int_{\Omega} X(\omega)^2 \,\mathrm{d}\mathcal{P}(\omega) \\
&= \int_{|X|>\alpha} X^2 \,\mathrm{d}\mathcal{P} + \int_{|X|\le\alpha} X^2 \,\mathrm{d}\mathcal{P} \\
&\ge \alpha^2 \int_{|X|>\alpha} \mathrm{d}\mathcal{P} + 0 \\
&= \alpha^2 \mathcal{P}[|X| > \alpha].
\end{aligned} \tag{4.6.7}$$

This proof has the following a nice visual memo, whereby the dark area ■ = $\alpha^2\mathcal{P}[|X|>\alpha]$, is clearly bounded from above by the light area □ = $\mathrm{E}\left[X^2\right]$



□

4.6.5. Corollary (Chebyshev's inequality). *Let* X *be a random variable with finite mean,* $\alpha > 0$*, then*

$$\mathcal{P}[|X - \mathrm{E}[X]| > \alpha] < \frac{\operatorname{var}[X]}{\alpha^2}. \tag{4.6.8}$$

4.6.6. Interpretation of Chebyshev's inequatlity. In spite of its simplicity, Chebyshev's inequality (4.6.8) is surprisingly far reaching. It can be interpreted as a relationship between the probability of the random variable X *deviating* from its average $\mathrm{E}[X]$ and the variance of said variable. Indeed, choosing $k > 0$, denoting $\sigma :=$ $\mathrm{var}[X]^{1/2}$ and taking $\alpha := k\sigma$ in inequality (4.6.8) we obtain that

$$\mathcal{P}\left[\left|X-\mu\right| > k\sigma\right] < \frac{1}{k^2}. \tag{4.6.9}$$

So, for example, if[2] $k = 6$ then

$$\mathcal{P}\left[\left|X-\mu\right| > 6\sigma\right] < \frac{1}{36} \approx 2.78\%. \tag{4.6.10}$$

This observed relationship between the variance and the deviations of a random variable around its average can be extended into a theory known as the *Law of Large Numbers* (see §7.1). The square root of the variance, traditionally denoted by sigma, is known as the *standard deviation.* What is striking about Chebyshev type inequalities is that they work for *any random variable X*.

4.7. General integration

Not all measures are probability measures. We look here to what happens if we want to integrate with respect to a measure that is not necessarily finite.

4.7.1. Positive measures. A real-valued set function $\mathcal{M}$ with domain a sigma-algebra $\mathscr{C}$ over a set A, satisfies

(i) Countable additivity: *if* $(A_n)_{n\in\mathbb{N}}$ *is a pairwise disjoint sequence in* $\mathscr{B}^d$ *then* $|[|\,d]\bigcup_{n\in\mathbb{N}} A_n = \sum_{n\in\mathbb{N}} |[|\,d]A_n$

(ii) Monotonicity: *for each* $A, B \in \mathscr{B}^d$ $A \subseteq B \Rightarrow |[|\,d]A \leq |[|\,d]B$

(iii) Positivity: $|[|\,d]\varnothing = 0$

but in general is might not be that $\mathcal{M}(A) = 1$. If $\mathcal{M}(A) \in \mathbb{R}$, then it can be "turned" into a probabilty measure $\mathcal{N}(C) := \mathcal{M}(C)/\mathcal{M}(A)$. But if $\mathcal{M}(C) \notin \mathbb{R}$, i.e., $\mathcal{M}(C) = \infty$, for some $C \in \mathscr{C}$, including the case $C = A$. Hence, strictly speaking a (positive) measure is not real-valued, but $[0,\infty]$-valued, allowing thus for sets of infinite measure. One such, very important, measure is the Lebesgue measure which discuss below in §4.7.4.

4.7.2. Integration with respect to a positive measure. The expectation theory developed in 4.5 can be adapted (almost) verbatim to cover integration with respect to any positive measure $\mathcal{M}$. This is a sketch:

1. Introduce the concept of measureable function $f : A \to \mathbb{R}$, i.e., one for which

$$f^{-1}(()B) \in \mathscr{C} \quad \forall\, B \in \mathscr{B} \tag{4.7.1}$$

where $\mathscr{B}$ is the Borel sigma-algebra in $\mathbb{R}$.

[2] In industrial quality control and business management strategies, the "six sigma" paradigm is used as the tolerance for a "perfect" product.

2. Simple functions are the measure-theoretic equivalent of finite discrete random variables (also known as simple random variables) defined in 3.1.3. If f is simple, say $f = \sum_{i=1}^{k} f_i \mathbb{1}_{C_i}$, for $\boldsymbol{f} = (f_1, \dots, f_k) \in \mathbb{R}^k$, $f_i < f_j$ for $i < j$, and nonoverlapping C_i's *with finite measure,* i.e.,

$$\mathcal{M}\{x \in A\colon f(x) \neq 0\} < \infty, \tag{4.7.2}$$

then we define

$$\int_A f(x)\mathcal{M}(\mathrm{d}x) := \int_A f(x)\mathrm{d}\mathcal{M}(x) := \int f\,\mathrm{d}\mathcal{M} := \sum_{i=1}^{k} f_i \mathcal{M}(C_i) = \sum_{i=1}^{k} f_i \mathcal{M}\big(f^{-1}(\{f_i\})\big). \tag{4.7.3}$$

Note that the notations $\mathrm{d}\mathcal{M}(x)$ and $\mathcal{M}(\mathrm{d}x)$ are interchangeable. Indeed, these are just symbols as d has no functional meaning.[3] The integral of the simple function f is a real number because of assumption (4.7.2). It is important at this point to check that the integral is monotone and linear.

3. Define the integral of a nonnegative, not necessarily simple, measurable function f as

$$\begin{aligned}\int_A f(x)\mathrm{d}\mathcal{M}(x) &:= \int_A f(x)\mathrm{d}\mathcal{M}(x) := \int f\,\mathrm{d}\mathcal{M} \\ &:= \sup\left\{\int g\,\mathrm{d}\mathcal{M}\colon g \text{ simple function and } g \le f\right\}.\end{aligned} \tag{4.7.4}$$

It is then important to show that this definition coincides with the earlier one if f happens to be simple. It is also useful to show that the sup can be obtained as one of an approximating sequence and to allow for nonintegrable nonnegative measurable functions, i.e., for which $\int f\,\mathrm{d}\mathcal{M} = \infty$.

4. Define the integral of a general function f as

$$\int f\,\mathrm{d}\mathcal{M} = \int [f]_+\,\mathrm{d}\mathcal{M} - \int [f]_-\,\mathrm{d}\mathcal{M}, \tag{4.7.5}$$

in case one of the right-hand side integrals is not infinite. If both are infinite, then $\int f\,\mathrm{d}\mathcal{M}$ is undefined. If one of the integrals in the right-hand side of (4.7.5) is infinite, then f is still considered to be nonintegrable, but it is handy to define

$$\int f\,\mathrm{d}\mathcal{M} = \pm\infty \tag{4.7.6}$$

with the appropriate choice of signs.

5. Derive all the properties of the integral, such as montonocity and linearity. Prove Fatou's Lemma, Monotone Convergence and Lebesgue's Dominated Convergence theorems.

[3] "Thinking" that d is some kind of differential is helpful, but it is also dangerous. This analogy has to be handled with care.

4.7.3. Remark (Is there a more natural way?) The approach is the most widely used in textbooks, but it is not the only possible one and by far not the most natural one. An alternative way to build integration theory, is discussed by Lieb and Loss, 2001, Ch. 1: their eclectic approach is much more related to probability theory and Riemann integration simultaneously. In a nutshell their definition of expectation of a nonnegative random variable X is

$$\mathrm{E}X := \int_0^\infty (1-\mathrm{cdf}_X(x))\,\mathrm{d}\,x \tag{4.7.7}$$

where the right-hand side is the (improper) Riemann integral, which is shown to exist for any random variable X. The disadvantage of this definition is that it is harder to prove "easy" properties such as linearity, but the advantage is that "hard" results like the Monotone Convergence Theorem or Fatou's lemma are easier to prove with somewhat more intuitive arguments. Lieb and Loss, 2001 is also an extremely well-written book which is a recommended reading at any time.

4.7.4. Lebesgue's measure. A particularly important case is Lebesgue integration, which is basically integration with respect to the "d-volume" ("area", for $d=2$, "length" for $d=1$ and "physical volume" for $d=3$).

Despite Lebesgue's measure l^d not being a probability measure (the total sum is infinite, not 1) all the above theory applies to it. Denote by $\mathscr{B}^d$ the Borel sigma-algebra on $\mathbb{R}^d$.[4] A Lebesgue measure $\mathrm{l}^d : \mathscr{B}^d \to \mathbb{R}$ is a positive measure which, in addition to (i)–(iii) in, satisfies:

(iv) Volumicity: $\mathrm{l}^d(\boldsymbol{a},\boldsymbol{b}) = \prod_{i=1}^d (b_i - a_i) \mathit{for\,all}\, \boldsymbol{a} < \boldsymbol{b}\ \mathit{in}\, \mathbb{R}^d$.

It can be shown that there is only one such measure, thereby justifying the article in "*the* Lebesgue measure". It can also be shown that the Borel sigma-algebra can be extended as to include all subsets of all sets $N \in \mathscr{B}^d$ such that $|[|d]N = 0$. The resulting sigma-algebra is strictly bigger than the Borel sigma-algebra and is called the *complete Lebesgue-Borel sigma-algebra.* The reason for doing this is that most results become much easier to formulate and some results like Fubini and Tonelli's theorems are actually easier to prove.

If d is clear from the context we just use l, instead of l^d and when the integration variable is explicitly written we remove the l altogether, i.e., we write

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \text{ instead of } \int f(\boldsymbol{x})\,\mathrm{d}|[|d]\boldsymbol{x},\text{ or } \int f\,\mathrm{d}\mathrm{l}^d,\text{ or } \int_{\mathbb{R}^d} f(\boldsymbol{x})\,\mathrm{d}|\boldsymbol{x}|; \tag{4.7.8}$$

and when the integration variable is omittable we write

$$\int_{\mathbb{R}^d} f \text{ or } \int f \text{ or } \int_{\mathbb{R}^d} f\,\mathrm{d}\mathrm{l}^d \tag{4.7.9}$$

[4] Notice that our notation for the Lebesgue measure is an upright, lower case letter "l", as in Lebesgue. It is not the number 1. Other authors use $\mathscr{L}$, λ, $\mathscr{L}$, L or ℓ. We prefer to use l, as this is easily confused with 1. And this confusion is useful. Because that's what Lebesgue measure is afterall: in some sense, it is the unit of measures on $\mathbb{R}^d$. Also thinking of l as 1 is consistent with the notation $\mathrm{d}|x| = \mathrm{d}\,x$.

instead of the more customary (but inconsistent)

$$\int f \, \mathrm{d}\boldsymbol{x}. \tag{4.7.10}$$

4.8. Product measures

4.8.1. Basics. Let $(A,\mathscr{C},\mathcal{M})$ and $(B,\mathscr{D},\mathcal{N})$ be two measure spaces. It is possible to define a *product measure* on $A \times B$ as follows, consider the sets of the type $C \times D$ with $C \in \mathscr{C}$ and $D \in \mathscr{D}$ and define

$$\mathcal{L}(C \times D) := \mathcal{M}(C)\mathcal{N}(D). \tag{4.8.1}$$

It can be shown that there such a measure exists, it is unique and is denoted by $\mathcal{M}\times\mathcal{N}$, or (less common) $\mathcal{M}\otimes\mathcal{N}$.

4.8.2. Theorem (Fubini's double integral). *Let $(A,\mathscr{C},\mathcal{M})$ and $(B,\mathscr{D},\mathcal{N})$ be two (sigma-finite) measure spaces, and let $\mathcal{L} = \mathcal{M}\times\mathcal{N}$. Suppose $f \in \mathrm{L}_1(\mathcal{L})$, then the functions*

$$\begin{array}{rccl} g: & A & \rightarrow & [-\infty,\infty] \\ & x & \mapsto & g(x) := \int_B f(x,y)\,\mathrm{d}\mathcal{N}(y) \\ h: & B & \rightarrow & [-\infty,\infty] \\ & y & \mapsto & h(y) := \int_A f(x,y)\,\mathrm{d}\mathcal{M}(y) \end{array} \tag{4.8.2}$$

are in $\mathrm{L}_1(\mathcal{M})$ and $\mathrm{L}_1(\mathcal{N})$ respectively, and

$$\int_{A\times B} f \,\mathrm{d}\mathcal{L} = \int_A g(x)\,\mathrm{d}\mathcal{M}(x) = \int_B h(y)\,\mathrm{d}\mathcal{N}(y) \tag{4.8.3}$$

which is also written (in symbol-economy mode) as

$$\int_{A\times B} f \,\mathrm{d}[\mathcal{M}\times\mathcal{N}] = \int_A\int_B f(x,y)\,\mathrm{d}\mathcal{N}(y)\,\mathrm{d}\mathcal{M}(x) = \int_B\int_A f(x,y)\,\mathrm{d}\mathcal{M}(x)\,\mathrm{d}\mathcal{N}(y). \tag{4.8.4}$$

4.8.3. Theorem (Tonelli's iterated integral). *Let $(A,\mathscr{C},\mathcal{M})$ and $(B,\mathscr{D},\mathcal{N})$ be two (sigma-finite) complete measure spaces, and let $\mathcal{L} = \mathcal{M}\times\mathcal{N}$. Let $f : A \times B \rightarrow \mathbb{R}$ be $\mathcal{L}$-measurable, such that*

$$\begin{array}{rccl} g: & A & \rightarrow & [-\infty,\infty] \\ & x & \mapsto & \int_B f(x,y)\,\mathrm{d}\mathcal{N}(y) \end{array} \tag{4.8.5}$$

satisfies $g \in \mathrm{L}_1(\mathcal{M})$ then $f \in \mathrm{L}_1(\mathcal{L})$ and (4.8.3) and (4.8.4) hold true.

4.8.4. Corollary (multidimensional Fubini–Tonelli–Lebesgue). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with $f \in \mathrm{L}_1(\mathbb{R}^d)$ and for $i = 1,\dots,d$ define the function*

$$\begin{array}{rccl} \check{f}_i: & \mathbb{R}^{d-1} & \rightarrow & \mathbb{R} \\ & \boldsymbol{x}'_i & \mapsto & \check{f}_i(\boldsymbol{x}'_i) := \int_{\mathbb{R}} f(\tilde{\boldsymbol{x}}_i(x))\,\mathrm{d}\,x \end{array} \tag{4.8.6}$$

where $\boldsymbol{x}_i'$ is the vector $\boldsymbol{x}$ without entry x_i and $\tilde{\boldsymbol{x}}_i(x)$ is the vector $\boldsymbol{x}$ with the i-th coordinate x_i replaced by x, as defined in 4.3.3. Then, for each fixed $i = 1,\dots,d$, the function $\check{f}_i$ is integrable and

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^{d-1}} \left(\check{f}_i(\boldsymbol{x}_i')\right)\mathrm{d}\boldsymbol{x}_i'. \tag{4.8.7}$$

4.8.5. Corollary (multiple integrals). *A Lebesgue-measurable function $f : \mathbb{R}^d \to \mathbb{R}$ is in $\mathrm{L}_1(\mathbb{R}^d)$ if and only if the following iterated integral is finite:*

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(\boldsymbol{x})\,\mathrm{d}x_1 \cdots \mathrm{d}x_d \text{ and equals } \int_{\mathbb{R}^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{4.8.8}$$

4.8.6. Corollary (the generalised Cavalieri principle). *Let $(A,\mathscr{C},\mathcal{M})$ and $(B,\mathscr{D},\mathcal{N})$ be two (sigma-finite) measure spaces and let $E \subset A \times B$ that is $\mathcal{M}\times\mathcal{N}$-measurable. Define the following x-section, for each $x \in A$,*

$$E_x := \left\{y \in B : (x,y) \in E\right\} \tag{4.8.9}$$

and y-section, for each $y \in B$,

$$E_y := \left\{x \in A : (x,y) \in E\right\} \tag{4.8.10}$$

and

$$s(x) := \mathcal{N}(E_x) \text{ and } r(y) := \mathcal{M}\left(E_y\right). \tag{4.8.11}$$

Then s and r are, respectively, $\mathcal{M}$- and $\mathcal{N}$- measurable functions, and

$$\mathcal{M}\times\mathcal{N}(E) = \int_A s(x)\,\mathrm{d}\mathcal{M}(x) = \int_B r(y)\,\mathrm{d}\mathcal{N}(y). \tag{4.8.12}$$

Proof Apply Fubini's Theorem to the function $(x,y) \mapsto \mathbb{1}_E(x,y)$. □

4.8.7. Remark. The classical Cavalieri Principle states that if two plane figures, E and F, within two parallel lines ℓ_1 and ℓ_2, have the property that each line ℓ parallel to ℓ_1 (and ℓ_2) meets both E and F with sets of equal lengths then E and F must have equal areas. The same is true if E and F are solid figures in space and all parallel lines are replaced by parallel planes.

4.8.8. Theorem (Riemann meets Lebesgue). *Let $(A,\mathscr{C},\mathcal{M})$ be a (sigma-finite) measurable space and $f : A \to \mathbb{R}$ be an $\mathcal{M}$-integrable nonnegative function. Define*

$$R_y := \left\{(x,y) \in A \times \mathbb{R} : f(x) \le y\right\} \text{ and } r(y) := \mathcal{M}\left(R_y\right). \tag{4.8.13}$$

Then the set $E \in A \times \mathbb{R}$ "enclosed" by the graph of f and the "horizontal set" $A \times \{0\}$, defined rigorously as

$$E := \bigcup_{y \in \mathbb{R}} R_y, \tag{4.8.14}$$

is $\mathcal{M}\times \mathrm{l}^1$-measurable, the function r is Riemann (and hence Lebsgue) integrable and

$$\mathcal{M}\times \|[|1]E = \int f(x)\,\mathrm{d}\mathcal{M}(x) = \int r(y)\,\mathrm{d}y, \tag{4.8.15}$$

where the last integral can be approximated by Riemann's sums.

Proof This is a corollary of the previous results. □

4.8.9. Remark. Theorem 4.8.8 has many interpretations. The geometric interpretation is the classical view that the integral of a nonnegative real-variate and real-valued function equals the area enclosed by the function's graph and the horizontal axis.

4.9. Probability measures on $\mathbb{R}$

4.9.1. Definition of induced probability measure, distribution measure. Let $X \in \mathrm{RV}(\mathcal{P})$, its *induced probability measure* or *distribution measure* is the function $\mathcal{P}_X : \mathscr{B}(\mathbb{R}) \to \mathbb{R}$ such that

$$\mathcal{P}_X(A) := \mathcal{P}[X \in A] \text{ for } A \in \mathscr{B}(\mathbb{R}). \tag{4.9.1}$$

4.9.2. Exercise. *Let $X \in \mathrm{RV}(\mathcal{P})$, prove that the real-valued set-function $\mathcal{P}_X(A)$ defined by (4.9.1) is effectively a probability measure in the sense of Definition 1.1.2.*

4.9.3. Probability distribution functions. A *probability distribution function* is a function $F : \mathbb{R} \to$ that:

(1) F is monotone increasing (losely),
(2) $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$,
(3) F is cadlag.

From theorem 4.2.2 we already know that a random variable X defined a probability distribution function as cdf_X. Conversely, given a probability distribution function F, it is possible to construct a probability space $\mathcal{P}$ and a random variable X on it such that its distribution is F. We shall see this further, before we point out that distribution functions and probability measures on $\mathbb{R}$ are basically the same thing.

4.9.4. Distribution functions and measures. Probabiliy distribution functions and cummulative distribution functions are one To see this first note that any probability distribution function F spawns a probability measure $\mathcal{Q}$ on $\mathbb{R}$ by posing

$$\mathcal{Q}(a,b] := F(b) - F(a), \mathcal{Q}(a,b) := \lim_{b-} F - F(a), \text{etc.} \tag{4.9.2}$$

and then extending this via countable additivity to all Borel subsets of $\mathbb{R}$. It is a useful exercise to convince yourself that this is indeed a probability measure on $\mathscr{B}(\mathbb{R})$.[∗] [∗]: Check!
Conversely given a probability measure, say $\mathcal{Q}$ on $\mathscr{B}(\mathbb{R})$, or a superalgebra thereof, e.g., the Lebesgue-measurable sets, then defininig

$$F(x) := \mathcal{Q}(-\infty, x], \tag{4.9.3}$$

we obtain a probability distribution function.[∗] [∗]: Check!
Note that this correspondence between distribution functions and probability measures on $\mathbb{R}$ is a one-to-one correspondence.

4.9.5. Given $X \in \mathrm{RV}(\mathcal{P})$, the set-function $\mathcal{P}_X(A)$ is in fact a probability measure on $\mathscr{B}(\mathbb{R})$.
Suppose now we are given a probability measure $\mathcal{Q}$ on $\mathscr{B}(\mathbb{R})$, it is possible to construct a probability space $(\Omega, \mathscr{F}, \mathcal{P})$ and a random variable X such that $\mathcal{P}_X = \mathcal{Q}$. Indeed, consider $\Omega := \mathbb{R}$, $\mathscr{F} := \mathscr{B}(\mathbb{R})$ and let $\mathcal{P} := \mathcal{Q}$, then taking $X := \mathrm{id}_{\mathbb{R}}$, i.e., $X(\omega) := \omega$ for all $\omega \in \Omega$ we obtain what we need.[∗]
Note that here uniqueness is not guaranteed, i.e., there can be two different random variables, $X \neq Y$, that yield $\mathcal{P}_X(A) = \mathcal{P}_Y(A)$ for all $A \in \mathscr{B}(\mathbb{R})$.
Finally note that this solves the original problem of given a distribution function F to find $\mathcal{P}$ and $X \in \mathrm{RV}(\mathcal{P})$ such that $F = \mathrm{cdf}_X$.

[∗]: Check!

Exercises and problems on general random variables and expectation

Exercise 4.1. Use indicator variables to solve the following problems.
(i) There are 32 teams in a knockout competition; before the (random) draw is made, you GUESS what the 16 matches will be. Show that the mean number of correct guesses is 16/31.
(ii) n married couples sit randomly in the $2n$ places round a circular table. Assuming $n \geq 2$, find the mean and variance of the number of couples who sit next to each other.

Exercise 4.2. Prove the Cauchy–Bunyakovskii–Schwarz inequality, i.e., that

$$\mathrm{E}[XY]^2 \leq \mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right]$$

by considering the expression $\mathrm{E}\left[(X + \theta Y)^2\right]$. (Hint: the mean of this square quantity can never be negative. So as a quadratic in θ, it cannot have two distinct real roots.)

Exercise 4.3. Prove the following generalisation of the Chebyshev inequality, which goes also by the same name in the literature: *for any $p \geq 1$, a random variable X and $\alpha > 0$ then*

$$\mathcal{P}[\alpha \leq |X|] \leq \frac{\mathrm{E}[|X|^p]}{\alpha^p}. \tag{P4.3.1}$$

Deduce that

$$\liminf_{p \to \infty} \|X\|_{\mathrm{L}_p(\Omega)} \geq \|X\|_{\mathrm{L}_\infty(\Omega)}, \tag{P4.3.2}$$

where, for all $p \in [0, \infty)$,

$$\|X\|_{\mathrm{L}_p(\Omega)} := \mathrm{E}[X^p]^{1/p} \text{ and } \|X\|_{\mathrm{L}_\infty(\Omega)} := \min\left\{\beta > 0 : \mathcal{P}\left[\beta \leq X\right] = 0\right\}. \tag{P4.3.3}$$

Prove also that, for all $p \in [0, \infty)$,

$$\limsup_{p \to \infty} \|X\|_{\mathrm{L}_p(\Omega)} \leq \|X\|_{\mathrm{L}_\infty(\Omega)}. \tag{P4.3.4}$$

Conclude that

$$\lim_{p \to \infty} \|X\|_{\mathrm{L}_p(\Omega)} \text{ exists and equals } \|X\|_{\mathrm{L}_\infty(\Omega)}. \tag{P4.3.5}$$

Problem 4.4. (a) Show that

$$\mathrm{E}X = \sum_{n=0}^{\infty} \mathcal{P}[X \geq n] \tag{P4.4.1}$$

when X takes nonnegative integer values only.

(b) The usual analogy between the discrete and continuous cases suggests the formula

$$\mathrm{E} X = \int_0^\infty \mathcal{P}(X > x)\,\mathrm{d}x = \int_0^\infty (1 - \mathrm{cdf}_X(x))\,\mathrm{d}x, \tag{P4.4.2}$$

for any random variable X. Prove this when X is continuously distributed and nonnegative.

(c) Using Fubini's Theorem (namely, the Riemann meets Lebesgue result) and the geometric fact that the integral of a function (and hence the expectation of a random variable) is the area between the graph and the horizontal set (or "axis"), prove that (P4.4.2) holds in general. Let your proof *generalise but not rely not on* (b).

(d) Generalise the result as to include any random variable X with finite expectation:

$$\mathrm{E}[X] = \int_0^\infty \mathcal{P}[x < X]\,\mathrm{d}x - \int_{-\infty}^0 \mathcal{P}[X < x]\,\mathrm{d}x = \int_0^\infty 1 - \mathrm{cdf}_X(x) - \mathrm{cdf}_X(-x)\,\mathrm{d}x. \tag{P4.4.3}$$

The integrand in (P4.4.3) us called a *hill-shaped rearrangement* of X.

(e) Denoting by $\mathcal{P}_X$ the probability distribution of X and using relation (P4.4.3), or the ideas that led you to it, show that

$$\mathrm{E}[X] = \int_{\mathbb{R}} x\,\mathrm{d}\mathcal{P}_X(x). \tag{P4.4.4}$$

This is known as the (basic) *expectation rule*.

Part (b) requires (in part) the use of densities, that are introduced in Chapter 5.
Parts (c), (d), and (e) do not rely on (b).

CHAPTER 5

Continuous distributions and random variables

An important subclass of random variables is that of absolutely continuously distributed random variables (also known as "continuous" random variables). For example, Gaussian (also known as normal) random variables play a central role in probabilities and statistics.

5.1. Density

5.1.1. Definition of density of a random variable. Let X be a random variable on a probability space $(\Omega, \mathscr{F}, \mathcal{P})$, and let F denote its distribution function, if there is a function f such that

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}\,u, \tag{5.1.1}$$

then X is said to be a *continuously distributed random variable*[1], with *density* f.
In general we denote the density function, if one exists, by pdf_X or f_X, or just f (when f or f_X do not clash with other notation).

5.1.2. Remark (densities and derivatives). Let X be a continuously distributed random variable, with distribution F and density f. The integral CDF–PDF relationship (5.1.1) can be written in a differentiable form as follows

$$\frac{\mathrm{d}}{\mathrm{d}\,x} F(x) = f(x), \tag{5.1.2}$$

whenever this notation makes sense (i.e., when F is differentiable).
In fact, F is *absolutely continuous*[2] and this relationship is valid for *almost all* points $x \in \mathbb{R}$, even if F is not differentiable everywhere. This means that whenever differentiation of F is possible relation (5.1.2) is valid, but one (especially newbies) should pay special attention to "sharp" points where the classical derivative is not defined. In most practical situation, "almost all" translates to "all but a finite number of points" (although things can be more complicated than that).

5.1.3. Iverson's brackets. In dealing with densities (and in many other places) it is often handy to use Iverson's brackets notation (defined in §5.1.3) which is particularly handy in manipulating sums and integrals. For example, suppose a function f is defined on $\mathbb{R}$ but takes non-zero values only in an interval, say (a, b), for some

[1] Some authors use the word "continuous" instead of "continuously distributed", but that is confusing because continuous means something different in analysis and, if the probability space is also a topological (or metric) space, some random variables on it end up being continuous and not continuous at the same time!

[2] If you don't know what an absolutely continuous function is you may safely ignore this comment.

$a, b \in \mathbb{R}$ and that it coincides with the polynomial $x \mapsto 3x^2$ only therein, then Iverson's bracket provides a quick way of writing this as

$$f(x) := [\![a < x < b]\!] 3x^2, \text{ for } x \in \mathbb{R}. \tag{5.1.3}$$

This avoids ambiguous definitions such as

$$f(x) := 3x^2, \text{ for } a < x < b, \tag{5.1.4}$$

(where it is not clear what goes on for $x \leq a$ or $x \geq b$) or bulky ones such as

$$f(x) := \begin{cases} 3x^2, & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1.5}$$

Two useful properties of the Iverson bracket are

$$[\![P \text{ and } Q]\!] = [\![P]\!][\![Q]\!] \text{ and } [\![P \text{ or } Q]\!] = [\![P]\!] + [\![Q]\!] - [\![P]\!][\![Q]\!]. \tag{5.1.6}$$

As a simple example, to integrate the function f defined in (5.1.3) over $\mathbb{R}$ we write

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} [\![a < x < b]\!] 3x^2\,\mathrm{d}x = \int_a^b 3x^2\,\mathrm{d}x = \left[x^3\right]_{x=a}^{x=b} = b^3 - a^3. \tag{5.1.7}$$

Suppose now we are required to integrate this function on $\left(-\infty, y\right]$ for a fixed $y \in \mathbb{R}$ we have

$$\begin{aligned} \int_{-\infty}^{y} f(x)\,\mathrm{d}x &= \int_{-\infty}^{y} [\![a < x < b]\!] 3x^2\,\mathrm{d}x \\ = \left[\!\!\left[a < y < b\right]\!\!\right] \int_a^{\min\{y,b\}} 3x^2\,\mathrm{d}x &= \begin{cases} 0 & \text{for } y \leq a, \\ y^3 - a^3 & \text{for } a < y < b, \\ b^3 - a^3 & \text{for } y \geq b. \end{cases} \end{aligned} \tag{5.1.8}$$

A concept related to the Iverson bracket is the *indicator function*. Given two sets $S \subset R$ we define the indicator function of S in R as

$$\mathbb{1}_S(x) = [\![x \in S]\!] \text{ for } x \in R. \tag{5.1.9}$$

5.1.4. Example.

PROBLEM. *Suppose X has density $f(x) = 3x^2$ for $0 < x < 1$, i.e., $\mathrm{pdf}_X(x) = 3x^2\mathbb{1}_{[0,1]}$. Find the densities of $Y = X^2$ and $W = |X - 1/2|$.*

Solution. Two remarks before solving this: firstly note that, as for all density functions, $f(x)$ *is* defined for all $x \in \mathbb{R}$, including outside $(0,1)$, it is zero. Note that the value of $f(x)$ for *finitely many* x in $\mathbb{R}$ does not matter. In this particular situation, we may safely ignore the values of $f(0)$ and $f(1)$ and write

$$f(x) = 3x^2\mathbb{1}_{[0,1]}(x) = \begin{cases} 3x^2, & \text{for } x \in (0,1), \\ 0, & \text{otherwise.} \end{cases} \tag{5.1.10}$$

We work via the distribution functions, denote by G, g and H, h the distribution, density of Y and W, respectively. Note (by splitting the integral in 3 pieces) that

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u)\,\mathrm{d}\,u = \int_{-\infty}^{x} 3u^2 \mathbb{1}_{[0,1]}(u)\,\mathrm{d}\,u \\ &= [\![0 < x]\!] \int_{0}^{\min\{1,x\}} 3u^2\,\mathrm{d}\,u = \begin{cases} 0, & \text{for } x \le 0 \\ x^3, & \text{for } 0 < x < 1 \\ 1, & \text{for } x \ge 1 \end{cases} \end{aligned} \tag{5.1.11}$$

For $0 < y < 1$, we have

$$\begin{aligned} G(y) &= \mathcal{P}\big[Y \le y\big] = \mathcal{P}\big[X^2 \le y\big] \\ &= \mathcal{P}\big[|X| \le \sqrt{y}\big] \\ &= \mathcal{P}\big[-\sqrt{y} \le X \le \sqrt{y}\big] = F(\sqrt{y}) - F(-\sqrt{y}) \\ &= y^{3/2}. \end{aligned} \tag{5.1.12}$$

Differentiation for $y \in (0,1)$

$$g(y) = f(\sqrt{y})\frac{1}{2}y^{-1/2} = 3y\frac{1}{2}y^{-1/2} = \frac{3}{2}y^{1/2}. \tag{5.1.13}$$

If $y < 0$ or $y > 1$ then $g(y) = 0$. In summary we have

$$\mathrm{pdf}_{X^2}(x) = \frac{3}{2}x^{1/2}\mathbb{1}_{[0,1]}(x). \tag{5.1.14}$$

To calculate h, consider H first. We have, for each $w \in \mathbb{R}$,

$$\begin{aligned} H(w) &= \mathcal{P}[W \le w] = \mathcal{P}\left[\left|X - \frac{1}{2}\right| \le w\right] \\ &= \mathcal{P}\left[\frac{1}{2} - w \le X \le \frac{1}{2} + w\right] \\ &= \begin{cases} \mathcal{P}(\varnothing) = 0 & \text{if } w < 0 \\ F(1/2 + w) - F(1/2 - w) & \text{if } w > 0. \end{cases} \end{aligned} \tag{5.1.15}$$

Hence, for $w > 0$ we have

$$H(w) = \begin{cases} 1, & \text{for } w > 1/2 \\ (1/2 + w)^3 - (1/2 - w)^3, & \text{for } 0 < w < 1/2. \end{cases} \tag{5.1.16}$$

Thus for $0 < w < 1/2$ we have

$$H(w) = 2\left(\frac{3}{4}w + w^3\right) = 2w^3 + \frac{3}{2}w, \tag{5.1.17}$$

and thus, by differentiation, we obtain

$$h(w) = \begin{cases} 6w^2 + \frac{3}{2}, & \text{for } 0 < w < 1/2 \\ 0, & \text{otherwise,} \end{cases} \tag{5.1.18}$$

of, using indicator functions,

$$\mathrm{pdf}_{|X-1/2|}(x) = \mathbb{1}_{[0,1/2]}(x)\left(6x^2 + \frac{3}{2}\right). \tag{5.1.19}$$

5.1.5. Proposition (properties of the density function). *Suppose a random variable X is continuously distributed and has a probability density function f, then the following hold*

- *(i)* *f is non-negative, i.e., $f(x) \geq 0$ for all $x \in \mathbb{R}$;*
- *(ii)* *f vanishes at infinity, i.e., $\lim_{x\to\pm\infty} f(x) = 0$;*
- *(iii)* *f has a unitary integral, i.e., $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$.*

5.1.6. Proposition (expectation rule). *Given a continuously distributed random variable X, with $f = \mathrm{pdf}_X$, its expectation is given by*

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x. \tag{5.1.20}$$

Moreover, for any Borel-measurable function $h : \mathbb{R} \to \mathbb{R}$, the following expectation rule holds:

$$\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\,\mathrm{d}x \tag{5.1.21}$$

5.1.7. Corollary (variance and densities). *Suppose X is a continuously distributed random variable with density $\mathrm{pdf}_X =: f$, then*

$$\operatorname{var} X := \int x^2 f(x)\,\mathrm{d}x - \left(\int x f(x)\,\mathrm{d}x\right)^2. \tag{5.1.22}$$

5.1.8. Example.

PROBLEM. *Let X have density $\mathrm{pdf}_X(x) = \mathbb{1}_{[0,1]} 3x^2$.*

(a) *Calculate the expectations $\mathrm{E}\,X$, $\mathrm{E}\left[X^2\right]$ and $\mathrm{E}\,|X - 1/2|$.*

(b) *Calculate the variances $\operatorname{var} X$, $\operatorname{var} X^2$ and $\operatorname{var}|X - 1/2|$.*

Solution. By the expectation rule we have

$$\mathrm{E}\,X = \int_{-\infty}^{\infty} x\,\mathrm{pdf}_X(x)\,\mathrm{d}x = 3\int_0^1 x^3\,\mathrm{d}x = \frac{3}{4}. \tag{5.1.23}$$

Also by expectation rule we have

$$\mathrm{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2\,\mathrm{pdf}_X(x)\,\mathrm{d}x = 3\int_0^1 x^4\,\mathrm{d}x = \frac{3}{5}. \tag{5.1.24}$$

Using (5.1.14) from Example 5.1.4, we can alternatively compute

$$\mathrm{E}\left[X^2\right] = \int_{-\infty}^{\infty} x\,\mathrm{pdf}_{X^2}(x)\,\mathrm{d}x = \frac{3}{2}\int_0^1 x^{3/2}\,\mathrm{d}x = \frac{3}{2} \times \frac{2}{5} = \frac{3}{5}. \tag{5.1.25}$$

Likewise, we have

$$
\begin{aligned}
\mathrm{E}\left|X-\frac{1}{2}\right| &= \int_{-\infty}^{\infty}\left|x-\frac{1}{2}\right| 3x^2 \mathbb{1}_{[0,1]}(x)\,\mathrm{d}\,x = 3\int_0^1 \left|x-\frac{1}{2}\right| x^2\,\mathrm{d}\,x \\
&= 3\left(\int_0^{1/2}\left(\frac{1}{2}-x\right)x^2\,\mathrm{d}\,x + \int_{1/2}^1\left(x-\frac{1}{2}\right)x^2\,\mathrm{d}\,x\right) \\
&= 3\left(\frac{1}{2\times 3\times 8}-\frac{1}{4\times 16}+\frac{1}{4}\left(1-\frac{1}{16}\right)-\frac{1}{2\times 3}\left(1-\frac{1}{8}\right)\right) \\
&= 3\left(\frac{1}{3\times 8}-\frac{1}{2\times 16}+\frac{1}{4}-\frac{1}{2\times 3}\right) \\
&= 3\left(\frac{7}{32}-\frac{1}{8}\right)=\frac{9}{32}.
\end{aligned}
\tag{5.1.26}
$$

Also here, we can use (5.1.19) to check our calculation through the alternative route:

$$
\begin{aligned}
\mathrm{E}\left|X-\frac{1}{2}\right| &= \int_{-\infty}^{\infty}\mathbb{1}_{[0,1/2]}(x)\left(6x^2+\frac{3}{2}\right)x\,\mathrm{d}\,x = 6\int_0^{1/2} x^3 + \frac{3}{2}\int_0^{1/2} x\,\mathrm{d}\,x \\
&= \frac{6}{4\times 16}+\frac{3}{2\times 2\times 4}=\frac{3+2\times 3}{32}=\frac{9}{32}.
\end{aligned}
\tag{5.1.27}
$$

(a) Let us calculate the variances.

$$
\operatorname{var} X = \mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2=\frac{3}{5}-\frac{9}{16}=\frac{3}{80}. \tag{5.1.28}
$$

5.2. Density functions of probability distributions

We have seen that for a given random variable X, its probability density function $f := \mathrm{pdf}_X$, if it exists, is a non-negative function, $f(x)\geq 0$ for all $x\in\mathbb{R}$, which is integrable with sum 1, $\int_{\mathbb{R}} f = 1$. In fact, any function satisfying these properties can be seen as the density of a random variable X on an appropriately chosen probability space.

5.2.1. Definition of density function. A *density function* is a function $f:\mathbb{R}\to\mathbb{R}$ such that $f(x)\geq 0$, for any $x\in\mathbb{R}$ and $\int_{\mathbb{R}} f(x)\,\mathrm{d}\,x = 1$.

5.2.2. Proposition. *Given a density function f there exists a probability measure $(\mathbb{R},\mathscr{B},\mathcal{P})$ such that $f=\mathrm{pdf}_{\mathrm{id}}$ where $\mathscr{B}$ is the sigma-algebra of Borel sets in $\mathbb{R}$ and* id *is the identity function:*

$$
\mathrm{id}(x)=x,\ \textit{for}\ x\in\mathbb{R}. \tag{5.2.1}
$$

Proof For each set $A\in\mathscr{B}$ define

$$
\mathcal{P}(A) := \int_A f(x)\,\mathrm{d}\,x. \tag{5.2.2}
$$

Then we have that $\mathcal{P}(A)\geq 0$, $\mathcal{P}(\varnothing)=0$ and

$$
\mathcal{P}(\mathbb{R}) = \int_{\mathbb{R}} f(x)\,\mathrm{d}\,x = 1. \tag{5.2.3}
$$

Furthermore, by the monotonicity of the integral in $\mathbb{R}$ we have for a countable mutually disjoint family $\{A_m\}_{m\in m\in\mathscr{I}}$, that

$$\mathcal{P}\left(\bigcup_{m\in\mathscr{I}} A_m\right) = \int_{\bigcup_{m\in\mathscr{I}} A_m} f(x)\,\mathrm{d}x = \sum_{m\in\mathscr{I}} \int_{A_m} f(x)\,\mathrm{d}x = \sum_{m\in\mathscr{I}} \mathcal{P}(A_m). \tag{5.2.4}$$

It follows that $\mathcal{P}\colon \mathscr{B} \to \mathbb{R}$ is a probability measure.
To close the proof consider

$$\mathrm{cdf}_{\mathrm{id}}(x) = \mathcal{P}\{\xi \in \mathbb{R} : \mathrm{id}(\xi) \le x\} = \mathcal{P}\{\xi \in \mathbb{R} : \xi \le x\} = \int_{-\infty}^{x} f. \tag{5.2.5}$$

It follows, by differentiation, that $\mathrm{pdf}_{\mathrm{id}}$ exists and

$$\mathrm{pdf}_{\mathrm{id}}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\,\mathrm{cdf}_{\mathrm{id}}(x) = f(x) \text{ for } x \in \mathbb{R}. \tag{5.2.6}$$

□

5.2.3. Definition of distribution function. Given a density function f, we associate to it a function $F(x) = \int_{-\infty}^{x} f(\xi)\,\mathrm{d}\xi$ is called the *distribution function*. Note that if $\mathcal{P}$ is the probability measure induced by f on $\mathbb{R}$, then $F(x) = \mathcal{P}(-\infty, x]$.

5.3. Examples of continuous random variables

5.3.1. Definition of continuous uniform distributions. Let $a < b$ be two real numbers. A random variable X is said to be *continuous and uniformly distributed* over $[a, b]$ if and only if X has a density and

$$\mathrm{pdf}_X(x) = \begin{cases} 1/b - a & \text{for } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases} \tag{5.3.1}$$

We denote by $\mathrm{U}[a, b]$ the set of all random variables that are continuously uniformly distributed from a through b and when X satisfies (5.3.1) we say "$X \in \mathrm{U}[a, b]$" or "X is $\mathrm{U}[a, b]$".[3]

5.3.2. Exponential distributions. The *exponential distribution with average* $1/\lambda$ is the function $\lambda \exp(-\lambda x)$ for $x > 0$ and 0 for $x \le 0$. A random variable X is said to be exponentially distributed if and only if

$$\mathrm{pdf}_X(x) = \lambda \exp(-\lambda x) \mathbb{1}_{\mathbb{R}^+}(x). \tag{5.3.2}$$

We indicate the set of all random variables that have the exponential distribution with $\mathrm{Exp}(\lambda)$.
A random variable $X \in \mathrm{Exp}(\lambda)$ models the time to wait for a random event that arises at rate λ to happen. For example, the lifetime of a filament bulb or the interval between one call and the next at a call centre. Intuitively, the higher the rate the lower must be the expected time. Indeed, it can be shown that if $X \in \mathrm{Exp}(\lambda)$, then $\mathrm{E}X = 1/\lambda$, which justifies the "average $1/\lambda$ in the definition.

5.3.3. Exercise. *Let* $X \in \mathrm{Exp}(\lambda)$, *for some* $\lambda \in \mathbb{R}^+$. *Calculate* $\mathrm{E}X$ *and* $\mathrm{var}\,X$.

Solution.

[3] Many texts use the ~ sign, hence the common notation "$X \sim \mathrm{U}[a, b]$".

5.3.4. Gaussian (also known as normal) distibution. Given $\sigma \in \mathbb{R}$. The *Gaussian distribution* (also known as *normal distribution*) is the function

$$g(x) := \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{-x^2}{2\sigma^2} \tag{5.3.3}$$

5.4. Joint densities and independence

We extend now the concept of density to higher dimensions, by considering random vectors and their distributions as outlined in §4.3, which should be revised before delving in this one

5.4.1. Definition of joint density. The random vector $\boldsymbol{X}$ is said to be (absolutely) continuously distributed, with density $f : \mathbb{R}^n \to \mathbb{R}$, if we have

$$\operatorname{cdf}_{\boldsymbol{X}}(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(\xi_1, \dots, \xi_n) \,\mathrm{d}\xi_n \cdots \mathrm{d}\xi_1. \tag{5.4.1}$$

As for random variables, we denote such f with $\operatorname{pdf}_{\boldsymbol{X}}$.

5.4.2. Proposition (properties of a joint density). *Suppose a random vector $\boldsymbol{X}$ has density $f := \operatorname{pdf}_{\boldsymbol{X}}$ we have that*

(i) f is non-negative, i.e., $f(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$;
(ii) f vanishes at infinity, i.e., $\lim_{|\boldsymbol{x}| \to \pm\infty} f(\boldsymbol{x}) = 0$;
(iii) f has a unitary integral, i.e., $\int_{\mathbb{R}^n} f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = 1$;
(iv) for each set $A \subseteq \mathbb{R}^n$, with A Borel-measurable, we have

$$\mathcal{P}\{\boldsymbol{X} \in A\} = \int_A f(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{5.4.2}$$

5.4.3. Remark (differentiable joint distribution implies density). Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$ and let $F := \operatorname{pdf}_{\boldsymbol{X}}$. If the distribution $F : \mathbb{R}^d \to \mathbb{R}$ is simultaneously differentiable in all variables then it has a density $f : \mathbb{R}^d \to \mathbb{R}$, given by

$$f(\boldsymbol{x}) = \partial_{x_1} \cdots \partial_{x_d} F(\boldsymbol{x}), \tag{5.4.3}$$

where we use the short partial derivative notation

$$\partial_{x_i} g(\boldsymbol{x}) = \frac{\partial g(\boldsymbol{x})}{\partial x_i} := \lim_{\epsilon \to 0} \frac{g(\boldsymbol{x} + \epsilon \boldsymbol{e}_i) - g(\boldsymbol{x})}{\epsilon} \tag{5.4.4}$$

$$\text{where } \boldsymbol{e}_i := \left([\![i = j]\!] \right)_{j=1,\dots,d}. \tag{5.4.5}$$

5.4.4. Theorem (marginal integrals). *If $\boldsymbol{X} = (X_i)_{i\in\{i\ldots 1\}d}$ be a random vector which is continuously distributed then for each fixed $i = 1,\ldots,d$, the component X_i, for is a continuously distributed random variable and*

$$\operatorname{pdf}_{X_i}(x) = \int_{\mathbb{R}^{d-1}} \operatorname{pdf}_{\boldsymbol{X}}(\tilde{\boldsymbol{x}}_i(x))\,\mathrm{d}\boldsymbol{x}'_i. \tag{5.4.6}$$

The integral on the right-hand side of (5.4.6) is called a marginal integral.

Proof Fix $i = 1,\ldots,d$, and consider $f_i : \mathbb{R} \to \mathbb{R}$ to be the i-th marginal

$$f_i(\xi) := \int_{\mathbb{R}^{d-1}} f(\tilde{\boldsymbol{x}}_i(\xi))\,\mathrm{d}\boldsymbol{x}'_i. \tag{5.4.7}$$

By the Marginal Distribution Theorem, then Fubini's Theorem, it follows that for each $x \in \mathbb{R}$ we have

$$\operatorname{cdf}_{X_i}(x) = \lim_{\boldsymbol{x}'_i \to +\infty} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(\xi_1,\ldots,\xi_d)\,\mathrm{d}\xi_1 \cdots \mathrm{d}\xi_d = \int_{-\infty}^{x} f_i(\xi)\,\mathrm{d}\xi. \tag{5.4.8}$$

That is X_i is continuously distributed with $\operatorname{pdf}_{X_i} = f_i$, as claimed. □

5.4.5. Definition of independent (vector of) random variables. We say that (a vector of) random variables $\boldsymbol{X} = (X_i)_{i\in\{i\ldots 1\}d}$ (is) are *independent* if the cumulative distribution function of $\boldsymbol{X}$ can be factored as a *tensor product* of some univariate functions $F_i : \mathbb{R} \to \mathbb{R}$, i.e.,

$$\operatorname{cdf}_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{d} F_i(x_i), \tag{5.4.9}$$

for some functions F_i. By the Marginal Distribution Theorem 4.3.5, it follows that if $\boldsymbol{X}$ is independent, then the functions $F_i = \operatorname{cdf}_{X_i}$ appearing in (5.4.9), for all $i = 1,\ldots,d$.

5.4.6. Theorem (independent joint densities). *An independent random vector $\boldsymbol{X} = (X_1,\ldots,X_d)$ is continuously distributed if and only if X_i is continuously distributed for each $i = 1,\ldots,d$, and in this case the joint density is the tensor product of the single densities, i.e.,*

$$\operatorname{pdf}_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{d} \operatorname{pdf}_{X_i}(x_i) \quad \forall\, \boldsymbol{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d. \tag{5.4.10}$$

Proof Suppose $\boldsymbol{X}$ has a density, say $f := \operatorname{pdf}_{\boldsymbol{X}}$. Then, by independence, for each $\boldsymbol{x} = (x_1,\ldots,x_d) \in \mathbb{R}^d$ we have

$$\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(\xi_1,\ldots,\xi_d)\,\mathrm{d}\xi_1 \cdots \mathrm{d}\xi_d = \operatorname{cdf}_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{d} \operatorname{cdf}_{X_i}(\boldsymbol{x}). \tag{5.4.11}$$

equivalent to its joint density factorising as the product of the individual densities of the components, i.e.,

$$f(x_1,\ldots,x_d) = f_1(x_1) \cdots f_d(x_d). \tag{5.4.12}$$

Indeed, if $\mathrm{pdf}_{\boldsymbol{X}} = f$ of the form (5.4.12) then, by Fubini's Theorem, we have

$$\begin{aligned}\mathrm{cdf}_{\boldsymbol{X}}(\boldsymbol{x}) &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_1(x_1)\cdots f_d(x_d)\,\mathrm{d}\,x_1 \cdots \mathrm{d}\,x_d \\ &= \int_{-\infty}^{x_1} f_1(\xi_1)\,\mathrm{d}\,\xi_1 \cdots \int_{-\infty}^{x_d} f_d(\xi_d)\,\mathrm{d}\,\xi_d \\ &= \mathrm{cdf}_{X_1}(x_1)\cdots \mathrm{cdf}_{X_d}(x_d),\end{aligned} \tag{5.4.13}$$

which means that $\mathrm{cdf}_{\boldsymbol{X}}$ is of the form (5.4.9).
Conversely, assuming $\mathrm{cdf}_{\boldsymbol{X}}$ satisfies (5.4.9) then, by the Marginal Distribution Theorem 4.3.5, we know that

$$\mathrm{cdf}_{X_i}(x) = \lim_{\boldsymbol{x}'_i \to \infty} \mathrm{cdf}_{\boldsymbol{X}}(\boldsymbol{x}) = \lim_{\boldsymbol{x}'_i \to \infty} \prod_{i=1}^{d} F_i() \tag{5.4.14}$$

□

5.4.7. Definition of convolution of two functions. Given any two integrable functions $f, g : \mathbb{R} \to \mathbb{R}$ their *convolution* is a function $h : \mathbb{R} \to \mathbb{R}$, usually denoted at $f * g$, given by

$$h(w) = f * g(w) \text{ for all } w \in \mathbb{R}. \tag{5.4.15}$$

5.4.8. Theorem (a sum's density is the densities's convolution). *If X and Y are continuously distributed independent random variables then $X + Y$ is also continuously distributed and*

$$\mathrm{pdf}_{X+Y}(w) = \mathrm{pdf}_X * \mathrm{pdf}_Y(w) = \int_{-\infty}^{\infty} \mathrm{pdf}_X(x)\,\mathrm{pdf}_Y(w-x)\,\mathrm{d}\,x. \tag{5.4.16}$$

5.4.9. Example (sum of independent uniform random variables). Take X and Y independent in $\mathrm{U}(0,1)$. Then their (common) density is

$$\begin{aligned}\mathrm{pdf}_X(x) &= f(x) = \mathbb{1}_{[0,1]}(x) := \begin{cases} 1 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases} \\ \mathrm{pdf}_Y(y) &= f(y).\end{aligned} \tag{5.4.17}$$

Note that since $\operatorname{spt} f = [0,1]$,[4] and using the Iverson brackets we have

$$\begin{aligned} f*f(w) &= \int_{-\infty}^{\infty} f(x)f(w-x)\,\mathrm{d}\,x \\ &= \int_{-\infty}^{\infty} \mathbb{1}_{[0,1]}(x)\mathbb{1}_{[0,1]}(w-x)\,\mathrm{d}\,x \\ &= \int_{-\infty}^{\infty} [\![0<x<1]\!][\![0<w-x<1]\!]\,\mathrm{d}\,x \\ &= \int_{-\infty}^{\infty} [\![0<x<1]\!][\![w-1<x<w]\!]\,\mathrm{d}\,x \\ &= \operatorname{length}(0,1)\cap(w-1,w) \\ &= [\![0<w \text{ and } w-1<1]\!](\min\{1,w\}-\max\{0,w-1\}). \\ &= \mathbb{1}_{[0,2]}(\min\{1,w\}-\max\{0,w-1\}). \end{aligned} \tag{5.4.18}$$

Explicitly, we may distinguish the following cases:

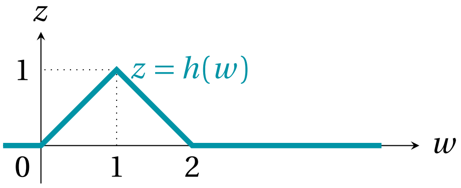
(i) If $w<0$ or $w>2$ then $h(w)=0$.

(ii) If $0<w<1$, then

$$h(w) = w-0 = w. \tag{5.4.19}$$

(iii) If $1<w<2$, then

$$h(w) = 1-(w-1) = 2-w. \tag{5.4.20}$$



In these examples, pay careful attention to the limits in the integration!

5.4.10. Change of variables. An important result in advanced calculus is the change of variables in integrals (also known as multivariate substitution). Let us “recall” this result. For this fix an integer $d \in \mathbb{N}$ and consider an integrable function $f : \mathbb{R}^d \to \mathbb{R}$. Let $\mathscr{C}$ and $\mathscr{D}$ be two domains in $\mathbb{R}^d$ related by a *diffeomorphism*, i.e., there exists a map

$$\begin{aligned} \hat{\boldsymbol{x}} : \quad \mathscr{C} &\to \mathscr{D} \\ \boldsymbol{u} &\mapsto \hat{\boldsymbol{x}}(\boldsymbol{u}) \end{aligned} \tag{5.4.21}$$

such that

- ⋆ $\hat{\boldsymbol{x}}$ is differentiable,
- ⋆ $\mathrm{D}\hat{\boldsymbol{x}}$ is continuous and $\det \mathrm{D}\hat{\boldsymbol{x}} \neq 0$,
- ⋆ $\hat{\boldsymbol{x}}$ is invertible,
- ⋆ $\hat{\boldsymbol{x}}^{-1}$ is continuously differentiable.

[4] The support of a function f, $\operatorname{spt} f$, is the complement of the biggest open set on which the function is zero. Intuitively, the support of f is the smallest closed set on which f is non-zero.

Then then following “change of variables formula” is satisfied

$$\int_{\mathscr{C}} f(\hat{\boldsymbol{x}}(\boldsymbol{u}))|\det \mathrm{D}\hat{\boldsymbol{x}}(\boldsymbol{u})|\,\mathrm{d}\boldsymbol{u} = \int_{\mathscr{D}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{5.4.22}$$

(This is formally written as the “substitution” $\boldsymbol{x} = \hat{\boldsymbol{x}}(\boldsymbol{u})$.) Similarly, noting that the inverse of a diffeomorphism is a diffeomorphism, and writing $\hat{\boldsymbol{u}} := \hat{\boldsymbol{x}}^{-1}$, we also have

$$\int_{\mathscr{D}} f(\hat{\boldsymbol{u}}(\boldsymbol{x}))|\det \mathrm{D}\,\hat{\boldsymbol{u}}(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} = \int_{\mathscr{C}} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}. \tag{5.4.23}$$

(This being the back-substitution $\boldsymbol{u} := \hat{\boldsymbol{x}}^{-1}(\boldsymbol{x}) = \hat{\boldsymbol{u}}(\boldsymbol{x})$
The result (5.4.22) has a useful application to random vectors with densities.

5.4.11. Theorem. *Suppose X and Y have joint density $f(x,y)$, and that $U = \hat{u}(X,Y)$, $V = \hat{v}(X,Y)$, where $(\hat{u},\hat{v}) : \mathscr{C} \to \mathscr{D}$ is a diffeomorphism with inverse $\left(\hat{x},\hat{y}\right) : \mathscr{D} \to \mathscr{C}$. Then the joint density of U and V is*

$$g(u,v) = f(\hat{x}(u,v),\hat{y}(u,v))J(u,v) \tag{5.4.24}$$

where $J(u,v)$ is the Jacobian of $\left(\hat{x},\hat{y}\right)$ at (u,v), i.e.,

$$J(u,v) = \left|\det\begin{bmatrix}\partial_u\hat{x}(u,v) & \partial_v\hat{x}(u,v)\\ \partial_u\hat{y}(u,v) & \partial_v\hat{y}(u,v)\end{bmatrix}\right| \tag{5.4.25}$$

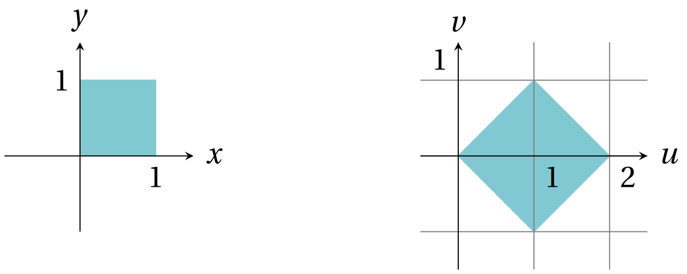
5.4.12. Example.

PROBLEM. *Let $X, Y \in \mathrm{U}[0,1]$ be independent, and take $U := X+Y$, $V := X-Y$. Find the joint density $\mathrm{pdf}_{(U,V)}$ and use it to decide whether U and V are independent.*

Solution. Here $\mathrm{pdf}_{(X,Y)} =: f = \mathbb{1}_{[0,1]^2}$. As $U+V = 2X$, $U-V = 2Y$, and $0 \le X \le 1$ and $0 \le Y \le 1$ with probability 1, also, with probability 1 we must have

$$0 \le U+V \le 2 \text{ and } 0 \le U-V \le 2. \tag{5.4.26}$$

It follows that $(U,V) \in A$, where A is the region in the (u,v)-plane bounded by the lines $u+v=0$, $u+v=2$, $u-v=0$, $u-v=2$.



To find the Jacobian, obtain $x = (u+v)/2$, $y = (u-v)/2$. The four partial derivatives are trivial, leading to $J(u,v) = 1/2$ for all u, v. Hence $\mathrm{pdf}_{(U,V)} = 1/2\mathbb{1}_A$, where A is the diamond described above.
A necessary condition for (U,V) to be independent is for the joint density to be a tensor product, i.e.,

$$\mathrm{pdf}_{(U,V)}(u,v) = \mathrm{pdf}_U(u)\mathrm{pdf}_V(v) \quad \forall (u,v) \in \mathbb{R}^2, \tag{5.4.27}$$

but this means that the support of $\mathrm{pdf}_{(U,V)}$ is a Cartesian product, which is not the case because of it being diamond-shaped. Hence (U,V) is not independent.
An alternative explanation on why U and V are not independent, is the fact that Y is positive and thus $U = X+Y \geq X-Y = V$, hence the choice of U gives us some information on the choice of V.

5.4.13. Example.

PROBLEM. *Let* $X, Y \in \mathrm{Exp}(\lambda)$, *independent, and suppose* $U := X+Y$, $V := X/Y$. *Calculate* pdf_{U+V}. *Are* U *and* V *independent?*

Solution.

$$\mathrm{pdf}_{(X,Y)}(x,y) = \lambda^2 \exp(-\lambda(x+y))\mathbb{1}_{\mathbb{R}^+\times\mathbb{R}^+}(x,y) =: f(x,y). \tag{5.4.28}$$

Now $x = vy$, so $u = y(1+v)$, leading to $y = u/(1+v)$ and $x = uv/(1+v)$. The Jacobian is then $J(u,v) := u/(1+v)^2$, and by Theorem 5.4.11, we have

$$\begin{aligned}\mathrm{pdf}_{(U,V)}(u,v) = f(x,y)J(u,v) &= f\left(\frac{uv}{1+v},\frac{u}{1+v}\right)\frac{u}{(1+v)^2}\\ &= \lambda^2\exp(-\lambda\frac{uv+v}{1+v})\mathbb{1}_{\mathbb{R}^+\times\mathbb{R}^+}\left(\frac{uv}{1+v},\frac{u}{1+v}\right)\frac{u}{(1+v)^2}\\ &= \lambda^2\exp(-\lambda u)\left[\!\left[\frac{uv}{1+v}>0 \text{ and } \frac{u}{1+v}>0\right]\!\right]\\ &= \frac{\lambda^2 u\,\mathrm{e}^{-\lambda u}}{(1+v)^2}\mathbb{1}_{\mathbb{R}^+\times\mathbb{R}^+}(u,v).\end{aligned} \tag{5.4.29}$$

The last passage being justified by the fact that

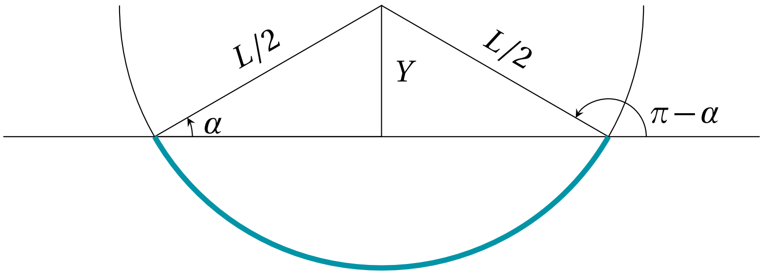
$$\big(x>0 \text{ and } y>0\big) \Leftrightarrow (u>0 \text{ and } v>0). \tag{5.4.30}$$

Notice that $g(u,v)$ splits into the product of $\lambda^2 u\exp(-\lambda u)$ on $u>0$ with $1/(1+v)^2$ on $v>0$, showing that U and V are independent.

5.4.14. Example (Buffon's needle).

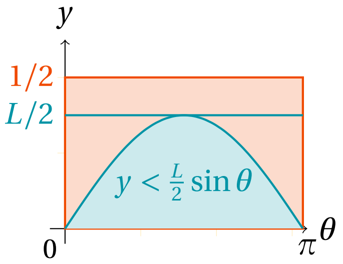
PROBLEM. *Parallel lines, unit distance apart, are marked on a flat board. A needle of length* L, *with* $L<1$, *is dropped at random on this board. With what probability will it meet a line?*

Solution. Write Y as the distance of the centre of the needle from the nearest line, and Θ as the orientation of the needle with respect to the lines. Model: Y and Θ are independent, with $Y \in \mathrm{U}[0,1/2]$ and $\Theta \in \mathrm{U}[0,\pi]$.
That means that the point (Y,Θ) is a random point in the rectangle $0<\theta<\pi$, $0<y<1/2$. Their joint density is the constant $2/\pi$ supported on this rectangle. We identify what region of this rectangle corresponds to the needle crossing the line.



Let α be such that $\sin(\alpha) = Y/(L/2)$. The region we seek is where $\alpha < \theta < \pi - \alpha$, i.e. where $Y < \frac{L}{2}\sin(\Theta)$. In symbols this region is

$$S := \left\{(y,\theta) \in \mathbb{R}^2 : \theta \in [0,\pi] \text{ and } 0 < y < \frac{L}{2}\sin(\theta)\right\} \tag{5.4.31}$$



The probability we seek is the integral of the density function $(2/\pi)\mathbb{1}_{[0,\pi]\times[0,1/2]}$ over this region, i.e.,

$$\begin{aligned}\int_S \frac{2}{\pi}\mathbb{1}_{[0,\pi]\times[0,1/2]}(\theta,y)\,\mathrm{d}y\,\mathrm{d}\theta &= \frac{2}{\pi}\int_0^\pi \int_0^{\frac{L}{2}\sin(\theta)} \mathrm{d}y\,\mathrm{d}\theta \\ &= \frac{2}{\pi}\int_0^\pi \frac{L}{2}\sin\theta\,\mathrm{d}\theta = \frac{L}{\pi}[-\cos\theta]_{\theta=0}^{\theta=\pi} = \frac{2L}{\pi}.\end{aligned} \tag{5.4.32}$$

5.5. Measure density

Not all random variables are continuously distributed. For example, a Bernoulli trial X with values 0 and 1, say, and failure rate q, has a

$$\mathrm{cdf}_X(x) = \begin{cases} 0 & \text{if } x \le 0 \\ q & \text{if } 0 < x \le 1 \\ 1 & \text{if } 1 < x. \end{cases} \tag{5.5.1}$$

The function cdf_X has jumps at 0 and 1 and thus X has no density function, say f, as that would require

$$\int_{-\delta}^{\delta} f \ge q \quad \forall\,\delta > 0, \tag{5.5.2}$$

which is not possible because for any integrable function g one must have[5]

$$\lim_{\delta\to 0}\int_{-\delta}^{\delta} g = 0. \tag{5.5.3}$$

Despite this lack of density, it is still possible to build a rigorous mathematical theory that extends the concept of function so that all random variables have a density. To do this the density function must be replaced by a density *distribution* (or *generalised function*).

[5] This property of integrals is called the *absolute continuity of integrals* and can be found in a good analysis text, e.g., Lieb and Loss, 2001.

5.5.1. Generalised functions. A generalised function is, roughly speaking, a mathematical construct that includes all "standard" (or "classical") functions[6] and extra objects that cannot be rigorously described as functions.
For example a mass measure $\mathcal{M}$ on $\mathbb{R}$ concentrated at a single point, say $0 \in \mathbb{R}$, has no density function, yet a physicist might like to think of it as being a "function", δ such that

$$\delta(x) = \begin{cases} 0 \text{ if } x \neq 0, \\ \infty \text{ if } x = \infty. \end{cases} \tag{5.5.4}$$

Of course, for a mathematician, if this δ were a function it would be Lebesgue-almost identical to 0 and hence integrable with $\int \delta = 0$. But the physicist will quickly prompt that the "delta function", being the density of $\mathcal{M}$, must enjoy as well the property that

$$\mathcal{M}(A) = \int_A \delta(x)\,\mathrm{d}\,x, \text{ for any measurable } A \subseteq \mathbb{R}. \tag{5.5.5}$$

5.5.2. Mass density random variables. Most of what we describe next works for general random variables. Nonetheless, unless otherwise stated, we will consider from now on only random variables which have a generalised probability distribution function consisting of the sum of a density function and a (possibly infinite) countable sum of modulated Dirac masses. That is, given a random variable X we assume that

$$f := \mathrm{pdf}_X = f^{\mathrm{reg}} + \sum_{i \in J} f^i \delta_{x_i} \tag{5.5.6}$$

for some function f^{reg} and a point mass distribution $f_i \in \mathbb{R}^+$ for all $i \in J$, a finite or countably infinite set. In particular, if φ is a smooth function, then we have

$$\int \varphi(x)\,\mathrm{d}\,F_X(x) = \int \varphi(x) f^{\mathrm{reg}}(x)\,\mathrm{d}\,x + \sum_{i \in J} f^i \varphi(x_i). \tag{5.5.7}$$

For example, in this case the expectation rule becomes

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}} g(x)\,\mathrm{d}\,F_X(x) = \int_{\mathbb{R}} g(x) f^{\mathrm{reg}}(x)\,\mathrm{d}\,x + \sum_{i \in J} f^i g(x_i). \tag{5.5.8}$$

5.5.3. Example (the Devil's Staircase as seen by the angels). This example is about a random variable that has no density function nor a density-mass only
Consider Ω to be the space of infinite sequences of independent fair coin tosses, where each $\omega \in \Omega$ is a sequence of symbols in $\{H, T\}$ (head or tail), i.e.,

$$\omega = (\omega(1), \omega(2), \ldots, \omega(k), \ldots) \text{ and } \omega(k) = H \text{ or } T \quad \forall\, k \in \mathbb{N}, \tag{5.5.9}$$

with the probability measure P induced by the "elementary" binomial toss probability p on $\{H, T\}$ where $p\{H\} = 1/2$ for heads and $p\{T\} = 1/2$ for tails. For example,

[6] Generalised functions are also knows as *distributions* in most of the modern analysis literature. Because the word "distribution" is somewhat overloaded we prefer to use "generalised functions". In this course, we do not have time, space nor need to develop the theory of generalised functions, so we spend a couple of words about them here and move on.

the probability of the event where the first toss is a head, the 5th a tail, and the 6th or 7th toss is a head, can be calculated as follows

$$\begin{aligned} &P\{\omega \in \Omega :\ \omega(1) = H \text{ and } \omega(5) = T \text{ and } (\omega(6) = H \text{ or } \omega(T) = H)\} \\ &\quad = p\{H\}p\{T\}P\{\omega(6) = H \text{ or } \omega(T) = H\} = \frac{1}{2}\frac{1}{2}\frac{3}{4} = \frac{3}{16}. \end{aligned} \tag{5.5.10}$$

The space (Ω, P) constructed here is a common model probability space found in many standard textbooks Billingsley, 1995; Jacod and Protter, 2003, e.g.

Consider playing the following game now. A chocolate bar C_0, of length 2 metres is divided into three pieces, A_1, B_1 and C_1, of equal length. You toss the coin then bar B_1 goes to the Bank, bar C_1 is saved for the next round and bar A_1 goes to you if the outcome is T, otherwise (if the toss produces H) A_1 goes to the bank. The next step is played with C_1 divided into three pieces of equal length, A_2, B_2, C_2, you toss the coin, B_2 to the bank, C_2 kept for next toss, and A_2 to you if T or to Bank if H. Consider now the random variable X that gives the length of chocolate obtained "after" infinitely many tosses.

To model this game, introduce the basic random variable $V : \{H, T\} \to \mathbb{R}$, where $V(H) := 0$, $V(T) := 1$, and let $X : \Omega \to \mathbb{R}$ be the random variable defined as follows

$$X(\omega) := 2\sum_{k=1}^{\infty} \frac{V(\omega(k))}{3^k}. \tag{5.5.11}$$

For example we have

$$\begin{aligned} X(T, T, T, T \ldots) = 1, X(T, H, H, H \ldots) = \frac{2}{3}, \\ X(H, T, T, T \ldots) = \frac{1}{3}, \text{ and } X(H, T, T, T \ldots) = \frac{2}{9}. \end{aligned} \tag{5.5.12}$$

PROPOSITION. *Let $K \in \mathbb{N}$. For each integer of the form*

$$j := 2\sum_{k=1}^{K-1} \sigma_k 3^k + l, \ \textit{with}\ 0 \le l \le 3 \ \textit{and}\ (\sigma_1, \ldots, \sigma_{K-1}) \in \{0, 1\}^{K-1}, \tag{5.5.13}$$

we have that

$$P\left\{X \le \frac{j}{3^K}\right\} = \left(\sum_{k=1}^{K} \sigma_k 2^k + \lceil l/2 \rceil\right)\frac{1}{2^K}. \tag{5.5.14}$$

Proof The complete proof of this result is not too hard and left as an exercise.
As an example take $K = 3$, then we may produce the following schedule

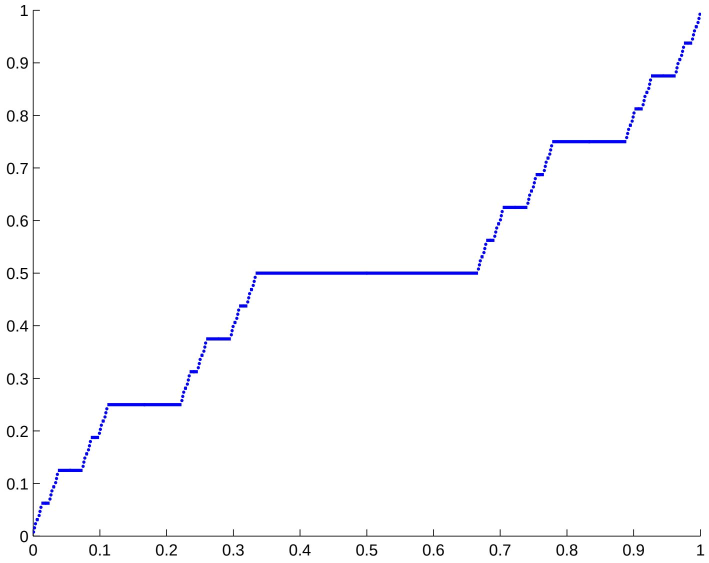
(σ_1, σ_2)	$(0,0)$				$(0,1)$				$(1,0)$				$(1,1)$			
l	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
j	0	1	2	3	6	7	8	9	18	19	20	21	24	25	26	27
$P\{X \le j/3^3\}$	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{7}{8}$	$\frac{7}{8}$	1

(5.5.15)

□

It follows that the cumulative distribution function, F_X, of X is the so-called Devil's Staircase, which is a well-known fractal structure. The function F_X is continuous (it has no jumps), it is differentiable almost everywhere, with derivative 0, but it has infinitely many points where it is differentiable only on one side. So its generalised derivative is not a function and neither a countable sum of mass points, but a *diffuse* *singular measure* concentrated on a set, denoted $\Gamma \subset [0,1)$, and known as the Cantor set.

An approximation of the Devil's Staircase can be plotted as follows



The Cantor set Γ can be shown to consist of all those real numbers between 0 and 1 with a base-3 expansion that has no 1's (for those numbers that admit two different expansions such as $1/3 = 0.1 = 0.0222\ldots$, it is enough to have one expansion with no 1's). A diagonal argument shows then that the Cantor set is uncountable. Furthermore, it is possible to show that Γ is closed and has Lebesgue measure 0. Thus the *distribution measure* P_X of X, defined by the "elementary blocks"

$$P_X(a,b] := P\{a < X \le b\} = F_X(b) - F_X(a), \tag{5.5.16}$$

is a *singular* measure with respect to the Lebesgue measure and the *Radon–Nikodym* Theorem does not apply.

Using the terminology introduced in this section, if f_X denotes the generalised density function of X, then the decomposition $f_X = f_X^{\text{reg}} + f_X^{\text{sng}}$ satisfies

$$f_X^{\text{reg}} = 0 \text{ and } \langle f_X^{\text{sng}} | \phi \rangle = \int_{-\infty}^{\infty} \phi(x)\,\mathrm{d}\,F_X(x), \tag{5.5.17}$$

where the last integral is the *Riemann–Stieltjes* integral. The fact that $f_X^{\text{reg}} = 0$ means that the only contribution to the generalised probability density function in this case comes from its singular part. For more details, you could consult Stroock (1999).

5.5.4. Expectation rule. Laws, c.d.f. and p.d.f. are useful tools. One the main purposes for them is to "get rid" of Ω. This means, for example, that by using the law of a random variable on Ω, integration on Ω can be replaced by integration on $\mathbb{R}$ (this is some kind of change-of-variable formula). This is an advantage, especially in practical applications of probability, because generally the space Ω is very hard to model and manipulate, whereas $\mathbb{R}$ is (as you know) quite an easy place to work in.

PROPOSITION (expectation rule). *If X is a real-valued random variable, with $P_X = \mathrm{law}_X$ and $F_X = \mathrm{dtrb}_X$. Let g be some function whose domain includes the set $X(\Omega)$. Then $g(X)$ is a random variable too and*

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}} g(x) P_X(\mathrm{d}\,x) = \int_{\mathbb{R}} g(x)\,\mathrm{d}\,F_X(x), \tag{5.5.18}$$

where the last integral is understood as a Riemann-Stieltjes integral.

Exercises and problems on continuous random variables and densities

Exercise 5.1. Suppose X and Y are independent continuous random variables in $\mathrm{U}[0,1]$, i.e., uniformly distributed over the unit interval. Draw suitable diagrams, and hence evaluate the following probabilities.

(a) $\mathcal{P}(\{X < 0.5\} \cap \{Y > 0.5\} \cap \{Y - X < 0.5\})$.
(b) $\mathcal{P}(X + Y \leq z)$ when $0 < z < 1$.
(c) $\mathcal{P}(X + Y \leq z)$ when $1 < z < 2$.
(d) $\mathcal{P}(W \leq w)$ when $0 < w < 1$ and W is the *fractional* part of $X + Y$, i.e.

$$W = \begin{cases} X + Y & \text{if } X + Y < 1; \\ X + Y - 1 & \text{if } X + Y \geq 1, \end{cases}$$

Exercise 5.2. A breakdown truck cruises along a straight road of unit length linking Newtown to Seaport; help for stranded motorists is also available in each town. Steve's car runs out of petrol. If x and y are the distances of his car and the truck from Newport, the distribution of (x, y) can reasonably be taken as uniform over the unit square. Show that, for $0 < t < 1/2$, the probability that Steve's nearest help is within distance t is $4t(1-t)$.

Exercise 5.3. Let $P = (x, y)$ be a point chosen "uniformly at random" in the circle, centre $(0,0)$ and radius 1. By setting up the appropriate probability space, find the probabilities that (a) P is distance at most $1/2$ from the origin (b) $y > 1/\sqrt{2}$ (c) $|x-y| < 1$ and $|x+y| < 1$.

Exercise 5.4. A random point (X, Y) has the uniform distribution over the unit square $\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 1\}$, and u is a given value with $0 < u < 1/2$. Events A and B are defined as

$$A := \{X \geq 2Y\} \text{ and } B := \{Y \leq X \leq Y + u\}. \tag{P5.4.1}$$

Sketch a diagram of the unit square, illustrating the events A, B and $A \cap B$. Find the probabilities of these three events, and then find the value of u such that A and B are independent.

Exercise 5.5. Suppose X and Y are independent in $\mathrm{Exp}(\lambda)$. Show that $X + Y$ has density

$$\mathrm{pdf}_{X+Y}(t) = \lambda^2 t \exp(-\lambda t)\,\mathbb{1}_{\mathbb{R}^+}(t). \tag{P5.5.1}$$

Hence find the density of the sum of *three* independent random variables, each in $\mathrm{Exp}(\lambda)$.

Exercise 5.6. Let $\{X_i\}$ be independent random variables, all with the identical distribution of density

$$\mathrm{pdf}_{X_i} = \lambda \exp(-\lambda x)[\![x > 0]\!] \tag{P5.6.1}$$

and write $S_n := X_1 + \cdots + X_n$. Show that the mgf of S_n is the same as that of a $\Gamma(n,\lambda)$ variable (whose density is $\lambda^n x^{n-1} \exp(-\lambda x)/(n-1)![\![x > 0]\!]$).

Exercise 5.7. A tree scatters 20 seeds at random. The distance of any seed from the tree has an Exponential distribution $\mathrm{Exp}(\lambda)$ with mean $\lambda = 4$ metres. What is the distribution of the distance of the closest seed to the tree?
Hint. Think of the minimum of twenty $\mathrm{Exp}(\lambda)$, with constant λ.

Exercise 5.8. X and Y are independent and both in $\mathrm{U}[0,1]$. Let

$$V := X \vee Y := \max\{X, Y\} \text{ and } W := X \wedge Y := \min\{X, Y\}. \tag{P5.8.1}$$

Find the densities of V and W. Are they independent? Show that $\mathrm{E}[VW] = 1/4$.

CHAPTER 6

Conditional expectation

Conditioning and independence is the single most important concept in Probability. We have already encountered conditional probabilities and independence of random variables. How about conditional random variables? Such a concept is indeed possible and turns out to be quite useful. In this chapter we develop it.

6.1. Expectation conditional to a discrete random variable

6.1.1. Expectation conditional to a given event. Let Y be a random variable on the probability space $(\Omega,\mathscr{F},\mathcal{P})$ and let $A\in\mathscr{F}$ be a given event. Let $\mathcal{Q}$ be the probability measure $\mathcal{P}$ conditional to A, i.e.,

$$\mathcal{Q}(B):=\mathcal{P}(B|A)\text{ for each }B\in\mathscr{F}. \tag{6.1.1}$$

The *expectation of Y given* (or *conditional to*) A is defined as the average of Y with respect to $\mathcal{Q}$. I.e.,

$$\mathrm{E}[Y|A]:=\int_\Omega Y(\omega)\,\mathrm{d}\,\mathcal{Q}(\omega). \tag{6.1.2}$$

If Y is discrete then

$$\mathrm{E}[Y|A]=\sum_{y\in Y(\Omega)} y\mathcal{P}\big(Y=y|A\big). \tag{6.1.3}$$

Note that the expectation of Y conditional to A can be also written as

$$\mathrm{E}[Y|A]=\int_A Y\,\mathrm{d}\mathcal{P}=\int_\Omega Y\mathbb{1}_A\,\mathrm{d}\mathcal{P}. \tag{6.1.4}$$

6.1.2. Expectation conditional to a partition. Let Y be a random variable on $(\Omega,\mathscr{F},\mathcal{P})$ and let $\mathscr{P}:=\{\Omega_\alpha\}_{\alpha\in\alpha\in\mathscr{A}}$ be a countable partition of Ω (so $\mathscr{A}$ is countable). We may define, for each $\alpha\in\mathscr{A}$, *conditional expectation*

$$\mathrm{E}[Y|\Omega_\alpha]=\sum_{y\in Y(\Omega)} y\mathcal{P}\big(Y=y|\Omega_\alpha\big), \tag{6.1.5}$$

and then the random variable

$$Z=\sum_{\alpha\in\mathscr{A}}\mathrm{E}[Y|\Omega_\alpha]\mathbb{1}_{\Omega_\alpha}, \tag{6.1.6}$$

that is

$$Z(\omega)=\mathrm{E}[Y|\Omega_\alpha]\text{ where }\alpha\in\mathscr{A}\text{ is the unique index such that }\omega\in\Omega_\alpha. \tag{6.1.7}$$

It follows that Z is a discrete random variable on Ω, which we call the *expectation of Y conditional to the partition* and denote by $\mathrm{E}[Y|\mathscr{P}]$. By the Partition of Total Probability Theorem 2.2.2 we see that

$$\begin{aligned}\mathrm{E}[Z] &= \sum_{\alpha\in\mathscr{A}} \mathrm{E}[Y|\Omega_\alpha]\mathcal{P}(\Omega_\alpha) = \sum_{\alpha\in\mathscr{A}}\sum_{y\in Y(\Omega)} y\mathcal{P}\big(Y=y|\Omega_\alpha\big)\mathcal{P}(\Omega_\alpha)\\ &= \sum_{y\in Y(\Omega)} y \sum_{\alpha\in\mathscr{A}} \mathcal{P}\big(Y=y|\Omega_\alpha\big)\mathcal{P}(\Omega_\alpha) = \sum_{y\in Y(\Omega)} y\mathcal{P}\big(Y=y\big) = \mathrm{E}[Y].\end{aligned} \tag{6.1.8}$$

6.1.3. Random-variable induced sigma-algebra. Let X be a random variable on the probability space $(\Omega, \mathscr{F}, \mathcal{P})$, consider the following subcollection of $\mathscr{F}$,

$$\sigma(X) := \left\{A\in\mathscr{F} : A = X^{-1}(B) \text{ for some } B\in\mathscr{B}\right\}, \tag{6.1.9}$$

where $\mathscr{B}$ is the Borel sigma-algebra in $\mathbb{R}$. It can be checked that $\sigma(X)$ is a sigma-algebra and $\sigma(X)\in\mathscr{F}$ (we say it is a sigma-subalgebra of $\mathscr{F}$). If X is a discrete random variable, then $\sigma(X)$ can be generated by the *atomic elements* given by X

$$\mathscr{P}_X := \left\{A\in\mathscr{F} : A = X^{-1}(\{x\}) \text{ for some } x\in\mathbb{R}\right\}, \tag{6.1.10}$$

where by the *sigma-algebra generated* by a collection $\mathscr{C}$ we mean the smallest sigma-algebra containing $\mathscr{C}$. Note that $\mathscr{P}_X$ constitutes a partition of the probability space Ω.

6.1.4. Expectation conditional to a discrete random variable. Let X and Y be two random variables, and let X be discrete. Then X induces a partition as follows

$$\Omega_k := \{X = x_k\}, \text{ for each } k\in\mathscr{K} \tag{6.1.11}$$

where

$$X(\Omega) = \{x_k\}_{k\in k\in\mathscr{K}}, \text{ for a countable set } \mathscr{K}\subseteq\mathbb{N}_0. \tag{6.1.12}$$

Denoting by $\mathscr{P} := \{\Omega_k\}_{k\in k\in\mathscr{K}}$, we define the *conditional expectation* of Y with respect to (or given) X as

$$\mathrm{E}[Y|X] := \mathrm{E}[Y|\mathscr{P}]. \tag{6.1.13}$$

By the discussion in §6.1.2 it follows that $\mathrm{E}[Y|X]$ is a random variable and that

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y]. \tag{6.1.14}$$

6.1.5. Remark (invariance under injective value transformations). It is worth noting that $\mathrm{E}[Y|X]$ depends on the partition induced by X, as defined in §6.1.4, but *not on the actual values of X*. I.e., if $h:\mathbb{R}\to\mathbb{R}$ is an injection (e.g., a strictly monotone function) then $\mathrm{E}[Y|X] = \mathrm{E}[Y|h(X)]$.

6.1.6. Proposition. *Let X, Y be random variables with X discrete. Define the function*

$$g(x) = \begin{cases}\mathrm{E}[Y|X=x] & \textit{for each } x\in X(\Omega)\\ 0 & \textit{otherwise.}\end{cases} \tag{6.1.15}$$

Then $\mathrm{E}[Y|X] = g(X)$.

Proof It is a straightforward application of the definitions. □

6.1.7. Theorem (conditional expectation is orthogonal projection). *Let X, Y be random variables with X discrete and $\sigma(X)$ the sigma-algebra generated by X. Then we have*

$$\mathrm{E}[Z\,\mathrm{E}[Y|X]] = \mathrm{E}[ZY] \text{ for any } \sigma(X)\text{-measurable random variable } Z, \tag{6.1.16}$$

and

$$\mathrm{E}[\mathbb{1}_A \mathrm{E}[Y|X]] = \mathrm{E}[\mathbb{1}_A Y] \quad \forall A \in \sigma(X). \tag{6.1.17}$$

Proof Exercise. □

6.1.8. Remark (Which orthogonal projection?) The orthogonal projection we mean in Theorem 6.1.7 has to be understood in the context of a particular inner product space, known as the space of random variables with finite mean square

$$\mathrm{L}_2(\mathcal{P}) := \left\{X \in \mathrm{RV}(\mathcal{P}) : \mathrm{E}\left[X^2\right] < \infty\right\}. \tag{6.1.18}$$

6.2. Conditioning with respect to a general random variable

A rigorous treatment of general conditioning requires technical background that is beyond the scope of this course. We briefly describe here the heuristics and main ideas underlying the theory. For a fuller treatment of the topic we refer to Jacod and Protter (2003) or Capiński and Kopp (2004).

6.2.1. Conditional expectation as orthogonal projection. The idea behind the *conditional expectation* of a general random variable Y with respect to a *general* random variable X, is to make identity (6.1.17) work also for a random variable X that is not discrete.

6.3. Conditioning and densities

6.3.1. Definition of marginal. Given (X,Y) with joint density $f(x,y)$, then $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$ are the *marginal* densities of X and Y respectively.

6.3.2. Definition of conditional density. Given (X,Y) with joint density $f(x,y)$, then

$$f(y|x) = f(x,y)/f_X(x)$$

is the *conditional* density of Y, given $X = x$.

6.3.3. Proposition. $f(x,y) = f(y|x)f_X(x)$. It shows how to get a joint density, if one variable is defined in terms of the other. Compare this with $P(A \cap B) = P(B|A)P(A)$, as a way to find the probability of two events, when one depends on the other.

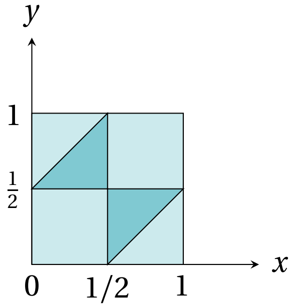
6.3.4. Example (the stick–triangle question revisited).

PROBLEM. *A stick of unit length is broken into three pieces by*
(i) Selecting one point at random, snap the stick
(ii) On the longer part, selecting one point at random and snap it again.
Find the probability we can make a triangle.

Solution. Take x as the position of the first point, y that of the second. Then $f_X(x) = 1$ on $0 < x < 1$.
If $\frac{1}{2} < x < 1$, then $f(y|x) = 1/x$ on $0 < y < x$; and if $0 < x < \frac{1}{2}$, then $f(y|x) = 1/(1-x)$ on $x < y < 1$. So we find the joint density of (X, Y) using the Corollary: i.e.

$$f(x,y) = \begin{cases} 1/(1-x) & 0 < x < \frac{1}{2}, x < y < 1 \\ 1/x & \frac{1}{2} < x < 1, 0 < y < x \end{cases}$$

The same diagram as before shows where (X, Y) must fall to give a triangle:



Thus the chance we get a triangle is the integral of $f(x, y)$ over the shaded region. It breaks down as

$$\begin{aligned} P &= \int_0^{1/2} \int_{1/2}^{x+1/2} \frac{1}{1-x} \,\mathrm{d}y\,\mathrm{d}x + \int_{1/2}^{1} \int_{x-1/2}^{1/2} \frac{1}{x} \,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^{1/2} \frac{x}{1-x} \,\mathrm{d}x + \int_{1/2}^{1} \frac{1-x}{x} \,\mathrm{d}x \\ &= 2\log(2) - 1 \approx 0.3863 \end{aligned} \tag{6.3.1}$$

6.3.5. Definition of conditional expectation. The *conditional expectation* $E(X|Y = y)$ is defined as

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f(x|y) dx.$$

Then we define $E(X|Y)$ as the value of $E(X|Y = y)$ as y varies.

6.3.6. Theorem. *$E(X) = E(E(X|Y))$. Indeed, $E(h(X)) = E(E(h(X)|Y))$.*

6.3.7. Example.

PROBLEM. *Suppose Y is $U(0, 1)$, and then X is $U(0, Y)$. Find the mean and variance of X.*

Solution. Plainly, $E(X|Y = y) = y/2$, so $E(X|Y) = Y/2$, hence $E(X) = E(Y/2) = E(Y)/2 = 1/4$.
Given $Y = y$, X is $U(0, y)$ from which $E(X^2) = y^2/3$ (easy), i.e. $E(X^2|Y = y) = y^2/3$. Thus $E(X^2|Y) = Y^2/3$, and so $E(X^2) = E(Y^2/3) = 1/9$. Then $\text{Var}(X) = 1/9-(1/4)^2 = 7/144$.

6.3.8. Definition of covariance. The *covariance* of two random variables X and Y is

$$\text{cov}[X, Y] = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]. \tag{6.3.2}$$

6.3.9. Proposition (covariance). *Given two random variables X and Y we have*

$$\text{cov}[X, Y] = \mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]. \tag{6.3.3}$$

Proof Exercise.[$*$] □ [$*$]: Check!

6.3.10. Theorem (independent variables have 0-covariance). *If X and Y are independent, their covariance is zero.*

6.3.11. Remark. The converse of Theorem 6.3.10 is not true. There are examples of X and Y not independent, yet $\text{cov}(X, Y) = 0$.

Exercises and problems on conditional expectation

Exercise 6.1. Suppose X is $\mathrm{B}(n, R)$, where R is drawn from $\mathrm{U}[0, 1]$. Find $\mathrm{E}[X]$ and $\text{var}(X)$.

Exercise 6.2. Let $(X_i)_{i\in\mathbb{N}}$ be an independent sequence of identically distributed random variables. Define

$$S_n := \sum_{i=1}^{n} X_i = \begin{cases} 0 & \text{when } n = 0 \\ X_1 + \cdots + X_n & \text{when } n > 0. \end{cases} \tag{P6.2.1}$$

Show that for each random variable N with $N(\Omega) \subseteq \mathbb{N}_0$ we have

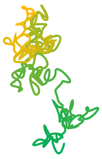
$$\mathrm{E}[S_N] = \mathrm{E}\,N\,\mathrm{E}\,X_1 \text{ and } \text{var}[S_N] = \mathrm{E}\,N\,\text{var}\,X_1 + \text{var}\,N\,\mathrm{E}[X_1]^2. \tag{P6.2.2}$$

Problem 6.3 (factorisation under conditional expectation). Let X and Y be two random variables, and $h : \mathbb{R} \to \mathbb{R}$, show that

$$\mathrm{E}[h(X)Y|X] = h(X)\mathrm{E}[Y|X]. \tag{P6.3.1}$$

CHAPTER 7

Ergodicity



Of the greatest achievements in Probability Theory are the convergence theorems that go by the name of the Weak and Strong Laws of Large Numbers and the Central Limit Theorem. They all deal with sums and averages of idependent and identically distributed random variables. Broadly speaking a result concerning how averages behave in the limit is called an *ergodic theorem*. Unsurprisingly, ergodicity is the main link between Probability Modelling and Statistics. In this chapter we brush over these concepts and discuss the various concepts of convergence needed to understand (and appreciate) these results.

7.1. From Chebyshev to the Weak Law of Large Numbers

One of the most useful applications of Chebyshev's inequality (4.6.8) we will now study what happens when we average an infinite sequence of independent random variables $(X_n)_{n\in\mathbb{N}}$ that have all the same expectation, call it μ, and the same variance, call it σ^2. We are interested in undertanding what happens when n is a large number, i.e., $n \to \infty$ and we look at the sequence average $\sum_{k=1}^{n} X_k/n$.

7.1.1. Definition of sequence average. Whenever one works with a sequence $(X_n)_n$ of random variables, one must be careful to distiguish the n-th *mean* (i.e., the expectation $\mathrm{E}[X_n]$ of each random variable in the sequence) and the n-th *sequence average* $\sum_{i=1}^{n} X_i/n$ often denoted as $\bar{X}_n$ and called simply *average*.

7.1.2. Definition of i.i.d.. Let (X_n) be a sequence of random variables. We say that the sequence is *independent* if and only if any finite subsequence forms an independent random vector in the sense of Definition 4.3.1.
We write that X_n *are i.i.d.* when the random variables X_n are independent and identically distributed (i.e., they have the same distribution function).

7.1.3. Average of two random variables. Let $\bar{X}_2 := (X_1 + X_2)/2$, then we have

$$\mathrm{E}\left[\bar{X}_2\right] = \frac{2\mu}{2} = \mu \tag{7.1.1}$$

and, noting that $X_1 - \mu$ and $X_2 - \mu$ must be independent (because X_1 and X_2[*]) and [*]: Check!

using Theorem 6.3.10 we get

$$\begin{aligned}\operatorname{var}\left[\bar{X}_2\right] &= \mathrm{E}\left[\left|\frac{X_1+X_2}{2}-\mu\right|^2\right] \\ &= \mathrm{E}\left[\left|\frac{(X_1-\mu)+(X_2-\mu)}{2}\right|^2\right] \\ &= \frac{1}{4}\left(\operatorname{var}[X_1]+\operatorname{var}[X_2]+2\underbrace{\operatorname{cov}\left[X_1-\mu,X_2-\mu\right]}_{=0 \text{ by independence}}\right) \\ &= \frac{\sigma^2}{2}.\end{aligned} \tag{7.1.2}$$

By Chebyshev's inequality (4.6.8), we may now conclude that for any $k>0$ we have

$$\mathcal{P}\left[\left|\bar{X}_2-\mu\right|>\frac{k\sigma}{\sqrt{2}}\right]<\frac{1}{k^2}. \tag{7.1.3}$$

Applying Chebyshev to $\bar{X}_1 := X_1$ we have, for the same $k>0$

$$\mathcal{P}\left[\left|\bar{X}_1-\mu\right|>k\sigma\right]<\frac{1}{k^2}. \tag{7.1.4}$$

Comparing inequalities (7.1.3) and (7.1.4) we see that by averaging two identical random variables one obtains a better "confidence threshold" in the sense that the bracket around the mean μ for $\bar{X}_1$ or $\bar{X}_2$ to be in with a given probability ($1/k^2$), is narrower for $\bar{X}_2$ than for $\bar{X}_1$ (by a factor of $\sqrt{2}$). What happens then if we add 3, 4, or n i.i.d.random variables?

7.1.4. Exercise (Average of many random variables). *For a fixed $n\in\mathbb{N}$, consider n random variables, $X_1,\dots,X_n$, that are independent and have the same mean μ an variance σ^2. Consider their sequence (sample) average*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{7.1.5}$$

Show that

$$\mathrm{E}\left[\bar{X}_n\right]=\mu \textit{ and } \operatorname{var}\left[\bar{X}_n\right]=\frac{\sigma^2}{n}. \tag{7.1.6}$$

Chebyshev's inequality now implies that for any $\epsilon>0$ we have

$$\mathcal{P}\left[\left|\bar{X}_n-\mu\right|>\epsilon\right]<\frac{\sigma^2}{\epsilon^2 n}. \tag{7.1.7}$$

7.1.5. Definition of convergence in probability. A random variable sequence $(X_n)_{n\in\mathbb{N}}$ on the probability space $(\Omega,\mathscr{F},\mathcal{P})$ *converges in probability* to the random variable X if and only if

$$\forall\,\epsilon>0: \lim_{n\to\infty}\mathcal{P}[|X_n-X|>\epsilon]=0. \tag{7.1.8}$$

There is a special notation for convergence in probability which is

$$X_n \xrightarrow{\text{prob}} X, \text{ or to emphasise the role of } \mathcal{P},\ X_n \xrightarrow[\mathcal{P}]{\text{prob}} X, \text{ as } n\to\infty. \tag{7.1.9}$$

7.1.6. Theorem (Weak Law of Large Numbers (WLLN)). *If $(X_n)_{n\in\mathbb{N}}$ are i.i.d. with mean $\mu := \mathrm{E}[X_n]$ and variance $\sigma^2 := \operatorname{var} X_n$, and sequence average $\bar{X}_n := \sum_{i=1}^n X_i/n$, then $\bar{X}_n \xrightarrow{\text{prob}} \mu$ as $n \to \infty$.*

Proof Since we would like to give meaning to the "average" of the whole infinite sequence $(X_n)_{n\in\mathbb{N}}$, inequality (7.1.7) comes to help, as it implies that for each $\epsilon > 0$ and each $\delta > 0$ we can find $N = \hat{N}(\delta,\epsilon) \in \mathbb{N}$ such that

$$\mathcal{P}\left[\left|\bar{X}_n - \mu\right| > \epsilon\right] < \delta \quad \forall\, n \geq N. \tag{7.1.10}$$

Keeping ϵ fixed and letting δ be arbitrarily small we obtain that

$$\lim_{n\to\infty} \mathcal{P}\left[\left|\bar{X}_n - \mu\right| > \epsilon\right] = 0. \tag{7.1.11}$$

The sequence $(Y_n)_{n\in\mathbb{N}}$ thus converges in probability to the (constant or sure) random variable μ. □

7.1.7. Example (WLLN and frequency statistics). In a sequence of Bernoulli trials, take $X_n = 1$ if trial n is Success, $X_n = 0$ otherwise. Here $\bar{X}_n$ can be written R_n/n, the fraction of Successes in the first n trials. The Weak Law of Large Numbers of Theorem 7.1.6 tells us that, for a fixed $\epsilon > 0$ (no matter how small), the probability that R_n/n differs from p, the probability of success in a single trial, by more than ϵ goes to zero as n increases. In other words, the average R_n/n is converging in probability to p, which conforms with the intuitive idea that given a coin, we can estimate the probability of it yielding H upon a toss by actually tossing it times as many as we can afford, say n, and counting the amount of times it actually yielded head, say r_n. The WLLN tells us then that the likelihood of R_n/n being close to p by no more than ϵ is smaller the higher is n.

7.1.8. Example. Suppose $(X_n)_{n\in\mathbb{N}}$ are independent, with

$$\mathcal{P}[X_n{=}1] = 1/n \text{ and } \mathcal{P}[X_n{=}0] = 1 - 1/n. \tag{7.1.12}$$

For some fixed $\epsilon \in (0,1)$, we see that

$$\mathcal{P}[X_n > \epsilon] = \mathcal{P}[X_n = 1] \to 0, \tag{7.1.13}$$

i.e., $\bar{X}_n \xrightarrow{P} 0$.
However,

$$\begin{aligned}\mathcal{P}[X_n = 0 \wedge X_{n+1} = 0 \wedge \ldots \wedge X_{n^2} = 0] &= \frac{n-1}{n}\,\frac{n}{n+1}\cdots\frac{n^2-1}{n^2}\\ &= \frac{n-1}{n^2}\\ &\to 0, \text{ as } n \to \infty,\end{aligned} \tag{7.1.14}$$

so, looking the complementary event, we have

$$\mathcal{P}[\text{At least one of } X_n, X_{n+1}, \ldots, X_{n^2} = 1] \to 1. \tag{7.1.15}$$

It follows that no matter how large N is picked, there will always be an $n \geq N$ (chosen amongst $N, N+1, \ldots, N^2$) for which $X_n \neq 0$ with some positive probability, hence $X_n \neq 0$. So it is impossible for the *sequence* (X_n) to converge to zero, even though $X_n \xrightarrow{\text{prob}} 0$.

7.1.9. Definition of almost sure convergence. A random variable sequence $(X_n)_{n\in\mathbb{N}}$ *converges almost surely* (or $\mathcal{P}$*-almost surely*, when necessary to distinguish measures) to a random variable X, written $X_n \xrightarrow{\text{a.s.}} X$ (or $X_n \xrightarrow[\mathcal{P}]{\text{a.s.}} X$), if and only if

$$\mathcal{P}[\omega : X_n(\omega) \to X(\omega)] = \mathcal{P}[X_n \to X] = 1. \tag{7.1.16}$$

7.1.10. Remark. Example 7.1.8, shows a sequence that converges in probability, but not almost surely.

7.1.11. Theorem (Strong Law of Large Numbers). *If (X_n) are iid with mean μ, and $\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$, then $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$.*

7.1.12. Theorem. *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$.*

7.1.13. Example (Kelly Strategy).

PROBLEM. *A casino offers a favourable bet: the chance of Red on the roulette wheel is $p > 0.5$, but the payout is at even money. What should you do, if $p < 1$?*

Solution. Take $X_0 > 0$ as your initial fortune, X_n as your fortune after n bets. An optimist might place his entire fortune on Red, every bet. Then X_{n+1} is either $2X_n$ or 0, with respective probabilities p and $1 - p = q$, so $E(X_{n+1}|X_n) = 2pX_n$.
Thus $E(X_{n+1}) = E(E(X_{n+1}|X_n)) = E(2pX_n) = 2pE(X_n)$, so by induction $E(X_n) = (2p)^n X_0 \to \infty$. On average, your fortune increases without bound. Goody.
BUT $P(X_n \neq 0) = P(\text{All bets win}) = p^n \to 0$, i.e. $P(X_n = 0) \to 1$. You are certain to become bankrupt, eventually. Overall, this optimist uses a bad strategy – even though, "on average", his fortune shoots up.
John L Kelly suggested: bet the fixed fraction, x of your current fortune each time. Here X_{n+1} will consist of $(1 - x)X_n$ (the amount left in the kitty), plus either $2xX_n$ (if you win), or zero (if you lose). Thus X_{n+1} is $(1 + x)X_n$ with probability p, or $(1 - x)X_n$ with probability q.
If, in n bets, you have m wins and $n - m$ losses, then

$$X_n = (1 + x)^m (1 - x)^{n-m} X_0.$$

Divide by X_0, take logs, divide by n to get

$$\frac{1}{n}\log(\frac{X_n}{X_0}) = \frac{m}{n}\log(1 + x) + \frac{m - n}{n}\log(1 - x).$$

call up the SLLN to see that the RHS converges (almost surely) to $p\log(1+x)+q\log(1-x) = f(x)$, say. And realise that the LHS is the mean growth rate of your fortune, per bet.
So we use simple calculus to find the maximum of $f(x)$ – this gives the best long-term mean growth rate. This looks a good criterion to use. The calculus leads to choosing $x = p - q$, which has the nice interpretation of saying: to maximise the growth rate of your capital, bet that fraction of your current wealth that measures the SIZE of your advantage $(p - q)$.

7.1.14. Definition of convergence in law (also known as convergence in distribution). Let X_n have distribution function F_n, and X have distribution function F. Suppose $F_n(x) \to F(x)$ at all points x where F is continuous. We say that X_n converges to X *in distribution*, and write $X_n \xrightarrow{D} X$.

7.1.15. Example. The Central Limit-Theorem 7.3.3 can be reworded as saying that Z_n converges in distribution to Z, where Z is Standard Normal.

7.1.16. Example. Look at the individual probabilities of a Binomial distribution, $\mathrm{B}(n, \lambda/n)$; keep λ fixed, but let $n \to \infty$. These probabilities are seen to converge to the corresponding probabilities in a Poisson distribution, Poiss(λ), demonstrating that the Binomial sequence converges to the Poisson limit, in distribution.

7.1.17. Theorem (convergence in distibution is necessary for convergence in probablity). *If* $X_n \xrightarrow{P} X$, *then* $X_n \xrightarrow{D} X$. Putting things together: convergence almost surely implies convergence in probability which, in turn, implies convergence in distribution. And convergence in distribution is all we need to find limiting probabilities.

7.2. Moments

7.2.1. Definition of moment. The *n-th moment* of a random variable X is $\mathrm{E}[X^n]$, for $n \in \mathbb{N}$. Its *moment generating function (mgf)* is

$$\mathrm{E}\left[\mathrm{e}^{tX}\right] = M_X(t) = \mathrm{mgf}_X(t) = \int \mathrm{e}^{tx} f(x)\,\mathrm{d}x. \tag{7.2.1}$$

7.2.2. Remark (how to generate moments from mgf). Using the power series expansion

$$\exp z = \sum_{n=0}^{\infty} \frac{z^n}{n!} \tag{7.2.2}$$

and the linearity of E, we see that

$$\mathrm{mgf}_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathrm{E}[X^n] = 1 + t\,\mathrm{E}[X] + t^2 \frac{\mathrm{E}\left[X^2\right]}{2} + \cdots. \tag{7.2.3}$$

Hence mgf_X lets us "read off" the values of X's n-th moment, $\mathrm{E}[X^n]$, for all $n \in \mathbb{N}_0$. Indeed, it is an exercise to see that

$$\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathrm{mgf}_X(t)\right]_{t=0} = \mathrm{E}[X^n]. \tag{7.2.4}$$

7.2.3. Example. If X is Exp(λ), then

$$\mathrm{mgf}_X(t) = \int_0^{\infty} \mathrm{e}^{tx} \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \frac{\lambda}{\lambda - t} \text{ for } t < \lambda. \tag{7.2.5}$$

Hence $\mathrm{mgf}_X(t) = (1 - \frac{t}{\lambda})^{-1}$ which we expand as a power series in t as $\sum \frac{t^n}{\lambda^n}$, and can read off $E(X^n)$ as $n!/\lambda^n$.

7.2.4. Theorem (algebra of mgf). *The following properties are true*

$$\operatorname{mgf}_{aX}(t) = \operatorname{mgf}_X(at) \textit{ for any } a \in \mathbb{R}. \tag{7.2.6}$$

$$X \textit{ and } Y \textit{ independent} \Rightarrow \operatorname{mgf}_{X+Y}(t) = \operatorname{mgf}_X(t)\operatorname{mgf}_Y(t). \tag{7.2.7}$$

7.2.5. Proposition (mgf Gaussian distribution). *Let Z have density $f(z) = \frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2}$ on the real line, i.e., Z has the standard Gausian (also known as normal) distribution. Then its mgf is*

$$\operatorname{mgf}_Z(t) = \int \exp(tz) f(z) dz = \exp(t^2/2). \tag{7.2.8}$$

Proof The proof is left as an exercise. □

7.3. Central limit theorem

7.3.1. Standardised sums. In the next paragraphs we will consider (partial) sums of a sequence of random variables, X_n, $n \in \mathbb{N}$, all sharing the same distribution; in particular, we denote

$$\mathrm{E}[X_n] =: \mu \text{ and } \operatorname{var}[X_n] =: \sigma^2, \tag{7.3.1}$$

where μ and σ are independent of n. One such sum, for a given $n \in \mathbb{N}$ is

$$S_n := \sum_{i=1}^n X_i. \tag{7.3.2}$$

We have

$$\mathrm{E}[S_n] = n\mu \tag{7.3.3}$$

and

$$\begin{aligned} \operatorname{var} S_n &= \mathrm{E}\left[S_n^2\right] - n^2\mu^2 \\ &= \mathrm{E}\left[\sum_{i=1}^n X_i^2 + \sum_{1 \le i \ne j \le n} X_i X_j\right] - \left(n\mu^2 + (n^2 - n)\mu^2\right) \\ &= \sum_{i=1}^n \left(\mathrm{E}\left[X_i^2\right] - \mu^2\right) + \sum_{1 \le i \ne j \le n} \left(\mathrm{E}[X_i]\mathrm{E}\left[X_j\right] - \mu^2\right) \\ &= n\sigma^2. \end{aligned} \tag{7.3.4}$$

7.3.2. Theorem (convergence of centered averages mgf). *Suppose $\{X_n\}$ are independent, all with the same mean μ and same variance σ^2. Define*

$$Z_n := \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}. \tag{7.3.5}$$

Then for each $t \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \operatorname{mgf}_{Z_n}(t) = \exp(t^2/2). \tag{7.3.6}$$

Proof Write $M_n(t) := \mathrm{mgf}_{Z_n}$, then

$$\begin{aligned} M_n(t) &= \mathrm{E}\exp(tZ_n) \\ &= \mathrm{E}\big[\exp(t(X_1+X_2+\ldots+X_n-n\mu)/(\sigma\sqrt{n})\big] \\ &= \mathrm{e}^{-nt\mu/(\sigma\sqrt{n})}\,\mathrm{E}\big[\exp(C_{1,n}t(X_1+X_2+\ldots+X_n))\big] \end{aligned} \tag{7.3.7}$$

where $C_{1,n} := 1/(\sigma\sqrt{n})$. Apply the logarithm function on both sides:

$$\log(M_n(t)) = \frac{-nt\mu}{\sigma\sqrt{n}} + \log(\mathrm{E}\exp(C_{1,n}t(X_1+X_2+\ldots+X_n)). \tag{7.3.8}$$

By independence, and Theorem 7.2.4, the second term can be worked out as

$$\log\big(\mathrm{E}\big[\exp(C_{1,n}tX)\big]^n\big) = n\log(E(\exp(C_{1,n}tX)) = n\log(M(C_{1,n}t)), \tag{7.3.9}$$

where, thanks to the i.i.d.assumption, we have posed

$$M(s) := \mathrm{mgf}_{X_i}(s) \text{ for all } i \in \mathbb{N}. \tag{7.3.10}$$

This implies that

$$M(s) = 1 + \mu s + \frac{\mu^2+\sigma^2}{2}s^2 + \beta(s) \text{ where } \left|\beta(s)\right| \le C_{2,M}s^3, \tag{7.3.11}$$

and thus

$$M(C_{1,n}t) = 1 + \mu C_{1,n}t + \frac{\mu^2+\sigma^2}{2}C_{1,n}^2t^2 + \tilde\beta_n(t) \tag{7.3.12}$$

where

$$\left|\tilde\beta_n(t)\right| \le C_{2,M}C_{1,n}^3t^3 = \frac{C_{2,M}t^3}{\sigma^3}n^{-3/2}. \tag{7.3.13}$$

By Taylor's expansion we have

$$\log(1+x) = x - x^2/2 + \gamma(x) \text{ with } \left|\gamma(x)\right| \le C_{3,\log}x^3. \tag{7.3.14}$$

Defining x and $\delta_n(t)$ as follows

$$\begin{aligned} \tilde x_n(t) &:= \mu C_{1,n}t + \frac{1}{2}\left(\mu^2+\sigma^2\right)C_{1,n}^2t^2 + \tilde\beta_n(t) \\ &= \frac{\mu t}{\sigma}n^{-1/2} + \left(\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\right)t^2n^{-1} + \tilde\beta_n(t) \\ &=: \frac{\mu t}{\sigma}n^{-1/2} + \delta_n(t), \end{aligned} \tag{7.3.15}$$

and noting (7.3.13) we see that this implies that for some $C_{4,M,t} > 0$, a sufficiently large $N_{M,t} \in \mathbb{N}$ and any $n \ge N_{M,t}$, we have

$$\begin{aligned} |\delta_n(t)| &\le C_{4,M,t}n^{-3/2}, \\ \text{and } |\tilde x_n(t)| &\le C_{4,M,t}n^{-1/2}. \end{aligned} \tag{7.3.16}$$

Thus, for $n \geq N_{M,t}$ we have

$$\begin{aligned} n\log(M(C_{1,n}t)) &= n\log(1+\tilde{x}_n(t)) \\ &= n\left(\tilde{x}_n(t) - \frac{1}{2}\tilde{x}_n(t) + \gamma(\tilde{x}_n(t))\right) \\ &= n\Bigg(\frac{\mu t}{\sigma}n^{-1/2} + \left(\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\right)t^2 n^{-1} + \tilde{\beta}_n(t) \\ &\qquad - \frac{1}{2}\left(\frac{\mu^2 t^2}{\sigma^2}n^{-1} + \delta_n(t)^2 + 2\frac{\mu t}{\sigma}n^{-1/2}\delta_n(t)\right) + \gamma(\tilde{x}_n(t))\Bigg) \\ &= \frac{\mu t}{\sigma}n^{1/2} + \frac{t^2}{2} + \underbrace{n\left(\tilde{\beta}_n(t) - \left(\frac{\delta_n(t)}{2}+\frac{\mu t}{\sigma}\right)\delta_n(t) + \gamma(\tilde{x}_n(t))\right)}_{=:\,\epsilon_n(t)}. \end{aligned} \tag{7.3.17}$$

The last term can be bounded as follows

$$\begin{aligned} \epsilon_n(t) &= n\left|\frac{\mu t}{\sigma}n^{-1/2} + \left(\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\right)t^2 n^{-1} + \tilde{\beta}_n(t)\right| \\ &\leq \left(\frac{C_{2,M}t^3}{\sigma^3} + \frac{C_{4,M,t}\mu t}{\sigma}\right)n^{-1/2}. \end{aligned} \tag{7.3.18}$$

From (7.3.8), (7.3.9), (7.3.17), (7.3.18), and the continuity of log and exp, for each fixed t,

$$M_n(t) = \exp\log M_n(t) = \exp\left(\frac{t^2}{2} + \epsilon_n(t)\right) \to \exp\frac{t^2}{2} \text{ as } n \to \infty. \tag{7.3.19}$$

□

7.3.3. Theorem (Central Limit-Theorem). *Under the conditions of Theorem 7.3.2, the distribution function of Z_n converges to the distribution function of a Standard Normal variable, as $n \to \infty$.*

Proof The proof of the Central Limit Theorem is similar in spirit to that of 7.3.2. It uses a complex-valued variant of the mgf, known as the characteristic function (or Fourier transform) of X. This is a more powerful, but slightly more technical, concept than the mgf and will not be covered in this course. □

7.3.4. Example. Let Z have density $f(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$ on the real line, i.e. Z has the *Standard Normal* distribution. Then its mgf is

$$M_Z(t) = \int \exp(tz)f(z)dz = \exp(t^2/2).$$

Exercises and problems on ergodicity

Exercise 7.1. You invest a fraction x of your capital each year in Bonds that safely return 6% interest, and the rest in a speculative venture that will lose everything with probability 20%, or else give you a 50% increase. What choice of x maximises the long-term average growth rate of your capital? What is that growth rate?

Exercise 7.2. (i) Suppose $V \geq 0$. By taking $X = \sqrt{V}$ and $Y = 1/\sqrt{V}$, use this last result to show that $E(1/V) \geq 1/E(V)$.
(ii) Suppose the buying price of shares on the i^{th} trading day is X_i, for $i = 1, 2, \ldots, N$, and let V be the random variable that takes each of these N values with equal probability. Use (i) to show that spending a fixed amount K on each of the N trading days will purchase more shares in total than spending amount KN once, buying shares at their mean price $\overline{X} = (X_1 + X_2 + \cdots + X_N)/N$.

Exercise 7.3. For a fixed $n \in \mathbb{N}$, consider n random variables, $X_1, \ldots, X_n$, that are independent and have the same mean μ an variance σ^2. Consider their sequence (sample) average

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{P7.3.1}$$

Show that

$$\mathrm{E}\left[\bar{X}_n\right] = \mu \text{ and } \mathrm{var}\left[\bar{X}_n\right] = \frac{\sigma^2}{n}. \tag{P7.3.2}$$

Exercise 7.4. The time to process a Library card application has mean one minute and standard deviation 20 seconds. Use the Central Limit Theorem to estimate the probability it takes over two hours to process 110 applications.

CHAPTER 8

Discrete stochastic processes

Stochastic is an often heard catch-word. A common misconception is that “stochastic” is synonymous with “chaotic” or “unpredictable”. In fact, stochastic processes are predictable, albeit in a probabilistic sense. The word *stochastic* comes from the (ancient) Greek στόχος, which means *to guess, to aim*, or *to target.* Roughly speaking *stochastic process* is thus a sequence (with discrete, or continuous parameter) that “guesses (or chooses)” at each step its next “state” with some random rule. Another name for stochastic process is *random process.*

8.1. Branching processes

8.1.1. Informal discussion. A *discrete time process* is a sequence of random variables, usually parametrised with respect to $\mathbb{N}$ or $\mathbb{N}_0$, although any countable set will work.

Consider a process $(X_k)_{k\in\mathbb{N}}$ modelling a given population’s count of certain individuals (say *cells*) parametrised with discrete time, modelling *generation.* Cells live exactly for one generation. Before dying each cell may give birth to child-cells: it will generate 1 cell with likelihood p_1, 2 cells with likelihood p_2, and so on, and no cells with likelihood p_0. This is given by a discrete probability distribution:

$$p_n \geq 0 \quad \forall\, n \in \mathbb{N}_0, \text{ and } \sum_{n=1}^{\infty} p_n = 1. \tag{8.1.1}$$

(Note that in “practical” situation, one may assume that $p_n = 0$ for $n \geq N$, for some given N, but we will leave the door open to the possibility of an unbounded number of children.) Suppose we start with $X_1 = 1$, i.e., 1 cell at time 1, we are interested in the evolution of this system over time.

8.1.2. Definition of branching process. A *branching process*, is a discretely indexed family of $\mathbb{N}_0$-valued random variables X_k, with index $k = 1, 2, \ldots$, such that

(i) It begins with one individual, $\mathcal{P}[X_0 = 1] = 1$.
(ii) The number of offspring, say X, of any member of the population has the same distribution, *independently* for all members. Write $\mathcal{P}[X = n] =: p_n$.
(iii) The trivial case $p_1 = 1$ is excluded.

Note that the trivial case would lead to $X_k = 1$ for all $k \geq 1$.

8.1.3. Definition of random counter. A *random counter* is a random variable that (almost surely) takes values in $\mathbb{N}_0$. Random counters form a subset (but not a subspace) of $\operatorname{DRV}(\mathcal{P})$ which is indicated by $\operatorname{RV}(\mathcal{P};\mathbb{N}_0)$. The number of offsprings of a given individual is a random counter.

8.1.4. Definition of probability generating function (also known as pgf). Let X be a random counter and denote

$$p_n := \mathcal{P}[X=n]. \tag{8.1.2}$$

Define its *probability generating function* (*pgf*)

$$\mathrm{pgf}_X(z) := \sum_{n=0}^{\infty} p_n z^n. \tag{8.1.3}$$

8.1.5. Example (geometrically branching process). Suppose X is geometric $G_0(p)$, i.e. $p_n = pq^n$ for $n \geq 0$. Then

$$\mathrm{pgf}_X(z) = \sum_{n=0}^{\infty} pq^n z^n = p\sum_{n=0}^{\infty}(qz)^n = p/(1-qz), \tag{8.1.4}$$

so long as $|qz| < 1$.

8.1.6. Lemma (independent sum and product of pgf). *Let* $\{X_1, \ldots, X_n\}$ *be an independent set of random counters then*

$$\mathrm{pgf}_{X_1+\cdots+X_n}(z) = \mathrm{pgf}_{X_1}(z)\cdots\mathrm{pgf}_{X_n}(z). \tag{8.1.5}$$

Proof The proof is very similar to that of the Convolution Theorem 3.3.5. Let us do first the case of two random variables (the general case follows readily by induction on n).

$$\mathrm{pgf}_{X+Y}(z) = \sum_{n=0}^{\infty} \mathcal{P}[X+Y=n]z^n \tag{8.1.6}$$

But

$$\{X+Y=n\} = \bigcup_{i=0}^{n}\{X=i\}\cap\{Y=n-i\} \tag{8.1.7}$$

and using disjoint additivity of $\mathcal{P}$ we get

$$\mathcal{P}[X+Y=n] = \sum_{i=0}^{n}\mathcal{P}[X=i]\mathcal{P}[Y=n-i]. \tag{8.1.8}$$

It follows that

$$\mathrm{pgf}_{X+Y}(z) = \sum_{n=0}^{\infty}\sum_{i=0}^{n} p_i q_{n-i} z^n \tag{8.1.9}$$

where

$$p_n := \mathcal{P}[X=n]\, q_n := \mathcal{P}[Y=n]. \tag{8.1.10}$$

Recalling the Cauchy product of two series—Graham, Knuth and Patashnik, 1994, eq. (7.20) or Wilf, 1994, eq. (2.1.2), and the Mertens Theorem (Apostol, 1974), we obtain

$$\mathrm{pgf}_{X+Y}(z) = \sum_{n=0}^{\infty} p_n z^n \sum_{n=0}^{\infty} q_n z^n = \mathrm{pgf}_X(z)\mathrm{pgf}_Y(z), \tag{8.1.11}$$

as desired. The general result follows by induction on n with $X = \sum_{i=1}^{n-1} X_i$ and $Y = X_n$.
□

8.1.7. Proposition (pgf of random sums of random counters). *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d.random variables, with*

$$g(z) = \mathrm{pgf}_{Y_n}(z) \quad \forall\, k \in \mathbb{N}. \tag{8.1.12}$$

Let N be a random variable such that $\{N, Y_1, Y_2, \ldots\}$ is independent, with $\mathrm{pgf}_N =: \phi$, *then*

$$\mathrm{pgf}_{Y_1+\cdots+Y_N}(z) = \mathrm{pgf}_N(g(z)). \tag{8.1.13}$$

Proof The proof follows ideas from the Discrete Convolution Theorem as well.

$$\begin{aligned}
\mathrm{pgf}_{Y_1+\cdots+Y_N}(z) &= \sum_{m=0}^{\infty} \mathcal{P}[Y_1+\cdots+Y_N = m]\, z^m \\
&= \sum_{m=0}^{\infty} \mathcal{P}\left(\bigcup_{n=0}^{\infty} \{N=n\} \cap \{Y_1+\cdots+Y_n = m\} \right) z^m \\
\text{(disjoint additivity and independence)} \quad &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \mathcal{P}[N=n]\mathcal{P}[Y_1+\cdots+Y_n=m]\, z^m \\
\text{(Fubini–Tonelli for series)} \quad &= \sum_{n=0}^{\infty} \mathcal{P}[N=n] \sum_{m=0}^{\infty} \mathcal{P}[Y_1+\cdots+Y_n=m]\, z^m \\
\text{(PGF's definition)} \quad &= \sum_{n=0}^{\infty} \mathcal{P}[N=n]\, \mathrm{pgf}_{Y_1+\cdots+Y_n}(z) \\
\text{(Lemma 8.1.6)} \quad &= \sum_{n=0}^{\infty} \mathcal{P}[N=n]\, g(z)^n \\
\text{(definition of } \mathrm{pgf}_N) \quad &= \mathrm{pgf}_N(g(z)).
\end{aligned} \tag{8.1.14}$$

□

8.1.8. Theorem (branching and PGF). *In a standard branching process $(X_k)_{k\in\mathbb{N}}$ let $g_k(z) := \mathrm{pgf}_{X_k}(z)$. Then $X_0 = 1$, so that $g_0(z) = z$, $g_1(z) = g(z)$ and, for $k \geq 1$,*

$$g_{k+1}(z) = g_{k-j}(g_{j+1}(z)) \quad \forall\, j = 0, \ldots, k. \tag{8.1.15}$$

In particular, the cases $j = 0, k-1$ imply

$$g_{k+1}(z) = g_k(g_1(z)) = g_1(g_k(z)). \tag{8.1.16}$$

Proof Note that $X_{k+1} = Y_1 + \cdots + Y_{X_k}$ where Y_i is the number of offsprings of the i-th individual in the population at time k. By Corollary 8.1.7 it follows that

$$g_{k+1}(z) = \mathrm{pgf}_{X_{k+1}}(z) = \mathrm{pgf}_{X_k}(g(z)) = g_k(g_1(z)). \tag{8.1.17}$$

Another way of viewing X_{k+1} is as $Z_1+\cdots+Z_{X_{k-1}}$ where Z_i is the number of descendants of the i-th individual of generation $k-1$, after 2 steps. By Corollary 8.1.7 we have $\mathrm{pgf}_{Z_i}(z) = \mathrm{pgf}_{X_2}(z) = g(g(z)) = g_2(z)$ and thus (Corollary 8.1.7 again)

$$g_{k+1}(z) = \mathrm{pgf}_{X_{k+1}}(z) = \mathrm{pgf}_{X_{k-1}}(\mathrm{pgf}_{X_2}(z)) = g_{k-1}(g_2(z)). \tag{8.1.18}$$

Proceeding by induction on j it follows that

$$g_{k+1}(z) = \mathrm{pgf}_{X_{k+1}}(z) = \mathrm{pgf}_{X_{k-j}}(\mathrm{pgf}_{X_{j+1}}(z)) = g_{k-j}(g_{j+1}(z)). \tag{8.1.19}$$

Since k is arbitrary, the result follows. □

8.1.9. Example. Suppose the offspring distribution is as in Example 8.1.5, i.e., $G_0(p)$, so that $g(z) = p/(1-qz)$. Plainly, $g_1(z) = g(z)$, so

$$g_2(z) = g_1(g_1(z)) = p\bigg/\left(1-q\frac{p}{1-qz}\right)$$

which simplifies. Indeed, following it up, we can show inductively that

$$g_k(z) = \frac{p\left((q^k-p^k)-qz(q^{k-1}-p^{k-1})\right)}{(q^{k+1}-p^{k+1})-qz(q^k-p^k)} \tag{8.1.20}$$

which, on expanding as a power series in z, lets us read off the coefficient of z^n, i.e. the probability that generation k has exactly n members.

8.1.10. Lemma (statistics via PGF). *If* $X \in \mathrm{RV}(\mathcal{P};\mathbb{N}_0)$ *and* $\mathrm{pgf}_X = g$ *then*

$$\mathcal{P}[X{=}0] = g(0); \tag{8.1.21}$$

$$\mathrm{E}[X] = g'(1); \tag{8.1.22}$$

$$\mathrm{var}\, X = g''(1) + g'(1) - g'(1)^2. \tag{8.1.23}$$

Proof The proof is left as an exercise. □

8.1.11. Theorem (statistics of a branching process). *Let* μ *and* σ^2 *denote the mean and variance of the number of offspring of an individual in a branching process* $(X_k)_{k\in\mathbb{N}}$. *Then*

$$\mathrm{E}[X_k] = \mu^k, and\, \mathrm{var}\, X_k = \begin{cases} k\sigma^2 & \textit{when } \mu = 1, \\ \frac{\sigma^2\mu^{k-1}(\mu^k-1)}{\mu-1} & \textit{otherwise.} \end{cases} \tag{8.1.24}$$

Proof The proof is a combination of Theorem 8.1.8 and Lemma 8.1.10 and is left as exercise. □

8.1.12. Corollary (ergodic behavior of a branching process). *Let* μ *and* σ^2 *denote the mean and variance of the number of offspring of an individual in a branching process* $(X_k)_{k\in\mathbb{N}}$.

$$\begin{array}{lll} \textit{if } \mu < 1, & \textit{then } \mathrm{E}[X_k] \to 0, & \textit{and } \mathrm{var}\, X_k \to 0. \\ \textit{if } \mu = 1, & \textit{then } \mathrm{E}[X_k] = 1\, \forall k \in \mathbb{N}_0, & \textit{and } \mathrm{var}\, X_k \to \infty. \\ \textit{if } \mu > 1, & \textit{then } \mathrm{E}[X_k] \to \infty, & \textit{and } \mathrm{var}\, X_k \to \infty. \end{array} \tag{8.1.25}$$

8.1.13. Theorem (extinction of a branching process). *Let* g *and* μ *respectively denote the PGF and mean of the number of offspring of an individual in a branching process* $(X_k)_{k\in\mathbb{N}}$. *Then* g *is positive, strictly monotone and convex function which has* 1 *as a fixed point, and possibly another one.*
Define $d_k := \mathcal{P}[X_k{=}0]$, *that is the probability that the branching process dies out at the* k*-th generation. Then* $d_0 = 0$, $d_{k+1} = g(d_k)$ *for all* $k \in \mathbb{N}_0$, $d_k \nearrow d$, *as* $k \to \infty$, *where* $d \leq 1$ *is the smallest positive fixed point of* g *(i.e., root of* $g(z) - z$*), and is the probability the process eventually dies out.*
If $\mu \leq 1$ *then* $d = 1$, *i.e., the process dies out surely in some finite time.*
If $\mu > 1$ *then* $d < 1$, *i.e., the process survives with some positive probability.*

Proof Since $X_0 = 1$ surely, then $d_0 = 0$. Look now at $d_1 = \mathcal{P}(X_1{=}0)$, this is the first coefficient of the power series of $g(z)$, and thus $g(0) = g(d_0)$. Similarly $d_2 = g_2(0) = g(g(0)) = g(d_1)$. By induction on k it follows that

$$d_{k+1} = g(d_k). \tag{8.1.26}$$

The behaviour of the sequence $(d_k)_{k\in\mathbb{N}_0}$, thus recursively defined by iterating g, can be understood from the behaviour of g and its fixed points. Recall that z is a fixed point of g if and only if

$$z = g(z). \tag{8.1.27}$$

Note that for $z \in \mathbb{R}_0^+$ all terms are nonnegative in the series

$$g'(z) = \sum_{n=1}^{\infty} p_n z^{n-1} \text{ and } g''(z) = \sum_{n=2}^{\infty} n(n-1) p_n z^{n-2}, \tag{8.1.28}$$

and thus $g' > 0$ (because $p_1 > 0$) and $g'' \geq 0$ on $\mathbb{R}_0^+$, making g strictly increasing and convex therein. Since $d_0 = 0$ and $d_1 > 0$ this and monotonicity of g inductively imply that

$$d_{k+1} > d_k. \tag{8.1.29}$$

Convexity of g means that g has at most 2 fixed points on $\mathbb{R}_0^+$ (think of the roots of $1 - g'(z)$). Since $g(1) = \sum_{n=0}^{\infty} p_n = 1$, 1 is a fixed point of g. All we have to find out now, is what happens for the other possible fixed point. The discussion now depends on $g'(1)$ which happens to equal μ.

(a) If $\mu < 1$ then the graph of g meets the diagonal at 1 with a slope smaller than 1, making 1 a stable fixed point, and another (unstable) fixed point occurs at $z_* > 1$ if $p_k > 0$ for some $k \geq 2$. In this case, $d = 1$ and we have $d_k \nearrow 1 = d$, as $k \to \infty$

(b) If $\mu = 1$ then the graph of g meets the diagonal at 1 tangentially and there are no other fixed points. (Note that this forces $p_2 > 0$.) Also in this case $d = 1$ and we have $d_k \nearrow 1 = d$, as $k \to \infty$.

(c) If $\mu > 1$ then the graph of g meets the diagonal at 1 with a slope larger than 1 and a stable fixed point must occur at $z_* < 1$. (Note that this forces $p_2 > 0$.) In this case, we have $d = z_*$ and the sequence d_k is bounded above by d and $\lim_{k\to\infty} d_k = d$.

The last two assertions in the Theorem's statement follow from this discussion as well.

□

8.1.14. Example. Suppose $g(z) = p/(1 - qz)$ (see Example 8.1.5). Then $d_1 = p$, $d_2 = p/(1 - pq)$, etc., and d satisfies $d = p/(1 - dx)$. This quadratic in d factors into $(qd - p)(d - 1) = 0$, so either $d = p/q$ or $d = 1$, whichever is smaller. If $q \leq p$, i.e., $p \geq 1/2$, then $d = 1$, i.e., extinction is certain; if $q > p$, i.e. $p < 1/2$, the chance of extinction in finite time is p/q.

8.1.15. Corollary. *Let $(X_k)_{k\in\mathbb{N}_0}$ be a branching process. Respectively denoting by*

$$d_k := \mathcal{P}[X_k{=}0] \text{ and } c_k := \mathcal{P}[X_k{=}0 \text{ and } X_{k-1} > 0], \tag{8.1.30}$$

the probability of extinction by generation k and the probability of becoming extinct in generation k, we have that

$$c_k = d_k - d_{k-1} \quad \forall\, k \geq \mathbb{N}. \tag{8.1.31}$$

Proof This is an immediate consequence of the definitions and the Total Probability Theorem 2.2.2

$$d_k = \mathcal{P}[X_k{=}0 \text{ and } X_{k-1} = 0] + \mathcal{P}[X_k{=}0 \text{ and } X_{k-1} > 0] = d_{k-1} + c_k, \tag{8.1.32}$$

where we have used the fact that $X_{k-1}{=}0$ implies $X_k{=}0$. □

8.2. Random walks

8.2.1. Definition. Given p, consider an i.i.d.sequence $(X_n)_{n\in\mathbb{N}}$ with

$$\mathcal{P}[X_n{=}1] = p \text{ and } \mathcal{P}[X_n{=}{-1}] = q := 1 - p \text{ for } n \in \mathbb{N}. \tag{8.2.1}$$

Write $S_0 = 0$, and $S_n = X_1 + X_2 \cdots + X_n$ for $n > 0$. Then $(X_n)_{n\in\mathbb{N}_0}$ is a *simple random walk*. It is *symmetric* if $p = q = 0.5$. The random variables X_n are called (random) *increments* or *steps* of the random walk.

8.2.2. Remark. The random variable $S_1 = X_1$ is a Bernoulli trial take the values ±1 only, in symbols

$$S_1 \in \mathrm{B}(\{-1,1\}, p). \tag{8.2.2}$$

Then $S_2 = X_1 + X_2$, being the sum of two independent Bernoulli trials, is a binomial with success rate p and values $\{-2,0,2\}$, in symbols,

$$S_2 \in \mathrm{B}(\{-2,0,2\}, p). \tag{8.2.3}$$

Similarly (by induction) it follows that

$$S_n \in \begin{cases} \mathrm{B}(\{-2m,\dots,-2,0,2,\dots,2m\}, p) & \text{when } n = 2m, \\ \mathrm{B}(\{-2m-1,\dots,-1,1,\dots,2m+1\}, p) & \text{when } n = 2m+1. \end{cases} \tag{8.2.4}$$

In particular, we have

$$\mathcal{P}[S_n{=}0] = \begin{cases} \binom{2m}{m} p^m q^m & \text{when } n \text{ is even, say } n = 2m \\ 0 & \text{when } n \text{ is odd} \end{cases} \tag{8.2.5}$$

8.2.3. Exercise. *Let* $(S_n)_{n\in\mathbb{N}_0}$ *be a symmetric random walk (i.e., with rate* $p = 1/2$*) calculate* $\psi_n := \mathcal{P}[S_n{=}0]$ *for* $n = 0,1,2,3,4,5,6$ *then deduce a recursive relation between* ψ_{n+2} *and* ψ_n *for all* $n \in \mathbb{N}_0$*.*

Solution. For n odd the result is trivial

$$\psi_1 = \psi_3 = \psi_5 = 0, \tag{8.2.6}$$

and more generally

$$\psi_{2m+1} = \mathcal{P}[S_{2m+1}{=}0] = 0 \quad \forall\, m \in \mathbb{N}_0. \tag{8.2.7}$$

More interestingly, let $n = 2m$ and introduce $\psi_n := \mathcal{P}[S_n{=}0] = 2^{-n}\binom{n}{n/2}$.

$$\begin{gathered} \psi_0 = 1, \quad \psi_2 = \frac{1}{2^2} \times \frac{2}{1\times 1} = \frac{1}{2}, \\ \psi_4 = \frac{3}{4} \times \frac{1}{2}, \quad \psi_6 = \frac{5}{6} \times \frac{3}{4} \times \frac{1}{2}, \dots, \end{gathered} \tag{8.2.8}$$

and more generally

$$\psi_{2(m+1)} = \frac{(2(m+1))!}{2^{2(m+1)}(m+1)!(m+1)!} = \frac{2(m+1)(2m+1)m!}{2^2(m+1)^2 2^m m!m!} \\ = \frac{(2m+1)m!}{2(m+1)2^m m!m!} = \frac{2m+1}{2(m+1)}\psi_{2m} \quad (8.2.9)$$

This can be written, including the odd indexes, also as

$$\psi_{n+2} = \frac{n+1}{n+2}\psi_n \quad \forall\, n \in \mathbb{N}_0. \qquad (8.2.10)$$

8.2.4. Return times and generating functions. A natural quesion to ask is whether it is certain that the walk will *return* to 0 at some time n for the first time after 0? And if yes, what is the expected time for this *first return* to occur?
To answer this let us look at the chance that the first return to 0 occurs at time $n \geq 1$, and write

$$\tau_n := \mathcal{P}[S_1 \neq 0 \wedge \cdots \wedge S_{n-1} \neq 0 \wedge S_n = 0]. \qquad (8.2.11)$$

We can think of modelling first return time as the random variable

$$\Omega \ni \omega \mapsto T(\omega) := \min\{n \in \mathbb{N} : S_n(\omega) = 0\}. \qquad (8.2.12)$$

(This is a random variable, but you might want to think about it for a minute.) The generating function of T is $\tau(z) = \sum_{n=1}^{\infty} \tau_n z^n$. Note that $\mathcal{P}[T=0] = 0$, so let's define $\tau_0 = 0$.

8.2.5. Remark (PGF or not?) The power series defining the function τ on its disc of convergence is the *probability generating function* of the random variable T. In Theorem 8.2.6 a related generating function $\varphi(z)$ will be used. However, $\varphi(z)$ is not a *probability* generating function, because $\varphi(1) \neq 1$ in general (a random walk can cross 0 many times). In other words there can be no random variable Y for which $\varphi = \mathrm{pgf}_Y$. Nevertheless φ is useful and, with appropriate care (i.e., checking that its radius of convergence is positive), the series defining φ can still be thought and manipulated as a PGF. When we think of an analytic function, such as φ, as the sum of a power series defined from a given sequence we talk about a *generating function* associated to that given sequence. This is a handy correspondence between discrete stuff, the sequence, and continuous stuff, the generating function. The science and art of generating functions, goes by the name “generatingfunctionology” (or “geefology”) an account of which is given in Wilf, 1994. A nice hands-on introduction to geefology can be found also in Graham, Knuth and Patashnik, 1994.

8.2.6. Lemma (return time PGF properties). *Consider the function*

$$\varphi(z) := \sum_{n=0}^{\infty} \varphi_n z^n \text{ where } \mathcal{P}[S_n = 0] =: \varphi_n, \text{ for each } n \in \mathbb{N}_0, \qquad (8.2.13)$$

models the probability of the random walk's passing (maybe not first) through 0 *at the time* n*. Then*

$$\tau(z)=\frac{\varphi(z)-1}{\varphi(z)}, \tag{8.2.14}$$

$$\varphi(z)=\frac{1}{\sqrt{1-4pqz^2}}, \tag{8.2.15}$$

and

$$\tau(z)=1-(1-4pqz^2)^{1/2}. \tag{8.2.16}$$

Proof Since $S_0=0$ surely, then $\varphi_0=1$. From (8.2.5) we know that $\varphi_{2m+1}=0$ for all $m\in\mathbb{N}_0$. Furthermore

$$\begin{aligned}\varphi_{2m}&=\mathcal{P}[S_{2m}=0]\\&=\mathcal{P}[S_{2m}=0\wedge S_{2m-2}\neq 0]+\mathcal{P}[S_{2m}=0\wedge S_{2m-2}=0]\\&=\mathcal{P}[S_{2m}=0\wedge S_{2m-2}\neq 0\wedge S_{2m-4}\neq 0]\\&\quad+\mathcal{P}[S_{2m}=0\wedge S_{2m-2}\neq 0\wedge S_{2m-4}=0]\\&\quad+\mathcal{P}[S_{2m}=0\wedge S_{2m-2}=0]\\&=\mathcal{P}[S_{2m}=0\wedge S_{2m-2}\neq 0\wedge\cdots\wedge S_2\neq 0\wedge S_0=0]\\&\quad+\mathcal{P}[S_{2m}=0\wedge S_{2m-2}\neq 0\wedge\cdots\wedge S_4\neq 0\wedge S_2=0]\\&\quad+\cdots\\&\quad+\mathcal{P}[S_{2m}=0\wedge S_{2m-2}\neq 0\wedge S_{2m-4}=0]\\&\quad+\mathcal{P}[S_{2m}=0\wedge S_{2m-2}=0]\\&=\tau_{2m}\varphi_0+\tau_{2m-2}\varphi_2+\cdots+\tau_4\varphi_{2m-4}+\tau_2\varphi_{2m-2}\end{aligned} \tag{8.2.17}$$

where we have used the fact that for any $j<m$, we have

$$\begin{aligned}&\mathcal{P}\big[S_{2m}=0\wedge S_{2m-2}\neq 0\wedge\cdots\wedge S_{2j+2}\neq 0\wedge S_{2j}=0\big]\\&\qquad=\mathcal{P}\big[S_{2m}=0\wedge S_{2m-2}\neq 0\wedge\cdots\wedge S_{2j+2}\neq 0|S_{2j}=0\big]\mathcal{P}\big[S_{2j}=0\big]\\&\qquad=\mathcal{P}\big[S_{2(m-j)}=0\wedge S_{2(m-j-1)}\neq 0\wedge\cdots\wedge S_2\neq 0|S_0=0\big]\mathcal{P}\big[S_{2j}=0\big]\\&\qquad=\tau_{2(m-j)}\varphi_{2j}.\end{aligned} \tag{8.2.18}$$

In summary we have, for $n>0$

$$\varphi_n=\sum_{i=0}^{n}\tau_n\varphi_{n-i} \tag{8.2.19}$$

and for $n=0$, simply $\varphi_0=1$. Thus, by the definition of Cauchy product, we have $\tau(z)\varphi(z)=\sum_{n=1}^{\infty}\varphi_n z^n$ and thus

$$1+\tau(z)\varphi(z)=\varphi_0+\sum_{n=1}^{\infty}\varphi_n z^n=\varphi(z). \tag{8.2.20}$$

Thus, noting that $\varphi(z)\neq 0$ for all z, relationship (8.2.14) is established.

To find an analytic expression for $\varphi(z)$ requires some geefology given by Lemma 8.2.7 (and the notation therein) by noting that (8.2.5) implies

$$\varphi(z) = \sum_{m=0}^{\infty} \binom{2m}{m} p^m q^m z^{2m} = \sum_{m=0}^{\infty} \frac{1}{4^m} \binom{2m}{m} (4pqz^2)^m$$
$$= \psi(2\sqrt{pq}z) = \frac{1}{1-4pqz^2} \quad (8.2.21)$$

and the analytic expression of φ (8.2.15) is established.
From (8.2.14) and (8.2.15) it follows that

$$\tau(z) = 1 - \frac{1}{\varphi(z)} = 1 - \sqrt{1-4pqz^2}, \quad (8.2.22)$$

which proves (8.2.16) and concludes the proof. □

8.2.7. Lemma (power series of arcsin's derivative). *Let*

$$\psi(w) := \sum_{n=0}^{\infty} \psi_n w^n \text{ where } \psi_n := \begin{cases} 2^{-2m}\binom{2m}{m} & \text{when } n = 2m \\ 0 & \text{when } n = 2m+1. \end{cases} \quad (8.2.23)$$

Then the series for $\psi(w)$ converges absolutely to the derivative of arcsin w

$$\psi(w) = \frac{1}{\sqrt{1-w^2}} \quad \forall\, w \in \mathbb{C} : |w| < 1. \quad (8.2.24)$$

Proof Denote $\sigma(w) := 1/\sqrt{1-w^2}$, we want to show that ψ converges absolutely and $\psi(w) = \sigma(w)$. It is instructive to simplify first the expression for ψ_n and find a recursive relation:

$$\begin{gathered} \psi_0 = 1, \psi_2 = \frac{1}{2}, \psi_4 = \frac{3}{4} \times \frac{1}{2}, \dots, \psi_{2(m+1)} = \frac{2m+1}{2(m+1)} \psi_{2m} \quad \forall\, m \in \mathbb{N}_0, \\ \text{and } \psi_{2m+1} = 0 \quad \forall\, m \in \mathbb{N}_0. \end{gathered} \quad (8.2.25)$$

It follows that the coefficients are bounded and hence the power series converges absolutely within the disk of radius 1, $\mathrm{B}_1^{\mathbb{C}}(0)$. Furthermore, noting that the function $\sigma(w)$ is analytic on $\mathrm{B}_1^{\mathbb{C}}(0)$, hence writable as the series $\sigma(w) = \sum_{n=0}^{\infty} \sigma_n w^n$, and the derivative of arcsin w, we have

$$\arcsin w = \sum_{n=0}^{\infty} \frac{\sigma_n}{n+1} w^{n+1} \,\forall\, w \in \mathrm{B}_1^{\mathbb{C}}(0), \quad (8.2.26)$$

and thus, for $|t| < \pi/2$ we have

$$t = \arcsin(\sin t) = \sum_{n=0}^{\infty} \frac{\sigma_n}{n+1} (\sin t)^{n+1}. \quad (8.2.27)$$

Differentiating once

$$0 = \sum_{n=0}^{\infty} \sigma_n (\sin t)^n \cos t \quad \forall\, t \in (-\pi/2, \pi/2), \quad (8.2.28)$$

and differentiating once more (treating the first coefficient carefully)

$$\begin{aligned}
0 &= -\sigma_0 \sin t + \sum_{n=1}^{\infty} \sigma_n \left(n(\sin t)^{n-1} (\cos t)^2 - (\sin t)^{n+1} \right) \\
&= -\sigma_0 \sin t + \sum_{n=1}^{\infty} \sigma_n \left(n(\sin t)^{n-1} \left(1 - (\sin t)^2\right) - (\sin t)^{n+1} \right) \\
&= -\sigma_0 \sin t + \sum_{n=1}^{\infty} \sigma_n \left(n(\sin t)^{n-1} - (n+1)(\sin t)^{n+1} \right) \\
&= \sum_{n=1}^{\infty} n\sigma_n (\sin t)^{n-1} - \sum_{n=0}^{\infty} (n+1)\sigma_n (\sin t)^{n+1} \\
&= \sum_{n=0}^{\infty} (n+1)\sigma_{n+1} (\sin t)^n - \sum_{n=1}^{\infty} n\sigma_{n-1} (\sin t)^n \\
&= \sigma_1 + \sum_{n=1}^{\infty} ((n+1)\sigma_{n+1} - n\sigma_{n-1})(\sin t)^n .
\end{aligned} \tag{8.2.29}$$

By the identity principle of power series, and the fact that sin is one-to-one from $(-\pi/2, \pi/2)$ to $(-1, 1)$ it follows that $\sigma_1 = 0$ and

$$\sigma_{n+1} = \frac{n}{n+1} \sigma_{n-1} \quad \forall\, n \geq 1. \tag{8.2.30}$$

By recursion we obtain

$$\begin{aligned}
\sigma_{2m+1} &= \psi_{2m+1} = 0 \quad \forall\, m \geq 0, \text{ and} \\
\sigma_{2m} &= \frac{2m-1}{2m} \sigma_{2m-2} \quad \forall\, m \geq 1,
\end{aligned} \tag{8.2.31}$$

which is the same recursion that gives ψ_n. Noting that the seeds coincide as well, $\sigma_0 = \sigma(0) = 1 = \psi_0$, it follows that

$$\sigma_n = \psi_n \quad \forall\, n \in \mathbb{N}_0. \tag{8.2.32}$$

This means that $\psi(w) = \sigma(w)$ as wanted. □

8.2.8. Theorem (random walk first return time). *Let $(S_n)_{n \in \mathbb{N}_0}$ be a random walk with rate p and initial state S_0, and denote by T the time of first return to* 0*. Then the event that S_n equals zero for some n equals the event that T is finite (written $T < \infty$). Furthermore*

(a) If the random walk is symmetric, i.e., $p = 1/2$, then

$$\mathcal{P}[T < \infty] = 1 \textit{ but } \mathrm{E}[T] = \infty. \tag{8.2.33}$$

(b) If the random walk is asymmetric, i.e., $p \neq 1/2$, then

$$\mathcal{P}[T < \infty] = 1 - \left| p - q \right| = 1 - \left| 2p - 1 \right| < 1 \textit{ and } \mathrm{E}[T] = \infty. \tag{8.2.34}$$

The intuition behind these results is that for a symmetric random walk it is reasonable to expect an eventual first return to the origin with high probability, whereas this is to be less and less likely as the parameter p is moved away from 1/2 creating a *drift* effect for S_n, towards $+\infty$ if $p > 1/2$ or $-\infty$ if $p < 1/2$. On the other hand, and somewhat less intuitive, is the fact that the expected time for first return to the origine is infinite, especially for a symmetric random walk. This is an illustration of the fact that mathematical analysis will usually yield more insight in probability than plain “common sense”.

Proof Reviewing the definition of T in (8.2.12). Let $\omega \in \Omega$ and suppose that for some $n \geq 1$, we have $S_n(\omega) = 0$, then $T(\omega) \leq n$, whence $T(\omega) < \infty$. Conversely, suppose $T(\omega) < \infty$ then $S_{T(\omega)}(\omega)=0$ hence

$$T(\omega) < \infty \Leftrightarrow \exists\, n \geq 1 : S_n(\omega)=0. \tag{8.2.35}$$

To proceed note that

$$\mathcal{P}[T < \infty] = \mathcal{P}\left(\bigcup_{n=1}^{\infty} \{T=n\}\right) = \sum_{n=1}^{\infty} \mathcal{P}[T=n] = \tau(1) = 1 - \sqrt{1-4pq}, \tag{8.2.36}$$

and noting that

$$1-4pq = (p+q)^2 - 4pq = (p-q)^2 \tag{8.2.37}$$

we conclude

$$\mathcal{P}[T < \infty] = 1 - \left|p-q\right|. \tag{8.2.38}$$

If $p \neq 1/2$ then $\mathcal{P}[T=\infty] > 0$, so $\mathrm{E}[T] = \infty$ and the case $p \neq 1/2$ is thus proved. If $p = 1/2$ then using the fact that $\tau = \mathrm{pgf}_T$ we have

$$\mathrm{E}[T] = \tau'(1) = \left[\frac{\mathrm{d}}{\mathrm{d}z}\left[1-\sqrt{1-z^2}\right]\right]_{z=1} = \left[\frac{z}{\sqrt{1-z^2}}\right]_{z=1} = \infty. \tag{8.2.39}$$

This concludes the proof of the case $p = 1/2$. □

8.2.9. Successive returns to the origin. Let $(S_n)_{n\in\mathbb{N}_0}$ a random walk with rate p. Now that we know the probability of returning to the origin once is $1-\left|2p-1\right|$, we may ask what is that of hitting it twice in a row. Write $T_1 := T$ and introduce a random variable that measures the instant of *second return*

$$T_2 := \min\left\{n \in \mathbb{N} : (S_n=0) \text{ and for one and only one } n_1 < n \text{ also } S_{n_1}=0\right\} \tag{8.2.40}$$

By the total probability theorem we have

$$\begin{aligned} \mathcal{P}(T_2=l) &= \sum_{n=1}^{l-1} \mathcal{P}[T_2=l|T=n]\mathcal{P}[T=n] \\ \text{(independence of steps)} \quad &= \sum_{n=1}^{l-1} \mathcal{P}[T=l-n]\mathcal{P}[T=n] \\ \text{(using } \mathcal{P}[T=0]=0) \quad &= \sum_{n=0}^{l} \mathcal{P}[T=l-n]\mathcal{P}[T=n] \\ &= \sum_{n=0}^{l} \tau_{l-n}\tau_n, \end{aligned} \tag{8.2.41}$$

which implies, by Cauchy product and Mertens, that

$$\mathrm{pgf}_{T_2}(z) = \tau(z)^2. \tag{8.2.42}$$

Similarly, by induction we can show that if T_m denotes the time of m-th return we have the relation

$$\mathrm{pgf}_{T_m} = \tau^m. \tag{8.2.43}$$

If follows that if the walk is symmetric the probability of having 2, 3, or any number $m \geq 4$ of returns to the origin is $1^m = 1$. Whereas for an asymmetric random walk, this probability is $\left(1-\left|2p-1\right|\right)^m$ decreases geometrically as $m \to \infty$.

8.2.10. Theorem (infinite returns to the origin). *Let E be the event that the walk returns to the origin infinitely often. Then*

$$\mathcal{P}(E) = \begin{cases} 1 & \textit{if } p = 1/2, \\ 0 & \textit{if } p \neq 1/2. \end{cases} \tag{8.2.44}$$

Proof For $m \in \mathbb{N}_0$, define E_m to be the event where the random walk returns to the origin at least m times; in symbols

$$E_m := \{\omega \in \Omega : \#\{n \in \mathbb{N} : S_n(\omega)=0\} \geq m\}. \tag{8.2.45}$$

Therefore we can write

$$E = \bigcap_{n\in\mathbb{N}_0} E_n. \tag{8.2.46}$$

Also, denoting by T_m the time of m-th return to the origin we have

$$\{T_m < \infty\} = E_m. \tag{8.2.47}$$

From 8.2.9, we deduce that

$$\mathcal{P}(E_m) = \mathcal{P}[T_m < \infty] = \tau(1)^m = \left(1-\left|2p-1\right|\right)^m. \tag{8.2.48}$$

If $p = 1/2$, by De Morgan's laws and subadditivity we have,

$$\mathcal{P}(E^{\mathrm{c}}) = \mathcal{P}\left(\bigcup_{n\in\mathbb{N}_0} {E_n}^{\mathrm{c}}\right) \leq \sum_{n\in\mathbb{N}_0} (1-\mathcal{P}(E_n)) = 0, \tag{8.2.49}$$

from which we obtain $\mathcal{P}(E) = 1$, as claimed.
If $p \neq 1/2$, then using (8.2.48) and the geometric series, we have

$$\sum_{m=1}^{\infty} \mathcal{P}(E_m) = \sum_{m=1}^{\infty} \left(1-\left|2p-1\right|\right)^m = \frac{1-\left|2p-1\right|}{\left|2p-1\right|} < \infty. \tag{8.2.50}$$

By the first Borel–Cantelli Lemma 1.1.9 we obtain that

$$\mathcal{P}\left(\limsup_{m\to\infty} E_m\right) = 0, \tag{8.2.51}$$

and upon noting that $E = \limsup_{m\to\infty} E_m$ we conclude, $\mathcal{P}(E) = 0$ as claimed. □

8.3. Gambler's ruin

8.3.1. A simple gambling game. A gambler enters a game with two outcomes: win one (+1) with chance $p \in (0,1)$ or lose one (−1) with chance $q := 1 = p$. The game is played repeatedly at each time $n \in \mathbb{N}$. Denote by S_n the gambler's fortune at time $n \in \mathbb{N}_0$. Assume the gambler has an initial fortune $a \in \mathbb{N}$ ($S_0 = a$) and will stop playing if either ruined ($S_n = 0$ for some n) or wins a given target $c - a \in \mathbb{N}$ ($S_n = c$ for some n).
The *Gambler's ruin problem* modifies the simple random walk in two ways. First, the start point is not 0, but some positive integer a (the initial fortune). Second, the walk stops as soon as either $S_n = 0$ (the gambler is ruined), or $S_n = c$ where $c > a$ (the

gambler reaches his goal of winning a net amount $c-a$); stopping is modelled by defining $S_{n+1}(\omega)=0$, or c, if $S_n(\omega)=0$, or c, respectively.

8.3.2. Lemma (Gambler's fortune evolution). *A gambler with initial fortune $a\in\mathbb{N}$, seeks a target win of $c-a$, for $c>a$. For $s=0,\dots,c$ write r_s as the probability the gambler will be ruined if their current fortune is amount s. Then $r_0=1$, $r_c=0$ and, for $0<s<c$,*

$$r_s = q\,r_{s-1}+p\,r_{s+1}. \tag{8.3.1}$$

Proof Define the random variable $T_{a,s}$ as the first time the fortune equals $s\in\mathbb{N}_0$, i.e.,

$$T_{a,s}(\omega):=\min\{n\in\mathbb{N}_0:\ S_n(\omega)=s\}\ \text{for}\ \omega\in\Omega, \tag{8.3.2}$$

with the convention that $\min\varnothing=\infty$. Then $T_{a,a}=0$ surely, $T_{a,0}$ is the ruin time, $T_{a,c}$ is the win time. Both $T_{a,0}$ and $T_{c,a}$ are *exit times* or *stopping times*, in the sense that after T_0 the walk stays at 0 or c respectively. The event $\left\{T_{a,0}<T_{a,c}\right\}$ equals the event that the gambler is ruined.

Suppose now that for some $s=1,\dots,c-1$ we have $S_0=s$ surely, then the probability that the gambler is ruined is

$$\begin{aligned}
r_s :&= \mathcal{P}\left[T_{a,0}<T_{a,c}|S_0=s\right]\\
\text{(total probability theorem 2.2.2)}\quad &= \mathcal{P}\left[T_{a,0}<T_{a,c}|S_0=s\wedge S_1=s+1\right]\mathcal{P}[S_1=s+1|S_0=s]\\
&\quad+\mathcal{P}\left[T_{a,0}<T_{a,c}|S_0=s\wedge S_1=s-1\right]\mathcal{P}[S_1=s-1|S_0=s]\\
&= \mathcal{P}\left[T_{a,0}<T_{a,c}|S_1=s+1\right]\mathcal{P}[S_1=s+1|S_0=s]\\
&\quad+\mathcal{P}\left[T_{a,0}<T_{a,c}|S_1=s-1\right]\mathcal{P}[S_1=s-1|S_0=s]\\
&= r_{s+1}p+r_{s-1}q.
\end{aligned} \tag{8.3.3}$$

In the penultimate step we have used the fact that a random walk has "no memory" in the sense that for any $m\in\mathbb{N}_0$, the position after time m depends on the position at time m but not on the at previous times $0,\dots,m-1$, i.e.,

$$\begin{aligned}
&\mathcal{P}\left[T_{a,0}<T_{a,c}|S_m=s\wedge S_{m-1}=s_1\wedge\cdots\wedge S_0=s_m\right]\\
&\qquad\qquad=\mathcal{P}\left[T_{a,0}<T_{a,c}|S_m=s\right]\quad\forall\,(s_1,\dots,s_m)\in\mathbb{N}^m.
\end{aligned} \tag{8.3.4}$$

□

8.3.3. Theorem (Gambler's ruin). *Write $\theta=q/p$. Then,*

$$r_a=\begin{cases}(\theta^c-\theta^a)/(\theta^c-1) & \text{if } p\neq 1/2,\\ 1-a/c & \text{if } p=1/2.\end{cases} \tag{8.3.5}$$

Furthermore, denoting by T_a the time (i.e., number of rounds) that a game starting with fortune a lasts, we have

$$\mathrm{E}[T_a]=\begin{cases}a(c-a) & \text{when } p=1/2,\\ (c-a-c\,r_a)/(p-q) & \text{when } p\neq 1/2.\end{cases} \tag{8.3.6}$$

Proof From Lemma 8.3.2 we know that the sequence $(r_s)_{s\in\{s\ldots 0\}c}$ solves the second order linear difference equation

$$q\,r_s-r_{s+1}+p\,r_{s+2}=0\quad\forall\,s\in\mathbb{N}_0. \tag{8.3.7}$$

Let us apply Lemma 8.3.5 to find out an explicit formula for r_s. The characteristic polynomial is $\pi(x) = q - x + px^2$ whose roots are

$$\xi_{1,2} = \frac{1 \pm \sqrt{1-4qp}}{2p} = \frac{1 \pm |2p-1|}{2p} = \begin{cases} 2p/(2p) & = 1, \text{ or} \\ (2-2p)/(2p) = q/p & = \theta. \end{cases} \tag{8.3.8}$$

If $p \neq 1/2$, then $\xi_1 = 1 \neq \theta = \xi_2$ and two distinct particular solutions are given by

$$r_s^1 = 1 \text{ and } r_s^2 = \theta^s \tag{8.3.9}$$

so that the general solution is given by

$$r_s = \alpha_1 + \alpha_2 \theta^s. \tag{8.3.10}$$

Recalling the "boundary conditions" $r_0 = 1$ and $r_c = 0$ we obtain a 2×2 system for $\alpha_{1,2}$:

$$\alpha_1 + \alpha_2 = 1, \text{ and } \alpha_1 + \theta^c \alpha_2 = 0, \tag{8.3.11}$$

whence $\alpha_2 = 1/(1-\theta^c)$, $\alpha_1 = 1 - \alpha_2 = \theta^c/(\theta^c - 1)$ and

$$r_s = \frac{\theta^c - \theta^s}{\theta^c - 1}. \tag{8.3.12}$$

Taking $s = a$ we obtain the first case in (8.3.5).
If $p = q = 1/2$ then π has a single root, $\xi_{1,2} = 1$ of multiplicity 2, and the particular solutions are given by

$$r_s^1 = 1 \text{ and } r_s^2 = s \tag{8.3.13}$$

which, proceeding similarly to above, implies the general solution

$$r_s = 1 - \frac{s}{c}, \tag{8.3.14}$$

and thus $r_a = 1 - a/c$.
The game ends when the gambler either loses ($S_n = 0$) or wins ($S_n = c$), therefore the game's stopping time is $T_a = T_{a,0} \wedge T_{a,c}$. If $a = 0$ or $a = c$ then $T_a = 0$ certainly and $t_0 := \mathrm{E}[T_0] = \mathrm{E}[T_c] =: t_c = 0$. Let us calculate $\mathrm{E}[T_a] =: t_a$ for $a = 1, \dots, c_1$. In this case, one round at least must be played after which we have

$$S_1 = \begin{cases} a+1 & \text{with chance } p, \\ a-1 & \text{with chance } q. \end{cases} \tag{8.3.15}$$

After that the game will last $\tilde{T}_{a+1}$ or $\tilde{T}_{a-1}$ respectively, where $\tilde{T}_k$ is the equivalent of T_k but starting at time 1 instead of 0. By independence T_k and $\tilde{T}_k$ are identically distributed and hence $t_k = \mathrm{E}[T_k] = \mathrm{E}[\tilde{T}_k]$, for each $k = 0, \dots, c$. An exercise shows that

$$\mathrm{E}\left[T_a | \mathscr{F}'\right] = 1 + p\tilde{T}_{a+1} + q\tilde{T}_{a-1}, \tag{8.3.16}$$

where $\mathscr{F}'$ is the subalgebra of $\mathscr{F}$ given by quotienting it with respect to the first round. It follows that

$$t_a = \mathrm{E}\left[\mathrm{E}\left[T_a | \mathscr{F}'\right]\right] = 1 + p\,t_{a+1} + q\,t_{a-1} \tag{8.3.17}$$

which is an affine relationship for the vector $(t_0,\dots,t_c)$ with $t_0=0=t_c$. In other words $(t_1,\dots,t_{c-1})$ solves the linear system

$$\begin{bmatrix} -1 & q & & & \\ p & -1 & q & & \\ & \ddots & \ddots & \ddots & \\ & & p & -1 & q \\ & & & p & -1 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_{c-2} \\ t_{c-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}. \tag{8.3.18}$$

Relationship (8.3.6) follows by applying 8.3.6 and is left as an exercise. □

8.3.4. Exercise. *Justify the recursion (8.3.16) rigorously by introducing the probability space $\Omega=\Omega_1\times\Omega'$, where $\Omega_1=\{\pm1\}$ and $\Omega'=\Omega=\Omega_1^{\mathbb{N}}$ with the measure $\mathcal{P}_1(\omega)=p$ or q if $\omega=1$ or $\omega=-1$ respectively and then averaging in the first round ω_1 of $\omega=(\omega_1,\omega')\in\Omega$.*

8.3.5. Lemma (linear difference equations). *Suppose $\pi(x)=p_0+p_1x+\dots+p_dx^d$ is a polynomial of degree d. Denote by ξ_i one of π's k (complex) roots, with multiplicity $\mu_i\in\mathbb{N}$. Then for each $m=1,\dots,\mu$, defining*

$$r_s^{i,m} := \frac{s!}{(s-m+1)!}(\xi_i)^s \quad \textit{for each } s\in\mathbb{N}_0, \tag{8.3.19}$$

we have that

$$p_0 r_s^{i,m} + p_1 r_{s+1}^{i,m} + \dots + p_d r_{s+d}^{i,m} = 0 \quad \forall\, s\in\mathbb{N}_0. \tag{8.3.20}$$

Conversely, a solution $(r_s)_{s\in\mathbb{N}}$ of the linear difference equation

$$p_0 r_s + p_1 r_{s+1} + \dots + p_d r_{s+d} = 0, \tag{8.3.21}$$

is a linear combination of the $\left(r^{i,m}{}_s\right)_{s\in\mathbb{N}}$ as $i=1,\dots,I$ and $m=1,\dots,\mu$.
Proof The proof is left as an exercise.
Hint. A root ξ of polynomial π, of multiplicity μ, satisfies

$$\left[\frac{\mathrm{d}^{m-1}}{\mathrm{d}x^{m-1}}\pi(x)\right]_{x=\xi} = 0 \quad \forall\, m=1,\dots,\mu. \tag{8.3.22}$$

□

8.3.6. Theorem (inversion of tridiagonal Toeplitz matrix by Yamamoto and Ikebe, 1979). *Suppose $p,q\in\mathbb{R}^+$ with $p+q=1$ and let $\theta=q/p$. The inverse $\boldsymbol{B}=\left[b_i^j\right]_{i=1,\dots,n}^{j=1,\dots,n}$ of the $n\times n$ tridiagonal Toeplitz matrix*

$$\boldsymbol{A} = \begin{bmatrix} -1 & q & & & \\ p & -1 & q & & \\ & \ddots & \ddots & \ddots & \\ & & p & -1 & q \\ & & & p & -1 \end{bmatrix} \tag{8.3.23}$$

has entries

$$b_i^j = -\begin{cases} \frac{p^{j-1}\left(1-\theta^i\right)\left(\theta^j-\theta^{n+1}\right)}{q^j(1-\theta)(1-\theta^{n+1})} & \textit{for } i\le j \\ \frac{p^{j-1}\left(1-\theta^j\right)\left(\theta^i-\theta^{n+1}\right)}{q^j(1-\theta)(1-\theta^{n+1})} & \textit{for } i>j \end{cases} \tag{8.3.24}$$

when $\theta \neq 1$ or

$$b_i^j = -\begin{cases} \frac{2i(n+1-j)}{n+1} \textit{ for } & i \le j \\ \frac{2j(n+1-i)}{n+1} & \textit{for } i > j. \end{cases} \tag{8.3.25}$$

Proof The proof, omitted, can be found in Yamamoto and Ikebe, 1979 as a special case of Theorem 2; see also Remark 2. □

8.4. Markov chains

8.4.1. Definition of Markov chain. Let S be a countable set (think of $S = \{1 \dots s\}$, for some $s \in \mathbb{N}$, or $S = \mathbb{N}$ or $S = \mathbb{Z}$). Suppose a given system can take a *state* in S, the *state space*. Let $(X_n)_{n\in\mathbb{N}_0}$ be a sequence of random variables, taking values in S. Since in many Markov chain models the subindex n in X_n, denotes a given time and X_n the state of the system at n-th time instant, so we refer to it as an *instant*.
We say that $(X_n)_{n\in\mathbb{N}_0}$ is a *Markov Chain* if the following *Markov property* is satisfied: for any choice of $m \le n \in \mathbb{N}$, (possibly repeating) states $s_1, \dots, s_m \in S$, and strictly increasing instants $n_1 < \dots < n_m \in \mathbb{N}_0$, we have that

$$\mathcal{P}\left[X_{n_m} = s_m | X_{n_1} = s_1 \wedge \cdots \wedge X_{n_{m-1}} = s_{m-1}\right] = \mathcal{P}\left[X_{n_m} = s_m | X_{n_{m-1}} = s_{m-1}\right]. \tag{8.4.1}$$

A Markov chain $(X_n)_{n\in\mathbb{N}_0}$ is *time homogeneous* if

$$\mathcal{P}\left[X_{m+n} = j | X_n = i\right] =: p_{ij}^{(n)} \tag{8.4.2}$$

depends on i, j and n only, and not on m. Our chains will always be time-homogeneous. Unless otherwise stated a Markov chain, in this text, is always assumed to be time-homogeneous. Let $i, j \in S$ be two states for a Markov chain $(X_n)_{n\in\mathbb{N}_0}$, we define the *(one-step) transition probability*

$$p_{ij} := p_{ij}^{(1)} \quad \forall\, i, j \in S. \tag{8.4.3}$$

If S is finite *transition matrix* of the Markov chain $(X_n)_{n\in\mathbb{N}_0}$ is defined as

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & \cdots & p_{1s} \\ \vdots & \ddots & \vdots \\ p_{s1} & \cdots & p_{ss} \end{bmatrix} \tag{8.4.4}$$

If S is infinite, say $S = \mathbb{N}$ (or $S = \mathbb{Z}$), we can talk about an infinite (or four-sided infinite) matrix.

8.4.2. Exercise. *Let $\boldsymbol{P} = [p_{ij}]_{i,j\in S}$, show that each rown sums up to 1, i.e., $\sum_{j\in S} p_{ij} = 1$ for all $s \in S$, and all the entries are non-negative, i.e., $p_{ij} \ge 0$.*

8.4.3. Example. The sizes of successive generations in a branching process. Random walks.

8.4.4. Example (the original Markov chain). Markov's original example did this with Pushkin's tone poem, "Eugene Onegin". Let v denote a vowel, and c a consonant, in a piece of prose. Replace actual vowels and consonants by these symbols. Thus[1] The one-to-one transcription from Cyrillic to extended Latin) as follows:

> Latyn' iz mody vyšla nyne:
> Tak, esli pravdu vam skazat',
> On znal dovol'no po-latyne,
> Čtob èpigrafy razbirat',

would become

> cvcvc vc cvcv cvccv cvcv
> cvc vccv ccvccv cvc ccvcvc
> vc ccvc cvcvccv cvcvcvcv
> ccvc vcvccvcv cvccvcvc

8.4.6. Example. With states $\{1,2,3\}$, let the transition matrix be

$$\boldsymbol{P} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 2/3 & 0 & 1/3 \end{bmatrix}. \tag{8.4.5}$$

To interpret it, look along each row in turn: if now in state 1, we are equally likely to be in any one of the three states next time; if now in 2, we move to 1 or 3 with equal probabilities; and if now in 3, we remain there with probability 1/3, otherwise move to 1.

8.5. Markov chain's algebra

8.5.1. Theorem (Chapman–Kolmogorov). *Let $\boldsymbol{P}$ be a Markov chain's transition matrix, and let $\boldsymbol{P}^{(n)}$ be the transition matrix from state* 0 *to state n, then*

$$p_{ij}^{(m+n)} = \sum_{k=1}^{s} p_{ik}^{(m)} p_{kj}^{(n)}. \tag{8.5.1}$$

8.5.2. Corollary. *Let $\boldsymbol{P}$ be a Markov chain's transition matrix, $m \in \mathbb{N}$, and let $\boldsymbol{P}^{(m)}$ be the transition matrix from state* 0 *to state m, then*

$$\boldsymbol{P}^{(m)} = \boldsymbol{P}^{m}, \tag{8.5.2}$$

where $\boldsymbol{P}^{m}$ denotes the m-th algebraic power of the matrix $\boldsymbol{P}$, i.e.,

$$\boldsymbol{P}^{m} = \underbrace{\boldsymbol{P} \cdots \boldsymbol{P}}_{m \text{ times}} \tag{8.5.3}$$

[1] According to Charles H. Johnston "Few foreign masterpieces can have suffered more than Eugene Onegin from the English translator's failure to convey anything more than – at best – the literal meaning." So we stick to the Russian original example, due to Andrey Markov (1907), although passing the text into c's and v's may still suffer from same lost-in-translation effect decried by Johnston's (whom most likely Markov did not read). If you don't know the Russian alphabet, a good online transliterator is Podolak, 2012. (Unfortunately it does not have a "translate to c-v" option.)

8.5.3. Corollary (Chapman–Kolmogorov and transition). *Given a Markov chain* $(X_n)_{n\in\mathbb{N}_0}$ *with transition matrix* $\boldsymbol{P}$*, let* $\boldsymbol{p}^{(n)}=(p_i^{(n)})_{i\in S}$ *be the distribution of the state of the chain at time* n*, i.e.* $p_i^{(n)}=\mathcal{P}[X_n=i]$*. Then* $\boldsymbol{p}^{(n)}=\boldsymbol{p}^{(0)}\boldsymbol{P}^n$*.*

8.5.4. Definition. Write $\boldsymbol{w}=(w_1,w_2,\dots)$; $\boldsymbol{w}$ is a *stationary distribution* for the chain if each $w_i\geq 0$, $\sum_i w_i=1$, and $\boldsymbol{w}\,\boldsymbol{P}=\boldsymbol{w}$. In other words, $\boldsymbol{w}$ is a left eigenvector of $\boldsymbol{P}$ associated to the eigenvalue 1.

8.5.5. Interpretation. When $\boldsymbol{w}$ is a stationary distribution, suppose $\boldsymbol{w}^{(n)}=\boldsymbol{w}$. Then, by Corollary 8.5.3 $\boldsymbol{p}^{(n+1)}=\boldsymbol{p}^{(n)}\boldsymbol{P}=\boldsymbol{p}^{(n)}$ by definition of stationarity. Thus, if *ever* the distribution of states is given by $\mathrm{p}^{(n)}$, the values of X_n continue to change, but their *distribution* never changes.

8.5.6. Example. Markov's original example had transition matrix

$$\begin{array}{rcc} \text{Vowel} & 0.128 & 0.872 \\ \text{Consonant} & 0.663 & 0.337 \\ \text{followed by}\rightarrow & \text{Vowel} & \text{Consonant} \end{array} =\boldsymbol{P}. \tag{8.5.4}$$

The successive powers $\boldsymbol{P}^2,\boldsymbol{P}^3,\boldsymbol{P}^4$ are

$$\begin{bmatrix}0.595 & 0.405\\ 0.308 & 0.692\end{bmatrix},\qquad \begin{bmatrix}0.345 & 0.655\\ 0.498 & 0.502\end{bmatrix},\qquad \begin{bmatrix}0.478 & 0.522\\ 0.397 & 0.603\end{bmatrix} \tag{8.5.5}$$

so that, for example,

$$\mathcal{P}(X_4=\text{Vowel}|X_0=\text{Consonant})=p_{21}^{(4)}=0.397. \tag{8.5.6}$$

The sequence (P^n) converges:

$$\lim_{n\to\infty}P^n=\begin{bmatrix}0.432 & 0.568\\ 0.432 & 0.568\end{bmatrix}, \tag{8.5.7}$$

so, whatever the initial state, the long-run chance a letter is a vowel is 0.432. You should now verify that $(0.432, 0.568)$ is a stationary distribution for P.

8.6. Reducible and irreducible Markov chains

8.6.1. Definition of leading-to and communicating states. If there is some $a\geq 0$ with $p_{ij}^{(a)}>0$, we say that i *leads to* j (or that j is a *reachable state* from i), and write $i\rightsquigarrow j$. If $i\rightsquigarrow j$ and $j\rightsquigarrow i$, we write $i\leftrightsquigarrow j$, and say that i and j *communicate.* We also take $i\leftrightsquigarrow i$ for each state i.

8.6.2. Theorem (splitting in communication classes). *The relation* $\leftrightsquigarrow$ *is an equivalence relation, splitting* S *into disjoint equivalence classes.*

Proof Recall an *equivalence relation* is a relation (or graph) which is *reflexive, symmetric* and *transititive.* The relation $\leftrightsquigarrow$ is reflexive and symmetric by construction. It
[∗]: Check! is transitive because $\rightsquigarrow$ is transitive.[∗]
As well known from elementary texts (Lakkis, 2011, Ch.7, e.g.), an equivalence relation splits S into a partition formed of equivalence classes, which we call in our particular set-up *communication classes.* □

8.6.3. Definition. If there is just one communication class, the chain is *irreducible,* otherwise it is *reducible.* A class is *closed* if it cannot be left.

8.6.4. Definition of probability of first arrival. Let $\boldsymbol{P}$ be the transition matrix of a Markov chain $(X_n)_{n\in\mathbb{N}_0}$ on the state space S. Given two states $i, j \in S$. The *probability offirst arrival at time n* to j from i is the likelihood to reach j for the *first time* at time n. In symbols we may write this as

$$f_{ij}^{(n)} := \mathcal{P}\big[X_1 \neq j, \ldots, X_{n-1} \neq j, X_n = j | X_0 = i\big] \text{ for } n \geq 1. \tag{8.6.1}$$

Write

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)} \tag{8.6.2}$$

as the chance we ever hit j, starting from i.

8.6.5. Definition. State i is *recurrent* if $f_{ii} = 1$, or *transient* if $f_{ii} < 1$.

8.6.6. Theorem. *If i is transient, then $\sum_n p_{ii}^{(n)}$ converges, if i is recurrent, that sum diverges. If i and j are in the same class, they are recurrent or transient together.*

8.6.7. Definition. Let d_i be the greatest common divisor of $\{n : p_{ii}^{(n)} > 0\}$. Then d_i is the *period* of state i. If $d_i = 1$, we say that i is *aperiodic*, otherwise i is *periodic* with period d_i.

8.6.8. Theorem. *All states in the same class have the same period.*
Hint. There are often two quick ways to show that state i is aperiodic. First, check that $p_{ii} > 0$. If that doesn't work, look for paths from i back to itself in both two steps, and in three steps.

8.6.9. Why care about irreducible aperiodic chains? If an irreducible chain has period $d > 1$, its states split into d subclasses, $C_1, C_2, \ldots, C_d$ having the property that steps rotate through these subclasses in a strict order. Write $Q = P^d$; then the chain with transition matrix Q has $C_1, C_2, \ldots, C_d$ as d aperiodic subclasses. This gives a good excuse to concentrate on *irreducible aperiodic* chains.

8.6.10. Theorem (Erdős–Feller–Pollard). *For all states in an irreducible aperiodic chain:*

(a) If the chain is transient, $p_{ij}^{(n)} \to 0$ as $n \to \infty$

(b) If the chain is recurrent, $p_{ij}^{(n)} \to \pi_j$, where either

(i) every $\pi_j = 0$ – a null recurrent chain, or

(ii) every $\pi_j > 0$, $\sum_j \pi_j = 1$ and $\pi\boldsymbol{P} = \pi$, i.e. π is a stationary vector—a positive recurrent chain.

(c) In case (b), let μ_i be the mean time to return to i, given that we are now in that state. Then in a positive recurrent chain, $\mu_i = 1/\pi_i < \infty$, in a null recurrent chain, $\mu_i = \infty$.

8.6.11. Corollary. *For an irreducible chain:*

(i) $P(X_n = i) \to 0$ if the chain is transient or null, whatever its period;

(ii) $P(X_n = i) \to \pi_i > 0$ if the chain is aperiodic and positive recurrent.

Apply these results to a simple random walk. Recall the matrix of Example 9.6, $P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 2/3 & 0 & 1/3 \end{pmatrix}$. First, check that it is irreducible and aperiodic (easy), then spell out $\pi.P = \pi$ as

$$\frac{\pi_1}{3} + \frac{\pi_2}{2} + \frac{2\pi_3}{3} = \pi_1$$
$$\frac{\pi_1}{3} = \pi_2,$$

and we don't need the third equation. The solution is seen to be $\pi = \frac{1}{25}(12,4,9)$, the chain is recurrent with this as limiting vector. Suppose the states are the non-negative integers, and that p_{ij} takes the value zero when $j > i+1$, and the value $1/(i+2)$ when $0 \le j \le i+1$. Analyse it. We have

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & \dots \\ 1/3 & 1/3 & 1/3 & 0 & \dots \\ 1/4 & 1/4 & 1/4 & 1/4 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{8.6.3}$$

First, see that it is irreducible: for, for any pair i, j a possible path from i to j is $i \to 0 \to 1 \to \cdots \to j$.

Then note it is aperiodic, because (e.g.) $p_{00} > 0$.

Now look for a stationary vector, π. The equations are:

$$\begin{aligned} \pi_0/2 + \pi_1/3 + \pi_2/4 + \pi_3/5 + \cdots &= \pi_0 \\ \pi_0/2 + \pi_1/3 + \pi_2/4 + \pi_3/5 + \cdots &= \pi_1 \\ \pi_1/3 + \pi_2/4 + \pi_3/5 + \cdots &= \pi_2 \\ \pi_2/4 + \pi_3/5 + \cdots &= \pi_3 \text{ etc.} \end{aligned}$$

Taking the differences of consecutive equations, $0 = \pi_0 - \pi_1$, $\pi_0/2 = \pi_1 - \pi_2$, $\pi_1/3 = \pi_2 - \pi_3$ etc., from which we find $\pi_1 = \pi_0$, $\pi_2 = \pi_0/2$, $\pi_3 = \pi_0/6$. It is easy to guess that $\pi_i = \pi_0/i!$, and then to verify this guess by induction. Choosing $\pi_0 = e^{-1}$ makes $\sum \pi_i = 1$, so there is a stationary probability vector, and the chain is positive recurrent. In the long run, $P(X_n = j) \to e^{-1}/j!$, the Poiss(1) distribution.

8.6.12. Example (Ehrenfest urn model of diffusion).

PROBLEM. *The diffusion of gas molecules between two glass bulbs, linked by a thin tube, is modelled as two urns containing n balls altogether, r being in the left urn and the rest in the right one. Successive movements are when one of the n balls is chosen at random, and moved to the other; this gives a Markov chain on* $\{0,1,2,\dots,n\}$*, the state being the number of balls in the left urn. Describe the long-term behaviour.*

Solution. Check it is irreducible; realise the period is 2. Write down the equations for a stationary vector anyway:

$\pi_1/n = \pi_0$; $\pi_0 + 2\pi_2/n = \pi_1$; $(n-1)\pi_1/n + 3\pi_3/n = \pi_3$ and so on. Easy to see that we can get π_1 in terms of π_0; then π_2 in terms of π_1 and π_0, hence in terms of π_0 alone; and so on – every π_j in terms of π_0. Then use the fact that the sum of these π_j should equal unity to find π_0, hence everything else.

Or, by a piece of inspiration, check that $\pi_k = \binom{n}{k}\frac{1}{2^n}$ satisfies inductively.

Once you have the solution, you then kick yourself, as it is "perfectly obvious" that, in the long run, each ball has, independently, probability 1/2 of being in the left urn, so the number there must indeed follow this binomial distribution.

Exercises and problems on discrete stochastic processes

Exercise 8.1. Given the offspring pgf $0.4+0.2z+0.4z^2$, find the pgfs for the sizes of generations one and two, starting with one individual. Find the probabilities of dying out

(i) *by* the second generation

(ii) *in* the second generation.

Exercise 8.2. A branching process corresponds to a cell either splitting into two identical cells, or dying, with respective probabilities p and $q = 1-p$. What happens, in the long run, for $p < 0.5$, $p = 0.5$ and $p > 0.5$?

Exercise 8.3. An organism has a Poisson number of offspring with mean $1+\epsilon$, where $\epsilon > 0$ is small. Show that the probability this branching process does not die out is close to 2ϵ.

Exercise 8.4. Let x be the smallest positive root of $z = f(z)$. A population begins with K members, all have offspring according to the pgf $f(z)$. What is its chance of extinction?

Exercise 8.5. For what values of λ is $f(z) = 0.2+(1.2-\lambda)z+(\lambda-0.4)z^3$ a pgf? For what values of λ is the corresponding branching process certain to die out?

Exercise 8.6. You enter a casino with $100, you will leave when you are bankrupt, or have reached $500, whichever comes earlier. Consider the three possible strategies:

(i) Bet $10 each time;

(ii) Bet $100 each time;

(iii) Bet your complete fortune each time (or, if a lesser winning bet would reach your target, this lesser sum).

You only bet on the "even" chances, i.e. those that pay out at even money. Find your chances of reaching your target, for each strategy in turn, for (a) a Las Vegas casino, with winning chance 18/38, and (b) a UK casino, with winning chance 73/148.

The remaining questions refer to Markov chains, and assume the corresponding notation.

Exercise 8.7 (Markov chain)**.** Given

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \end{bmatrix},$$

find P^2, the values of $p_{ii}^{(3)}$, and the unique stationary probability vector. Repeat for

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Exercise 8.8 (Markov chain). Construct the transition matrix of some twelve-state Markov chain with a transient class of four states, having period three, and accessible from an aperiodic transient class; a recurrent class of period two, not accessible from any other class; and an aperiodic recurrent class with three states.

Exercise 8.9 (irreducible Markov chains). Show that an irreducible aperiodic finite chain cannot be transient or null recurrent.

Exercise 8.10 (doubly stochastic Markov chain). A Markov chain is said to be *doubly stochastic* if the columns also sum to unity. Show that an irreducible aperiodic doubly stochastic chain with *finitely* many states is positive recurrent, and spends equal time in all states in the long run. Give random walk examples of doubly stochastic chains with infinitely many states that are (i) transient and (ii) null recurrent.

Exercise 8.11 (Markov chain). On the state space $\{1\dots7\}$, consider the Markov chain given by the transition matrix

$$\boldsymbol{P} = \begin{bmatrix} & & 1/2 & 1/2 & & & \\ & & & & & 1/3 & 2/3 \\ & & & & 2/3 & & 1/3 \\ & & & & 1 & & \\ 3/4 & & & & & & 1/4 \\ & 1/2 & & & & & 1/2 \\ & 2/3 & & & & 1/3 & \end{bmatrix} \text{(0 entries omitted).} \qquad \text{(P8.11.1)}$$

Find the two classes, the period of each class, and say whether the class is transient or recurrent. Identify any closed class, and find the stationary probability vector for it. Hence find the limit, as $n \longrightarrow \infty$, of $\mathcal{P}\left[X_n = j\right]$ for every state j. What is the mean time between visits to state 7?

Exercise 8.12 (Markov chain). Consider a Markov chain on the state space $\{1\dots6\}$ of transition matrix

$$\boldsymbol{P} := \begin{bmatrix} & 1/3 & 1/3 & 1/3 & & \\ 1/3 & & 1/3 & 1/6 & 1/6 & \\ 1/3 & 1/3 & & & & 1/3 \\ & & & & 1 & \\ & & & 2/3 & & 1/3 \\ & & & & 1 & \end{bmatrix} \text{(0 entries omitted).} \qquad \text{(P8.12.1)}$$

(a) Find the classes, and their periods. For each class is it recurrent or transient? Justify your answer.

(b) Evaluate $p_{12}^{(2)}$ and $p_{13}^{(3)}$.

(c) Use indicator variables to show that the mean number of visits to state 3, starting at state 1, is $\sum_{n=1}^{\infty} p_{13}^{(n)}$.

(d) Now alter the transition matrix, to make $p_{45} = 1/3$ and $p_{46} = 2/3$. Find a stationary probability vector of the form $\left(0,0,0,x,y,z\right)$, and deduce that the mean time between successive visits to state 4 is 11/3.

Exercise 8.13 (positive recurrence). The state space is $\mathbb{N}_0 = \{0,1,2,\dots\}$, and $p_{0j} = a_j$ where $a_j > 0$ for all j and $\sum_{a=0}^{\infty} j = 1$. Suppose that, for $j \geq 1$, $p_{jj} = \theta$ and $p_{j,j-1} = 1-\theta$ for some $0 < \theta < 1$. Show that this chain is irreducible and aperiodic. Show that, if

$\sum ja_j = \mu$ is finite, the chain is positive recurrent. Find its limiting distribution, and the mean time between successive visits to state 10.
Alter the matrix (after its top two rows) to make the non-zero entries $p_{jj} = \theta$, $p_{j,0} = 1-\theta$ for $j \geq 2$. Solve the same problem, showing that this chain is always positive recurrent, irrespective of the value of μ.

Bibliography

Apostol, Tom M. (1974). *Mathematical analysis*. Second. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., pp. xvii+492.

Billingsley, Patrick (1995). *Probability and measure*. Third. Wiley Series in Probability and Mathematical Statistics. A Wiley-Interscience Publication. New York: John Wiley & Sons Inc., pp. xiv+593. ISBN: 0-471-00710-2.

Bouniakowsky, V. (1859). "Sur quelques inégalités concernant les intégrales ordinaires et les intégrales aux différences finies". In: *Mémoires de l'Académie Impériale des Sciences de St.-Pétersbourg, VIIème Série* 1.9, pp. 1–18.

Capiński, Marek and Ekkehard Kopp (2004). *Measure, integral and probability*. Second. Springer Undergraduate Mathematics Series. London: Springer-Verlag London Ltd., pp. xvi+311. ISBN: 1-85233-781-8.

Cauchy, Augustin Louis (1821). *Œuvres*. Vol. 2. 3. Paris: Gauthier-Villars. URL: `http://openlibrary.org/works/OL6347037W/%C3%85%C2%92uvres_comple%C3%8C%C2%80tes_d'Augustin_Cauchy`.

Evans, Lawrence Craig (2009). *An Introduction to Stochastic Differential Equations*. Online Version 1.2. Berkeley, California, USA. URL: `http://math.berkeley.edu/~evans/SDE.course.pdf`.

Graham, Ronald L., Donald E. Knuth and Oren Patashnik (1994). *Concrete mathematics*. Second. A foundation for computer science. Reading, MA: Addison-Wesley Publishing Company, pp. xiv+657. ISBN: 0-201-55802-5.

Grinstead, Charles M. and J. Laurie Snell (1997). *Introduction to Probability*. American Mathematical Society. ISBN: 978-0-8218-0749-1. URL: `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`.

Jacod, Jean and Philip Protter (2003). *Probability essentials*. Second. Universitext. Berlin: Springer-Verlag, pp. x+254. ISBN: 3-540-43871-8.

Knuth, Donald E. (1992). "Two notes on notation". In: *Amer. Math. Monthly* 99.5, pp. 403–422. ISSN: 0002-9890. DOI: `10.2307/2325085`. URL: `http://arxiv.org/abs/math/9205211v1`.

Lakkis, Omar (2011). *Introduction to Pure Mathematics*. Online lecture notes. published freely online under Creative Commons license. University of Sussex. URL: `https://dl.dropboxusercontent.com/u/15751353/omar_lakkis-mathematics/Notes/General/Introduction_to_Pure_Mathematics--Lakkis--2011.pdf`.

Lieb, Elliott H. and Michael Loss (2001). *Analysis*. Second. Vol. 14. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, pp. xxii+346. ISBN: 0-8218-2783-9.

Muller, Greg and various authors (2007). *The Axiom of Choice is Wrong.* post with discussions. URL: http://cornellmath.wordpress.com/2007/09/13/the-axiom-of-choice-is-wrong/ (visited on 13/09/2007).

Podolak, Martin (2012). *CyrAcademisator.* URL: http://podolak.net/en/russian-studies/cyracademisator (visited on 14/04/2012).

Strang, Gilbert (2010). *Are Random Triangles Acute or Obtuse?* URL: http://www.youtube.com/watch?v=XxHIrVTLubE&feature=relmfu (visited on 24/02/2010).

Stroock, Daniel W. (1999). *A concise introduction to the theory of integration.* Third. Boston, MA: Birkhäuser Boston Inc., pp. xiv+253. ISBN: 0-8176-4073-8.

Villani, Cédric (2003). *Topics in optimal transportation.* Vol. 58. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, pp. xvi+370. ISBN: 0-8218-3312-X.

Wilf, Herbert S. (1994). *generatingfunctionology.* Second. Freely available online edition. Boston, MA: Academic Press Inc., pp. x+228. ISBN: 0-12-751956-4. URL: http://www.math.upenn.edu/~wilf/DownldGF.html.

Yamamoto, Tetsuro and Yasuhiko Ikebe (1979). "Inversion of band matrices". In: *Linear Algebra Appl.* 24, pp. 105–111. ISSN: 0024-3795. DOI: 10.1016/0024-3795(79)90151-4. URL: http://dx.doi.org/10.1016/0024-3795(79)90151-4.

Index