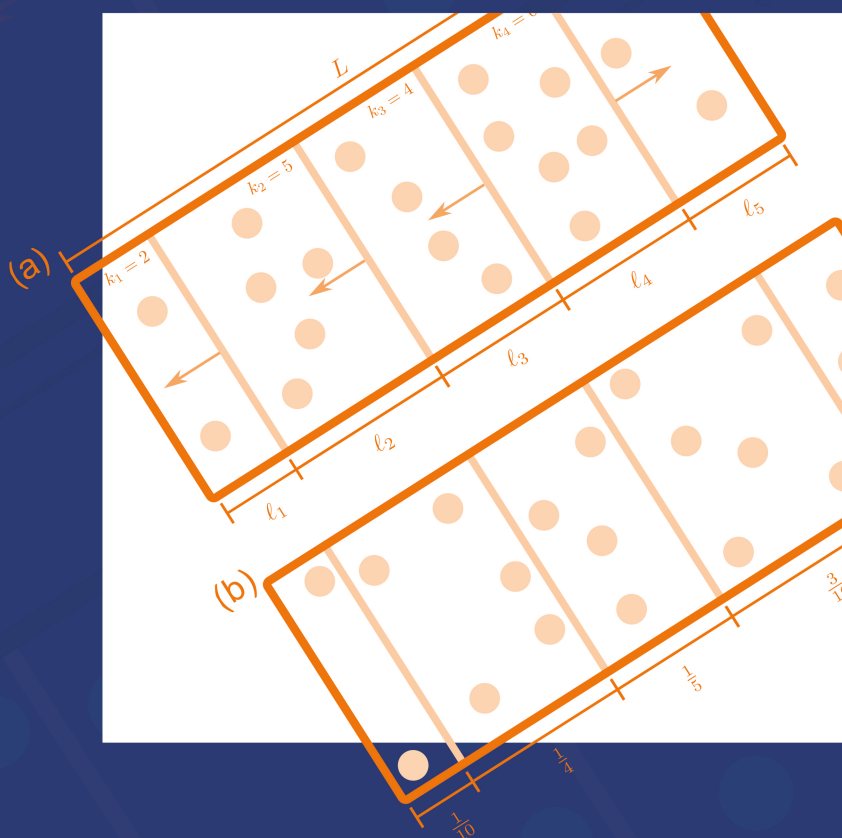


# 3

## Lectures on probability, information and large scale behaviour

Matteo Marsili







*This page was intentionally left blank.*



*This page was intentionally left blank.*



# Lecture on probability, information and large scale behaviour

Matteo Marsili

Published by SISSA Medialab S.r.l.

Via Bonomea 265

34136 Trieste, Italy

<https://medialab.sissa.it/>

Cover: Giacomo Sanna — Dotik

Typesetting: Elia A. Calderan and Giorgia del Bianco — SISSA Medialab S.r.l.



This book is distributed under the terms of the Creative Commons Attribution License ([CC-BY 4.0](#)), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

The third party material in this work are not included in the Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license, you will need to obtain permission directly from the copyright holder.

© 2025 Matteo Marsili



This book was published Open Access with funding support from the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3).

Title: Lecture on probability, information and large scale behaviour

Authors: Matteo Marsili

Keywords: 1. Probability 2. Information theory 3. Laws of large numbers 4. Limit theorems 5. Large deviations 6. Statistical mechanics 7. Statistical inference

First published 2025

ISBN: 9788898587063 (electronic edition)

DOI: [10.22323/9788898587063](https://doi.org/10.22323/9788898587063)

The unity of all science consists alone  
in its method, not in its material.

---

*Karl Pearson,*  
The Grammar of Science,  
(1892), p. 16



# Contents

<b>Overview</b>	<b>xiii</b>
<b>I Learning to count</b>	<b>1</b>
<b>1 A safe definition of probability</b>	<b>11</b>
1.1 Chance and randomness . . . . .	11
1.2 The concept of probability . . . . .	12
1.3 Kolmogorov's axioms . . . . .	13
1.4 The fallacy of intuition . . . . .	15
1.4.1 The Bertrand paradox . . . . .	16
<b>2 But what is probability?*</b>	<b>19</b>
2.1 de Finetti and subjective probabilities . . . . .	19
2.2 Probability as a theory of plausible reasoning . . . . .	20
2.2.1 A digression into logics . . . . .	21
2.2.2 Quantifying plausibility . . . . .	23
<b>3 Classical probability</b>	<b>31</b>
3.1 Combinatorics . . . . .	32
3.1.1 The Stirling's approximation to $n!$ . . . . .	34
3.2 Different ways of counting . . . . .	35
3.2.1 Balls in boxes and draws with and without replacement	37
3.2.2 Sub-sampling . . . . .	37
3.2.3 Distinguishable and indistinguishable balls in $n$ boxes	38
3.3 An extension of the sub-additivity rule . . . . .	40
<b>4 Stochastic dependence</b>	<b>45</b>
4.1 Statistical dependence is not causation . . . . .	50
<b>5 Random variables</b>	<b>53</b>
5.1 Many random variables . . . . .	55
5.2 Examples of random variables . . . . .	57

5.3	Expectation . . . . .	63
5.4	Correlation and factor graphs* . . . . .	66
<b>6</b>	<b>On urn models and sampling*</b>	<b>71</b>
6.1	Sampling and undersampling . . . . .	75
<b>7</b>	<b>Generating functions</b>	<b>79</b>
7.1	Warm-up: Fibonacci numbers . . . . .	79
7.2	Asymptotics of $a_n$ from the structure of singularities . . . . .	81
7.3	Counting with functions* . . . . .	83
7.3.1	Operations on sets . . . . .	84
7.4	Labeled objects . . . . .	90
7.5	Generating functions for integer random variables . . . . .	91
7.5.1	Sums of variables and convolutions . . . . .	92
7.5.2	Sums of a random number of random variables . . . . .	95
7.5.3	Cumulant generating function . . . . .	99
<b>8</b>	<b>More on balls and boxes*</b>	<b>103</b>
<b>9</b>	<b>Random walks</b>	<b>111</b>
9.1	The reflection principle . . . . .	114
9.2	Returns and first returns . . . . .	115
9.3	Last visit and the arc-sine law . . . . .	117
9.4	Random walks with drift . . . . .	119
9.4.1	Returns to the origin . . . . .	120
9.4.2	Last visit to the origin . . . . .	122
<b>10</b>	<b>Branching processes</b>	<b>125</b>
10.1	The main equation . . . . .	126
10.2	The extinction probability . . . . .	127
10.3	The total progeny and universality . . . . .	130
10.4	An application to random networks* . . . . .	136
<b>11</b>	<b>Markov chains</b>	<b>141</b>
11.1	Stochastic matrices . . . . .	142
11.2	Classification of states . . . . .	143
11.3	The invariant distribution . . . . .	146
11.4	Time reversibility . . . . .	149
<b>12</b>	<b>Exercises on the first part of the course</b>	<b>151</b>

<b>II</b>	<b>Typical and atypical</b>	<b>161</b>
<b>13</b>	<b>Almost surely et el.</b>	<b>167</b>
13.1	Limits in probability . . . . .	168
13.1.1	Almost certain convergence . . . . .	168
13.1.2	Convergence in probability . . . . .	170
13.1.3	Convergence in mean square . . . . .	171
13.1.4	Convergence in distribution . . . . .	171
13.2	Borel-Cantelli lemmas . . . . .	172
<b>14</b>	<b>Laws of Large Numbers</b>	<b>177</b>
14.1	The weak law . . . . .	179
14.2	The strong law . . . . .	180
14.3	Typical samples and the Asymptotic Equipartition Property . . . . .	182
14.3.1	Should we expect the expected value? . . . . .	189
<b>15</b>	<b>Limit theorems and universality</b>	<b>193</b>
15.1	Limit theorems for Sums of i.i.d. random variables . . . . .	193
15.1.1	Relation to the Law of Large Numbers . . . . .	194
15.1.2	Characteristic functions . . . . .	195
15.1.3	Derivation of the fundamental equation . . . . .	196
15.1.4	Stable distributions and universality . . . . .	203
15.1.5	Sums as stochastic processes . . . . .	206
15.2	Limit theorems for extremes . . . . .	210
15.2.1	Some applications* . . . . .	215
<b>16</b>	<b>Information theory</b>	<b>225</b>
16.1	Shannon entropy and Shannon theorem . . . . .	228
16.1.1	Entropy for continuous variables . . . . .	232
16.1.2	Relative entropy . . . . .	233
16.1.3	Mutual information . . . . .	235
16.2	The data processing inequality . . . . .	238
16.3	The entropy of Markov Chains . . . . .	239
16.3.1	Irreversibility and the arrow of time . . . . .	240
16.4	Data compression and coding theory . . . . .	242
<b>17</b>	<b>Large deviation theory</b>	<b>251</b>
17.1	LDT with finite support . . . . .	253
17.2	LDT for thin tails . . . . .	259
17.3	LDT and the Legendre transform . . . . .	262
17.4	How much do we learn?* . . . . .	266
17.5	Weakly correlated variables . . . . .	268
17.5.1	Large deviations for Markov Chains . . . . .	273
17.6	LDT for fat tails . . . . .	274

<b>18 States of knowledge</b>	<b>279</b>
18.1 Maximum entropy . . . . .	279
18.1.1 Generalised thermodynamics . . . . .	283
18.1.2 Maximum entropy learning* . . . . .	285
18.1.3 Continuous variables . . . . .	287
18.1.4 What can we learn? . . . . .	291
<b>19 Statistical mechanics</b>	<b>293</b>
19.1 Statistical mechanics as maximum entropy inference . . . . .	297
19.2 The classical ideal gas . . . . .	300
19.2.1 The Szilárd information engine* . . . . .	302
19.3 The Ising model . . . . .	306
19.4 The Random Energy Model . . . . .	311
19.4.1 A gas of weakly interacting particles and the Grand Canonical ensemble . . . . .	313
19.5 A teaser in stochastic thermodynamics* . . . . .	318
<b>20 Statistical inference</b>	<b>321</b>
20.1 Hypothesis testing . . . . .	324
20.2 Parameter estimation and the Fisher Information . . . . .	328
20.2.1 The Data Processing Inequality and Sufficient statistics	333
20.2.2 The Fisher Information . . . . .	335
20.2.3 The Cramer-Rao bound . . . . .	337
20.2.4 Distinguishability of distributions and Fisher information	338
20.2.5 Exponential families . . . . .	340
20.3 Model selection . . . . .	342
20.3.1 Akaike Information Criterion (AIC) . . . . .	343
20.3.2 Bayesian Information Criterion (BIC) . . . . .	345
20.3.3 Minimum Description Length . . . . .	349
20.4 The high dimensional limit and beyond . . . . .	351
20.5 Beyond statistical inference: learning and intelligence . . . . .	353
<b>21 Exercises for the second part</b>	<b>357</b>
<b>Index</b>	<b>375</b>



To my life's mentor, Daisaku Ikeda.

---



# Overview

The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure, even though perhaps also to our bafflement, to wide branches of learning. (E. Wigner [1])

These lecture notes are divided in two parts. The first part will focus on *modelling*, i.e. how to translate a real world problem, into a mathematical statement, and on the tools to derive from this a quantitative answer. In particular, we shall see that modelling is often a reduction of the problem to a paradigmatic problem of probability theory (e.g. drawing balls from urns, distributing balls in boxes, random walks, etc) for which one can develop a theory and an intuition. The second step, often deals with counting, so a major part of the first part of the course will be about *learning to count*. We shall mostly rely on the definition of probability as given by Kolmogorov axioms. This definition, however, is axiomatic and it does not tell us anything about what probability really is. A more modern derivation of probability — as the language that should be used to extend logic to discuss about the plausibility of statements — will be given following the book of E. T. Jaynes [2], to which we shall refer as JAYNES.

For the rest, this first part of the course is heavily based on the book by Feller [3], to which we shall refer as FELLER.

The second part of the course will focus on more advanced subjects. First we will delve into the relation between probability and information theory. Second we'll focus on asymptotic properties. These are key to understanding the collective behaviour of real systems. We will realise that a *typical behaviour* emerges with its rules and laws. The Central Limit Theorem is probably the clearest example that formalises the idea that the collective (or large scale)

behaviour of a system, under specific assumptions, may be independent of microscopic details. This is precisely the same logic of the more advanced applications of the renormalisation group in statistical physics.

We shall then focus on *atypical behaviour*, i.e. the most likely way in which very unlikely events, such as *large deviations*, occur.

We shall find that these concepts provide an unifying language<sup>1</sup> for a broad range of different disciplines, from statistical physics in physics, to statistical inference and computer science (coding and complexity theory). Phenomena such as phase transitions manifest in a different way in different disciplines, but they build on the same theoretical foundations, though they are discussed with a different language. The main goal of the second part of the course is to discuss the concepts which underlie all these fields.

The main text we shall follow in the second part is the one by Cover and Thomas [4], to which we shall frequently refer as COVER. A further textbook to which we shall refer is the one by Gnedenko [5], using the acronym GNEDENKO. Some of the chapters and sections are excursions to applications of ideas discussed in other chapters and they are marked with an asterisk \*. They may serve to gain deeper understanding and intuition.

As for the reason for this particular structure, I realised after my studies that my understanding of many problems had suffered from the fact that I had been taught subjects in the wrong order.<sup>2</sup> Take the concept of entropy: I first learned about it from thermodynamic transformations in heat engines. Then I was taught that this is a measure of volume in phase space in statistical mechanics. But I understood what entropy really is when I studied typical sequences, Shannon's theorem and information theory. This also made it clear why it is related to volume in phase space and heat, that remained somewhat mysterious until that point. The same applies to limit theorems or large deviation theory, a subject which is often clouded in sophisticated mathematical language and is approached with unnecessary awe by beginners.

Probability is often regarded as a branch of mathematics. If we agree with E. T. Jaynes, that probability is the *language of science*, then it should be spoken properly by all scientists, not only by measure theorists. Much of classical probability (and most of the books that I refer to in this course) can indeed be understood with a rather limited knowledge of mathematics, which should

---

<sup>1</sup>Interestingly, many non-trivial statistical phenomena can be understood in simple settings. Having a good grasp of what happens in simple settings, such as those addressed in the sequel, provides a guide for attacking more complex problems, besides conforming to Occam's prescription *Frustra fit per plura quod potest fieri per pauciora*.

<sup>2</sup>I was lucky enough to follow the course in probability theory given by G. Jona-Lasinio in my undergraduate studies at the University of Rome, *La Sapienza*.

be familiar to all undergraduate students who passed the basic courses in mathematics.<sup>3</sup>

How do we learn? We feel stranger in a city where we cannot go from A to B. We learn by turning corners and realising “ahah, I could have come here also from this other path!” I believe the same is true for a subject. At the beginning we start populating an empty map and with time, we start seeing the connections. At the end we can navigate autonomously and we enjoy walking on our own.<sup>4</sup> Surprise is a driving force in the discovery process, in which we instinctively take beauty as a truth certificate.<sup>5</sup> Aesthetic amusement, I believe, is what makes us dig deeper, and a relevant part of what makes us humans.

## Acknowledgments

This book grew out of the lecture notes that I’ve been distributing to students of a course that I have been teaching in the last twenty seven years. It is the result of my interaction with students in all these years, initially PhD students in SISSA, Trieste, then students of the International Master in the Physics of Complex Systems and lately students of the Diploma programme in Quantitative Life Sciences at the Abdus Salam International Centre for Theoretical Physics (ICTP). The text reflects their comments, questions and corrections. It reflects their thirst for understanding which is one of the most valuable things on earth. I’m deeply indebted to all of them.

Even though generations of students have helped me correct errors in these notes, many errors may still remain. My gratitude is extended to all those who would point out further mistakes.

I couldn’t have developed these ideas without the constant and critical interaction with my colleagues in the scientific community. It’s hard to list names, but those who read these words may know whether I’m referring to them or not. I regard the atmosphere of critical attitude I have breathed in my career is a sacred fire that we all cherish and preserve for future generations. The sense of belonging to this community is one of the things I value the most.

---

<sup>3</sup>Mathematics deals with proving true statements. That is almost never possible in science. Science is about falsifiability of theories. It reduces to a disciplined method to show that something (the prediction of a theory) is wrong, which is a much easier task. Our current theories are those that survived all these attacks, but this does not mean that they are true.

<sup>4</sup>Walking on your own, in this course, means challenging yourself with the exercises.

<sup>5</sup>A point that is maybe best expressed by John Keats verses:

*Beauty is truth, truth beauty, — that is all  
Ye know on earth, and all ye need to know.*

A self-critical attitude is necessary to distill true wisdom, but the wellspring of even the most rigorous theorem lies in beliefs. We wouldn't set out to prove something if we didn't believe it can be true. The beliefs that drove me in this process have been heavily shaped by the teachings of Nichiren and by their practice, as explained by Daisaku Ikeda, to whom goes my deepest gratitude.

## **Part I**

# **Learning to count**





Probability is at the basis of scientific and quantitative analysis. It formalises the approach with which we go from a question on a real world problem to a quantitative estimate, a number.<sup>6</sup> The stages in this process are the following:

**Real world.** The problem we are interested typically refers to a situation that occurs in the real world. This is full of details, from the atomic composition of the entities involved up to the properties of the environment they are immersed in. Experiments can be carried out and quantitative measures can be taken of the relevant quantities.

**A description.** The problem we're interested in is described in common language, with a text of finite length. This description is silent about many details of the real world and (hopefully) only concentrates on the relevant details that are necessary to arrive at a quantitative answer.

**A mathematical model.** The description has to be translated in mathematical language, introducing the appropriate variables and the appropriate assumptions.

**A calculation.** The solution to the problem entails a mathematical calculation. If it reproduces experimental results then we're allowed to believe that we reached some understanding, i.e. that the description captures relevant ingredients and that experiments measure relevant quantities.

## From the real world to mathematics

Much of physics is the science of approximation: to zeroth order, a cow is a sphere and the only two number that suffice to describe it are its radius and mass. To first order we can add the head and the legs, making the mathematical description more complex, and so on. The appropriate level of description varies with the type of questions we're interested in. Theoretical physicists inhabit the land of spherical cows, whereas if you want to build bridges and airplanes that fly, you need to take into account many more details. The description of a real world situation entails innumerable details, but these hopefully can be arranged in a hierarchy of relevance and we can cut it to achieve the desired level of approximation. This is true in physics, but it is by no means trivial. Take the free fall of bodies: since Galileo Galilei we know that

---

<sup>6</sup>From this point of view, probability theory could be thought of as the *theory of the theories of everything*.

a body falling from an height  $h$  takes a time  $t = \sqrt{2h/g}$  to reach the ground, where  $g = 9.81m/s^2$ .

It is definitely remarkable that there is such a specific relation between  $t$  and  $h$ . What is more remarkable is that  $t$  does not depend on any other detail. So blue bodies fall exactly in the same manner as red bodies. There is a sharp separation between relevant details (the height) and irrelevant ones (the color, the smell, etc) which is non-trivial. As Wigner says [1] (read this essay!), the fact that such precise relations exist and that they have a mathematical form is a gift. There are two other aspects which are worth to point out in this respect.

First, not all possible questions have such sharp answers, i.e. depend on few variables. Much of science is precisely about identifying those questions which allow for such sharp answers. These are typically very unnatural questions, you would rarely ask in your daily life. Imagine what Galileo's contemporaries were thinking of him spending his days letting objects fall from a tower.

The second aspect is that the statement above refers to a quite idealised situation. Have you ever tried to use the relation  $t = \sqrt{2h/g}$  to measure  $g$ ? If you do, you will see that every time you get a different number. The more you control the conditions under which you do the experiment, the more the dispersion of the numbers you get decreases. We use the term *experimental errors* to describe this fact, but there is no error in how bodies fall. The error refers to our lack of experimental ability and to account the influence of aspects that we deem *irrelevant*, as they don't enter the relation  $t = \sqrt{2h/g}$ . The conditions under which  $t = \sqrt{2h/g}$  holds with good precision are somewhat far from the typical ones that hold in the real world, they are quite un-natural conditions. The second aspect, is that in the end every statement about the real world is a probabilistic statement: the time it takes for the body to fall will be close to  $\sqrt{2h/g}$  most of the time, i.e. with high probability.

It is important to reflect on the appropriateness of this approach as we move our attention away from physics, to life sciences.<sup>7</sup> We'd dream to find a cure for cancer or Parkinson's disease. These are not questions that we have chosen. There is no reason to believe that the occurrence of cancer depends on few causes or variables, or that a single pill can cure it. There is no guarantee that the same sharp separation between relevant and irrelevant variables holds there and there may be no idealised conditions under which this is true. Often therapies are developed and tested on *model* organisms of increasing complexity, from yeast and worms, to rats and monkeys. Yet what cures a disease in worms may not work in rats. Even quantifying the

---

<sup>7</sup>Indeed, even in physics there are strong coupling problems where a perturbative approach of successively refined approximations does not work.

relevance of variables in a specific problem is an issue. In these domains, even more, all statements are of probabilistic nature, and discipline in going from a real world problem to a quantitative result is a key issue.

## Translating a description into mathematics

Summarising, the best we can do is to develop a discipline to translate real world problems into mathematical problems.

*Kolmogorov's axioms* define a general scheme for describing a problem in probability in a mathematically precise manner. This entails defining the *sample space* — i.e. the set of all possible outcomes — and how the probability is assigned to each of them. This is an important point which will be treated in detail in the next lecture. For the time being, let us appeal to an intuitive notion of what probability is.

The first step is to “read carefully” the statement of the problem, both what is written and what is not written. Let's illustrate this with few examples.

What is the probability of a single pair at poker?

Let's analyse this question. In this statement there's a lot of missing information of three different types:

**Irrelevant details.** Implicitly the statement refers to a real world situation where the game of poker is played by some players, each with different expertise and dressed differently... the cards are of a certain brand and... All these details are not contained in the description of the problem. The answer is assumed to be the same irrespective of these details. It means that they are irrelevant.

**Common knowledge.** It is implicitly assumed that we know that poker is played with a set of 52 cards, of 4 different groups, each numbered from 1 to 13; and that a hand at poker consists of 5 cards drawn from the 52. We assume that a “single pair at poker” is a concept which is common knowledge.

**Implicit information.** The statement does not say anything about how the 5 cards are chosen. Yet this is a relevant information. There is no reason to believe that a particular group of 5 cards will be more or less likely than some other group (otherwise it would have been stated in the problem). So there is a symmetry in the problem which implies that the probability of each group of 5 cards must be the same. This is a very useful information, because it reduces the problem to that of counting the number of ways in which a single pair at poker can arise.

So the probability of a single pair at poker is the fraction of all possible hands at poker that result in a single pair. Among the  $\binom{52}{5} = 2598960$  possible hands, there are<sup>8</sup>

$$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3 = 1098240$$

ways to choose the 5 cards. So the probability is 0.422569.

The key lessons to learn from this exercise is *i*) what is written in the statement of a problem is important, but what is not written may be even more important and *ii*) in many cases computing probabilities amounts to counting outcomes.

## Prototype models of probability theory

What is the probability that at least two people have a birthday on the same day of the year in a room with  $n$  people?

Again there is a lot of implicit information. In particular, there is no information that suggests that individuals are more likely to be born in certain days, so we shall assume that every day is equally likely as a birthday.<sup>9</sup>

There is no information about the relation among the individuals, so we should assume that there is no relation. Knowing the birthday of Mr X does not tell us anything on when Mrs Y was born. So the correct way to translate our ignorance on the relation between the people is to treat their birthdays as independent variables.<sup>10</sup> If there were twins among the  $n$  people this would not be true. But if this were true, it would have been specified in the statement, so we disregard this possibility.

Finally, we should consider that one every 4 years is a leap year. We shall neglect this fact for simplicity, and work under the approximation that every year has 365 days. This is an approximation. Whether this is appropriate or

---

<sup>8</sup>There are  $\binom{13}{1}$  ways of choosing the number which appear twice and  $\binom{4}{2}$  ways of choosing the two equal cards among the four possible ones. The other three cards must be different, which account for the factor  $\binom{12}{3}$  and each of them can be of  $\binom{4}{1}$  possible types.

<sup>9</sup>If one looks closely at statistics this is not actually true. There are certain times of the year when there are more newborn than in other days, depending on the geographical location. We assume we don't have this information, and again ignorance entails symmetry that means equiprobability. Here what we treat as common knowledge is somewhat arbitrary. If the question would be asked at a conference on demography probably this assumption might not be tenable.

<sup>10</sup>In loose terms, there are many ways in which the birthdays of different people could be related, but there is only one way in which they can be unrelated. If there is nothing that suggests in which direction this relation should go in the statement of the problem, then it's reasonable to assume that there is no relation.

not depends on the context. It is definitely appropriate for the point we want to make here.

With these premises, the problem becomes formally equivalent to one of drawing at random  $n$  balls in 365 boxes and asking what is the probability that at least two balls fall in the same box. There are many other problems that can be formally mapped into problems of distributions of *balls in boxes*, so it makes sense to study random distribution of balls in boxes in its own right. Balls and boxes is the first prototype model of probability theory that we have encountered but there are many others. In many instances, the answer to a problem in probability entails finding ways to map it into one of these prototype problems. We'll see more examples below.

The answer entails counting all configurations where one or more boxes contains two or more balls. It's definitely easier to count the number of configurations where no box has more than one ball and to subtract this from the total number of ways to draw the balls. If one thinks of computing this number as the number of ways we can accommodate the first ball, times the number of ways we can accommodate the second, etc we realise that this problem is equivalent to one of drawing  $n$  times balls from an urn with  $r = 365$  distinguishable balls, *without replacement*.<sup>11</sup> The number of ways in which we can draw  $n$  balls from an urn of  $r$  balls without replacement is

$$r(r-1) \cdot (r-n+1) = \frac{r!}{(r-n)!}.$$

The problem of counting how many possible configurations of birthdays there can be in total, instead, is equivalent to drawing  $n$  times from an urn of  $r = 365$  distinguishable balls *with replacement*, because each birthday can be chosen to be any of the  $r$  days. So this number is  $r^n$  and the probability is

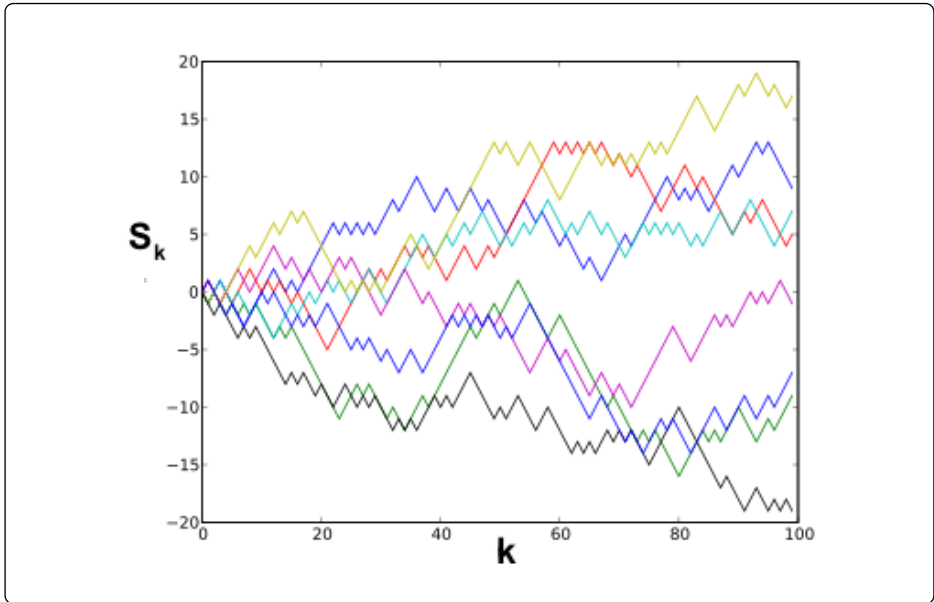
$$1 - \frac{r!}{(r-n)!} r^{-n}, \quad r = 365$$

A different problem we shall discuss is how to estimate these numbers, when  $n$  and  $r$  is large. We anticipate that this probability is of order one when  $n \simeq \sqrt{r} \simeq 19$ . *Drawing balls from urns with or without replacement*, or with more complicated procedures is a further prototype model of probability of intrinsic interest.

The show at a theatre in Moskow costs 5 rubles.  $2n$  people show up.  $n$  of them have only notes of 10 rubles, whereas the rest has

---

<sup>11</sup>Notice: in this further description of the problem, boxes become balls!



**Figure 1.** The variable  $S_k$  as a function of  $k$  in different possible realisations of the arrival of customers at the theatre.

notes of 5 rubles. The cashier initially has no notes. What is the probability that the cashier has no change to give to some customer?

Here the fact that the theatre is in Moscow is irrelevant, it suggests that this exercise probably first appeared in a Russian book on probability theory. The  $n$  people can show up in any possible order, so each of them is equiprobable. What matters to answer the question is whether, at any time, the number of customers with 5 rubles that have arrived up to that time is at least as large as the number of customers with 10 rubles that have arrived so far, or not.

So the key variable is the difference  $S_k$  between the customers with 5 rubles and those with 10 rubles that have arrived up to time  $k$ , i.e. when the  $k^{\text{th}}$  customer has arrived. Then the problem can be conveniently represented graphically by drawing a plot of  $S_k$  as a function of  $k = 1, \dots, 2n$ . All possible paths in this plot correspond to a different order in which customers can arrive and there is no reason to assume that a particular path is more likely than some other. So all paths are equally likely. A random paths with this property is called a *random walk*, which is yet another cornerstone models of probability theory.

If the cashier has no change to give at some point in time, it means that  $S_k < 0$  for some  $k = 1, \dots, 2n$ . So the problem above is equivalent to computing the probability that a random walk of  $2n$  steps, that returns to the origin at time  $2n$  (because  $S_{2n} = 0$ ), never visits the negative half plane. This is a classical problem in random walk theory that we will discuss.

## Using invariance, random variables and generating functions

As an overview of the concepts that we shall discuss in the first part of the course, consider the following problem:

Mr X checks emails every minute with probability  $p$ . He receives on average  $\lambda$  emails per minute. What is the probability that Mr X finds no email the next time he checks?

The answer to this question involves a few conceptual steps that are useful building blocks in dealing with a large number of problems.

**Probability distributions.** The probability that in a given minute Mr X receives  $k$  emails is given by the Poisson distribution

$$P\{Z = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

As we will see, this is the only distribution that is consistent with the (implicit) assumption of time translation invariance (i.e. that any time is as likely as any other time for the arrival of an email) and of independence (the arrival of an email now does not imply that emails are more or less likely to arrive in the near future). It is important to learn which distribution is appropriate for which situation.

**Random variables.** The number of emails that Mr X finds is a random variable  $N$

$$N = Z_1 + \dots + Z_T$$

where  $T$  is the number of minutes that have passed since last time he checked emails, and  $Z_t$  are the emails that arrived in the  $t^{\text{th}}$  minute.  $Z_t$  are also random variables. The probability that  $Z_t = k$  is the Poisson distribution discussed above, and all  $Z_t$  are independent. Indeed also  $T$  is a random variable. With probability  $p$  it is equal to 1, with probability  $p(1 - p)$  it is equal to 2, ... with probability  $p(1 - p)^{t-1}$  it is equal to  $t$  (this is a *geometric distribution*).

We're interested in the event that  $N = 0$ , that only occurs if all  $Z_t = 0$ . Decomposing the problem as in the equation above, paves the way to finding a solution in a simple manner.

**Computing with functions.** A convenient way to compute the answer is to consider the functions

$$\phi(s) = \sum_{t=1}^{\infty} P\{T = t\} s^t = \frac{p s}{1 - (1-p)s}, \quad \psi(s) = \sum_{k=0}^{\infty} P\{Z = k\} s^k = e^{(s-1)\lambda}$$

that are conveniently written as *expected values*:  $\phi(s) = \mathbb{E}[s^T]$  and  $\psi(s) = \mathbb{E}[s^Z]$ . These functions are called *generating functions*. Then we can write

$$\mathbb{E}[s^N] \equiv \sum_{t=1}^{\infty} P\{T = t\} \mathbb{E}[s^Z]^t = \mathbb{E}[\mathbb{E}[s^Z]^T] = \phi(\psi(s))$$

and, after a moment of reflection, it is clear that the sought answer is given by setting  $s = 0$  in this expression, i.e.

$$P\{N = 0\} = \phi(\psi(0)) = \frac{p}{e^{\lambda} - 1 + p}. \quad (1)$$

Indeed many counting problems can be solved very efficiently by introducing appropriately defined (generating) functions (i.e.  $\phi$  and  $\psi$  here). How to count with functions will be another important subject of the first part of the course.

The aim of the first part of the course is to acquire familiarity with all the concepts and techniques involved in the derivations above (as well as with others), in order to be able to tackle and solve complex problems.

### Exercise

Consider the following problems. Find what is the missing information in their statements and of which type? What does "surprising" means in the first problem?

1. "In a parking lot there are 12 places arranged in a row. A person observes that 8 places are taken and the 4 free places are adjacent to each other. Is this surprising?"
2. "Mr Brown has  $n$  keys. Only one of them opens the door. What is the probability that he needs to try  $k$  of them to open the door?"
3. "What is the probability that in a family with five children, none is a girl?"



# Chapter 1

## A safe definition of probability

Probability begins and ends with probability.

(John Maynard Keynes,  
*The Application of Probability to Conduct*)

### 1.1 Chance and randomness

In the classical textbook of GNEDENKO we find a definition of probability as “that branch of mathematics that deals with the regularities in chance phenomena”. But what are “chance phenomena”? Are there real chance phenomena? Is a financial crisis a chance phenomenon? And what about an earthquake? Conversely, are there phenomena that are *really* deterministic? Any experiment is to some extent affected by uncontrollable effects that we may call “chance”. Indeed, laws of physics describe ideal situations in which the predicted outcome only occur when a number of factors are carefully controlled.

We believe events generally happen because of causal mechanisms, yet we might not be able to specify or know all the conditions that are necessary for an event  $A$  to occur.<sup>1</sup> Only if a specified set of conditions  $\Omega$  contains all those ingredients that are necessary for the event  $A$  to occur, we may say that  $A$  is certain. Since this is rarely the case, we are left with statements about the likelihood of events under specified conditions, that take the form:

---

<sup>1</sup>For the moment, you can think of an event  $A$  as a statement, e.g. a description of what happens, e.g. an earthquake of magnitude between 6.3 and 6.5 in a give period of time (e.g. next month) and region. Likewise, we think of the conditions  $\Omega$  as a set of statements specifying all the information we have on the factors potentially relevant for the event: e.g. the time series of previous earthquakes, whether nuclear tests are going to be performed or not in that region, levels of humidity and temperature etc.

The probability of the event  $A$  given the conditions  $\Omega$  is  $p$

By convention we assume  $0 \leq p \leq 1$ .

On the other hand, “chance events” which are typical of classical probability, such as the toss of a coin (e.g.  $A = \{\text{head}\}$ ), are not intrinsically stochastic. We could include in  $\Omega$  all those informations that allows us to solve the equation of motion of the coin from its start, and that would allow us to say exactly whether  $A$  occurs or not. Therefore chance events exist because of our negligence in the description of the problem or of our lack of information. Indeed probability has a lot to do with information as we’ll see. But for the moment being, let it suffice to say that in the description of any phenomena, there are a set of conditions that we specify and control. All the rest, at finer level of description, is what we call “random” or “chance”. Chance is a useful label to attach to all those details that we ignore or consider irrelevant with respect to the questions we’re interested in.

It’s important to understand what are the rules of chance, because it is important to make sure that the predictions we derive are robust and meaningful.

## 1.2 The concept of probability

Probability is a primitive concept. In order to see this, Marinari and Parisi [6] offer the following reasoning: let us focus on a simple event like the occurrence of a particular outcome in an experiment (e.g. head in coin tossing). One might try to define the probability  $p$  of an event as the limit of the frequency  $f_n$  of its occurrence, when the experiment is repeated many times and the number  $n$  of trials goes to infinity:

$$p \equiv \lim_{n \rightarrow \infty} f_n \quad (1.1)$$

This means that  $\forall \epsilon > 0$  there is an  $\bar{n}$  such that  $\forall n > \bar{n}$ ,

$$|f_n - p| < \epsilon. \quad (1.2)$$

However this cannot be true, because the fact that  $f_n$  is close to  $p$  by less than a distance  $\epsilon$  is itself a random event. However large  $n$  may be, realisations in which the inequality (1.2) is violated are possible. So at most one can say that such deviations become very unlikely as  $n$  gets large, i.e.

$$\lim_{n \rightarrow \infty} \text{Prob}\{|f_n - p| > \epsilon\} = 0.$$

Therefore Eq. (1.1) is a definition of probability which relies on the concept of probability. Indeed probability cannot be defined in terms of other concepts. Probability is a primitive concept. One way to deal with primitive concepts is to give them an axiomatic definition.

We note in passing that Eq. (1.1) is a non-trivial fact, a result called the Law of Large Numbers, that we shall derive later in the course, once we have a proper definition of probability.

### 1.3 Kolmogorov's axioms

The theory of probability is based on three objects

$$(\Omega, \mathcal{A}, \mathcal{P})$$

- $\Omega$  is the sample space. If we are dealing with an experiment, its elements  $\omega$  are all its possible outcomes. If we are dealing with a forecast for a future time  $\omega$  is a possible *state of the world* at that time.  $\Omega$  can be a finite set, or a set of countably infinite elements, or a continuum measurable set.
- $\mathcal{A}$  is a  $\sigma$ -field. In words, it is a collection of subsets  $E \subseteq \Omega$  of the sample space – that are called *events*<sup>2</sup> — satisfying the following three properties:

$$i) \quad \Omega \in \mathcal{A}$$

$$ii) : \text{ if } A \in \mathcal{A} \text{ then } \bar{A} = \Omega/A \in \mathcal{A}$$

$$iii) \quad \mathcal{A} \text{ is closed under countable unions: this means that if } A_1, A_2, \dots \in \mathcal{A} \text{ then also } A_1 \cup A_2 \cup \dots \in \mathcal{A}.$$

Since  $A_1 \cap A_2 \cap \dots = \overline{\bar{A}_1 \cup \bar{A}_2 \cup \dots}$ , these three properties imply that  $\mathcal{A}$  is closed also under countable intersections.<sup>3</sup>

---

<sup>2</sup>**Notation:** If  $A, B \subseteq \Omega$  are two events,  $A \cup B$  is the union, that corresponds to points  $\omega$  that either belong to  $A$  or to  $B$ .  $\cup$  is equivalent to the OR logical operation or to addition (+) in mathematics. Likewise  $A \cap B$  is the intersection, that contains points  $\omega$  that belong both to  $A$  and to  $B$ .  $\cap$  is equivalent to the AND logical operation or to product ( $\times$ ) in mathematics. I denote by  $A/B$  the set of points  $\omega \in \Omega$  that belong to  $A$  but that do not belong to  $B$ . This operation is analogous to the difference between sets. The set  $\bar{A} = \Omega/A$  is the complement of set  $A$ , that includes all points  $\omega \in \Omega$  that do not belong to  $A$ . The complement operation  $\bar{\phantom{x}}$  is analogous to the logical negation (NOT). The empty set is denoted as  $\emptyset = \bar{\Omega}$ .

<sup>3</sup>In all cases we will discuss  $\mathcal{A}$  will be the family of all possible subsets of  $\Omega$ .

- $\mathcal{P}$  is the *probability measure*, which is a real function defined on  $\mathcal{A}$

$$\mathcal{P} : \mathcal{A} \rightarrow [0, 1] \quad (1.3)$$

$$A \in \mathcal{A} \rightarrow P(A) \in [0, 1] \quad (1.4)$$

which satisfies the following properties

$$\text{positivity : } P(A) \geq 0, \forall A \in \mathcal{A} \quad (1.5)$$

$$\text{normalization : } P(\Omega) = 1 \quad (1.6)$$

$$\text{additivity : } P\left(\bigcup_i A_i\right) = \sum_i P(A_i), \quad (1.7)$$

$$\forall A_i \in \mathcal{A} : \forall i \neq j \ A_i \cap A_j = \emptyset$$

$$\text{continuity : } \forall A_1 \supseteq A_2 \supseteq \dots, A_n \rightarrow \emptyset, \quad (1.8)$$

$$\text{then } P(A_n) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (1.9)$$

As consequences of these axioms we have

- for all  $A \subseteq \Omega$ ,  $P(A) \leq 1$  and  $P(\bar{A}) = P(\Omega) - P(A) = 1 - P(A)$ , because  $A$  and its complement  $\bar{A} = \Omega/A$  are disjoint, and we can apply the additivity rule.
- Subadditivity: for all  $A, B \in \mathcal{A}$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

because  $B/A$  and  $A$  are disjoint, as well as  $B/A$  and  $A \cap B$ . Using  $B = (B/A) \cup (A \cap B)$  and  $A \cup B = (B/A) \cup A$  and the additivity rule gives the result.

- if  $A \subseteq B$  then whenever  $A$  occurs  $B$  also occurs. In other words, this means that “ $A$  implies  $B$ ”, or  $A \Rightarrow B$ . For all  $A, B \in \mathcal{A}$  such that  $A \Rightarrow B$ ,  $P(A) \leq P(B)$ .

It is important to remark that the probability of an event  $A$  depends on the state of knowledge. This is encoded in  $\Omega$  because  $\Omega$  specifies all the outcomes that are possible. If the state of knowledge changes, because an event  $\Omega' \subseteq \Omega$  is known to be true, then the state of knowledge changes to  $\Omega'$ . The more we know, the more the sample space  $\Omega$  shrinks. So, strictly speaking, we should use the notation  $P(A|\Omega)$  for the probability of event  $A$  under the conditions specified by  $\Omega$ . Therefore, all probabilities are *conditional* to a given state of knowledge  $\Omega$ , i.e. all probabilities are *conditional probabilities*. Yet, when  $\Omega$  is fixed, we shall disregard the dependence on it and write simply  $P(A)$  for the

probability of event  $A$ . We shall come back to this point when we will discuss conditional probability.

Kolmogorov's axioms provide the dictionary and the basic grammar rules to discuss probability.

One annoying aspect of Kolmogorov axioms is illustrated by the following example (taken from FELLER, p. 8): imagine we want to know the probability that a person lives more than 1000 years. A statistician's approach would extrapolate from formulas extracted from mortality tables and come with a probability of one in  $10^{10^{36}}$ , a number which clearly makes no sense. Still if one is going to take seriously the problem, one has to decide whether this event (a person living 1000 years) is possible or not. In the first case, it has to be assigned a positive probability. If we assume it to be impossible, then the event does not belong to  $\Omega$ . But then we should find out what is the maximal age a person can live, based on first principles, in order to define  $\Omega$ . This problem needs to be solved before we even start talking about probabilities.

We shall discuss other definitions of probability theory that overcome these difficulties. Our focus here is on computing, i.e. quantifying the plausibility of statements. For this, Kolmogorov's axioms are enough.

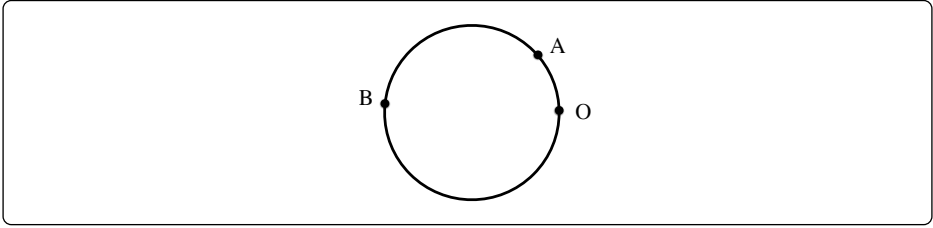
## 1.4 The fallacy of intuition

Probabilistic thinking is hard wired in us, as there are regions of our brain that are activated when, for example, we have to take decision in uncertain circumstances or that have uncertain consequences. So we have a lot of intuition about probability. Yet there are also well documented biases in our probabilistic thinking [7], so it is important not to rely blindly on our intuition.

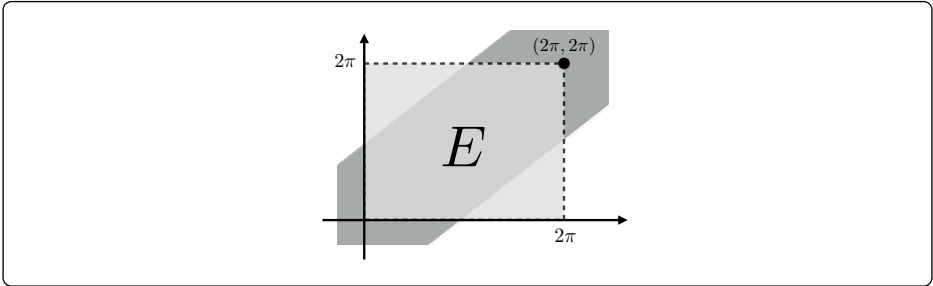
In order to be sure about what your intuition suggests, you can use Kolmogorov's axioms: ask yourself: what is  $\Omega$ ? What is  $\mathcal{P}$ ? Then do the calculation. Let's consider a couple of examples:

Take two points  $A$  and  $B$  at random on a circle. This divides the circle in two arcs. What is the probability that the arc that contains the origin is larger than the other?

Intuitively there is no reason to think that one of the arcs should be larger than the other, so you would conclude that  $p = 1/2$ . But let's check. A point on a circle is identified by the angle, i.e. by a number  $X \in (0, 2\pi]$ . So two points correspond to a pair  $(X_A, X_B) \in (0, 2\pi]^2$ . The sample space is then  $\Omega = (0, 2\pi]^2$  and every subset  $E \subseteq \Omega$  is a possible event. Each outcome is a priori equally likely, so  $\mathcal{P}$  is the uniform measure on  $(0, 2\pi]^2$ . Event  $E \subseteq \Omega$  has a probability  $p = |E|/|\Omega|$  where  $|E|$  is the area of the set  $E$ .



**Figure 2.** Taking two points, A and B, at random on a circle, what is the probability that the arc that contains the origin is larger than the other?



**Figure 3.** The sample space of the problem discussed in the text.

Now you can translate the statement above in a precise mathematical form. Actually it is easier to draw the sample space and identify the set corresponding to the event

$$E = \{\text{the arc that contains the origin is larger than the other}\}.$$

If you do that, you discover that the right answer is  $p = 3/4$ , contrary to intuition.

How can this be possible? On second thought you realise this is expected. Indeed there is a symmetry and the choice of the origin breaks it. So there are three points drawn at random in reality,  $A$ ,  $B$  and  $O$ . So there are three intervals and the interval containing the origin in the problem formulation corresponds to the union of two of them. It is natural to expect that it should be longer than the other.

### 1.4.1 The Bertrand paradox

A further aspect where our intuition may need to be checked is the notion of *drawing at random*. Let us consider the following problem:

Choose a chord  $AB$  at random on the circle of radius one. What is the probability  $p$  that  $AB$  is longer than the side of the inscribed triangle (which is  $\sqrt{3}$ )?

Bertrand gave three different answers:

- by symmetry chose the chord to be horizontal and draw the vertical diameter. If the chord intersects the diameter in its middle half, then  $AB > \sqrt{3}$ . This means  $p = 1/2$ .
- by symmetry, chose one of the end points ( $A$ ) to coincide with a vertex of the triangle. The second point is identified by choosing the angle  $\theta \in [0, \pi]$  of the chord with the tangent in  $A$ . Then  $AB > \sqrt{3}$  if  $\theta \in [\pi/3, 2\pi/3]$ . This means  $p = 1/3$ .
- the chord is uniquely identified by its middle point. So one can chose this middle point at random in the circle. If the point falls inside the circle inscribed in the triangle, then  $AB > \sqrt{3}$ . Since the area of the inscribed circle is 4 times smaller than that of the outer circle,  $p = 1/4$ .

The problem is that the sentence “the chord is chosen at random” has not a clear meaning, and it is indeed interpreted differently, with a different  $\Omega$  and  $\mathcal{P}$ , in the three cases above.<sup>4</sup> There is no paradox.

### Exercise 1.1

1.  $A$  is the event of a single pair at poker. What is  $\Omega$ ? What is  $\mathcal{P}$ ?
2.  $A$  is the event that at least two people have a birthday on the same day of the year in a room with  $n$  people. What is  $\Omega$ ? What is  $\mathcal{P}$ ?
3. The show at a theatre in Moskow costs 5 rubles.  $2n$  people show up in a random order.  $n$  of them have only notes of 10 ruble, whereas the rest has notes of 5 ruble.  $A$  is the event that the cashier has no change to give to some customer. What is  $\Omega$ ? What is  $\mathcal{P}$ ?
4.  $N$  gentlemen go to theatre each leaving his hat at the wardrobe. On exit they are assigned their hats in a random order.  $A$  is the event that none of the gentlemen get his own hat back. What is  $\Omega$ ? What is  $\mathcal{P}$ ?

<sup>4</sup>Notice that we assume that, in each of the three cases, drawing at random implies an uniform probability distribution over  $\Omega$ , as if ignorance is naturally translated into equiprobability. This is not an innocent assumption. For example, why should an interval  $[\theta, \theta + d\theta]$  have the same probability when  $\theta$  is close to the endpoints ( $\theta = 0, \pi$ ) and in the middle ( $\theta = \pi/2$ )? Jaynes discusses this issue in some detail in this paper: *Prior probabilities* [8].

5. A fair coin is tossed until for the first time the same result appears twice consecutively.  $A$  is the event that the experiment ends before the 6th toss.  $B$  is the event that an even number of tosses is required. Compute the probabilities of  $A$  and of  $B$ .
6. Consider two dice. Let  $A = \{\text{sum of the faces is odd}\}$  and  $B = \{\text{at least one ace}\}$ . Describe the events  $A \cup B$ ,  $A \cap B$  and  $A \cap \bar{B}$ . Assuming that each outcome is equiprobable, find the probabilities of all these events.
7. An insurance is interested in the age distribution of couples  $(x, y)$ , where  $x$  is the age of the husband and  $y$  is the age of the wife (both are integers, in years). What is the sample space? What is the event  $A$  that the husband is older than 40,  $B$  that the husband is older than the wife and  $C$  that the wife is older than 40? Draw them. Show that  $A \cap \bar{C} \subset B$ .
8. Verify the relations and try to express them in words:
  - (a)  $\overline{A \cup B} = \bar{A} \cap \bar{B}$
  - (b)  $A \cup A = A \cap A = A$
  - (c)  $(A \cup B)/(A \cap B) = (A \cap \bar{B}) \cup (\bar{A} \cap B)$
  - (d)  $\overline{A \cup \bar{B}} = \bar{A} \cap B$
9. Find simpler expressions for
  - (a)  $(A \cup B) \cap (A \cup \bar{B})$ ,
  - (b)  $(A \cup B) \cap (\bar{A} \cup B) \cap (A \cup \bar{B})$
  - (c)  $(A \cup B) \cap (A \cup C)$
10. Let  $A, B$  and  $C$  be three events. Find expressions for the events:
  - (a) Only  $A$  occurs
  - (b) All three events occur
  - (c) at least two occur
  - (d) two and no more occur
  - (e) not more than two occur
  - (f) none occurs
  - (g) at least one occurs



## Chapter 2

# But what is probability?\*

“ [...] le regole della logica probabilistica [...] — come quelle della logica formale nel campo delle proposizioni — ci insegnano a ragionare nel campo delle valutazioni di probabilità mantenendo intatta la coerenza del pensiero con se stesso.” (B. De Finetti [9])

Kolmogorov’s axioms are enough for computing probabilities in an unambiguous manner. Yet, they don’t provide any insight on what probability really is.

### 2.1 de Finetti and subjective probabilities

Bruno de Finetti, argued that probability is nothing else than the degree of confidence that an individual has that some event will actually occur or that a fact is true. Probability is subjective by definition.

One way to quantify the probability of an event  $A$  is to devise a lottery. A ticket of the lottery grants a payoff of one pound to its holder if  $A$  occurs and nothing otherwise. The price  $P(A)$  of a ticket of this lottery measures the degree of confidence that a buyer has on the likelihood of event  $A$ . Clearly  $P(A) \geq 0$ . For an event that is almost certain,  $P(A)$  should be close to one (pound), and if  $A$  is very unlikely then  $P(A)$  should be small.

When there is more than one event, the price of the tickets of the corresponding lotteries should be fixed in a consistent way, in order to ensure that the system of lotteries is fair, and that no-one can extract a positive gain without taking any risk.<sup>1</sup> Prices should be such that an agent would be indifferent between being on the sell or the buy side. This implies that:

---

<sup>1</sup>This is known as the *no-arbitrage hypothesis* in finance.

- If two events  $A$  and  $A'$  are equally likely, the tickets of the corresponding lotteries should be the same,  $P(A) = P(A')$ .
- If  $A$  and  $A'$  are exclusive events (i.e.  $A \cap A' = \emptyset$ ), then the price of the combined lottery for  $A \cup A'$ , that grants a win of one pound if either  $A$  or  $A'$  occur, should be equal to the sum of the prices of the lotteries for  $A$  and  $A'$ , i.e.  $P(A \cup A') = P(A) + P(A')$  pounds.<sup>2</sup>
- If  $\bar{A}$  is the event that  $A$  does not occur, a gambler that buys one ticket of the lottery  $A$  and one of the lottery  $\bar{A}$  is sure to win one pound. Hence  $P(A) + P(\bar{A}) = 1$  pound.

These rules are consistent with the Kolmogorov axioms that define the probability  $P(A)$ , and indeed de Finetti has shown that they are identical. In addition, they give a meaning to the probability  $P(A)$  of an event  $A$  as the amount that an individual is willing to bet on its occurrence. Different individuals may assign different probabilities to the same event, so probability is *subjective*. Yet, each of them should assign probabilities to different events in a way which is consistent with the rules of probability.

Buying a share of a stock in the financial market is like buying a ticket of a lottery. Indeed, Bruno de Finetti's idea of relating probability to monetary outcomes of uncertain events is the basis of the theory of finance.<sup>3</sup> He laid the foundations of asset pricing theory (i.e. the theory that says how the price of a stock or a financial instrument should be fixed) and of portfolio theory.

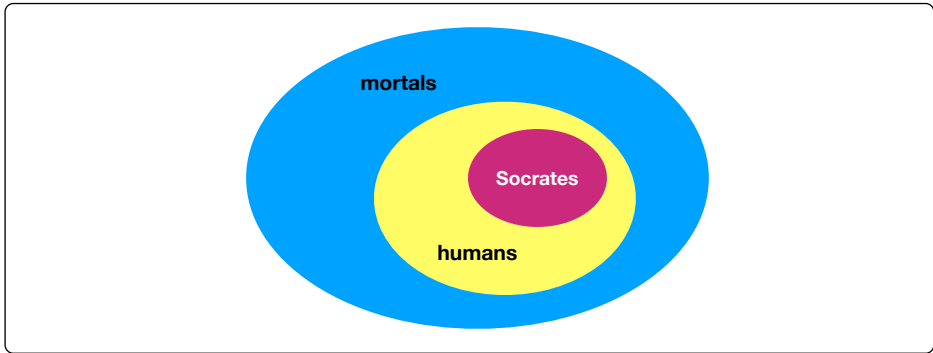
## 2.2 Probability as a theory of plausible reasoning

Several authors have argued that probability is a way to formalise in a quantitative manner our way of *reasoning* about the plausibility of statements. We follow the discussion in the first two chapters of JAYNES, to which we refer for a more detailed discussion. Here we only sketch the main ideas.

The first observation is that deductive logic (see Figure 4) is hardly applicable to the real world, as there are few cases where we can say that a statement  $A$  implies another statement  $B$  in the strong sense (if  $A$  is true, then  $B$  is true). In real life and in science, we're almost always arguing about how the fact

<sup>2</sup>Because if  $P(A \cup A') > P(A) + P(A')$ , buying a ticket for  $A \cup A'$  and selling one ticket for both  $A$  and  $A'$ , would ensure a gain  $P(A \cup A') - P(A) - P(A') > 0$  irrespective of what happens.

<sup>3</sup>de Finetti worked at the Generali insurance company in Trieste for some time, where he faced the problem of computing prices for insurance contracts.



**Figure 4.** The typical syllogism on which deductive logic is based: "all humans are mortal", "Socrates is a human" then "Socrates is mortal".

that something ( $A$ ) is true affects the plausibility of something else ( $B$ ).<sup>4</sup> As J. Clerk Maxwell put it:

[...] the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of probability which is, or ought be, in a reasonable man's mind.

### Exercise 2.1

Does the quote "What I cannot create, I do not understand", attributed to R.P. Feynman, implies that he though he could create what he understood?

## 2.2.1 A digression into logics

The basic entities that are the object of probability theory are statements, such as

$A$  = it will start to rain by 10 am at the latest

<sup>4</sup>This applies also to mathematics. While theorems express logical relations between statements they are almost always derived or discovered starting from reasonable conjectures. The worth of any theory  $A$  is to make predictions on other statements, e.g. if  $A$  is true then  $B$  is true. We can falsify  $A$  if we find that  $B$  is false. But if  $B$  is true we can only say that  $A$  is more plausible (see later). We cannot prove that  $A$  is true.

which are either true or false. Logic provides a language to combine different statements into more complex ones:<sup>5</sup>

*Logical product or conjunction*  $AB$  = both  $A$  and  $B$  are true

*Logical sum or disjunction*  $A + B$  = either  $A$  or  $B$  are true

*Negation or denial*  $\bar{A}$  =  $A$  is false

that satisfy a set of properties

*Idempotence*  $AA = A, A + A = A$

*Commutativity*  $AB = BA, A + B = B + A$

*Associativity*  $A(BC) = (AB)C = ABC$

$$\text{and } A + (B + C) = (A + B) + C = A + B + C$$

*Distributivity*  $A(B + C) = AB + AC, A + (BC) = (A + B)(A + C)$

*Duality* If  $C = AB$ , then  $\bar{C} = \bar{A} + \bar{B}$  and if  $D = A + B$ , then  $\bar{D} = \bar{A}\bar{B}$ <sup>6</sup>

It can be shown that with these operations we can generate all possible statements.

$A$	T	F
$f_1(A)$	T	T
$f_2(A)$	T	F
$f_3(A)$	F	T
$f_4(A)$	F	F

**Table 2.1.** All possible statements derived from  $A$ .

### Exercise 2.2

Think of all possible statements involving two events  $A, B$  (see table 2.2). How many are them? Express them in terms of sum, negation and product.

<sup>5</sup>Although the notation is different, the product  $AB$  for statements is the same as the intersection  $A \cap B$  for sets, and the sum  $A + B$  is the same as the union  $A \cup B$ .

<sup>6</sup>Note that  $\bar{A}\bar{B} \neq \overline{AB}$ .

$A, B$	T,T	F,T	T,F	F,F
$f_1(A, B)$	T	T	T	T
$f_2(A, B)$	T	F	T	T
$f_3(A, B)$	F	T	T	T
...	...	...	...	...

**Table 2.2.** All possible statements involving two events  $A, B$ .

Yet, the language of logics is redundant as indeed the same statement can be expressed in many ways. For example<sup>7</sup>

$$C = (A + B)(\bar{A} + AB) + \bar{A}\bar{B}(A + \bar{B}) = \bar{A} + B.$$

$n$  elementary statements  $A_1, \dots, A_n$  can only have  $2^n$  possible truth assignments, which means that there are  $2^{2^n}$  different statements that can be obtained by combining them.

The redundancy of the language based on the three logical operators above suggests that any statement can be expressed in terms of a fewer number of operators. It can be proved that all operators can be expressed in terms of the NAND operator  $A \uparrow B \equiv \overline{AB} = \bar{A} + \bar{B}$ . For example, you can check that

$$\begin{aligned}\bar{A} &= A \uparrow A \\ AB &= (A \uparrow B) \uparrow (A \uparrow B) \\ A + B &= (A \uparrow A) \uparrow (B \uparrow B).\end{aligned}$$

### 2.2.2 Quantifying plausibility

JAYNES approach to probability is normative: how should a robot assign plausibility to different statements in a “correct” way? First, the plausibility of any statement  $A$  depends on the state of knowledge of the robot, i.e. those statements that the robot knows to be true. If  $B$  is the statement that encodes the state of knowledge, plausibility should be a function of  $A|B$ , i.e. of *statement  $A$  given  $B$* . JAYNES shows that there is a unique way of defining probability that satisfies the following desiderata:

<sup>7</sup>Note that  $C$  is false whenever  $A$  is true and  $B$  is false. In this sense  $C$  can be read as the statement  $C = \{A \Rightarrow B\}$  that  $A$  implies  $B$ , because  $C$  being true means that if  $A$  is true then  $B$  must be true. The statement  $C$  does not say anything about  $B$  if  $A$  is false. Alternatively,  $C$  can also be expressed as the statement  $C = \{A = AB\}$ , which is true if whenever  $A$  is true  $B$  is also true. Logical implication seen in this way has the interesting property that all false statements imply any other statement, as well as their opposite, because if  $A$  is false then  $AB$  is also false (and hence  $A = AB$ , i.e.  $C$  is true) and  $A\bar{B}$  is also true. It’s suggestive to think about the implications of this fact for the “surprising” propagation of fake news. See [10].

I *Degrees of plausibility are represented by real numbers.*

II *Qualitative correspondence with common sense.*

For example, if  $A|C$  is less plausible than  $A|C'$  and if  $B|AC$  and  $B|AC'$  are equally plausible, then  $AB|C'$  should be more plausible than  $AB|C$  and  $\bar{A}|C'$  should be less plausible than  $\bar{A}|C$ .

III *Consistency*

1. If a conclusion can be reached in more than one way, then every possible way must lead to the same quantitative estimate of the plausibility.
2. The robot always takes into account all of the evidence which is relevant to a question that it is aware of. It does not base its conclusion on a subset of the information available, neglecting the rest.
3. The robot always represents equivalent states of knowledge by equivalent plausibility assignments. That is, if the robot's information about two statements is the same (except perhaps for the labelling of the propositions), then it must assign the same plausibilities in both.

We shall not repeat the derivation here and refer to JAYNES for it. We only state the key steps, which consists in deriving rules for computing the plausibility of composed statements such as the product  $AB|C$  and the sum  $A + B|C$  from the plausibility of elementary statements, e.g.  $A|C$  and  $B|C$ . The *product* and the *sum rule* are enough to compute the plausibility of any composite statement. We shall avoid using the word *probability* until the very end, and discuss instead about a generic measure of plausibility. As we shall see, for any measure of plausibility which is consistent with the desiderata above, it is possible to derive a function, that we call probability, that satisfies the product and sum rules of probability that we're used to.

The first requirement implies that there should be a function  $g(\cdot)$  that assigns to any statement  $A|B$  a real value  $g(A|B)$  that we call the plausibility of  $A$  given the state of knowledge  $B$ .<sup>8</sup>

**The product rule.** Let us start by considering the statement  $AB|C$ . JAYNES argues that its plausibility  $g(AB|C)$  should be a function

$$g(AB|C) = F[g(A|BC), g(B|C)] = F[g(B|AC), g(A|C)] \quad (2.1)$$

---

<sup>8</sup>Note that plausibility does not depend on what the statement is about.

of either  $g(A|C)$  and  $g(B|AC)$ , or of  $g(B|C)$  and  $g(A|BC)$ . Indeed, the reasoning about  $AB$  given  $C$  can be decomposed in two steps: first estimate the plausibility of  $B$  given  $C$  and then that of  $A$  given  $BC$ . Equivalently we can reason first about  $A$  given  $C$  and then about  $B$  given  $AC$ . The result must be the same for consistency, with the same function  $F(\cdot, \cdot)$ . Note also that Eq. (2.1) implies that  $g(A|BC)$  does not depend on the plausibility of other statements involving  $A, B$  and  $C$ . For example, whether  $A|\bar{B}C$  is more or less plausible should not affect the result, because if  $B$  is not true, then  $AB$  is also not true. Next, the function  $F$  should be a nondecreasing function of both arguments, for common sense.

Now, consider the plausibility of  $ABC$  given  $D$ . This can be expressed in two ways

$$\begin{aligned} g(ABC|D) &= F[g(BC|D), g(A|BCD)] = F[F[g(C|D), g(B|CD)], g(A|BCD)] \\ &= F[g(C|D), g(AB|CD)] = F[g(C|D), F[g(B|CD), g(A|BCD)]] \end{aligned}$$

which implies that the function  $F$  satisfies

$$F[F[x, y], z] = F[x, F[y, z]] \quad (2.2)$$

where  $x = g(C|D)$ ,  $y = g(B|CD)$  and  $z = g(A|BCD)$ . It is easy to check that the function<sup>9</sup>

$$F(x, y) = w^{-1}(w(x)w(y))$$

where  $w(x)$  is a monotone increasing function of  $x$ , satisfies Eq. (2.2). For a proof that this solution is also unique, under general assumptions of continuity that derive from *common sense* (II), we refer to JAYNES.<sup>10</sup>

The key point of the derivation is that the function

$$\gamma(A|B) = w[g(A|B)]$$

satisfies the product rule

$$\gamma(AB|C) = w(g(AB|C)) \quad (2.3)$$

$$= w(F(g(A|BC), g(B|C))) \quad (2.4)$$

$$= w(g(A|BC)) w(g(B|C)) \quad (2.5)$$

$$= \gamma(A|BC)\gamma(B|C). \quad (2.6)$$

<sup>9</sup>To prove this, use the fact that  $w(F(x, y)) = w(x)w(y)$  and check that applying  $w(\cdot)$  to Eq. (2.2) yields  $w(x)w(y)w(z)$  on both sides.

<sup>10</sup>Note that  $F(x, y) = F(y, x)$  is invariant under exchange of its arguments. Should this be expected?

Consider the case where  $B|C = T$  is true.<sup>11</sup> Then  $AB|C = A|C$  and  $B$  does not change the state of knowledge, i.e.  $BC = C$ . Eq. (2.6) then becomes  $\gamma(A|C) = \gamma(A|C)\gamma(B|C)$  that can only be satisfied for all  $A$  and  $C$  if  $\gamma(B|C) = 1$ . Hence we conclude that

$$B|C \text{ is true} \Rightarrow \gamma(B|C) = 1. \quad (2.7)$$

Next, if  $A|C = F$  is false,<sup>12</sup> then  $AB|C = A|C$  is also false and  $\gamma(AB|C) = \gamma(A|C)$ . Also  $\gamma(A|BC) = \gamma(A|C)$  because  $A|C$  is false irrespective of whether  $B$  is true or not. Therefore  $\gamma(A|C) = \gamma(A|C)\gamma(B|C)$  that can only be satisfied for all  $B$  and  $C$  if  $\gamma(A|C) = 0$ . Therefore

$$A|C \text{ is false} \Rightarrow \gamma(A|C) = 0. \quad (2.8)$$

Common sense implies that, the plausibility of any statement must be higher than that of a false statement and lower than that of a true statement. Hence  $\gamma(A|C) \in [0, 1]$  for all statements.

**The sum rule.** Let us now consider the two statements  $A|B$  and its negation  $\bar{A}|B$ . It is clear that if one of the two becomes more plausible the other decreases in plausibility, by common sense. So there should be a decreasing function  $S(x)$  such that  $\gamma(\bar{A}|B) = S(\gamma(A|B))$ . If  $A|B = T$  is true, we know that  $\bar{A}|B = F$  is false and eqs. (2.7) and (2.8) imply that  $S(1) = 0$  and  $S(0) = 1$ . For  $x \in [0, 1]$  the function  $S(x)$  takes also values in  $[0, 1]$ .

An equation for  $S(\cdot)$  can be derived by the following steps

$$\begin{aligned} \gamma(AB|C) &= \gamma(A|C)\gamma(B|AC) \\ &= \gamma(A|C)S(\gamma(\bar{B}|AC)) \\ &= \gamma(A|C)S\left(\frac{\gamma(A\bar{B}|C)}{\gamma(A|C)}\right) \end{aligned}$$

where we used  $\gamma(A\bar{B}|C) = \gamma(A|C)\gamma(\bar{B}|AC)$  in the last line. An equivalent equation can be derived by inverting  $A$  and  $B$  in the derivation, which leads to

$$\gamma(A|C)S\left(\frac{\gamma(A\bar{B}|C)}{\gamma(A|C)}\right) = \gamma(B|C)S\left(\frac{\gamma(\bar{A}B|C)}{\gamma(B|C)}\right) \quad (2.9)$$

This equations holds whatever  $A, B$  or  $C$  are. If we specialise to a situation where  $\bar{B} = AD$  with  $D$  an arbitrary statement, then<sup>13</sup>  $A\bar{B} = \bar{B}$ . Also  $B =$

<sup>11</sup>Here  $T$  is the true statement. It is analogous of the sure event  $\Omega$ .

<sup>12</sup> $F$  is the false statement. It is analogous to the impossible event  $\emptyset$ .

<sup>13</sup> $\bar{B} = AD$  means that  $\bar{B}$  implies  $A$ , i.e.  $\bar{B} \Rightarrow A$ , therefore  $A\bar{B} = \bar{B}$ . On the other hand, if  $A$  is false, then  $B$  cannot be false, i.e.  $\bar{A} \Rightarrow B$  and hence  $\bar{A}B = \bar{A}$ .



$\overline{AD} = \bar{A} + \bar{D}$ , so that  $\bar{A}B = \bar{A}(\bar{A} + \bar{D}) = \bar{A}$ . This reduces Eq. (2.9) to

$$xS\left(\frac{S(y)}{x}\right) = yS\left(\frac{S(x)}{y}\right). \quad (2.10)$$

with  $x = \gamma(A|C)$  and  $y = \gamma(B|C)$ . With  $y = 1$  the equation above yields  $x = S(S(x))$  which is consistent with the self-reciprocal property of the negation, i.e. the negation of  $\bar{A}$  is  $A$ , and with the conditions  $S(0) = 1$  and  $S(1) = 0$ .

A lengthy algebraic derivation (see JAYNES) shows that  $S(x)$  must be of the form

$$S(x) = (1 - x^m)^{1/m} \quad (2.11)$$

with  $m > 0$  a positive real number. This means that, for all measures of plausibility  $g(A|B)$  there exist a function

$$P(A|B) = w(g(A|B))^m \quad (2.12)$$

that satisfies the following product and sum rules

$$P(AB|C) = P(A|C)P(B|AC) \quad (2.13)$$

$$P(\bar{A}|C) = 1 - P(A|C) \quad (2.14)$$

We can forget about the plausibility function  $g(A|B)$  and just work with this function, that we call *probability*. It has the property

$$P(A|C) = 1 \quad \text{if } A \text{ is true given } C$$

and  $P(A|C) = 0$  if  $A$  is false given  $C$ .

### Exercise 2.3

Show that Eq. (2.11) is a solution of Eq. (2.10). Why should  $m$  be positive?

Knowing how the probability transforms under the operations of conjunction and negation allows one to compute the probability of combinations of statements. For example, one can derive the rule

$$P(A + B|C) = P(A|C) + P(B|C) - P(AB|C)$$

from the two properties above.<sup>14</sup> In particular, if the events  $A$  and  $B$  are mutually exclusive, i.e.  $P(AB|C) = 0$  then we obtain the additivity rule.

<sup>14</sup>Start with  $\overline{A + B} = \bar{A}\bar{B}$ , so  $P(A + B|C) = 1 - P(\bar{A}\bar{B}|C)$ . Then use  $P(\bar{A}\bar{B}|C) = P(\bar{A}|C)P(\bar{B}|\bar{A}C)$  and  $P(\bar{A}|C) = 1 - P(A|C)$ . The rest is left as an Exercise.

In brief, what Jaynes shows is that all the rules that are encoded in Kolmogorov axioms can be derived as an extension of logic, that defines how a robot should assign probabilities to statements, in a way to satisfy the desiderata above. This is remarkable.

As a final consequence of consistency, consider the situation where the robot has to assign probabilities to  $n$  exclusive statements  $A_1, \dots, A_n$ . If the state of knowledge  $B$  does not distinguish between the statements<sup>15</sup> (i.e.  $B$  does not say anything on  $A_i$  that it does not say on  $A_j$ ) then  $P(A_i|B) = P(A_j|B)$ . In this situation, the way in which the statements are labeled from 1 to  $n$  is completely arbitrary because any robot that looks at a problem where the labels are a permutation of the original ones, should give the same numerical values. Finally if the events are also exhaustive, i.e. if  $P(A_1 + \dots + A_n|B) = 1$ , then the permutation symmetry invoked above implies  $P(A_i|B) = 1/n$ . Indeed symmetries are a key element to compute probabilities [8].

As we observed, if  $A$  implies  $B$  given a state of knowledge  $C$ , then we cannot conclude anything about  $A$  if  $B$  is true. Yet, the probability of  $A$  should increase in the case where  $B$  is true. Indeed

$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|C)} = \frac{1}{P(B|C)}P(A|C) \geq P(A|C) \quad (2.15)$$

where we used the product rule in the first equation,  $P(B|AC) = 1$  in the second (because  $A$  implies  $B$  given  $C$ ) and  $P(B|C) \leq 1$ . Finding out that a quasi-obvious statement  $B$  that implies  $A$  is true, does not increase by much the likelihood of  $A$ , whereas a non-trivial statement  $B$ , with a small probability  $P(B|C)$ , increases the plausibility of  $A$  considerably. Likewise the fact that  $A$  is false, does not imply that a statement  $B$  that is implied by  $A$  is also false, but it decreases its plausibility by an amount that can be computed with a derivation similar to Eq. (2.15):

$$P(B|\bar{A}C) = \frac{P(\bar{A}|BC)P(B|C)}{P(\bar{A}|C)} = P(B|C) - \frac{P(A|C)}{P(\bar{A}|C)}P(\bar{B}|C) \leq P(B|C). \quad (2.16)$$

If  $B|C$  is likely true *a priori*, i.e. if  $P(B|C) \simeq 1$ , showing that a new fact  $A$  that could explain it is wrong does not affect its likelihood significantly. On the other hand, if  $A$  is very likely true, showing that it is false decreases considerably the probability that all its consequences  $B$  are true. This is why the changes in our state of knowledge that occur when well established theories are falsified are often called *paradigm shifts*.

---

<sup>15</sup>For example  $B = \{\text{the dice is fair}\}$  and  $A_i = \{\text{a throw of the dice results in } i\}$  with  $i = 1, \dots, 6$ .

**Exercise 2.4**

Derive the expression (2.16).

Jaynes formalisation of probability is conceptually more transparent than that based on Kolmogorov's axioms. Yet, from a computational point of view, Kolmogorov axioms are more practical (although they introduce unnecessary axioms and assumptions) and lead to the same conclusions.<sup>16</sup>

---

<sup>16</sup>If you're interested to know more, see the discussion in JAYNES, appendix A.



## Chapter 3

# Classical probability

Probability theory is nothing but common sense reduced to calculation. (Laplace, 1819)

Let us consider again the two problems we discussed earlier:<sup>1</sup>

1.  $A$  is the event of a single pair at poker.
2.  $A$  is the event that at least two people have a birthday on the same day of the year in a room with  $n$  people.

One way to estimate the probability of  $A$  is to identify those elementary events  $\omega \in \Omega$  that are “evidently” equiprobable. By this we mean that there is no indication in the statement of the problem or in our state of knowledge that would hint at the fact that some  $\omega$  are more or less likely than others.<sup>2</sup> This means that  $P(\omega) = P(\omega')$  for all  $\omega, \omega' \in \Omega$  and that  $P(\omega) = \frac{1}{|\Omega|}$  for all  $\omega \in \Omega$ , because of normalisation

$$\sum_{\omega \in \Omega} P(\omega) = |\Omega|P(\omega) = 1.$$

Here  $|\Omega|$  is the number of elements of  $\Omega$ . Therefore, for any event  $A \subseteq \Omega$ , the probability can be written as

$$P(A) = \sum_{\omega \in A} P(\omega) = \frac{|A|}{|\Omega|}$$

---

<sup>1</sup>This part is discussed in FELLER II, which you’re strongly suggested to study. In particular, if you want to make sure you master this material, challenge yourself with the problems at the end of the chapter.

<sup>2</sup>Another way to state the same fact, is that when the state of knowledge is such that the answer to a question is invariant with respect to any relabelling (or permutation) of the elementary events  $\omega \in \Omega$ , then  $P(\omega) = P(\omega')$  for all  $\omega, \omega' \in \Omega$ .

where  $|A|$  is the number of “favourable” cases and  $|\Omega|$  is the total number of cases. We stress the fact that what allows us to compute probability exactly is the (permutation) symmetry present in the problem (as in the first example above) or assumed (as in the second).

In these cases computing probabilities becomes a counting problem. This is the realm of *classical probability*.

### 3.1 Combinatorics

Counting problems are often combinatorial problems. The mathematical objects that occurs frequently are:

**Permutations.** The number of different permutations<sup>3</sup> of  $n$  objects (e.g. the numbers  $1, 2, \dots, n$ ) is given by the product of all integers up to  $n$

$$n! = n(n-1)(n-2) \cdot 2 \cdot 1,$$

which is called the *factorial* of  $n$ . Indeed, let us label the  $n$  objects by integers  $x_i$  from 1 to  $n$ . Then we can write a permutation as a sequence  $(x_1, x_2, \dots, x_n)$ , where  $x_i$  is the label of the object in position  $i$  ( $x_i \neq x_j \forall i \neq j$ ). Then  $x_1$  can be chosen in  $n$  ways,  $x_2$  in  $n-1$  ways, and so on.

**Ordered samples.** The number of ways to draw  $r$  out of  $n$  objects is given by

$$(n)_r = n(n-1) \cdots (n-r+1) = \frac{n!}{(n-r)!}.$$

Indeed, using the same notation as above, a draw of  $r$  of the  $n$  objects corresponds to an  $r$ -tuples  $x_1, x_2, \dots, x_r$  with  $x_i \in \mathbb{N}$ ,  $1 \leq x_i \leq n$  and  $x_i \neq x_j$  for all  $i \neq j$ . The number of distinct  $r$ -tuples is given by the expression above because  $x_1$  can be chosen in  $n$  ways,  $x_2$  in  $n-1$  ways, ... and  $x_r$  in  $n-r+1$  ways.

**Combinations.** The number of subsets of  $r$  objects of a set of  $n$  elements is

$$\binom{n}{r} = \frac{(n)_r}{r!}$$

Indeed, from each subset of  $\{x_1, \dots, x_n\}$  of size  $r$ , it is possible to form  $r!$  ordered samples of size  $r$ , by permuting the  $r$  elements in all possible ways.

---

<sup>3</sup>I.e. ways in which the  $n$  objects can be ranked.

The combinatorial coefficient has several properties that are discussed in FELLER, Chapter II. Here we remind only two main important facts:

- The binomial coefficient can be generalised when  $n$  is replaced by any real number. Indeed

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} \quad (3.1)$$

is a valid expression even if  $n \in \mathbb{R}$ . If  $n < k$  is an integer, one of the terms in the numerator is zero, so  $\binom{n}{k} = 0$  for  $n < k$  and  $n$  integer. Yet this is not true if  $n$  is not an integer. So for example<sup>4</sup>

$$\binom{-1/2}{k} = \frac{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{1}{2} - k + 1\right)}{k!} \quad (3.2)$$

$$\begin{aligned} &= \frac{(-1)^k 1 \cdot 3 \cdots (2k-3)(2k-1)}{2^k k!} = \frac{(-1)^k (2k-1)!!}{2^k k!} \\ &= \frac{(-1)^k}{4^k} \binom{2k}{k} \end{aligned} \quad (3.3)$$

is non-zero for  $k > -1/2$ .

- *The binomial theorem*: for any  $a, b \in \mathbb{C}$  and any  $n \in \mathbb{R}$

$$(a+b)^n = \sum_{k=0}^{\infty} \binom{n}{k} a^k b^{n-k}. \quad (3.4)$$

For integer values of  $n$ , the sum in Eq. (3.4) is limited to  $n$ , because  $\binom{n}{k} = 0$  for all  $n, k \in \mathbb{N}$ ,  $k > n$ . This allows us to derive non-trivial identities, such as, for example

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{4^k} \binom{2k}{k} z^k = \sum_{k=0}^{\infty} \binom{-1/2}{k} z^k = \frac{1}{\sqrt{1+z}}.$$

---

<sup>4</sup>Here we define the double factorial

$$(2k-1)!! = 1 \cdot 3 \cdot 5 \cdots (2k-1)$$

as the product of odd integers up to  $2k-1$ . Analogously

$$(2k)!! = 2 \cdot 4 \cdot 6 \cdots (2k) = 2^k k!$$

is the product of all even integers up to  $2k$ . Notice that  $(2k)! = (2k-1)!! \cdot (2k)!!$ .

### 3.1.1 The Stirling's approximation to $n!$

Stirling's formula provides an approximation of  $n!$  for large  $n$  that reads:

$$n! \simeq n^n e^{-n} \sqrt{2\pi n} (1 + O(n^{-1})). \quad (3.5)$$

In Feller II you find a derivation of this result. Here we give a different derivation based on the saddle point method.

Let us start from an important identity

$$n! = \int_0^\infty dx x^n e^{-x} \equiv \Gamma(n+1) \quad (3.6)$$

where the function  $\Gamma(z)$  defined by the second equality (with  $n+1$  replaced by a complex number  $z$ ) is called *gamma function*. Eq. (3.6) indeed provides a generalisation of the factorial for integers (i.e. an analytic continuation) to all complex values of  $n$ . For  $n = 0$  the integral is easily evaluated and we discover that  $0! = 1$ . Eq. (3.6) can be proved to reproduce the factorial  $n! = n \cdot (n-1) \cdots 2 \cdot 1$  because for  $n > 0$  integration by parts yields  $\Gamma(n+1) = n\Gamma(n)$ .

In order to derive Eq. (3.5), observe that the integrand above is maximal for  $x = n$ . Hence set  $x = n(z+1)$  so that

$$n! = n^n e^{-n} n \int_{-1}^\infty dz e^{n[\log(1+z)-z]}$$

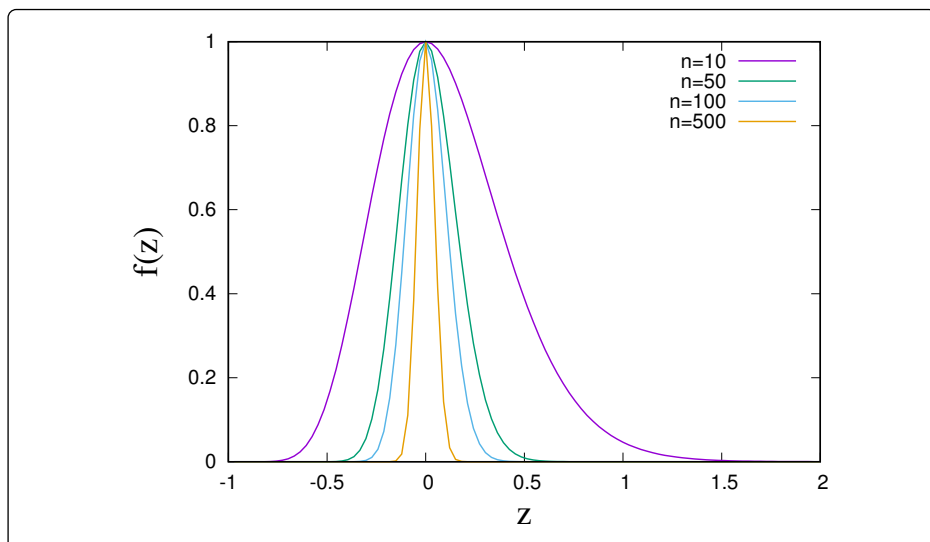
The function in the integral is shown in Figure 5. For  $n$  large, this function is sharply peaked around  $z = 0$ . So the integral is dominated by the region  $z \approx 0$  where  $\log(1+z) - z$  is maximal. This allows us to approximate this function by its power expansion around  $z = 0$ , i.e.  $\log(1+z) - z \simeq -z^2/2 + O(z^3)$ . Then, by making the further change of variables  $\sqrt{nz} = u$ , we find that

$$\begin{aligned} \int_{-1}^\infty dz e^{n[\log(1+z)-z]} &\simeq \frac{1}{\sqrt{n}} \int_{-\sqrt{n}}^\infty du e^{-u^2/2 + u^3/(3\sqrt{n}) + \dots} \\ &\simeq \frac{1}{\sqrt{n}} \int_{-\infty}^\infty du e^{-u^2/2} \left( 1 + \frac{u^3}{3\sqrt{n}} + \dots \right) \end{aligned} \quad (3.7)$$

$$= \sqrt{\frac{2\pi}{n}} (1 + O(1/n)) \quad (3.8)$$

which gives Stirling's approximation, Eq. (3.5).





**Figure 5.** The function  $f(z) = e^{n[\log(1+z)-z]}$  for  $n = 10, 50, 100$  and  $500$ .

### Exercise

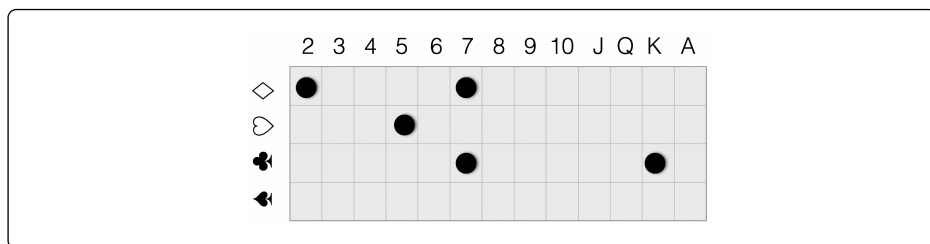
Show that replacing  $-\sqrt{n}$  by  $-\infty$  in the lower limit of integration of Eq. (3.8) involves an error which is of order  $e^{-n/2}/\sqrt{n}$ , and hence is negligible with respect to the leading term of order  $1/n$ . Explain the result.

### Exercise

The first order correction in the Stirling's formula (3.5) should be of order  $1/\sqrt{n}$  according to Eq. (3.8). Yet it turns out to be only of order  $1/n$ . Can you explain why? Can you compute the coefficient of the  $1/n$  correction?

## 3.2 Different ways of counting

What is the probability that no student in a class has his/her birthday on the same day as another one? As already mentioned, this question can be translated into that of random distributions of  $r$  balls (the students) into  $n$  boxes (birthdays,  $n = 365$ ). The probability of this event is then computed counting the number of ways in which the event  $A = \{\text{at most one birthday per day}\}$  can occur. Counting the way in which we can choose the birthdays of Amelie,



**Figure 6.** The number of ways in which a pair at poker can occur can be counted either labelling cards from one to five or counting the ways in which the different boxes in the scheme above can be occupied so as to result in a pair.

George, ..., Carla in such a way to satisfy  $A$ , leads to  $|A| = n(n-1) \dots (n-r+1)$ . It does not matter in which order we take the students, we get always the same number.<sup>5</sup>

What is the probability that we get a pair at poker? We can count in the same way, as you pick up the cards one by one. There are 52 ways in which we can get the first card, to get a pair the second must be one of the three with the same number. The third can be chosen among the 48 with a different number, the fourth in 44 ways and the last in 40 ways. Then we have to consider that the two equal cards can be any, not necessarily the first two. This suggests

$$|A|_1 = \binom{5}{2} \times 52 \times 3 \times 48 \times 44 \times 40 = 131788800$$

where the index 1 refers to the way of counting. We can also count in a different way. There are 13 ways of choosing the number of the pair and  $\binom{4}{2}$  ways of choosing their type. Then there are  $\binom{12}{3}$  ways to choose the numbers of the three remaining cards and  $4^3$  ways in which we can choose their type. Hence

$$|A|_2 = 13 \times \binom{4}{2} \times \binom{12}{3} \times 4^3 = 1098240.$$

We get two different numbers! What is going on? The problem is that we're counting in different ways. In the first, we're considering an ordered sample whereas in the second we're not. So we should apply the same counting when we compute  $|\Omega|$ . In the first case we should take  $|\Omega_1| = (52)_5$  whereas in the second  $|\Omega_2| = \binom{52}{5} = |\Omega_1|/5!$ . Does this fix the problem?

<sup>5</sup>The element of the sample space that we consider is an ordered sample of birthdays  $\omega = (b_1, b_2, \dots, b_r)$ . Yet the order does not matter, i.e. every ordered sample has the same probability. This is why it is enough to compute the number of ordered samples to calculate the probability.

### 3.2.1 Balls in boxes and draws with and without replacement

Take a random number in  $[0, 1)$ . What is the probability that the first five digits are all different? This problem can be stated as that of birthdays, in terms of distributions of  $r$  balls (the digits) into  $n$  boxes (the integers  $0, 1, \dots, 9$ ), and  $A$  is the event where all boxes contain at most one ball. This is also equivalent to drawing  $r = 5$  distinguishable balls (the digits) from an urn with  $n = 10$  balls (the integers  $0, 1, \dots, 9$ ) with replacement, and asking what is the probability that all balls are different.<sup>6</sup> There are  $|\Omega| = n^r$  possible draws with replacement and in  $|A| = (n)_r = n!/(n-r)!$  of them all balls are different. Now  $|A|$  is the number of possible draws without replacement. Hence  $P(A) = (n)_r/n^r = 189/625 = 0.3024$  (for  $r = 5$  and  $n = 10$ ). If  $r = n$ , this probability is  $P(A) = n!/n^n \simeq \sqrt{2\pi n}e^{-n}$  that for  $n = 10$  is already very small (0.00036). So the same problem can be addressed mapping it to different prototype problems of probability.

#### Exercise 3.1

Consider the limiting behaviour of the probability  $P(A)$  for  $n \rightarrow \infty$  when  $r = cn^\alpha$  and  $\alpha < 1$  so that  $n \gg r \gg 1$ . Show that

$$\lim_{n \rightarrow \infty, r = cn^\alpha} P(A) = \begin{cases} 1 & \alpha < 1/2 \\ e^{-c^2/2} & \alpha = 1/2 \\ 0 & \alpha > 1/2 \end{cases}$$

### 3.2.2 Sub-sampling

A lake contains an unknown number  $n$  of fishes. In order to estimate it a sub-population of  $m$  fishes is caught and marked.<sup>7</sup> Then they are released in the lake. In a second catch,  $r$  fishes are caught and  $k$  of them turn out to be marked. If we can compute the probability to find that  $k$  out of  $r$  fishes are marked, as a function of  $n$ , then we can estimate  $n$  by requiring that this probability be as large as possible.

There are two ways to compute this probability. In the first, among all possible ways to draw  $r$  balls without replacement from an urn with  $n$  balls — that are  $(n)_r$  — we are interested in those where  $k$  are of a sub-type (marked) and the rest is not. There are  $\binom{r}{k}$  sequences of draws which result in sub-sets of  $k$  marked and  $r - k$  unmarked balls. Furthermore, there are  $(m)_k$  ways to

<sup>6</sup>Note: balls now are what boxes were in the other case.

<sup>7</sup>This problem is discussed in FELLER, II.6.

draw the  $k$  marked balls and  $(n - m)_{r-k}$  ways to draw the others. Hence:

$$P\{k|n, m, r\} = \binom{r}{k} \frac{(m)_k (n - m)_{r-k}}{(n)_r} = \frac{\binom{m}{k} \binom{n-m}{r-k}}{\binom{n}{r}}.$$

This is called the *Hypergeometric distribution*. In the second method, we invert the order of the argument, starting with the observation that there are  $(n)_m$  ways to choose the  $m$  marked fishes, and proceeding in an analogous manner.

### Exercise 3.2

Complete the argument.

This results in

$$P\{k|n, m, r\} = \frac{\binom{r}{k} \binom{n-r}{m-k}}{\binom{n}{m}}.$$

You can easily check that, in this case, two different ways of counting give the same result (because both are correct).

Imagine that this experiment is done because the number  $n$  of fishes in the lake is unknown and we want to estimate it. Then we can ask which value of  $n$  maximises the log-likelihood<sup>8</sup>  $\log P\{k|n, m, r\}$ . Since

$$\frac{d}{dn} \log P\{k|n, m, r\} \simeq \frac{P\{k|n, m, r\} - P\{k|n - 1, m, r\}}{P\{k|n, m, r\}}$$

we need to find the value of  $n$  for which

$$\frac{P\{k|n, m, r\}}{P\{k|n - 1, m, r\}} = \frac{(n - m)(n - r)}{n(n - m - r + k)} \approx 1$$

which yields  $n \approx rm/k$ . This is a reasonable estimate of the unknown value of  $n$ .

### 3.2.3 Distinguishable and indistinguishable balls in $n$ boxes

Consider again distributions of  $r$  balls in  $n$  boxes. If the  $r$  balls are distinguishable and distributed independently in the boxes, each element of the sample space is defined by the “coordinate”  $x_i$  of ball  $i = 1, \dots, r$ , where  $x_i = 1, \dots, n$

<sup>8</sup> $P\{k|n, m, r\}$  as a function of  $k$  is a probability distribution.  $n, m$  and  $r$  are parameters. As a function of the parameters (e.g. of  $n$ )  $P\{k|n, m, r\}$  is a likelihood. Quoting David McKay: *Never say “the likelihood of the data”. Always say “the likelihood of the parameters”*.

indicates the box containing ball  $i$ . The number of possible outcomes is then  $|\Omega| = n^r$ .

In discussing the physics of gases, where balls are particles and boxes are states, this is called the Maxwell-Boltzmann distribution in statistical physics.

This way of counting treats balls as *distinguishable* because we can attach a label to each of them. For example, a sample point  $\omega$  where ball 5 is in box 1 (i.e.  $x_5 = 1$ ) and ball 2 is in box 7 (i.e.  $x_2 = 7$ ) is different from the one  $\omega'$  that only differs by the interchange of the two balls (i.e. where  $x_5 = 7$  and  $x_2 = 1$  and all other  $x_i$  are the same).

If balls are *indistinguishable* these two sample points cannot be distinguished, i.e.  $\omega = \omega'$ . We cannot attach labels to indistinguishable balls.<sup>9</sup> We can only know how many balls are in each box. So an outcome  $\omega \in \Omega$  is defined in terms of *occupation numbers*  $\omega = (m_1, \dots, m_n)$ , where  $m_k$  specifies how many balls fall in box  $k = 1, \dots, n$ .<sup>10</sup> This is the correct way to count in quantum physics, because quantum particles are indistinguishable. For bosons each state (box) can be occupied by more than one particle (ball) and this leads to Bose-Einstein statistics ( $m_k \geq 0$ ). For fermions, instead, at most one particle can occupy a state ( $m_k = 0, 1$ ). This leads to Fermi-Dirac statistics.

### Exercise 3.3

In how many ways can you put  $r$  indistinguishable particles in  $n \geq r$  indistinguishable boxes?

The number of elements in the sample space in these two cases is, respectively

$$|\Omega_{\text{BE}}| = \binom{n+r-1}{r}, \quad |\Omega_{\text{FD}}| = \binom{n}{r}. \quad (3.9)$$

The second is just the number of ways in which the  $r$  occupied boxes can be chosen out of the  $n$  boxes. The first is slightly more complex to derive. Each element  $\omega \in \Omega_{\text{BE}}$  can be represented as a string of  $r$  balls  $\bullet$  and  $n-1$  sticks  $|$ , which delimitate one box and the next one. For example,

$$\omega = \bullet \bullet \bullet | \bullet \bullet | \dots \bullet | \bullet \bullet$$

<sup>9</sup>Notice the difference between identical and indistinguishable. Even identical balls can be distinguished by attaching labels to them. Indistinguishability means that this is not possible, i.e. that balls do not have an identity.

<sup>10</sup>And clearly

$$\sum_{k=1}^n m_k = r.$$

is a configuration with  $m_1 = 3, m_2 = 1, m_3 = 0, \dots, m_n = 2$ . The number of possible  $\omega$ 's of this type is the number of ways we can combine  $r$  among  $n - 1 + r$  objects, as in Eq. (3.9). This different way of counting gives rise to peculiar phenomena such as an effective repulsion of fermions or an effective attraction of bosons, as compared to classical particles (Maxwell-Boltzmann statistics). Indeed there is no interaction between particles. It's only that in quantum physics we need to count differently.

### Exercise 3.4

In order to see this, compute the probabilities of the events  $A = \{m_k = 1, k \leq r, m_k = 0, k > r\}$  and  $B = \{\exists k; m_k > 1\}$  both in the case of distinguishable (Maxwell-Boltzmann) and of indistinguishable balls, and for the latter for both the Bose-Einstein and the Fermi-Dirac statistics. Show that for  $r = 2$

$$P_{FD}(B) < P_{MB}(B) < P_{BE}(B).$$

as if Bose-Einstein (Fermi-Dirac) particles were subject to an effective attraction (repulsion).

### Exercise 3.5

Which statistics would you use to handle the problem of a single pair at poker?

## 3.3 An extension of the sub-additivity rule

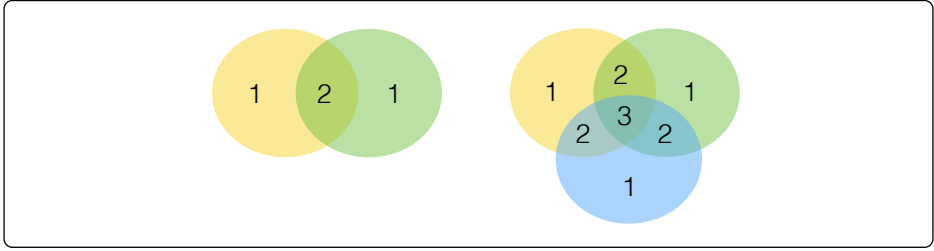
Consider a sequence  $A_1, \dots, A_n$  of  $n$  events. The event  $A_{>0}$  that at least one of the events occurs is

$$A_{>0} = \bigcup_{i=1}^n A_i$$

For any subset<sup>11</sup> of  $k \leq n$  of events, let  $P\{A_{i_1} \cap \dots \cap A_{i_k}\}$  be the probability that all events with indices  $i_1, \dots, i_k$  occur, and let

$$S_k = \sum_{i_1 < i_2 < \dots < i_k} P\{A_{i_1} \cap \dots \cap A_{i_k}\}.$$

<sup>11</sup>For this part, see FELLER IV.



**Figure 7.** The number of times that elementary events  $\omega \in A_{>0}$  are counted in  $S_1$  for  $n = 2$  and  $3$ .

Note that the events in the sum are not disjoint, so we cannot interpret  $S_k$  as a probability. It is just a sum of the probabilities of the joint occurrence of  $k$  events, in all possible ways.

The generalisation of the additivity rule for events that are not necessarily disjoint, is given by:

$$P\{A_{>0}\} = \sum_{\nu=1}^n (-1)^{\nu+1} S_{\nu}. \quad (3.10)$$

Notice that for disjoint events  $S_{\nu} = 0$  for all  $\nu > 1$ . So Eq. (3.10) reduces back to the additivity rule of Kolmogorov's axioms.

As a corollary, the probability that none of the events  $A_i$  occur, is

$$P\{A_0\} = 1 - P\{A_{>0}\} = \sum_{\nu=0}^n (-1)^{\nu} S_{\nu}, \quad (3.11)$$

with the understanding that  $S_0 = 1$ . This is a generalization of the rule  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ . For three events we have

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C),$$

where each term “corrects” the counting of the previous one.

### Exercise 3.6

Consider the case where  $\forall k$ ,

$$P\{A_{i_1} \cap \dots \cap A_{i_k}\} = \prod_{j=1}^k P(A_{i_j}), \quad \forall i_1, \dots, i_k,$$

which, as we shall discuss later, means that the events  $A_i$  are indepen-

dent. Then show that

$$P(A_0) = \prod_{i=1}^n (1 - P(A_i)).$$

Indeed, the proof of Eq. (3.10) is based on the generalization of the intuition used for  $n = 2$  and  $3$ , and is done by counting how many times a sample point  $\omega \in A_{>0}$  “is counted” in the expression on the right hand side of Eq. (3.10). Let  $m$  be the number of events  $A_i$  that contain  $\omega$ . Then  $\omega$  contributes only to terms  $S_\nu$  with  $\nu \leq m$ , and for each of these there are  $\binom{m}{\nu}$  terms in the sum which defines  $S_\nu$  where  $\omega$  contributes. Therefore the total number of times that  $\omega \in A_{>0}$  is counted in the r.h.s. is

$$\sum_{\nu=1}^m \binom{m}{\nu} (-1)^{\nu+1} = 1 - \sum_{\nu=0}^m \binom{m}{\nu} (-1)^\nu = 1 - (1 - 1)^m = 1.$$

Hence each sample point  $\omega \in A_{>0}$  is counted exactly once both on the left and on the right hand side of Eq. (3.10).

### Exercise 3.7

$N$  gentlemen go to theatre each leaving his hat at the wardrobe. On exit they are assigned their hats in a random order. What is the probability that none of the gentlemen get his own hat back? How likely is this event for  $N \rightarrow \infty$ ? (*Hint*: take  $A_i$  as the event that Mr  $i$  gets his hat back).

### Exercise 3.8

Compute the probability of the different hands in 5-cards poker (see table).

Hand	Probability	Number of Hands
Single Pair	0.422569	1098240
Two Pair	0.047539	123552
Triple	0.0211285	54912
Full House	0.00144058	3744
Four of a Kind	0.000240096	624
Straight (excl. Straight Flush and Royal Flush)	0.00392465	10200
Flush (but not a Straight)	0.0019654	5108
Straight Flush (but not Royal)	0.0000138517	36
Royal Flush	0.00000153908	4



**Exercise 3.9**

Using saddle point integration, show that, for  $m \in (-1, 1)$

$$Z(m) = \int_{-\infty}^{\infty} dh \frac{e^{nhm}}{(\cosh h)^n} \simeq \sqrt{\frac{2\pi}{n(1-m^2)}} e^{n[\log 2 - H(m)]} [1 + O(n^{-1})]$$

with

$$H(m) = -\frac{1+m}{2} \log \frac{1+m}{2} - \frac{1-m}{2} \log \frac{1-m}{2}.$$

See FELLER IV and II for more exercises and examples.



## Chapter 4

# Conditional probability and stochastic dependence

Two events  $A, B \in \mathcal{A}$  are independent<sup>1</sup> if the probability of their simultaneous occurrence is the product of the probabilities that each of them occurs:

$$P(A \cap B) = P(A)P(B). \quad (4.1)$$

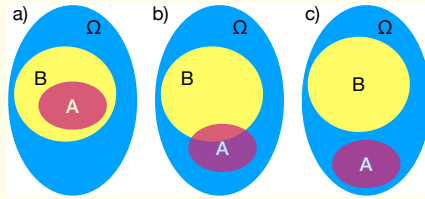
One situation where  $A$  and  $B$  are independent is when it is possible to decompose  $\Omega = \Omega_1 \otimes \Omega_2$  in such a way that  $\forall \omega \in \Omega, \exists \omega_1 \in \Omega_1, \omega_2 \in \Omega_2$  such that  $\omega = (\omega_1, \omega_2)$  and  $p(\omega) = p_1(\omega_1)p_2(\omega_2)$ . Then if  $A = \{\omega_1 \in A_1\}$  only involves conditions on  $\omega_1$  and  $B = \{\omega_2 \in B_2\}$  only involves conditions on  $\omega_2$ , then

$$\begin{aligned} P(A \cap B) &= \sum_{\omega_1 \in A_1} \sum_{\omega_2 \in A_2} p_1(\omega_1)p_2(\omega_2) = \left[ \sum_{\omega_1 \in A_1} p_1(\omega_1) \right] \left[ \sum_{\omega_2 \in A_2} p_2(\omega_2) \right] \\ &= P(A)P(B) \end{aligned}$$

Consider the example of the throw of two dice, i.e.  $\Omega = \{(d_1, d_2), d_i = 1 \dots, 6\}$  and assume that all outcomes are equiprobable:  $P(d_1, d_2) = P(d_1)P(d_2) = 1/36$ . Then the events  $A = \{d_1 = 6\}$  and  $B = \{d_2 = 1\}$  are trivially independent. Yet generally independence might be less evident, and it might not imply a structure on the sample space  $\Omega$ . Indeed independence is a property of  $\mathcal{P}$  and, in general, one needs to compute the probability of the events and of their intersection in order to verify it.

---

<sup>1</sup>This material can be found in FELLER, Chapters V, VIII, IX and X. Here I give a more concise discussion. You can refer to FELLER for a more extended discussion.

**Exercise 4.1**

Can the events in Figures a), b) or c) be independent? What is  $P(B|A)$  in the three cases?

This is illustrated by the example of the two dice above: consider events  $C = \{d_1 + d_2 = 7\}$  and  $D = \{d_1 + d_2 = 8\}$ . Are events A and C independent? And what about events A and D? And what if one of the dice is biased?

In order to have more intuition on what independence means, let us define conditional probability. The probability of event A conditional on the occurrence of event B, is defined as

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)} \quad (4.2)$$

Equivalently, we can write  $P(A \cap B) = P(A|B)P(B)$ , i.e. that the probability that both A and B occur is the probability that B occurs, times the probability that A occurs given that B occurs.<sup>2</sup> In words, if A, B are independent then  $P(A|B) = P(A)$ , i.e. the occurrence of B does not tell us anything on whether A will also occur or not.<sup>3</sup> Notice also that

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

This means that you can compute  $P(A \cap B)$  in either way, starting from the probability of A and then asking what is the probability of B given A, or *vice-versa*. This is also the way in which we think logically.

Conditioning on an hypothesis H, is equivalent to substituting  $\Omega$  with<sup>4</sup> H. Indeed, probabilities should in general be written as  $P(A) = P(A|\Omega)$ . Conditional on H all the rules of probability apply, e.g.

$$P(A \cup B|H) = P(A|H) + P(B|H) - P(A \cap B|H).$$

<sup>2</sup>For short,  $P(A|B)$  is called the probability of A given B.

<sup>3</sup>For the example of two dice above, compute  $P(C|A \cap B)$ . Is this equal to  $P(C)$ ?

<sup>4</sup>Indeed, in classical probability, you can easily check that  $P(A|H) = |A \cap H|/|H|$ . So the number of elements  $|\Omega|$  in the sample space disappears.

For more than two events, we can decompose the joint probability of  $A_1, A_2, \dots, A_n$  by iteratively applying the rule of conditional probability

$$\begin{aligned} P(\cap_{i=1}^n A_i) \\ = P(A_n | A_{n-1} \cap A_{n-2} \cap \dots \cap A_1) P(A_{n-1} | A_{n-2} \cap \dots \cap A_1) \dots P(A_2 | A_1) P(A_1) \end{aligned}$$

This can be useful as sometimes conditional probabilities are easier to compute than joint probabilities. Note that  $P(\cap_{i=1}^n A_i)$  can be expanded in the same way in terms of the conditional probabilities of events  $A_i$  taken in any order.

Independence can be defined for any sequence of events: the events  $A_i$ ,  $i = 1, \dots, n$  are independent if for any subset  $I \subseteq \{1, \dots, n\}$  of indices

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i) \quad (4.3)$$

or equivalently, if for any subset  $I$  and index  $j \notin I$

$$P(A_j | \cap_{i \in I} A_i) = P(A_j).$$

This states that no combination of events  $A_i$  can give information on the likelihood of a different event  $A_j$ .

#### Exercise 4.2

Show that if  $A$  and  $B$  are independent then also  $\bar{A}$  and  $\bar{B}$  are independent. This means that if  $B$  carries no information on  $A$ , then neither its negation does. Show by induction that the same is true for any set  $A_i$  of  $n$  events, i.e. if  $A_i$  are independent, then also their complements are.

It is important to remark that independence is different from pairwise independence, which amounts to

$$P(A_i \cap A_j) = P(A_i)P(A_j), \quad \forall i \neq j,$$

in the sense that pairwise independence does not imply independence.<sup>5</sup> Likewise  $P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$  does not imply independence of the  $n$  events.

#### Exercise 4.3

Find a simple example showing that  $P(A \cap B \cap C) = P(A)P(B)P(C)$  does not imply independence of the three events  $A$ ,  $B$  and  $C$ .

<sup>5</sup>To see this, consider the events  $A$ ,  $B$  and  $C$  in the example above of the two dice. Check that  $A$ ,  $B$ ,  $C$  are pairwise independent but  $P(C | A \cap B) \neq P(C)$ .

Furthermore, notice also that pairwise independence is *not* a transitive property. If  $A$  and  $B$  are independent and if  $B$  and  $C$  are independent, this does not imply that  $A$  and  $C$  are independent.

#### Exercise 4.4

Find an example showing this.

Independence of  $n$  events is a very demanding condition. In order to see this, let us consider a finite sample space  $\Omega$ . For any event  $A \subset \Omega$  the probability

$$P(A) = \sum_{\omega \in A} p(\omega)$$

is a linear combination of the probabilities  $p(\omega)$  of the elements  $\omega \in \Omega$ . Let us take  $n$  events  $A_1, \dots, A_n$  and let us ask whether we can find a probability measure  $\mathcal{P} = \{p(\omega)\}$  such that  $A_1, \dots, A_n$  are independent. The independence of  $n$  events imposes

$$\mathcal{N}_{\text{eq}} = \binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = 2^n - 1 - n \quad (4.4)$$

linear equations on the probabilities  $p(\omega)$ . Barring non-generic choices<sup>6</sup> of the events  $A_i$ , this number needs to be smaller than the number of variables  $p(\omega)$ , which is  $|\Omega| - 1$  (considering normalisation). So the size of the sample space needs to be larger than

$$|\Omega| \geq 2^n - n$$

in order for  $n$  events to be independent. In order to have  $n = 3$  independent events, the sample space needs to contain at least 5 elements, for  $n = 4, 5, 10, 20$  and 100 events we need  $|\Omega| \geq 12, 27, 1014, 1048556$  and  $|\Omega| \geq 1.27 \cdot 10^{30}$ , respectively.

#### Exercise 4.5

Are Eqs. (4.3) really independent? Let  $A_1, A_2$  and  $A_3$  be such that  $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$ . Show that  $A_i$  and  $A_j$  are inde-

<sup>6</sup>Notice that the *sure event*  $\Omega$  is independent of any other event  $A$ . Indeed  $P(A \cap \Omega) = P(A) = P(A)P(\Omega)$ , because  $A \cap \Omega = A$  and  $P(\Omega) = 1$ . This is also true of the impossible event  $\emptyset$ , because  $A \cap \emptyset = \emptyset$  and  $P(\emptyset) = 0$ . All sure or impossible events are independent of all events, because they are true or false no matter what. So they cannot provide information on other events. Notice that two exclusive events  $A \cap B = \emptyset$  cannot be independent as long as they have positive probability.

pendent if  $P(A_k) = P(A_k|A_i \cap A_j)$  for  $i \neq j \neq k = 1, 2, 3$ .

#### Exercise 4.6

Check explicitly if there can be two independent events  $A, B \neq \Omega, \emptyset$  for  $|\Omega| = 3$ ?

Notice that the number of different events  $A \subseteq \Omega$  equals  $2^{|\Omega|}$ , including  $\Omega$  and  $\emptyset$ .<sup>7</sup> Therefore, as the size of the sample space  $|\Omega|$  increases, the number of possible events increases exponentially (as  $\mathcal{N}_A = 2^{|\Omega|}$ ) but only at most  $n \simeq \log_2 |\Omega|$  of these can be simultaneously independent. This suggests that statistical dependence is the norm, whereas independence is the exception.

A useful decomposition of the probability of an event is the following:

**Total probability rule.** Let  $C_i, i = 1, \dots, n$  be a complete set of events. By this we mean that they are exclusive ( $C_i \cap C_j = \emptyset \forall i \neq j$ ) and that

$$\bigcup_{i=1}^n C_i = \Omega$$

Then, for any event  $A \subset \Omega$ , we can decompose its probability as

$$P(A) = \sum_{i=1}^n P(A|C_i)P(C_i). \quad (4.5)$$

The proof is nothing but the application of the additivity axiom. We can consider  $C_i$  as causes, and hence decompose the probability of  $A$  into that of  $A$  conditional on the occurrence of each cause  $C_i$ . This rule is useful, because it makes it possible to compute  $P(A)$  once one finds a suitable set of “causes”  $C_i$  for which  $P(A|C_i)$  and  $P(C_i)$  are easy to compute.

**Bayes theorem of causes.** In other circumstances, we are interested in the inference of the probability of a particular cause  $C_i$  given that we know that an event  $A$  has occurred. This is given by Bayes theorem

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{\sum_j P(A|C_j)P(C_j)} \quad (4.6)$$

<sup>7</sup>Each event can be represented by a sequence  $(a_1, \dots, a_{|\Omega|})$  where, each  $a_\omega$  can be chosen in two ways:  $a_\omega = 1$  if  $\omega \in A$  and  $a_\omega = 0$  otherwise. So the number of sequences is  $2^{|\Omega|}$ .

This is the basis of statistical inference:  $C_i$  may represent alternative theories and  $A$  an experimental observation.  $P(A|C_i)$  can be computed from the theory. Yet the interesting quantity is  $P(C_i|A)$  that quantifies the probability that theory  $C_i$  is correct given the observation of  $A$ . Bayes theorem tells us how to compute it. In statistics  $P(A|C_i)$  is called the *likelihood* (of  $C_i$ ),  $P(C_i)$  is the *prior*, i.e. the probability of  $C_i$  before observing event  $A$ , and  $P(C_i|A)$  is called the *posterior*, i.e. the updated probability of  $C_i$  after event  $A$  has been observed. The denominator  $P(A) = \sum_j P(A|C_j)P(C_j)$  is called the *evidence*.

## 4.1 Statistical dependence is not causation

It is important to notice that if  $A$  depends on  $B$  then we cannot conclude that  $A$  causes  $B$ , nor *viceversa*. Writing  $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$  does not implies that  $C$  causes  $B$  and  $B, C$  cause  $A$ . Indeed  $P(A \cap B \cap C)$  can be written also as  $P(B|A \cap C)P(C|A)P(A)$  or in four other ways, each of which refers to a different permutation of the three events. One of these ways may reflect a causal structure, but this is not necessarily the case. Indeed, there may be another event  $D$  that is causing all of them.<sup>8</sup> There is a whole field of causal inference which deals with this issue. For the moment, it is important to stress that statistical dependence  $P(A \cap B) \neq P(A)P(B)$  is about *observational* probabilities, that concerns the probability of simultaneous occurrence of  $A$  and  $B$ . No causal conclusion can be drawn from statistical dependence.

### Exercise 4.7

Let there be three coins, one with head on both sides, one with tail on both sides and the other with head on one side and tail on the other. If you see one of these coins on a table with the upward face which is head, what is the probability that the other face is also head?

### Exercise 4.8

The next upgrade of the machine at Cern is going to explore an higher energy range. Skeptics say that as soon as the new machine is turned on, a black hole will form at Cern and the planet will collapse (so the new

<sup>8</sup>Imagine that we observe an increase in the price of butter and later on an increase in the price of cheese. Can we infer that the former causes the latter? Not in general. Indeed both may be caused by the increase in the price of milk, because it takes less time to make butter from milk than cheese. Statistics shows that countries with a higher consumption of chocolate also receive more Nobel prizes, and that fertility is higher in regions of Germany where storks are more abundant. Does this mean that eating chocolate is a good strategy for winning the Nobel prize or that babies are brought by storks?



machine should not be turned on!). Scientists reply that, according to our present theories, this event has zero probability. Is this a satisfactory answer?

### Exercise 4.9

Consider families with two children  $\Omega = \{bb, bg, gb, gg\}$  where the first (second) character stands for the sex (boy or girl) for the elder (younger) child. Imagine that all four possibilities occur with the same probability  $P(\omega) = 1/4$ .

- Given that a family has a boy, what is the probability that the other child is also a boy?
- Given that the older child of a family is a boy, what is the probability that the younger is also a boy?
- Given a boy taken at random what is the probability that the other child in his family is also a boy?
- Given that a family has a boy who was born on Tuesday, what is the probability that the other child is also a boy?

### Exercise 4.10

The Monty Hall problem: suppose you're on a game show, and you're given the choice of three doors: behind one door is a prize; behind the others there is nothing. You pick a door and the host, who knows what's behind the doors, opens another door, which is empty. He then asks: "Do you want to pick the other door?" Is it to your advantage to switch your choice?

### Exercise 4.11

Let  $A$  and  $B$  be two event. Show that the probability that both of them occur, given that at least one of them occurs, is smaller than the probability that both of them occur given that you know which of the two events occurs.



## Chapter 5

# Random variables

A random variable is not a variable. It is a function.

A random variable (RV) is a function<sup>1</sup>

$$X : \Omega \rightarrow \mathbb{F} \quad (5.1)$$

$$: \omega \rightarrow X(\omega) \in \mathbb{F} \quad (5.2)$$

where  $\mathbb{F}$  is a field, for example the real numbers  $\mathbb{R}$  (real RV), the integers  $\mathbb{N}$  (integer RV), the complex plane  $\mathbb{C}$ , etc.

We assume that statements like  $X \in [a, b]$  that concern the random variable  $X$  are all events that belong to  $\mathcal{A}$ . Then the definition of  $\mathcal{P}$  on  $\mathcal{A}$  induces a *probability distribution* on the values that the random variable takes. For variables defined on a set  $\mathbb{F}$  of finite or countably many elements  $x$  the probability distribution is defined as

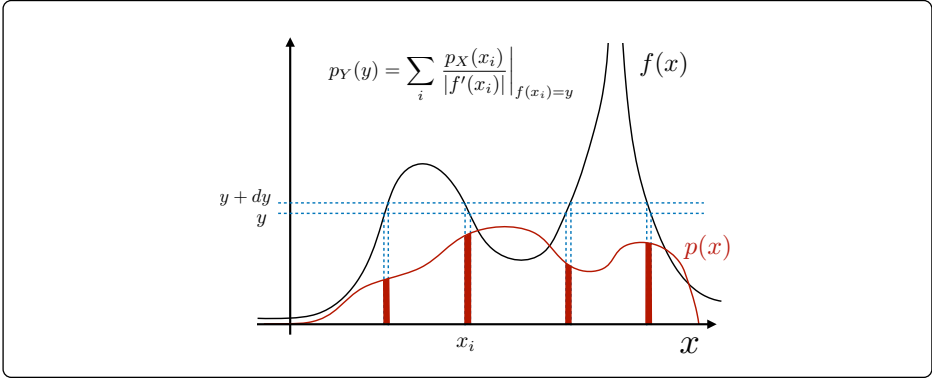
$$p_x = P\{X(\omega) = x\} \equiv P\{\omega \in \Omega : X(\omega) = x\} \quad (5.3)$$

where  $p_x \geq 0$  and  $\sum_x p_x = 1$  by normalisation. For real RV, the probability distribution is defined as follows. For any interval  $[a, b] \subset \mathbb{R}$ , we define

$$P\{X(\omega) \in [a, b]\} = \int_a^b dx p(x)$$

---

<sup>1</sup>As a general rule, we shall use uppercase letters for random variables and the corresponding lowercase letter for the values they take. When not needed, we shall suppress the dependence on  $\omega$ . Yet it is important that you always remember that random variables are functions, not variables.



**Figure 8.** A graphical representation of the relation between the pdf of  $X$  and that of  $Y = f(X)$ .

where  $p(x)$  is called *probability density function* (pdf). More precisely, the pdf of a continuous random variable  $X$  is defined as

$$p(x) = \lim_{dx \rightarrow 0} \frac{1}{dx} P\{X(\omega) \in [x, x + dx)\}. \quad (5.4)$$

The term “density” in pdf appears because  $p(x)$  is *not* the probability of the event  $\{X(\omega) = x\}$ . The probability of this event is zero for any  $x \in \mathbb{R}$ .<sup>2</sup> Eq. (5.4) states that the probability to find a random variable in an interval  $[x, x + dx)$ , for an infinitesimally small  $dx$ , is  $p(x)dx$ .

The cumulative distribution is defined as<sup>3</sup>

$$P\{X(\omega) < x\} = \int_{-\infty}^x dx' p(x') \equiv P(x).$$

The normalisation implies

$$\int_{-\infty}^{\infty} p(x) dx = \lim_{x \rightarrow \infty} P(x) = 1. \quad (5.5)$$

<sup>2</sup>This does not mean that it is impossible for a real random variable to take any value  $x$ , of course.

<sup>3</sup>**Notation:** when  $X$  is a discrete random variable, we use the values  $x$  that it takes as indices for the probability, as in Eq. (5.3). When  $X$  is continuous we use  $x$  as the argument of either the pdf Eq. (5.4) or of the cumulative distribution. For the pdf we use lowercase letters — as  $p$  in Eq. (5.4). We shall use uppercase letters for cumulative distributions.

**Change of variables:** let  $f(x)$  be a monotonic function. Then the pdf of the variable  $Y(\omega) = f[X(\omega)]$  is

$$p_Y(y) = \frac{p_X(x)}{|f'(x)|} \Big|_{x=f^{-1}(y)} \quad (5.6)$$

this merely reflect the fact that the probability of corresponding intervals<sup>4</sup>  $[x, x + dx]$  and  $[y, y + dy]$ , with  $y = f(x)$  and  $dy = f'(x)dx$ , must be the same, i.e.  $p_X(x)dx = p_Y(y)dy$ . This can be extended to non-monotonic functions  $f(x)$ , splitting the domain of  $x$  in subdomains where  $f(x)$  is monotonic and adding all the contributions to  $p_Y(y)$  that come from each domain (see Figure 8).

The factor  $|f'(x)|$  in the denominator Eq. (5.6) appears because the pdf is not a probability, but a probability density. If for example  $X(\omega)$  is a length, then the pdf has dimensions of inverse length. Probability are numbers, so they are adimensional. The pdf of  $X$  is not adimensional, it has dimensions of  $1/X$ .

## 5.1 Many random variables

The same definition extends to the case of  $n$  random variables  $\underline{X}(\omega) = (X_1, \dots, X_n)$ , where each component  $X_i(\omega)$  is a random variable. The *joint* pdf is defined by

$$p(\underline{x})d\underline{x} = P\{X_i(\omega) \in [x_i, x_i + dx_i], i = 1, \dots, n\}, \quad (5.7)$$

where  $d\underline{x} = dx_1 dx_2 \dots dx_n$ . If we're interested only in the distribution of one of the random variables, say  $X_1(\omega)$ , we can derive its pdf integrating over all other random variables:

$$p_{X_1}(x) = \int_{-\infty}^{\infty} dx_2 \dots \int_{-\infty}^{\infty} dx_n p(x, x_2, \dots, x_n). \quad (5.8)$$

This is called the *marginal* distribution of  $X_1$ . Likewise we can compute the marginal distribution of any subset of random variables by integrating the joint distribution on all the others.

We can also define the distribution of a variable, say  $X_1$ , conditional to other variables, say  $X_2$ . In order to do this, we need to consider the events

<sup>4</sup>This refers to an increasing function. For a decreasing function  $dy < 0$ , so the interval on  $Y$  is between  $y + dy$  and  $y$ . This is the reason of the absolute value in Eq. (5.6).

$\{X_1(\omega) \in [x_1, x_1 + dx_1]\}$  and  $\{X_2(\omega) \in [x_2, x_2 + dx_2]\}$  and apply the rules of conditional probability. Then

$$\begin{aligned} P\{X_1(\omega) \in [x_1, x_1 + dx_1] | X_2(\omega) \in [x_2, x_2 + dx_2]\} &= \frac{p(x_1, x_2) dx_1 dx_2}{p(x_2) dx_2} \\ &\equiv p(x_1 | x_2) dx_1, \end{aligned}$$

hence the pdf of  $X_1$  conditional to  $X_2$  is:

$$p(x_1 | x_2) = \frac{p(x_1, x_2)}{p(x_2)}. \quad (5.9)$$

This generalises in obvious ways to the joint pdf of any subset of variables conditional to another subset of variables.

Two random variables  $X_1$  and  $X_2$  are independent if their joint pdf factorises, i.e. if  $p(x_1, x_2) = p(x_1)p(x_2)$ .  $n$  random variables are mutually independent if

$$p(x_{i_1}, x_{i_2}, \dots, x_{i_\ell}) = \prod_{k=1}^{\ell} p(x_{i_k}) \quad (5.10)$$

for any  $\ell = 2, 3, \dots, n$  and for any indices  $i_1 < i_2 < \dots < i_\ell$ . Again pairwise independence (i.e.  $p(x_i, x_j) = p(x_i)p(x_j)$  for all  $i < j$ ) is not enough to ensure mutual independence among  $n$  random variables. When the sample space is finite, the same argument that we invoked before to estimate the maximal number of independent events suggests that the maximal number of independent variables is, generically, upper bounded by  $\log_2 |\Omega|$ .

### Exercise 5.1

What is your estimate of the maximal number of pairwise independent random variables defined on a sample space with a finite number  $|\Omega|$  of elements?

### Exercise 5.2

Show that it is possible to obtain a constant as a linear combination of  $n$  random variables  $X_i(\omega)$  when  $n = |\Omega| < +\infty$ . In other words, in this case there are combinations of random variables that are not random at all. [Hint: you can think of a random variable as a vector in an  $|\Omega|$  dimensional space.]

## 5.2 Examples of random variables

**A binary random variable.** The simplest random variable takes just two values:  $X(\omega) \in \{0, 1\}$ . The distribution is given by

$$p = P\{\omega : X(\omega) = 1\} \text{ and } P\{\omega : X(\omega) = 0\} = 1 - p.$$

**Bernoulli trials.** Consider the repetition of  $n$  independent experiments (trials), each of which results in either a success or a failure. This setup corresponds to the case of  $n$  independent binary random variables  $X_k$ ,  $k = 1, \dots, n$ , where  $X_k = 1$  stands for success in the  $k^{\text{th}}$  trial and  $X_k = 0$  for failure.<sup>5</sup> As before  $P\{\omega : X_k(\omega) = 1\} = p$  for all  $k$ . Then one can define the random variable  $S_n = \sum_{k=1}^n X_k$  which is the number of successes in  $n$  trials. The probability

$$B(k|n, p) = P\{S_n = k\} = \binom{n}{k} p^k (1-p)^{n-k} \quad (5.11)$$

is called *binomial distribution* easily computed observing that all outcomes  $(X_1, X_2, \dots, X_n)$  with  $k$  successes have probability  $p^k (1-p)^{n-k}$  and there are  $\binom{n}{k}$  of them.

**Multinomial distribution.** The binomial distribution naturally generalises to cases where the outcomes  $X_i$  of each of the  $n$  trials can be more than two. If each trial corresponds to a random variable  $X_i = 1, 2, \dots, d$  that can take  $d$  different values, and  $P\{X_i = \ell\} = p_\ell$  (with  $\sum_\ell p_\ell = 1$ ), then the number of times the outcome  $X_i = \ell$  is observed across trials<sup>6</sup>

$$K_\ell = \sum_{i=1}^n \delta_{X_i, \ell},$$

is a random variable with distribution<sup>7</sup>

$$P\{K_\ell = k_\ell, \ell = 1, \dots, d\} = \frac{n!}{\prod_{\ell=1}^d k_\ell!} \prod_{\ell=1}^d p_\ell^{k_\ell}, \quad \sum_{\ell=1}^d k_\ell = n. \quad (5.12)$$

Again the probability that in a sequence  $X_1, \dots, X_n$  there are  $k_\ell$  variables with the value  $X_i = \ell$  is  $p_\ell^{k_\ell}$ . This accounts for the second factor. The

<sup>5</sup>In this case the sample space  $\Omega = \bigotimes_{k=1}^n \Omega_k$  is the direct product of the sample spaces for each trial and  $X_k : \Omega_k \rightarrow \{0, 1\}$  depends only on the outcome of trial  $k$ .

<sup>6</sup>Here  $\delta_{i,j}$  is the Kroneker delta, which is one if  $i = j$  and zero otherwise. In this equation it is used to count the number variables  $X_i$  which are equal to  $\ell$ .

<sup>7</sup>Note that the random variables  $K_\ell$  are not independent, because  $\sum_{\ell=1}^d K_\ell = n$ .

combinatorial factor accounts for the number of sequences that satisfy this condition for each  $\ell$ .

### Exercise 5.3

Let  $p_i$  be the probability that a child of a family was born on day  $i$  of the year ( $i = 1, 2, \dots, n = 365$ ). Compute the probability that the two children of a family have their birthday on different days. Show that this is maximal when  $p_i = 1/n$ . Show that, in a family of three children the probability that all have different birthdays is also maximal when  $p_i = 1/n$ . Argue that this should be true for a family with any number  $r$  of children and try to prove it.

### Exercise 5.4

Show that, if  $\{K_1, \dots, K_d\}$  follows a multinomial distribution with parameters  $p_1, \dots, p_d$ , then for any  $\ell = 1, \dots, d$  the marginal distribution of  $K_\ell$  is given by the binomial distribution with parameter  $p_\ell$ , over  $n$  trials. Is this what you would expect?

**Poisson distribution.** Often one is interested in the same problem as above, but in the limit where  $n \rightarrow \infty$ ,  $p \rightarrow 0$  with  $np = \lambda$  fixed. The typical example is the decay of nuclei in a sample of radioactive material. Each of these events can occur with the same probability in an infinitesimal time interval  $dt$ , and the probability  $p = rdt$  of this to occur is proportional to  $dt$ . In a fixed time interval of duration  $T$  there are  $n = T/dt$  such intervals. In this case,  $\lambda = rT$  where  $r$  is the rate of decay per unit time. The probability of observing  $k$  events is given by the Poisson distribution

$$P(k|\lambda) = \lim_{n \rightarrow \infty} B(k|n, \lambda/n) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (5.13)$$

which is derived from the binomial distribution Eq. (5.11) in a straightforward manner.

### Exercise 5.5

1. Derive Eq. (5.13) and check that Eq. (5.13) is correctly normalised.
2. What is the same limit of the multinomial distribution Eq. (5.12) when  $n \rightarrow \infty$  with  $p_\ell = \lambda_\ell/n$ .



### The Geometric and the Negative Binomial distributions (waiting

**times).** In a sequence of Bernoulli trials, one can ask the question of how many trials one should “wait” before observing the first success. This is an integer random variable — a *waiting time* —  $T$  whose distribution is

$$P\{T = k + 1\} = p(1 - p)^k. \quad (5.14)$$

This is because there need to be  $k$  failures (which have probability  $(1 - p)^k$ ) in order for the first success (which accounts for the factor  $p$ ) to occur at trial  $T = k + 1$ . Eq. (5.14) is called the *geometric distribution* and it is a special case of the *negative binomial distribution*

$$P\{T = k + r\} = \binom{-r}{k} p^r (1 - p)^k = \binom{r + k - 1}{k} p^r (1 - p)^k, \quad (5.15)$$

for  $r = 1$ . This is evident from the second expression, that is obtained from the first by a simple manipulation of binomial coefficients. The negative binomial gives the probability that the  $r^{\text{th}}$  success in a sequence of Bernoulli trials occurs at “time”  $T = r + k$ . In this case,  $k$  is the number of failures and there can be  $\binom{r+k-1}{k}$  ways in which they can occur before the  $r^{\text{th}}$  success.<sup>8</sup> Waiting times are a useful concept that we shall use frequently in what follows.

#### Exercise 5.6

What is the probability that  $T \geq k + k_0 + r$  given that  $T \geq k_0 + r$  if  $T$ 's distribution is given by Eq. (5.15)? Show that for  $r = 1$  this equals the probability that  $T \geq k + r$ . This means that knowing that  $T \geq k_0 + r$  does not make the event that we should wait  $k$  more steps for the first success any more or less likely. This is a sign of lack of memory in the Bernoulli trial process — i.e. knowing what has happened up to a certain point does not affect what will happen in the future. Show that for  $r > 1$  this is not so. Is the event  $T \geq k + k_0 + r$  given that  $T \geq k_0 + r$  more or less likely than  $T \geq k + r$  for  $r > 1$ ?

**The Gaussian distribution.** The pdf of a Gaussian variable  $X(\omega) \in \mathbb{R}$  is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad (5.16)$$

<sup>8</sup>See FELLER VI.8 for more details.

where<sup>9</sup>  $m$  (the mean) and  $\sigma$  (the standard deviation) are real parameters. Clearly  $\sigma > 0$ .

An important property of Gaussian variables is that the sum of two independent Gaussian variables  $Z = X + Y$  is also a Gaussian variable with mean  $\mu_Z = \mu_X + \mu_Y$  and variance  $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$ , where  $\mu_X, \mu_Y$  and  $\sigma_X$  and  $\sigma_Y$  are the mean and standard deviations of  $X, Y$ , respectively. This has important consequences.

### Exercise 5.7

Prove this.

The de Moire-Laplace theorem shows that the Binomial distribution, in the limit  $n \rightarrow \infty$ , is asymptotically equal to

$$B(k|n, p) \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}} \quad (5.17)$$

this means that, for large  $n$ , a binomial random variable  $S_n$  is well approximated<sup>10</sup> by a Gaussian random variable with parameters  $m = np$  and  $\sigma = \sqrt{np(1-p)}$ .

### Exercise 5.8

Derive Eq. (5.17) using Stirling's formula.

**The Multivariate Gaussian distribution.** The Gaussian distribution generalizes to a vector  $\underline{X}(\omega) = (X_1, \dots, X_n)$  of  $n$  random variables in the following manner

$$p(\underline{x}) = \sqrt{\frac{\det A}{(2\pi)^n}} e^{-\frac{1}{2}(\underline{x}-\underline{m})' \hat{A}(\underline{x}-\underline{m})} \quad (5.18)$$

where  $\underline{m} \in \mathbb{R}^n$  is a vector and  $\hat{A}$  is an  $n \times n$  positive definite symmetric matrix (prime denotes transpose). This is called the multivariate Gaussian distribution.

<sup>9</sup>To check that Eq. (5.16) is correctly normalised, you can use the fact that, by a change variables to polar coordinates,

$$\left[ \int_{-\infty}^{\infty} dz e^{-\frac{z^2}{2}} \right]^2 = \int_0^{2\pi} d\theta \int_0^{\infty} r e^{-\frac{r^2}{2}} dr = 2\pi.$$

<sup>10</sup>We could equivalently say that a binomial random variable converges to a Gaussian when  $n \rightarrow \infty$ . Yet we shall discuss later what “converges” means for random variables.

**Exercise 5.9**

Using the spectral decomposition of the matrix  $\hat{A}$  in eigenvalues and eigenvectors, prove that Eq. (5.18) is correctly normalised.

Notice that if  $\hat{A}$  is diagonal, i.e. if its matrix elements  $A_{i,j}$  vanish for all  $i \neq j$ , then the  $n$  random variables  $X_i$  are independent. However, there is a linear combination  $\underline{Y} = \hat{V}(\underline{X} - \underline{m})$  of the  $n$  variables that transforms them into a vector of independent variables  $Y_i$ . Indeed, if one takes  $\hat{V}'$  as the matrix<sup>11</sup> of eigenvectors of  $\hat{A}$ , one can write  $\hat{A} = \hat{V}'\hat{\Lambda}\hat{V}$ , where  $\hat{\Lambda}$  is the diagonal matrix of eigenvalues. Then one can rewrite Eq. (5.18) as

$$p(\underline{y}) = \prod_{i=1}^n \sqrt{\frac{\lambda_i}{2\pi}} e^{-\frac{\lambda_i}{2} y_i^2}$$

which is the joint distribution of  $n$  independent random variables  $Y_i$ . In words, any multivariate Gaussian distribution can be transformed into the distribution of  $n$  independent Gaussian variables by a “simple rotation”.

**The uniform distribution.** The RV  $X(\omega) \in [0, 1]$  with  $p(x) = 1$  for  $x \in [0, 1]$  and  $p(x) = 0$  otherwise is called a uniform random variable. The random number generator in your computer *simulates* realisations of uniform random variables.<sup>12</sup> Therefore, this is your starting point for generating pseudo-random variables with any continuous pdf  $p(x)$ . One of the methods relies on the fact that for any random variable  $X$  with pdf  $p(x)$ , the random variables

$$U(\omega) = \int_{-\infty}^{X(\omega)} p(x) dx \quad (5.19)$$

is a uniform random variable (please check). If this relation can be inverted to find  $X$  as a function of  $U$ , then it can be used to generate random variables with pdf  $p(x)$  starting from a pseudo-random number generator of uniform random variables  $U$ .

**The exponential distribution** applies to RV  $X(\omega) \in [0, \infty)$  with  $p(x) = ae^{-ax}$  for  $x \geq 0$  and  $p(x) = 0$  for  $x < 0$ .

<sup>11</sup>Where the prime indicates matrix transpose.

<sup>12</sup>The book Numerical Recipes ([11], also available online) provides a practical and concise discussion of how this is done. Please read it.

**Exercise 5.10**

How would you generate an exponential random variable using Eq. (5.19)?

This arises in problems related to waiting times. Consider a process like the decay of a certain radioactive sample. The probability of the event  $A_{t,dt}$  that an  $\alpha$  particle is emitted in any interval  $[t, t + dt)$  is independent of  $t$  by time translation invariance. If we also assume that the occurrence of a decay at time  $t$  does not provide any information about decays at later times then,  $P(A_{t,dt})$  must also be proportional to  $dt$ , i.e.,  $P(A_{t,dt}) = a dt$ . The probability that  $k$  events occur in  $[t, t + \tau)$  is then a Poisson distribution with parameter  $a\tau$ .

**Exercise 5.11**

Can you demonstrate explicitly the last two sentences?

Then the probability that no event occur in  $[t, t + \tau)$  is  $e^{-a\tau}$ . This is also the probability that the time  $T$  one has to wait for the next event is larger than  $t$ , i.e.,  $P\{T > \tau\} = e^{-a\tau}$ . Therefore the pdf of  $T$  is

$$p(t) = -\frac{d}{dt}P\{T > t\} = ae^{-at}. \quad (5.20)$$

A characteristic of exponential random variables is that it describes *memory-less* processes. This is well explained by the fact that, if buses arrive at your bus stop at random times, and the time you have to wait has an exponential distribution, it makes no sense to ask to by-standers how long they have been waiting. That information will not tell you anything on whether the next bus will arrive sooner or later.

**Exercise 5.12**

To show this, compute the conditional probability that you will have to wait at least  $t$  more minutes, given that someone at the bus stop has seen no bus arriving in the past  $\tau$  minutes. Compare this with the unconditional probability that you will have to wait at least  $t$  more minutes.

## 5.3 Expectation

The expected value of a random variable  $X$  is defined as

$$\mathbb{E}[X] = \int dx p(x)x \quad \text{or} \quad \mathbb{E}[X] = \sum_x p_x x$$

for real or discrete random variables, respectively. In general the expectation operator  $\mathbb{E}[\cdot]$  is defined on any functions and combinations of RV. E.g.

$$\mathbb{E}[f(X)] = \int dx p(x)f(x)$$

It is a linear operator  $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$  for any constants  $a, b$  and RV  $X(\omega)$  and  $Y(\omega)$ . The  $n^{\text{th}}$  moment of  $X(\omega)$  is defined as

$$M_n = \mathbb{E}[X^n] = \int dx p(x)x^n$$

In particular the first moment  $\mathbb{E}[X]$  — the *mean* — gives a measure of the value around which  $X(\omega)$  is distributed and the variance

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

gives a measure of the variability of  $X$ , because the standard deviation  $\sigma[X] = \sqrt{\mathbb{V}[X]}$  quantifies the dispersion of  $X(\omega)$  around its expected value.

### Exercise 5.13

Compute the expected value of the waiting time of an exponential random variable with pdf given by Eq. (5.20). Now go back to the bus stop problem. If a bus comes on average every 10 minutes, how much time do I expect that I will have to wait?

### Exercise 5.14

Compute the mean and the variance for the binomial (Eq. (5.11)), the Poisson (Eq. (5.13)) and the Gaussian distribution (Eq. (5.16)). How is the Poisson distribution special?

The expected value can also be defined for more than one random variable, for example:

$$\mathbb{E}[X_1 X_2] = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 p(x_1, x_2) x_1 x_2. \quad (5.21)$$

In general, the expected value of a function of  $n$  random variables  $X_1, \dots, X_n$  is given by

$$\mathbb{E}[f(X_1, \dots, X_n)] = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n p(x_1, \dots, x_n) f(x_1, \dots, x_n).$$

In other words, every random variable on which the expectation is taken needs to be integrated over with the corresponding probability. The expected value can be decomposed in conditional expected values, just like the joint distribution  $p(x_1, \dots, x_n)$  can be decomposed in conditional distributions. For example, for two variables

$$\mathbb{E}[f(X_1, X_2)] = \int_{-\infty}^{\infty} dx_1 p(x_1) \mathbb{E}[f(X_1, X_2) | X_1 = x_1]$$

where  $\mathbb{E}[f(X_1, X_2) | X_1 = x_1] = \int_{-\infty}^{\infty} dx_2 p(x_2 | x_1) f(x_1, x_2)$  is the expected value conditional to  $X_1 = x_1$ . So we can write<sup>13</sup>

$$\mathbb{E}[f(X_1, X_2)] = \mathbb{E}[\mathbb{E}[f(X_1, X_2) | X_1]]$$

where the inner conditional expected value is taken with respect to  $X_2$  with  $X_1$  fixed, and the outer one with respect to  $X_1$ .

### Exercise 5.15

The tower property of conditional expectation is

$$\mathbb{E}[\mathbb{E}[f(X, Y, Z) | X, Y] | X] = \mathbb{E}[f(X, Y, Z) | X].$$

Show it and interpret it.

Going back to Eq. (5.21), if  $X_1$  and  $X_2$  are independent, then the joint pdf factorises  $p(x_1, x_2) = p_1(x_1)p_2(x_2)$  and the integrals also factorise. Therefore if  $X_1$  and  $X_2$  are independent<sup>14</sup>

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2].$$

Notice that the converse is not true, i.e.  $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2]$  does not imply that  $X_1$  and  $X_2$  are independent.

<sup>13</sup>The expected value  $\mathbb{E}[f(X_1, X_2)]$  is a number, but  $\mathbb{E}[f(X_1, X_2) | X_1]$  is a RV, because it is a function of the random variable  $X_1$ .

<sup>14</sup>Note that  $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2]$  if  $X_1$  (or  $X_2$ ) is a constant. A random variable which is constant is independent of any other random variable.

**Exercise 5.16**

Find a simple counter-example with  $X_1$  and  $X_2$  that take values in  $\{-1, 0, 1\}$ , where

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2]$$

but  $X_1$  and  $X_2$  are not independent.

The same consideration applies to expected values of functions of more than one random variable: the expected value of the product of  $n$  independent random variables  $X_1, \dots, X_n$  equals the product of the expected values, or more generally

$$\mathbb{E}[f_1(X_1)f_2(X_2)\cdots f_n(X_n)] = \mathbb{E}[f_1(X_1)]\mathbb{E}[f_2(X_2)]\cdots\mathbb{E}[f_n(X_n)],$$

for any set of functions  $f_i(x)$ . But the converse is not true in general.

The covariance between two variables is defined as

$$\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]$$

whereas the correlation is defined as

$$\text{Corr}(X_1, X_2) = \frac{\mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]}{\sqrt{\mathbb{V}[X_1]\mathbb{V}[X_2]}}.$$

One important point to keep in mind is that if  $X_1$  and  $X_2$  are independent, then they are also uncorrelated, i.e.  $\text{Cov}(X_1, X_2) = 0$ , but the converse is not true, unless the variables have a multivariate Gaussian distribution.

**Exercise 5.17**

Prove that if  $X_1$  and  $X_2$  have a multivariate Gaussian distribution and  $\text{Cov}[X_1, X_2] = 0$  then they are also independent. Find a counter-example that shows that this is not true in general.

**Exercise 5.18**

Show that the covariance matrix  $C_{i,j} = \text{Cov}(X_i, X_j)$  of a ensemble of  $n$  random variables  $\underline{X} = (X_1, \dots, X_n)$ , is a non-negative definite matrix, whatever is their distribution  $p(\underline{X})$ . (*Hint*: the variance of any linear combination

$$U = \sum_{i=1}^n u_i X_i$$

cannot be negative).

Other examples of the use of the expected value that we shall use frequently in the sequel are the *generating function*  $g(s) = \mathbb{E}[s^X]$  for integer valued random variables, and the *characteristic function*  $\phi(k) = \mathbb{E}[e^{ikX}]$  for real random variables.

## 5.4 Correlation and factor graphs\*

Imagine that a variable  $Z$  is *caused* by  $X$  and  $Y$ , where  $X$  and  $Y$  are two independent causes. One way to think about this is that there is a “mechanism”, such that  $Z$  is a function of  $X$  and  $Y$ . One way to write this is  $Z = f(X, Y, U)$  where  $U$  is an unobserved independent random variable. Then, among the six ways in which we can write the joint pdf  $p(x, y, z)$  in terms of conditional pdf, there is one

$$p(x, y, z) = p(z|x, y)p(x|y)p(y) = p(z|x, y)p(x)p(y)$$

that reflects this causal structure. Notice that this is the only way in which the conditional dependence is simplified, because  $p(x|y) = p(x)$ . For example, in  $p(x, y, z) = p(y|x, z)p(z|x)p(x)$  the term  $p(y|x, z)$  is different from

$$p(y|z) = \int dx p(y|x, z)p(x) \quad (5.22)$$

because even if  $X$  and  $Y$  are independent, conditioning on  $Z$  introduces a statistical dependence between them. This allows us to *identify* the causal relation, i.e. to say that  $X$  and  $Y$  cause  $Z$  and not that  $Y$  is caused by  $X$  and  $Z$ .

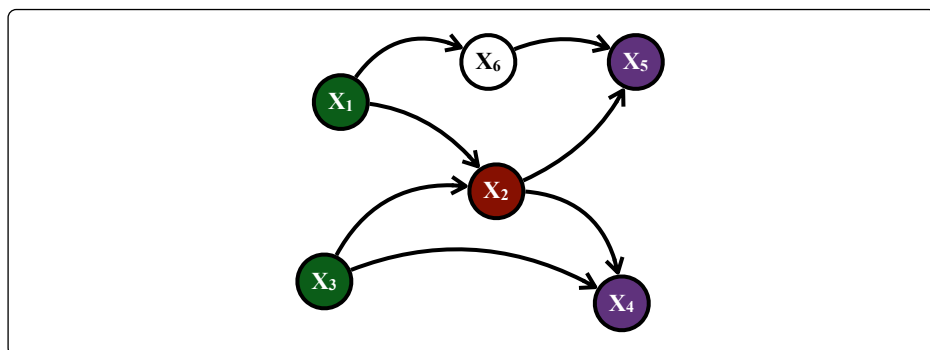
### Exercise 5.19

In order to convince you about Eq. (5.22), write  $p(y|x, z)$  in terms of  $p(z|x, y)$ ,  $p(x)$  and  $p(y)$ .

A causal dependence between  $n$  variables  $X_1, \dots, X_n$  can be represented as a *structural causal model*  $X_i = f(X_{\partial_i}, U_i)$ , where  $X_{\partial_i} = \{X_j, j \in \partial_i\}$  is a shorthand for the set of variables that “cause”  $X_i$  (and  $\partial_i$  is the set of indices of these variables). A structural causal model can be represented as a *directed a-cyclical graph* (DAG), where “directed” means that each link has a direction and “a-cyclic” means that there are no loops. For example, the DAG in Figure 9 corresponds to

$$p(x_1, \dots, x_6) = p(x_1)p(x_2|x_1, x_3)p(x_3)p(x_4|x_2, x_3)p(x_5|x_2, x_6)p(x_6|x_1)$$





**Figure 9.** A directed acyclic graph representing the statistical dependence between six variables.

In each factor, the conditional probability of  $X_i$  only contains the variables that “cause”  $X_i$ , which are called the *parent* variables. The best way to identify causal relations is to act on the variables. So suppose one can act on the variables and fix one of the variables to a specific value, let’s say  $X_2 = x_2$ . It is clear that this intervention will affect only the variables which are downstream of  $X_2$  (e.g.  $X_4$  and  $X_5$ ) and not those that are upstream (e.g.  $X_1, X_3$  and  $X_6$ ). This means that the marginal pdf of  $X_i$  changes only for those variables that are causally dependent on  $X_2$  but not for those which are not causally related. Note, in particular, that fixing  $X_2 = x_2$  to a constant makes  $X_4$  and  $X_5$  independent. Graphically, fixing a variable corresponds to removing the corresponding node from the DAG. This may break the DAG into disconnected components. Two variables that belong to different disconnected components are independent.

These arguments are developed further in the field of *causal inference*.<sup>15</sup> For our purposes, let us suffice to say once again that statistical dependence should not be interpreted as causation.

### Exercise 5.20

Compute the  $m^{\text{th}}$  moment of the Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ , for a generic  $m$ .

<sup>15</sup>See very interesting lectures by Bernhard Schölkopf at the 2020 Machine Learning Summer School.

**Exercise 5.21**

Show that for a multivariate Gaussian distribution (Eq. (5.18))

$$\mathbb{E} \left[ e^{\vec{b} \cdot \underline{X}} \right] = e^{\vec{b} \cdot \vec{m} + \frac{1}{2} \vec{b}' \hat{A}^{-1} \vec{b}}$$

**Exercise 5.22**

Show that if  $\underline{X} = (X_1, \dots, X_n)$  follows a multivariate Gaussian distribution (Eq. (5.18)), then the marginal distribution of  $X_i$  is also a Gaussian with mean  $m_i$  and variance equal to  $\{\hat{A}^{-1}\}_{i,i}$  (i.e. the  $i, i$  element of the inverse of  $\hat{A}$ ).<sup>a</sup>

<sup>a</sup>The world of Gaussian variables is like Eden. It's beautiful and perfect. The only way to get out of it is to commit a sin.

**Exercise 5.23**

Let  $X$  be a random variable distributed in the range  $[a, \infty)$  with pdf  $p(x)$ . Show that

$$\mathbb{E}[X] = a + \int_a^\infty [1 - P(x)] dx, \quad P(x) = \int_a^x p(x) dx.$$

**Exercise 5.24**

Show that if  $X$  and  $Y$  are two independent random variables with cumulative distribution  $P(x)$ , then the cumulative distribution of the minimum is

$$P\{\min(X, Y) < x\} = 1 - [1 - P(x)]^2.$$

**Exercise 5.25**

If  $X_1$  and  $X_2$  are two independent RV with the same cumulative distribution  $P(x)$ , show that

$$\mathbb{E}[|X_1 - X_2|] = 4\text{Cov}[X, P(X)].$$

**Exercise 5.26**

Let  $\theta$  be a uniformly distributed random variable in  $[0, 2\pi]$ . Show that the two random variables  $X = \cos \theta$  and  $Y = \sin \theta$  are uncorrelated but are not independent. Find their marginal distributions.

**Exercise 5.27**

Show that, if  $X$  is a Gaussian variable, then

$$\mathbb{E} [\operatorname{erf}(aX + b)] = \operatorname{erf} \left( \frac{a\mathbb{E}[X] + b}{\sqrt{1 + 2a^2\mathbb{V}[X]}} \right),$$

where

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz$$

is the error function.



## Chapter 6

# On urn models and sampling\*

The fundamental idea of probability theory is expressed in terms of urn models. FELLER

Even though we cannot define the probability  $P(A)$  of an event  $A$  as the limit of the frequency of its occurrence in a sequence of many independent trials, this remains a possible way to estimate  $P(A)$ .

The process of repeating the experiment becomes conceptually equivalent to repeated draws with replacement from an urn with many balls, a (unknown) fraction  $p = P(A)$  of which is black and the rest is white. So a “success” in the experiment, i.e. the occurrence of the event  $A$ , is equivalent to a draw of a black ball. In this schematisation,  $p$  is an objective, physical property of the system (the fraction of black balls). Let’s first argue that indeed the frequency of draws of black balls will converge to  $p$ .

Let  $K_n$  be the number of black balls drawn after  $n$  draws with replacement, i.e. when the ball which is drawn is put back into the urn. It is clear that the probability of  $K_n(\omega) = k$  is given by the binomial distribution:

$$P\{k|n, p\} = \binom{n}{k} p^k (1 - p)^{n-k}.$$

We expect and we can explicitly check by De Moivre-Laplace limit of the binomial that, as  $n$  gets large, the frequency  $K_n(\omega)/n \rightarrow p$ , in the sense that the probability that  $|K_n/n - p| > \epsilon$  gets soon very small. Please note that this is a non-trivial statement because  $K_n(\omega)/n$  depends on  $\omega$  (it is a random variable!), whereas  $p$  is independent of  $\omega$ , it is a constant.

It is also instructive to check this numerically. It is easy to write a computer code<sup>1</sup> — let’s call it `A.for` — that will generate the sequence  $k_n$  (for  $n \leq 1000$

---

<sup>1</sup>Do it and run it!

and  $p = 0.4$ , for example),<sup>2</sup>

```
seed=81701
p=0.4
do n=1,1000
    if (ran(seed).lt.p) k=k+1
    print *,n,k
end do
end
```

If we plot  $\frac{k}{n}$  we expect to see a trajectory drawing closer to  $p = 0.4$ .

This seems a good simulation of what we expect from our experimental process, with one big difference:

*when we do the experiment we do not know  $p$ . Actually, we do the experiment precisely because we want to measure  $p$ !*

In order to imagine the situation where  $p$  is unknown, think of the situation where the line `p=0.4` is replaced by `p=ran(seed)` in the program above. Imagine someone compiles the program, with an unknown value of `seed`, and gives you the executable, but not the source file. Then, you can run the code and plot  $k/n$  vs  $n$ . You will observe  $k/n$  converge to some value, which will be close to the unknown  $p$ . This means that over time we will learn the value of  $p$  to a better and better approximation.

Indeed, by Bayes rule you can find out what is the probability of  $p$  being in any interval  $[x, x + dx]$ , and find that this is sharply peaked at  $x \approx k/n$ , for large  $n$ . In order to do this, you need a prior distribution on the value of  $p$ . How to choose a prior is a quite interesting and non-trivial issue that is discussed in detail in ref. [8]. We'll get back to it, for the moment I will assume that  $P_0(p) = Ap^{a-1}(1-p)^{a-1}$  (i.e.  $a = 1$  corresponds to the uniform prior). Then

$$\begin{aligned} P\{p \in [x, x + dx] | n, k\} &= \frac{P\{k | n, p = x\} A x^{a-1} (1-x)^{a-1} dx}{\int_0^1 P(k | n, y) A y^{a-1} (1-y)^{a-1} dy} \\ &= \frac{\Gamma(n+2a)}{\Gamma(k+a)\Gamma(n-k+a)} x^{k+a-1} (1-x)^{n-k+a-1} dx \end{aligned}$$

---

<sup>2</sup>The variable `seed` initialises the random number generator `ran()`. This is morally the analog of the element  $\omega$  of the sample space, in the sense that different choices of `seed` generate different random sequences. The `if` statement uses the fact that `ran(seed)` produces a uniform random variable, which is less than  $p$  with probability  $p$ .

Then, our estimate of the fraction  $p$  of black balls, is the expected value on this distribution:<sup>3</sup>

$$\mathbb{E}[p|n, k] = \frac{\Gamma(n+2a)}{\Gamma(k+a)\Gamma(n-k+a)} \int_0^1 x^{k+a-1}(1-x)^{n-k+a-1} x dx = \frac{k+a}{n+2a}.$$

Summarising, in our sampling process the distribution of possible values of  $p$  becomes more and more sharply peaked around a value that is close to the frequency  $p \simeq k/n$  as  $n \rightarrow \infty$ .

Note that the best estimate of the probability that the next draw will be a success, using the results above, is given by

$$P\{K_{n+1} = k+1 | K_n = k, n\} = \frac{k+a}{n+2a}. \quad (6.1)$$

So we can compute the probability that the next drawn ball will be black without knowing  $p$ , incorporating all the information we have so far about it. Indeed, we can “simulate” the whole process without knowing  $p$ . In other words, if you generate a sequence  $k_n$ , starting from  $k_0 = 0$ , using the above rule — i.e setting  $k_{n+1} = k_n + 1$  with probability  $(k_n + a)/(n + 2a)$  and  $k_{n+1} = k_n$  otherwise — then this should be statistically indistinguishable from  $n$  repeated draws from an urn with unknown composition. To make the statement concrete, consider the program B. for

```
a=1.0
seed=81701
do n=1,1000
  if ((n+2*a)*ran(seed).lt.k+a) k=k+1
  print *,n,k
end do
end
```

The output of this program is statistically undistinguishable from the output of A. for, which means that if both are compiled and given the same name there is no way we can tell whether the output comes from one or the other.

---

<sup>3</sup>The expected value of the estimate of  $p$  does not coincide with  $p$  unless  $a = 0$  (in statistics jargon, the estimator is biased).  $a = 0$  corresponds to a prior  $P_0(p) = A/[p(1-p)]$  that is non normalizable (which is called an improper prior). This seems odd, but there are good reasons to believe that indeed  $a = 0$  correctly encodes our state of ignorance on  $p$ . Indeed, if you had observed no success ( $k = 0$ ) after  $n$  trials, you would infer that  $p = 0$ . The article by Jaynes cited above provides arguments for  $a = 0$ .

**Exercise 6.1**

Show that the random variable  $P_n = \frac{K_n + a}{n + 2a}$  in this process satisfies the equation  $\mathbb{E}[P_{n+1} | P_n] = P_n$ . Sequences of random variables that satisfy this property are called *martingales*. Iterating this equation, show that  $\mathbb{E}[P_n] = \frac{1}{2}$ .

Interestingly, you can interpret even B. for as reproducing a sequence of draws from an urn. Take an urn which initially has  $a$  black and  $a$  white balls.<sup>4</sup> At each draw, the drawn ball is put back into the urn and an additional ball of the same colour is added (by some device internal to the urn). Equivalently, you can think that the ball “magically” duplicates just after being put back in the urn. So, if a black (white) ball is drawn, it is put back in the urn and a further black (white) ball is added. The number of balls increases with  $n$ . If out of  $n$  draws  $k$  black balls have been observed, then the urn will contain  $k + a$  black balls out of a total of  $n + 2a$  balls. Hence the probability that the next draw results in a black ball is exactly given by Eq. (6.1). And this is precisely the process that B. for simulates. This model is called a *Polya urn*.<sup>5</sup>

Arbitrary<sup>6</sup> as it might seem, this construction is just a different conceptual model for our sequence of experiments. Remember that the observer is running the experiments *precisely* because he/she wants to learn about a system that is not known. So there is no *a priori* reason to prefer one to the other.

The fact that A. for and B. for produce statistically indistinguishable outputs is striking, for the two programs code for processes which are completely different. One is a sequence of independent draws, whereas in the second the outcome in a draw depends on the whole sequence of previous draws. In the first the urn is always the same, whereas the second is a process where draws modify the composition of the urn and hence the probability of future events.

*There is no way to know whether you're learning about the unknown composition of the urn or if you're filling up an urn in a history*

<sup>4</sup>Isn't it curious that the parameter  $a$  that specifies the number of balls in the urn before the first draw is also the one that defines the prior distribution  $P_0$ ? This does not seem like a coincidence, because that is precisely what is known about the urn before the draws.

<sup>5</sup>See FELLER, Chapter V.

<sup>6</sup>One may question about the arbitrariness of such constructions. JAYNES argues that even the simple scheme of draw from an urn with replacement is not at all unambiguous, as the state of the urn is affected by the drawing. One implicitly assumes that the urn is *shuffled after each draw enough to ensure that at the next draw the urn is in its original state*. Yet how much shaking and what “ensure” really means is never really spelled out. Definitely the observer, after all these operations, is not in the same state of knowledge as at the time of the first draw.



*dependent manner.*<sup>7</sup>

Now you should be in a position to answer the following questions:

1. what is the fraction of black balls in the limit of infinite draws in the second process?
2. What is the limiting fraction if one starts from an urn with two black and one white balls? or with  $b$  black and  $w$  white balls?

## 6.1 Sampling and undersampling

This setup can be generalized to experiments that can give any number of outcomes. In many cases, when we do experiments, we do not even know how many different outcomes we can get. Consider for example a botanist that is classifying samples into species of plants in a yet unexplored island. He/she has a criterium to decide whether the next sample is a further exemplar of one of the species he/she has already seen or if it is a new species. In this case, as well in cases where the system we're studying is complex, the number of outcomes can be very large, and much larger than the number  $M$  of available samples.<sup>8</sup>

So consider the general situation of an experiment repeated  $M$  times and let  $k_x$  be the number of times the state (or outcome)  $x$  is observed, with  $x = 1, \dots, \Omega$  and  $\sum_x k_x = M$ . We may think of the experiment as sampling an underlying distribution. But when  $M \ll \Omega$  we're very far from sampling correctly this distribution (we call this under-sampling regime).

What can one learn from this data? What is the typical behavior of a sampling process? What type of frequency distribution can we expect?

Generalizing the discussion above, it is possible to estimate in a Bayesian manner, the probability  $p_x$  of outcomes  $x$ , given the outcome of earlier experiments (the number  $k_x$  of times the outcome  $x$  has been found). The

<sup>7</sup>This means that, on the basis of the data alone, it is not possible to exclude "magical" explanations that maintain that the outcome of an experiment is influenced in mysterious ways by the outcome of previous experiments. Common sense suggests that this is not the right explanation.

<sup>8</sup>As another example, consider the case where each observation can be a gene expression array that tells you whether each gene is on or off. There may be hundreds of genes, so the number of possible outcomes can be as large as  $\Omega \simeq 2^{n. \text{ of genes}}$ . Since the number of genes is of the order of tens of thousands, this is an astronomical number, in principle. In practice, the observed gene expression profiles are only those compatible with a biologically functional organism, so they may be much less. The number of samples in typical experiments can be of order  $M \sim 10^2 \nabla \cdot 10^3$ , which is much less than  $\Omega$ .

probability of  $\vec{k}$  is

$$P\{\vec{k}|\vec{p}\} = M! \prod_{x=1}^{\Omega} \frac{p_x^{k_x}}{k_x!} \delta_{\sum_x k_x, M}.$$

Let the prior (pdf) on  $\vec{p}$  be<sup>9</sup>

$$p_0(\vec{p}) = \frac{\Gamma(\alpha\Omega)}{\Gamma(\alpha)^\Omega} \prod_x p_x^{\alpha-1} \delta\left(\sum_x p_x - 1\right)$$

Then the posterior pdf is given by Bayes rule

$$p(\vec{p}|\vec{k}) = \Gamma(M + \alpha\Omega) \prod_x \frac{p_x^{k_x + \alpha - 1}}{\Gamma(k_x + \alpha)} \delta\left(\sum_x p_x - 1\right)$$

Imagine now that I can do a further experiment. The probability to observe outcome  $x$  is

$$\mathbb{E}[p_x|\vec{k}] = \int d\vec{p} p(\vec{p}|\vec{k}) p_x = \frac{k_x + \alpha}{M + \alpha\Omega} \quad (6.2)$$

If the outcome of the experiment  $M + 1^{\text{st}}$  and that of subsequent experiments is consistent with what we have learned so far, then we can view  $\vec{k}^{(M)}$  as being the result of a draw from a Polya urn that initially contains  $\alpha$  balls of each color  $x$ . A sequence of draws of a ball from the urn is executed where, after each draw, the chosen ball is put back into the urn together with a further ball of the same color. After  $M$  draws, the probability of drawing a ball of color  $x$  is precisely given by Eq. (6.2).

Therefore Polya urn schemes describe experiments whose outcomes, at each time, are consistent with the statistics accumulated that far: these *self-consistent* experiments produce a peculiar distribution of  $\vec{k}$  which is obtained generalizing the arguments in Feller Vol. 1 for the simple case  $\Omega = 2$ : each sequence  $x_1, x_2, \dots, x_M$  with a certain number  $k_x$  of outcomes  $x_\ell = x$  has the same probability, which is given by

$$\begin{aligned} P\{x_1, x_2, \dots, x_M\} &= \frac{\prod_{x: k_x > 0} \alpha(\alpha + 1) \cdots (\alpha + k_x - 1)}{\Omega \alpha(1 + \Omega \alpha) \cdots (M - 1 + \Omega \alpha)}, & k_x &= \sum_{\ell=1}^M \delta_{x_\ell, x} \\ &= \frac{\Gamma(\alpha\Omega)}{\Gamma(\alpha\Omega + M)} \prod_x \frac{\Gamma(k_x + \alpha)}{\Gamma(\alpha)} \end{aligned}$$

<sup>9</sup>Here  $\delta(x)$  is Dirac's delta function, defined by the relation

$$f(x_0) = \int_{-\infty}^{\infty} f(x) \delta(x - x_0) dx$$

for any function  $f(x)$  and any  $x_0 \in \mathbb{R}$ .

The number of sequences of this type is given by the multinomial factor  $M!/(\prod_x k_x!)$ , so that

$$P\{\vec{k}|M\} = \frac{M!\Gamma(\alpha\Omega)}{\Gamma(M + \Omega\alpha)\Gamma(\alpha)\Omega} \prod_x \frac{\Gamma(k_x + \alpha)}{k_x!}, \quad \sum_x k_x = M \quad (6.3)$$

which is also what one gets from integrating the likelihood over the prior on  $\vec{p}$ .

The number  $N_M$  of different colours discovered up to  $M$ , i.e. the number of  $x$  with  $k_x > 0$ , can be estimated for large  $M, \Omega$  as follows: write  $M = m\Omega$  and  $N_M = n(m)\Omega$  then

$$P\{N_{M+1} = N_M + 1\} = \frac{\alpha(\Omega - N_M)}{M + \alpha\Omega} = \frac{\alpha(1 - n)}{m + \alpha} \quad (6.4)$$

$$\simeq \frac{dN}{dM} = \frac{dn}{dm}. \quad (6.5)$$

By integration of the resulting differential equation, we find

$$n(m) = 1 - \left(1 + \frac{m}{\alpha}\right)^{-\alpha}. \quad (6.6)$$

So,  $N_M \simeq M$  for  $M \ll \alpha\Omega$  and  $N_M$  saturates at  $\Omega$  when  $M \gg \alpha\Omega$ .

The probability that, for a particular  $x$ , we find  $k_x = k$  decays as  $k^{\alpha-1}e^{-\nu k}$ , where  $\nu$  is adjusted so that  $E[k] = M/\Omega$ . The number of  $x$  with  $k_x = k$ , on average, is<sup>10</sup>

$$P\{k_x = k|M\} = Ak^{\alpha-1}e^{-\nu k}.$$

when  $1 \ll M \ll \Omega$  we expect  $\nu \ll 1$  so the distribution of frequency types is very broad.

Indeed, broad frequency distributions are observed rather ubiquitously when one samples complex systems. For example, the abundance of species in a given environment, the number of species with a given gene, the number of firms with a given number of employees all follow broad frequency distributions. Our discussion suggests that this is the hallmark of an under-sampled system.

We'll come back to this, from a different angle, when we'll talk about statistics and inference.

<sup>10</sup>In order to obtain this result, you should sum Eq. (6.3) on all values of  $k_{x'}$  for  $x' \neq x$  with the constraint  $\sum_x k_x = M$ . This constraint can be introduced with the integral representation of the delta function  $\delta_{\ell,j} = \int_{-\pi}^{\pi} \frac{dq}{2\pi} e^{iq(\ell-j)}$ . The sums on  $k_{x'}$  factorise inside the integral, which can then be evaluated by saddle point.



# Chapter 7

## Generating functions

A very useful tool to handle infinite sequences  $a_n$  ( $n = 0, 1, \dots$ ) is to construct the function

$$A(s) = \sum_{n=0}^{\infty} a_n s^n, \quad s \in \mathbb{C} \quad (7.1)$$

which is called a *generating function* (GF). This is useful because all the properties of the sequence are encoded in the analytic behaviour of  $A(s)$ , as a function of  $s$ . Eq. (7.1) is a formal power series that need not necessarily converge for any value of  $s$ . The variable  $s$  has no meaning. It is used as a device.

In this chapter we shall first see how the structure of singularities of  $A(s)$  can inform us on the asymptotic behaviour of the sequence  $a_n$ . Then we shall see how generating functions can be used to solve counting problems. Finally we shall apply generating function to probability. We shall restrict attention to cases where  $a_n \geq 0$ , which are those we shall be interested in. As a teaser, we start by an example that shows the power of generating functions.

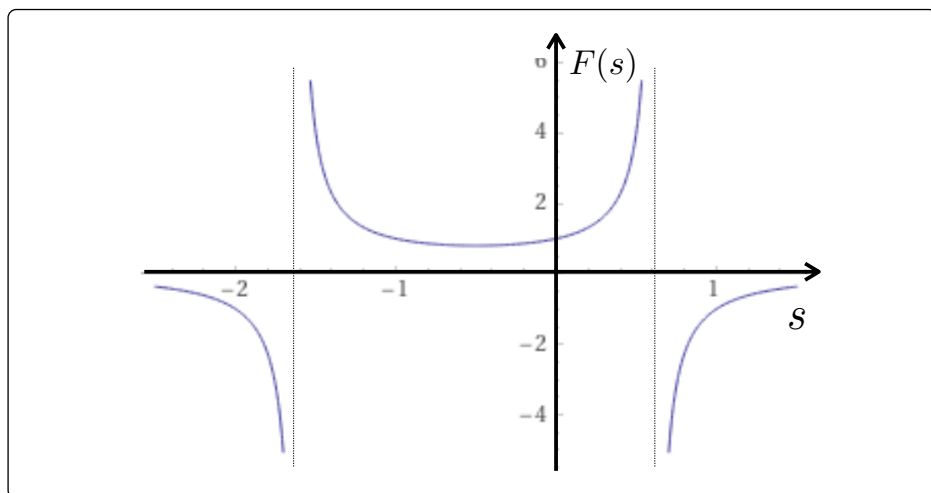
### 7.1 Warm-up: Fibonacci numbers

Fibonacci numbers are defined by the recurrence relation

$$f_{n+1} = f_n + f_{n-1}$$

for  $n > 0$  and  $f_0 = f_1 = 1$ . Let  $F(s)$  be the corresponding generating function. If you multiply the recurrence relation by  $s^{n+1}$  and sum on all  $n > 0$  you find  $F(s) - f_0 - f_1 s = s[F(s) - f_0] + s^2 F(s)$ , i.e.,

$$F(s) = \frac{1}{1 - s - s^2}.$$



**Figure 10.** The function  $F(s)$  has two singularities, at  $s = 1/\phi$  and at  $s = -\phi$ .

This has two poles (the zeros in the denominator) at  $s = -\phi$  and at  $s = 1/\phi$ , where  $\phi = \frac{1+\sqrt{5}}{2} \simeq 1.61803 \dots$  is the golden ratio. Expanding  $F(s)$  in simple fractions, and then in geometric series, we find

$$F(s) = \frac{1}{\sqrt{5}} \left[ \frac{\phi}{1 - \phi s} + \frac{\phi^{-1}}{1 + \phi^{-1} s} \right] \quad (7.2)$$

$$= \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} [\phi^{n+1} - (-\phi^{-1})^{n+1}] s^n. \quad (7.3)$$

This gives the remarkable result

$$f_n = \frac{1}{\sqrt{5}} [\phi^{n+1} - (-\phi^{-1})^{n+1}], \quad n \geq 0.$$

This formula allows to compute  $f_{100} = 573147844013817084101$  with few operations, without the need to iterate the recursion relation 100 times. In spite of the fact that  $\phi$  is an irrational number, this expression yields an integer number for all  $n$ . For  $n \rightarrow \infty$ , the asymptotic behaviour of Fibonacci numbers is given by the first term in the expression above, since the second term is exponentially smaller with respect to the first. Hence  $f_n \simeq \phi^{n+1}/\sqrt{5}$  for  $n$  large.

**Exercise 7.1**

Can you compute the generating function when  $f_0 = 0$  and  $f_1 = 1$ ? What is the difference? What if  $f_0 = 1$  and  $f_1 = 2$ ?

The same techniques can be applied for any sequence  $a_n$  that is defined by a recursion relation, transforming the latter into an equation for the generating function  $A(s)$  in Eq. (7.1). This equation can be solved to obtain  $A(s)$  in explicit form. Finally the expansion on powers of  $s$  yield  $a_n$  as the coefficient of  $s^n$ . The  $n^{\text{th}}$  coefficient in the Maclaurin power expansion of  $A(s)$  can be computed in different ways:

$$a_n = \frac{1}{n!} \left. \frac{d^n A(s)}{ds^n} \right|_{s=0} = \int_0^{2\pi} \frac{dq}{2\pi} A(e^{iq}) e^{-iqn} = \frac{1}{2\pi i} \oint \frac{ds}{s} A(s) s^{-n}, \quad (7.4)$$

where the last integral is on a contour around the origin in the complex plane  $s \in \mathbb{C}$ . In particular,  $a_0 = A(0)$  and if  $A(s) = s^m B(s)$  where  $B(s)$  is analytic at  $s = 0$ , then  $a_n = 0$  for all  $n < m$ .

Even if it is not possible to find an exact formula for  $a_n$ , it is still possible to extract its asymptotic behaviour as  $n \rightarrow \infty$ , by studying the analytic properties of  $A(s)$  close to its singularities, as we're going to see next.

## 7.2 Asymptotics of $a_n$ from the structure of singularities

Let us start by the simplest case where  $A(s)$  has only isolated single poles<sup>1</sup>

$$A(s) = \frac{N(s)}{D(s)}, \quad (7.5)$$

where  $D(s)$  is a polynomial of degree  $d$  and  $N(s)$  is a polynomial of degree<sup>2</sup>  $n \leq d$ . Let  $s_1, \dots, s_d$  be the zeroes of  $D(s)$ , which we assume to be all different  $s_i \neq s_j$  for  $i \neq j$ . This means that  $D(s)$  can be written as the product of  $(s - s_i)$  over the different roots. Close to  $s_i$ ,  $A(s)$  diverges as  $c_i/(s - s_i)$ , where the constant  $c_i$  can be computed as the limit of  $(s - s_i)A(s)$ , as  $s \rightarrow s_i$ . Using L'Hôpital rule, we find

$$c_i = \lim_{s \rightarrow s_i} \frac{(s - s_i)N(s)}{D(s)} = \frac{N(s_i)}{D'(s_i)},$$

<sup>1</sup>This argument is presented in much more detail in FELLER XI.4.

<sup>2</sup>How does the case  $n > d$  reduces to this?

where  $D'(s_i)$  is the first derivative of  $D(s)$  computed in  $s_i$ . Then we can write

$$A(s) = \sum_{i=1}^d \frac{N(s_i)}{D'(s_i)} \frac{1}{s - s_i}. \quad (7.6)$$

You can check that this expression for  $A(s)$  has the same poles of Eq. (7.5), with the same asymptotic behaviour as  $s \rightarrow s_i$  for all  $i$ . Now each term in Eq. (7.6) can be expanded as a geometric series, so

$$A(s) = \sum_{n=0}^{\infty} \left[ - \sum_{i=1}^d \frac{N(s_i)}{D'(s_i)} s_i^{-n-1} \right] s^n \quad (7.7)$$

which shows that

$$a_n = - \sum_{i=1}^d \frac{N(s_i)}{D'(s_i)} s_i^{-n-1}. \quad (7.8)$$

For  $n \rightarrow \infty$  the term that dominates the sum is the one corresponding to the root  $s_i$  which is closest to the origin. Without loss of generality, we can assume that  $|s_1| \leq |s_2| \leq \dots \leq |s_d|$ . Then  $a_n \sim e^{an}$  has an exponential behaviour for large  $n$ , with a rate given by  $a = -\log |s_1|$ .

The first lesson that we learn can be summarised in the following.<sup>3</sup>

**First Principle of Coefficient Asymptotics.** The location of the singularities of a function dictates the exponential growth of the coefficients of its power expansion. More precisely, if the closest singularity of  $A(s)$  to the origin is at  $s_1$ , then  $a_n \sim |s_1|^{-n}$  in the sense that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log a_n = -\log |s_1|.$$

If instead of a single pole, the singularity closest to the origin is a double pole  $A(s) \sim (s - s_1)^{-2}$ , then it is easy to see<sup>4</sup> that the corresponding leading asymptotic behaviour is given by  $a_n \sim ne^{an}$ . This suggests that the type of singularity determines the sub-exponential asymptotic behaviour of  $a_n$ . In

<sup>3</sup>If  $a_n \simeq An^\beta z^n$ , a simple recipe to compute  $z$  is to observe that

$$\lim_{n \rightarrow \infty} \frac{a_{n+k}}{a_n} = z^k.$$

For example, if  $a_n$  satisfies a recursion relation, such as  $a_{n+2} = ba_{n+1} + ca_n$ , then dividing this equation by  $a_n$  and taking the limit  $n \rightarrow \infty$ , one finds that  $z$  is given by the solution of the equation  $z^2 = bz + c$ , with the largest value of  $|z|$ .

<sup>4</sup>Hint: compare the singularities of  $A(s)$  in Eq. (7.7) and of its derivative.



order to get more intuition, let us consider the generating function  $A(s) = (s_1 - s)^{-\alpha-1}$ . The binomial theorem gives directly the power expansion in  $s$ , as

$$A(s) = \sum_{n=0}^{\infty} \binom{-\alpha-1}{n} (-s)^n s_1^{-\alpha-1-n}.$$

In order to simplify the expression of  $a_n$  we shall use the properties of the  $\Gamma$  function

$$a_n = (-1)^n \binom{-\alpha-1}{n} s_1^{-\alpha-1-n} \quad (7.9)$$

$$= (-1)^n \frac{(-\alpha-1)(-\alpha-2) \dots (-\alpha-n)}{n!} s_1^{-\alpha-1-n} \quad (7.10)$$

$$= \frac{\Gamma(n+1+\alpha)}{\Gamma(\alpha+1)n!} s_1^{-\alpha-1-n} \quad (7.11)$$

$$\simeq \frac{s_1^{-\alpha-1}}{\Gamma(\alpha+1)} n^\alpha s_1^{-n}, \quad n \gg 1 \quad (7.12)$$

where we used Stirling's approximation for both  $n!$  and the  $\Gamma$  function in the last relation.

This has been generalised by Flajolet and Odlyzko [12] to the

**Second Principle of Coefficient Asymptotics.** The nature of the singularity closest to the origin of the generating function  $A(s)$  determines the sub-exponential behaviour of the sequence  $a_n$ . More precisely, setting  $s_1 = 1$ , if

$$A(s) = F\left(\frac{1}{1-s}\right) \quad \text{with} \quad F(u) \sim u^\alpha (\log u)^\gamma (\log \log u)^\delta \quad \text{as } u \rightarrow \infty$$

then<sup>5</sup>

$$a_n \sim \frac{1}{\Gamma(\alpha)} \frac{F(n)}{n}$$

as  $n \rightarrow \infty$ .

## 7.3 Counting with functions\*

Much of classical probability is about counting. There are smart ways to count objects using algebra, and it's worthwhile doing a digression.

---

<sup>5</sup>The symbol  $\sim$  means that the limit of the ratio of the right hand side and the left hand side of the relation equals one when  $n \rightarrow \infty$ .

Imagine you are interested in a class  $\mathcal{A}$  of objects of a certain kind. Each object  $A$  has an integer size  $|A| = n$  that counts its components. For example,  $A$  may be a graph of  $n$  nodes, and  $\mathcal{A}$  the set of all graphs satisfying some rules.

The typical questions that we address is: how many objects are there in  $\mathcal{A}$  of size  $n$ ? For example, how many words of 7 letters from a given alphabet can you form? How many trees of  $n$  nodes are there?

One way to do this, given the set  $\mathcal{A}$  of all possible objects  $A \in \mathcal{A}$ , is to construct the function

$$A(z) = \sum_{A \in \mathcal{A}} z^{|A|} = \sum_n a_n z^n. \quad (7.13)$$

As before,  $z$  is a real or complex variable that does not have any meaning, per se. It serves just as a counting device. It is not necessary that Eq. (7.13) be a well defined function in general. Yet, in all cases we shall discuss  $A(z)$  is a convergent series in a neighbourhood of the origin  $z = 0$ , i.e. the radius of convergence is finite.<sup>6</sup>

As the second equality in Eq. (7.13) shows, the number  $a_n$  of objects of size  $n$  is given by the coefficient of the  $n^{\text{th}}$  term in the power expansion of  $A(z)$ . This is very useful in case  $A(z)$  can be computed analytically. This subject is dealt with in great detail in Flajolet's book [13], to which we refer as FLAJOLET in what follows. The purpose of this discussion is to introduce you to the subject and to show the power of generating functions as counting devices.

### 7.3.1 Operations on sets

The objects we want to count may satisfy some properties that can be expressed in terms of basic operations. These operations correspond to algebraic operations on generating functions. Let us see some of them (see FLAJOLET for more):

**Union.** The objects we're interested in are of two possible types, i.e. they belong to a class  $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$  of objects that can be split in two disjoint ( $\mathcal{A} \cap \mathcal{B} = \emptyset$ ) classes. Then

$$\begin{aligned} C(z) &\equiv \sum_{C \in \mathcal{C}} z^{|C|} = \sum_n c_n z^n \\ &= \sum_{A \in \mathcal{A}} z^{|A|} + \sum_{B \in \mathcal{B}} z^{|B|} = A(z) + B(z), \\ \text{i.e.} \quad c_n &= a_n + b_n \end{aligned}$$

---

<sup>6</sup>Given what we have discussed in the previous section, this restricts our attention to series for which  $|a_n|$  diverges at most exponentially.

**Product.** Each object  $C = (A, B)$  can be decomposed in sub-objects of smaller sizes,  $|A| + |B| = |C|$ . Correspondingly the class of objects  $\mathcal{C} = \mathcal{A} \otimes \mathcal{B}$  can be written as the direct product of the classes of the sub-objects. This implies, for generating functions

$$C(z) = \sum_{A \in \mathcal{A}, B \in \mathcal{B}} z^{|A|+|B|} = A(z)B(z).$$

Therefore  $c_n$  is given by

$$c_n = a_0 b_n + a_1 b_{n-1} + \dots + a_n b_0.$$

which is called the *convolution* of the sequences  $a_n$  and  $b_n$ .

**Sequence.** The class of objects  $\mathcal{C}$  we're interested in are the repetition of more elementary objects in a class  $\mathcal{A}$ . We write  $\mathcal{C} = \text{Seq}(\mathcal{A})$  to denote the fact that the generic element  $C = (A_1, A_2, \dots, A_k)$  is a sequence of elements  $A_j \in \mathcal{A}$ . Clearly<sup>7</sup>  $\mathcal{C} = \emptyset + \mathcal{A} + \mathcal{A} \otimes \mathcal{A} + \mathcal{A} \otimes \mathcal{A} \otimes \mathcal{A} + \dots$  that, using the two relations above, imply

$$C(z) = 1 + A(z) + A^2(z) + \dots = \frac{1}{1 - A(z)} \quad (7.14)$$

Note that we admit as a possible object in  $\mathcal{C}$  the sequence with zero elements of  $\mathcal{A}$ .

**Powerset.** Imagine we want to consider all possible subsets of objects  $A$  of a given set  $\mathcal{A}$ . The set of all these objects — called the *power-set* — can formally be written as

$$\mathcal{C} \equiv \text{PSet}(\mathcal{A}) = \bigotimes_{A \in \mathcal{A}} [\emptyset \cup \{A\}]$$

Indeed expanding the product, each term corresponds to a “monomial” with some of the objects  $A \in \mathcal{A}$  occurring only once. The corresponding

---

<sup>7</sup>The first element of the sequence is the empty set, i.e. a set containing no element. We interpret the empty set as the set with one element of size zero. Hence the generating function of the empty set is 1. With this definition  $\mathcal{A} = \emptyset \otimes \mathcal{A}$  because all objects in the l.h.s. correspond to one object on the r.h.s. where we add one element of size zero.

FLAJOLET avoids reference to the empty set, defining a *neutral* set which is composed of one *neutral* element of size zero.

generating function is given by

$$\begin{aligned}
 C(z) &= \prod_{A \in \mathcal{A}} (1 + z^{|A|}) = \prod_n (1 + z^n)^{a_n} \\
 &= \exp \left[ \sum_n a_n \log(1 + z^n) \right] \\
 &= \exp \left[ \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \sum_n a_n z^{kn} \right] = \exp \left[ \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} A(z^k) \right]
 \end{aligned}$$

**Multiset.** The multi set  $\mathcal{C} = \text{MSet}(\mathcal{A})$  is the set of all collections of objects taken from  $\mathcal{A}$  with repetition. One can write

$$\text{MSet}(\mathcal{A}) = \bigotimes_{A \in \mathcal{A}} \text{Seq}(\{A\})$$

from which<sup>8</sup>

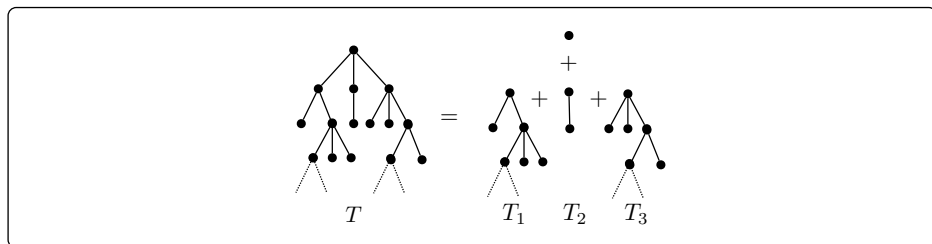
$$\begin{aligned}
 C(z) &= \prod_{A \in \mathcal{A}} (1 - z^{|A|})^{-1} = \prod_n (1 - z^n)^{-a_n} \\
 &= \exp \left[ - \sum_n a_n \log(1 - z^n) \right] \\
 &= \exp \left[ \sum_{k=1}^{\infty} \frac{1}{k} \sum_n a_n z^{kn} \right] = \exp \left[ \sum_{k=1}^{\infty} \frac{A(z^k)}{k} \right]
 \end{aligned}$$

Let us illustrate with some examples how these concepts can be useful to count:

- The set of all binary words is  $\mathcal{C} = \text{Seq}(\{0, 1\})$ . The generating function of  $\mathcal{A} = \{0, 1\}$  is  $A(z) = 2z$ , because there are two objects of size 1 in  $\mathcal{A}$ . Then  $C(z) = 1/(1 - 2z) = \sum_n 2^n z^n$ . Indeed there are exactly  $2^n$  binary words of size  $n$ .
- Consider the set  $\mathcal{T}$  of all rooted plane trees. A rooted plane tree of size  $n$  is a connected graph of  $n$  points  $\bullet$  and  $n - 1$  links joining them. One of the vertices is the root, from which the tree starts. Plane means that

---

<sup>8</sup>The difference between sequence, power-set and multi-set is the same as the difference between Boltzmann, Fermi-Dirac and Bose-Einstein statistics. Indeed each element of a sequence  $\text{Seq}(\{A\})$  is a collection of elements in  $\mathcal{A}$  in any order, whereas in the power-set  $\text{PSet}(\mathcal{A})$  each element of  $\mathcal{A}$  can occur only once, and in the multi-set  $\text{MSet}(\mathcal{A})$  only the number of times different elements of  $\mathcal{A}$  occur matter, as for indistinguishable particles.



**Figure 11.** Rooted plane trees.

ordering is specified for the sub-trees of each vertex. You can describe a tree as being a node linked to a number of (sub)trees, in the sense that if you remove a node, you're left with a collection of smaller trees. Therefore

$$\mathcal{T} = \{\bullet\} \otimes \text{Seq}(\mathcal{T})$$

This means that  $T(z) = z/[1 - T(z)]$ . This can be cast in a quadratic equation for  $T(z)$  whose solution is

$$T(z) = \frac{1 - \sqrt{1 - 4z}}{2} = \sum_n C_{n-1} z^n, \quad C_n = \frac{1}{n+1} \binom{2n}{n}$$

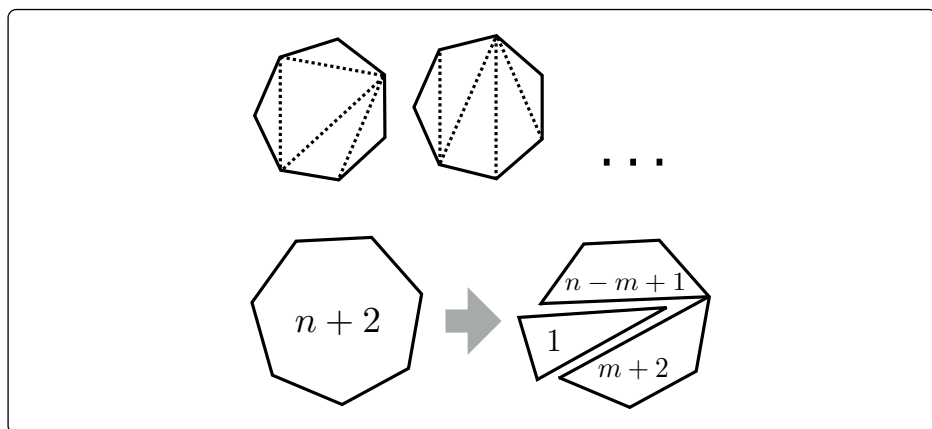
where  $C_n$  are the Catalan's numbers. Therefore the number of trees of size  $n$  is given by the  $n - 1^{\text{st}}$  Catalan number.

### Exercise 7.2

How does  $T_n$  grows with  $n$  for  $n \rightarrow \infty$ ? In particular, if  $T_n \sim n^\beta e^{\alpha n}$ , what are  $\alpha$  and  $\beta$ ?

- Consider a convex polygon with  $n+2$  edges. By drawing non intersecting diagonals this can be reduced to the union of triangles, which is called a *triangulation* of the polygon. How many triangulations does a  $n + 2$  sides polygon admit?

Let us consider the set  $\mathcal{T}$  of all triangulations of all polygons. For any given polygon with  $n + 2$  edges, removing one of the sides and joining the endpoints with another point reduces the polygon into the union of two smaller polygons, say of size  $m + 2$  and  $n - m + 1$  edges (note indeed that overall one edge has been added). Each of the sub-polygons admits a certain number of triangulations. So the number of triangulations of the original polygon can be related to the number of triangulations of the sub-polygons, with the addition of a further triangle. This implies



**Figure 12.** Triangulation of polygons.

that  $\mathcal{T} = \emptyset \cup \mathcal{T} \otimes \{\triangle\} \otimes \mathcal{T}$ , which considers the possibility that the polygon with  $n = 0$  has no triangles. Therefore

$$T(z) = 1 + zT^2(z) = \sum_n C_n z^n$$

where the second equality result from solving the second order equation and expanding the result in series. Again the result involves Catalan numbers!

### Exercise 7.3

Consider the set  $\mathcal{B}_{11}$  of binary sequences that terminate whenever a pair of ones occurs for the first time. Show that for this set

$$\mathcal{B}_{11} = \{11\} \cup \{0\} \otimes \mathcal{B}_{11} \cup \{10\} \otimes \mathcal{B}_{11}.$$

Find the equation for the generating function and show that the number of sequences of length  $n$  is given by  $b_n^{(11)} = f_{n-2}$  for  $n \geq 2$  where  $f_n$  are the Fibonacci numbers ( $b_0^{(11)} = b_1^{(11)} = 0$ ).

Next consider the set  $\mathcal{B}_{00}$  of sequences that terminate whenever a pair of zeroes occurs for the first time. What is the number  $b_n^{(00)}$  of such sequences of length  $n$ ? Now consider the set  $\mathcal{B}_=$  of sequences that terminate whenever a pair of equal digits occurs for the first time. Is  $\mathcal{B}_= = \mathcal{B}_{00} \cup \mathcal{B}_{11}$ ? What is the number of such sequences of length  $n$ ?

Next consider the set  $\mathcal{B}_{10}$  of binary sequences that terminate whenever the subsequence 10 occurs for the first time. Derive an

equation for the set  $\mathcal{B}_{10}$  and, from this, the associated generating function. Is the number of sequence of length  $n$  in  $\mathcal{B}_{10}$  equal to that in  $\mathcal{B}_{11}$ ? Does an equation like the one above for  $\mathcal{B}_{11}$  holds for  $\mathcal{B}_{10}$ ? Finally, what is the number  $b_n^\neq$  of sequences of length  $n$  that terminate whenever two consecutive digits are different? Is  $b_n^\neq = b_n^{(01)} + b_n^{(10)}$ ?

- Integer partitions and compositions. A positive integer can be decomposed in the sum of other positive integers

$$n = x_1 + x_2 + \dots, x_k, \quad x_\ell \geq 1$$

in a number of ways. Integer *partitions* correspond to the case when the summands are non-decreasing ( $x_1 \leq x_2 \leq \dots \leq x_k$ ) whereas *compositions* to the general case where  $x_\ell$  can appear in any order. How many partitions  $p_n$  (compositions  $c_n$ ) does an integer  $n$  admits? First, the set of integer can be constructed from a single element  $\{\bullet\}$  as  $\mathcal{I} = \text{Seq}(\{\bullet\}) \setminus \emptyset$ . Since the generating function of the set  $\{\bullet\}$  is just  $z$ , we have

$$I(z) = \frac{z}{1-z}.$$

This is consistent with the fact that there is one integer of size  $n$ ,  $i_n = 1$ .

Compositions are given by

$$\mathcal{C} = \text{Seq}(\mathcal{I}) \setminus \emptyset = \{(\bullet), (\bullet\bullet), (\bullet, \bullet), (\bullet \bullet \bullet), (\bullet, \bullet\bullet), (\bullet\bullet, \bullet), (\bullet, \bullet, \bullet), \dots\}$$

Correspondingly their generating function is

$$C(z) = \frac{I(z)}{1-I(z)} = \frac{z}{1-2z} = \sum_{n=1}^{\infty} 2^{n-1} z^n$$

Therefore the number of compositions of the integer  $n$  is equal to  $2^{n-1}$ . A simple way to recover this result is by noting that, representing the integer as a string of  $n$  symbols  $\bullet \bullet \bullet \dots \bullet$ , the number of partitions correspond to the number of ways this string can be split inserting commas in the  $n-1$  spaces between the symbols. Since in each of the  $n-1$  spaces a comma can be present or not, this makes  $2^{n-1}$  possible ways to split  $n$  as a sum of integers.

Partitions are more complicated objects and they are given by the multi-set  $\mathcal{P} = \text{MSet}(\mathcal{I})$ . Correspondingly

$$P(z) = e^{I(z) + \frac{1}{2}I(z^2) + \frac{1}{3}I(z^3) + \dots} \quad (7.15)$$

$$= \prod_{m=1}^{\infty} \frac{1}{1 - z^m} = (1 + z + z^2 + \dots)(1 + z^2 + z^4 + \dots) \dots \quad (7.16)$$

There is no explicit expression for the number of partitions of  $n$ , but Hardy and Ramanujan [14] derived the celebrated asymptotic result

$$p_n \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right).$$

Note that if we are interested in partitions in distinct parts, i.e. when  $x_1 < x_2 < \dots < x_k$ , then the powerset has to be used instead of the multi set  $\mathcal{P}_{\neq} = \text{PSet}(\mathcal{I})$ . Correspondingly

$$P_{\neq}(z) = e^{I(z) - \frac{1}{2}I(z^2) + \frac{1}{3}I(z^3) + \dots} = \prod_{m=1}^{\infty} (1 + z^m) \quad (7.17)$$

$$p_n^{\neq} \sim \frac{3^{3/4}}{12n^{3/4}} \exp\left(\pi\sqrt{\frac{n}{3}}\right) \quad (7.18)$$

where the last asymptotic expression was also derived by Hardy and Ramanujan.

## 7.4 Labeled objects

Generating functions are also useful if we want to count objects that are labeled. Take for instance a connected graph of  $n$  nodes where each node has a different label, say the integers from 1 to  $n$ . How many such graphs are there? The key point that we have to take into account is that, besides counting different graphs as before, we also need to count the number of different ways in which each graph can be labeled.

We'll not enter into much details here, but just mention the main fact and give a flavor of the method. Consider for example combining two objects, one from a set  $\mathcal{A}$  the other from set  $\mathcal{B}$  and let  $a_n$  and  $b_n$  be the number of labeled objects of size  $n$  in the two sets.

In order to count labeled objects  $C \in \mathcal{C} = \mathcal{A} \otimes \mathcal{B}$  one needs to account for the fact that each object  $C$  of size  $n$  will be composed of an object  $A \in \mathcal{A}$



of size  $k \in [0, n]$  and an object  $B \in \mathcal{B}$  of size  $n - k$ . There are  $\binom{n}{k}$  ways to label  $C$ , so

$$c_n = \sum_{k=0}^n \binom{n}{k} a_k b_{n-k}$$

Then if we construct *exponential* generating functions

$$\tilde{A}(z) = \sum_n \frac{a_n}{n!} z^n, \quad \tilde{B}(z) = \sum_n \frac{b_n}{n!} z^n \quad (7.19)$$

the generating function of labeled objects  $C \in \mathcal{C}$  will be given by the simple multiplicative formula

$$\tilde{C}(z) = \tilde{A}(z)\tilde{B}(z).$$

This indicates that for labeled objects one needs to use *exponential generating functions* such as the ones defined in Eq. (7.19). These satisfy further relations if one considers more complicated constructions.

The simplest example is that of permutations. In order to compute the number of permutations of integers, think of a permutation as a sequence of labeled symbols  $\bullet_k$  for  $k = 1, \dots, n$ . The generating function of permutations is again<sup>9</sup>

$$\tilde{P}(z) = \frac{z}{1-z} = \sum_n \frac{p_n}{n!} z^n$$

so that the number of permutations of  $n$  is given by  $p_n = n!$ .

## 7.5 Generating functions for integer random variables

Let<sup>10</sup>  $p_n = P\{X(\omega) = n\}$  be the probability distribution of an integer random variable  $X(\omega) \in \mathbb{N}$ . Consider the associated generating function

$$P(s) \equiv \sum_{n=0}^{\infty} p_n s^n = \mathbb{E}[s^X]. \quad (7.20)$$

The properties of this function close to  $s = 1$  give us a lot of information about  $X$ . First when we set  $s = 1$  in Eq. (7.20) we find  $P(1) = 1$ , by normalisation.

<sup>9</sup>We can construct  $\mathcal{P}$  from the relation  $\mathcal{P} = \{\bullet_1\} \cup \{\bullet_\ell\} \otimes \mathcal{P}$ , where  $\bullet_\ell$  is an object with a new label. This corresponds to  $\tilde{P}(z) = z + z\tilde{P}(z)$ . Note that the same generating function counts integers when objects are unlabeled permutations when they are labeled, i.e.  $\tilde{P}(z) = I(z)$ .

<sup>10</sup>This material is discussed in FELLER, Chapter XI.

Second, if we take a derivative and compute it at  $s = 1$  we get

$$P'(1) = \sum_{n=0}^{\infty} p_n n s^{n-1} \Big|_{s=1} = \mathbb{E}[X s^{X-1}] \Big|_{s=1} = \mathbb{E}[X]$$

Likewise, if we take  $k$  derivatives we find<sup>11</sup>

$$\frac{d^k P(s)}{ds^k} \Big|_{s=1} = \sum_{n=0}^{\infty} p_n (n)_k s^{n-k} \Big|_{s=1} = \mathbb{E}[(X)_k].$$

So, for example, the variance of  $X$  is given by

$$\mathbb{V}[X] = P''(1) + P'(1)[1 - P'(1)]. \quad (7.21)$$

The cumulative distribution  $q_n = P\{X > n\}$  of an integer random variable is also a sequence for which we can define a generating function  $Q(s)$ . This is related to  $P(s)$  because the probability that  $X \geq n$  equals the probability that  $X = n$  plus the probability that  $X > n$ , i.e.,  $q_{n-1} = p_n + q_n$  for  $n > 0$ . Multiplying this by  $s^n$  and summing over  $n > 0$  we obtain  $sQ(s) = P(s) - p_0 + Q(s) - q_0$ . Finally, observing that  $p_0 + q_0 = 1$  and rearranging terms, we find

$$Q(s) = \frac{1 - P(s)}{1 - s}. \quad (7.22)$$

Using de l'Hopital rule, we find  $Q(1) = \lim_{s \rightarrow 1} Q(s) = P'(1) = \mathbb{E}[X]$ .

### 7.5.1 Sums of variables and convolutions

Let  $X$  and  $Y$  be two independent integer random variables with distributions  $p_n = P\{X = n\}$  and  $r_n = P\{Y = n\}$ . Then the probability  $s_n = P\{X + Y = n\}$  is given by

$$s_n = p_0 r_n + p_1 r_{n-1} + \dots + p_n r_0.$$

This operation is called a *convolution*, i.e.  $s_n$  is a convolution of the sequences  $p_n$  and  $r_n$ . Then the generating function of the sum  $X + Y$  is given by the product of the generating functions of  $X$  and  $Y$ <sup>12</sup>

$$S(s) = \sum_{n=1}^{\infty} s_n s^n = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X] \mathbb{E}[s^Y] = P(s)R(s). \quad (7.23)$$

---

<sup>11</sup>Remember that

$$(n)_k = n(n-1)(n-2) \cdots (n-k+1).$$

<sup>12</sup>Here the symbol  $s$  is slightly abused.  $s_n$  is the sequence,  $S(s)$  is the generating function and  $s$  is the variable it depends on.  $R(s)$  is the generating function of  $Y$ .

because, for independent variables the expected value factorizes. This is so important that it makes sense to make all passages explicitly.

If  $P\{X = n, Y = m\} = P\{X = n\}P\{Y = m\} = p_n r_m$  then

$$\mathbb{E}[s^{X+Y}] = \sum_{n,m} s^{n+m} p_n r_m = \left[ \sum_n s^n p_n \right] \left[ \sum_m s^m r_m \right] = \mathbb{E}[s^X] \mathbb{E}[s^Y].$$

The same applies to the sum of many independent random variables. A particular case is that of independent and identically distributed (*i.i.d.*)<sup>13</sup> random variables  $X_1, \dots, X_n$ . Then the sum

$$\Sigma_n = \sum_{i=1}^n X_i$$

has the generating function

$$P_{\Sigma_n}(s) = \mathbb{E}[s^{X_1+\dots+X_n}] = \mathbb{E}[s^X]^n = P_X(s)^n \quad (7.24)$$

by exactly the same argument. Taking derivatives, you can easily find that the mean and the variance of the sum are related to that of the variable  $X$  by

$$\mathbb{E}[\Sigma_n] = n\mathbb{E}[X], \quad \mathbb{V}[\Sigma_n] = n\mathbb{V}[X]. \quad (7.25)$$

For example, let us consider binary random variables  $X_i = 0, 1$  with  $P\{X_i = 1\} = p$  and  $P\{X_i = 0\} = 1 - p$ . The generating function of  $X_i$  is  $P(s) = 1 - p + ps$ . The variable  $\Sigma_n$  obtained by summing  $n$  i.i.d. binary random variables, by Eq. (7.24), has generating function given by

$$B_{n,p}(s) = (1 - p + ps)^n = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} s^k$$

which is indeed the generating function of the binomial distribution. From this it is very easy to compute the mean and the variance by taking derivatives, as well as to obtain the generating function of the Poisson distribution

$$P_\lambda(s) = e^{-(1-s)\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} s^k \quad (7.26)$$

by taking the limit  $n \rightarrow \infty$  with  $p = \lambda/n$ .

<sup>13</sup>The abbreviation i.i.d. will be used frequently, so please memorise it.

**Exercise 7.4**

Derive Eqs. (7.25) and (7.26). Compute the mean and variance of the Binomial and Poisson distribution by taking derivatives of the generating function.

**Exercise 7.5**

Consider a coin tossing experiment that terminates when two consecutive heads occur for the first time. Compute the expected length  $\mathbb{E}[X_{\text{HH}}]$  of sequences so generated and its variance. Next, let  $X_{\text{HT}}$  be the length of the sequence of coin tosses that terminate when a head is followed by a tail for the first time. Is the distribution of  $X_{\text{HT}}$  the same as the distribution of  $X_{\text{HH}}$ ? If not, compute the expected length  $\mathbb{E}[X_{\text{HT}}]$  and the variance  $\mathbb{V}[X_{\text{HT}}]$ . What is the probability that  $X_{\text{HH}} < X_{\text{HT}}$ ? Comment the result.

The argument also runs in the reverse direction. If the generating function of a random variable  $X$  can be written as the  $n^{\text{th}}$  power of a generating function  $Q$ ,<sup>14</sup> i.e.  $P(s) = Q(s)^n$ , then  $X = Z_1 + \dots + Z_n$  can be written as the sum of  $n$  i.i.d. random variables  $Z_i$  with distribution  $P\{Z_i = k\} = q_k$ . For example, if  $X$  has a Poisson distribution with mean  $\lambda$ , Eq. (7.26) implies that, for any  $n > 0$ ,  $X = Z_1, \dots, Z_n$  can be considered as the sum of  $n$  i.i.d. random variables  $Z_i$  with generating function

$$Q(s) = P(s)^{1/n} = e^{-(1-s)\lambda/n},$$

which implies that the variables  $Z_i$  are themselves Poisson random variables with mean  $\lambda/n$ . The variables  $Z_i$  can, in their turn, be “divided” as  $Z_i = Y_{i,1} + \dots + Y_{i,m}$  into a sum of  $m$  other i.i.d. random variables  $Y_{i,j}$ , each of which has a Poisson distribution with mean  $\lambda/(nm)$ , and so on... Because of this property, the Poisson distribution is called *infinitely divisible*.<sup>15</sup>

Another interesting example is the negative binomial distribution Eq. (5.15)

<sup>14</sup>This requires that  $Q(s)$  has a power expansion in  $s$  with all non-negative coefficients  $q_k$ .

<sup>15</sup>There are other examples of infinitely divisible distributions (see later and FELLER XII.2). One can gain intuition on infinite divisibility of the Poisson distribution recalling that it describes the number  $X_T$  of events that occur in a time interval  $[0, T)$  of a Poisson process. The interval  $[0, T)$  can be divided in an arbitrary number  $n$  of non-overlapping intervals  $[t_i, t_{i+1})$  of size  $T_i = t_{i+1} - t_i$ , with  $t_1 = 0$ , and  $t_{n+1} = T$ . Clearly  $X_T = X_{T_1} + X_{T_2} + \dots + X_{T_n}$  and each of the variables  $X_{T_i}$  has a Poisson distribution with parameter  $T_i \mathbb{E}[X_T]/T$ .

which has the generating function

$$N_r(s) = \sum_{n=0}^{\infty} \binom{-r}{n} p^r (p-1)^n s^n = \left( \frac{p}{1 - (1-p)s} \right)^r. \quad (7.27)$$

This is evidently the generating function of a random variable that is the sum of  $r$  random variables with a geometric distribution (Eq. (5.15) with  $r = 1$ ). Indeed, Eq. (5.15) describes the number of failures one has to wait before the  $r^{\text{th}}$  success in Bernoulli trials. This is clearly the failures that one has to wait before the first success plus the failures between the first and the second success, and so on. The number of failures between each pair of consecutive successes is a random variable with a geometric distribution, so  $N_r(s)$  is the generating function of the sum of  $r$  such random variables.

Notice also that, if we generalise to  $r \in \mathbb{R}_+$ , then also the negative binomial distribution is infinitely divisible, i.e. a negative binomial random variable can be written as the sum of  $n$  i.i.d. random variables with parameter  $r/n$ , for any  $n$ .

Another interesting limit of the negative binomial is obtained for  $r \rightarrow \infty$  with  $p = 1 - \lambda/r$ , i.e. when successes become more and more likely as  $r$  increases. Then we recover the Poisson distribution

$$\lim_{r \rightarrow \infty: p=1-\frac{\lambda}{r}} N_r(s) = e^{-(1-s)\lambda}.$$

Can you figure out why this result should be expected?<sup>16</sup>

### 7.5.2 Sums of a random number of random variables

A further very practical use of generating functions is in problems that involve a sum of a random number of integer random variables, as the one discussed at the end of the introductory Section:

Mr X checks emails every minute with probability  $p$ . He receives on average  $\lambda$  emails per minute. What is the probability that Mr X finds no email the next time he checks?

If  $X_i$  is the number of emails received in the  $i^{\text{th}}$  minute, we're interested in the sum

$$\Sigma_T = X_1 + \dots + X_T \quad (7.28)$$

<sup>16</sup>Hint: invert successes with failures.

where  $T$  is the number of minutes elapsed before Mr X looks at his emails.  $T$  is a random variable with a geometric distribution

$$g_t = P\{T = t\} = p(1 - p)^{t-1}. \quad (7.29)$$

Here the factor  $(1 - p)^{t-1}$  is the probability that Mr X does not check emails for the first  $t - 1$  minutes, whereas  $p$  is the probability that he checks email at time  $t$ . Let

$$G(s) = \sum_{t=1}^{\infty} g_t s^t = \mathbb{E}[s^T]$$

be the generating function of  $T$ , and  $F(s) = \mathbb{E}[s^X]$  be the generating function of the variables  $X$ . Then the generating function of the variable  $\Sigma_T$  is given by

$$H(s) = \mathbb{E}[s^{\Sigma_T}] = \sum_{t=1}^{\infty} g_t \mathbb{E}[s^{X_1 + \dots + X_t}] \quad (7.30)$$

$$= \sum_{t=1}^{\infty} g_t \mathbb{E}[s^X]^t = \sum_{t=1}^{\infty} g_t [F(s)]^t \quad (7.31)$$

$$= G(F(s)). \quad (7.32)$$

In the particular case of Mr X,  $F(s) = e^{-\lambda(1-s)}$  is the generating function of a Poisson random variable and  $G(s) = \frac{ps}{1-(1-p)s}$ . Therefore the generating function of  $\Sigma_T$  is given by

$$H(s) = \frac{pe^{-\lambda(1-s)}}{1 - (1-p)e^{-\lambda(1-s)}}.$$

The probability that Mr X will find no emails is  $H(0) = \frac{p}{e^\lambda - 1 + p}$ , as stated in Eq. (1).

### Exercise 7.6

Why is the probability that  $\Sigma_T = 0$  related to  $H(s)$  for  $s = 0$ ?

The mean and the variance of  $\Sigma_T$  are obtained by taking derivatives of  $H(s)$ . Using Eq. (7.32) this can be related to the mean and the variance of  $X$  and  $T$  as

$$\mathbb{E}[\Sigma_T] = \mathbb{E}[T]\mathbb{E}[X], \quad \mathbb{V}[\Sigma_T] = \mathbb{E}[T]\mathbb{V}[X] + \mathbb{V}[T]\mathbb{E}[X]^2. \quad (7.33)$$

The equation for the mean is the natural generalisation of the case where the number of random variables in the sum is fixed (Eq. (7.25)). Yet the variance

acquires an additional term (the second), due to the fluctuations of  $T$ . Indeed this expression reduces to Eq. (7.25) when  $T = n$  is fixed.

### Exercise 7.7

Derive Eq. (7.33)

### Exercise 7.8

In a repeated Bernoulli trial scheme, where success occurs with probability  $p$  and failure with probability  $1 - p$ , let  $T$  be the waiting time for the first occurrence of two consecutive successes. Using the same idea of the exercise on binary sequences  $\mathcal{B}_{11}$ , i) derive a recursion relation for  $p_n = P\{T = n\}$ , ii) compute the generating function and iii) confirm the asymptotic behaviour  $p_n \sim n^{-\beta} e^{-\alpha n}$  and compute  $\alpha$  and  $\beta$ .

When  $T$  is a Poisson random variable given by Eq. (7.28) where  $X_i$  are i.i.d. RV with generating function  $F(s)$ , the variable  $\Sigma_T$  has a *compound Poisson* distribution. This case is of particular interest, because then the random variable  $\Sigma_T$  has an infinitely divisible distribution. Indeed its generating function is given by

$$H_\lambda(s) = e^{-\lambda[1-F(s)]}. \quad (7.34)$$

and  $H_{\lambda/n}(s)$  clearly is also a generating function of a probability distribution for any  $n$ . The converse is also true, as shown in FELLER XII.2: every infinitely divisible distribution of integer random numbers is a compound Poisson process, i.e. its generating function has the form of Eq. (7.34). FELLER XII.2 provides a practical criterium for a generating function  $H(s)$  to be infinitely divisible. This requires that i)  $H(1) = 1$  and

$$\log \frac{H(s)}{H(0)} = \sum_{k=1}^{\infty} a_k s^k$$

with ii)  $a_k \geq 0$  for all  $k > 0$  and iii)  $\lambda = \sum_{k=1}^{\infty} a_k < +\infty$ . In this case,  $a_k/\lambda = P\{X = k\}$  is the probability distribution of an integer random variables  $X$ , such that  $H(s)$  is the generating function of the variable  $\Sigma_T$ , where  $T$  is a Poisson random variable with mean  $\lambda$ .

### Exercise 7.9

Show that the negative binomial distribution is a compound Poisson

process, because it satisfies Eq. (7.34) with  $\lambda = -r \log p$  and

$$f_n = P\{X_i = n\} = \frac{1}{\log p^{-1}} \frac{(1-p)^n}{n},$$

for  $n > 0$  and  $f_0 = 0$ . This is known as the *logarithmic distribution*.

The dependence on  $\lambda$  of infinitely divisible generating functions can be derived by observing that the equation  $H_{\lambda+\lambda'}(s) = H_\lambda(s)H_{\lambda'}(s)$  implies that  $H_0(s) = H_0(s)^2 = 1$ . Furthermore, with  $\lambda' = d\lambda$  one finds, to leading order in  $d\lambda$ ,

$$H_{\lambda+d\lambda}(s) = H_\lambda(s) \left[ 1 + \left. \frac{\partial H_\lambda(s)}{\partial \lambda} \right|_{\lambda=0} d\lambda \right].$$

Upon integration in  $\lambda$ , this leads to Eq. (7.34) with  $F(s) = 1 - \left. \frac{\partial H_\lambda(s)}{\partial \lambda} \right|_{\lambda=0}$ .

Some further intuition on the nature of infinitely divisible distributions can be gained by the following construction: consider a random variable  $X_\lambda$  that depends on a continuous parameter  $\lambda$  such that  $X_0 = 0$  and, for an infinitesimal  $d\lambda$ ,

$$X_{\lambda+d\lambda} = \begin{cases} X_\lambda & \text{with probability } 1 - d\lambda \\ X_\lambda + Z & \text{with probability } d\lambda \end{cases} \quad (7.35)$$

where  $Z \in \mathbb{N}$  is an integer random variable with  $\mathbb{E}[s^Z] = F(s)$  which is independent of  $X_\lambda$ . Then

$$H_{\lambda+d\lambda}(s) = \mathbb{E}[s^{X_{\lambda+d\lambda}}] = (1 - d\lambda)\mathbb{E}[s^{X_\lambda}] + d\lambda\mathbb{E}[s^{X_\lambda+Z}] \quad (7.36)$$

$$= H_\lambda(s) - H_\lambda(s)[1 - F(s)]d\lambda \quad (7.37)$$

where we used the fact that  $Z$  and  $X_\lambda$  are independent random variables. Eq. (7.34) is obtained by integrating this equation in  $d\lambda$  from 0 to  $\lambda$  with initial condition  $H_0(s) = 1$  (i.e.  $X_0 = 0$ ). In words, every infinitely divisible random variable  $X_\lambda$  is an increasing random function of  $\lambda$ , which increases in random steps, each of which is drawn independently from a distribution with generating function  $F(s)$ .



### 7.5.3 Cumulant generating function

There is a simpler way than Eq. (7.21) to compute  $\mathbb{V}[X]$  from a generating function  $P(s)$ . Set  $s = e^z$  and take the logarithm of the generating function<sup>17</sup>

$$\psi(z) \equiv \log \Psi(z), \quad \Psi(z) \equiv P(s = e^z) = \mathbb{E}[e^{zX}]. \quad (7.38)$$

Notice that the function<sup>18</sup>  $\Psi$  admits the power expansion

$$\Psi(z) = \sum_{m=0}^{\infty} \frac{\mathbb{E}[X^m]}{m!} z^m$$

which means that the  $n^{\text{th}}$  derivative of  $\Psi$  equals  $\mathbb{E}[X^m]$ . By normalisation  $\Psi(0) = 1$  so  $\psi(0) = 0$ . The mean and the variance of  $X$  can be obtained from the first two derivatives of  $\psi$ :

$$\left. \frac{d\psi(z)}{dz} \right|_{z=0} = \frac{1}{\Psi(0)} \mathbb{E}[X] = \mathbb{E}[X] \quad (7.39)$$

$$\left. \frac{d^2\psi(z)}{dz^2} \right|_{z=0} = \frac{1}{\Psi(z)} \left. \frac{d^2\Psi(z)}{dz^2} \right|_{z=0} - \left( \frac{1}{\Psi(z)} \left. \frac{d\Psi(z)}{dz} \right|_{z=0} \right)^2 \quad (7.40)$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{V}[X]. \quad (7.41)$$

The function  $\psi(z)$  is called the *cumulant generating function*, because the coefficients  $C_m$  of the expansion of  $\psi(z)$  in powers of  $z$  are called the cumulants

$$\psi(z) = \log \mathbb{E}[e^{zX}] = \sum_{m=0}^{\infty} \frac{C_m}{m!} z^m. \quad (7.42)$$

Clearly  $C_1 = \mathbb{E}[X]$  and  $C_2 = \mathbb{V}[X]$ . Higher order cumulants can be related to moments by comparing the coefficient of  $z^n$  in the expansion of  $\Psi(z)$  with the coefficient of  $z^n$  in the expansion of

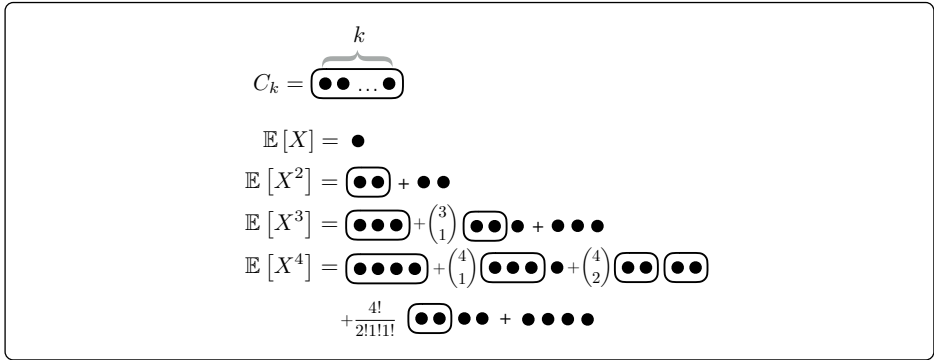
$$e^{\psi(z)} = \sum_{\ell=0}^{\infty} \frac{\psi(z)^\ell}{\ell!}.$$

Using Eq. (7.42) in each of the  $\ell$  factors  $\psi(z)$  that appear in this equation, one obtains

$$\mathbb{E}[X^n] = \sum_{\ell=0}^{\infty} \sum_{k_1=0}^{\infty} \dots \sum_{k_\ell=0}^{\infty} \frac{1}{\ell!} \frac{n!}{\prod_m k_m!} \prod_{m=1}^{\ell} C_{k_m} \delta_{n, \sum_m k_m}. \quad (7.43)$$

<sup>17</sup>Note that convergence of the series that defines  $P(s)$  for  $s \in [0, 1]$  implies that  $\psi(z)$  is well defined for  $z \leq 0$ .

<sup>18</sup>Which is why  $\Psi$  is called the *moment generating function*.



**Figure 13.** The relation between moments and cumulants.

Curiously, there is a diagrammatic method to derive the moment of order  $n$  in terms of the cumulants of order  $k \leq n$ , in terms of distribution of balls in unordered boxes: represent the cumulant of order  $k$  as  $k$  balls in a box. Then the  $n^{\text{th}}$  moment is equal to the sum of all ways to group  $n$  balls in unordered (because of the factor  $1/\ell!$ ) boxes containing  $k_1, \dots, k_\ell$  balls (with  $k_1 + \dots + k_\ell = n$  because of the Kronecker delta). For the first moment there is only one ball to group, so  $\mathbb{E}[X] = C_1$ , for the second moment, there are two balls and two ways to group them, either in a box of two or as two isolated balls. Correspondingly  $\mathbb{E}[X^2] = C_2 + C_1^2$ . For  $n = 3$ ,  $\mathbb{E}[X^3] = C_3 + 3C_2C_1 + C_1^3$ , where the factor 3 comes from the fact that there are three ways to choose the isolated point. Hence  $C_3 = \mathbb{E}[X^3] - 3\mathbb{E}[X]\mathbb{E}[X] - \mathbb{E}[X]^3$ . Derive the fourth order cumulant, as an exercise.

### Exercise 7.10

Show that all cumulants of a Poisson distribution with mean  $\lambda$  are equal to  $\lambda$ .

The cumulant generating function (CGF) is also very practical when dealing with sums of random variables. Indeed if  $X_1$  and  $X_2$  are two independent random variables with CGF  $\psi_1(z)$  and  $\psi_2(z)$ , respectively, then the CGF of  $X_1 + X_2$  is the sum of the CGFs:

$$\psi_{1+2}(z) = \log \mathbb{E}[e^{z(X_1+X_2)}] = \log \mathbb{E}[e^{zX_1}] + \log \mathbb{E}[e^{zX_2}] = \psi_1(z) + \psi_2(z).$$

This extends to sums of  $n$  i.i.d. random variables  $\Sigma_n = X_1 + \dots, X_n$  with CGF  $\psi(z)$ . The CGF of the sum is  $\psi_{\Sigma_n}(z) = n\psi(z)$ . From this, computing derivatives at  $z = 0$ , it is straightforward to see that  $\mathbb{E}[\Sigma_n] = n\mathbb{E}[X_i]$  and  $\mathbb{V}[\Sigma_n] = n\mathbb{V}[X_i]$ .

This extends also to sums  $\Sigma_T$  of a random number of random variables. With the notation used earlier, let

$$F(s) = e^{\phi(z)}, \quad G(s) = e^{\gamma(z)}, \quad H(s) = e^{\eta(z)}, \quad s = e^z$$

Then

$$\eta(z) = \log H(s) = \log G(F(s)) = \gamma(\phi(z)). \quad (7.44)$$

For example, this makes the derivation of Eq. (7.33) much simpler.<sup>19</sup>

For an infinitely divisible distribution with generating function  $H_{\lambda+\lambda'}(s) = H_\lambda(s)H_{\lambda'}(s)$ , the CGF  $\eta_\lambda(z) = \log H'_\lambda(e^z)$  satisfies the additive relation

$$\eta_{\lambda+\lambda'}(z) = \eta_\lambda(z) + \eta_{\lambda'}(z). \quad (7.45)$$

---

<sup>19</sup>The trick of deriving cumulants from derivatives of the logarithm of a generating function is of widespread use in statistical mechanics, as we shall see.



# Chapter 8

## More on balls and boxes\*

To understand a complex system, you must first understand its simplest possible instance (H. Simon, 1969)

In order to consolidate what we have learned so far, consider the following problem: imagine to distribute  $r$  balls in  $n$  boxes. What<sup>1</sup> is the probability  $p_m(r, n)$  that exactly  $m$  cells are empty?

Let us first focus on the case  $m = 0$ . If  $A_i$  is the event that box  $i$  is empty, then the event we're interested in is the one where none of these events occur, i.e.,

$$A_0 = \cap_{i=1}^n \bar{A}_i = \overline{\bigcup_{i=1}^n A_i}.$$

This same problem can be formulated in terms of waiting times. Imagine balls are drawn in the boxes one at a time. Then we can ask how many balls need to be added in order for the condition that no box is empty, is met for the first time. The number of balls we have to “wait” for  $A_0$  to occur is a random variable  $T$ , which is a waiting time. Clearly  $p_0(r, n) = P\{T \leq r\}$  which means that the distribution of  $T$  is given by

$$f_r^{(n)} \equiv P\{T = r\} = p_0(r, n) - p_0(r - 1, n) \quad (8.1)$$

because of the relation between the events  $\{T \leq r - 1\} \subset \{T \leq r\}$  and  $\{T = r\} = \{T \leq r\} / \{T \leq r - 1\}$ .

The probability that box  $i$  is empty is  $P(A_i) = (1 - 1/n)^r$  and the probability that boxes  $i_1, \dots, i_\nu$  are empty is  $P\{A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_\nu}\} = \left(1 - \frac{\nu}{n}\right)^r$ . Then, the

---

<sup>1</sup>These problems are discussed also in FELLER IV.2, that you are suggested to read.

generalised sub-additivity rule (Eq. (3.11)), with

$$S_\nu = \sum_{i_1 < i_2 < \dots < i_\nu} P\{A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_\nu}\} = \binom{n}{\nu} \left(1 - \frac{\nu}{n}\right)^r$$

leads to

$$p_0(r, n) \equiv P\{A_0\} = \sum_{\nu=0}^n (-1)^\nu \binom{n}{\nu} \left(1 - \frac{\nu}{n}\right)^r. \quad (8.2)$$

In the rest of this chapter we'll see how the answers to the following simple questions can be extracted from this complicated formula:

- 1) Is Eq. (8.2) consistent with the expectation that  $p_0(r, n) = 0$  for all  $r < n$ ?
- 2) How many balls we expect we have to draw to fill all the boxes with at least one ball?
- 3) Can we approximate the distribution of  $T$  for large  $n$ ?

A deep understanding of a problem is not only intellectually satisfying but it also allows to solve practical problems, such as

- 4) How can we draw a value of  $T$  from the distribution (8.1)?

One way to address the first question is to build the generating function

$$P_n(s) = \sum_{r=0}^{\infty} p_0(r, n) s^r = \sum_{\nu=0}^n (-1)^\nu \binom{n}{\nu} \left[1 - \left(1 - \frac{\nu}{n}\right)s\right]^{-1}.$$

A better expression can be derived using the identity  $q^{-1} = \int_0^\infty dt e^{-qt}$ , so that one can sum the binomial expansion and find

$$P_n(s) = \int_0^\infty dt e^{-(1-s)t} \left(1 - e^{-st/n}\right)^n = \frac{n}{s} B(n(1-s)/s, n+1)$$

where  $B(a, b)$  is the Beta function<sup>2</sup> and we made the change of variables  $u = e^{-st/n}$  in the last equation. Using the recursion  $B(a, b+1) = \frac{b}{a+b} B(a, b)$  iteratively  $n+1$  times, and the identity  $B(a, 1) = 1/a$ , we find

$$P_n(s) = n! s^n \prod_{k=1}^n (n - ks)^{-1}. \quad (8.3)$$

---

<sup>2</sup>The Beta function is defined as

$$B(a, b) = \int_0^1 dx x^{a-1} (1-x)^{b-1}$$

for  $a, b \in \mathbb{C}$  and  $\text{Re}(a), \text{Re}(b) > 0$ . It is related to the Gamma function by  $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ .

From this and Eq. (7.4), it is clear that all coefficients of  $s^r$  in the expansion of  $P_n(s)$  vanish, i.e.  $p_0(n, r) = 0$ , for  $n < r$ . This answers question 1) above.

### Exercise 8.1

Let  $A(r, n)$  be the number of ways to distribute  $r$  particles in  $n$  boxes so that no box is empty. Show, by a combinatorial argument, that

$$A(r, n+1) = \sum_{k=1}^r \binom{r}{k} A(r-k, n).$$

Show that the generating function  $A_n(s) = \sum_{r=0}^{\infty} A(r, n)s^r$  is given by  $A_n(s) = P_n(sn)$  and check that Eq. (8.3) is consistent with the recursion equation above. [Hint: using Eq. (7.4) show that

$$A_{n+1}(s) = \frac{1}{1-s} A_n\left(\frac{s}{1-s}\right) - A_n(s),$$

and show that  $A_n(s) = P_n(sn)$  satisfies this equation].

How many balls are needed to fill all boxes? In a realisation of drawing balls in  $n$  boxes, let  $T$  be the smallest number of balls for which all boxes contain at least one ball. As stated before,  $T$  has the distribution in Eq. (8.1). Its associated generating function can be computed using Eq. (8.3):

$$F_n(s) = \sum_{r=0}^{\infty} P\{T = r\} s^r = (1-s)P_n(s) = (n-1)! s^n \prod_{k=1}^{n-1} (n-ks)^{-1}.$$

One important property of this function is that its value at  $s = 1$  gives  $F_n(1) = \sum_{r=0}^{\infty} P\{T = r\} = 1$ . This implies that  $P\{T = r\}$  is correctly normalised, i.e. that sooner or later all boxes will be filled with at least one ball. Another property is that the expected value of the waiting time  $T$  is given by

$$\mathbb{E}[T] = F'_n(1) = 1 + n \sum_{k=1}^{n-1} \frac{1}{k} \simeq n \log n + n\gamma + \frac{3}{2} + O(1/n) \quad (8.4)$$

where  $\gamma = 0.5772156649 \dots$  is Euler constant. This implies that, for large  $n$ , one needs approximately  $\log n + \gamma$  balls per box, in order to satisfy the condition  $A_0$ . This answers question 2).

The analysis of  $F_n(s)$  can give us more information on the distribution of  $T$ . Yet extracting the asymptotic behaviour of  $f_t^{(n)}$  from  $F_n(s)$ , when  $n \rightarrow \infty$ ,

is not easy<sup>3</sup> because the interesting range of values of  $T$  changes with  $n$ .

What makes this problem complicated is that the events  $A_i$  are not independent. In order to see this, let us focus on a single box and let  $B_i$  be the number of balls that fall in this box after  $r$  draws. For each ball the probability that it will fall in box  $i$  is  $1/n$ . So  $B_i$  has a binomial distribution that, for  $n \gg 1$  is well approximated by a Poisson distribution

$$P\{B_i = k\} = \binom{r}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{r-k} \simeq \frac{(r/n)^k}{k!} e^{-r/n}$$

If all the variables  $B_i$  were independent Poisson variables, then the total number of particles  $R = B_1 + \dots + B_n$  would also be a Poisson random variable. A simple way to see this is to compute the generating function of  $R$

$$\mathbb{E}[s^R] = \prod_{i=1}^n \mathbb{E}[s^{B_i}] = [e^{-(1-s)r/n}]^n = e^{-(1-s)r},$$

which is the generating function of a Poisson random variable with mean  $r = \mathbb{E}[R]$ . The origin of the difficulty of the original problem is that the number of balls is fixed, i.e.  $B_1 + \dots + B_n = r$ . This is what makes the random variables  $B_i$  dependent. This intuition is confirmed by the fact that, if we compute the expected value of  $p_0(R, n)$ , when  $R$  is drawn from a Poisson distribution with mean  $n\rho$ , we find

$$\mathbb{E}[p_0(n, R)] = \sum_{r=0}^{\infty} p_0(n, r) \frac{(\rho n)^r}{r!} e^{-\rho n} = (1 - e^{-\rho})^n. \quad (8.5)$$

This formula is much simpler than Eq. (8.2) and it has a simple interpretation, if we revert our argument. The expected value of balls that fall in each box  $i$  is  $\mathbb{E}[B_i] = \mathbb{E}[R/n] = \rho$ .  $B_i$  has a Poisson distribution with expected value<sup>4</sup>  $\rho$  then  $1 - e^{-\rho} = P\{B_i > 0\}$  is the probability that box  $i$  is not empty. When the number of balls  $R$  is drawn from a Poisson distribution, all boxes become independent.<sup>5</sup> Hence the probability that no box is empty takes the simple form of Eq. (8.5).

<sup>3</sup>This function has  $n - 1$  simple poles at  $s_k = n/k$  for  $k = 1, \dots, n - 1$ . So  $f_t^{(n)} \sim (1 - 1/n)^t$  for  $n$  small. Yet when  $n \rightarrow \infty$  the poles become densely concentrated in the neighbourhood of  $s = 1$ , which complicates the analysis.

<sup>4</sup>This is because the Poisson distribution is an *infinitely divisible* distribution, as explained before and in FELLER XXII.2.

<sup>5</sup>This technique of removing the dependence of integer random variables due to a constraint  $B_1 + \dots + B_n = r$ , by substituting  $r$  with a Poisson random variable  $R$  with mean  $r$  is called *poissonisation*. It is a very useful trick worth being remembered. When  $r \gg 1$  this



### Exercise 8.2

Let

$$p(x_1, \dots, x_n | r) = A_r \prod_{i=1}^n p(x_i) \delta_{\sum_i x_i, r} \quad (8.6)$$

be the probability that  $B_i = x_i$  for all  $i$ , where  $A_r$  is a normalisation constant. By Eq. (8.6), the variables  $B_i$  would be independent with  $P\{B_i = x_i\} = p(x_i)$ , where it not for the constraint. Prove that if  $r$  is replaced by a random variable  $R$  with distribution  $q(r) = A_r^{-1}$ , then

$$\sum_{r=0}^{\infty} p(x_1, \dots, x_n | r) q(r) = \prod_{i=1}^n p(x_i).$$

Check that  $q(r)$  is correctly normalised, and that when  $B_i$  are Poisson random variables with mean  $\rho$  then  $q(r)$  is the Poisson distribution with mean  $n\rho$ .

### Exercise 8.3

Let  $K_1, \dots, K_d, K_i \in \mathbb{N}$  have a multinomial distribution with probabilities  $p_1, \dots, p_d$ , and  $\sum_{i=1}^d K_i = n$ . Show that if  $n$  is replaced by a Poisson random variable  $N$  with mean  $\nu d$ , then the variables  $K_1, \dots, K_n$  become independent Poisson random variables with mean  $\mathbb{E}[K_i] = \nu p_i$ .

Now we can go back to the issue of estimating the probability distribution of  $T$  for  $n$  large. Since  $\mathbb{E}[T] \sim n(\log n + \gamma)$  it makes sense to make a change of variables  $T = n(\log n + X)$  and study the distribution of the random variable  $X$  instead. Treating again  $R$  as a Poisson random variable, we have

$$\begin{aligned} P\{T \leq n[\log n + x]\} &= \mathbb{E}[p_0(n, R)] \quad \text{with } \mathbb{E}[R] = n(\log n + x) \\ &= \left(1 - \frac{e^{-x}}{n}\right)^n \\ &\simeq e^{-e^{-x}}, \quad (n \rightarrow \infty) \end{aligned} \quad (8.7)$$

where we used Eq. (8.5) with  $\rho = \log n + x$ . Note that  $P\{T \leq n(\log n + x)\} = P\{X \leq x\}$  yields the cumulative distribution of the variable  $X$ . Hence the pdf

approximation is accurate because the fluctuations  $\delta R$  of  $R$  are of order  $\sqrt{r}$ , and hence they are small compared to the mean  $\mathbb{E}[R] = r$ . In the present case, this implies that the statistical dependence between different boxes becomes weaker and weaker as  $r$  increases. Similar considerations apply in general.

of  $X$ , asymptotically for  $n \rightarrow \infty$ , is obtained taking a derivative of Eq. (8.7):

$$p(x) = e^{-x-e^{-x}}. \quad (8.8)$$

This is called the *Gumbel* distribution<sup>6</sup> and it will be discussed at length when we discuss the distribution of the maximum of many independent random variables. Its occurrence in this problem is not accidental. Indeed, let  $T_i$  be the number of balls that you have to draw in order to occupy box  $i$  for the first time. Then clearly

$$T = \max_{i=1,\dots,n} T_i \quad (8.9)$$

is precisely given by the maximum of  $n$  independent random variables. In order to see why the problem in Eq. (8.9) leads asymptotically to the Gumbel distribution, notice that

$$\begin{aligned} P\left\{\max_{i=1,\dots,n} T_i \leq t\right\} &= P\{T_i \leq t \ \forall i = 1, \dots, n\} \\ &= P\{T_i \leq t\}^n = [1 - P\{T_i > t\}]^n. \end{aligned} \quad (8.10)$$

The event that box  $i$  gets the first ball at each draw has probability  $1/n$ , therefore  $P\{T_i > t\} = (1 - 1/n)^t \simeq e^{-t/n}$ . Inserting this in Eq. (8.10) with  $t = n(\log n + x)$ , one finds

$$P\{T \leq n(\log n + x)\} \simeq [1 - e^{-\log n - x}]^n \simeq e^{-e^{-x}} \quad (8.11)$$

which is Eq. (8.7), as anticipated.

Let us now compute the probability  $p_m(r, n)$  that after the draw of  $r$  balls, exactly  $m$  boxes remain empty. If  $m$  cells are empty,  $n - m$  cells must be occupied and there are  $\binom{n}{m}$  ways to chose the  $m$  empty cells. Now the probability

<sup>6</sup>Notice that, if  $X$  has a Gumbel distribution,  $\mathbb{E}[X] = \gamma$  and  $\mathbb{V}[X] = \frac{\pi^2}{6}$ , which can be computed by taking derivatives of

$$\psi(\lambda) = \log \int_{-\infty}^{\infty} dx e^{-\lambda x - e^{-x}} = \log \Gamma(\lambda)$$

at  $\lambda = 1$ .

So, for large  $n$ ,  $\mathbb{E}[T] \simeq n(\gamma + \log n)$ , which agrees with Eq. (8.4) to leading order, and  $\mathbb{V}[T] = \frac{\pi^2}{6} n^2$  (i.e. the fluctuations of  $T$  are of order  $\delta T \propto n$ ). Notice also that, assuming  $T = n(\log n + X)$ , with  $X$  distributed according to Eq. (8.8), the probability that  $T < n$  is non-zero. Yet  $T < n$  corresponds to  $X < -\log n$ , and by Eq. (8.7) the probability  $P\{T < n\} \simeq P\{X < -\log n\} \simeq e^{-n}$  is negligible. This is because  $p(x)$  falls off very fast as  $x \rightarrow -\infty$ .

that no ball falls in the  $m$  empty cells is  $(1 - m/n)^r$  and the probability that the remaining  $n - m$  cells are all occupied is  $p_0(r, n - m)$ . Therefore

$$\begin{aligned} p_m(r, n) &= \binom{n}{m} \left(1 - \frac{m}{n}\right)^r p_0(r, n - m) \\ &= \binom{n}{m} \sum_{\nu=0}^{n-m} (-1)^\nu \binom{n-m}{\nu} \left(1 - \frac{m+\nu}{n}\right)^r. \end{aligned} \quad (8.12)$$

Again *poissonisation*, i.e. replacing  $r$  by a Poisson random variable with mean  $n\rho$  and taking the expected value, yields a much more transparent expression

$$\mathbb{E}[p_m(R, n)] = \binom{n}{m} e^{-m\rho} (1 - e^{-\rho})^{n-m}, \quad (8.13)$$

which has the same, simple, interpretation as Eq. (8.5). In particular, setting  $\rho = \log(n/\lambda)$  and taking  $n \rightarrow \infty$ , one finds again the Poisson distribution<sup>7</sup>

$$\lim_{n \rightarrow \infty, \rho = \log \frac{n}{\lambda}} \mathbb{E}[p_m(R, n)] = \frac{\lambda^m}{m!} e^{-\lambda}. \quad (8.14)$$

Hence  $\lambda = ne^{-r/n}$  approximates the expected number of empty boxes, for  $r \gg n \gg 1$ .

#### Exercise 8.4

Derive Eq. (8.13).

#### Exercise 8.5

In a town of 2000 inhabitants, each citizen gives a party on his/her birthday. Estimate the probability that there are no days without a party.

The understanding that we have reached is not only theoretical, but it allows us to sample values of  $T$ , for  $n \gg 1$ , much more efficiently than by drawing balls one by one until no empty box is left. This latter process would take  $\mathbb{E}[T] \sim n \log n$  steps. Can we do better?

<sup>7</sup>FELLER IV.2 finds the same approximation, observing that the terms that dominate the sum in Eq. (8.12) for  $n \rightarrow \infty$  with  $\lambda = ne^{-r/n}$  fixed, are those for finite  $\nu$ . Hence in  $p_0(r, n)$ , for  $r \gg n \gg 1$

$$\binom{n}{\nu} \left(1 - \frac{\nu}{n}\right)^r \simeq \frac{1}{\nu!} (ne^{-r/n})^\nu = \frac{\lambda^\nu}{\nu!}$$

The sum on  $\nu$  can be extended to  $+\infty$  and it yields  $p_0(r, n) \simeq e^{-\lambda}$ .

Here is one idea. Imagine there are already  $r$  balls and  $n_0$  boxes are empty. The probability that the next ball will fall in an empty box is  $n_0/n$ . Then the number of additional balls that need to be drawn before a further box is occupied (i.e. for  $n_0 \rightarrow n_0 - 1$ ) has the distribution

$$P\{\delta R = m\} = \frac{n_0(n - n_0)^{m-1}}{n^m},$$

which is the geometric distribution. Draw an integer  $\delta R$  from this distribution, increase  $r \rightarrow r + \delta R$  and decrease  $n_0 \rightarrow n_0 - 1$ . Repeat this process until there are no empty boxes ( $n_0 = 0$ ). Then  $T = \sum_{i=1}^n \delta R_i$  is the sum of the number of added balls in this process, starting from  $n_0 = n$  to  $n_0 = 0$ . This takes  $n$  steps. It's better than the naïve algorithms above by a factor  $\log n$ .

We can do better though.

Start from  $n_0 = n$  and  $T = 0$ . We know that if we draw  $R$  as a Poisson random variable with mean  $n\rho$ , then the number of balls that fall in each box is an independent Poisson random number with mean  $\rho$ , and the probability that box  $i$  is empty is  $e^{-\rho}$ . Therefore the number of empty boxes ( $n'_0$ ) after the  $R$  balls are drawn is reduced, and it is a Bernoulli random variable with  $p = e^{-\rho}$  over  $n_0$  trials. Clearly  $\rho$  cannot be too big otherwise you hit the  $n'_0 = 0$  condition. Set  $T \rightarrow T + R$  and  $n_0 \rightarrow n'_0$  to the new value and you're left with a very similar problem. If you draw another value of  $R$  in the same way and distribute these balls among the boxes, each of the  $n'_0$  empty boxes will remain empty with probability  $e^{-\rho}$ . Then you can again draw the new value of  $n''_0$  from a binomial distribution, as before, and set  $T \rightarrow T + R$ . This step can be repeated until  $n_0 = 0$ . In each step you need to draw only two random numbers ( $R$  and  $n_0$ ) and this process will end in  $O(\log n)$  steps, which is much more efficient than the previous algorithms.

### Exercise 8.6

Implement these algorithms on your computer and verify that the distribution of  $X$  is well approximated by Eq. (8.8) for  $n$  sufficiently large.

## Chapter 9

# Random walks

The random walk is Nature's way of exploring possibilities – from particle collisions to evolutionary mutations. (Freeman J. Dyson, *Infinite in all directions*, 1988).

Let  $X_1, X_2, \dots, X_n, \dots$  be a sequence of i.i.d. random variables that take values  $X_i = \pm 1$  with equal probability  $P\{X_i = \pm 1\} = \frac{1}{2}$ . A random walk is defined as the sum

$$S_n = \sum_{i=1}^n X_i, \quad S_0 = 0.$$

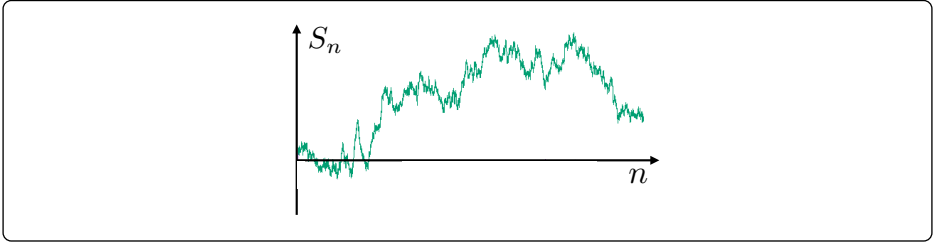
The name comes from considering  $n$  as a “time” variable, i.e. the number of steps in the walk and  $S_n$  as the position of a particle at time  $n$ . Hence  $X_i$  is the  $i^{\text{th}}$  step of the walk. If  $X_i = -1$ , the walker moves to the left by one step and if  $X_i = +1$  it moves to the right. The trajectory  $(n, S_n)$  for  $n \geq 0$  can be displayed on a graph as shown in Figure 14. Each of these trajectories or *paths* correspond to a realisation  $\omega$  of the sample space.<sup>1</sup> The sample space  $\Omega$  of a random walk of  $n$  steps contains  $|\Omega| = 2^n$  paths  $\omega$ , each of which has the same probability  $P(\omega) = 2^{-n}$ .

Many problems in classical probability can be reduced to studying properties of random walks, as in the case of the problem of the Moskow theatre in the introduction. Random walks are the prototype model to study *diffusion processes*, which describe the erratic motion of particles under the effects of random perturbations.

The position  $S_n$  of a random walker that starts at  $S_0 = 0$  changes by one unit, as  $n$  increases by one unit. Hence  $S_n$  always keeps the same parity of  $n$ .

---

<sup>1</sup>Indeed, we should write  $S_n(\omega)$  because  $S_n$  is a random variable. We omit to specify  $\omega$  for the sake of a lighter notation.



**Figure 14.** The trajectory of a random walk.

In other words

$$P\{S_n = k\} = 0 \quad \text{if } n + k \text{ is odd.}$$

The probability that  $S_n = k$ , if  $n + k$  is even, is given by the number of paths that reach  $k$  in  $n$  steps times the probability of each path, i.e.<sup>2</sup>

$$P\{S_n = k\} = \binom{n}{\frac{n+k}{2}} 2^{-n} \quad \text{if } n + k \text{ is even.}$$

Notice that  $\mathbb{E}[S_n] = n\mathbb{E}[X_i] = 0$  and  $\mathbb{V}[S_n] = n\mathbb{V}[X_i] = n$ . Hence the standard deviation of  $S_n$  increases as  $\sqrt{n}$  with  $n$ . This is *the first important characteristic of random walks* that you should remember. More specifically,  $S_n$  has a binomial distribution and, because of the de Moivre-Laplace theorem,  $S_n$  is well approximated, for large  $n$ , by

$$S_n \sim \sqrt{n}Z, \quad n \rightarrow \infty$$

where  $Z$  is a Gaussian variable with zero mean and unit variance (this is called a *Standard* variable). The limit  $n \rightarrow \infty$  can be realised by dividing a finite continuous time interval  $[0, t]$  in infinitesimal elements of size  $dt$ . This allows us to define random walks in continuous time  $t$  by the limit

$$W_t = \lim_{dt \rightarrow 0} \sqrt{dt} S_{n=t/dt}. \quad (9.1)$$

$W_t$  is a random<sup>3</sup> function of  $t$  which is called the *Wiener* process, and is the analogue of the random walk in discrete time  $n$ . Clearly  $\mathbb{E}[W_t] = 0$  and

<sup>2</sup>Note that  $S_n = n_+ - n_-$  is the difference between the number  $n_+$  of steps in the positive direction ( $X_i = +1$ ) and the number  $n_-$  of steps in the negative direction ( $X_i = -1$ ), whereas  $n = n_+ + n_-$ . Hence in a walk of  $n$  steps with  $S_n = k$ , the number of steps  $X_i = +1$  is  $\frac{n+k}{2}$ . There are  $\binom{n}{\frac{n+k}{2}}$  ways of choosing the  $n_+$  steps with  $X_i = +1$ .

<sup>3</sup>To be precise,  $W_t(\omega)$  is a function of  $t$  for every realisation  $\omega \in \Omega$ . The limit in Eq. (9.1) is in distribution.

$\mathbb{V}[W_t] = t$ . We shall discuss further properties of the Wiener process later on in the course.

Coming back to discrete time  $n$ , let us discuss two further applications of random walks:

**Counting votes:** in an election between two candidates, candidate  $P$  gets  $p$  votes and candidate  $Q$  receives  $q < p$  votes (i.e.  $P$  wins the election). What is the probability that, during the counting,  $P$  always leads? The answer, as we'll see later, is surprisingly simple

$$P\{P \text{ always leads } Q\} = \frac{p - q}{p + q}. \quad (9.2)$$

**Kolmogorov-Smirnov (KS) test:** let  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  be two samples resulting from two series of  $n$  independent experiments. For example, think of the case where the  $a_i$ 's measure the response of a group of  $n$  patients treated with a given drug and the  $b_i$ 's are collected measuring the same quantity in a test group of  $n$  untreated patients. One way to find out whether the treatment is effective or not, is to consider the  $a_i$ 's and the  $b_i$ 's as independent draws from two distributions and to ask whether the two distributions are really different. Let  $H$  be the *hypothesis* that the two samples are drawn i.i.d. from the same distributions  $P\{a_i < x\} = P\{b_i < x\} = P(x)$ . The KS test is based on computing

$$\Delta = \sup_{x \in \mathbb{R}} |A(x) - B(x)|$$

where  $A(x) = |\{i : a_i < x\}|$  and  $B(x) = |\{i : b_i < x\}|$  are the number of points in the two samples that are smaller than  $x$ . First observe that, under hypothesis  $H$ ,  $\mathbb{E}[A(x)] = \mathbb{E}[B(x)] = nP(x)$ . The plot of  $A(x) - B(x)$  on all the points  $x$  that coincide either with  $a_i$ 's or with  $b_i$ 's sorted in increasing order, looks like that of a random walk  $S_k$  on  $2n$  steps, with the condition  $S_{2n} = 0$ . This is a special type of random walk which is called a *random bridge*. If  $H$  is correct, then  $\Delta$  is the maximum excursion of a random bridge. The distribution of  $\Delta$  can be computed for  $n \gg 1$ , and it is given by<sup>4</sup>

$$P\{\Delta < \sqrt{n}\xi\} \simeq 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2\xi}.$$

Using this, we can find whether the value of  $\Delta$  we compute is likely or not, i.e. whether the hypothesis  $H$  can be confirmed or whether it should be rejected.

---

<sup>4</sup>The proof of this statement will not be given here. We refer to [15].





point, to the number  $N_{A' \rightarrow B}$  of paths from  $A'$  to  $B$ , which is easy to compute. As an application, let us see how we can derive Eq. (9.2) for the ballot theorem. Each path contributing to the event  $\{P \text{ always leads } Q\}$  must pass from the point  $A = (1, 1)$  and reach the point  $B = (p + q, p - q)$ , without crossing the horizontal axis. The number of these paths is the total number of paths  $A \rightarrow B$  minus those crossing the axis, i.e.

$$N_{A \rightarrow B} - N_{A' \rightarrow B}^{\neq} = N_{A \rightarrow B} - N_{A' \rightarrow B} = \binom{p+q-1}{p-1} - \binom{p+q-1}{p}$$

where  $A' = (1, -1)$  is the reflected point  $A$ . A simple manipulation of binomial coefficients shows that this number is  $(p - q)/(p + 1)$  times the total number of paths  $N_{O \rightarrow B} = \binom{p+q}{p}$  from the origin  $O = (0, 0)$  to  $B$ . This yields Eq. (9.2).

## 9.2 Returns and first returns

A return to the origin at time  $n$  is the event  $\{S_n = 0\}$ . Let  $u_n = P\{S_n = 0\}$  be its probability. Because of the parity of random walks, returns cannot occur at odd times, i.e.  $u_{2n-1} = 0$ . At even times

$$u_{2n} = \binom{2n}{n} 2^{-2n} \simeq \frac{1}{\sqrt{\pi n}} \quad (9.3)$$

where the last asymptotic expression holds for  $n \rightarrow \infty$  and is a consequence of Stirling's formula. Clearly  $u_0 = P\{S_0 = 0\} = 1$ .

Returns to the origin of random walks are an example of *recurrent* events.<sup>5</sup> These are events that can occur many times and conditional to the occurrence of an event at time  $n$ , the occurrence of future events is independent of the occurrence of past events.

Among returns, the first one is of special importance. We say that a first return occurs at time  $2n$  if  $S_k \neq 0$  for all  $k < 2n$  and  $S_{2n} = 0$ . We can also define a *first return time*  $T_f$  whose distribution is given by

$$f_{2n} = P\{T_f = 2n\} = P\{S_k \neq 0, 0 < k < 2n; S_{2n} = 0\}. \quad (9.4)$$

We set  $f_0 = 0$ , because the random walker can only return after it leaves the origin.

The first return distribution is related to the probability  $u_n$  of returns by the equation

$$u_n = \sum_{\nu=0}^n f_{\nu} u_{n-\nu}, \quad n > 0. \quad (9.5)$$

---

<sup>5</sup>See FELLER XII.

This holds for any recurrent event and it says that in order for a recurrent event to occur at time  $n > 0$ , it must first occur at some time  $\nu$  (that can be equal to  $n$ ) and then it has to occur again after  $n - \nu$  steps. In principle this equation allows us to compute  $f_n$  once  $u_n$  is known.

There is a more direct way to compute  $f_n$  that uses the following surprising result: *the probability that the walker never returns to the origin up to time  $2n$  equals the probability that the walker is at the origin at  $2n$ , i.e.,*

$$P\{T_f > 2n\} = P\{S_k \neq 0, 0 < k \leq 2n\} = P\{S_{2n} = 0\} = u_{2n}. \quad (9.6)$$

This is a remarkable result which is a consequence of the ballot theorem (i.e. of the reflection principle). As we have seen, the number of walks not crossing the horizontal axis that reach a point  $B = (2n, 2b)$  equals the difference

$$N_{(1,1) \rightarrow (2n,2b)} - N_{(1,-1) \rightarrow (2n,2b)}.$$

in order to compute the number of paths that never get back to the origin up to  $2n$ , staying on the positive semi-plane, we need to sum this difference over all  $b = 1, 2, \dots$ . Yet  $N_{(1,1) \rightarrow (2n,2b+2)} = N_{(1,-1) \rightarrow (2n,2b)}$  by translation invariance in the vertical direction. Hence in the sum over  $b$  the term  $N_{(1,-1) \rightarrow (2n,2b)}$  cancels  $N_{(1,1) \rightarrow (2n,2(b+1))}$  in the next term of the sum. The only remaining term is

$$N_{(1,1) \rightarrow (2n,2)} = \binom{2n-1}{n} = \frac{1}{2} \binom{2n}{n} = 2^{2n-1} u_{2n}.$$

In order to consider also paths that do not go back to the origin staying below the horizontal axis, this number has to be multiplied by two, which proves Eq. (9.6).

Now it is clear that, for  $n > 0$

$$f_{2n} = P\{T_f > 2n - 2\} - P\{T_f > 2n\} = u_{2n-2} - u_{2n} \quad (9.7)$$

$$= \binom{2n-2}{n-1} 2^{-2n+2} - \binom{2n}{n} 2^{-2n} \quad (9.8)$$

$$= \frac{1}{2n-1} \binom{2n}{n} 2^{-2n} \quad (9.9)$$

$$\sim \frac{1}{2\sqrt{\pi}} n^{-3/2}, \quad n \rightarrow \infty \quad (9.10)$$

Because of the second equality in (9.7)

$$\sum_{n>0} f_n = u_0 = 1$$

which means that the random walker will surely return to the origin. We say that the random walk is *persistent*.<sup>6</sup> The asymptotic expression (9.10) shows that the probability of a first return vanishes as  $n \rightarrow \infty$  very slowly. Indeed the expected value of the first return time diverges

$$\mathbb{E}[T_f] = \sum_{n>0} f_n n = +\infty.$$

So the random walker will surely return to the origin, but the expected time for this to happen diverge.<sup>7</sup>

### 9.3 Last visit and the arc-sine law

Let us now focus on the last visit to the origin of a random walk of  $2n$  steps. Let  $L_{2n}$  be the number of steps taken by the random walk when this event occurs. The probability of this event can be computed as

$$\alpha_{2n,2k} = P\{L_{2n} = 2k\} = P\{S_{2k} = 0\}P\{S_j \neq 0, 2k < j \leq 2n\} \quad (9.11)$$

the second factor can be computed using Eq. (9.6) and it equals  $P\{S_{2n-2l} = 0\} = u_{2n-2k}$ . Therefore

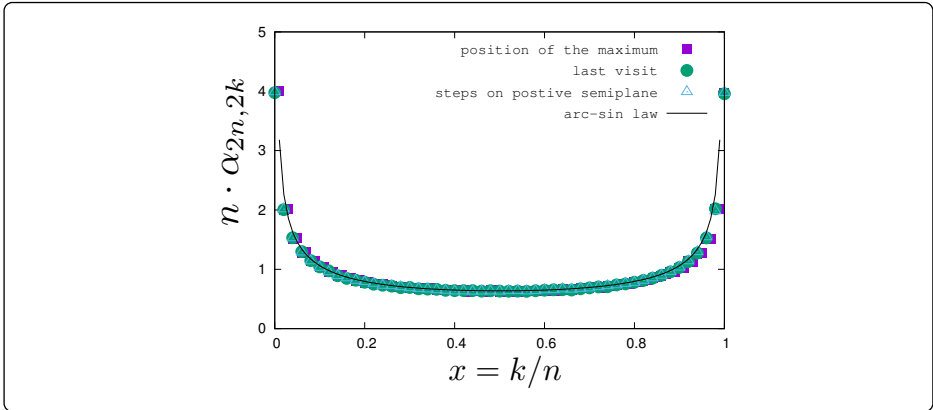
$$\alpha_{2n,2k} = u_{2k}u_{2n-2k} = \binom{2k}{k} \binom{2n-2k}{n-k} 2^{-2n}. \quad (9.12)$$

Surprisingly this probability is symmetric for  $k \rightarrow n-k$ , i.e.  $\alpha_{2n,2n-2k} = \alpha_{2n,2k}$ . This means that the random walker's last visit to the origin is as likely to occur close to the origin as close to its end point. What is more surprising is that these are the regions where the last visit is most likely to occur, whereas it is less likely to occur in the middle. This is evident in the large  $n$  limit where

$$\alpha_{2n,2k} \simeq \frac{1}{\pi} \frac{1}{\sqrt{k(n-k)}}, \quad n \rightarrow \infty. \quad (9.13)$$

<sup>6</sup>In general, a recurrent event that surely occurs in the future is called *persistent*. If there is a finite probability that the event will never occur, the recurrent event is called *transient*.

<sup>7</sup>In each realisation, the random walk returns to the origin in a finite time. Indeed, there can be other ways to estimate the time of the first return as e.g.  $\mathbb{E}[T_f^\alpha]^{1/\alpha}$  which would be finite for  $\alpha < 1/2$  or  $e^{\mathbb{E}[\log T_f]}$  which is also finite. The divergence of  $\mathbb{E}[T_f]$  is due to the fact that  $n$  grows while  $f_n$  vanishes too slowly for the series to converge. This means that the expected value does not always represent what we expect. We'll come back to this point.



**Figure 17.** The arc-sin law for last returns, the position of the maxima and the number of steps spent by the random walk on the positive semi-plane. Result of numerical simulations of  $10^6$  random walks of  $2n = 100$  steps.

### Exercise 9.1

Derive Eq. (9.13).

Eq. (9.12) is called the *arc-sine law* for the last visit to the origin. The reason for this name comes from the fact that, for  $x \in (0, 1)$  and in the limit  $n \rightarrow \infty$ , the probability that  $L_{2n} < 2nx$  is given by

$$P\{L_{2n} < 2nx\} = \sum_{k < nx} \alpha_{2n, 2k} \quad (9.14)$$

$$\simeq \sum_{k < nx} \frac{1}{\pi} \frac{1}{\sqrt{k(n-k)}} \quad (9.15)$$

$$\simeq \frac{1}{\pi} \int_0^x \frac{dz}{\sqrt{z(1-z)}} \quad (9.16)$$

$$= \frac{2}{\pi} \arcsin \sqrt{x} \quad (9.17)$$

Here we first transformed the sum on  $k$  into an integral on  $z = k/n$  and then used the transformation  $z = \sin^2 \theta$ . The arc-sin law holds not only for the last visit, but also for other quantities such as the position of the maximum of a random walk of  $2n$  steps or the number of steps spent by the random walk on the positive semi-plane,<sup>8</sup> as shown in Figure 17.

<sup>8</sup>See FELLER III for more details.

## 9.4 Random walks with drift

Let us now consider random walks

$$S_n = \sum_{i=1}^n X_i, \quad S_0 = 0$$

with  $X_i = \pm 1$  being i.i.d. random variables with  $P\{X_i = +1\} = p$  and  $P\{X_i = -1\} = 1 - p \equiv q$ . In this case different paths  $\omega = (X_1, \dots, X_n)$  can have different probabilities

$$P\{\omega\} = p^{\frac{n+S_n(\omega)}{2}} q^{\frac{n-S_n(\omega)}{2}}.$$

Therefore computing probabilities of events related to random walks is no longer a counting problem. In particular, the reflection principle cannot be used for  $p \neq 1/2$ . When  $p > 1/2$ , paths  $S_n$  that increase with  $n$  are more likely than those that decrease with  $n$ . We say that the random walk has a *drift* (in the positive direction, in this case).

Gambling is a typical example of a situation described by random walks with a drift. Consider a gambler that plays repeatedly a game where he/she can win one euro with probability  $p$  and loose one euro with probability  $q = 1 - p$ . Then  $S_n$  corresponds to the total gain (if positive) or loss (if negative) after  $n$  games. We can study the fate of the gambler using generating functions. Let us start by discussing the distribution of the waiting time  $T$  for a gain. This is the time when  $S_n = 1$  for the first time, i.e.

$$P\{T = n\} = P\{S_k \leq 0, 0 \leq k < n; S_n = 1\} \equiv \phi_n$$

Let us introduce the generating function for  $T$

$$\Phi(s) = \mathbb{E}[s^T] = \sum_{n=0}^{\infty} \phi_n s^n.$$

In order to compute  $\Phi(s)$  let us analyse the first step. With probability  $p$  the gambler wins the first game and then  $T = 1$ . With probability  $q$  the gambler looses. Then he/she has to wait a time  $T_1$  to get back to  $S_{T_1+1} = 0$  and then wait another  $T_2$  steps for the first gain. Hence  $T = 1 + T_1 + T_2$ . Now  $T_1$  and  $T_2$  are two independent random variables and they have exactly the same

distribution (and generating function) as  $T$ . Therefore

$$\Phi(s) = p\mathbb{E}[s^1] + q\mathbb{E}[s^{1+T_1+T_2}] \quad (9.18)$$

$$= ps + qs\mathbb{E}[s^{T_1}]\mathbb{E}[s^{T_2}] \quad (9.19)$$

$$= ps + qs\Phi(s)^2 \quad (9.20)$$

$$= \frac{1 - \sqrt{1 - 4pq}s^2}{2qs} \quad (9.21)$$

$$= -\frac{1}{2q} \sum_{n=1}^{\infty} \binom{1/2}{n} (-4pq)^n s^{2n-1} \quad (9.22)$$

Note that the solution to the quadratic equation has two roots, of which we choose the one in Eq. (9.21) because it is the only one consistent with  $P\{T = 0\} = \Phi(0) = 0$ . Note also that the expansion only generates odd powers of  $s$ . This is consistent with the fact that  $P\{T = 2n\} = 0$  for all  $n$ . The expression of  $\phi_n$  can be read from the last equation above and it can be further simplified using trite manipulations of the binomial coefficients

$$\phi_{2n-1} = -\frac{1}{2q} \binom{1/2}{n} (-4pq)^n = \frac{1}{2n-1} \binom{2n}{n} \frac{(pq)^n}{2q}, \quad \phi_{2n} = 0. \quad (9.23)$$

Note that

$$\sum_{n=0}^{\infty} \phi_n = \Phi(1) = \frac{1 - \sqrt{(1-2p)^2}}{2(1-p)} = \begin{cases} \frac{p}{1-p} & p < 1/2 \\ 1 & p \geq 1/2 \end{cases}$$

This means that for  $p \geq 1/2$  the gambler is sure to gain sooner or later. For  $p < 1/2$  there is a probability  $\frac{1-2p}{1-p}$  that this will never happen. So  $\phi_n$  is not a probability distribution for  $p < 1/2$  because it is not normalised. The term *defective probability distribution* is used to describe these cases.

You can also check that, provided the gambler will sooner or later gain, the expected time he/she has to wait is given by

$$\mathbb{E}[T|T < +\infty] = \frac{\Phi'(1)}{\Phi(1)} = \frac{1}{|1-2p|}.$$

This time is finite for  $p \neq 1/2$  and it diverges as  $p \rightarrow 1/2$ .

### 9.4.1 Returns to the origin

The probability of a return to the origin at time  $2n$  is now

$$u_{2n} = P\{S_{2n} = 0\} = \binom{2n}{n} (pq)^n = \binom{-1/2}{n} (-4pq)^n$$

Hence the associated generating function is

$$U(s) = \sum_{n=0}^{\infty} u_n s^n = \frac{1}{\sqrt{1-4pqs^2}} \quad (9.24)$$

We can find the distribution of the first return  $T_f$  to the origin by the same argument used for the first gain. If the first step is  $X_1 = -1$ , that occurs with probability  $q$ , then  $T_f = 1 + T$  where  $T$  is the time for the first gain, starting from the point  $A = (-1, 1)$ . If the first step is  $X_1 = +1$ , then  $T_f = 1 + T'$  where  $T'$  is the time for the first loss, starting from the point  $A' = (1, 1)$ . By symmetry, the generating function of  $T'$  is obtained from that of  $T$  by interchanging  $p$  and  $q$ . Therefore

$$F(s) = \mathbb{E}[s^{T_f}] = qs\mathbb{E}[s^T] + ps\mathbb{E}[s^{T'}] \quad (9.25)$$

$$= qs \frac{1 - \sqrt{1-4pqs^2}}{2qs} + ps \frac{1 - \sqrt{1-4pqs^2}}{2ps} \quad (9.26)$$

$$= 1 - \sqrt{1-4pqs^2}. \quad (9.27)$$

Expanding this in powers of  $s$ , one finds

$$f_{2n} = P\{T_f = 2n\} = \frac{1}{2n-1} \binom{2n}{n} (pq)^n, \quad f_{2n-1} = 0, \quad (n > 0)$$

that reverts to the result we found earlier when  $p = q = 1/2$ . Note that

$$F(1) = 1 - |1 - 2p|$$

which means that  $f_n$  is a defective probability distribution for  $p \neq 1/2$ . With probability  $|1 - 2p|$  the random walk will never return back to the origin. The random walk is called *transient* in this case ( $p \neq 1/2$ ) whereas it is *persistent* for  $p = 1/2$ . Provided the random walk returns to the origin, the expected time this takes is

$$\mathbb{E}[T_f | T_f < +\infty] = \frac{F'(1)}{F(1)} = 1 + \frac{1}{|1 - 2p|},$$

i.e. for  $p \neq 1/2$  either the random walk comes back to the origin in a finite time or it does not come back at all.

Eq. (9.27) can also be derived from the generating function  $U(s)$ , using the relation Eq. (9.5) between  $u_n$  and  $f_n$ . Multiplying both sides of this equation by  $s^n$  and summing over  $n > 0$  one finds  $U(s) - 1 = F(s)U(s)$ , which leads to

$$F(s) = 1 - \frac{1}{U(s)} \quad (9.28)$$

This equation relates the generating function of the time of the first occurrence of an event to that of its occurrence at a specific time, for a broad class of *recurrent events*.<sup>9</sup> In the present case of returns to the origin of random walks, this equation combined with Eq. (9.24) immediately delivers Eq. (9.27).

### Exercise 9.2

The simplest recurrent event is a success in repeated Bernoulli trials. In this case  $u_n = p$  for  $n > 0$  and  $u_0 = 1$ . Find  $f_n$  using Eq. (9.28). Check explicitly that Eq. (9.5) is satisfied.

### 9.4.2 Last visit to the origin

The generating function of the probability  $v_{2n} = P\{T_f > 2n\}$  that the random walk does not return to the origin up to time  $2n$  can be derived using Eq. (7.22), that relates the cumulative distribution of a random variable to the distribution itself,<sup>10</sup> i.e.

$$V(s) = \sum_{n=0}^{\infty} v_{2n} s^{2n} = \frac{1 - F(s)}{1 - s^2} = \frac{\sqrt{1 - 4pqs^2}}{1 - s^2}. \quad (9.29)$$

Here  $s^2$  appears instead of  $s$  in the denominator because we sum only on even powers of  $s$ , assuming  $v_{2n-1} = 0$ .<sup>11</sup> Notice that  $V(s) = U(s)$  for  $p = 1/2$ , in agreement with Eq. (9.6), but this is not true for  $p \neq 1/2$ . This has a consequence for the probability  $\alpha_{2n,2k}$  that the last visit of a random walk of  $2n$  steps occurs at time  $2k$ . Again we can write  $\alpha_{2n,2k} = u_{2k} v_{2n-2k}$ . Since this depends on two indices, we introduce a double generating function  $A(s, z)$  with  $s$  “counting”  $n$  and  $z$  “counting”  $k$ :

$$A(s, z) = \sum_{n=1}^{\infty} \sum_{k=0}^n \alpha_{2n,2k} s^{2n} z^{2k} \quad (9.30)$$

$$= \sum_{n=1}^{\infty} \sum_{k=0}^n u_{2k} (sz)^{2k} v_{2n-2k} s^{2n-2k} \quad (9.31)$$

$$= U(sz)V(s) = \frac{\sqrt{1 - 4pqs^2}}{(1 - s^2)\sqrt{1 - 4pqs^2 z^2}} \quad (9.32)$$

<sup>9</sup>We refer to FELLER XIII for more details.

<sup>10</sup>This relation stems from the equation  $f_{2n} = v_{2n-2} - v_{2n}$ .

<sup>11</sup>If instead we take  $v_{2n+1} = P\{T_f > 2n+1\} = P\{T_f > 2n\} = v_{2n}$  then we get an additional factor  $(1 + s)$  in the right hand side of Eq. (9.29), and recover Eq. (7.22). The choice  $v_{2n+1} = 0$  is motivated by the fact that we’re using the sequence  $v_n$  only for even values of  $n$ .



First, let us check that  $\alpha_{2n,2k}$  is correctly normalised. Observe that the coefficients of  $s^{2n}$  in the expansion in powers of  $s^2$  of  $A(s, 1)$  equal  $\sum_{k=0}^n \alpha_{2n,2k}$ . Setting  $z = 1$ , we find

$$A(s, 1) = \frac{1}{1 - s^2} = 1 + s^2 + s^4 + \dots$$

that confirms that  $\alpha_{2n,2k}$  is correctly normalised for all  $n$ . Second, notice that the symmetry  $\alpha_{2n,2k} = \alpha_{2n,2n-2k}$  that this probability satisfies for  $p = 1/2$  is no longer satisfied when  $p \neq 1/2$ .<sup>12</sup> Finally, let us compute the expected value of the time  $L_{2n}$  of the last visit, in a walk of  $2n$  steps. This is obtained observing that the partial derivative of  $A(s, z)$  with respect to  $z$ , evaluated at  $z = 1$ , has a power series in  $s^2$  with coefficients that are exactly the desired quantities

$$\left. \frac{\partial}{\partial z} A(s, z) \right|_{z=1} = \sum_{n=1}^{\infty} \sum_{k=0}^n \alpha_{2n,2k} 2k s^{2n} \quad (9.33)$$

$$= \sum_{n=1}^{\infty} \mathbb{E}[L_{2n}] s^{2n} \quad (9.34)$$

$$= \frac{4pq}{1 - 4pq} \left[ \frac{1}{1 - s^2} - \frac{1}{1 - 4pqs^2} \right] \quad (9.35)$$

$$\Rightarrow \mathbb{E}[L_{2n}] = \frac{4pq}{1 - 4pq} [1 - (4pq)^n]. \quad (9.36)$$

Notice that,

$$\lim_{p \rightarrow 1/2} \mathbb{E}[L_{2n}] = n$$

which is consistent with the symmetry  $k \rightarrow n - k$ . Yet for  $p \neq 1/2$

$$\mathbb{E}[L_{2n}] \leq \frac{4pq}{1 - 4pq}.$$

The last visit to the origin when  $n \rightarrow \infty$  is likely to occur at a finite time, close to the origin.

---

<sup>12</sup>In order to check this, observe that this symmetry implies that  $A(sz, 1/z) = A(s, z)$ . As an Exercise, show that this symmetry is satisfied only for  $p = 1/2$ .



# Chapter 10

## Branching processes

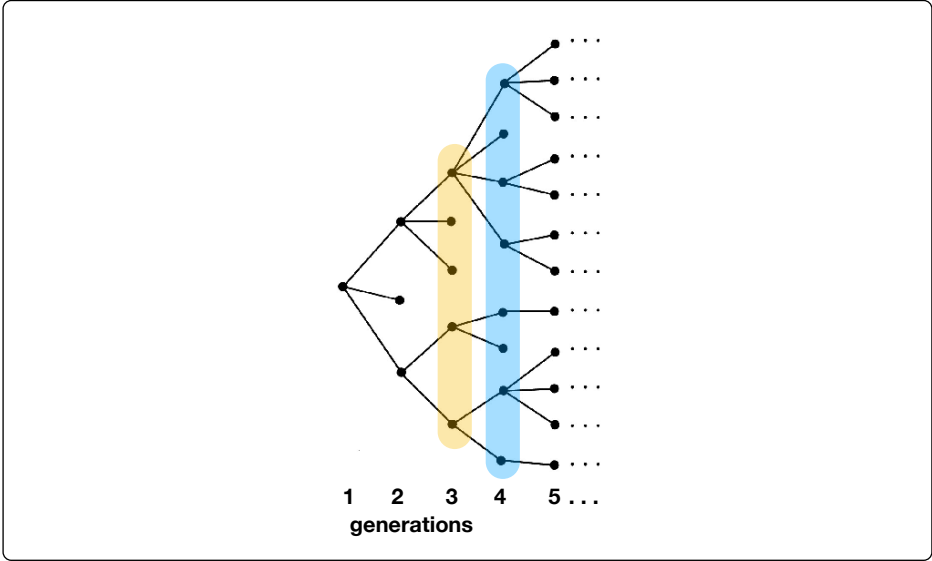
Branching processes<sup>1</sup> describe the evolution of a population of individuals (or units) that reproduce from one generation to the next. For example, in Italy, individuals inherit their family name from the father. Then, neglecting migrations, the number of male individuals in Italy with the same family name is the sum of the offsprings of male individuals at the previous generation. A further classical example is nuclear reactions: each atom when bombarded by neutrons may become unstable and release more neutrons that may induce the decay of other atoms, and so on. The way in which a viral epidemics such as Covid-19 or influenza spreads in a population is also an example of a branching process. Each infected individual can transmit the virus to more individuals. The mechanism of transmission and the contact network between individuals determines whether the epidemics will stop or whether it will become endemic in the population. Epidemic phenomena are not limited to diseases. It also applies to computer viruses, behaviours, fashions, habits and many other phenomena. In all cases, a relevant question is to understand whether the process will come to an end or continue indefinitely in an explosive manner. We shall address this issue in a very simple setting.

In order to describe a branching process, let  $Z_n \in \mathbb{N}$  be the number of individuals at generation  $n$  that descend from the same ancestor at generation zero. The next generation is composed of all the offsprings of individuals of the  $n^{\text{th}}$  generation

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_i^{(n)}, \quad (10.1)$$

---

<sup>1</sup>Branching processes are discussed in FELLER XII.3/4/5.



**Figure 18.** A branching process. The 2<sup>nd</sup> and the 3<sup>rd</sup> generations are highlighted.

where  $X_i^{(n)}$  is the number of offsprings of the  $i^{\text{th}}$  individual of the  $n^{\text{th}}$  generation. We consider a very simplified situation where  $X_i^{(n)}$  are i.i.d.  $\forall i = 1, \dots, Z_n$  and  $n = 0, 1, \dots$ , with

$$P\{X_i^{(n)} = k\} = p_k. \quad (10.2)$$

At generation 0,  $Z_0 = 1$  because we assume that the whole population starts with one individual, the ancestor. Note that  $Z_n$  is an integer random variable for  $n > 0$ . In addition Eq. (10.1) shows that  $Z_{n+1}$  is a sum of a random number of random variables. Progress is then possible by introducing generating functions.

## 10.1 The main equation

Let

$$P(s) = \mathbb{E} \left[ s^{X_i^{(n)}} \right] = \sum_{k=0}^{\infty} p_k s^k$$

be the generating function of  $X_i^{(n)}$ . Then Eq. (10.1) readily yields a recursion equation for the generating function of  $Z_n$ :

$$P_n(s) = \mathbb{E} [s^{Z_n}] .$$

This reads

$$P_{n+1}(s) = \mathbb{E} \left[ s^{X_1^{(n)} + \dots + X_{Z_n}^{(n)}} \right] = \mathbb{E} [P(s)^{Z_n}] = P_n(P(s)). \quad (10.3)$$

Starting from  $P_0(s) = s$ , because  $Z_0 = 1$ , we get  $P_1(s) = P(s)$ ,  $P_2(s) = P(P(s))$ , and so on. This allows us to compute the generating function for all values of  $n$ , in principle. In practice, extracting the asymptotic behaviour of a branching process from this equation is not easy.

We can also compute  $Z_{n+1}$  as the sum of the number  $Z_n^{(j)}$  of individuals generated after  $n$  generations by each of the offsprings  $j$  of the ancestor, i.e.

$$Z_{n+1} = \sum_{j=1}^{X_0} Z_n^{(j)}$$

where  $X_0$  is the number of offsprings of the ancestor. Introducing again generating functions, we find<sup>2</sup>

$$P_{n+1}(s) = P(P_n(s)). \quad (10.5)$$

### Exercise 10.1

Let's assume that the number  $X_i^{(n)}$  of unstable atoms generated as the result of the decay of one atom, in a nuclear reactor, is a Poisson random variable. Its mean  $\lambda = \mathbb{E} [X_i^{(n)}]$  can be adjusted to control the reaction in order to keep the expected value of unstable atoms at the next generation  $\mathbb{E} [Z_{n+1} | Z_n]$  at a constant value  $z$ . What is the protocol  $\lambda(Z_n)$  that should be adopted to achieve this goal?

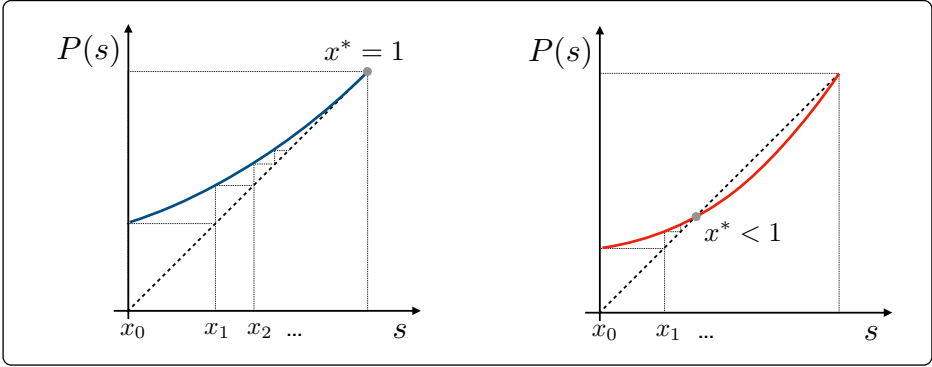
## 10.2 The extinction probability

A branching process is extinct at generation  $n$  if  $Z_n = 0$ . The probability of this event is  $x_n = P\{Z_n = 0\} = P_n(0)$  and, because of Eq. (10.5), it satisfies the

<sup>2</sup>Note that  $Z_n$  is also given by the sum of all individuals at generation  $\nu$  of the population  $Z_{n-\nu}^{(j)}$  generated from them after  $n - \nu$  generations. This leads to the general equation

$$P_n(s) = P_\nu(P_{n-\nu}(s)), \quad (10.4)$$

which holds for any  $\nu = 0, 1, \dots, n$ . This equation relies on the fact that, conditional on  $Z_\nu$ , the “future” of the branching process (i.e. what happens for  $k > \nu$ ) is independent of the “past” (i.e. what happened for  $k < \nu$ ). Processes that enjoy this property are called *Markov processes* and they all satisfy an equation like (10.4), that is called the *Chapman-Kolmogorov equation*.



**Figure 19.** The recursion equation (10.6) and its limit.

recursion relation

$$x_{n+1} = P_{n+1}(0) = P(P_n(0)) = P(x_n), \quad n \geq 0 \quad (10.6)$$

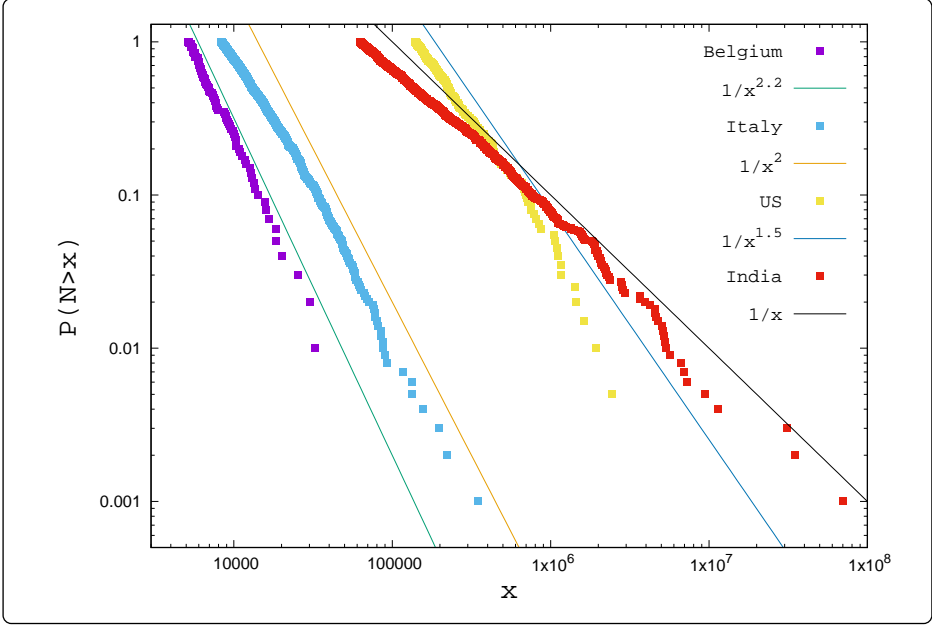
with the initial condition  $x_0 = 0$ , because the process is not extinct at  $n = 0$  ( $Z_0 = 1$ ). Hence  $x_1 = P(0)$ ,  $x_2 = P(P(0))$  etc. It is evident that  $x_n$  must be an increasing sequence, because if  $Z_n = 0$  then for sure  $Z_{n+1} = 0$ . Indeed, this can be proven by induction:  $x_1 > x_0$ , because  $P(s)$  is an increasing function, and if  $x_n > x_{n-1}$ , then  $x_{n+1} = P(x_n) > P(x_{n-1}) = x_n$ , again because  $P(s) \nearrow s$ . Since  $x_n \leq 1$  is a bounded sequence, the limit of  $x_n$  for  $n \rightarrow \infty$  exists and it satisfies

$$x^* = \lim_{n \rightarrow \infty} x_n = P(x^*). \quad (10.7)$$

It is possible to gain insight on the behaviour of  $x_n$  by a graphical analysis, as shown in Figure 19. This plots  $P(s)$  as a function of  $s$ . This is an increasing function and all its derivatives are non-negative.  $P(s)$  intersects the  $45^\circ$  line at  $s = 1$ . If this is the only intersection, as in Figure 19 (left), then  $x^* = 1$ . In this case the branching process will surely come to an end, i.e.  $x^* = 1$ . If there is another solution of the equation  $P(s) = s$ , then  $x^* < 1$  is the smallest of the two solutions. In this case, with probability  $1 - x^*$  the branching process will continue indefinitely.

Whether the branching process will get extinct ( $x^* = 1$ ) or not ( $x^* < 1$ ) is determined by the slope of  $P(s)$  at  $s = 1$ . If the slope  $P'(1)$  is smaller than one then  $x^* = 1$  and if  $P'(1) > 1$  then  $x^* < 1$ . The slope  $P'(1)$  coincides with the expected number of offsprings per individual

$$P'(1) = \mathbb{E} [X_i^{(n)}] = \mu \quad (10.8)$$



**Figure 20.** Family names can be considered as an example of a branching process. In different countries (Belgium, Italy, US and India) the fraction of surnames that belong to more than  $x$  individuals in the population behaves as a power law  $P(N > x) \sim x^{-\gamma}$  with an exponent  $\gamma$  that, apparently, is smaller the larger is the country. Can you relate this behaviour with the theory of branching processes being discussed here? (The data is taken from different sources on internet in Nov. 2024. In each case the list of the  $M$  most frequent names –  $M = 100$  for Belgium, 200 for US and 1000 for Italy and India – was reported with the number of individuals with that surname.)

that we shall denote by  $\mu$  henceforth. Therefore

$$\mu \leq 1 \Rightarrow x^* = 1 \quad (10.9)$$

$$\mu > 1 \Rightarrow x^* < 1 \quad (10.10)$$

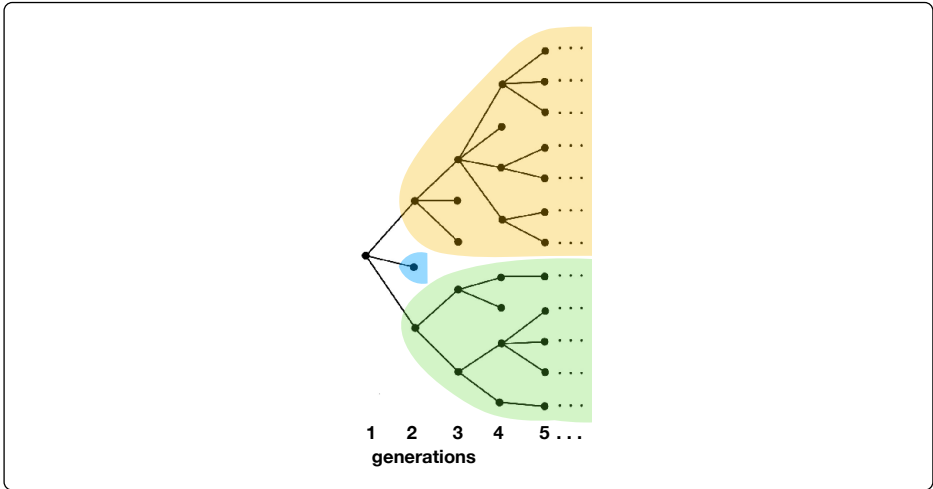
The rationale for this result appears more clearly if we use Eq. (10.5) to find how the expected value of the population grows with the generations:

$$\mathbb{E}[Z_{n+1}] = P'_{n+1}(1) = P'_n(1)P'(1) = \mu \mathbb{E}[Z_n].$$

Iterating this recursion, starting with  $\mathbb{E}[Z_0] = 1$ , we find that

$$\mathbb{E}[Z_n] = \mu^n.$$

Summarising, when  $\mu \leq 1$  the population does not grow exponentially and in-



**Figure 21.** The total progeny of the ancestor can be split into the progenies of its offsprings (plus itself).

deed it will surely become extinct ( $x^* = 1$ ).<sup>3</sup> When  $\mu > 1$  there is a probability  $1 - x^*$  that the process will continue forever, and in this case the population's size will explode, with an exponential behaviour.

### Exercise 10.2

Show that for  $\mu < 1$  the variance  $\mathbb{V}[Z_n]$  remains of the same order of the expected value  $\mathbb{E}[Z_n]$  whereas when  $\mu > 1$ ,  $\mathbb{V}[Z_n] \propto \mathbb{E}[Z_n]^2$ . (Hint: find a recursion relation for  $a_n = \mathbb{V}[Z_n] / \mathbb{E}[Z_n]$ ).

## 10.3 The total progeny and universality

The total number of individuals up to generation  $n$

$$Y_n = Z_0 + Z_1 + \dots + Z_n$$

is called the total *progeny* of the ancestor up to generation  $n$ . If  $Y_{n-1}^{(j)}$  is the

<sup>3</sup>This is true because for a non-negative integer random variable we have the inequality

$$P\{Z_n > 0\} \leq \mathbb{E}[Z_n]$$

which you can easily prove.

So if  $\mathbb{E}[Z_n] \rightarrow 0$  as  $n \rightarrow \infty$ , so does  $P\{Z_n > 0\} = 1 - x_n$ .



total progeny of the  $j^{\text{th}}$  offspring of the ancestor after  $n - 1$  generations, then

$$Y_n = 1 + \sum_{j=1}^{X_0} Y_{n-1}^{(j)}.$$

The associated generating function  $R_n(s) = \mathbb{E}[s^{Y_n}]$  then satisfies

$$R_n(s) = \mathbb{E}\left[s^{1+Y_{n-1}^{(1)}+\dots+Y_{n-1}^{(X_0)}}\right] = s\mathbb{E}\left[R_{n-1}(s)^{X_0}\right] = sP(R_{n-1}(s)) \quad (10.11)$$

for all  $n > 0$ , with  $R_0(s) = s$ , because  $Y_0 = 1$ . By the same argument used to show that  $x_n$  is an increasing sequence in  $n$ , one can show that for each  $s \in (0, 1)$ ,  $R_n(s)$  is a monotonic sequence in  $n$ . Hence the limit of  $R_n(s)$  as  $n \rightarrow \infty$  exists and it satisfies

$$\rho(s) = \lim_{n \rightarrow \infty} R_n(s) = sP(\rho(s)). \quad (10.12)$$

The coefficient of  $s^n$  in the power expansion of  $\rho(s)$  is the probability that the total (asymptotic) progeny of the ancestor  $Y_\infty$  equals  $n$ , i.e.

$$\rho(s) = \sum_{n=1}^{\infty} P\{Y_\infty = n\} s^n.$$

Notice that for  $s = 1$ , the equation for  $\rho(1)$  reduces to Eq. (10.7). This means that  $\rho(1) = x^*$  equals the extinction probability. When  $\mu \leq 1$ , we have  $x^* = 1$ , which implies that  $P\{Y_\infty = n\}$  is correctly normalised. When  $\mu > 1$  instead  $\rho(1) = x^* < 1$  which means that the distribution  $P\{Y_\infty = n\}$  is defective. Indeed  $\rho(1)$  does not account for the probability  $P\{Y_\infty = \infty\} = 1 - x^*$  of an infinite population.

The expected size of the total progeny is infinite for  $\mu > 1$ . Yet we can compute the expected value of  $Y_\infty$  conditional on  $Y_\infty < +\infty$ , as  $\mathbb{E}[Y_\infty | Y_\infty < +\infty] = \frac{\rho'(1)}{\rho(1)}$ . Now

$$\rho'(1) = P(\rho(1)) + P'(\rho(1))\rho'(1) = \frac{P(x^*)}{1 - P'(x^*)} = \frac{x^*}{1 - P'(x^*)} \quad (10.13)$$

For  $\mu < 1$ ,  $x^* = 1$  and  $P'(x^*) = \mu$ . Then the expected value of  $Y_\infty$  diverges as  $\mu \rightarrow 1^-$ , i.e.

$$\mathbb{E}[Y_\infty | Y_\infty < +\infty] = \frac{1}{1 - \mu} \quad \mu < 1. \quad (10.14)$$

For  $\mu > 1$ , instead, it is not possible to derive a closed form equation. However it is possible to understand the limit behavior for  $\mu \rightarrow 1^+$ . In this limit we expect that  $x^* \approx 1$  so we can expand Eq. (10.7) around  $x^* = 1$

$$x^* = P(x^*) \simeq P(1) + P'(1)(x^* - 1) + \frac{1}{2}P''(1)(x^* - 1)^2 + \dots$$

which means that

$$x^* \simeq 1 - 2\frac{\mu - 1}{P''(1)} + O((\mu - 1)^2).$$

This allows us to estimate the denominator in Eq. (10.13) using

$$P'(x^*) \simeq P'(1) + P''(1) \left[ -2\frac{\mu - 1}{P''(1)} \right] + O((\mu - 1)^2) \simeq \mu - 1 + O((\mu - 1)^2)$$

Therefore we find that

$$\mathbb{E}[Y_\infty | Y_\infty < +\infty] \simeq \frac{1}{\mu - 1} \quad \mu \rightarrow 1^+ \quad (10.15)$$

diverges in the same way, on both sides of  $\mu = 1$ . This is an example of *universal* behaviour, because the singularity in Eq. (10.15) is the same, irrespective of the details of the branching process.

Let us make a specific example and consider a branching process with

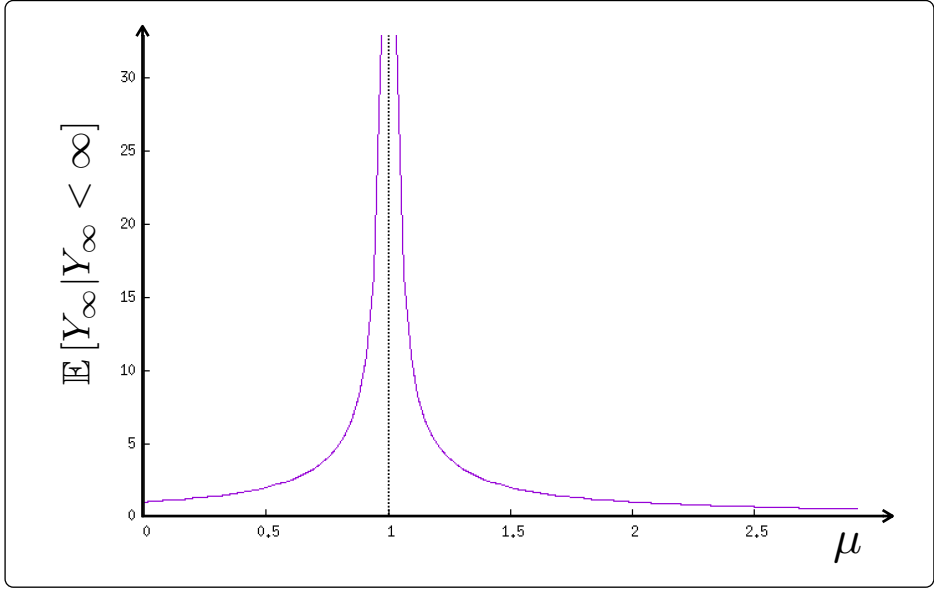
$$p_k = \begin{cases} p & \text{for } k = 2 \\ 1 - p = q & \text{for } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then  $P(s) = q + ps^2$  and the expected number of offsprings per individual is  $\mu = P'(1) = 2p$ . The extinction probability is given by the solution of the quadratic equation  $s = q + ps^2$ , which is

$$x^* = \frac{1 - \sqrt{1 - 4pq}}{2p} = \begin{cases} 1 & \text{for } p \leq 1 \\ \frac{1-p}{p} & \text{for } p > 1/2. \end{cases}$$

As expected,  $x^* = 1$  for  $\mu \leq 1$ , and  $x^* < 1$  when  $\mu > 1$ . The generating function of the total progeny is

$$\rho(s) = sP(\rho(s)) = s[q + p\rho(s)^2] = \frac{1 - \sqrt{1 - 4pqs^2}}{2ps}.$$

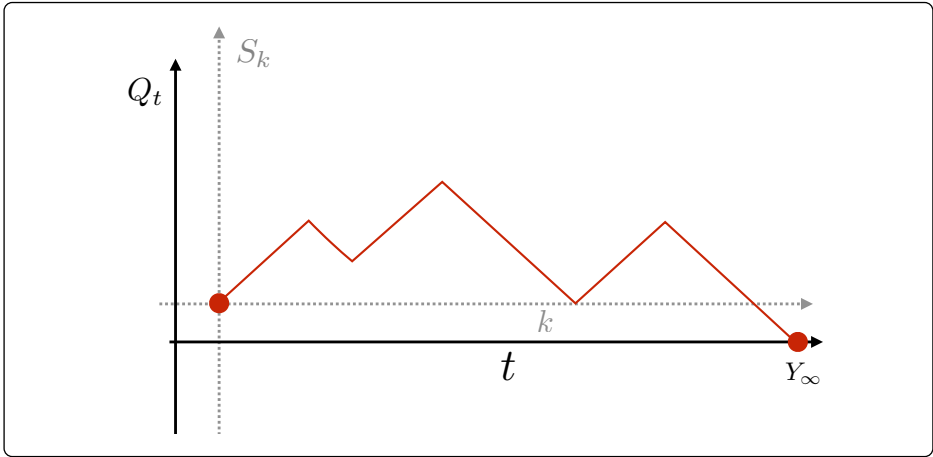


**Figure 22.** The expected value of the total progeny, for a finite branching process.

This is the same as the generating function  $\Phi(s)$  for the first loss in a random walk (see Eq. (9.21) with  $p \leftrightarrow q$ ). This coincidence is not accidental, as shown by the following argument.

In order to count the progeny, we can form a queue starting from the ancestor. Each time we count an individual in the queue we remove it but we also add its offsprings at the end of the queue. Let  $Q_{t+1}$  be the length of the queue when we have counted  $t$  individuals. Then  $Q_{t+1} = Q_t - 1$  if the  $t^{\text{th}}$  individual leaves no offsprings (which occurs with probability  $q$ ). Otherwise (with probability  $p$ ) the  $t^{\text{th}}$  individual leaves two offsprings and  $Q_{t+1} = Q_t + 1$ . Therefore  $Q_t - 1 = S_{t-1}$  behaves exactly as a random walk. The total progeny  $Y_\infty = \min\{t : Q_t = 0\}$  is the time  $t$  when the queue is empty for the first time. In terms of the random walk  $S_k$  (with  $k = t - 1$ )  $Y_\infty = T$  is exactly the waiting time for the first loss.<sup>4</sup> The analogy between queuing problems and random walks goes beyond this specific example. It applies in general with  $p_k$  being the probability that  $k$  new customers join the queue while the first in the queue is served. This corresponds to a random walk that takes a negative step  $X_{n+1} = X_n - 1$  with probability  $p_{k=0}$  and that otherwise takes  $k - 1$  steps in the positive direction ( $X_{n+1} = X_n + k - 1$ ) with probability  $p_k$ . The behaviour of  $Y_\infty$  described above implies that the expected time for the

<sup>4</sup>Compute  $\mathbb{E}[Y_\infty | Y_\infty < +\infty]$  for a branching process with  $p_k = p(1-p)^k$ ,  $k = 0, 1, 2, \dots$



**Figure 23.** The total progeny of a branching process and the time for first loss in a random walk.

first loss of this random walk, which is closely related to the first return to the origin, has the same generic behaviour (i.e. it is *universal*) as a function of  $\mathbb{E}[X_{n+1} - X_n - 1] = \mu - 1$  (see Figure 22).

### Exercise 10.3

Can one use the theory of branching process to predict the evolution of a pandemic such as Covid-19, given the past data on the number of reported cases? What are the main problems in applying these ideas to a real epidemics?

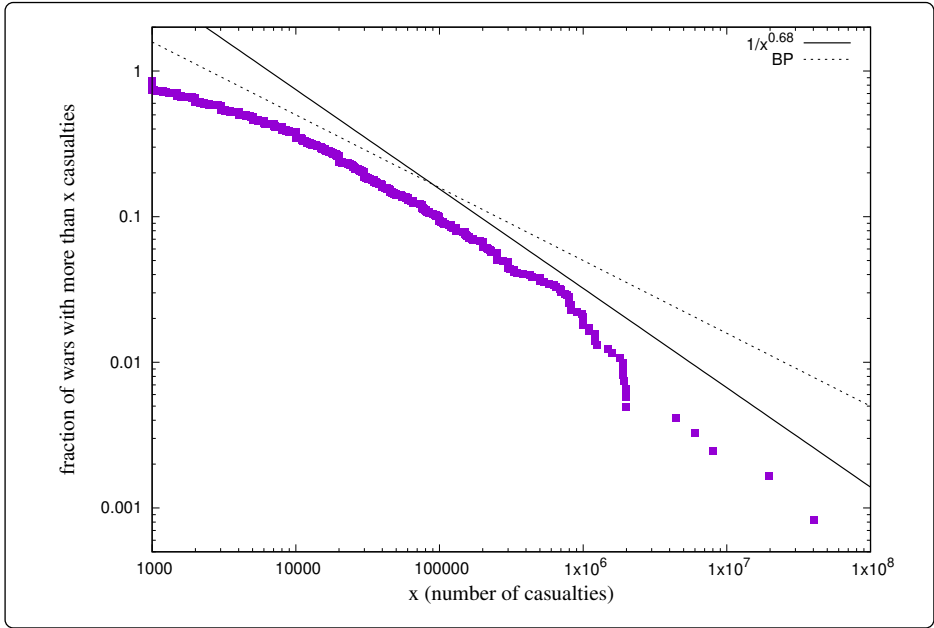
The distribution of the total progeny  $Y_\infty$  for  $\mu = 1$  also has an *universal* asymptotic behaviour  $P\{Y_\infty = n\} \sim n^{-3/2}$ , irrespective of  $p_k$ . This asymptotic behaviour is consistent with a singularity at  $s = 1$  of  $\rho(s)$ . Therefore, we shall study the behavior of  $\rho(s)$  for  $s \sim 1$ . In order to do this, we set  $s = 1 - \epsilon$  with  $\epsilon \ll 1$ . We also expect  $\rho(s) \simeq \rho(1) = 1$ . Hence we set  $\rho(s) = 1 - \eta(\epsilon)$ , with  $\eta \rightarrow 0$  as  $\epsilon \rightarrow 0$ . Now expand the equation  $\rho(s) = sP(\rho(s))$  to leading order:

$$1 - \eta = (1 - \epsilon)P(1 - \eta) \quad (10.16)$$

$$= (1 - \epsilon) \left[ P(1) - P'(1)\eta + \frac{1}{2}P''(1)\eta^2 + \dots \right] \quad (10.17)$$

$$= 1 - \epsilon - \eta + \epsilon\eta + \frac{1}{2}P''(1)\eta^2 + \dots \quad (10.18)$$

Keeping only the leading order terms, this equation becomes  $0 \simeq -\epsilon + \frac{1}{2}P''(1)\eta^2 + \dots$ . This shows that  $\eta(\epsilon) \sim \sqrt{\epsilon}$ , i.e. that  $\rho(s) \simeq 1 - c\sqrt{1 - s}$  for



**Figure 24.** Lewis Fry Richardson, besides landmark contributions to meteorology, also studied the statistics of “deadly quarrels”. These may be thought of as a branching process where, each casualty on one side causes a random number of casualties on the other. In this simplified view, the statistics of deadly quarrels suggests that the branching process is critical, i.e. that the logic driving deadly quarrels is *an eye for an eye* [Data from the Conflict Catalogue by Peter Brecke].

$s \rightarrow 1^-$ , with  $c = \sqrt{2/P''(1)}$ . This type of singular behaviour of  $\rho(s)$  implies  $P\{Y_\infty = n\} \sim n^{-3/2}$ , as anticipated.

#### Exercise 10.4

This derivation assumes that  $P(s)$  has finite first and second derivative at  $s = 1$ . This means that the number  $X_i$  of offsprings of each individual has finite expected value and variance. Suppose that  $p_k = P\{X_i = k\} \sim k^{-\gamma-1}$  with  $\gamma \in (1, 2)$ , so that the variance is infinite and  $P(s) \simeq s - c(1-s)^\gamma$  for  $s \simeq 1$  (again with  $\mu = 1$ ). Using the same argument, show that in this case  $P\{Y_\infty = n\} \sim n^{-1/\gamma-1}$ .

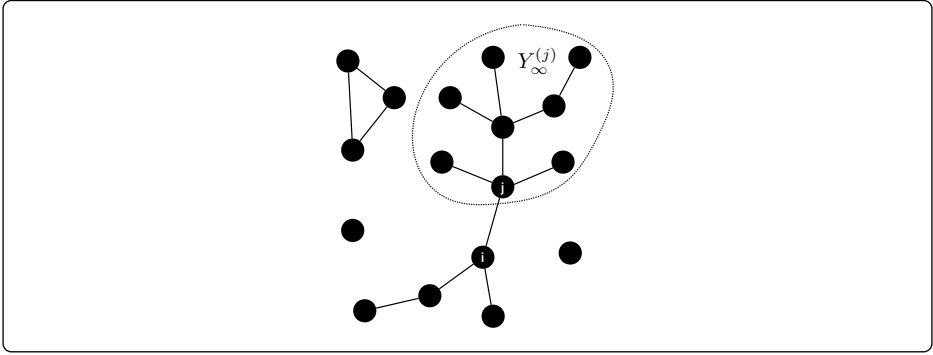


Figure 25. A random network.

## 10.4 An application to random networks\*

A network<sup>5</sup>  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is composed of a set  $\mathcal{V}$  of  $n$  nodes or vertices and a set  $\mathcal{E}$  of edges, each of which connects two nodes  $i, j \in \mathcal{V}$  with  $i \neq j$ . Random networks are networks where the edges are drawn at random between pairs of nodes and at most one edge connects any two nodes.<sup>6</sup> The prototype example are Erdős-Rényi random graphs, which are generated drawing at random edges between each pair of nodes with a probability  $p$ . In the limit  $n \rightarrow \infty$ , with  $p = \lambda/n$ , each node ends up having a number of edges  $E_i$  — which is called the *degree* — which has a Poisson distribution

$$P\{E_i = k\} = \frac{\lambda^k}{k!} e^{-\lambda}. \quad (10.19)$$

This construction can be generalised to random networks with a generic *degree distribution*  $P\{E_i = k\} = \pi_k$ ,  $k = 0, 1, \dots$ <sup>7</sup>

For a given  $\pi_k$ , an interesting question is whether the network is composed of a single component or by many, or whether a component of infinite size exists or not, in the limit  $n \rightarrow \infty$ .

<sup>5</sup>This section follows [16].

<sup>6</sup>Networks have no double edge and no tadpole, which is an edge joining a node to itself. Graphs with this properties are called simple graphs.

<sup>7</sup>In order to construct such a network, first draw at random the degree  $E_i$  of each node  $i \in \mathcal{V}$  from the distribution  $\pi_k$ . Each node  $i$  comes with  $E_i$  “half edges” that have to be connected. In order to do this, build a list of all half edges and recursively pick two of them at random, connect the corresponding nodes and remove the two half edges from the list. Continue this procedure until the list is empty. If two nodes are connected by more than one edge or if an edge connects a node to itself, restart the procedure from scratch until you get a simple graph. Notice that  $\sum_i E_i$  should be an even number.

In order to address this question, consider picking a node  $i$  at random, and consider the number of nodes that can be reached from this node following one of its links. Let  $j$  denote this neighbour. The number of nodes that can be reached from  $i$  through  $j$  is obtained following all the other links of  $j$  thus reaching all neighbours of  $j$  and then to the neighbours of neighbours of  $j$  and so on. The similarity with branching processes should now be clear:  $j$  is the ancestor, the neighbours of  $j$  form the first generation, the neighbours of neighbours the second generation and so on. Notice that the size of the first generation  $Z_1 = E_j - 1$  equals the number of neighbours of  $j$  minus one, which is the link that joins  $j$  to  $i$ . We can relate the distribution of  $Z_1$ , i.e. the distribution of the number of offsprings in the branching process, to the degree distribution  $\pi_k$ . The key insight is that the probability to choose a node  $j$  is proportional to its number of links. Hence

$$P\{E_j = k\} = \frac{k\pi_k}{\mathbb{E}[E_j]},$$

where the denominator  $\mathbb{E}[E_j] = \sum_k k\pi_k$  ensures normalisation. The distribution of the number of offsprings in the associated branching process, therefore is given by

$$p_k = \frac{(k+1)\pi_{k+1}}{\mathbb{E}[E_j]} \quad (10.20)$$

which accounts for the fact that node  $j$  should have at least one link for it to be reached. Then it is clear that the size of the network that can be reached from  $j$  is the total progeny  $Y_\infty^{(j)}$  of the branching process with offspring distribution  $p_k$ . The total number of nodes that are connected to  $i$ , i.e. the size of the component of the network to which  $i$  belongs, is obtained summing over all neighbours  $j$

$$N_i = 1 + Y_\infty^{(1)} + \dots + Y_\infty^{(E_i)} \quad (10.21)$$

and its generating function is

$$G(s) = \mathbb{E} \left[ s^{1+Y_\infty^{(1)}+\dots+Y_\infty^{(E_i)}} \right] = s\mathbb{E} [\rho(s)^{E_i}] = \Pi(\rho(s)) \quad (10.22)$$

where  $\rho(s) = \mathbb{E} [s^{Y_\infty^{(j)}}] = sP(\rho(s))$  satisfies the equation of the total progeny of a branching process with probability  $p_k$  and generating function  $P(s) = \sum_k p_k s^k$ . Therefore we see that if the expected number

$$\mu = \sum_k k p_k = \frac{\mathbb{E}[E_j(E_j - 1)]}{\mathbb{E}[E_j]} \quad (10.23)$$

of neighbours of a neighbour of  $i$  (excluding  $i$ ) is less than one, node  $i$  will surely belong to a finite component. This means that a random network with  $\mu \leq 1$  will surely have no component of infinite size in the limit  $n \rightarrow \infty$ . When  $\mu > 1$  instead, there is a finite probability that the branching process will not get extinct and hence that node  $i$  belongs to a component of infinite size.

For an Erdős-Rényi random graphs  $p_k = \pi_k$  coincides with the degree distribution and hence  $\mu = \lambda$ . In general, the condition  $\mu > 1$  for the existence of a giant component in a random graph demands that the expected value of the square of the degree should be larger than twice the expected degree, i.e.  $\mathbb{E}[E_j^2] > 2\mathbb{E}[E_j]$ .

Note also that the solution  $G(s)$  also provides access to the distribution of the sizes of the components of the network. For example, at  $\mu = 1$  this theory predicts that the fraction of components of size  $s$  should be proportional to  $s^{-3/2}$ .

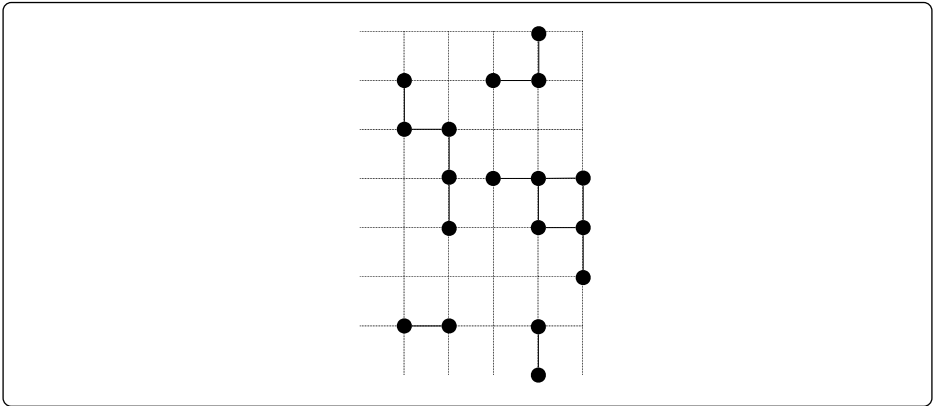
As a final comment, we observe that this theory assumes that following the links away from a given node  $i$  one never gets back to the original site  $i$ . In other words, this theory assumes that the network has no loops. This is wrong, because a random graph of  $n$  nodes can have loops. Consider for example the case where  $\mu > 1$ . Then the construction discussed above suggests that the number of nodes at distance  $d$  from  $j$  grows as  $\mu^d$ . Yet this number cannot exceed the total number  $n$  of nodes. Therefore it is clear that when  $d \simeq \frac{\log n}{\log \mu}$  some of the nodes reached in this construction must necessarily have been reached already. This argument suggests that loops of size  $\log n$  exist in random graphs. This implies that locally a random network looks like a tree, for  $n \rightarrow \infty$ . Furthermore, when  $\mu < 1$  components are all of a finite size, hence loops are rare. So the theory, though not exact, offers a good approximation of the statistics of component sizes of random networks for  $\mu < 1$  and for the emergence of the giant component when  $\mu \rightarrow 1^-$ .

### Exercise 10.5

Show that a random graph where all the nodes have degree  $E_i = 2$  is a collection of loops. Compute the probability that a random node  $i$  belongs to a loop of size  $\ell$ .

The transition at  $\mu = 1$  between a sparse network composed of many disconnected components and a dense network characterised by a giant component, is an example of a percolation transition. Percolation is a mathematical model defined on a  $d$ -dimensional lattice (e.g. a hyper-cubic lattice) of linear





**Figure 26.** A percolation network on a  $d = 2$  dimensional lattice.

size  $L$ . Each pair of neighbour nodes in this lattice are connected by a link with probability  $p$  and the two nodes are disconnected otherwise. When  $p$  is small, all clusters of connected nodes are of finite size so that there is no path of links that connects one side of the lattice to the other side, when  $L \rightarrow \infty$ . Such a path emerges when  $p$  reaches a critical *percolation threshold*  $p_c$  and the cluster of links “percolates” from one side of the system to the other for all  $p \geq p_c$ .

The difference between random graphs and percolation is that in the latter links can only occur between neighbours on the lattice. Since the number of neighbours increases with the dimension  $d$ , this restriction plays a weaker and weaker role as  $d$  increases. Indeed random graphs may be thought of as describing the  $d \rightarrow \infty$  limit of percolation. Indeed the statistical behaviour of the two problems share many similarities. For example the distribution of cluster sizes at  $p_c$  also follows a power law distribution  $p(s) \sim s^{-\tau}$  and the exponent takes its “mean-field” value  $\tau = 3/2$  for  $d \geq 6$ , which coincides with the exponent that governs the component size distribution of random networks.



# Chapter 11

## Markov chains

Up to now we have discussed sequences of independent random variables  $X_1, \dots, X_n$ , for which<sup>1</sup>

$$P\{x_1, \dots, x_n\} = \prod_{i=1}^n P\{x_i\}.$$

In general, the joint distribution satisfies

$$P\{x_1, \dots, x_n\} = P\{x_n | x_{n-1}, x_{n-2}, \dots, x_1\} P\{x_{n-1} | x_{n-2}, \dots, x_1\} \dots P\{x_2 | x_1\} P\{x_1\}.$$

Markov processes are sequences of random variables where the index  $n$  can be considered as a time variable, and where the conditional probability

$$P\{x_{t+1} | x_t, x_{t-1}, \dots, x_1\} = P\{x_{t+1} | x_t\}, \quad t = 1, 2, \dots \quad (11.1)$$

does not depend on the values  $x_\tau$  of the process, for  $\tau < t$ . In other words, for a Markov process, conditional on the present ( $x_t$ ), the future ( $x_{t+1}, x_{t+2}, \dots$ ) is independent of the past ( $x_{t-1}, x_{t-2}, \dots$ ). This means that the present state  $x_t$  contains all the information needed to determine the future evolution. The probability of a sequence, for a Markov process reads

$$P\{x_1, \dots, x_n\} = \left[ \prod_{t=1}^{n-1} P\{x_{t+1} | x_t\} \right] P\{x_1\}. \quad (11.2)$$

A Markov process where  $x_t$  takes values in a discrete set  $\mathcal{S}$  is called a *Markov chain*.<sup>2</sup> The elements of  $\mathcal{S}$  are also called *states* and we shall denote

<sup>1</sup>We use  $x_i$  as a shorthand for the event  $\{X_i = x_i\}$ .

<sup>2</sup>A full account of Markov chains is given in FELLER XV, of which what follows is a synthesis.

them by integers,<sup>3</sup> i.e.  $\mathcal{S} \subseteq \mathbb{N}$ . Because of Eq. (11.2), a Markov chain is fully determined by an initial distribution  $\alpha_i = P\{X_1 = i\}$  and by a matrix

$$p_{i,j} = P\{X_{t+1} = j | X_t = i\}, \quad i, j \in \mathcal{S}. \quad (11.3)$$

The matrix  $\hat{P} = \{p_{i,j}\}$  is called the *transition matrix*, because its elements give the probabilities of transitions between any two states.<sup>4</sup>

We have already encountered examples of Markov chains. A random walk  $S_n$  is a Markov chain, because  $S_{n+1}$  depends only on the position  $S_n$  of the walker at the previous step, not on how it got there (i.e. on  $S_t$  for  $t < n$ ). Branching processes  $Z_n$  are also Markov chains. In both cases, the state space  $\mathcal{S}$  is infinite. For simplicity, we shall limit our discussion to cases where  $\mathcal{S}$  is finite, and refer to FELLER for an extended treatment.

## 11.1 Stochastic matrices

The transition matrix  $\hat{P}$  satisfies positivity and normalisation, i.e.

$$p_{i,j} \geq 0 \quad \forall i, j \in \mathcal{S}, \quad \sum_{j \in \mathcal{S}} p_{i,j} = 1, \quad (11.4)$$

where the latter implies that from state  $i$  the Markov chain will move to another state  $j$  (possibly equal to  $i$ , if  $p_{i,i} > 0$ ). The two conditions (11.4) define the set of *stochastic matrices*. If  $\hat{P}$  and  $\hat{Q} = \{q_{i,j}\}$  are two stochastic matrices (on  $\mathcal{S}$ ), then their product  $\hat{P}\hat{Q}$  is also a stochastic matrix. Indeed,  $\{\hat{P}\hat{Q}\}_{i,j} \geq 0$  because it is the sum of non-negative terms and

$$\sum_{j \in \mathcal{S}} \{\hat{P}\hat{Q}\}_{i,j} = \sum_{j,k \in \mathcal{S}} p_{i,k} q_{k,j} = \sum_{k \in \mathcal{S}} p_{i,k} \sum_{j \in \mathcal{S}} q_{k,j} = \sum_{k \in \mathcal{S}} p_{i,k} = 1. \quad (11.5)$$

Therefore the set of stochastic matrices defined on a set of states  $\mathcal{S}$  with the matrix multiplication is a semi-group.<sup>5</sup>

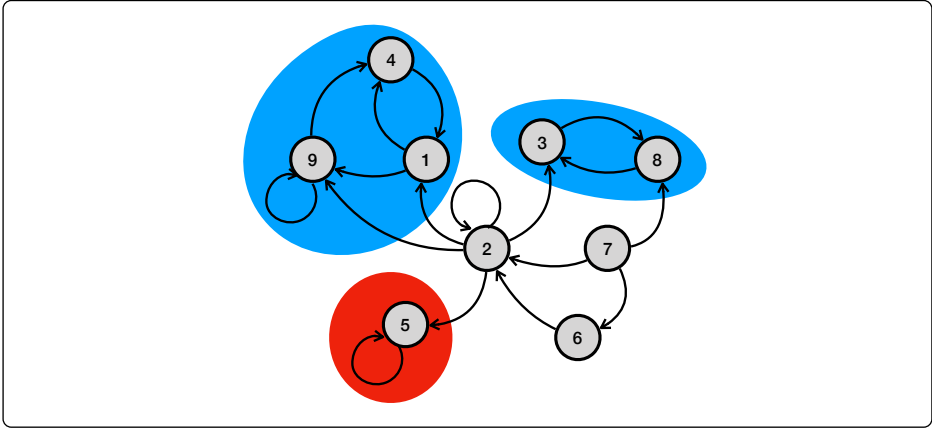
This property allows us to generate a Markov chain  $\hat{P}\hat{Q}$  by combining two Markov chains. By extension, the combination of any number of Markov chains is a Markov chain. In particular, combining a Markov chain  $\hat{P}$  with itself  $n$  times yields a Markov chain with transition matrix  $\hat{P}^n$ . Its matrix elements

$$p_{i,j}^{(n)} = \sum_{k_1 \in \mathcal{S}} \dots \sum_{k_{n-1} \in \mathcal{S}} p_{i,k_1} p_{k_1,k_2} \dots p_{k_{n-1},j} = P\{X_{t+n} = j | X_t = i\} \quad (11.6)$$

<sup>3</sup>The sample space of a Markov chain of  $n$  steps is  $\Omega = \mathcal{S}^n$ .

<sup>4</sup>In the most general case, the transition probability can also depend on time  $t$ . We limit our discussion to *homogeneous* Markov chains, for which this is not the case.

<sup>5</sup>It is not a group, because of the absence of an inverse.



**Figure 27.** Decomposition of states in a Markov chain (taken from FELLER XV.4):  $\mathcal{S} = \mathcal{T} \cup_i \mathcal{C}_i$ , where  $\mathcal{T} = \{2, 6, 7\}$  is the set of transient states,  $\mathcal{C}_1 = \{1, 4, 9\}$  and  $\mathcal{C}_2 = \{3, 8\}$  are closed sets and  $\mathcal{C}_3 = \{5\}$  is an absorbing state.

have a simple interpretation of transition probabilities between states in  $n$  steps. This probability is the sum over all paths  $i \rightarrow k_1 \rightarrow k_2 \rightarrow \dots k_{n-1} \rightarrow j$  from  $i$  to  $j$  through the intermediate states  $k_\ell$ .

## 11.2 Classification of states

A state  $i$  is connected to  $j$  if it is possible to go from  $i$  to  $j$  in one step. We denote this as

$$i \rightarrow j \Leftrightarrow \text{if } p_{i,j} > 0. \quad (11.7)$$

This directional relation can be visualised in a network where the nodes are the states  $\mathcal{S}$  and the possible transitions  $p_{i,j} > 0$  are represented as directed links, as shown in Figure 27. State  $j$  can also be reached from state  $i$  by a directed path of more than one step, that starts in  $i$  and reaches  $j$ . For example, in Figure 27, state 9 can be reached from 7 (e.g. by the path  $7 \rightarrow 2 \rightarrow 1 \rightarrow 9$ ), but there is no path from state 9 to 7 (i.e. 7 cannot be reached from 9). This leads to the definition of a *closed set*  $\mathcal{C} \subseteq \mathcal{S}$  which is a subset of states such that

i) no state  $j \notin \mathcal{C}$  can be reached from any state  $i \in \mathcal{C}$ , and<sup>6</sup>

ii) all states  $j \in \mathcal{C}$  can be reached from any state  $i \in \mathcal{C}$ .

<sup>6</sup>This differs from the definition in FELLER that considers only condition i). Without ii) the union of two closed sets would also be a closed set.

If a closed set is composed of a single state  $\mathcal{C} = \{\ell\}$  then  $\ell$  is called an *absorbing state*. States that do not belong to closed sets are called *transient* for reasons that will become soon clear. So the states of a Markov chain can be decomposed as

$$\mathcal{S} = \mathcal{T} \cup \bigcup_a \mathcal{C}_a,$$

where  $\mathcal{T}$  is the set of transient states. In order to mathematically define transient states, we resort to ideas similar to those used for random walks. Let  $T_{i \rightarrow j}$  be the time that a Markov chain that starts at  $X_0 = i$  at time  $t = 0$ , visits state  $j$  for the first time. This is a *first passage time*. Its distribution<sup>7</sup>

$$f_{i,j}^{(n)} = P\{T_{i \rightarrow j} = n\} = P\{X_n = j, X_t \neq j \forall 0 < t < n | X_0 = i\} \quad (11.9)$$

is the distribution of first passage times from  $i$  to  $j$ . Likewise one can define first return distributions as  $f_{i,i}^{(n)}$  (with  $f_{i,i}^{(0)} = 0$  by convention). The probability that a Markov chain that starts from state  $i$  ever returns to  $i$  is

$$f_{i,i} = \sum_{n=0}^{\infty} f_{i,i}^{(n)}.$$

If  $f_{i,i} = 1$  the Markov chain will surely return to state  $i$ , so we call state  $i$  *persistent*. If  $f_{i,i} < 1$  instead, state  $i$  is *transient*: with probability  $1 - f_{i,i} > 0$  the Markov chain will never return to state  $i$ . The first passage time distribution is related to the probability  $p_{i,j}^{(n)}$ , by the equation

$$p_{i,j}^{(n)} = \sum_{\nu=0}^n f_{i,j}^{(\nu)} p_{j,j}^{(n-\nu)}, \quad n > 0. \quad (11.10)$$

---

<sup>7</sup>As an example, consider a Markov chain between states  $\mathcal{S} = \{0, 1, \dots, N\}$  that satisfies the *martingale property*

$$\mathbb{E}[X_n | X_0] = X_0 \leftrightarrow \sum_{k \in \mathcal{S}} p_{i,k}^{(n)} k = i, \quad (11.8)$$

for all  $i \in [0, N]$  and  $n > 0$ . Eq. (11.8) for  $n = 1$  and  $i = 0$  cannot be true unless  $p_{0,0} = 1$  and  $p_{0,k} = 0$  for all  $k > 0$ . Likewise, for  $i = N$  and  $n = 1$ , the only possibility to satisfy Eq. (11.8) is to have  $p_{N,N} = 1$  and  $p_{N,k} = 0$  for all  $k < N$ . Hence 0 and  $N$  are absorbing states. If there is no further closed set, then the Markov chain  $X_n$  will either converge to 0 or to  $N$  as  $n \rightarrow \infty$ . This means that  $p_{i,k}^{(n)} \rightarrow 0$  as  $n \rightarrow \infty$  for all  $0 < k < N$ . Then Eq. (11.8) implies that the probabilities that the Markov chain is “absorbed” at either states 0 or state  $N$  are given by

$$\lim_{n \rightarrow \infty} p_{i,k}^{(n)} = \frac{k}{N}, \quad \lim_{n \rightarrow \infty} p_{k,0}^{(n)} = 1 - \frac{k}{N}.$$

This shows that the martingale property imposes very strong constraints on the process. Irrespective of the details of the dynamics on the transient states, this property allows us to determine the asymptotic probability that the Markov chain will be absorbed in either one of the two absorbing states.

This equation reads as follows: in order for the Markov chain to be at  $j$  if it started  $n$  steps before from  $i$ , it must visit state  $j$  for the first time at some intermediate time  $\nu \in [1, n]$ , and then return back to  $j$  after  $n - \nu$  steps. Each  $\nu$  determines a disjoint set of paths  $i \rightarrow j$ , so the probability of the event  $\{X_n = j | X_0 = i\}$  can be computed as the sum on the probabilities of paths going through  $j$  at time  $\nu$  for the first time.

### Exercise 11.1

Consider an urn containing  $N$  particles,  $X_0$  of which are black and the remaining  $N - X_0$  are white. At each step, draw  $N$  balls with replacement and let  $X_1$  be the number of black balls drawn. Build a new urn that contains  $X_1$  black balls and  $N - X_1$  white balls. Continue the process in the same way, with  $X_{n+1}$  being the number of black balls drawn with replacement from an urn with  $N$  balls,  $X_n$  of which are black. Show that  $X_n$  is a martingale and that, as  $n \rightarrow \infty$ , all balls in the urn will be either black or white.

If we take  $i = j$  and sum  $p_{j,j}^{(n)}$  over  $n \geq 0$ , we obtain

$$\sum_{n=0}^{\infty} p_{j,j}^{(n)} = 1 + \sum_{n=1}^{\infty} p_{j,j}^{(n)} = 1 + f_{j,j} \sum_{m=0}^{\infty} p_{j,j}^{(m)} \quad (11.11)$$

$$= \frac{1}{1 - f_{j,j}}, \quad (11.12)$$

where we used Eq. (11.10) for the sum on  $n > 0$  and we changed the sum over  $n$  and  $\nu$  into a sum over  $m = n - \nu$  and  $\nu$ . Therefore the series in Eq. (11.11) converges if  $f_{j,j} < 1$  (i.e. if  $j$  is a transient states). Eqs. (11.10) and (11.11) imply that the probability to find a Markov chain on a transient state  $j$  vanishes as  $n \rightarrow \infty$ . First because the convergence of the series in Eq. (11.11) implies that

$$\text{if } j \text{ is transient} \Rightarrow \lim_{n \rightarrow \infty} p_{j,j}^{(n)} = 0. \quad (11.13)$$

Second, taking the limit  $n \rightarrow \infty$  in Eq. (11.10), shows also that  $p_{i,j}^{(n)} \rightarrow 0$  as  $n \rightarrow \infty$  for all  $i \in \mathcal{S}$ , i.e. the Markov chain will not visit a transient state  $j$ , irrespective of where it starts from. The Borel-Cantelli lemma, that we shall discuss later in the course, states that convergence in Eq. (11.11) is a sufficient condition to ensure that state  $j$  will be visited only a finite number of times. As a consequence, after a transient period, the Markov chain will “enter” one of the closed sets  $\mathcal{C}_a$ . The dynamics will be confined to  $\mathcal{C}_a$  for all subsequent

times. It is easy to see that the Markov chain restricted to states  $i \in \mathcal{C}_a$  is itself a Markov chain, because all transitions to states  $j \notin \mathcal{C}_a$  are impossible (i.e.  $p_{i,j} = 0$ ).

### 11.3 The invariant distribution

In order to discuss the dynamics of a Markov chain on states belonging to the same closed set, let us focus on Markov chains with an unique closed set  $\mathcal{C} = \mathcal{S}$  that is identical to the whole set of states (i.e.  $f_{i,j} = 1$  for all  $i, j \in \mathcal{S}$ ). We restrict attention to the case where  $\mathcal{S}$  is finite ( $|\mathcal{S}| < +\infty$ ). Markov chains of this type are called *irreducible*.<sup>8</sup>

For an irreducible Markov chain, the probability  $p_{i,j}^{(n)}$  to visit state  $j$  after  $n$  steps, starting from  $i$ , converges to a limit

$$\lim_{n \rightarrow \infty} p_{i,j}^{(n)} = u_j, \quad \forall i, j \in \mathcal{S}. \quad (11.14)$$

For the proof of this statement relies on Perron-Frobenius theorem, which states that the maximal (in modulus) eigenvalue of a real square matrix with positive entries is real and is unique. This applies to our case because  $p_{i,j}^{(n)} > 0$  for all  $i, j \in \mathcal{S}$ , for sufficiently large  $n$ , because for an irreducible Markov chain every state  $j$  can be reached from any other state  $i$ , by a sufficiently long path. In addition, Perron-Frobenius theorem states that the corresponding eigenvector has strictly positive components and that the largest eigenvalue  $\lambda_1$  is bounded by

$$\min_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} p_{i,j} \leq \lambda_1 \leq \max_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} p_{i,j}$$

For a stochastic matrix this implies  $\lambda_1 = 1$ . Eq. (11.15) indeed coincides with the statement that the matrix  $\hat{P}$  has an eigenvalue equal to one with left eigenvector equal to  $u_i$ . Normalisation of  $p_{i,j}$  implies that the corresponding right eigenvector has all components equal to one.<sup>9</sup>

<sup>8</sup>Irreducible Markov chains are also called *ergodic*. The term ergodic also denotes recurrent states that occur with positive asymptotic probability (see FELLER).

<sup>9</sup>The spectral representation of  $\hat{P}$  gives detailed information on the Markov chain. Indeed, if  $u_i^{(m)}$  and  $v_i^{(m)}$  are the left and right eigenvectors corresponding to the  $m^{\text{th}}$  largest (in modulo) eigenvalues  $\lambda_m$ , then one can write

$$p_{i,j}^{(n)} - u_j = \sum_{m>1} v_i^{(m)} u_j^{(m)} \lambda_m^n \sim \lambda_2^n \rightarrow 0$$

as  $n \rightarrow \infty$ . So the convergence of  $p_{i,j}^{(n)}$  to  $u_j$  for large  $n$  is dominated by the second largest eigenvalue  $\lambda_2$  of  $\hat{P}$ . As a consequence, we expect that  $X_n$  is distributed according to  $u_j$  for times  $n \gg 1/|\log |\lambda_2||$ .



Notice that the limit does not depend on  $i$ . This means that the Markov chain loses *memory* of the initial state, when  $n \rightarrow \infty$ .

The left eigenvector  $u_j$  is a probability distribution that is called the *invariant distribution*. This is the asymptotic probability to find the Markov chain in state  $j$ , for  $n \rightarrow \infty$ .

Taking the limit  $n \rightarrow \infty$  on both sides of the equation

$$p_{i,j}^{(n+1)} = \sum_{k \in S} p_{i,k}^{(n)} p_{k,j}$$

gives

$$u_j = \sum_{k \in S} u_k p_{k,j}. \quad (11.15)$$

This equation shows that the distribution  $u_j$  is *invariant* under the action of  $\hat{P}$ , i.e. it is time translation invariant.

Let us now show that the probability  $u_i$  to be asymptotically at a recurrent state  $i$  is inversely proportional to the expected time it takes to return to that site. Indeed, for any recurrent states, we can use Eq. (11.10) to derive a relation between the generating function  $F_{i,i}(s)$  of first return times  $T_{i \rightarrow i}$  to  $i$  and the generating function

$$U_{i,i}(s) = \sum_{n=0}^{\infty} p_{i,i}^{(n)} s^n$$

of the probability  $p_{i,i}^{(n)}$  of returns to  $i$  at time  $n$ . This relation is analogous to Eq. (9.28) and it reads  $F_{i,i}(s) = 1 - 1/U_{i,i}(s)$ . This allows us to compute the expected return time to  $i$  as

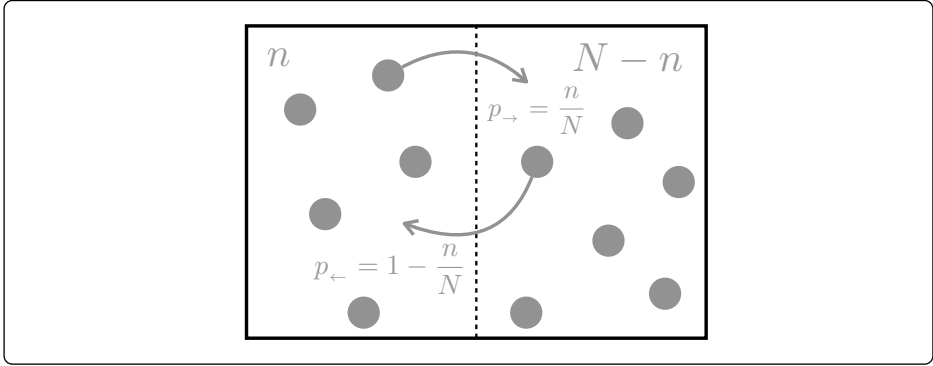
$$\begin{aligned} \mathbb{E}[T_{i \rightarrow i}] &= F'_{i,i}(1) = \lim_{s \rightarrow 1^-} \frac{d}{ds} \left[ 1 - \frac{1}{U_{i,i}(s)} \right] \\ &= \lim_{s \rightarrow 1^-} \frac{U'_{i,i}(s)}{U_{i,i}(s)^2}. \end{aligned} \quad (11.16)$$

The leading singularity for  $s \rightarrow 1$  of  $U_{i,i}(s)$  is given by

$$U_{i,i}(s) \simeq \frac{u_i}{1-s} + \dots$$

because  $p_{i,i}^{(n)} \rightarrow u_i > 0$  converges to a finite limit for  $n \rightarrow \infty$ , if  $i$  is a recurrent state. Then the limit in Eq. (11.16) yields

$$\mathbb{E}[T_{i \rightarrow i}] = \frac{1}{u_i} \quad (11.17)$$



**Figure 28.** The Ehrenfest model of diffusion.

which is what we set out to prove. The equation  $u_i = 1/\mathbb{E}[T_{i \rightarrow i}]$  has an intuitive meaning: the probability that a Markov chain is found in a recurrent state  $i$  is inversely proportional to the time it takes to return to  $i$ .

As an example, let us consider the *Ehrenfest model of diffusion*. This describes the equilibration of a gas of  $N$  particles in a box divided into two equal parts. Let  $n$  be the number of particles in the left side of the box. At each transition, one particle, chosen at random, moves from one side to the other. Hence the transition probability is

$$p_{n,n'} = \begin{cases} \frac{n}{N} & \text{for } n' = n - 1 \\ 1 - \frac{n}{N} & \text{for } n' = n + 1 \\ 0 & \text{otherwise} \end{cases}$$

The invariant distribution satisfies the equation

$$u_n = u_{n+1} \frac{n+1}{N} + u_{n-1} \left(1 - \frac{n-1}{N}\right), \quad 0 < n < N \quad (11.18)$$

and<sup>10</sup>  $u_0 = u_1/N$ ,  $u_N = u_{N-1}(1 - \frac{1}{N})$ . A solution for  $u_n$  can be found expressing  $u_{n+1}$  in terms of  $u_n$  and  $u_{n-1}$ . Then starting, from  $n = 0$ , we have

$$u_1 = Nu_0, \quad u_2 = \frac{N(N-1)}{2}u_0, \quad u_3 = \frac{N(N-1)(N-2)}{6}u_0, \quad \dots$$

<sup>10</sup>Note that, in order to read Eq. (11.18) you need to “invert” time: the probability to be in state  $n$  is the probability to be in state  $n - 1$  at the previous step and then to add one particle  $n - 1 \rightarrow n$ , plus the probability to be at  $n + 1$  and then to remove one particle ( $n + 1 \rightarrow n$ ). On the right hand side of Eqs. (11.18) and (11.15) you find the contribution from those states that can lead to state  $n$ .

Finally  $u_0$  can be determined by imposing normalisation. This leads to

$$u_n = \binom{N}{n} 2^{-N}.$$

This result is consistent with intuition. After a very long time you expect each particle to be on the left with probability  $1/2$ , hence  $n$  should have a binomial distribution.

These discussion extends to Markov chains that have more than one closed set in obvious ways. To each closed set  $\mathcal{C}_a$  we can associate an invariant distribution  $u_j^{(a)}$ , which vanishes on all states  $j \notin \mathcal{C}_a$ . The probability  $p_{i,j}^{(n)}$  will now converge to the invariant distributions  $u_j^{(a)}$  of closed set  $\mathcal{C}_a$  with a probability  $q_i^{(a)}$  that depends on the initial state  $i$

$$\lim_{n \rightarrow \infty} p_{i,j}^{(n)} = \sum_a q_i^{(a)} u_j^{(a)}.$$

If  $i \in \mathcal{C}_a$  then  $q_i^{(a)} = 1$  and  $q_i^{(a')} = 0$  for all  $a' \neq a$ .

## 11.4 Time reversibility

Imagine to observe a sequence of states  $\dots, X_n, \dots, X_{n+k}, \dots$  generated from a Markov chain. If we cannot distinguish it from the time reversed process  $\dots, X_{n+k}, \dots, X_n, \dots$ , then the Markov chain is *reversible*, i.e. it is invariant under time inversion. A Markov chain that starts from a state  $i$  will keep memory of that state for a finite time, so it makes sense to address this question only when  $n \rightarrow \infty$  and the sequence we're observing does not bear memory of its initial conditions. In this case, all transient states will not appear in the sequence, so it makes sense to restrict our discussion on time reversibility to irreducible chains.

The transition matrix of the (time) reversed chain can be computed using Bayes formula

$$q_{j,i} = P\{X_n = i | X_{n+1} = j\} \quad (11.19)$$

$$= \frac{P\{X_{n+1} = j | X_n = i\} P\{X_n = i\}}{P\{X_{n+1} = j\}} \quad (11.20)$$

$$= \frac{u_i p_{i,j}}{u_j} \quad (11.21)$$

where we used the fact that, for  $n$  large,  $P\{X_n = i\}$  converges to the invariant distribution  $u_i$ .

If  $q_{j,i} = p_{j,i}$  then there is no way in which the reversed process can be distinguished from the forward one. The reversibility condition  $q_{j,i} = p_{j,i}$  can also be stated in terms of the *detailed balance* condition

$$u_j p_{j,i} = u_i p_{i,j}, \quad \forall i, j \in \mathcal{S} \quad (11.22)$$

This equation states that a Markov chain is reversible if, asymptotically, the probability to observe transitions from any state  $i$  to any other state  $j$  equals the probability to observe the reverse transition  $j \rightarrow i$ . If the detailed balance condition is violated, the process is not reversible. This clearly happens if there are two states  $i$  and  $j$  for which transitions  $i \rightarrow j$  are possible but the reversed ones  $j \rightarrow i$  are not (i.e.  $p_{j,i} = 0$ ). In summary, in order to find out whether a Markov chain is reversible or not, the first step is to compute the invariant distribution  $u_i$  and the second is to check whether Eq. (11.22) holds for all  $i, j \in \mathcal{S}$  or not.

### Exercise 11.2

Is the Ehrenfest model of diffusion reversible?

### Exercise 11.3

Can a Markov chain on  $|\mathcal{S}| = 2$  states be irreversible?

## Chapter 12

# Exercises on the first part of the course

1. Consider two dice. Let

$$A = \{\text{sum of the faces is odd}\}$$

and

$$B = \{\text{at least one ace}\}.$$

Describe the events  $A \cup B$ ,  $A \cap B$  and  $A \cap \overline{B}$ . Assuming that each outcome is equiprobable, find the probabilities of all these events.

2. Find simpler expressions for

(a)  $(A \cup B) \cap (A \cup \overline{B})$ ,

(b)  $(A \cup B) \cap (\overline{A} \cup B) \cap (A \cup \overline{B})$

(c)  $(A \cup B) \cap (A \cup C)$

3. In 14000 tosses of a fair coin, one observes 7428 heads. Estimate the probability to observe a larger number of heads to two decimal digits?
4. Let  $A_1, \dots, A_n$  be mutually independent events and let  $P\{A_k\} = p_k$ . What is the probability that none of the events occur? Show that this probability is always less than  $e^{-\sum_k p_k}$ . Show that the same inequality holds if the events  $A_1, \dots, A_n$  are mutually exclusive with  $P\{A_k\} = p_k$ . Show that the probability that none of the events occur is always less than  $e^{-\sum_k p_k}$ .

5. Three dice are rolled. If no two show the same face, what is the probability that one is an ace?
6. Suppose that 5 men out of 100 and 25 women out of 10000 are color-blind. A colour-blind person is chosen at random. What is the probability of his being male?
7. In a trow of  $6n$  dice what is the probability to observe each face exactly  $n$  times?
8. Three dice are thrown.  $A$  is the event that two and only two dice show the same face. Compute the probability of  $A$ . Consider the event  $B$  that the sum of the outcomes is even. Are  $A$  and  $B$  independent?
9. Compute the probability that the sum of  $n$  dice is even and the probability that it is divisible by three.
10. A fair coin is tossed until for the first time the same result appears twice consecutively.  $A_n$  is the event that this occurs at the  $n^{\text{th}}$  toss. Compute the probability of  $A_n$ . Prove that the probability that the event  $A_n$  never occurs is zero. Consider the same problem in the case of the throw of a dice with  $k$  faces.
11. Consider an experiment where balls are consecutively put at random in  $n$  boxes. Compute the probability of the event

$$A_r = \{\text{box 1 is empty after } r \text{ draws}\}$$

Find a representation of the elements  $\Omega$  of the sample space that allows you to compute the probability of  $A_r$  and compute it. Is  $A_r \subset A_{r+1}$  or  $A_{r+1} \subset A_r$  or none of the two?

Let  $B_r = A_{r-1} \cap \overline{A_r}$  with  $A_0 = \Omega$ . Compute  $P\{B_r\}$  and compute

$$\lim_{r \rightarrow \infty} P\left(\bigcup_{k=1}^r B_k\right).$$

12. Show that if  $a, b > 0$  are integers, then the number of paths of  $n$  steps of a random walk that are always above  $-b$  and end at  $a$  is

$$|\{\omega : S_k(\omega) > -b, \forall k, S_n = a\}| = \binom{n}{\frac{n+a}{2}} - \binom{n}{\frac{n+a}{2} + b}$$

Show that if  $b > a > 0$  are integers, then the number of paths of  $n$  steps that are always below  $b$  and end at  $a$  is

$$|\{\omega : S_k(\omega) < b, \forall k, S_n = a\}| = \binom{n}{\frac{n+a}{2}} - \binom{n}{\frac{n-a}{2} + b}$$

13. Let  $S_n$  and  $S_m$  be two independent binomial random variables, denoting the number of successes in  $n$  and  $m$  experiments respectively. In both cases, the probability of success in a single trial is  $p$ . How would you show that  $S_n + S_m = S_{n+m}$ ? Show it.
14. Let  $X_\lambda$  and  $X_\mu$  be two independent Poisson random variables with parameters  $\lambda$  and  $\mu$  respectively. Compute the distribution of the variable  $X_\lambda + X_\mu$ . Could the result have been guessed?
15. The term echo chamber is used for situations where the opinion of an individual is reinforced by the opinion of others, who are themselves influenced by him/her. Consider a situation where Mr X may have two opinions  $\sigma_X = \pm 1$  about a particular issue. Contrast the situation where Mr X is in isolation and  $P\{\sigma_X\} = e^{h\sigma_X} / (2 \cosh h)$  to the one where he interacts with Ms Y. In the second case, Ms Y's opinion  $\sigma_Y \in \{\pm 1\}$  on the same issue is influenced by that of Mr X, so that the joint distribution is

$$P\{\sigma_X, \sigma_Y\} = \frac{1}{Z} e^{h\sigma_X + J\sigma_X\sigma_Y}$$

with  $Z$  a normalisation constant. Show that there is no echo chamber effect, in the sense that the probability that  $\sigma_X = 1$  is the same in both cases. Show that there is an echo chamber effect in the case when Mr X and Ms Y can also be undecided, i.e. if  $\sigma_X, \sigma_Y$  can also take value 0 besides  $\pm 1$ . Contrast the case where  $P\{\sigma_X\} = e^{h\sigma_X} / (1 + 2 \cosh h)$  when Mr X is in isolation to the case where he interacts with Y with the same joint distribution of  $\sigma_X, \sigma_Y$  as above (with a different  $Z$ ) (see [17] for the general case).

16. Consider the random variable  $X(\omega) : \Omega \rightarrow [0, \infty)$  with pdf  $p(x) = Ax^{q-1}e^{-x}$ . Compute  $A$ , the mean and the variance. Compute the expected value of  $e^{-sx}$ . Do the same for a random variable  $X(\omega) \in \mathbb{R}$  with  $p(x) = Ae^{x-e^x}$ .
17. Let  $X_1$  and  $X_2$  be two independent uniform random variables. What is the pdf of  $X_1$  conditional on the event  $A = \{X_1 < X_2\}$ ? (*Hint*: consider the event  $B = \{X_1 \in [x, x + dx)\}$ ). Imagine now there are  $n$  uniform random variables. What is the pdf of  $X_1$  conditional on the event  $A = \{X_1 < X_i \forall i = 2, \dots, n\}$ ?
18. The show at a theater in Moskow costs 5 rubles.  $2n$  people show up in a random order.  $n$  of them have only notes of 10 rubles, whereas the rest has notes of 5 rubles.  $A$  is the event that the cashier has no change to

give to some customer. Translate this problems in probability. What is  $\Omega$ ? What is  $\mathcal{P}$ ? Compute the probability of  $A$ .

19.  $N$  gentlemen go to theater each leaving his hat at the wardrobe. On exit they are assigned their hats in a random order.  $A$  is the event that none of the gentlemen get his own hat back. Translate this problems in probability. What is  $\Omega$ ? What is  $\mathcal{P}$ ? Compute the probability of  $A$ .
20. Records: let  $X_1, X_2, \dots, X_n, \dots$  be a sequence of i.i.d. random variables drawn from a continuous distribution with pdf  $p(x)$ . Let  $A_n = \{X_n > X_i \forall i < n\}$  be the event that  $X_n$  is a record. Show that the events  $A_n$  are independent.
21. A group of  $n$  couples (husband and wife) arrives in a hotel. All the  $2n$  people get distributed at random in  $n$  double rooms. What is the probability that Mrs Smith and Mr Smith are assigned the same room? What is the probability that no wife is assigned a room with her husband?
22. A smoker has two boxes of  $n$  matches in the two pockets of his coat. Each time the smoker picks a match from a pocket chosen at random. Consider the event  $A_r$  that when he picks the last match from one of the boxes, the other still contains  $r$  matches. Compute the probability of  $A_r$ .

Consider the case where the smoker has only one box with  $2n$  matches:  $n$  of them are red and the other  $n$  are blue. Let  $A_r$  now be the event that when he picks the last match of one color, there are still  $r$  matches of the other color in the box. What is the difference with the previous case?

Consider the case where the two boxes contain  $N \geq n$  matches and a match is chosen each time with equal probability. When the  $n$  matches of the same colour are chosen, what is the probability that  $k$  more matches are left in the box of matches of the other colour?

23. Prove that

$$P\{S_k \geq 0, 0 < k < 2n; S_{2n} = 0\} = 2f_{2n+2}$$

*Hint:* look at the paths.

24. How far from the origin do you expect a point drawn at random inside a  $d$  dimensional sphere to be? Consider the case  $d \rightarrow \infty$ .



25. Law of succession of Laplace: imagine that in  $n$  independent trials of an experiment, the event  $A$  has always occurred. What is the probability that the event  $A$  will occur in the  $n + 1^{\text{th}}$  trial?

*Hint:* the first  $n$  trials can be used to estimate the probability  $p = P\{A\}$  of  $A$  in a single trial. Assume that, *a priori*  $P\{p \in [x, x + dx]\} = dx$  if  $x \in [0, 1)$  and 0 otherwise.

How you would write a program that simulates the experiment?

26. In  $n$  successive Bernoulli trials, each with probability of success  $p$ , what is the probability that the last success occurs at trial  $n - k$ ? What is the expected value of  $k$ ? What is the expected value of  $p^{-k}$ ?
27. A particle moves on a one dimensional lattice at discrete time steps. Let  $x_t$  be its position and let

$$P\{x_{t+1} = y | x_t\} = \begin{cases} p & \text{if } y = x_t + 2 \\ 1 - p & \text{if } y = x_t - 1 \\ 0 & \text{else} \end{cases}$$

and let  $x_{t=0} = z$ . Find the probability that the particle will ever reach the origin (this has an interpretation in terms of gambling: you play a game where you win two euros with probability  $p$ , and you lose one euro otherwise. If you enter the game with  $z$  euros, what is the probability that you will lose all?).

28. Let  $N$  have a Poisson distribution with mean  $\lambda$  and let  $N$  balls be placed randomly in  $n$  cells. Show that the probability of finding exactly  $m$  cells empty is

$$p_{n,m} = \binom{n}{m} e^{-\lambda m/n} (1 - e^{-\lambda/n})^{n-m}$$

29. Bivariate generating function: let  $p_{n,k} = P\{X = n, Y = k\}$  be the joint distribution of the variables  $X$  and  $Y$ . Consider the joint generating function

$$P(s, z) = E[s^X z^Y] = \sum_{n,k} p_{n,k} s^n z^k$$

Show that the generating function of the marginal distributions  $P\{X = n\}$  and  $P\{Y = k\}$  are given by  $P_X(s) = P(s, 1)$  and  $P_Y(z) = P(1, z)$  respectively. Find an expression for the covariance  $E[(X - E[X])(Y - E[Y])]$  in terms of  $P(s, z)$ .

30. A student is assigned every day a new problem with probability  $p$ . In one day, s/he can solve at most one problem with probability  $q$  (not necessarily  $q = 1 - p$ ). If s/he does not solve the problem s/he will try to solve it again the next day, so problems may pile up on her/is desk and need to work on them as long as the pile is not empty. At day  $t = 0$ , s/he is assigned one problem. What is the probability that s/he will ever have at least one day free?
31. In a sequence of coin tossing, let  $a_n$  be the probability that the pattern HHH does not occur in the first  $n$  draws. Estimate the leading asymptotic behaviour of  $a_n$  for  $n$  large.
32. The event  $A_t$  that a farmer goes and collects eggs from his chickens at day  $t = 0, 1, 2, \dots$  are independent for each  $t$ , and  $P\{A_t\} = p$ . Each day  $t$ , chickens produce a number  $X_t$  of eggs that is an i.i.d. Poisson random variable with mean  $\lambda$ . What is the probability that the next time the farmer goes and collects the eggs he find none?
33. Consider a branching process where each individual can have  $X = 0, 1$  or  $2$  offsprings, with probabilities  $(1 - p)^2$ ,  $2p(1 - p)$  or  $p^2$  respectively. Find the extinction probability  $x$  and verify that  $x = 1$  if  $E[X] = \mu \leq 1$ . Compute the generating function of the total progeny and discuss the asymptotic behavior of the probability  $R_n$  that the total progeny of one individual is of size  $n$ . Discuss in particular the case where  $\mu = 1$ .
34. Consider the sequence of numbers

$$l_0 = 2, l_1 = 1, l_{n+2} = l_{n+1} + l_n, \quad n \geq 0$$

Find the generating function and an explicit expression of  $l_n$ .

35. Define a sequence of integers,  $\{P_n\}$  by the initial conditions  $P_1 = 1$ ,  $P_2 = 2$ , and the recurrence  $P_n = 2P_{n-1} + P_{n-2}$  for  $n \geq 3$ . To what real number does the sequence  $(P_{n-1} + P_n)/P_n$  converge?
36. Consider a random walk in  $d$  dimensions  $\vec{S}_n = (S_n^{(1)}, \dots, S_n^{(d)})$  where each component  $S_n^{(a)}$  is an independent random walk of  $n$  steps. Compute the probability  $p_n$  that the random walk returns to the origin (i.e.  $S_n^{(a)} = 0 \forall a = 1, \dots, d$ ) after  $n$  steps. Estimate the asymptotic behavior of  $p_n$ .
37. The inspection time paradox. Imagine a process that occurs at times  $t_i$ , with  $i = \dots, -2, -1, 0, 1, 2, \dots$ . Let the inter-event intervals  $\tau = t_i - t_{i-1}$  be i.i.d. random variables with pdf  $p(\tau)$ . Let  $T \in \mathbb{R}$  be a random time

and  $T^* = \min\{t_i : t_i > T\}$  be the time of the next event. Show that, depending on  $p(\tau)$ , the expected time  $T_r = T^* - T$  for the next event can be smaller, equal or larger than the expected inter-event time  $\mathbb{E}[\tau]$ . In particular, show that

$$\mathbb{E}[T_r] = \frac{\mathbb{E}[\tau^2]}{2\mathbb{E}[\tau]}$$

so that  $\mathbb{E}[T_r] > \mathbb{E}[\tau]$  whenever  $\mathbb{V}[\tau] > \mathbb{E}[\tau]^2$ .

38. In a random population, the number  $k$  of friends that each individual has is an i.i.d. random variable with distribution  $P(k)$ . Show that in such a population, the expected number of friends of an individual is smaller than that of his/her friends.
39. Completion time with resetting: consider a process that takes a random time  $T$  to complete, i.e. to reach a final state  $x_1$  from an initial state  $x_0$ . As an example, think of a random walk starting from  $x_0 \neq 0$  and let the process be complete when it hits the origin  $x_1 = 0$ . Let us consider the generic case where  $T$  is a continuous random variable with pdf  $p(t)$ . Consider introducing resetting at random times. This means that, as long as the process is not completed, in any interval  $[t, t+dt)$  the process re-starts from  $x_0$  with probability  $r dt$  (and it continues with probability  $1 - r dt$ ), for an infinitesimal  $dt$ . More precisely, Let  $T_r$  be the time of completion with resetting, one naïvely expects that  $\mathbb{E}[T_r] \geq \mathbb{E}[T]$ . Show that this is not true, using the relation

$$T_r = \begin{cases} T & \text{if } T \leq R \\ R + T'_r & \text{if } T > R \end{cases}$$

between the time  $T$  to completion without reset, the reset waiting time  $R$ , and the time to completion with reset  $T_r$  where  $T'_r$  has the same distribution as  $T_r$ . Find an equation for  $\tau_r(s) = \log \mathbb{E}[e^{-sT_r}]$  and show that

$$\mathbb{E}[T_r] = \frac{1 - \mathbb{E}[e^{-rT}]}{r\mathbb{E}[e^{-rT}]}.$$

One may naïvely expect that introducing resetting delays completion. Using the small  $r$  expansion of this expression, show that if  $\mathbb{V}[T] > \mathbb{E}[T]^2$ , introducing resetting decreases the expected completion time. Notice, in particular, that the expected time to reach the origin of a random walk starting at  $x_0$  is infinite, but it becomes finite when resetting is introduced. Explain why this is so.

40. What is the number of shortest walks that start at one corner and end at the opposite corner of a chessboard?
41. Mr X does a test for a rare disease that hits one individual in a million, on average. The test is very reliable: the test is positive in 99% of the cases with the disease and in 1% of the cases that do not. What is the probability that Mr X has the disease, given that his test is positive?
42. You are at a metro stop at peak hour, trains arrive every 3 minutes on average in each direction. What is the probability that before the next train you will see  $k$  trains coming in the other direction? Imagine that instead trains arrive exactly every 3 minutes at the stop in each direction. What is the probability that before the next train you will see  $k$  trains coming in the other direction?
43. Each package of Pokemon cards contains 1 of  $N$  possible legendary Pokemon. How many packs do you expect you have to buy to get all  $N$ ? We assume all  $N$  are equally likely with each purchase.
44. A mailman delivers  $n$  letters at random to  $n$  recipients. The probability that the first letter goes to the right person is  $1/n$ , so the probability that it doesn't is  $1 - 1/n$ . Thus the probability that no one gets the right letter is  $(1 - 1/n)^n \approx 1/e = 37\%$  for  $n$  large. This argument is clearly wrong for  $n = 2$ , why? Find the correct expression for this probability and show that the prediction is right for  $n \rightarrow \infty$ .
45. Suppose there are  $A$  defects among  $N$  items. We sample  $n$  items at random. What is the probability  $p_a$  of finding  $a$  defects in this sample? Show that if  $A = pN$  and  $N \rightarrow \infty$  with  $n, p$  and  $a$  fixed, then

$$p_a = \binom{n}{a} p^a (1-p)^{n-a}.$$

46. Let  $X_1, X_2, \dots, X_n, \dots$  be a sequence of i.i.d. random variables with  $P\{X_i = +1\} = P\{X_i = -1\} = 1/2$ . Consider the random variable

$$Q_{1:n} = X_1 + X_1 X_2 + X_1 X_2 X_3 + \dots + X_1 X_2 \dots X_n = X_1 (1 + Q_{2:n}).$$

Show that  $P\{Q_n = x\} = P\{S_n = x\}$ , for  $x = 0, \pm 1, \pm 2, \dots, \pm n$ , where  $S_n = X_1 + X_2 + \dots + X_n$  is a random walk.

47. A monkey is standing one step from the edge of a cliff (i.e. if he takes one step in the direction of the cliff he falls) and takes repeated independent steps; forward (i.e. towards the edge of the cliff), with probability  $p$ , or backward, with probability  $q = 1 - p$ .

- What is the probability  $x$  that the monkey, sooner or later, will fall off the cliff?
  - Let  $p < q$ . Show that  $x$  is the extinction probability not only of the monkey, but also of a branching process with reproduction distribution  $p_0 = p, p_2 = q, p_k = 0 \forall k \neq 0, 2$ .
  - Conditional on the event that he falls, sooner or later from the cliff, what is the expected waiting time?
  - What is the variance of this waiting time?
48. Balls in unequal boxes. Let  $n$  balls be drawn independently in  $N$  boxes. For each ball, its probability to fall in box  $i$  is  $p_i = \lambda_i/N$ , with  $\sum_{i=1}^N p_i = 1$ .
- Let  $n = 2$  and compute the probability that two balls fall in the same box. Is this smaller or larger than the case where  $p_i = 1/N$  for all  $i$ ?
  - Compute the generating function of the number  $n_i$  of balls in box  $i$ . What is the distribution of  $n_i$  in the limit  $N \rightarrow \infty$  with  $n = N$  when  $p_i = \lambda_i/N$ ?
  - Let  $I$  be a subset of the integers  $1, 2, \dots, n$ . Compute the generating function of

$$n_I = \sum_{i \in I} n_i.$$

Show that the random variables  $n_i$  are not independent.

- Consider the limit  $N \rightarrow \infty$  with  $n = N$  of this random variable where  $I$  is a subset of a finite number of elements, with  $p_i = \lambda_i/N$  and  $\lambda_i$  finite for all  $i \in I$ . Show that in this limit  $n_i$  become independent random variables.
49. At each bus stop, one passenger drops from the bus and, with probability  $p_k = \frac{a^k}{k!} e^{-a}, k = 0, 1, 2, \dots$  passengers get on the bus. The bus starts with one passenger. What is the probability that the bus will never be empty (assuming the number of stops is very large)? Write down an equation for the generating function of the probability that the bus will be empty for the first time at stop  $t$ .
50. Let  $p_k, k \geq 0$  be the probability that each individual of a population, at each generation, contribute with  $k$  offsprings to the next generation.

Let

$$P(s) = \sum_{k=0}^{\infty} p_k s^k$$

be the corresponding generating function, and consider the branching process where, from a single individual at generation  $n = 0$ , a population of individuals is produced at any successive generation  $n + 1$  by each individual of the current generation  $n$  producing offsprings according to  $p_k$ , independently. Let's call this the branching process  $p$ .

i) Show that for any  $z \in (0, 1)$ , the distribution of  $k$

$$q_k(z) = \frac{p_k z^k}{P(z)}$$

is normalised to one.

Define the  $q$  branching process using  $q_k$  as the probability of generating  $k$  offsprings from each individual. Compute the corresponding generating function  $Q(s)$ . Consider the case where the  $p$  branching process is overcritical, i.e. that the extinction probability  $x_p$  is less than one. ii) Show that for  $z = x_p$  the  $q$  branching process defined by  $q_k$  is under critical i.e.  $x_q = Q(x_q) = 1$  and  $\mu_q = Q'(1) < 1$ . Let  $Y_{\infty}^{(p)}$  be the total progeny of the branching process  $p$  and  $Y_{\infty}^{(q)}$  be the total progeny of the branching process  $q$ . iii) Show that, conditional on  $Y_{\infty}^{(p)} < \infty$ , their distribution is the same, i.e.

$$P\{Y_{\infty}^{(p)} = n | Y_{\infty}^{(p)} < +\infty\} = P\{Y_{\infty}^{(q)} = n\}.$$

51. The tradition of the dynasty of Mr K demands that each family generates children until they have at least one daughter and one son. Each children reaches reproductive age with probability  $p$  and only males carry the surname K. Show that if  $p \leq 2/3$  the surname K will surely disappear. What is the expected size of the total progeny of Mr K, conditional on it being finite?

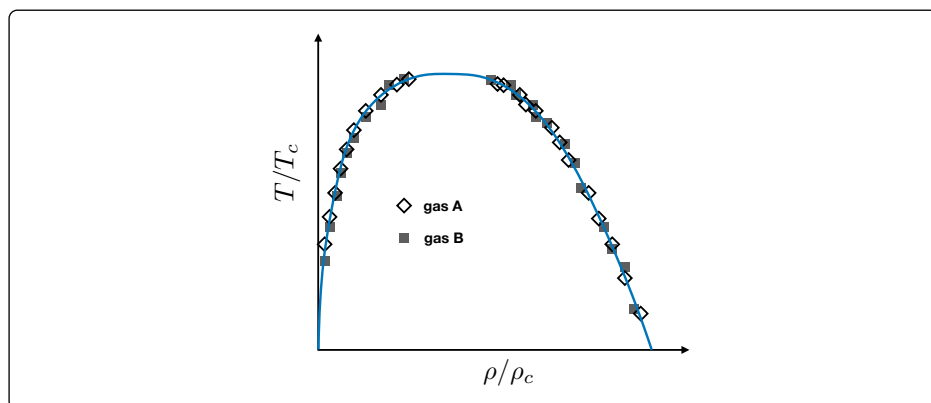
## **Part II**

# **Typical and atypical**





A gas is formed of  $\sim 10^{23}$  interacting molecules. Its detailed description would require to integrate Newton's law for all of them. Yet we can describe the macroscopic behaviour of a gas in terms of few variables (e.g. temperature, pressure, density) with a remarkable precision. Likewise the specific heat of a piece of glass is well defined and it turns out to be the same for any other piece of the same glass, in spite of the fact that in each piece the arrangement of atoms is different. These are two examples of the *typical* behaviour that emerges in systems of many degrees of freedom. This behaviour is remarkably robust and largely independent of microscopic details. Indeed, gases with different chemical composition obey the same laws in terms of appropriate macroscopic variables, to a very good degree of accuracy, especially in particular conditions (i.e. close to critical points) as sketched in Figure 29 (see [18] for more details).



**Figure 29.** (Sketch of) the equation of state (that relates temperature and density) of real gases close to the critical point. The data points represent two different gases A (e.g. Argon) and B (e.g. methane).

*Statistical mechanics* — a discipline developed by Ludwig Boltzmann and others in order to derive the macroscopic behaviour of physical systems from the (classical or quantum) microscopic description — has shown that the macroscopic behaviour is an exquisitely statistical phenomenon, whose origin has to do more with probability than with physics. Newton's laws of motion do not rule out that all molecules of the gas in a room concentrate in a small corner, leaving the rest of the room empty. This is possible but it is highly unlikely. *Typically* the molecules occupy uniformly the volume available to them. The probability of seeing a substantial deviation from this *typical* behavior is so small that we don't expect it has ever happened since the Big Bang. Statistical mechanics allows us to classify the *typical behaviour* of many particles into

different *phases of matter*, separated by *phase transitions*. Quantities such as the entropy and the temperature, whose meaning was elusive until Boltzmann, found a precise formalisation. Microscopic details are even more irrelevant at particular points of the phase diagram of a macroscopic physical system, that separate different phases of matter. The *critical phenomena* that govern the behaviour of a number of properties at second order phase transition points are so *universal* that the same quantitative laws may describe systems as different as binary alloys, ferromagnets and liquids.<sup>1</sup>

How can information be optimally represented? How can a message be efficiently coded in bits? What is the maximal achievable compression of a text in a given language, that optimises the use of a given storage capacity? How can a message be coded in such a way that it can be retrieved even if it is corrupted by noise when it is transmitted? These are apparently very different questions, but they hinge on understanding the *typical* structure of the messages we're interested in.<sup>2</sup> Claude Shannon and others, have shown that understanding their structure allows us to represent messages most efficiently, and to give a bound on the number of bits needed to compress a message. He was undecided on how to call this bound. Apparently [19], John Von Neumann told him: "You should call it *entropy*, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

Others used similar ideas to understand how to "dress" messages with structures that can make them robust with respect to noise. Such an error correction algorithm works efficiently as long as the noise level is below a certain threshold, that marks a *phase transition* to a regime where the noise is so strong that the original message is "lost in transmission". Information theory has been primarily developed in computer science and electrical engineering, but it's applications go well beyond these fields. For example, learning is a distinguishing feature of life, as opposed to inanimate matter. We more and more realise that understanding efficient information processing is key for a quantitative approach to how living systems learn, adapt and respond. For example, evolution has selected species that learn about their environment,

---

<sup>1</sup>Boltzmann himself, in a speech in 1904, remarked that "The wide perspectives opening up if we think of applying this science [statistical mechanics] to the statistics of living beings, human society, sociology and so on, instead of only to mechanical bodies, can here only be hinted at in a few words", suggesting that this general idea could be applied not only to physics but also to other domains.

<sup>2</sup>A sentence in English is a sequence of letters of the alphabet, but not all sequences of letters are meaningful English sentences. Typical sentences in English have a peculiar structure.

in spite of the fact that information processing is costly. Why is this so?

Both statistical mechanics and information theory deal with the *direct problem*, where the model that describes the interaction between particles, or the way in which messages are generated, is (assumed to be) known, and the statistical behaviour can be derived from it. *Statistics* deals with learning a model from observed behaviour (i.e. data). This entails solving an *inverse problem* with respect to that of statistical mechanics: given an observed collective behaviour, what is the model that would *typically* generate it? Choosing which model best describes a data-set is conceptually similar to finding which “phase” a physical systems belongs to, under certain conditions. Phase transitions separate statistical hypotheses and models as they separate behaviours of matter in physics.

Inference in physics is often so much constrained by what we know that statistics does not need to go much further than mean and variance.<sup>3</sup> In life sciences, high-throughput experiments produce massive amounts of data on systems we know very little about (e.g. multi-electrode recordings in the brain, gene expression profiles in cells, contacts in social networks). What can we learn from these data? How much information is there? How relevant are the variables we’re measuring? Can we reconstruct mathematical models that reproduce these data? How much data do we need to do that?

In some way or another, all these questions are related to understanding what is the *typical* behaviour that arises in “large” systems composed of many variables. There is a lot that one can learn from the *direct approach*, studying sequences of independent and identically distributed random variables. We can understand why statistical regularities arise, why macroscopic behaviour depends only on few relevant variables, what is the rôle of the entropy, and when *universal* features emerge. We will see that when the interaction (i.e. statistical dependence) among the variables is turned on, *phase transitions* will emerge to separate distinct statistical behaviours (i.e. phases).

It is also important to study *atypical behaviour*, i.e. to understand how unlikely are deviations from the typical behaviour and how atypical deviations are expected to occur *typically*. For example a living cell needs to *deviate* from it’s thermodynamic equilibrium with the environment — that would coincide with its death — by spending energy in very precise ways in order to meet some constraints. So it’s typical state can be considered as a *large deviation* with respect to thermodynamic equilibrium, i.e. as an *atypical* state where these constraints are enforced. This also applies to hypothesis testing, which is

---

<sup>3</sup>Yet, even in physics Machine Learning is being used more and more in fields like astrophysics, condensed matter, biophysics and even string theory, in order to cope with the huge amounts of data coming from experiments.

a traditional subject of statistics. An hypothesis can be rejected if it is possible to prove that the observed data would be very atypical if the hypothesis were correct.

Asymptotic properties of ensembles of many random variables, typical and atypical behaviour are the central themes of this part of the course. *Information theory* provides key insights as well as a language to properly describe this behaviour. We will learn how *universality* and *phase transitions* arise, and how these concepts can be applied to Statistical mechanics and coding theory, as well as to statistical inference and learning.

A detailed treatment of some of these subjects is available in other texts, specially COVER, to which we shall refer frequently.

## Chapter 13

# Almost surely et el.

The epistemological value of probability theory is based on the fact that chance phenomena, considered collectively and on a grand scale, create non-random regularity. (AN Kolmogorov, 1954)

The general type of question addressed in what follows concerns the laws of probability for events or random variables which involve  $N$  events  $A_1, \dots, A_N$  or random variables  $X_1, \dots, X_N$  in the limit  $N \rightarrow \infty$ .

There is a class of results, known as 0 – 1 laws, that concern events  $E_N$  which depend on  $N$  events or random variables, and state that, under some conditions,  $P(E_N) \rightarrow 0$  or  $1$  as  $N \rightarrow \infty$ . Events  $E_N$  for which  $P(E_N) \rightarrow 1$  as  $N \rightarrow \infty$  are said to occur *almost surely*, meaning that their probability is equal to one. It is customary to use the abbreviation

a.s. = almost surely

Almost refer to the fact that the complement of this event need not be the empty set, i.e. it may be possible to find realisations  $\omega \in \Omega$  for which the event  $E_N$  does not occur.  $E_N$  occurs almost surely if the probability of all the sample points  $\omega$  for which it does not happen tends to zero, as  $N \rightarrow \infty$ .

### Exercise 13.1

As an example, we say that an unbiased random walk (in one dimension) almost surely returns to the origin. Formally, if  $E_N = \bigcup_{k \leq N} \{S_k = 0\}$  where  $S_k$  is a random walk, then  $E_N$  occurs almost surely. There are clearly many realisations of the random walk which will never return to the origin. Yet their probability tends to zero as  $N \rightarrow \infty$ . Show that  $P\{E_N\} \rightarrow 1$  as  $N \rightarrow \infty$ .

The aim of this chapter is to familiarise with the concepts and the logic involved in the limit behaviour of probability laws. We start by discussing the issue of convergence.

## 13.1 Limits in probability

The convergence of a sequence  $x_n$  to a limit  $x$

$$\lim_{n \rightarrow \infty} x_n = x$$

has an unambiguous meaning. It means that  $\forall \epsilon > 0$  there is an  $N(\epsilon) \in \mathbb{N}$  such that for all  $n > N(\epsilon)$  the sequence  $x_n$  is away from  $x$  by at most  $\epsilon$ , i.e.  $|x_n - x| < \epsilon$ . In other words, *deviations of  $x_n$  from the limit  $x$  larger than  $\epsilon$  occur only a finite number of times, for any  $\epsilon$ .*

When  $X_n$  is a random variable, this definition cannot be used.<sup>1</sup> For a sequence of random variables  $X_n(\omega)$ , there are different ways in which the statement  $X_n(\omega) \rightarrow X(\omega)$  can be interpreted, because we're dealing with the convergence of functions.

### 13.1.1 Almost certain convergence

For fixed  $\omega$ , the statement  $X_n(\omega) \rightarrow X(\omega)$  reduces to convergence of sequences, so it is well defined. If this happens for all  $\omega \in \tilde{\Omega}$  where  $P\{\tilde{\Omega}\} = 1$ , we say that

$$X_n(\omega) \rightarrow X(\omega) \text{ a.s.}$$

One way to state a.s. convergence is:

$X_n \rightarrow X$  a.s. if for any  $\epsilon, \delta > 0$  there is a  $N(\epsilon, \delta)$  such that

$$P\{|X_n - X| < \epsilon, \forall n > N(\epsilon, \delta)\} \geq 1 - \delta.$$

In other words, the probability that there are no deviations larger than  $\epsilon$  from the limit, beyond a certain value of  $n$ , can be made arbitrarily close to one.

If  $X_m \rightarrow X$  a.s., then for any  $\epsilon > 0$ , the events

$$A_n = \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\} \tag{13.1}$$

should occur at most a finite number of times. This condition can be rephrased by saying that the probability that  $A_n$  occurs *infinitely often* is zero, i.e.

$$P(A_n \text{ i.o.}) = 0.$$

---

<sup>1</sup>This material is also discussed in Chapter 2 of [20].

Here “i.o.” stands for “infinitely often”, which is a term of common use in probability, that is worth discussing in more detail.

For a given  $\omega \in \Omega$ ,  $A_n$  occurs infinitely often if  $\omega \in A_n$  for an infinite sub-sequence of indices  $n$ . If such a sub-sequence exists, then for any  $m$  there must be at least one  $A_n$  with  $n \geq m$  that occurs. In other words we can write<sup>2</sup>

$$\{A_n \text{ i.o.}\} = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n. \quad (13.2)$$

Here the union  $\bigcup_{n=m}^{\infty} A_n$  indicates the event that at least one event  $A_n$  with  $n \geq m$  occurs and the intersection indicates that this occurs for all  $m$ .

### Exercise 13.2

Show that for any finite  $M$

$$A_M = \bigcap_{m=1}^M \bigcup_{n=m}^M A_n \neq \bigcap_{m=1}^M \bigcup_{n=1}^m A_n = A_1$$

This shows that it is important to take the limits in Eq. (13.2) in a well defined order

$$\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \equiv \lim_{M \rightarrow \infty} \lim_{N \rightarrow \infty} \bigcap_{m=1}^M \bigcup_{n=m}^N A_n.$$

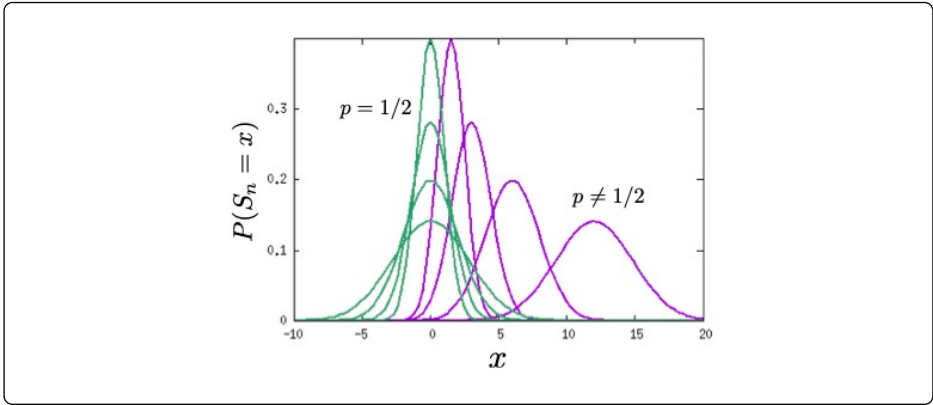
If the set of  $\omega$  for which this happens has probability one, i.e. if there is a.s. an infinite sub-sequence of indices  $n$  for which the events  $A_n$  occur, then we say that  $A_n$  occurs infinitely often. So while i.o. refers to events, a.s. refers to how the probability measure is defined.

In order to help intuition, let us consider as an example the sequence of events  $A_n = \{S_n = 0\}$  where  $S_n$  is the random walk discussed in a previous chapter. Then  $\{A_n \text{ i.o.}\}$  is the event that the random walk returns to the origin infinitely often. As we discussed, an unbiased random walk ( $p = 1/2$ ) surely returns to the origin and it does so an infinite number of times, almost surely. Hence  $P(A_n \text{ i.o.}) = 1$  for  $p = 1/2$ . Inspection of the distribution of  $S_n$  suggests why this is so. Indeed the probability distribution of the position of the random

<sup>2</sup>In other texts you will find the notation

$$\limsup_{n \rightarrow \infty} A_n = \{A_n \text{ i.o.}\}$$

to denote the set of points  $\omega$  for which  $A_n$  occurs infinitely often.



**Figure 30.** Probability distribution of the position  $S_n$  of a random walk for increasing values of  $n$ . An unbiased random walk ( $p = 1/2$ ) returns infinitely often to the origin almost surely, whereas a biased one ( $p \neq 1/2$ ) does not.

walk remains centred around the origin for all  $n$  and is has its maximum at  $S_n = 0$ . For  $p \neq 1/2$  instead the distribution “moves” away from the origin as  $n$  increases (see Figure 30). The reason why  $A_n$  occurs only a finite number of times, as we shall prove later, is that the probability that  $S_n = 0$  vanishes very fast as  $n$  increases, for  $p \neq 1/2$ .

### 13.1.2 Convergence in probability

If  $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} P\{|X_n - X| > \epsilon\} = 0$$

then we say that  $X_n(\omega) \rightarrow X(\omega)$  in probability. Almost certain convergence implies convergence in probability.<sup>3</sup>

<sup>3</sup>Indeed the event

$$D_n = \bigcup_{m=n}^{\infty} A_m, \quad A_m = \{|X_m - X| > \epsilon\}$$

that at least one  $X_m$  deviates more than  $\epsilon$  from  $X$  for some  $m \geq n$ , is telescopic, i.e.  $D_n \supseteq D_{n+1}$  for all  $n$ , because  $D_{n+1}$  implies  $D_n$ . Therefore, their intersection Eq. (13.2) equals the limit

$$\{A_n \text{ i.o.}\} = \lim_{n \rightarrow \infty} D_n.$$

If  $X_n \rightarrow X$  a.s., then

$$\lim_{n \rightarrow \infty} P(A_n) \leq \lim_{n \rightarrow \infty} P(D_n) = 0.$$

because  $A_n \subseteq D_n$ .



### 13.1.3 Convergence in mean square

If

$$\lim_{n \rightarrow \infty} \mathbb{E}[(X_n - X)^2] = 0$$

then we say that  $X_n(\omega) \rightarrow X(\omega)$  in mean square. Convergence in mean square implies convergence in probability. This can be shown with the help of Chebyshev inequality applied to the random variable  $X_n - X$ .

**Chebyshev inequality:** For any real random variable  $Z$  and for any constant  $a > 0$ , we have

$$P\{|Z| > a\} \leq \frac{\mathbb{E}[Z^2]}{a^2}. \quad (13.3)$$

The proof of the inequality (13.3) is straightforward, i.e.

$$\mathbb{E}[Z^2] = \int_{-\infty}^{\infty} dz p(z) z^2 \geq \int_{|z|>a} dz p(z) z^2 \geq a^2 \int_{|z|>a} dz p(z)$$

where the second inequality derives from the fact that  $z^2 \geq a^2$  for all  $z$  in the domain of integration.<sup>4</sup>

Using Chebyshev inequality for the random variable  $Z = X_n - X$ , with  $a = \epsilon$ , we have that if  $X_n \rightarrow X$  in mean square, then

$$P(|X_n - X| > \epsilon) \leq \frac{1}{\epsilon^2} \mathbb{E}[(X_n - X)^2] \rightarrow 0$$

as  $n \rightarrow \infty$ , i.e.  $X_n \rightarrow X$  in probability.

### 13.1.4 Convergence in distribution

If for all continuous and bounded functions  $f(x)$

$$\lim_{n \rightarrow \infty} E[f(X_n)] = E[f(X)]$$

then  $X_n(\omega) \rightarrow X(\omega)$  in distribution. Note that this is equivalent to

$$\lim_{n \rightarrow \infty} \int dx [p_n(x) - p(x)] f(x) = 0, \quad \forall f(x).$$

---

<sup>4</sup>We note in passing that the same proof works if the exponent 2 is replaced by  $p > 1$ , and it leads to

$$P\{|Z| > a\} \leq \frac{\mathbb{E}[|Z|^p]}{a^p}, \quad p > 1.$$

This implies that the distribution of  $X_n$  converges to that of  $X$  on all points  $x$  except at most a set of zero measure. Mean square convergence and convergence in probability imply convergence in distribution. When  $X_n$  converges in distribution to a constant  $c$ , then  $X_n \rightarrow c$  also in probability. We omit the proofs of these statements.

## 13.2 Borel-Cantelli lemmas

The Borel-Cantelli lemma is so simple and general that is worth being remembered.

**Borel-Cantelli Lemma.** *Let  $A_1, A_2, \dots$  be an infinite sequence of events. If*

$$\sum_{j=1}^{\infty} P(A_j) < +\infty \quad (13.4)$$

*then, almost surely, at most a finite number of events occur.*

This result is often stated by saying that, if Eq. (13.4) holds, then the probability that events  $A_n$ ,  $n = 1, 2, 3, \dots$  occur infinitely often is zero, i.e.  $P\{A_n \text{ i.o.}\} = 0$ .

The proof of the Borel-Cantelli lemma is simple: if  $A_n$  occurs infinitely often, then for any fixed  $N > 0$ , there must be at least one event  $A_n$  with  $n \geq N$  that occurs. This means that,  $\forall N > 0$

$$\{A_n \text{ i.o.}\} \subseteq \bigcup_{j=N}^{\infty} A_j.$$

But then, sub-additivity of probability implies

$$P\{A_n \text{ i.o.}\} \leq P\left(\bigcup_{j=N}^{\infty} A_j\right) \leq \sum_{j=N}^{\infty} P(A_j). \quad (13.5)$$

The latter expression can be made as small as one wishes, by taking  $N$  large enough. Indeed if the series in Eq. (13.4) converges, then the partial sum in the right hand side of Eq. (13.5) vanishes as  $N \rightarrow \infty$ .

The Borel-Cantelli lemma is a very general result. Notice that no assumption on the events (e.g. on their independence) is needed.

As an application, consider again returns to the origin of a biased random walk. Let  $S_n = \sum_{i=1}^n X_i$  with  $X_i$  being i.i.d. binary random variables with  $P(X_i = +1) = p = 1 - P(X_i = -1)$ . Then

$$P\{A_n\} = P\{S_{2n} = 0\} = \binom{2n}{n} [p(1-p)]^n \simeq \frac{1}{\sqrt{n}} e^{-a(p)n},$$

with  $a(p) = -\log[4p(1-p)]$ . For all  $p \neq 1/2$ ,  $a(p) > 0$  and the condition Eq. (13.4) of the Borel-Cantelli lemma applies. This means that, almost surely, a biased random walker returns to the origin only a finite number of time. In order to deal with the case  $p = 1/2$  we need a converse of the Borel-Cantelli lemma.

For  $p = 1/2$ , we can apply the Borel-Cantelli lemma to the  $d$  dimensional random walk. This is defined by  $d$  independent random walks

$$S_n^{(a)} = \sum_{k=1}^n X_k^{(a)}, \quad a = 1, \dots, d$$

with  $X_k^{(a)}$  i.i.d. with distribution  $P\{X_k^{(a)} = \pm 1\} = \frac{1}{2}$ . Then  $(S_n^{(1)}, \dots, S_n^{(d)})$  is a point on a  $d$  dimensional hyper-cubic lattice. Consider the return to the origin in  $2n$  steps

$$A_n^{(d)} = \{S_{2n}^{(1)} = 0, \dots, S_{2n}^{(d)} = 0\}$$

Then

$$P(A_n^{(d)}) \sim n^{-d/2}.$$

For  $d > 2$  the series in Eq. (13.4) converges and therefore the random walk returns to the origin at most a finite number of times, i.e. it is transient. For  $d \leq 2$  the random walk is recurrent, i.e.  $A_n$  occurs i.o., but that's harder to show.

It is clear that the converse of the Borel-Cantelli lemma does not hold, unless we add more hypotheses. Take for example

$$A_n = \{X \in (0, 1/n)\}$$

where  $X$  is a uniform random variable in  $(0, 1]$ . Then  $P(A_n) = 1/n$  and the series in Eq. (13.4) diverges. Yet for any value of  $X \in (0, 1]$ , the event  $A_n$  occurs only for  $n < 1/X$ , so  $P\{A_n \text{ i.o.}\} = 0$ . The problem with this example is that the events  $A_n$  are strongly dependent (indeed  $A_n$  implies all  $A_m$  for  $m < n$ ).

**The converse of Borel-Cantelli's Lemma.** *Let  $A_1, A_2, \dots$  be an infinite sequence of events. If  $A_n$  are independent and*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad (13.6)$$

*then an infinite number of events occurs almost surely, i.e.*

$$P(A_n \text{ i.o.}) = 1$$

Proof: if  $A_n$  does not occur i.o., then there is a maximal  $n$  such that none of the events  $A_k$  occur for  $k \geq n$ . Therefore, the complement of the event  $\{A_n \text{ i.o.}\}$  can be written as

$$\overline{\{A_n \text{ i.o.}\}} = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \bar{A}_k.$$

The probability that  $A_n$  does not occur i.o. can be written as

$$1 - P(A_n \text{ i.o.}) = P\left(\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \bar{A}_n\right) \leq \sum_{N=1}^{\infty} \prod_{n=N}^{\infty} P(\bar{A}_n) \quad (13.7)$$

$$= \sum_{N=1}^{\infty} \prod_{n=N}^{\infty} [1 - P(A_n)] \quad (13.8)$$

$$\leq \sum_{N=1}^{\infty} \exp\left\{-\sum_{n=N}^{\infty} P(A_n)\right\} = 0 \quad (13.9)$$

where the first inequality arises from the sub-additivity of the probability and the independence of events  $A_n$ . The second from the fact that  $1 - x \leq e^{-x}$ , and the last equality from the fact that, for a divergent series, every partial sum diverges, i.e.  $\sum_{n=N}^{\infty} P(A_n) = \infty$ . Hence every term in the sum is zero.

For example, take a sequence  $\vec{x} = (x_1, \dots, x_m)$  of  $m$  binary digits ( $x_i = 0$  or  $1$ ) and an infinite sequence of Bernoulli trials  $X_1, \dots, X_n, \dots$  with  $p = 1/2$ . Let

$$A_n = \{X_{(n-1)m+1} = x_1, \dots, X_{nm} = x_m\}$$

be the event that the Bernoulli sequence reproduces  $\vec{x}$  exactly at positions  $(n-1)m+1, \dots, nm$  (so  $A_1$  is the event that the first  $m$  values of  $X_i$  coincide with  $x_i$ ). The events  $A_n$  are independent and  $P(A_n) = 2^{-m}$  is independent of  $n$ . Hence the series in Eq. (13.6) diverges. This means that the sequence  $X_n$  contains almost surely the sequence  $\vec{x}$  an infinite number of times.<sup>5</sup>

<sup>5</sup>The sequence  $X_n$  can be generated by flipping a coin repeatedly and the sequence  $\vec{x}$  can be the binary representation of Hamlet. You don't need to be Shakespeare to produce Hamlet, you only need to be patient enough...

The assumption of independence can be relaxed to pairwise independence [21]. The proof of the converse of the Borel-Cantelli lemma for pairwise independent events is based on using the Chebyshev inequality for the random variable  $S_m - \mathbb{E}[S_m]$  where  $S_m$  is the number of events  $A_n$  that occur for  $n \leq m$ . Then, using  $a = \epsilon \mathbb{E}[S_m]$ , we have

$$P\left\{\left|\frac{S_m}{\mathbb{E}[S_m]} - 1\right| > \epsilon\right\} \leq \frac{\mathbb{V}[S_m]}{\epsilon^2 \mathbb{E}[S_m]^2} \quad (13.10)$$

where mean and variance of  $S_m$  are given by

$$\mathbb{E}[S_m] = \sum_{n=1}^m P(A_n), \quad \mathbb{V}[S_m] = \sum_{n=1}^m P(A_n)[1 - P(A_n)]$$

because of pairwise independence  $P(A_n \cap A_{n'}) = P(A_n)P(A_{n'})$  if  $n \neq n'$ . The last equation also implies that  $\mathbb{V}[S_m] \leq \mathbb{E}[S_m]$  that can be used to transform Eq. (13.10) into

$$P\left\{\left|\frac{S_m}{\mathbb{E}[S_m]} - 1\right| > \epsilon\right\} \leq \frac{1}{\epsilon^2 \mathbb{E}[S_m]}$$

Since  $\mathbb{E}[S_m] \rightarrow \infty$  as  $m \rightarrow \infty$ , this shows that the ratio of the number of events  $A_n$  that occur up to  $m$  to its expected value converges to one as  $m \rightarrow \infty$

$$\frac{S_m}{\mathbb{E}[S_m]} \rightarrow 1 \quad (13.11)$$

in probability. In order to prove the converse of the Borel-Cantelli lemma this result should be turned into almost sure convergence. Because then the number of events  $A_n$  that occurs (up to  $m$ ) diverges a.s. like  $\mathbb{E}[S_m]$  as  $m \rightarrow \infty$ , i.e.  $\{A_n \text{ i.o.}\}$ . The trick to do this, is to consider subsequences  $m_k$  such that  $\mathbb{E}[S_{m_k}] \geq k^2$ , so that the event

$$B_k = \left\{\left|\frac{S_{m_k}}{\mathbb{E}[S_{m_k}]} - 1\right| > \epsilon\right\}$$

satisfies the condition Eq. (13.4) of the Borel-Cantelli lemma. Therefore  $P\{B_k \text{ i.o.}\} = 0$ , which means that the limit (13.11) holds a.s. on subsequences  $m_k$ . The last step requires to show that this must also hold on the whole sequence. This is intuitive since  $S_m$  is an increasing function of  $m$ : if a subsequence  $S_{m_k}$  diverges,  $S_m$  has to diverge too.

**Exercise 13.3**

In a sequence  $X_1, X_2, \dots, X_n, \dots$  of i.i.d. random variables drawn from a continuous distribution with pdf  $p(x)$ , a record  $A_n = \{X_n > X_i \forall i < n\}$  is the event that  $X_n$  is larger than all the previous values  $X_i$ , for  $i < n$ . Can Eq. (13.11) be used to estimate the number  $S_m$  of records that occur before  $m$ , asymptotically for  $m \rightarrow \infty$ ?

Since the proof requires only to control the second moment of  $S_m$ , the condition of pairwise independence in the inverse of Borel-Cantelli lemma can be relaxed further by replacing it with the milder condition

$$\lim_{n \rightarrow \infty} \frac{\sum_{i \neq j=1}^N P(A_i \cap A_j)}{[\sum_{i=1}^N P(A_i)]^2} = 1.$$

## Chapter 14

# Laws of large numbers and the Asymptotic Equipartition property

It is common practice, when measuring a physical quantity to run several independent experiments and then compute the average of the outcomes in each of them. Each experiment may be affected by uncontrolled factors that impact on the measurement introducing “errors” that are sometimes positive, sometimes negative. When we take the arithmetic mean these errors *average out*.<sup>1</sup> Although we give it for granted, this is a remarkable fact, because if it were not for this, quantitative science could not be possible. This fact has its theoretical roots in the *law of large numbers* (LLN). The LLN states that, given a sequence  $X_1, X_2, \dots, X_n, \dots$  of i.i.d. random variables with a finite expected value  $\mu = \mathbb{E}[X_i]$ , the (arithmetic) mean converges to the expected value

$$\frac{1}{n} \sum_{i=1}^n X_i(\omega) \rightarrow \mu = \mathbb{E}[X_i] , \quad (14.1)$$

when  $n \rightarrow \infty$ . There are different ways in which the limit could be interpreted, but before coming to that, let us make a few remarks:

- the quantity on the left of the limit in (14.1) is a random variable, whereas the limit  $\mu$  is not. This type of results often go under the name of concentration properties, referring to the fact that the distribution of the mean concentrates on a single point.

---

<sup>1</sup>If they don't we talk about *systematic errors*, i.e. of effects that persist in all the experiments.

- in a physical system such as a gas, physical quantities such as the energy, are the sum over an astronomically large number ( $n \sim 10^{23}$ ) of particles. *Intensive* quantities (e.g. the energy density) are related to averages over this many variables. Macroscopic physical systems correspond to situations in which the limit is realised in practice.<sup>2</sup> If it were not for the law of large numbers, the specific heat of a disordered materials (such as a piece of concrete or a glass) would depend on the specific spatial arrangement of all atoms. Instead, the energy is the sum of many local contributions which vary from point to point because of impurities, but these variations “average out”. The quantities which satisfy laws of large numbers in physics are called *self-averaging*.
- The same argument should apply to the per capita Gross Domestic Product of a country, which is the sum of the contributions to economic activity of all its citizens. For countries such as India or China ( $n \sim 10^7$ ) we should expect that the per-capita GDP does not fluctuate. Yet apparently [22] this is not true. Macro-economic fluctuations are much larger than what the LLN would allow. Why?
- The law of large numbers is used when we want to estimate expected values of random variables

$$\mathbb{E}[f(X)] = \sum_{\omega \in \Omega} p_{\omega} f(X(\omega)). \quad (14.2)$$

The way we do it is to take  $T$  samples  $\omega_t$ ,  $t = 1, \dots, T$ , that in the best of the possible worlds can be thought of as independent draws from the distribution  $p_{\omega}$ . Then we compute the mean and argue that

$$\frac{1}{T} \sum_{t=1}^T f[X(\omega_t)] = \sum_{\omega \in \Omega} \frac{n_{\omega}}{T} f[X(\omega_t)] \approx \mathbb{E}[f(X)] , \quad (14.3)$$

where  $n_{\omega}$  is the number of times that outcome  $\omega$  occurs in the sample. When  $T \gg |\Omega|$  is much larger than the number of possible outcomes  $\omega$ , then  $\frac{n_{\omega}}{T}$  provides a good approximation of  $p_{\omega}$  (as we shall see), and

---

<sup>2</sup>In a physical system,  $X_i$  can be one coordinate of particle  $i$ , or it's magnetic moment. Generally  $X_i$  cannot be considered as independent random variables because particles interact. Yet these interactions are short ranged. This means that each  $X_i$  depends on a number of other variables  $X_j$  which is finite. These are generally called systems of weakly dependent random variables and they obey the LLN if the interaction is weak enough. Although *statistical dependencies* introduced by interactions are negligible, they play a key role in allowing the system to *equilibrate*, i.e. to converge to an equilibrium state. We'll come back to this point.



Eq. (14.3) is not so surprising. Yet Eq. (14.3) works remarkably well also when  $T \ll |\Omega|$ . For example, Eq. (14.3) is routinely used in statistical physics to estimate averages. There  $\omega$  is a point in phase space that specifies the coordinates of each particle.<sup>3</sup> So the size of  $\Omega$  can be astronomically large (typically exponentially large in the number of particles). In these conditions, the number  $n_\omega$  of times that a particular configuration  $\omega$  is visited is zero most of the time and sometimes one. The ratio  $n_\omega/T$  does not provide a good approximation of  $p_\omega$ . How can a handful of measures  $\omega_1, \dots, \omega_T$  allow us to compute expected values? If Eq. (14.2) is true, something quite peculiar must happen.

As we shall see, there are particular features of samples of many independent random variables that are not very random, in the sense that their probability distribution *concentrates* on a small neighbourhood of a single point in the space of distributions.

## 14.1 The weak law of large numbers (WLLN)

A sequence  $X_1, \dots, X_n, \dots$  of independent and identically distributed (i.i.d.) random variables with  $\mathbb{E}[X_i] = \mu$  satisfies the WLLN if the limit (14.1) holds in probability. This means that, for all  $\epsilon > 0$ , we have

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right\} = 0.$$

Khinchin has shown that a finite expected value  $\mathbb{E}[X_i] = \mu$  is a sufficient condition for the WLLN to hold (see GNEDENKO). Here we limit ourselves to a much simpler proof based on Chebyshev inequality,<sup>4</sup> that assumes that the variance  $\mathbb{V}[X_i] = \mathbb{E}[(x - \mu)^2] = \sigma^2$  is finite. To prove the WLLN, we apply Chebyshev inequality to the variable

$$Z = \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]$$

and observe that for i.i.d. random variables,  $\mathbb{E}[Z^2]$  in Eq. (13.3) reads

$$\mathbb{V} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \frac{\mathbb{V}[X_i]}{n},$$

<sup>3</sup>Eq. (14.3) holds if the *ergodic hypothesis* — that states that ensemble averages are equivalent to time averages — is true.

<sup>4</sup>We shall see an argument that shows that  $|\mathbb{E}[X]| < +\infty$  is a sufficient condition for the WLLN when we discuss limit theorems for sums.

because  $\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = 0$  for  $i \neq j$ . Therefore Eq. (13.3) implies that

$$P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq \epsilon\right\} \leq \frac{\mathbb{V}[X_i]}{n\epsilon^2}$$

which converges to zero as  $n \rightarrow \infty$ .

The same proof shows that the WLLN holds whenever the variables  $X_i$  are uncorrelated, or when the correlation  $\mathbb{E}[(X_i - \mu)(X_j - \mu)]$  is small enough. Indeed

$$V\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{V[X_i]}{n} + \frac{1}{n^2} \sum_{i \neq j} E[(X_i - \mu)(X_j - \mu)]$$

and the law of large numbers holds when the last term vanishes as  $n \rightarrow \infty$ .

The WLLN states that *the probability* of excursions larger than  $\epsilon$  of the mean away from the expected value, converges to zero for large  $n$ . It is a statement about the limit of the probability of excursions away from the mean, it is not a statement about the probability that the sequence of means converges to the expected value.

## 14.2 The strong law of large numbers (SLLN)

The strong law of large numbers (SLLN) states the almost certain convergence of the mean to the expected value, whereas the WLLN states the convergence in probability. The SLLN says that for almost all  $\omega \in \Omega$  the mean converges to the expected value, i.e. that *the probability that the limit* of the mean is the expected value is one. For a given  $\omega$ , the mean converges to the expected value if,  $\forall \epsilon > 0$  there exist a  $\nu(\epsilon, \omega)$  such that

$$\left|\frac{1}{n} \sum_i X_i - \mu\right| < \epsilon,$$

for all  $n \geq \nu(\epsilon, \omega)$ . Saying that this holds almost surely, is equivalent to saying that one can make the probability of points  $\omega$  for which the above condition holds for all  $n$  large enough, as close as desired to one. In other words, a sequence  $X_1, \dots, X_n, \dots$  of independent random variables with  $\mathbb{E}[X_i] = \mu$  satisfies the SLLN if, for all  $\epsilon, \delta > 0$  there is an  $N(\epsilon, \delta)$  such that

$$P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| < \epsilon, \forall n > N(\epsilon, \delta)\right\} \geq 1 - \delta.$$

The SLLN can be also stated as follows: for any  $\epsilon > 0$ , define the events

$$A_n = \left\{ \left| \frac{1}{n} \sum_i X_i - \mu \right| > \epsilon \right\}.$$

Then the SLLN is equivalent to saying that a.s. at most a finite number of events  $A_n$  occur,<sup>5</sup> i.e.  $P\{A_n \text{ i.o.}\} = 0$ . By contrast, the WLLN states that  $P(A_n) \rightarrow 0$  as  $n \rightarrow \infty$ . Kolmogorov has shown that the existence of the expected value, i.e.  $E[X] = \mu$  is a sufficient condition for the SLLN (see GNEDENKO).

For example, we can prove the SLLN for the convergence of the frequency to the probability in Bernoulli trials:

$$X_i = \begin{cases} 1 & \text{w. p. } p \\ 0 & \text{w. p. } 1 - p, \end{cases} \quad S_n = \sum_{i=1}^n X_i.$$

The SLLN, in this case, is equivalent to saying that for all  $\epsilon > 0$ , the probability that the event

$$A_n = \left\{ \left| \frac{S_n}{n} - p \right| > \epsilon \right\}$$

occurs infinitely often, is zero. De Moivre - Laplace approximation of the binomial distribution, states that  $S_n/n - p$  is asymptotically well approximated by a Gaussian variable with mean zero and variance  $p(1 - p)/n$ . Therefore<sup>6</sup>

$$P(A_n) \cong \sqrt{\frac{2}{\pi}} \int_{\sqrt{\frac{n}{p(1-p)}} \epsilon}^{\infty} dx e^{-x^2/2} < \sqrt{\frac{2}{\pi}} e^{-\frac{n}{2p(1-p)} \epsilon^2}$$

and the Borel-Cantelli lemma ensures us that the SLLN holds, i.e. that  $P\{A_n \text{ i.o.}\} = 0$ , because  $\sum_{n>0} P(A_n) < +\infty$ .

---

<sup>5</sup>Remember that for a fixed  $\omega$ , convergence implies the existence of an integer  $\nu(\epsilon, \omega)$  such that none of the events  $A_n$  occur for  $n > \nu(\epsilon, \omega)$ , i.e. that  $\omega \notin A_n$  for all  $n > \nu(\epsilon, \omega)$ . The threshold  $\nu(\epsilon, \omega)$  is different for each  $\omega$  for which the mean converges to  $\mu$ . Yet the fact that the number of excursions larger than  $\epsilon$  of the mean away from  $\mu$  is finite holds for all these  $\omega$ 's. This is why the SLLN is equivalent to the statement  $P\{A_n \text{ i.o.}\} = 0$ .

<sup>6</sup>This is because, for all  $z \geq 1$ , the inequality

$$\int_z^{\infty} dx e^{-x^2/2} \leq \frac{1}{z} e^{-z^2/2} \leq e^{-z^2/2}$$

holds, as you can show as an Exercise.

**Exercise 14.1**

It is possible to prove something stronger in the same manner. Take

$$B_n = \left\{ \left| \frac{S_n - np}{\sqrt{np(1-p)}} \right| > \sqrt{2a \log n} \right\},$$

and show that for  $a > 1$  the SLLN holds.

### 14.3 Typical samples and the Asymptotic Equipartition Property

As a particular application of the law of large numbers, consider the probability of a sequence  $\underline{X} = (X_1, \dots, X_n)$  of i.i.d. random variables. Let us first consider the case where  $X_i$  are drawn from a discrete distribution  $p(x)$ . In other words, we assume that the variables  $X_i \in \chi$  take a finite number of values ( $|\chi| < +\infty$ ) and that  $p(x) > 0$  or all  $x \in \chi$ . Then the probability of a sequence is

$$p(\underline{X}) = \prod_{i=1}^n p(X_i).$$

Taking the logarithm and dividing by  $n$ , we have

$$\frac{1}{n} \log p(\underline{X}) = \frac{1}{n} \sum_{i=1}^n \log p(X_i).$$

The variables  $\log p(X_i)$  are themselves random variables and their variance is finite. Therefore they satisfy the law of large numbers, which means that, for  $n \rightarrow \infty$ , the mean converges to the average

$$\mathbb{E} [\log p(X)] = \sum_{x \in \chi} p(x) \log p(x) \equiv -H[X].$$

So we have that (Theorem 3.1.1 in COVER)

*Let  $\underline{X} = (X_1, \dots, X_n)$  be independent draws from a discrete distribution  $p(x)$  ( $X_i \in \chi$  with  $|\chi| < +\infty$ ). Then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log p(\underline{X}) = -H[X] \quad (14.4)$$

*in probability.*

In brief, this theorem states that, for  $N$  large, all sequences of random variables that are independently drawn from  $p(x)$  have essentially<sup>7</sup> the same probability,

$$p(\underline{X}) \sim e^{-nH[X]}.$$

Sequences with this probability are called *typical* sequences. Non-typical sequences occur with a probability which is exponentially small (in  $n$ ) with respect to typical ones. This does not only include those sequences with a probability  $p(\underline{X})$  which is much smaller than  $e^{-nH[X]}$ , but also sequences whose probability is higher than that of typical sequences. For example, if  $p(x_0) > p(x)$  for all  $x \neq x_0 \in \mathcal{X}$ , then the sequence  $X_i = x_0$  for all  $i = 1, \dots, n$  has a probability  $p(x_0)^n$  that is exponentially larger than the probability  $e^{-nH[X]}$  of typical sequences. Yet it is very unlikely to see this sequence because typical sequences are much more in number (see below).

The fact that typical sequences have the same probability is called *Asymptotic Equipartition Property* (AEP).<sup>8</sup> In brief, the AEP states that: For any  $\epsilon > 0$ , one can define the set of  $\epsilon$ -typical sequences as

$$A_n^\epsilon = \left\{ \underline{X} : \left| \frac{1}{n} \log p(\underline{X}) + H[X] \right| < \epsilon \right\}$$

Then an equivalent way to state the AEP is that

1. By definition, all  $\epsilon$ -typical sequences are equally likely:  $P(\underline{X}) \sim e^{-nH[X]}$  for all  $\underline{X} \in A_n^\epsilon$
2. As a consequence of the law of large numbers, a random sequence is almost surely an  $\epsilon$ -typical sequence

$$P\{A_n^\epsilon\} > 1 - \epsilon.$$

3. As a consequence, the number of  $\epsilon$ -typical sequences is

$$|A_n^\epsilon| \sim e^{nH[X]}.$$

The last statement comes from the fact that

$$1 \approx P\{A_n^\epsilon\} = \sum_{\underline{X} \in A_n^\epsilon} P(\underline{X}) \sim e^{-nH[X]} |A_n^\epsilon|$$

where we used the fact that all  $\underline{X} \in A_n^\epsilon$  have the same probability, by definition.

<sup>7</sup>Up to the leading exponential behaviour. Here and below we shall use the symbol  $a_n \sim e^{\alpha n}$  to denote asymptotic equality of  $\frac{1}{n} \log a_n$  and  $\alpha$ .

<sup>8</sup>A more detailed treatment of this issue is given in COVER, chapter 3, which is a suggested reading.

The functional<sup>9</sup>

$$H[X] = - \sum_{x \in \chi} p(x) \log p(x) = \mathcal{H}[p] \quad (14.5)$$

is called the *entropy* of the random variable  $X$ . The last equality above emphasises that the entropy is a function of the probability distribution  $p(x)$  of  $X$ . We shall discuss the entropy in more detail in what follows. For the moment, let it suffice to say that it takes values in the interval  $0 \leq H[X] \leq \log |\chi|$ . The lower limit is achieved when  $X = x_0$  is a constant, and  $p(x) = 1$  if  $x = x_0$  and  $p(x) = 0$  for all  $x \neq x_0$ . In this limit, there is only one possible sequence  $\underline{X}$ , and  $H[X] = 0$ . The upper limit is achieved when  $p(x) = 1/|\chi|$  is a uniform distribution.<sup>10</sup>

Therefore, the number of typical samples  $|A_n^\epsilon| \sim e^{nH[X]}$  is much smaller than the number of all possible samples, which is  $|\chi|^n = e^{n \log |\chi|}$ , whenever the distribution differs from the uniform one  $p(x) = 1/|\chi|$ , for which one has  $H[X] = \log |\chi|$ . To put it differently, the probability that any sequence  $\underline{X}$  is typical is exponentially small in  $n$ .

In order to shed more light on the nature of typical samples, consider a sample  $\underline{X}$  of  $n$  observations of a random variable  $X \in \chi$ , drawn independently from a distribution  $p(x) = P\{X_i = x\}$ . Let  $m_x$  be the count of the points in the sample for which  $X_i = x$  (so  $\sum_x m_x = n$ ). Let us consider the case  $n \gg |\chi|$ , so that  $m_x \gg 1$  for all  $x \in \chi$ .  $m_x$  is related to the empirical distribution

$$\hat{p}_{\underline{X}}(x) = \frac{m_x}{n} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i, x}. \quad (14.6)$$

of the sample  $\underline{X}$ , which is also called the *type* of  $\underline{X}$ . The probability of the sample  $\underline{X}$  can be written as

$$P(\underline{X}) = \prod_{i=1}^n p(X_i) = \prod_{x \in \chi} p(x)^{m_x} = e^{n \sum_x \hat{p}_{\underline{X}}(x) \log p(x)}. \quad (14.7)$$

This shows that  $P(\underline{X})$  depends on  $\underline{X}$  only through  $m_x$  or through its type  $\hat{p}_{\underline{X}}$ . Therefore, the probability to observe a given vector of counts

<sup>9</sup>A functional is a function of a function, which maps the space of functions to the real axis.  $H[X]$  is a functional, because a random variable  $X(\omega)$  is a function. We use square brackets for functionals and parentheses (...) for functions.

<sup>10</sup>This is easily seen by solving the maximisation problem

$$\max_p \mathcal{H}[p],$$

subject to the normalisation constraint  $\sum_{x \in \chi} p(x) = 1$ .

$\underline{m} = \{m_x, x \in \chi\}$  is

$$P\{\underline{m}\} = \frac{n!}{\prod_{x \in \chi} m_x!} \prod_{x \in \chi} p(x)^{m_x},$$

where the combinatorial factor counts the number of samples  $\underline{X}$  with the same  $\underline{m}$ . Using Stirling's approximation  $n! \sim n^n e^{-n}$ , we obtain  $P\{\underline{m}\} \sim e^{-d(\underline{m})n}$  where the constant

$$d(\underline{m}) = \sum_{x \in \chi} \frac{m_x}{n} \log \frac{m_x}{np(x)} \quad (14.8)$$

is non-negative and it vanishes only for<sup>11</sup>  $m_x = np(x)$ . Therefore, when  $n \rightarrow \infty$ , the probability that the empirical distribution  $\hat{p}_{\underline{X}}(x)$  does not matches the true distribution  $p(x)$  becomes exponentially small. Put differently, all typical samples have the same empirical distribution (or type), that coincides with  $p(x)$  asymptotically for  $n \rightarrow \infty$  (i.e.  $\hat{p}_{\underline{X}}(x) \approx p(x)$ , for all  $x \in \chi$ ).

Loosely speaking, this explains why we can estimate expected values with empirical averages, as in Eq. (14.1). This is possible because all typical samples share the same empirical distribution of  $f(X)$ , and this approximates very well the true distribution for large  $n$ .

### Exercise 14.2

A sport newspaper gives every Friday the probabilities of the outcomes (1, X or 2) of the 13 football matches in the Italian league that are in the *schedina*, a popular betting scheme in Italy. Take the examples where the probabilities of the three different outcomes are 50%, 30% and 20%

<sup>11</sup>The proof follows from the inequality  $\log \frac{1}{z} \geq 1 - z$  applied to  $d(\underline{m})$  with  $z = np(x)/m_x$ , which gives

$$d(\underline{m}) \geq \sum_{x \in \chi} \frac{m_x}{n} \left[ 1 - \frac{np(x)}{m_x} \right] = 0,$$

because of the normalisation of  $p(x)$  and because  $\sum_{x \in \chi} m_x = n$ . The constant  $d$  in Eq. (14.8) can also be expressed in terms of the type

$$d = \sum_{x \in \chi} \hat{p}_{\underline{X}}(x) \log \frac{\hat{p}_{\underline{X}}(x)}{p(x)} \equiv D_{KL}(\hat{p}_{\underline{X}} \| p)$$

where the functional  $D_{KL}$  is called relative entropy or Kullback-Leibler divergence, and it will be discussed in later chapters. For the time being, let it suffice to say that  $D_{KL}(q \| p) \geq 0$  and  $D_{KL}(q \| p) = 0$  only if  $q = p$ .

in any possible order. So the forecast may look like this:

Games	1	X	2
A vs B	50%	30%	20%
C vs D	30%	50%	20%
E vs F	20%	50%	30%
G vs H	50%	30%	20%
$\vdots$	$\vdots$	$\vdots$	
U vs V	30%	50%	20%
W vs X	30%	50%	20%
Y vs Z	50%	20%	30%

The simplest schedina consists of a sequence  $(\omega_1, \dots, \omega_{13})$  of forecasts, one for each game, with  $\omega_i \in \{1, X, 2\}$ .

How would you know whether a particular schedina is a typical one? Do you expect people will play a typical schedina?

The same result holds also for continuous variables, and it goes under the name of

**Glivenko-Cantelli theorem.** *Let  $X$  be a real random variable with distribution  $F(x) = P\{X \leq x\}$  and  $X_1, \dots, X_n$  be a sample of  $n$  i.i.d. draws from  $F(x)$  then the fraction*

$$F_n(x) = \frac{1}{n} |\{i \in [1, n] : X_i \leq x\}|$$

*of  $X_i$  that is smaller than  $x$  converges a.s. to  $F(x)$  for  $n \rightarrow \infty$ . More precisely*

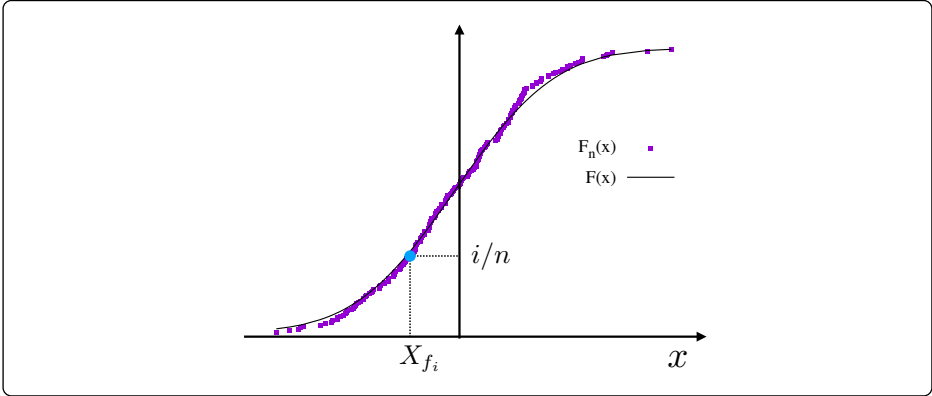
$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0 \quad \text{a.s.}$$

The proof is a consequence of the SLLN. So, here is a simple recipe to estimate the distribution  $F(x)$  from a sample  $X_1, \dots, X_n$ : i) sort the sample in ascending order,  $X_{f_1} < X_{f_2} < \dots, X_{f_n}$ , where  $\{f_i\}$  is a permutation of the integers up to  $n$ , ii) plot  $i/n$  versus  $X_{f_i}$ . Since  $i/n = F_n(X_{f_i})$ , this produces a plot that for large  $n$  approximates the distribution  $F(x)$ , by the Glivenko-Cantelli theorem, as shown in Figure 31.

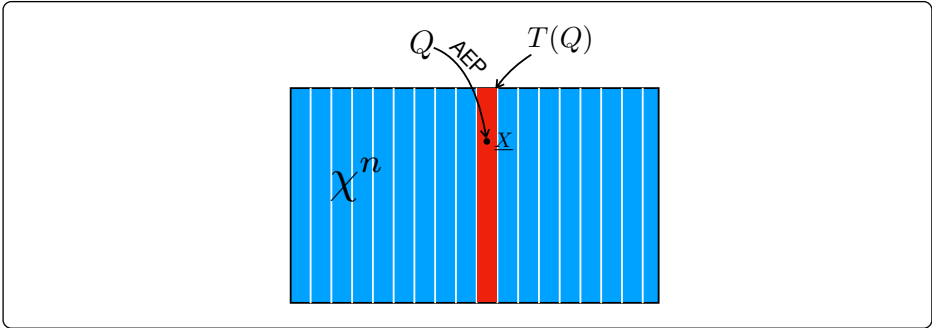
It is worth to spend few more words on types (see Eq. (14.6)).<sup>12</sup> Types provide an alternative description for problems that involve sequences  $\vec{X} = \{X_1, \dots, X_n\}$  of  $n$  i.i.d. random variables  $X_i \in \chi$  where  $\chi$  is a finite set. As we have seen, the probability of a sequence  $\vec{X}$  is a function of its type  $\hat{p}_{\vec{X}}$ . The set

<sup>12</sup>See COVER 11.1.





**Figure 31.** The empirical distribution of a sample of  $n = 200$  draws from a Gaussian distribution (full line).



**Figure 32.** The space  $\chi^n$  of sequences  $\vec{X}$  of  $n$  variables  $X_i \in \chi$  is divided into subsets  $T(P)$  of sequences with the same type  $\hat{p}_{\vec{X}} = P$ . By the AEP, sequences of  $n$  independent draws from a distribution  $Q$  falls with very high probability in the type classes  $T(P)$  with  $P \approx Q$ .

of all sequences with type  $\hat{p}_{\vec{X}} = P$  is called the *type class*

$$T(P) = \{\vec{X} : \hat{p}_{\vec{X}} = P\}$$

Type classes partition the space of all sequences into disjoint subsets. The number of sequences with a given type  $P$  is computed using Stirling's approximation of the multinomial distribution, as we did above, and it is given by  $|T(P)| \sim e^{nH[P]}$ . The AEP can be rephrased saying that sequences of  $n$  independent draws from a distribution  $Q$  falls with very high probability in type classes  $T(Q)$  (see Figure 32).

The AEP is the result of a trade-off between the probability of sequences and the number of sequences with that probability. There are sequences

$\underline{X}$  which are much more probable than typical ones,<sup>13</sup> yet they are too few. Likewise there are type classes which are way more numerous than the typical one, yet their sequences are too unlikely.

This trade-off is the same as the one between energy and entropy in physics. In order to make this point more clear, let us consider the simplest case of sequences  $\underline{X} = (X_1, \dots, X_n)$  of i.i.d. binary variables  $X_i = 0, 1$  with  $P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$ . Without loss of generality, let us take  $p > 1/2$ . The probability of  $\underline{X}$  is  $P(\underline{X}) = p^{n\hat{p}(\underline{X})}(1-p)^{n[1-\hat{p}(\underline{X})]}$  with  $\hat{p}(\underline{X}) = \frac{1}{n} \sum_i X_i$  being the fraction of 1's in the sequence  $\underline{X}$ . Let us introduce the variable, that we shall call<sup>14</sup> *energy*,

$$\epsilon(\underline{X}) = -\frac{1}{n} \log P(\underline{X}) = \hat{p}(\underline{X})\epsilon_0 + [1 - \hat{p}(\underline{X})]\epsilon_1 \quad (14.9)$$

with  $\epsilon_0 = -\log p$  and<sup>15</sup>  $\epsilon_1 = -\log(1-p)$ . Clearly  $\epsilon(\underline{X}) \in [\epsilon_0, \epsilon_1]$  ( $p > 1/2$ ). Notice that if  $\hat{p}$  attains a finite limit when  $n \rightarrow \infty$ , so does  $\epsilon$ .

Let us also introduce the function

$$\sigma(\epsilon) = \frac{1}{n} \log \binom{n}{n \frac{\epsilon_1 - \epsilon}{\epsilon_1 - \epsilon_0}}.$$

that we shall call<sup>16</sup> the *entropy*. After a moment of reflection, it can be realised that  $\sigma(\epsilon)$  is the logarithm of the number of sequences with *energy*  $\epsilon$ . Using Stirling's approximation, you can easily check that

$$\sigma(\epsilon) \cong -\frac{\epsilon_1 - \epsilon}{\epsilon_1 - \epsilon_0} \log \frac{\epsilon_1 - \epsilon}{\epsilon_1 - \epsilon_0} - \frac{\epsilon - \epsilon_0}{\epsilon_1 - \epsilon_0} \log \frac{\epsilon - \epsilon_0}{\epsilon_1 - \epsilon_0}$$

attains a finite limit when  $n \rightarrow \infty$ . We can now compute the probability that a random sequence drawn from this distribution has *energy*  $\epsilon$ , which is just the product of the probability  $e^{-n\epsilon}$  of all sequences with that *energy*, times the number  $e^{n\sigma(\epsilon)}$  of sequences with that energy, i.e.

$$p(\epsilon) = e^{n(\sigma(\epsilon) - \epsilon)}.$$

<sup>13</sup>As for example the sequence with

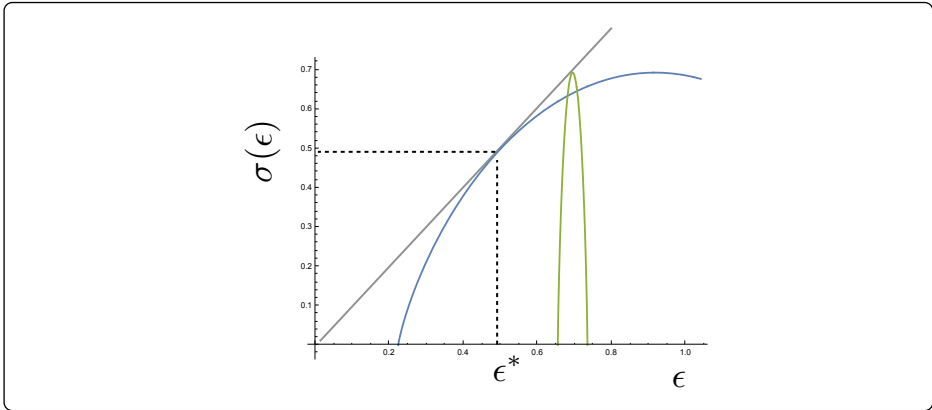
$$X_i = \arg \max_x p(x)$$

for all  $i$ .

<sup>14</sup>Arbitrarily, for the time being.

<sup>15</sup>Note that  $\epsilon_1 = -\log(1 - e^{-\epsilon_0})$ .

<sup>16</sup>Again arbitrarily, for the time being.



**Figure 33.** The AEP for sequences of binary variables. The *entropy* is plotted against the *energy* for  $p = 0.8$  (blue) and  $p = 0.52$  (green).

This distribution, for large  $n$ , is sharply peaked around the maximum of the function  $\sigma(\epsilon) - \epsilon$ , which occurs at the point where  $\sigma'(\epsilon^*) = 1$ , and a straightforward calculation (do it) leads to the result

$$\epsilon^* = -p \log p - (1 - p) \log(1 - p) = H[X].$$

Summarising, all random sequences  $\underline{X}$  drawn from  $P(\underline{X})$  will have the same *energy*  $\epsilon(\underline{X}) \simeq \epsilon^*$ , i.e. the same probability  $P(\underline{X}) \sim e^{-nH[X]}$ . Because of Eq. (14.9), all types of random sequences take the same value  $\hat{p}_{\underline{X}}(x = 1) \simeq p$ . Furthermore, the point  $\epsilon^*$  is also the point where  $\sigma(\epsilon^*) = \epsilon^*$ . This means that the number of sequences with *energy*  $\epsilon^*$  is inversely proportional to their probability, as the AEP states.<sup>17</sup>

Figure 33 displays the interplay between *energy* (straight black line) and *entropy* (concave lines). Sequences with *energy*  $\epsilon < \epsilon^*$  are more probable than typical ones, but they are too few and so  $\sigma(\epsilon) - \epsilon < 0$  and their probability is exponentially small. Sequences with *energy* larger than  $\epsilon^*$  are more numerous, but they are not likely enough so that again  $\sigma(\epsilon) - \epsilon < 0$ .

### 14.3.1 Should we expect the expected value?

A further example to gain intuition on typical sequences is the following: imagine a lottery whose ticket costs 1 euro, and yields a reward of 2 euros with probability  $1/2$  and of  $q \in (0, 1/2)$  euros otherwise. If the invested capital is

<sup>17</sup>Note that the condition  $\sigma(\epsilon^*) = \epsilon^*$  is necessary in order to ensure normalisation of  $p(\epsilon)$ , as shown by carrying out the normalisation integral by saddle point.

$W_0$ , after playing the game the capital is expected to be

$$\mathbb{E}[W_1] = \mathbb{E}[X_1] W_0 = (1 + q/2)W_0, \quad X_1 = \begin{cases} 2 & \text{w. p. } 1/2 \\ q & \text{w. p. } 1/2 \end{cases}$$

This looks like a convenient game because  $\mathbb{E}[W_1] > W_0$ , for any  $q > 0$ . Indeed, if the game is repeated  $n$  times, with  $X_i, i = 1, \dots, n$  being i.i.d. random variables as  $X_1$ , and  $W_0$  tickets are bought each time, then the LLN ensures a positive gain per game, which is equal to  $q/2$  times  $W_0$ .

### Exercise 14.3

In a faraway land long ago, girls were valued more than boys. So couples kept having babies until they had a girl. Assume that each newborn was a female with probability  $p$ . What is the expected fraction of females in a randomly chosen family? What was the fraction of females in the population, assuming it is composed of a very large number of families? (adapted from K. Binmore's *Playing for real*, p. 109).

Now the gain is clearly proportional to  $W_0$ , so the more one invests the higher the gain. In particular, the best thing to do seems to invest all the capital accumulated, at each time. In this way, the capital after  $n$  bets will be  $W_n = X_n \cdots X_1 W_0$  and one can “expect” that the capital will increase as

$$\mathbb{E}[W_n] = (1 + q/2)^n W_0$$

i.e. exponentially. Great!

However, it is easy to show that if  $0 < q < 1/2$  then this strategy leads to bankruptcy, i.e.

$$P\{W_n > a\} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

for all  $a > 0$ . Indeed

$$\frac{1}{n} \log(W_n/W_0) = \frac{1}{n} \sum_{i=1}^n \log X_i \rightarrow \mathbb{E}[\log X_i] = \frac{1}{2} \log(2q)$$

almost surely, as  $n \rightarrow \infty$ , by the SLLN. This means that, almost surely,

$$W_n \sim W_0 e^{-cn}, \quad c = \frac{1}{2} |\log(2q)| > 0$$

i.e. the capital will *typically* vanish exponentially. Here typical means with very high probability, which tends to one as  $n \rightarrow \infty$ .

The origin of the discrepancy of this result with the behavior of  $\mathbb{E}[W_n]$  becomes evident if one takes some care in evaluating the expected value

$$\mathbb{E}[W_n] = W_0 \sum_{k=0}^n \binom{n}{k} 2^{-n} [2^k q^{n-k}] \quad (14.10)$$

$$\sim W_0 \int_0^1 dx e^{nf(x)} \sim W_0 e^{nf(x^*)} \quad (14.11)$$

where we compute the expected value by averaging over the number  $k$  of lucky outcomes (the term in [...] is the corresponding gain), and then we use Stirling's formula for the binomial and change the sum on  $k$  into an integral on  $x = k/n$ . The function  $f(x)$  is

$$f(x) = -x \log x - (1-x) \log(1-x) + (1-x) \log(q/2).$$

The integral can be performed with the saddle point method, i.e. by determining the value  $x^*$  for which  $f'(x^*) = 0$ . We find that the fraction of lucky outcomes that dominates  $\mathbb{E}[W_n]$  is

$$x^* = \frac{1}{1+q/2}.$$

Sequences with this frequency of successes occur only with exponentially small probability, but when they occur they yield an exponentially large gain. The calculation of the expected value  $\mathbb{E}[W_n]$  is determined by the interplay between these two exponential behaviours. Yet the asymptotic equipartition property, guarantees that almost surely only *typical* sequence with a frequency  $x = 1/2$  of successes will occur. This yields a gain  $W_n^{\text{typical}} \sim (2q)^n W_0$ , that vanishes as  $n \rightarrow \infty$ , because  $2q < 1$ .

#### Exercise 14.4

It is a bit frustrating that a game which seems profitable leads to bankruptcy. Maybe one should not invest all the capital. Consider the strategy of investing a fraction  $\lambda \in [0, 1]$  of the capital  $W_n$  each time. What is the best value of  $\lambda$ ?

The lesson is that, given a sequence  $X_1, \dots, X_n, \dots$  of i.i.d. random variables, there are combination  $f(X_1, \dots, X_n)$  whose typical value is close to the expected value  $\mathbb{E}[f]$ , in the sense that the distribution (density)  $p(f) = P\{f(X_1, \dots, X_n) = f\}$  is peaked around the expected value  $\mathbb{E}[f]$ . These are called *self-averaging* quantities. There are other combination of variables,

like the product  $W_n$ , for which this is not true. These are not self-averaging quantities. The expectation based on the expected value is correct only when  $f$  is a self-averaging quantity.

#### Exercise 14.5

On a different planet, the hypsbryx civilisation based its science on using the geometric mean, instead of the arithmetic mean, to measure physical quantities from experimental data. For a physical quantity  $X > 0$  how would the measure that the hypsbryx estimate from a series of experiments  $(X_1, \dots, X_n)$  compare with the one that we would measure on the earth, based on the same data (and the arithmetic mean)? Would it be the same, would it be smaller or bigger?

# Chapter 15

## Limit theorems and universality

The law of large numbers states that the arithmetic mean of many i.i.d. random variables  $X_i$  converges to the expected value<sup>1</sup> as the number  $n$  of variables on which the average is taken diverges. When  $n$  is finite but very large, how big are the deviations and how are they distributed? Limit theorems address this question and show that the deviations have a remarkable feature. Their distribution is *universal* in the sense that it is the same for all random variables  $X_i$  whose distribution  $p(x)$  satisfies certain asymptotic conditions for  $x \rightarrow \pm\infty$ .

As we shall see, an universal behaviour also characterises the extremes, i.e. maxima and minima, of many random variables.

### 15.1 Limit theorems for Sums of i.i.d. random variables

Limit theorems for sums of i.i.d. random variables should be treated within a course of its own. Here we give a non-rigorous derivation of the main results. We refer to GNEDENKO, Chapters VII, VII and IX for a detailed treatment.

Let us consider sums

$$S_n = \sum_{i=1}^n X_i$$

of  $n$  i.i.d. random variables  $X_i \in \mathbb{R}$  with a common pdf  $p(x)$ . The problem we want to address is the following:

---

<sup>1</sup>Whenever this is finite.

Find two sequences  $a_n, b_n \in \mathbb{R}$  such that

$$S_n = a_n + b_n Y$$

and the random variable  $Y$  has a non-degenerate distribution density  $p^*(x)$  in the limit  $n \rightarrow \infty$ .

Non-degenerate means non-trivial.  $Y$  should not be a constant  $y_0$ , i.e. its distribution should not be concentrated on a single point. In other words we are looking for constants  $a_n$  and  $b_n$  such that the centered and rescaled random variable

$$Y_n(\omega) = \frac{S_n(\omega) - a_n}{b_n} \quad (15.1)$$

converges in distribution to a proper random variable, which is not a constant  $y_0$ .<sup>2</sup>

Eq. (15.1) explicitly reminds us that the random variable  $Y : \Omega \rightarrow \mathbb{R}$  is a function of the realisation  $\omega \in \Omega$  in the sample space. Hence our objective is to disentangle the dependence of  $S_n(\omega)$  on  $n$  from its stochastic dependence (on  $\omega$ ), in an explicit manner.

### 15.1.1 Relation to the Law of Large Numbers

The law of large numbers implies convergence of  $S_n/n$  to a constant  $\mu = \mathbb{E}[X_i]$ . If this holds, then

- i)  $a_n = \mu n$  should grow linearly in  $n$  and
- ii)  $b_n/n \rightarrow 0$  should vanish as  $n \rightarrow \infty$ .

Limit theorems are a refinement of the Law of Large Numbers, in that they also specify

- the convergence behaviour, i.e. the size  $b_n/n$  of stochastic fluctuations of  $S_n/n$  around  $\mu$ ,
- the detailed shape of the distribution of these fluctuations, given by  $p^*(x)$  and
- what happens when the Law of Large Numbers does not hold.

In order to address these questions we introduce the

---

<sup>2</sup>To gain some intuition about the meaning of  $a_n$  and  $b_n$ , imagine taking two sequences  $\underline{X}^{(1)}$  and  $\underline{X}^{(2)}$  of  $n$  independent draws from  $p(x)$ , and to compute the sums  $S_n^{(1)}$  and  $S_n^{(2)}$ . Then the difference  $S_n^{(1)} - S_n^{(2)} = b_n(Y_n^{(1)} - Y_n^{(2)})$  is proportional to  $b_n$ . Therefore  $a_n$  provides a measure of the size of  $S_n$  whereas  $b_n$  estimates how much  $S_n$  may vary from one realisation of the sequence  $\underline{X}$  to another.



### 15.1.2 Characteristic functions

Let  $X$  be a random variable with pdf  $p(x)$ . The characteristic function (CF) of  $X$  is:

$$\phi(q) \equiv \mathbb{E}[e^{-iqX}] = \int dx p(x) e^{-iqx}$$

It is also useful to introduce the logarithm of the CF

$$\psi(q) = \ln \phi(q).$$

These are the analogue of the generating function and the cumulant generating function for discrete variables. Indeed they satisfy analogous properties:

- i)  $\phi(0) = 1$  and  $\psi(0) = 0$  by normalisation. Also, since  $|e^{-iqx}| = 1$ , we have

$$|\phi(q)| \leq \int dx p(x) |e^{-iqx}| = 1$$

and the real part of  $\psi(q)$  should be non-positive, i.e.  $\text{Re}[\psi(q)] \leq 0$ .

- ii) Power expansion in  $q$ . When  $\phi$  (and  $\psi$ ) are analytic at the origin (which means that all derivatives exist):

$$\phi(q) = \sum_{n=0}^{\infty} \frac{(-iq)^n}{n!} \mathbb{E}[X^n] \quad (15.2)$$

The coefficients of the power expansion of  $\phi$  yield the moments  $\mathbb{E}[X^n]$  of  $X$ . For this reason  $\phi$  is also called the *moment generating function*. Similarly  $\psi(q)$  admits the power expansion

$$\psi(q) = \sum_{n=0}^{\infty} \frac{(-iq)^n}{n!} C_n \quad (15.3)$$

where  $C_n$  is the  $n^{\text{th}}$  order *cumulant* of the distribution  $p(x)$ . These are related to the moments by the same relations that we have discussed for discrete variables

$$C_1 = \mathbb{E}[x], \quad C_2 = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{V}[X], \dots$$

- iii) Translation: the random variable  $X_a = X + a$ , where  $a \in \mathbb{R}$  is a constant, has pdf  $p_a(x) \equiv p(x - a)$ , and its CF is

$$\phi_a(q) = e^{-iqa} \phi(q), \quad \psi_a(q) = \psi(q) - iqa$$

- iv) **Scaling:** for any constant  $b$  the random variable  $X_b = bX$  has pdf  $p_b(x) \equiv p(x/b)/b$ , and its CF is

$$\phi_b(q) = \phi(bq), \quad \psi_b(q) = \psi(bq)$$

- v) **Convolution:** if  $X_1$  and  $X_2$  are *independent* variables with pdf  $p_1(x)$  and  $p_2(x)$  and characteristic function  $\phi_1(q)$  and  $\phi_2(q)$  respectively, then  $X_1 + X_2$  has pdf

$$p_{1+2}(x) = \int dy p_1(y) p_2(x - y)$$

and characteristic function

$$\phi_{1+2}(q) = \phi_1(q)\phi_2(q), \quad \psi_{1+2}(q) = \psi_1(q) + \psi_2(q).$$

For  $S_n = X_1 + \dots + X_n$ , where  $X_i$  are all independent, this generalizes to

$$\phi_{1+\dots+n}(q) = \prod_{i=1}^n \phi_i(q), \quad \psi_{1+\dots+n}(q) = \sum_{i=1}^n \psi_i(q).$$

If  $X_i$  are also identically distributed and  $\phi_i(q) = \phi(q) = e^{\psi(q)}$  for all  $i$ , then

$$\phi_{1+\dots+n}(q) = [\phi(q)]^n \quad \psi_{1+\dots+n}(q) = n\psi(q).$$

- vi) **Lévy's continuity theorem:** a sequence  $X_n$  of random variables converges in distribution to a random variable  $X$  if and only if the sequence  $\phi_n(q) = \mathbb{E}[e^{-iqX_n}]$  of the corresponding characteristic functions converges point-wise to a function  $\phi(q)$  which is continuous at the origin. Then  $\phi$  is the characteristic function of  $X$ .

Notice that a degenerate distribution  $p(x) = \delta(x - x_0)$  corresponds to a CF  $\phi(q) = e^{-iqx_0}$  and to  $\psi(q) = -iqx_0$ .

### 15.1.3 Derivation of the fundamental equation

Let us consider the characteristic function  $\phi_n(q)$  of  $Y_n$  defined in Eq. (15.1). Then:

$$\phi_n(q) = \mathbb{E}[e^{-iq(S_n - a_n)/b_n}] \tag{15.4}$$

$$= e^{iq a_n/b_n} \mathbb{E}[e^{-i(q/b_n)S_n}] \tag{15.5}$$

$$= e^{iq a_n/b_n} \prod_{k=1}^n \mathbb{E}[e^{-i(q/b_n)X_k}] \tag{15.6}$$

$$= e^{iq a_n/b_n} [\phi(q/b_n)]^n \tag{15.7}$$

Here we used properties *iii)* and *v)* for i.i.d. variables. In terms of the function  $\psi$  this means

$$\psi_n(q) = \frac{iq a_n}{b_n} + n\psi(q/b_n) \quad (15.8)$$

which is the starting point of our analysis. We are interested in finding sequences  $a_n, b_n$  such that the limit

$$\lim_{n \rightarrow \infty} \psi_n(q) = \lim_{n \rightarrow \infty} \left[ iq \frac{a_n}{b_n} + n\psi\left(\frac{q}{b_n}\right) \right] = \psi^*(q) \quad (15.9)$$

is non-degenerate. This means that  $\psi^*(q)$  is the logarithm of the CF of a proper random variable, i.e. we should avoid that the limit results in  $\psi^*(q) = -iqy_0$  for some  $y_0$ .

It is clear that  $b_n \rightarrow \infty$  as  $n \rightarrow \infty$ , because  $b_n$  quantifies the size of fluctuations of  $S_n$ , and summing more and more variables we expect the fluctuations to increase. Then the relevant information, as far as the limit in Eq. (15.9) is concerned, is contained in the behaviour of the characteristic function  $\phi(k)$  of  $X_i$  for small  $k = q/b_n$ . This, in turn, is related to the existence of moments of low order. Note that small  $|k|$  means large  $|x|$ , in the sense that the behaviour of  $\psi$  for  $|k| \ll 1$  is related to the behaviour of  $p(x)$  in the tails<sup>3</sup> (i.e. for  $|x| \gg 1$ ).

There are three main cases:

- 1)  $\mu = \mathbb{E}[X]$  finite and  $\sigma^2 = \mathbb{V}[X] < +\infty$ .

Then the leading terms in the expansion of  $\psi$  for  $|k| \ll 1$  are

$$\psi(k) = -i\mu k - \frac{\sigma^2}{2}k^2 + ck^{2+m} + \dots$$

where  $c$  and  $m > 0$  are constants and ... stands for higher order terms in  $k$ . Then Eq. (15.9) becomes:

$$\psi_n(q) = iq \frac{a_n - n\mu}{b_n} - \frac{\sigma^2 n}{2b_n^2} q^2 + \frac{cn}{b_n^{2+m}} q^{2+m} + \dots \quad (15.10)$$

With the choice

$$a_n = n\mu \quad \text{and} \quad b_n = \sigma\sqrt{n},$$

we find

$$\psi^*(q) = \lim_{n \rightarrow \infty} \left[ -\frac{1}{2}q^2 + \frac{cq^{2+m}}{\sigma^{2+m}n^{m/2}} + \dots \right] = -\frac{1}{2}q^2$$

---

<sup>3</sup>The regions  $|x| \gg 1$  of a pdf  $p(x)$  are called the *tails* of  $p(x)$ , and its behaviour in this region is called its *tail behaviour*. More specifically, the region  $x \rightarrow \infty$  is called the *right tail* while the *left tail* indicates the limit  $x \rightarrow -\infty$ .

This is the logarithm of the CF of a gaussian variable, i.e.

$$p^*(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

This result is called the *Central Limit Theorem* (CLT). It states that the limit distribution of the sum of  $n$  i.i.d. random variables with finite variance  $\mathbb{V}[X] = \sigma^2$ , when properly rescaled as

$$Y_n = \frac{S_n - \mu n}{\sqrt{n}\sigma}$$

converges to a Gaussian.

The speed of convergence is ruled by the first non-zero cumulant  $C_{2+m} \neq 0$  of order larger than 2. The correction to the limit is of the order  $n^{-m/2}$ . So if  $m = 1$  the deviation from the Gaussian vanishes as  $1/\sqrt{n}$ . For distributions which are symmetric around the mean  $p(\mu+x) = p(\mu-x)$ ,  $C_3 = 0$  and  $C_4 \neq 0$  so the error vanishes as  $1/n$ . We'll discuss this further later.

- 2)  $\mu = \mathbb{E}[X]$  finite and  $\mathbb{V}[X] = \infty$ .

This case occurs when

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} dx p(x) x^2 = +\infty.$$

This integral can diverge only if the integrand  $p(x)x^2$  is not integrable for  $|x| \rightarrow \infty$ . This implies that for either or both the left and the right tails

$$p(x) \sim |x|^{-\alpha-1} \quad (15.11)$$

with  $\alpha \in (0, 2)$ . The condition  $|\mathbb{E}[X]| < +\infty$  further restricts the range of values of  $\alpha$ , because it requires that  $\alpha > 1$ . The small  $k$  expansion of  $\phi(k)$  in Eq. (15.2) cannot be used because  $\mathbb{E}[X^m] = +\infty$  for all  $m > \alpha$ . Rather, the function  $\phi(k)$  develops a singular behaviour for  $k \ll 1$ , of the form  $\phi(k) = 1 - i\mu k - c|k|^\alpha + \dots$ , where ... stands for higher powers of  $k$ . A non-rigorous argument leading to this result is based on dimensional analysis. If  $x$  has dimension of  $[X]$  (e.g. length) then  $k$  must have dimensions  $[X^{-1}]$ , because the argument of the exponential  $(-ikx)$  must be dimensionless (i.e. a number). For small  $k$ , the leading singular term is

$$\psi(k) + ik\mu \simeq \phi(k) - 1 + ik\mu = \int_{-\infty}^{\infty} dx p(x) [e^{-ikx} - 1 + ikx].$$

Since  $p(x) \sim |x|^{-\alpha-1}$  for  $|x| \rightarrow \infty$ , the integral has dimension  $[X]^{-\alpha}$  and hence it has to be proportional to  $|k|^\alpha$ . The constant  $c$  can be determined by the integral

$$c = \lim_{k \rightarrow 0} \frac{\psi(k) + ik\mu}{|k|^\alpha} = \lim_{k \rightarrow 0} \frac{1}{|k|^\alpha} \int_{-\infty}^{\infty} dx p(x) [e^{-ikx} - 1 + ikx] \quad (15.12)$$

that can be evaluated changing variables to  $z = |k|x$  and using the asymptotic behaviour  $p(x) \sim |x|^{-\alpha-1}$ . In general, the latter can be different in the left and the right tail, i.e.

$$\lim_{x \rightarrow \pm\infty} |x|^{\alpha+1} p(x) = C_{\pm}.$$

This implies that the result of the limit (15.12) depends on the sign of  $k$ , and a tedious calculation shows that

$$\psi(k) \simeq -i\mu k + c \left[ 1 - i\beta \frac{k}{|k|} \tanh\left(\frac{\pi}{2}\alpha\right) \right] |k|^\alpha + \dots,$$

where  $c > 0$  is a constant and

$$\beta = \frac{C_+ - C_-}{C_+ + C_-}.$$

A non-degenerate limit in Eq. (15.9) is obtained with

$$a_n = n\mu \quad \text{and} \quad b_n = (cn)^{1/\alpha}$$

which means that

$$\psi^*(q) = -|q|^\alpha \left[ 1 - i\beta \frac{q}{|q|} \tan\left(\frac{\pi}{2}\alpha\right) \right] \quad (15.13)$$

These are called *Levy stable distributions*, the parameter  $\beta \in [-1, 1]$  is called the *asymmetry* parameter and  $\alpha$  the *characteristic index*. There is not an explicit form for the corresponding pdf  $p^*(x)$ , except for particular cases.<sup>4</sup>

---

<sup>4</sup>The special case  $\alpha = 2$  needs special care. For example, if

$$p(x) = \frac{2}{3} \min(1, x^{-3}), \quad x \geq 0$$

then the distribution of  $S_n$  converges to a Gaussian, as in the CLT, but with  $a_n = \sqrt{n \log n}$ .

**Exercise 15.1**

Show that the Fourier transform of  $f(x) = (1 + x^2)^{-(\alpha+1)/2}$  behaves as

$$\hat{f}(k) \simeq \hat{f}(0) - c|k|^\alpha + \dots$$

as  $k \rightarrow 0$ . Find an expression for the constant  $c$ . *Hint:* use the identity

$$A^{-\gamma} = \frac{1}{\Gamma(\gamma)} \int_0^\infty dt t^{\gamma-1} e^{-At}.$$

**Exercise 15.2**

Limit theorems for products: let  $X_1, \dots, X_n$  be a sequence of positive i.i.d. random variables ( $X_i > 0$ ). Find sequences  $a_n$  and  $b_n$  such that the variable

$$G_n = \left( \prod_{i=1}^n X_i \right)^{1/n} = a_n + b_n Y_n$$

has a non-trivial limit, such that  $Y_n$  converges (in distribution) to a non-degenerate random variable  $Y$  for  $n \rightarrow \infty$ . Find the distribution of  $Y$  depending on the distribution  $p(x)$  of  $X_i$  and the conditions on  $p(x)$  for the limit to exist.

**Exercise 15.3**

What if the two tails have a different behaviour, i.e.

$$p(x) \simeq C_\pm |x|^{-\alpha_\pm-1} \quad x \rightarrow \pm\infty$$

with  $\alpha_+ \neq \alpha_-$ ? One way to attack the problem is to split the sum  $S_n = S_n^+ + S_n^-$  into the sums over the positive and negative variables

$$S_m^+ = \sum_{i: X_i > 0} X_i, \quad S_m^- = \sum_{i: X_i < 0} X_i.$$

Both sums have a limit behaviour of the form  $S_n^\pm \simeq a_n^\pm + b_n^\pm Y_\pm$  where  $Y_\pm$  has a Lévy distribution with parameters  $\alpha^\pm$  and  $\beta = \pm 1$ . This suggests that the sum has a different asymptotic behaviour  $p^*(x) \sim |x|^{-\alpha_\pm-1}$  in the two tails  $x \rightarrow \pm\infty$ . Yet the leading

asymptotic behaviour of the whole sum is dominated by the tail with the smallest value of  $\alpha = \min\{\alpha_-, \alpha_+\}$  and correspondingly,  $\beta = \pm 1$ . Can check that this intuition is correct by a numerical analysis?

As a corollary these results show that the law of large numbers, in its weak form, holds as long as  $\mathbb{E}[X] = \mu$  is finite. First because  $b_n/n \rightarrow 0$  implies that  $S_n/n \rightarrow \mu$  in distribution. Second, because it can be shown that if a random variable converges in distribution to a constant, then it also converges in probability.

- 3)  $|\mathbb{E}[x]| = +\infty$  and  $\mathbb{E}[(x - \mathbb{E}[x])^2] = \infty$ .

This occurs if for  $|x| \gg 1$ ,

$$p(x) \sim |x|^{-\alpha-1} \quad \text{with } 0 < \alpha \leq 1.$$

Then, for  $k \ll 1$ , the cumulant generating function has a leading behaviour  $\psi(k) \sim |k|^\alpha$ . A non-degenerate limit in Eq. (15.9) is obtained with considerations analogous to the previous case, with

$$a_n = 0 \quad \text{and} \quad b_n = (cn)^{1/\alpha}.$$

and the limit of  $\psi_n(q)$  is again given by Eq. (15.13), with  $0 < \alpha < 1$ . The case  $\alpha = 1$  is special because instead of Eq. (15.13) one has:

$$\psi^*(q) = -|q| \left[ 1 + i\beta \frac{q}{|q|} \frac{\pi}{2} \ln |q| \right].$$

For  $\beta = 0$  this is called Cauchy distribution, and it has an explicit form

$$p^*(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

#### Exercise 15.4

Let  $Q_n = T_1 + \dots + T_n$  be the time for the  $n^{\text{th}}$  return to the origin of a random walk (with  $p = 1/2$ ). Show that  $Y = Q_n/n^2$  has a limit distribution, as  $n \rightarrow \infty$  with cumulant generating function  $\psi^*(q)$  as predicted by Eq. (15.13). *Hint:* derive the generating function  $\phi(q)$  of  $T$  from the expression of the generating function we derived in earlier chapters.

**Exercise 15.5**

Consider a two dimensional random walk  $\vec{S}_n = (S_n^x, S_n^y)$  where  $S_n^x$  and  $S_n^y$  are two independent (unbiased) random walks. Let  $T > 0$  be the first return to the  $x$ -axis, i.e. the first time for which  $S_T^y = 0$ . Show that the cumulant generating function of  $S_T^x$  is given by  $\psi(q) = \log[1 - |\sin q|]$ .

Using this, show that for  $m \gg 1$ , the position on the  $x$ -axis corresponding to the  $m^{\text{th}}$  time when  $S_n^y = 0$  is well approximated by  $mY$  where  $Y$  follows the Cauchy distribution. You can run numerical simulations to confirm this conclusion.

Note that, for  $\alpha < 1$ , the law of large numbers does not hold because  $S_n$  has fluctuations  $b_n \sim n^{1/\alpha}$  which grow faster than<sup>5</sup>  $n$ . Therefore the sum of  $n$  elements grows faster than  $n$ , which is strange at first sight. This occurs because the sum is dominated by few elements which are themselves of order  $X_i \sim n^{1/\alpha}$ . Remember that the Glivenko-Cantelli theorem shows that for large  $n$ , if  $X_{[k]}$  is the  $k^{\text{th}}$  largest value among  $X_1, \dots, X_n$ , then

$$\frac{k}{n} \approx \int_{X_{[k]}}^{\infty} dx p(x) \sim X_{[k]}^{-\alpha}$$

when  $p(x) \sim x^{-\alpha-1}$ . Inverting this relation shows that the  $k^{\text{th}}$  largest value in the sequence is of the order

$$X_{[k]} \sim \left(\frac{n}{k}\right)^{1/\alpha}.$$

Hence for  $\alpha < 1$  the largest element  $X_{[1]} \sim n^{1/\alpha}$  is of the same order of magnitude as the whole sum.

A measure of how a sum is unevenly dominated by its different terms is given by the *participation ratio*, which in our case is defined as

$$Y = \sum_{i=1}^n \left(\frac{X_i}{S_n}\right)^2. \quad (15.14)$$

When all terms contribute more or less equally,  $X_i \sim S_n/n$  and each of the terms of the sum is of order  $1/n^2$ . Therefore the participation ratio vanishes as  $n \rightarrow \infty$ . The same is true in our case when  $X_i$  are i.i.d.

<sup>5</sup>This is why we can set  $a_n = 0$ . Taking  $a_n = \mu n$  would not change the limit in Eq. (15.9).



random variables with a pdf with tail behaviour as in Eq. (15.11) with  $\alpha > 1$ . When  $S_n$  is dominated by a finite number of terms, instead, some of the terms in Eq. (15.14) are finite and hence  $Y$  remains finite even when  $n \rightarrow \infty$ . This applies to our discussion of sums of i.i.d. random variables with  $\alpha < 1$ . In this case  $Y$  remains a random variable even in the limit  $n \rightarrow \infty$  with  $\mathbb{E}[Y] \rightarrow 1 - \alpha$ .  $Y$  is not self-averaging as  $n \rightarrow \infty$  for  $\alpha < 1$ .<sup>6</sup>

### 15.1.4 Stable distributions and universality

From the previous discussion it should be clear that the sum of two independent Gaussian variables is also a Gaussian variable, whose mean and variance are the sum of the means and the variances of the original variables. Likewise, you can check that the mean of two variables with a Cauchy distribution is also distributed as a Cauchy distribution. In general, given two i.i.d. random variables  $X_1$  and  $X_2$ , with pdf  $p^*(x)$ , if there are constants  $a$  and  $b$  such that the random variable  $(X_1 + X_2 - a)/b$  has the same distribution  $p^*(x)$ , then  $p^*(x)$  is called a *stable distribution*.

The cumulant generating function of a stable distribution has to satisfy the equation

$$\psi^*(q) = \frac{iq a}{b} + 2\psi^*(q/b) \quad (15.15)$$

The Gaussian and the Lévy distributions are all stable distributions. Stable distributions are also infinitely divisible, in the sense that if  $X$  has distribution  $p^*(x)$  then, for any  $n \in \mathbb{N}$ , there are constants  $a_n$  and  $b_n$  such that  $a_n + b_n X$  is the sum of  $n$  i.i.d. random variables  $X_i$  with distribution  $p^*(x)$ .

#### Exercise 15.6

Can  $\psi^*(q) = -q^4$  be a the cumulant generating function of a stable distribution?

One way of deriving the distribution of the sum of  $n$  random variables with distribution  $p(x)$ , in the limit  $n \rightarrow \infty$ , is to first sum the variables in pairs, and then consider the sum of the pairs. This procedure can be iterated for  $k$

<sup>6</sup>See [23] for more details.

steps,<sup>7</sup> so that the total sum  $S_n$  is the sum of  $n/2^k$  independent variables  $S_{2^k}$ , each of which is the sum of  $2^k$  independent variables  $X_i$ . The distribution of  $S_{2^{k+1}}$  can be derived from that of  $S_{2^k}$  because  $S_{2^{k+1}} = S_{2^k} + S'_{2^k}$ . It is clear that we can obtain the asymptotic distribution of  $S_n$  by studying the distribution of  $S_{2^k}$  as  $k \rightarrow \infty$ . Yet, the location and scale of the distribution of  $S_{2^k}$  will change with  $k$ , so in order to obtain a non-degenerate distribution in the limit it is necessary to *rescale* the variables  $S_{2^k}$  in an appropriate manner, defining

$$Y_k = \frac{S_{2^k} - a_{2^k}}{b_{2^k}}$$

in such a way that the location and scale of  $Y_k$  are independent of  $k$ . For example,  $a_{2^k}$  and  $b_{2^k}$  can be computed imposing that  $P\{Y_k > 0\} = P\{Y_k < 0\}$  and  $P\{|Y_k| \leq 1\} = 1/2$  for all  $k$ . If  $p^{(k)}(y)$  is the pdf of the variables  $Y_k$ , then the distribution  $p^{(k+1)}$  of  $Y_{k+1}$  can be obtained using the recursion relation

$$Y_{k+1} = \frac{Y_k + Y'_k - A_k}{B_k}$$

where  $Y_k$  and  $Y'_k$  are i.i.d. with pdf  $p^{(k)}$ ,  $A_k = (a_{2^{k+1}} - 2a_{2^k})/b_{2^k}$  and  $B_k = b_{2^{k+1}}/b_{2^k}$ . This defines a transformation<sup>8</sup>

$$p^{(k+1)} = \mathcal{R}(p^{(k)}) \quad (15.16)$$

$$p^{(k+1)}(x) = \int_{-\infty}^{\infty} dx_1 dx_2 p^{(k)}(x_1) p^{(k)}(x_2) \delta\left(x - \frac{x_1 + x_2 - A_k}{B_k}\right)$$

where  $A_k$  and  $B_k$  are determined by the conditions

$$\int_0^{\infty} dx p^{(k+1)}(x) = \int_{-\infty}^0 dx p^{(k+1)}(x), \quad \int_{-1}^1 dx p^{(k+1)}(x) = \frac{1}{2}.$$

---

7

$$\begin{aligned} S_n &= X_1 + X_2 + \dots + X_n \\ &= S_2^{(1)} + S_2^{(2)} + \dots + S_2^{(n/2)} \\ &\vdots \\ &= S_{2^k}^{(1)} + S_{2^k}^{(2)} + \dots + S_{2^k}^{(n/2^k)} \\ &\vdots \end{aligned}$$

The sum  $S_n$  of  $n$  random variables can be described in terms of “block” variables  $S_{2^k}$  at different “scales”  $k$ .

<sup>8</sup>In terms of the cumulant generating function, this transformation takes the simpler form

$$\psi^{(k+1)}(q) = \mathcal{R}(\psi^{(k)}) \equiv -iqA_k + 2\psi^{(k)}(q/B_k)$$

$\mathcal{R}$  defines a transformation  $p^{(k)} \mapsto p^{(k+1)}$  in the space of distributions with the same scale and location, that starts from the original distribution  $p^{(0)} = p$ . This is the simplest example of a *Renormalisation Group* transformation, that relates the statistical description of a system (here  $S_n$ ) at two different “scales”  $\ell$  and  $\ell'$ , i.e. in terms of “block” variables  $S_\ell$  and  $S_{\ell'}$  (here  $\ell = 2^k = \ell'/2$ ). This transformation is used in statistical physics to study the critical behaviour of systems at second order phase transition points.<sup>9</sup> This transformation is based on two steps i) *coarse graining*, i.e.  $Z_k = Y_k + Y'_k$  and ii) *rescaling*  $Y_{k+1} = (Z_k - A_k)/B_k$ .

The “flow” induced by  $\mathcal{R}$  in the space of probability distributions converges to fixed points

$$\lim_{k \rightarrow \infty} p^{(k)} = p^* = \mathcal{R}(p^*), \quad \lim_{k \rightarrow \infty} A_k = A^*, \quad \lim_{k \rightarrow \infty} B_k = B^*$$

that define the distribution of properly rescaled sums (with constants  $A^*$  and  $B^*$ ) of infinitely many random variables with pdf  $p(x)$ . In terms of CF, the equation  $p^* = \mathcal{R}(p^*)$  coincides with Eq. (15.15). Note that:

- only for appropriately chosen values  $A^*$  and  $B^*$  the transformation (15.16) has a fixed point  $p^*$ .
- The transformation  $\mathcal{R}$  preserves the tail properties of the distributions, i.e. if  $p^{(k)}(x)$  has a finite variance, then  $p^{(k+1)}(x)$  has also a finite variance. If  $p^{(k)}(x) \sim |x|^{-\alpha-1}$  with  $\alpha < 2$  then  $p^{(k+1)}(x)$  has the same behaviour.<sup>10</sup>

<sup>9</sup>In statistical physics the renormalisation group is a way of defining appropriately the thermodynamic limit by relating the description of a system at length-scale  $2L$  to that of systems of size  $L$ . In this limit, appropriately rescaled quantities should have the same fluctuation properties at different (large enough) scales. The renormalisation group is also used in particle physics in order to deal with the microscopic limit. Theories of interacting particles (e.g. electrons and photons) suffer from ultra-violet divergences (i.e. arising from processes taking place at very small scales). These divergences can be “cured” by the same “renormalisation” procedure, which relates the description of a system at two different scales. The parameters of the theory (e.g. the mass and the charge of the electron) can be adjusted so that the renormalisation transformation admits a fixed point which describes a well defined ultra-violet limit.

Renormalisability implies limits to what we can learn by studying systems at one scale, on systems at a different scale. When the theory of a macroscopic system is renormalisable, which means that this program can be successfully carried out, the behaviour of the system it describes is completely independent of microscopic details. This means that there is no measure on the macroscopic properties that can reveal microscopic properties. Likewise, in particle physics, if particles like electrons and photons are fully described by a renormalizable theory, then it is not possible to learn about more fundamental constituents of matter (e.g. quarks) by studying how electrons and protons interact. For more, see the essay of Tian Yu Cao in [24].

<sup>10</sup>This is more clearly seen from the fact that the singular behaviour of  $\psi^{(k)}(q)$  is the same as that of  $\psi^{(k+1)}(q)$  for  $q \rightarrow 0$ .

- The equation  $p^* = \mathcal{R}(p^*)$  admits different fixed points corresponding to different tail behaviors. The space of distributions is divided into the basins of attraction of each of them. For example, all distributions with a finite variance belong to the basin of the Gaussian fixed point. The tail behaviour of  $p(x)$  determines to which fixed point  $p^*$  the distribution  $p^{(k)}$  will converge to. In this sense, the distribution  $p^*$  is *universal* because it is attained asymptotically, starting from whatever distribution  $p$  with the same tail behaviour.
- Asymptotically the scale parameter  $B_k$  converges to  $B^*$ . This means that for  $n = 2^\ell$  the combined scale parameter from the original distribution  $p$  to  $p^{(\ell)}$  should be equal to

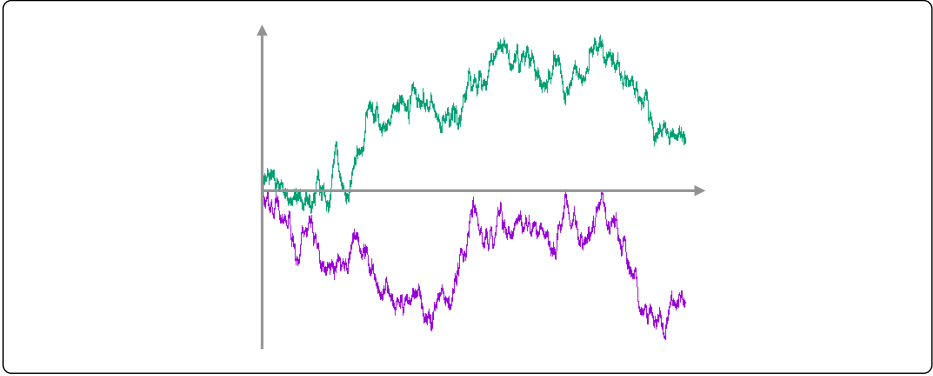
$$b_n = \prod_{k=1}^{\ell} B_k \simeq (B^*)^\ell = n^{1/\alpha}, \quad \frac{1}{\alpha} = \log_2 B^*.$$

The fact that the coefficient  $b_n$  takes the *scaling* form  $b_n = n^{1/\alpha}$  is a direct consequence of *scale invariance*, i.e. of invariance under the scale transformations  $\mathcal{R}$  (where “scale” here refers to the “size” of the sums in each block).

### 15.1.5 Sums as stochastic processes

Take  $X_i$  with  $\mathbb{E}[X_i] = 0$  and plot the sum  $S_n$  of  $n$  i.i.d. random variables  $X_i$  as a function of  $n$ . Imagine doing the plot for two values of  $n$  which are large enough that you can’t distinguish individual dots on the plots (say  $n = 10^4$  and  $10^5$ ) and remove the tick labels from both axes, in both plots. A concrete manifestation of the existence of a limit theorem for sums is that you should not be able to distinguish which plot was generated with the larger value of  $n$  and which was generated with the smaller one by just looking at the plots. Put differently, if for both plots you rescale the  $x$  axis by  $n$  and the  $y$ -axis by  $n^{1/\alpha}$  you’ll get two curves which are *statistically* indistinguishable<sup>11</sup> You’re encouraged to generate these curves numerically and to verify this statement.

<sup>11</sup>Indeed, if you zoom in Figure 34 in any interval of size  $b$  and rescale the  $x$ -axis by  $b$  and the  $y$ -axis by  $\sqrt{b}$ , then you get a curve that is statistically indistinguishable from the original one. If  $n$  is really very large, the same is true if you zoom into a part of the part and so on. Since you cannot distinguish the scale of the interval by the shape of the curve, these curves are called *self-similar*. Objects that enjoy this self-similarity property are called *fractals*. A curve such as  $\sin(x)$  does not enjoy this property, because there is a special scale  $x \sim 2\pi$  that corresponds to the period. If you zoom in on a small interval the curve will look like a straight line, which is different from the original curve.



**Figure 34.** Two independent rescaled sums  $S_k/\sqrt{n}$  of  $k$  random variables with finite variance and zero mean are plotted versus  $k/n$ , with  $k = 1, \dots, n$ , for  $n = 10^4$  and  $n = 10^5$ . Which is which?

So, the existence of a limit for  $n \rightarrow \infty$  of  $S_n/b_n$  implies that one can define these trajectories in a continuous time. More precisely, we can set  $t = n\Delta t$  and look for an appropriate rescaling

$$Z(t) = (\Delta t)^\beta S_{n=t/\Delta t}$$

such that  $Z(t)$  is finite when  $\Delta t \rightarrow 0$ . It is clear that  $b_n \sim n^{1/\alpha}$  implies that a finite limit is achieved for  $\beta = 1/\alpha$  and

$$Z(t) = t^{1/\alpha} Y$$

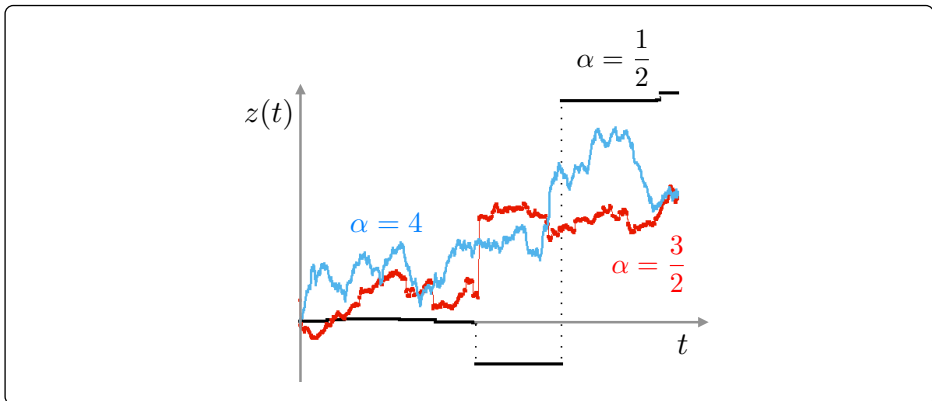
where  $Y$  has Gaussian distribution for  $\alpha = 2$  and a Levy distribution for  $0 < \alpha < 2$ . In the former case  $Z(t)$  is called the Wiener process, in the latter it is called a Levy process. What are the properties of the trajectory?

First, notice that  $Z(t)$  is a process with independent increments. This means that for  $t > t_0$ ,  $Z(t) - Z(t_0) = |t - t_0|^{1/\alpha} Y$  where  $Y$  is independent of  $Z(t_0)$  and has an *universal* distribution (either Gaussian or Levy).

Next, we can ask: is  $Z(t)$  a continuous function of  $t$ ?

Usually a function is continuous at  $t_0$  if for any  $\epsilon > 0$  one can find a  $\delta > 0$  such that for all  $t \in [t_0 - \delta, t_0 + \delta]$  we have  $|Z(t) - Z(t_0)| < \epsilon$ . Now, this definition cannot be used in the present case, because  $Z(t) = Z(t, \omega)$  is a random variable.

We should define what convergence means. For example if we adopt an  $L^2$  norm, we would say that the process is continuous if  $\mathbb{E} [|Z(t) - Z(t_0)|^2] \rightarrow 0$  as  $t \rightarrow t_0$ . Now clearly the expected value diverges for all  $\alpha < 2$ , suggesting that Levy processes are not continuous. However, if one instead uses an  $L_p$



**Figure 35.** Rescaled function  $Z(t)$  obtained from sums  $S_n$  of  $n \leq 10^3$  random variables with a symmetric distribution with asymptotic behaviour  $p(x) \sim |x|^{-\alpha-1}$  for  $\alpha = 1/2$  (black)  $\alpha = 3/2$  (red) and  $\alpha = 4$  (blue). The random variables  $X_i$  are generated taking  $X_i = \sigma_i U_i^{-1/\alpha}$  where  $\sigma_i = \pm 1$  with equal probability and  $U_i$  is an uniform random variable. Note that for  $\alpha < 1$  the largest jump accounts for a large part of the total excursion of  $S_n$ .

norm, requiring  $\mathbb{E}[|Z(t) - Z(t_0)|^p] \rightarrow 0$  as  $t \rightarrow t_0$ , one would conclude that processes with  $\alpha > p$  are continuous whereas those with  $\alpha \leq p$  are not. Now if you look at the plots of  $Z(t)$  in Figure 35, it seems this can't be true.

A different way to approach the problem is the following: consider an interval  $[t_0, t_1]$  and divide it into  $m$  smaller intervals of size  $\Delta t = (t_1 - t_0)/m$ . Let  $\delta z_i$  be the increment of  $Z(t)$  in the  $i^{\text{th}}$  interval ( $i = 1, \dots, m$ ). Then we say that the process is continuous if, for any  $\epsilon > 0$  we have

$$\lim_{m \rightarrow \infty} P\{|\delta z_i| < \epsilon \forall i = 1, \dots, m\} = 1.$$

If this is not true, then there is a finite probability that  $Z(t)$  has a discontinuous jump larger than  $\epsilon$  in the interval  $[t_0, t_1]$ . Since intervals are independent, it suffices to compute the probability  $P\{|\delta z| < \epsilon\}$  of the deviation in one interval. In general we have

$$\begin{aligned} P\{|\delta z_i| < \epsilon \forall i = 1, \dots, m\} &= P\{|\delta z| < \epsilon\}^m \\ &= [1 - P\{|\delta z| > \epsilon\}]^m \simeq e^{-mP\{|\delta z| > \epsilon\}} \end{aligned} \quad (15.17)$$

where we assumed that  $P\{|\delta z| > \epsilon\} \rightarrow 0$  as  $m \rightarrow \infty$ . In the regime of the central limit theorem ( $\alpha \geq 2$ ) we have

$$P\{|\delta z| > \epsilon\} \simeq \sqrt{\frac{2|t_1 - t_0|}{\pi \epsilon m}} e^{-\frac{m \epsilon^2}{2|t_1 - t_0|}}$$

which vanishes exponentially as  $m \rightarrow \infty$ . Therefore  $mP\{|\delta z| > \epsilon\} \rightarrow 0$  for  $m \rightarrow \infty$  and Eq. (15.17) yields  $P\{|\delta z_i| < \epsilon \forall i = 1, \dots, m\} \rightarrow 1$  in the same limit. From this we conclude that the Wiener process is continuous.

In the case of the Levy process, instead, we have

$$\delta z_i(\omega) = \left( \frac{|t_1 - t_0|}{m} \right)^{1/\alpha} Y_i(\omega)$$

where  $Y_i$  are all i.i.d. with Levy distribution. Hence, since  $Y$  has a pdf  $p(x) \simeq Cx^{-\alpha-1}$  for large  $x$ ,

$$P\{|\delta z_i| > \epsilon\} \simeq 2C \int_{\frac{\epsilon m^{1/\alpha}}{|t_1 - t_0|^{1/\alpha}}}^{\infty} x^{-\alpha-1} dx = \frac{2C}{\alpha} \frac{|t_1 - t_0|}{m \epsilon^\alpha}.$$

Therefore the limit in Eq. (15.17) is

$$\begin{aligned} \lim_{m \rightarrow \infty} P\{|\delta z_i| < \epsilon \forall i = 1, \dots, m\} &= \lim_{m \rightarrow \infty} \left[ 1 - \frac{2C}{\alpha \epsilon^\alpha} \frac{|t_1 - t_0|}{m} \right]^m \\ &= e^{-\frac{2C|t_1 - t_0|}{\alpha \epsilon^\alpha}} \end{aligned} \quad (15.18)$$

which is finite. Therefore with a finite probability there is at least one infinitesimally small interval where the process has increments larger than  $\epsilon$ . The process is not continuous. Notice that as  $\epsilon \rightarrow 0$  in every interval with probability very close to one you will find jumps larger than  $\epsilon$ , so the function is no-where continuous.

### Exercise 15.7

Notice the similarity of Eq. (15.18) with the probability that a Poisson random variable with mean

$$\lambda = \frac{2C|t_1 - t_0|}{\alpha \epsilon^\alpha}$$

takes value  $k = 0$ . Show that the probability to observe  $k$  jumps  $|\delta z_i| > \epsilon$  in the time interval  $[t_0, t]$  is given by the Poisson distribution  $\frac{\lambda^k}{k!} e^{-\lambda}$ . Is this a coincidence?

In the case of the Wiener process, we can further ask whether the function  $Z(t)$  is differentiable or not. The derivative of a function  $Z(t)$  at  $t$  is defined as

$$\frac{dZ}{dt} = \lim_{h \rightarrow 0} \frac{Z(t+h) - Z(t)}{h}.$$

One way to address this question is to compute the probability that the increment  $Z(t+h) - Z(t)$  is smaller than  $Kh$ , for an arbitrarily large constant  $K$ , i.e.

$$P\{|Z(t+h) - Z(t)| < Kh\}$$

where  $K$  is an arbitrarily large positive constant. Please verify that the limit of this probability when  $h \rightarrow 0$  is zero. This means that the Wiener process is no-where differentiable.

## 15.2 Limit theorems for extremes

Finding the minimum or the maximum of a number of random variables is a classical subject in probability theory that goes under the name of *extreme value theory*. Situations where maxima and minima occur are called *extreme events*. There are many situations where we may be interested in extreme events. For example, engineers need to ensure that the structure they build (e.g. a bridge) will resist perturbations (e.g. floods) for a long time (at least longer than their lifetime). Hence they need to estimate what is the maximal size of the perturbation they can expect over a certain period of time. World records in athletics is another example of extreme random variables.

Here we focus on the simple case of finding the maxima and minima

$$Z_n = \max\{X_1, \dots, X_n\}, \quad W_n = \min\{X_1, \dots, X_n\}$$

of  $n$  i.i.d. random variables  $X_1, \dots, X_n$  with common pdf  $p(x)$ . If we change the sign of each random variable  $X'_i = -X_i$ , then the problem of finding the minimum becomes that of finding the maximum, i.e.  $Z'_n = -W_n$ . Hence, it is enough to consider  $Z_n$  only. The problem can be stated as follows.

Find sequences  $a_n, b_n \in \mathbb{R}$  such that<sup>12</sup>

$$Z_n = a_n + b_n \Lambda_n \tag{15.19}$$

and  $\Lambda_n \rightarrow \Lambda$ , in distribution as  $n \rightarrow \infty$ , where  $\Lambda$  has a non-degenerate distribution. In other words,  $a_n$  and  $b_n$  should be such that the limit

$$H(x) = \lim_{n \rightarrow \infty} P\{\Lambda_n < x\}$$

---

<sup>12</sup>As for sums of random variables,  $a_n$  provides a measure of how large we expect the maximum to be, whereas  $b_n$  gives an estimate of how much the maximum can vary between two different realisations of the sequence of  $n$  independent random variables, because  $Z_n^{(1)} - Z_n^{(2)} = b_n(\Lambda_n^{(1)} - \Lambda_n^{(2)})$ .



yields the cumulative distribution of a non-degenerate random variable. The key idea is that

$$P\{\Lambda_n < x\} = P\{Z_n < a_n + b_n x\} \quad (15.20)$$

$$= P\{X_i < a_n + b_n x, \forall i\} \quad (15.21)$$

$$= \left(1 - \int_{a_n + b_n x}^{\infty} dx' p(x')\right)^n \quad (15.22)$$

In words, if  $Z_n < a_n + b_n x$  then all  $X_i$  must be less than  $a_n + b_n x$ , and the last equality is due to the fact that the variables are all i.i.d.. In order for the limit as  $n \rightarrow \infty$  to be non-trivial, the integral in the last equation must be proportional to  $n^{-1}$ . Indeed, if  $a_n$  and  $b_n$  are chosen so that

$$\lim_{n \rightarrow \infty} n \int_{a_n + b_n x}^{\infty} dx' p(x') = c(x) \quad (15.23)$$

then, for large  $n$  the right hand side of Eq. (15.22) is  $\simeq \left(1 - \frac{c(x)}{n}\right)^n$  and we have

$$H(x) = \lim_{n \rightarrow \infty} P\{\Lambda_n < x\} = e^{-c(x)}. \quad (15.24)$$

The pdf of the random variable  $\Lambda$  is obtained as

$$h(x) = \frac{dH(x)}{dx} = -c'(x)e^{-c(x)}. \quad (15.25)$$

The problem is then condensed in finding  $a_n$  and  $b_n$  such that the limit in Eq. (15.23) is well defined. This limit probes the tail of the distribution, i.e. that region of  $x$  such that the probability that  $X_i > x$  is of the order of<sup>13</sup>  $1/n$ .

There are three main cases:

- 1) *Power law distributions*  $p(x) \simeq Ax^{-\gamma-1}$  for  $x \gg 1$ . Then the integral in Eq. (15.23) can be done explicitly

$$n \int_{a_n + b_n x}^{\infty} dx' p(x') \simeq \frac{An}{\gamma} (a_n + b_n x)^{-\gamma}$$

Then the choice

$$\begin{aligned} a_n &= 0 \\ b_n &= \left(\frac{An}{\gamma}\right)^{1/\gamma} \sim n^{1/\gamma} \end{aligned}$$

---

<sup>13</sup>This is intuitive, because the maximum of  $n$  random variables  $X_i$  is expected to fall in this region with finite probability.

leads to the result

$$\begin{aligned} c(x) &= x^{-\gamma}, \quad \Rightarrow \quad H(x) = e^{-x^{-\gamma}} \theta(x) \\ h(x) &= \gamma x^{-\gamma-1} e^{-x^{-\gamma}} \quad (x > 0). \end{aligned} \quad (15.26)$$

Note that  $h(x)$  preserves the same behaviour of  $p(x)$  for large  $x$  and is *universal*, in the sense that it does not depend on any other details of  $p(x)$ .

It is interesting to compare the behaviour of sums and of extremes, for distributions with a power law tail.

- For  $\gamma > 2$ , the sum  $S_n$  is asymptotically described by the Central Limit Theorems whereas the distribution of the maximum retains the same tail behaviour of the distribution of  $X_i$  (see Eq. (15.26)). In addition, the sum is of the order  $S_n \sim n$  with fluctuations  $\delta S_n$  of order  $\sqrt{n}$ . The maximum  $Z_n$  is the largest element in the sum  $S_n$ , and since  $Z_n \sim n^{1/\gamma}$ , it is negligible with respect to both the sum and its fluctuations:

$$Z_n \ll \delta S_n \ll S_n \quad (\gamma > 2)$$

- For  $1 < \gamma < 2$  the sum and the maximum have pdf's with the same asymptotic behaviour. The sum is still of order  $n$  with fluctuations  $\delta S_n \sim n^{1/\gamma}$  which are of the same order of  $Z_n$ , i.e.

$$Z_n \sim \delta S_n \ll S_n \quad (1 < \gamma < 2)$$

- For  $\gamma < 1$  the maximum grows as fast as the whole sum, and both grow faster than  $n$  (as  $n^{1/\gamma}$ ):

$$Z_n \sim \delta S_n \sim S_n \quad (\gamma < 1)$$

The maximum accounts for a finite fraction of the whole sum.

2) *Distribution with a support bounded to  $x \leq \omega$ .* If for  $x \approx \omega$

$$p(x) \simeq \begin{cases} (\omega - x)^{\gamma-1} & x \leq \omega \\ 0 & x \geq \omega \end{cases},$$

the integral in Eq. (15.23) yields

$$n \int_{a_n + b_n x}^{\omega} dx' p(x') \simeq \frac{An}{\gamma} (\omega - a_n - b_n x)^{\gamma}$$

Then the choice

$$a_n = \omega$$

$$b_n = \left( \frac{\gamma}{An} \right)^{1/\gamma} \sim n^{-1/\gamma}$$

leads to the result

$$c(x) = (-x)^\gamma, \Rightarrow H(x) = e^{-(x)^\gamma} \quad (x < 0)$$

$$h(x) = \gamma(-x)^{-\gamma-1} e^{-(x)^\gamma}$$

and  $H(x) = 1$ ,  $h(x) = 0$  for  $x \geq 0$ .

Again, the singular behaviour of  $h(x)$  for  $x \rightarrow 0^-$  coincides with that of  $p(x) \sim (\omega - x)^{\gamma-1}$  for  $x \rightarrow \omega^-$ . This is the only relevant feature that the limit distribution retains of the original distribution  $p(x)$ .

Note that  $h(x)$  for  $\gamma = 1$  coincides with the exponential distribution for  $x \leq 0$ . This implies that the minimum of many random variables with a support which has a finite inferior limit  $\omega$  and whose pdf as  $x \rightarrow \omega^+$  is finite, is an exponential random variable. In particular the minimum of many exponential random variables is itself an exponential random variable.

- 3)  $p(x)$  with unbounded support, that falls off faster than any power for  $x \rightarrow \infty$ . This includes distributions for which the moments  $\mathbb{E}[|X|^m] < +\infty$  are finite for any  $m$ . Rather than carrying out the limit in general, we consider a specific example, the stretched exponential distribution

$$p(x) = \begin{cases} \nu x^{\nu-1} e^{-x^\nu} & x \geq 0 \\ 0 & x \leq 0 \end{cases}, \quad (15.27)$$

for which the limit can be carried out easily. Eq. (15.23) reads

$$n \int_{a_n + b_n x}^{\infty} dz \nu z^{\nu-1} e^{-z^\nu} = n e^{-(a_n + b_n x)^\nu} \quad (15.28)$$

$$= n e^{-a_n^\nu - \nu a_n^{\nu-1} b_n x - \frac{\nu(\nu-1)}{2} a_n^{\nu-2} b_n^2 x^2 + \dots}$$

We can get rid of the factor  $n$  by taking  $a_n = (\log n)^{1/\nu}$ . Fixing  $b_n$  such that the coefficient of  $x$  in the second term of the exponential equals one (i.e.  $\nu a_n^{\nu-1} b_n = 1$ ), yields

$$b_n = \frac{1}{\nu} (\log n)^{1/\nu-1}. \quad (15.29)$$

It is easy to check that all other terms in the expansion vanish as  $n \rightarrow \infty$ . Therefore, with this choice of  $a_n$  and  $b_n$ , the limit in Eq. (15.23) becomes  $c(x) = e^{-x}$ . For a general distribution in this class, the coefficient  $a_n$  can be chosen as the value of  $x$  for which the expected number of points larger than  $a_n$ , in a sample of  $n$  i.i.d. draws, equals one, i.e.

$$nP\{X > a_n\} = n \int_{a_n}^{\infty} dx p(x) = 1. \quad (15.30)$$

In this way,  $a_n$  provides a measure of the value that we expect for the maximum of  $n$  i.i.d. random variables. The coefficient  $b_n$  can be taken as a measure of the scale of fluctuations of  $X_i$  conditional to  $X_i > a_n$ , i.e.

$$b_n = \mathbb{E}[X - a_n | X > a_n] = n \int_{a_n}^{\infty} dy \int_y^{\infty} dx p(x). \quad (15.31)$$

Then it can be shown that the limit in Eq. (15.23) yields  $c(x) = e^{-x}$  and

$$H(x) = e^{-e^{-x}}, \quad h(x) = e^{-x-e^{-x}}. \quad (15.32)$$

Eq. (15.32) is known as the *Gumbel distribution*.<sup>14</sup>

It is interesting to note that  $b_n$  in Eq. (15.29) diverges if  $\nu < 1$  whereas  $b_n \rightarrow 0$  for  $\nu > 1$ . This means that the maximum  $Z_n$  is well approximated by the deterministic sequence  $a_n$  for  $\nu > 1$ , because  $|Z_n - a_n| \rightarrow 0$  in probability as  $n \rightarrow \infty$ .<sup>15</sup> Conversely, for  $\nu < 1$  the fluctuations of  $Z_n$  become larger for larger values of  $n$ . The case  $\nu = 1$ , which corresponds to the exponential, is special, because then  $b_n = 1$  independently of  $n$ .

This last case can be considered as the limit  $\gamma \rightarrow \infty$  of the first case. Indeed there is a general formula that, apart from an affine transformation, includes the three cases discussed above, which is the *Fisher-Tippett* distribution

$$P_{\xi}(x) = P\{\Lambda \geq x\} = e^{-(1+\xi x)^{1/\xi}} \quad \text{for } \xi x > -1, \quad (15.33)$$

whereas when  $\xi x \leq -1$ ,  $P_{\xi}(x) = 0$  for  $\xi > 0$  and  $P_{\xi}(x) = 1$  if  $\xi < 0$ . The first case corresponds to  $\xi = 1/\gamma > 0$  whereas the second to  $\xi = -1/\gamma < 0$ . The third case to the limit  $\xi \rightarrow 0$ .

<sup>14</sup>Also the other limit distributions have names, but they are less used.

<sup>15</sup>The limit does not always exist. E.g. for  $p(x) = 1/[x(\ln x)^2]$  there are no coefficients  $a_n$  and  $b_n$  for which  $(Z_n - a_n)/b_n$  has a non-degenerate distribution. Indeed there is a limit for  $\tilde{Z}_n = \max\{\ln X_1, \dots, \ln X_n\} = \ln Z_n$  of the form  $\tilde{Z}_n = \tilde{a}_n + \tilde{b}_n \tilde{Y}$ . Therefore, for  $n \gg 1$ , the maximum of  $X_i$  is well approximated by  $Z_n = e^{\tilde{a}_n + \tilde{b}_n \tilde{Y}}$ . This explains why there cannot be a non-degenerate limit of  $(Z_n - a_n)/b_n$ .

### 15.2.1 Some applications\*

**The Gillespie algorithm.** Consider the dynamics of a system that can be in any of  $n$  states and that can make transitions from states  $i$  to  $j$  at any time  $t \in \mathbb{R}$ .<sup>16</sup> More precisely, if the system is in state  $i$  at time  $t$ , it can jump to state  $j$  in the interval  $[t, t + dt)$  with a probability  $w_{i,j}dt$ . For infinitesimal  $dt$  the probability that two transitions occur in the same interval  $[t, t + dt)$  is proportional to  $dt^2$  and it is therefore negligible. The dynamics is a sequence of transitions between states at different times. The constant  $w_{i,j}$  is called the *transition rate*.

Imagine that we want to generate a trajectory of this system with a computer code. In order to do so — i.e. to *simulate* the dynamics — one can fix a small time increment  $\Delta t$  and then, with probability  $w_{i,j}\Delta t$  perform transition  $i \rightarrow j$ , for all  $j$ . The problem is that at most one transition should occur, which implies that  $\Delta t$  should be taken very small. This is computationally very inefficient.<sup>17</sup> This method gets more accurate the smaller is  $\Delta t$ , i.e. the more it is inefficient. There is a smarter way to simulate this process that relies on the realisation that the waiting time for the transition  $i \rightarrow j$  to occur is exponential

$$p_{i,j}(t) = w_{i,j}e^{-w_{i,j}t}$$

Hence, it is possible to draw  $n$  waiting times  $T_{i,j}$  for all  $j = 1, \dots, n$  from the corresponding exponential distribution and find the transition  $i \rightarrow j^*$  that will occur first, i.e.

$$j^* = \arg \min_{j=1, \dots, n} T_{i,j}.$$

Then one can execute the transition and advance time by  $\Delta t = T_{i,j^*}$ . This is exact but it still requires to draw  $n$  random variables for each transition. We can do better because we can compute the probability that  $j^*$  takes any value,

<sup>16</sup>A large number of systems can be described in these terms. For example, in a mixture of molecules, state  $i$  will correspond to a particular composition of the mixture, and state  $j$  to the composition that attains if a certain chemical reaction takes place. For completeness, let us mention that the probability  $p_i(t)$  to find the system in state  $i$  at time  $t$  satisfies the so called *Master Equation*

$$\frac{dp_i}{dt} = \sum_{k \neq i} [p_k w_{k,i} - p_i w_{i,k}].$$

This states that changes in  $p_i$  either occur because of transitions from other states  $k$  to  $i$  (the first term in [...]) or because of transitions out of  $i$ , to other states  $k$  (second term in [...]).

<sup>17</sup>In order to find out whether jump  $i \rightarrow j$  occurs in the time interval  $[t, t + \Delta t)$  in a simulation, draw a uniform random number  $R \in [0, 1]$  and compare it with  $w_{i,j}\Delta t$ . If  $R < w_{i,j}\Delta t$  then the jump occurs. Doing this for each  $j$  implies that  $n$  random numbers are needed to perform one transition.

by computing the probability that the corresponding time is smaller than all the others, i.e.

$$\begin{aligned} P\{j^* = j\} &= P\{T_{i,j} < T_{i,k}, \forall k \neq j\} \\ &= \int_0^\infty dt w_{i,j} e^{-w_{i,j}t} \prod_{k \neq j} e^{-w_{i,k}t} \\ &= \frac{w_{i,j}}{W_i}, \quad W_i = \sum_{k=1}^n w_{i,k} \end{aligned}$$

where we used the fact that  $P\{T_{i,k} > t\} = e^{-w_{i,k}t}$  for exponential random variables. Therefore with one draw from the distribution  $P\{j^* = j\}$  we can find the transition  $i \rightarrow j^*$  that will occur. In addition, the waiting time for the transition is the minimum of the waiting times for all the processes, i.e.  $T_{i,j^*} = \min_k T_{i,k}$ , and we can compute its probability distribution as

$$P\{T_{i,j^*} > t\} = P\{T_{i,k} > t, \forall k = 1, \dots, n\} = \prod_{k=1}^n e^{-w_{i,k}t} = e^{-W_i t}.$$

Hence  $\Delta t = T_{i,j^*}$  can be drawn from this distribution directly. We need just to draw two random numbers for each transition, instead of  $n$ , to simulate exactly the process. This reasoning is the basis of the *Gillespie algorithm*, which is routinely used to simulate stochastic processes.

**On the validity of the CLT.** Let  $X_i$  be i.i.d. random variables with  $E[X] = \mu$  and variance  $\sigma^2$ . Then the CLT says that the variable

$$Y_n = \frac{X_1 + X_2 + \dots + X_n - \mu n}{\sigma\sqrt{n}}$$

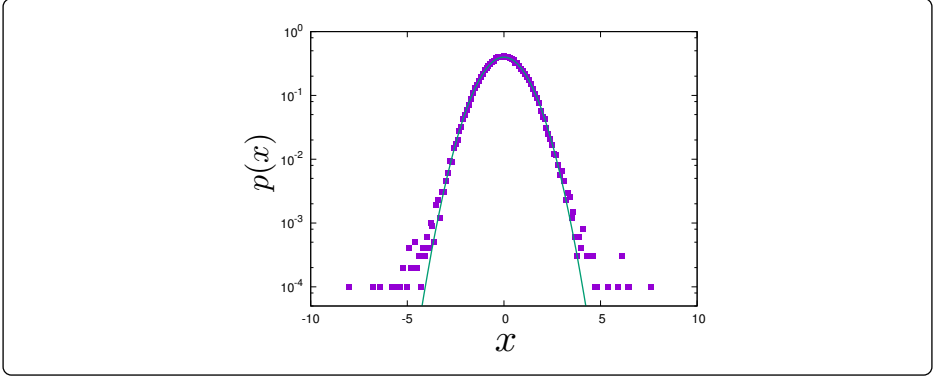
converges in distribution to a Gaussian variable with zero mean and unit variance. This means that  $p(x) = e^{-x^2/2}/\sqrt{2\pi}$  is a good approximation for the pdf  $p_n(x)$  of  $Y_n$ . Yet, for finite  $n$ , this is only true in an interval  $[-y_0, y_0]$  around the origin. How does the size of the interval  $y_0$  depends on  $n$ ?

Let us first discuss the case where all the moments of  $X_i$  are finite. Then

$$p_n(x) = \int_{-\infty}^{\infty} \frac{dq}{2\pi} e^{iqx + \psi_n(q)}$$

where

$$\psi_n(q) = \log E[e^{-iqY_n}] = -\frac{q^2}{2} + \frac{(-iq)^3}{3!\sqrt{n}} C_3 + \frac{(-iq)^4}{4!n} C_4 + \dots$$



**Figure 36.** The distribution of the variable  $Y_n$  deviates from the Gaussian in the tails. The pdf of  $Y_n$  is estimated with the empirical distribution of  $Y_n$  for  $10^5$  draws, for  $n = 10^2$  and  $p(x) \sim |x|^{-4}$  for  $|x| \gg 1$ .

and  $C_m$  is the  $m^{\text{th}}$  order cumulant of the variables  $X_i$ . In the integral defining  $p_n(x)$ , we can expand

$$e^{\psi_n(q)} = e^{-q^2/2} \left[ 1 + \frac{(-iq)^3}{3!\sqrt{n}} C_3 + \left( \frac{(-iq)^4}{4!} C_4 + \frac{(-iq)^6}{2(3!)^2} C_3^2 \right) \frac{1}{n} + \dots \right]$$

Now, each power of  $(-iq)$  in the integral acts as a derivative  $-\frac{d}{dx}$  taken outside the integral, therefore

$$\begin{aligned} p_n(x) &= \left[ 1 - \frac{C_3}{3!\sqrt{n}} \frac{d^3}{dx^3} + \left( \frac{C_4}{4!} \frac{d^4}{dx^4} + \frac{C_3^2}{2(3!)^2} \frac{d^6}{dx^6} \right) \frac{1}{n} + \dots \right] \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ &= \left[ 1 - \frac{C_3}{3!\sqrt{n}} H_3(x) + \left( \frac{C_4}{4!} H_4(x) + \frac{C_3^2}{2(3!)^2} H_6(x) \right) \frac{1}{n} + \dots \right] \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \end{aligned}$$

where we have used the relation  $\frac{d^m}{dx^m} e^{-x^2/2} = H_m(x) e^{-x^2/2}$  that defines the Hermite polynomial of degree  $m$ . For a given value of  $n$ , the approximation of the CLT is accurate as long as the correction terms are small. Since  $H_m(x) = x^m + \dots$ , this requires that  $|x|^3 \ll \sqrt{n}$  when  $C_3 \neq 0$ , i.e.  $|x| \ll y_0 \sim n^{1/6}$ . Notice that,  $n^{1/6} \simeq 2.15$  for  $n = 100$ , so one should be careful to apply the CLT to deviations of sums  $Y_n$  which are not small.

Things get better if  $C_3 = 0$  and  $C_4 \neq 0$ . Then the range of validity of the CLT is larger, i.e.  $y_0 \sim n^{1/4}$ . In general if the first non-zero cumulant is  $C_m$ , then the range of validity of the CLT depends on  $n$  as  $y_0 \sim n^{1/2-1/m}$ .

When instead the distribution of  $X_i$  has divergent moments, i.e. when  $p(x) \sim |x|^{-\alpha-1}$ , with  $\alpha > 2$ , we can argue in the following manner. Consider

the case where  $X_i > 0$  and let's focus on the right tail, for simplicity, i.e.  $p(x) \sim \alpha c x^{-\alpha-1}$  for  $x \gg 1$ . Clearly the event  $\{X_1 + \dots + X_n < x\}$  implies the event  $\{\max(X_1, \dots, X_n) < x\}$ , and hence

$$P\{X_1 + \dots + X_n < x\} \leq P\{\max(X_1, \dots, X_n) < x\}$$

assuming that we're in the range of validity of the CLT, and taking  $x = n\mu + \sqrt{n}\sigma y$ , we approximate the l.h.s. as

$$P\{X_1 + \dots + X_n < n\mu + \sqrt{n}\sigma y\} \simeq 1 - \frac{1}{\sqrt{2\pi}y} e^{-y^2/2}, \quad y \gg 1$$

whereas

$$\begin{aligned} P\{\max(X_1, \dots, X_n) < n\mu + \sqrt{n}\sigma y\} &\simeq [1 - c(\mu n)^{-\alpha}]^n \\ &\simeq e^{-c'n^{1-\alpha}} \simeq 1 - c'n^{1-\alpha} \end{aligned}$$

where we assumed  $\mu n \gg \sigma\sqrt{n}y$ , i.e.  $y \ll \sqrt{n}$ . Therefore the inequality above, after some straightforward manipulations, leads to  $y^2 \leq 2(\alpha - 1) \log n + \dots$  where  $\dots$  stands for subleading terms. Neglecting these, we arrive at

$$y \leq \sqrt{2(\alpha - 1) \log n}$$

This means that if  $y$  exceeds the value  $y_0 = \sqrt{2(\alpha - 1) \log n}$  then something in the above derivation necessarily goes wrong. The only real assumption that we made is that of the validity of the CLT. So we conclude that the CLT does not hold if  $y \gg \sqrt{2(\alpha - 1) \log n}$ . The interval in which the CLT holds, therefore, grows extremely slowly with  $n$  when the pfd of  $X$  has power law tails.

**Knowns and unknowns.** In many cases, we are interested in phenomena which are — or so we think — the result of an optimisation problem. For example, we think of the sequence  $\underline{s} = (s_1, \dots, s_n)$  of amino-acids of a protein as optimising a specific biological function in an organism. Yet in reality the optimisation may also involve many other variables  $\bar{s}$ , besides  $\underline{s}$ , which may not be observed or even known.<sup>18</sup> We may describe this situation as a generic optimisation problem of a function  $U(\bar{s})$  over a certain number of

<sup>18</sup>There are many examples of problems of this type. The choice of the city (i.e.  $\underline{s}$ ) in which John decides to live, does not only depend on the name  $\underline{s}$  of the city, but also on other factors ( $\bar{s}$ ) that enter in John's decision.

A plant selects its reproductive strategy depending on the environment where it lives. There



variables  $\vec{s} = (\underline{s}, \bar{s})$ , where only a fraction of the variables — the “knowns”  $\underline{s}$  — are observable, whereas the other variables  $\bar{s}$  — the “unknowns” — are unobservable. To what extent does the available knowledge allows us to predict the real behaviour of the system? How much should we know in order to be predictive?

Formally, we can define that part of the objective function that is known as  $u_{\underline{s}} = \mathbb{E} [U(\vec{s})|\underline{s}]$ , where the expected value is taken over the distribution that encodes all the knowledge available on  $U$ , for a given value of  $\underline{s}$ . We shall call  $u_{\underline{s}}$  *the model*, because it is the best possible description of the system on the basis of what is known. Accordingly, we can write

$$U(\vec{s}) = u_{\underline{s}} + v_{\bar{s}|\underline{s}} \quad (15.34)$$

where  $v_{\bar{s}|\underline{s}} = U(\vec{s}) - \mathbb{E} [U(\vec{s})|\underline{s}]$  is an unknown function of  $\bar{s}$  and  $\underline{s}$ . Because it is unknown, we assume it to be drawn randomly and independently for each  $\vec{s} = (\underline{s}, \bar{s})$  from a given distribution  $p(v)$ . The assumption that  $v_{\bar{s}|\underline{s}}$  are independent draws from  $p(v)$  is the simplest possible, but it also encodes a state of (almost) complete ignorance.<sup>19</sup> This is a very complex system, as the full specification of  $U(\vec{s})$  for each value of  $\underline{s}$  requires a number of parameters  $v_{\bar{s}|\underline{s}}$  that grows exponentially with the number of unknown variables  $\bar{s}$ .

The behaviour of the system is given by the solution

$$\vec{s}^* = (\underline{s}^*, \bar{s}^*) \equiv \arg \max_{\vec{s}} U(\vec{s}). \quad (15.35)$$

Notice that, since  $U(\vec{s})$  is a random function,  $\vec{s}^*$  and its observable component  $\underline{s}^*$  are random variables. The behaviour of the observable variables predicted by the model, on the other hand, is given by

$$\underline{s}_0 \equiv \arg \max_{\underline{s}} u_{\underline{s}}. \quad (15.36)$$

Therefore, the predictability of the model is quantified by the probability

$$p_{\underline{s}_0} = P\{\underline{s}^* = \underline{s}_0\} \quad (15.37)$$

are measurable characteristics e.g. of its flowers, that can be classified according to a discrete variables  $\underline{s}$ . The fitness of that species is optimised over a much broader set of variables  $\vec{s} = (\underline{s}, \bar{s})$  which include unobserved variables  $\bar{s}$ , that influence other traits of the phenotype.

We can think that Shakespeare, in writing Hamlet, chose a particular sentence  $\underline{s}$  in an optimal manner. Each sentence  $\underline{s}$  in the text has been chosen by Shakespeare, depending on the words  $\bar{s}$  that precede and follow it.

<sup>19</sup>I.e. if  $v_{\bar{s}|\underline{s}}$  were dependent and/or not identical, we should know how they depend and/or how they differ.

that the model reproduces the behaviour of the system. This probability can be derived for the following generic complex optimisation problem: we focus on the case where all the moments of  $v_{\bar{s}|\underline{s}}$  are finite and, without loss of generality, we take  $\underline{s} = (s_1, \dots, s_n)$  and  $\bar{s} = (s'_1, \dots, s'_m)$ , with the variables  $s_i, s'_i = \pm 1$  taking two values. If  $n$  and  $m$  were small, the problem would not be that complex. So we consider the limit where both  $n$  and  $m$  are very large (ideally  $n, m \rightarrow \infty$ ), with  $m = \mu n$ , which may be more appropriate for a complex system such as those discussed above.

For all  $\underline{s}$ , extreme value theory implies that

$$\max_{\bar{s}} v_{\bar{s}|\underline{s}} \cong a_m + \frac{\eta_{\underline{s}}}{\beta_m}, \quad (15.38)$$

where  $a_m$  is a constant,  $\beta_m$  depends on the tail behaviour of the distribution of  $v_{\bar{s}|\underline{s}}$  (see later) and, because of our assumption on  $v_{\bar{s}|\underline{s}}$ ,  $\eta_{\underline{s}}$  are i.i.d. Gumbel distributed, i.e.  $P\{\eta_{\underline{s}} < x\} = e^{-e^{-x}}$ . Therefore

$$P\{\underline{s}^* = \underline{s}\} = P\{\beta_m u_{\underline{s}} + \eta_{\underline{s}} \geq \beta_m u_{\underline{s}'} + \eta_{\underline{s}'}, \forall \underline{s}' \neq \underline{s}\} \quad (15.39)$$

$$\begin{aligned} &= \int_{-\infty}^{\infty} d\eta_{\underline{s}} e^{-\eta_{\underline{s}} - e^{-\eta_{\underline{s}}}} \prod_{\underline{s}' \neq \underline{s}} \int_{-\infty}^{\eta_{\underline{s}} + \beta_m(u_{\underline{s}} - u_{\underline{s}'})} d\eta_{\underline{s}'} e^{-\eta_{\underline{s}'} - e^{-\eta_{\underline{s}'}}} \\ &= \frac{1}{Z(\beta_m)} e^{\beta_m u_{\underline{s}}}, \quad Z(\beta_m) = \sum_{\underline{s}'} e^{\beta_m u_{\underline{s}'}} \end{aligned} \quad (15.40)$$

where, in the second line, we used the change of variables  $z_{\underline{s}} = e^{-\eta_{\underline{s}}}$  to ease the calculation of the integral.<sup>20</sup> Hence limit theorems on extremes dictate

<sup>20</sup>In the case where  $p(v)$  has a tail behaviour

$$p(v) \sim v^{-\gamma-1}$$

for  $v \rightarrow \infty$ , an analogous calculation leads to

$$\begin{aligned} P\{\underline{s}^* = \underline{s}\} &= \int_0^{\infty} dt e^{-t \left( 1 + \sum_{\underline{s}' \neq \underline{s}} (1 + \beta_m(u_{\underline{s}} - u_{\underline{s}'})) t^{1/\gamma} \right)^{\gamma}} \\ &= \int_0^{\infty} dt e^{-t \sum_{\underline{s}' \neq \underline{s}} (1 + \beta_m(u_{\underline{s}} - u_{\underline{s}'})) t^{1/\gamma}}. \end{aligned}$$

Given that  $\beta_m = \beta_0 e^{-m/\gamma} \ll 1$ , it is possible to expand the argument of the exponential, leading to

$$\begin{aligned} P\{\underline{s}^* = \underline{s}\} &= \int_0^{\infty} dt e^{-2^n t [1 - \gamma \beta_m (u_{\underline{s}} - \bar{u}) t^{1/\gamma} + \dots]} \\ &\simeq \frac{1}{2^n} \left[ 1 + c 2^{-\frac{n+m}{\gamma}} (u_{\underline{s}} - \bar{u}) + \dots \right] \end{aligned}$$

the form of  $P\{\underline{s}^* = \underline{s}\}$ , that for the class of models for which  $v_{\underline{s}|\underline{s}}$  has all finite moments, coincides with Eq. (15.40). This distribution is known under the name of *Gibbs-Boltzmann distribution* in statistical physics and of *Logit model* in statistics, choice theory and economics. We will see that there are other ways to derive the same result (Eq. (15.40)) from an assumption of maximal ignorance. Notice that when  $\beta_m \rightarrow \infty$  the distribution  $P\{\underline{s}_0 = \underline{s}^*\}$  gets more and more peaked around the maximum  $\underline{s}_0$  of  $u_{\underline{s}}$ , whereas when  $\beta_m \rightarrow 0$  it gets more and more uniform on all  $2^n$  states.

### Exercise 15.8

If  $p(v) \sim (\omega - v)^{\gamma-1}$  for  $v < \omega$  and  $p(v) = 0$  for  $v > \omega$ , then  $P\{\underline{s}^* = \underline{s}\} \simeq \delta_{\underline{s}^* = \underline{s}_0}$  for  $n, m$  large. Show it.

Let us specialise our discussion to the specific case of  $p(v) = \gamma v^{\gamma-1} e^{-v^\gamma}$  (as in Eq. (15.27) with  $\gamma \leftrightarrow \nu$  and  $v \leftrightarrow x$ ). Then Eq. (15.29) with  $n \mapsto 2^m$ , gives

$$\beta_m = \gamma [m \log 2]^{1-1/\gamma} \quad (15.41)$$

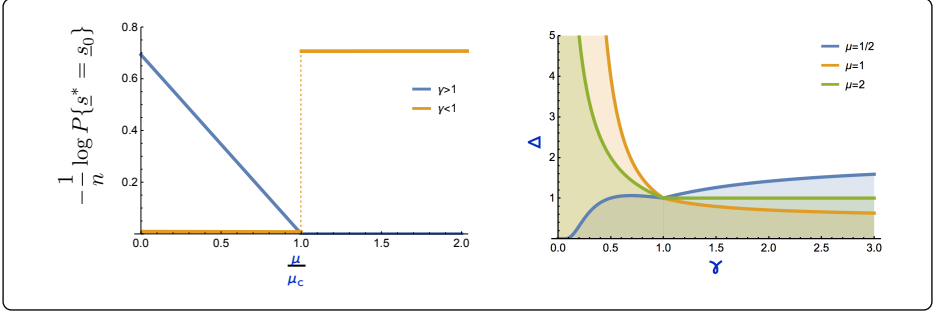
One may naïvely expect that the predictability of the model gets worse, i.e. that  $P\{\underline{s}_0 = \underline{s}^*\}$  decreases, when the number  $m$  of unknown variables increases. This is only true for  $\gamma < 1$ , as indeed  $\beta_m$  decreases as the number  $m$  of unknown unknowns increases in this case. When  $p(v)$  decays faster than exponential ( $\gamma > 1$ ), which includes the case of Gaussian variables,  $\beta_m$  diverges with the number  $m$  of unknowns. For  $\gamma > 1$ , if the number  $n$  of observed variables stays finite, we expect that  $P\{\underline{s}_0 = \underline{s}^*\} \rightarrow 1$  in the limit  $m \rightarrow \infty$  of an infinite number of unknown variables. For  $\gamma > 1$ , the more we don't know the better we can predict.

**An ensemble of random optimization problems.** We can make further progress if we assume that also  $u_{\underline{s}}$  is drawn from a distribution  $p(u)$  with the same behaviour,<sup>21</sup> i.e.  $P\{u_{\underline{s}} > u\} = e^{-(u/\Delta)^\gamma}$ . The number of states with  $u_{\underline{s}} > u$  is given by

$$\left| \left\{ \underline{s} : u_{\underline{s}} > u \right\} \right| \sim 2^n e^{-(u/\Delta)^\gamma}, \quad u > 0. \quad (15.42)$$

with  $c = \gamma \Gamma\left(2 + \frac{1}{\gamma}\right) \beta_0$  a constant. The integral is evaluated first changing variables to  $x = 2^n t$  and then expanding the exponential to leading order. Therefore, for  $p(v) \sim v^{-\gamma-1}$  no prediction is possible.

<sup>21</sup>This problem is similar to the *Random Energy Model* [25] studied in statistical mechanics as a toy model for disordered systems, where the energy  $E_{\underline{s}} = E_0 - \sqrt{n} u_{\underline{s}}$  is drawn independently from a Gaussian distribution with mean  $E_0$  and variance  $n$ .



**Figure 37.** Left: phase transition as a function of the ratio  $\mu = m/n$  between the number of unknown and known variables. The vertical axis reports the value of the exponential rate with which the probability  $P\{s_0 = s^*\} \rightarrow 0$  vanishes. Right: phase diagram in the  $(\gamma, \Delta)$  plane for different values of  $\mu$ . Notice that the upper region where  $P\{s_0 = s^*\}$  is finite expands as  $\mu$  increases for  $\gamma > 1$ , whereas it shrinks as  $\mu$  increases for  $\gamma < 1$ .

The parameter  $\Delta$  provides a scale of the known part of the function  $U$  with respect to the unknown part, i.e.  $u_{\underline{s}}/v_{\bar{s}|\underline{s}} \sim \Delta$ . For  $\Delta \gg 1$  we expect the optimisation to depend weakly on the variables  $\bar{s}$ , and to be dominated by the term  $u_{\underline{s}}$ . Hence we expect that  $P\{s_0 = s^*\} \rightarrow 1$  for large  $\Delta$ . The largest value of  $u_{\underline{s}}$  predicted by the limit theorems for extremes, is given by

$$u_0 = \max_{\underline{s}} u_{\underline{s}} \simeq \Delta(n \log 2)^{1/\gamma}.$$

Using this and Eq. (15.41), the probability that the model predicts the right outcome is

$$P\{s^* = s_0\} = \frac{1}{Z} e^{\beta_m u_0} \simeq \frac{1}{Z} e^{n(\log 2) \Delta \gamma \mu^{1-1/\gamma}}$$

is also exponential in  $n$ . The partition sum  $Z$  can be computed within a saddle point approximation. When the sum is dominated by the state  $s_0$ , then  $P\{s^* = s_0\}$  attains a finite value, otherwise  $P\{s^* = s_0\} \rightarrow 0$  as  $n \rightarrow \infty$ . The behaviour of  $P\{s_0 = s^*\}$  as a function of the parameters  $\gamma$ ,  $\Delta$  and  $\mu = m/n$  in the limit  $n \rightarrow \infty$ , has been studied in ref. [26]. We refer to this paper and report the main results in Figure 37.

The probability  $P\{s_0 = s^*\}$  features a phase transition between a phase where it is exponentially small in  $n$ , and one where it is finite. When  $\gamma > 1$ ,  $P\{s_0 = s^*\}$  is exponentially small in  $n$  for

$$\frac{m}{n} = \mu < \mu_c = \Delta^{-\gamma/(\gamma-1)}, \quad (\gamma > 1)$$

whereas for  $\mu > \mu_c$  it is finite. As shown in Figure 37 (left), the transition is continuous. For  $\gamma < 1$ , instead, the transition is reversed and it becomes discontinuous. More precisely,  $P\{s_0 = \underline{s}^*\} \rightarrow 0$  as  $n \rightarrow \infty$  when

$$\frac{m}{n} = \mu > \mu_c = (\gamma\Delta)^{\gamma/(1-\gamma)} \quad (\gamma < 1). \quad (15.43)$$

and  $P\{s_0 = \underline{s}^*\}$  suddenly jumps to finite values for  $\mu < \mu_c$ . The *phase diagram* of the system in the  $(\gamma, \Delta)$  plane is shown in Figure 37 (right).

The case of an exponential distribution ( $\gamma = 1$ ) is very peculiar, because then the transition point is independent of  $\mu$ . This invariance is argued to be a peculiar property of systems that learns, in the paper cited above.



# Chapter 16

## Information theory

All knowledge degenerates into probability.  
(David Hume, 1739)

How can we quantify information?<sup>1</sup> Let us take a specific example: Alice (A) is in a state of ignorance about a certain variable  $X$  which is known to Bob (B). She anticipates that the answer  $X \in \mathcal{X}$  can be one of  $n = |\mathcal{X}|$  possible ones. One way to quantify the information content of  $X$  is to count the number of binary questions (yes/no) that A needs to pose to B in order to know the answer  $X$ . Indeed, A's uncertainty will be dispelled after she hears the answers because she will know what  $X$  is. Therefore, the number  $N_Q$  of binary questions needed to dispel A's ignorance is an operative definition of the information content of  $X$ , and it is measured in *bits*.<sup>2</sup>

Take for example the case  $\mathcal{X} = \{a, b, c, d\}$ . Then A may ask a first question

$Q_1$ : is  $X \in \{a, b\}$  or not?

and depending on the answer, A may ask

$Q_2$ : if  $X \in \{a, b\}$  is  $X = a$  or not?

Else, if  $X \notin \{a, b\}$  is  $X = c$  or not?

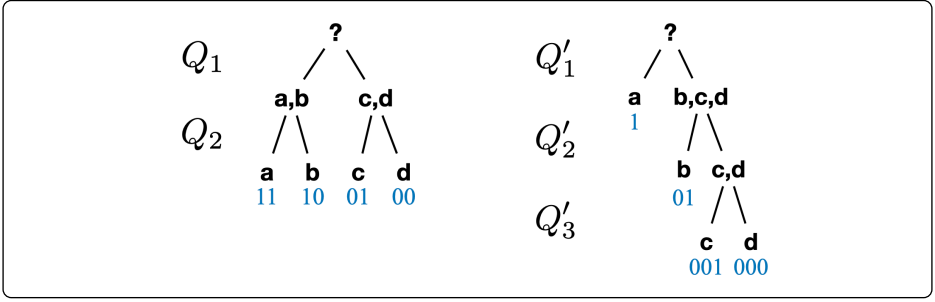
The answers to these two questions reveal the correct outcome  $X$ . Hence the information is  $N_Q = 2$  bits. Yet there are many other ways in which A could ask questions, and hence  $N_Q$  could vary accordingly.

For example A can modify her questions as follows:

---

<sup>1</sup>This chapter heavily draws from COVER, Chapter 2, and Chapter 4 of [27].

<sup>2</sup>A bit is a variable that takes two values, 0 (for no) or 1 (for yes).



**Figure 38.** Different ways of asking binary questions.

$Q'_1$ : is  $X = a$  or not?

only if  $X \neq a$  A will need to pose a further question. Then she may ask:

$Q'_2$ : is  $X = b$  or not?

Only if the result is no, she will need to ask

$Q'_3$ : is  $X = c$  or not?

in which case the number of binary questions can be  $N_Q(a) = 1$ ,  $N_Q(b) = 2$  or  $N_Q(c) = N_Q(d) = 3$ , depending on the value of  $X$ . Indeed,  $N_Q(X)$  is a random variable, because it is a function of  $X$ .

Formally, the state of uncertainty of A is encoded in the probability distribution  $P\{X = x\} = p_x$ , of  $X$ . We're looking for a measure of information content of  $X$  that can quantify the uncertainty of A *before* the questions are posed and the answers are heard. Therefore, it makes sense to define a measure of information content as the expected number  $\mathbb{E}[N_Q]$  of binary questions that are needed to elicit the value of  $X$ .

The expected value

$$\mathbb{E}[N_Q] = \sum_{x \in \mathcal{X}} p_x N_Q(x)$$

depends on the distribution  $p_x$ , that we assume is known to A, and on the way in which the answers are posed. For example, if A didn't know  $p_x$ , there is nothing that would distinguish the different outcomes, e.g.  $X = a$  from  $X = b$ , so there is nothing that suggests that  $p_a$  should be smaller or larger than  $p_b$ . Hence, she would have to assume that  $p_x = 1/4$  for all  $x$ . This distribution indeed encodes a state of maximal ignorance, as we shall see. Then asking questions  $(Q_1, Q_2)$  yields  $\mathbb{E}[N_Q] = 2$  whereas formulating questions  $(Q'_1, Q'_2, Q'_3)$



leads to a larger value of  $\mathbb{E}[N_Q] = 9/4$ . If instead  $p_a = 1/2, p_b = 1/4$  and  $p_c = p_d = 1/8$ , then again  $\mathbb{E}[N_Q] = 2$ , but

$$\mathbb{E}[N_Q] = p_a \cdot 1 + p_b \cdot 2 + p_c \cdot 3 + p_d \cdot 3 = \frac{7}{4}. \quad (16.1)$$

The optimal way of answering questions is different in the two cases. The minimal expected number of binary questions that A needs to pose to elicit  $X$  is a measure of her irreducible ignorance about  $X$ . Hence, we provisionally define

The information content  $H[X]$  of a random variable  $X$  is the *minimal* expected number of binary questions needed to elicit the value of  $X$ ,

$$H[X] = \min_Q \mathbb{E}[N_Q] \quad (16.2)$$

where the expected value is taken with respect to the distribution  $P\{X = x\} = p_x$  that defines the state of knowledge on  $X$ , and the minimum is taken over all possible ways of posing yes/no questions.

Note that the information content

$$H : X \rightarrow \mathbb{R}$$

is a *functional* that associates a real number  $H[X]$  to a function

$$X : \Omega \rightarrow \mathbb{R}.$$

This is why we use square brackets in  $H[\cdot]$ .

The way in which Alice poses question associates to each values of  $X$  a strings of binary variables that we can take to be 1 for yes and 0 for no. Such a transformation between values of  $X$  and strings of bits is called a *code*. Imagine that Alice asks Bob the same question many times (e.g. what's the weather today?) and that they communicate through a binary channel, i.e. a device that allow Bob and Alice to send either a 0 or a 1 to the other end at any time. Alice and Bob might be interested in finding the code which makes them exchange the shortest possible strings of bits. This problem is the same as the problem of finding the best way to ask questions.

Indeed, each protocol  $Q$  for asking questions corresponds to a scheme to encode the possible answers  $X$ . For example, the protocol  $Q$  above would correspond to the code

$$a \rightarrow 00; b \rightarrow 01; c \rightarrow 10; d \rightarrow 11.$$

The bit strings associated to a given value of  $X$  is called its *codeword*.<sup>3</sup> This code will require 2 bits for each answer transmitted from B to A. Protocol  $Q'$  corresponds to a different association of values of  $X$  to codewords, i.e.

$$a \rightarrow 1; b \rightarrow 01; c \rightarrow 001; d \rightarrow 000$$

Notice that each codeword has length  $\ell_Q(X) = N_Q(X)$  which is equal to the number of binary questions needed to elicit  $X$  under protocol  $Q$ .

Therefore the problem of finding the code that is *expected* to use the least number of bits (i.e. that minimises  $\mathbb{E}[\ell_Q]$ ) is exactly the same as the problem of finding the best way to pose questions. The fact that these two apparently different problems — A posing questions to B optimally and B transmitting answers to A efficiently — have the same solution, is interesting.

Note also that the optimal way  $Q^*$  of posing questions, and hence  $H[X]$ , depends only on the probabilities  $p_x$ , and not on what  $X$  is.<sup>4</sup> In particular, if an answer  $x$  is more likely than  $x'$ , then it is natural that<sup>5</sup>  $\ell_{Q^*}(x) \leq \ell_{Q^*}(x')$ . For example, the knowledge of  $p_x$  in the example above, carries some information on the answer, which can be quantified in the difference between  $\mathbb{E}[N_Q]$  in the two cases, and is  $1/4$  of a bit in that case.

## 16.1 Shannon entropy and Shannon theorem

The minimal number of binary questions needed to elicit  $X$ , or equivalently the expected length of the optimal code for  $X$ , is given by the *Shannon entropy*

$$H[X] = \mathbb{E}[\log_2 1/p_X] = - \sum_{x \in \mathcal{X}} p_x \log_2 p_x \quad (16.3)$$

of the random variable  $X$ , that we shall simply call *entropy*, henceforth. The entropy depends on the distribution  $p_x$ , and we will equivalently denote it as  $\mathcal{H}[p]$ , when referring to it as a functional of the probability distribution  $p_x$ .

It is easy to check that this is the correct answer in the examples above, where codewords have length exactly equal to  $\log_2 1/p_x$ , but one can argue that Eq. (16.3) works for all *discrete* random variables  $X$ , provided that we consider *messages*  $\underline{X} = (X_1, \dots, X_n)$  where each of the  $n$  characters  $X_i \in \mathcal{X}$ , are drawn i.i.d. from the distribution  $p_x$ . Then, in the limit  $n \rightarrow \infty$ , almost surely,

<sup>3</sup>In coding theory jargon,  $X$  are called *words*.

<sup>4</sup> $X$  could be football teams in the Premier League or species of bird on some island. As long as the probabilities  $p_x$  are the same, the information content is the same.

<sup>5</sup>Think of the first binary question you would ask to know which team won the last Premier League championship.

we need at most  $H[X]$  bits per character. This result, that goes under the name of *Shannon theorem*, is a direct consequence of the Asymptotic Equipartition Property. The idea of the proof is simple. Remember that the Asymptotic Equipartition Property ensures us that, for any  $\epsilon > 0$ , a message  $\underline{X}$  belongs to the  $\epsilon$ -typical set

$$A_n^{(\epsilon)} = \left\{ \underline{X} : \left| \frac{1}{n} \log P(\underline{X}) + H[X] \right| < \epsilon \right\}$$

almost surely, as  $n \rightarrow \infty$ . Imagine that Alice and Bob assign to all messages  $\underline{X} \in A_n^{(\epsilon)}$  a different integer  $Q(\underline{X})$  from one to  $|A_n^{(\epsilon)}|$ , and to messages  $\underline{X} \notin A_n^{(\epsilon)}$  integers  $Q(\underline{X})$  larger than  $|A_n^{(\epsilon)}|$ . Then each message will require a codeword of length  $\ell_Q(\underline{X}) = \log_2 Q(\underline{X})$ , which is given by the binary representation of  $Q(\underline{X})$ . Then, almost surely, Alice and Bob will need less than

$$\frac{1}{n} \max_{\underline{X} \in A_n^{(\epsilon)}} \log_2 Q(\underline{X}) = \frac{1}{n} \log_2 |A_n^{(\epsilon)}|$$

bits per character, as  $n \rightarrow \infty$ . In this limit, the Asymptotic Equipartition Property also implies that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 |A_n^{(\epsilon)}| = H[X].$$

because  $|A_n^{(\epsilon)}| \sim 2^{nH[X]}$ . Therefore, at most  $H[X]$  bits per character  $X$  need to be used to transmit the message, almost surely.

### Exercise 16.1

The Rényi entropy is defined as

$$H_a[X] = \frac{1}{1-a} \log \sum_{x \in \mathcal{X}} p_x^a$$

with  $a > 0$ . Show that  $H_a[X]$  is a generalisation of the Shannon entropy, which is recovered in the limit  $a \rightarrow 1$ . Show that if  $X$  and  $Y$  are independent

$$H_a[X, Y] = H_a[X] + H_a[Y].$$

Show that, if the conditional Rényi entropy is defined as

$$H_a[X|Y] = \frac{1}{1-a} \mathbb{E} \left[ \log \sum_{x \in \mathcal{X}} p^a(x|Y) \right]$$

then the chain rule

$$H_a[X, Y] = H_a[Y] + H_a[X|Y]$$

holds only for  $a \rightarrow 1$ .

### Exercise 16.2

Tsallis entropy is defined as

$$H_q[X] = \frac{1}{1-q} \left( 1 - \mathbb{E} \left[ p_X^{q-1} \right] \right).$$

Show that i)  $H_q[X]$  reduces to the Shannon entropy for  $q \rightarrow 1$ , and that ii)  $H_q$  is not additive for  $q \neq 1$ , i.e. if  $X$  and  $Y$  are independent random variables, then

$$H_q[X, Y] = H_q[X] + H_q[Y] + (1-q)H_q[X]H_q[Y].$$

There are other ways to derive this result. For example, the same result can be obtained observing that the optimal number of bits needed to code  $X$ , should be a function  $f(p_X)$  of  $p_X$ . Then the expected number of bits needed has to be of the form

$$H[X] = \mathbb{E} [f(p_X)] = \sum_{x \in \mathcal{X}} p_x f(p_x).$$

If  $X = (Y, Z)$  where  $Y \in \mathcal{X}_Y$  and  $Z \in \mathcal{X}_Z$  are independent random variables, then  $H[X] = H[Y] + H[Z]$ , because knowing  $Y$  does not give any clue on what  $Z$  could be. Hence

$$\sum_{Y \in \mathcal{X}_Y, Z \in \mathcal{X}_Z} p_y p_z f(p_y p_z) = \sum_{Y \in \mathcal{X}_Y, Z \in \mathcal{X}_Z} p_y p_z [f(p_y) + f(p_z)]$$

for any  $p_y$  and  $p_z$ . Therefore  $f(p_y p_z) = f(p_y) + f(p_z)$ , which means that  $f(p) = a \log p$ . If in addition we want to measure information in bits, then  $f(1/2) = 1$ , i.e.  $f(p) = -\log_2 p$ . The entropy quantifies how much B's reply can be surprising for A. Indeed if both A and B knows that  $p_x = 1$  if  $x = a$  and  $p_x = 0$  for all  $x \neq a$ , then B's reply cannot be surprising. Actually A doesn't even need to ask because both of them know that  $X = a$ . So no bit needs to be exchanged and, accordingly  $H[X] = 0$ . As we said,  $H[X]$  quantifies the uncertainty of Alice about Bob's answer *before* she hears the answer. After

she hears the answer, she knows that one answer occurs with probability one and the others with probability zero, i.e.  $H = 0$ . Then  $H$  measures how much Alice has decreased her degree of uncertainty.

Conversely, the entropy is maximal when  $X$  is maximally uncertain:  $p_x = 1/|\mathcal{X}|$ . Accordingly

$$0 \leq H[X] \leq \log |\mathcal{X}|.$$

The entropy can be generalised to any number of random variables  $X_1, \dots, X_n$  in a straightforward fashion, i.e.

$$H[X_1, \dots, X_n] = -\mathbb{E} [\log_2 P\{X_1, \dots, X_n\}].$$

Likewise, we can define the conditional entropy

$$H[X|Y] = -\mathbb{E} [\log_2 P\{X|Y\}] = - \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log_2 p(x|y)$$

as the entropy of the conditional distribution  $p(x|y)$ , averaged over  $y$ . The law of conditional probability imply that

$$H(X|Y) = H(X, Y) - H(Y). \quad (16.4)$$

In words, the conditional entropy is the expected reduction of the uncertainty about  $X$  if  $Y$  where known. Put differently,  $H(X|Y)$  quantifies the residual uncertainty on  $X$  that Alice expects to reach if she asks Bob about  $Y$ . In particular, for a sequence of random variables  $X_1, \dots, X_n$ , we have that

$$H[X_1, \dots, X_n] = \sum_{m=2}^n H[X_m|X_{m-1}, \dots, X_1] + H[X_1].$$

If the sequence is a Markov chain, then  $H[X_m|X_{m-1}, \dots, X_1] = H[X_m|X_{m-1}]$ , because  $X_m$  given  $X_{m-1}$  is independent of  $X_k$ , for all  $k < m-1$ . If the transition probability  $p_{i,j} = P\{X_n = j|X_{n-1} = i\}$  does not depend on  $n$ , and if the Markov chain is irreducible, then

$$H[X_2|X_1] = -\mathbb{E} [\log p_{X_1, X_2}]$$

is called the *entropy rate*, because  $H[X_1, \dots, X_n]/n \rightarrow H[X_2|X_1]$  as  $n \rightarrow \infty$ .

### Exercise 16.3

Derive Eq. (16.4).

### 16.1.1 Entropy for continuous variables

The generalisation<sup>6</sup> of the concept of entropy to continuous variables is problematic. Indeed, imagine that Alice asks to Bob what is the area  $X$  of a unit circle. She will need to ask an infinite number of binary questions in order to know that  $X = \pi$  exactly, because an irrational number is represented by an infinite number of bits. This does not match with the straightforward generalisation of Eq. (16.3)

$$h[X] = E[\log_2 1/p(X)] = - \int dx p(x) \log_2 p(x) \quad (16.5)$$

which is finite, barring pathological cases. Furthermore, Eq. (16.5) seems problematic, since you may get negative numbers! So what is the meaning of  $h[X]$ ?

#### Exercise 16.4

Compute  $h[X]$  in Eq. (16.5) for  $p(x) = 1/[x(\log x)^2]$  for  $x \geq e$  and  $p(x) = 0$  for  $x < e$ .

#### Exercise 16.5

Check that  $h[X] = -3$  for a uniform random variable  $X \in [0, 1/8]$ .

Coming back to Alice and Bob, Alice may be happy to know  $X$  to a pre-assigned precision  $\Delta$ . So imagine that Alice “quantizes” the random variable  $X$  into the random variable  $X^\Delta$  that takes values  $x_i$  which are defined as<sup>7</sup>

$$p(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} dx p(x), \quad (16.6)$$

for all integer  $i = 0, \pm 1, \pm 2, \dots$ . With this definition, the distribution of  $X^\Delta$  is defined as  $P\{X^\Delta = x_i\} = p(x_i)\Delta$ , which is the probability that  $X \in [i\Delta, (i+1)\Delta)$ . She can now give a precise estimate of the information content of Bob’s

<sup>6</sup>See Chapter 8 of COVER.

<sup>7</sup>Because of the mean value theorem for integrals,  $x_i \in [i\Delta, (i+1)\Delta]$  is inside the interval of integration.

answer, which is the entropy  $H[X^\Delta]$  of  $X^\Delta$ . For  $\Delta \ll 1$ , this can be expressed as

$$\begin{aligned} H[X^\Delta] &= - \sum_i p(x_i) \Delta \log_2 [p(x_i) \Delta] \\ &= - \sum_i \int_{i\Delta}^{(i+1)\Delta} dx p(x) \log_2 p(x_i) - \log_2 \Delta \simeq h[X] - \log_2 \Delta \end{aligned} \quad (16.7)$$

where the approximation gets more and more precise as  $\Delta \rightarrow 0$ . Here  $h[X]$  is defined in Eq. (16.5), and it is called *differential entropy*. Its meaning is that  $h[X] - \log_2 \Delta$  is the expected number of bits needed to specify  $X$  to a precision  $\Delta$ , for  $\Delta \rightarrow 0$ . The fact that  $h[X]$  may not be positive is not a problem. For example, a uniform random variable  $X \in [0, a]$  has  $h[X] = \log_2 a$  which is negative if  $a < 1$ . If  $a = 1/8$  and you want to determine  $X$  up to the  $n^{\text{th}}$  binary digit (i.e.  $\Delta = 2^{-n}$ ), you will need  $n - 3$  bits, because the first three bits will be zero anyhow.

One property of the entropy that we used, is that  $H[X]$  does not actually depend on what values  $X$  takes. It only depends on the value of the probabilities  $p_x = P\{X = x\}$ . In particular, if we do a bijective transformation  $X \rightarrow Y = f(X)$  — i.e. such that to every possible value of  $X$  there corresponds one and only one value of  $Y$  — then  $H[X] = H[Y]$ .

This is not true for the differential entropy, because even when  $f(x)$  is monotonous — and hence to every  $X$  there correspond one and only one  $Y = f(X)$  — the pdf transforms as  $p_Y(y) = p_X(x)/|f'(x)|_{x=f^{-1}(y)}$ . Therefore

$$h[Y] = h[X] + \mathbb{E} [\log_2 |f'(X)|] . \quad (16.8)$$

Hence, the differential entropy is not reparametrization invariant. A simple application of this is that, if  $a$  is a constant, then  $h[X + a] = h[X]$  and  $h[aX] = h[X] + \log_2 |a|$ .

### Exercise 16.6

Compute the differential entropy for a Gaussian with mean  $\mu$  and variance  $\sigma^2$ , for an exponential distribution  $p(x) = ae^{-ax}$ ,  $a, x > 0$ , and for a multi-dimensional Gaussian with mean  $\vec{\mu}$  and covariance  $\text{Cov}[X_i, X_j] = A_{i,j}$ .

## 16.1.2 Relative entropy

Imagine now that A has a wrong estimate  $q_x$  of the probability  $p_x$  of B's answers  $x$ . How much this impacts on the efficiency of the questions she's going to ask?

Given  $q$ , A is going to effectively encode B's answers in such a way that answer  $x$  will require  $\log_2 1/q_x$  bits, so the number of questions she will ask, on average, is

$$\mathbb{E} \left[ \log_2 \frac{1}{q} \right] = \sum_{x \in \mathcal{X}} p_x \log_2 \frac{1}{q_x}$$

the difference between this and the most efficient way of asking questions, which requires  $\mathcal{H}[p]$  bits, is

$$D_{KL}[p\|q] = \sum_{x \in \mathcal{X}} p_x \log_2 \frac{p_x}{q_x}$$

which is known as the *Kullback-Leibler divergence* or *relative entropy*. It tells us how costly is the error in the estimate of probabilities, in bits. In this sense,  $D_{KL}$  is a measure of how “far” Alice is from the true distribution. This is why  $D_{KL}$  is often considered as a distance, though it is not symmetric and it does not satisfy the triangle inequality.<sup>8</sup>

### Exercise 16.7

A coin can either be fair with  $P\{\text{head}\} = P\{\text{tail}\} = 1/2$ , or biased, with  $P\{\text{head}\} = p$  and  $P\{\text{tail}\} = 1 - p$ . Show that it is worse to assume that the coin is biased when it is not, than to assume that it is fair when it is biased.

Though it is not evident  $D_{KL}[p\|q] \geq 0$  and it vanishes only for  $q = p$ . The way to prove it, is to use the convexity of the logarithm  $\log_2 x \leq (x - 1)/\log 2$  in the definition of  $D_{KL}$ , i.e.

$$D_{KL}[p\|q] = - \sum_{x \in \mathcal{X}} p_x \log_2 \frac{q_x}{p_x} \quad (16.9)$$

$$\geq - \frac{1}{\log 2} \sum_{x \in \mathcal{X}} p_x \left[ \frac{q_x}{p_x} - 1 \right] = 0 \quad (16.10)$$

because of normalisation of  $p_x$  and  $q_x$ .

The Kullback-Leibler divergence (or relative entropy) generalises to continuous variables as

$$D_{KL}[p\|q] = \int dx p(x) \log \frac{p(x)}{q(x)} \quad (16.11)$$

<sup>8</sup>See Theorem 11.6.1 in COVER for an example where  $D_{KL}(p\|q)$  satisfies the opposite of the triangle inequality.



Contrary to the differential entropy, the relative entropy is reparametrization invariant. If  $p$  and  $q$  represent two possible distributions for the random variable  $X$ , their divergence remains the same if one changes parametrization  $Y = f(X)$ . As for discrete variables, it is easy to see that  $D_{KL}[p\|q] \geq 0$  with equality holding only if  $p = q$ .

### Exercise 16.8

Show that the Kullback-Leibler divergence is invariant under changes of variables.

### 16.1.3 Mutual information

Imagine you have two random variables  $X \in \mathcal{X}$  and  $Y \in \mathcal{Y}$  with joint distribution  $p(x, y)$  and marginals  $p(x)$  and  $p(y)$ .<sup>9</sup> One way to quantify their mutual dependence is to compute how much information is lost by assuming that they are independent. This is given by

$$I[X, Y] = D_{KL}[p(x, y)\|p(x)p(y)] \quad (16.12)$$

$$= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (16.13)$$

$$= H[X] + H[Y] - H[X, Y] \quad (16.14)$$

and it is called the *mutual information* between  $X$  and  $Y$ . The last equality, which follows from simple algebra, with the positivity of  $D_{KL}$  implies that  $H[X, Y] \leq H[X] + H[Y]$ . In other words, *the state of maximal ignorance about two random variables  $X$  and  $Y$  corresponds to the case where they are independent*.

In the same way, one can define the mutual information  $I[X, Y]$  between continuous variables as

$$I[X, Y] = D_{KL}[p(x, y)\|p(x)p(y)] = h[X] + h[Y] - h[X, Y] \quad (16.15)$$

where

$$p(x) = \int dy p(x, y), \quad p(y) = \int dx p(x, y),$$

are the marginal distributions. This implies that  $I[X, Y] \geq 0$  with equality if and only if  $X$  and  $Y$  are independent. So the mutual information provides a universal measure of statistical dependence. It is universal also because, the

<sup>9</sup>The abuse of the symbol  $p(\cdot)$  follows the notation of COVER. It should be understood that  $p(x)$  and  $p(y)$  are different functions of their arguments.

mutual information is invariant under any transformation  $(X, Y) \rightarrow (U, V)$  of the random variables, where  $U = f(X)$  and  $V = g(Y)$  with  $f(x)$  and  $g(y)$  monotonous functions. These transformations changes the “shape” of the distributions of the two variables, but leaves their statistical dependence invariant. This invariance becomes manifest if we apply the transformation  $f(x) = P\{X \leq x\}$  and  $g(y) = P\{Y \leq y\}$  which transforms  $X$  and  $Y$  into two uniform random variables  $U$  and  $V$ . The mutual information can then be expressed as

$$I[X, Y] = \int_0^1 du \int_0^1 dv c(u, v) \log_2 c(u, v), \quad (16.16)$$

where

$$c(u, v) = \frac{\partial^2}{\partial u \partial v} C(u, v).$$

and the function  $C(u, v)$  is the joint cumulative distribution of  $U$  and  $V$ , defined as

$$P\{X \leq x, Y \leq y\} = C(P\{X \leq x\}, P\{Y \leq y\}). \quad (16.17)$$

The function  $C(u, v)$  is called the *copula function* of the two random variables  $X$  and  $Y$ .<sup>10</sup>

### Exercise 16.9

Prove Eqs. (16.16) and (16.17).

In order to illustrate the meaning of  $I$  consider the following problem. We are interested in estimating a random variable  $X$  of which at present we know the distribution  $p(x)$ , and the corresponding entropy  $H[X]$  which quantifies our state of uncertainty about  $X$ . You can think of  $X$  as a parameter of a theory of a given system.<sup>11</sup> Now we have the possibility to perform an experiment,

<sup>10</sup>Eqs. (16.16) and (16.17). suggest an easy way to check whether two variables are dependent or not, based on a sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  of  $n$  joint observations. Let  $U(x)$  and  $V(y)$  be the fraction of points for which  $X_i \leq x$  and  $Y_j \leq y$ , respectively. Plot the points  $(U(X_i), V(Y_i))$  in the  $(u, v)$  plane. If  $X$  and  $Y$  are independent, the  $n$  points should be uniformly distributed in the unit square  $[0, 1]^2$ . Statistical dependence is spotted by the clustering of points in some region. This plot reveals not only whether  $X$  and  $Y$  are dependent or not, but also how they depend on each other. For example a monotonous dependence (e.g. if  $X$  increases  $Y$  tends to increase or decrease) corresponds to points clustering on one of the diagonals of the square. This is the kind of dependence which is usually quantified by covariance measures. Yet there are many other possibilities of how  $X$  and  $Y$  can depend on each other, some of which may not be detectable by covariance.

<sup>11</sup> $I[X, Y]$  is the reduction of Alice's uncertainty on  $X$  if, instead of asking Bob about  $X$ , she asks Carl about a different variable  $Y$ .

i.e. to measure a random variable  $Y$ , of which we know, before doing the experiment, its distribution. We also know the joint distribution  $p(x, y)$  of the two variables. How much information do we expect the experiment will convey on  $X$ ? The reduction in the uncertainty is given by

$$H[X] - H[X|Y] = I[X, Y]$$

as can be shown by a direct calculation. So the mutual information tells us how much we learn, on average, about  $X$  if we know  $Y$ . Note that the mutual information is symmetric

$$I[X, Y] = I[Y, X] = H[Y] - H[Y|X].$$

In other words, the amount of information that we can gain about a theory by performing an experiment, is exactly equal to the uncertainty that the theory provides on the outcome of the experiment.

### Exercise 16.10

Let there be  $n + 1$  boxes labeled  $\omega = 0, 1, \dots, n$ , with  $n$  even. One of the boxes contains a prize, the others are empty. The probability that the prize is in box  $\omega$  is  $p_0$  for  $\omega = 0$  and  $(1 - p_0)/n$  for all  $\omega > 0$ . We have two available strategies:

- 1) open the box  $\omega = 0$
- 2) open the last  $n/2$  boxes ( $\omega > n/2$ )

Which one is the most convenient? Which one conveys more information on where the prize actually is?

Draw a plot of the threshold  $p_0^*$  for which strategies 1 and 2 are equivalent, according to the two criteria. Show that the second is at least as informative as the first for  $p_0 = 1/(n + 1)$  and hence  $p_0^*(n) \geq 1/(n + 1)$ .

This is a toy model for a situation where a phenomenon can be explained by alternative theories, one of which is the prevailing one, whereas the others are very unlikely but are many. The two options correspond to two possible experiments, one that tries to refute or confirm the prevailing theory, the other that can exclude half of the unlikely ones. Check that even if  $p_0 = 0.99$  it might be more informative to exclude unlikely theories if  $n > 270$ .

(Adapted from problem 131 of Bialek's book, *Biophysics*).

Another important point is that knowledge of  $Y$  reduces *a priori* the uncertainty on  $X$ , since  $H[X|Y] \leq H[X]$ , but *a posteriori* this might not be the

case! Take, for example, two random variables  $X, Y \in \{1, 2\}$ , with a joint distribution:

$$p(x, y) = \begin{cases} 0 & \text{if } x = y = 1 \\ 3/4 & \text{if } x = 2 \text{ and } y = 1 \\ 1/8 & \text{if } x = 1 \text{ and } y = 2 \text{ or if } x = y = 2 \end{cases} \quad (16.18)$$

Then  $H[X] \simeq 0.544$  and  $H[X|Y] = 0.25$  bits, i.e.  $I[X, Y] \simeq 0.294$  bits. However if the outcome  $Y = 2$  occurs, the uncertainty on  $X$  actually increases, because  $H[X|Y = 2] = 1$  bit. It is instructive to check the opposite. Does the uncertainty on  $Y$  decreases no matter what value  $X$  turns out to take or not? This should give you a sense of what are the conditions under which the uncertainty may increase after a measurement.

### Exercise 16.11

Generalise this example to the case where  $P\{X = 2, Y = 1\} = a$  and  $P\{X = 1, Y = 2\} = b$  and  $P\{X = 2, Y = 2\} = 1 - a - b$ . What is the values of  $a$  and  $b$  for which no measurement of one of the variables can increase the uncertainty on the other? Are there values of  $a, b$  such that measuring any of the two variables will increase the uncertainty on the other?

## 16.2 The data processing inequality

Information is degraded at every passage, as we know from everyday life. Imagine that Alice communicates a message  $X$  to Bob, and Bob refers the message to Carl. The message  $Y$  that Bob receives may be corrupted by noise, so  $Y \neq X$ , likewise Carl receives a message  $Z$  that may be different from  $Y$ . Formally we represent the situation by saying that  $X, Y$  and  $Z$  are three random variables that form a *Markov chain*, denoted as<sup>12</sup>

$$X \rightarrow Y \rightarrow Z$$

which means that

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

As a consequence, conditional to  $Y$ ,  $X$  and  $Z$  are independent, because  $p(x, z|y) = p(x|y)p(z|y)$ . Note also that the directions of the arrows can be reversed by using Bayes rule, so  $X \rightarrow Y \rightarrow Z$  is equivalent to  $Z \rightarrow Y \rightarrow X$ .

<sup>12</sup>We mention in passing that this notion generalises to *Markov fields*, that specify the dependence between  $n$  random variables with a *graphical model* of  $n$  nodes which are connected by links (or hyperlinks) if the corresponding variables are dependent.

For a Markov chain  $X \rightarrow Y \rightarrow Z$ , the *Data-processing inequality* states that

$$I[X, Z] \leq I[X, Y]. \quad (16.19)$$

In words, the information that  $Y$  contains on  $X$  cannot be increased,<sup>13</sup> whatever transformation  $Y \rightarrow Z$  one can apply. This result is important in statistics, because it suggests that any manipulation of the data can only decrease the information content of the data.

The proof of the inequality (16.19) is simple. The mutual information between  $X$  and  $W = (Y, Z)$  can be written in two ways

$$I[X, W] = \mathbb{E} \left[ \log_2 \frac{p(X, Y, Z)}{p(X)p(Y, Z)} \right] \quad (16.20)$$

$$= \mathbb{E} \left[ \log_2 \frac{p(X, Z|Y)}{p(X)p(Z|Y)} \frac{p(X|Y)}{p(X|Y)} \right] \quad (16.21)$$

$$= I[X, Y] + I[X, Z|Y] \quad (16.22)$$

$$= I[X, Z] + I[X, Y|Z] \quad (16.23)$$

where

$$I[X, Y|Z] = \mathbb{E} \left[ \log_2 \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)} \right]$$

is the conditional mutual information of  $X$  and  $Y$  given  $Z$ . In Eq. (16.22) the term  $I[X, Z|Y] = 0$  vanishes, because  $X$  and  $Z$  are independent, conditional on  $Y$ . The inequality (16.19) follows from the fact that  $I[X, Y|Z] \geq 0$ .

## 16.3 The entropy of Markov Chains

Let us consider Markov chains, i.e. sequences  $\underline{X} = (X_1, \dots, X_N)$  of random variables generated by a transition probability matrix

$$P\{X_t = s | X_{t-1} = s'\} = p_{s,s'}$$

with  $s, s'$  being elements of a finite set  $\mathcal{S}$ . We restrict attention to irreducible chains, for which we know that the probability to observe state  $X_t = s$  converges to the invariant measure  $\mu_s = \sum_{s'} p_{s,s'} \mu_{s'}$ . We further assume that we

<sup>13</sup>There are other general inequalities that can be derived from basic laws. For example the mutual information between  $X_1$  and  $X_2$  cannot be larger than the average of the two entropies. See the book [28].

know that the sequence is sampled in the stationary state, i.e.  $P\{X_1 = s\} = \mu_s$ . Then, the probability of the sequence is given by

$$P\{\underline{X}\} = p_{X_N, X_{N-1}} p_{X_{N-1}, X_{N-2}} \cdots p_{X_2, X_1} \mu_{X_1}. \quad (16.24)$$

Note that the time index goes from right ( $t = 1$ ) to left ( $t = N$ ) in this equation. Taking the logarithm and dividing by  $N$ , please check that the law of large numbers implies

$$\lim_{N \rightarrow \infty} -\frac{1}{N} \log P\{\underline{X}\} = H[X_t | X_{t-1}] \equiv - \sum_{s, s'} p_{s, s'} \mu_{s'} \log p_{s, s'}. \quad (16.25)$$

Note that the entropy of the sequence is smaller than  $N$  times  $H[X_t] = -\sum_s \mu_s \log \mu_s$  which is the entropy of a sequence of i.i.d. random variables, because knowledge of  $X_{t-1}$  provides information on  $X_t$ . From the point of view of the Asymptotic Equipartition property, sequences of  $N$  random variables explore a smaller space than that of  $N$  i.i.d. random variables drawn from  $\mu_s$ .

### 16.3.1 Irreversibility and the arrow of time

Imagine that we do not know whether the sequence  $\underline{X}$  generated from a Markov chain with transition matrix  $p_{s, s'}$  has been given to us in the right order — with time going from 1 to  $N$  — or in the reverse one — with time going from  $N$  to 1. Can we figure this out? In order to do this, let us refine our notation and call  $P\{\underline{X}\} = P_{\rightarrow}\{\underline{X}\}$ , as defined in Eq. (16.24), to distinguish it from the backward probability

$$P_{\leftarrow}\{\underline{X}\} = p_{X_1, X_2} \cdots p_{X_{N-2}, X_{N-1}} p_{X_{N-1}, X_N} \mu_{X_N}. \quad (16.26)$$

#### Exercise 16.12

Show that the naïve generalisation of Eq. (16.25)

$$\log P_{\leftarrow}\{\underline{X}\} \simeq -NH[X_{t-1} | X_t]$$

is wrong.

Show also that  $H[X_{t-1} | X_t] = H[X_t | X_{t-1}]$  in the stationary state. In loose words, given the present, the past is as uncertain as the future in a Markov chain.

The probability of the sequence  $\underline{X}$  can also be expressed in terms of the reverse Markov chain with transition matrix  $q_{s, s'} = p_{s', s} \mu_s / \mu_{s'}$ , as

$$Q_{\leftarrow}\{\underline{X}\} = q_{X_1, X_2} \cdots q_{X_{N-2}, X_{N-1}} q_{X_{N-1}, X_N} \mu_{X_N} = P_{\rightarrow}\{\underline{X}\} \quad (16.27)$$

$$Q_{\rightarrow}\{\underline{X}\} = q_{X_N, X_{N-1}} q_{X_{N-1}, X_{N-2}} \cdots q_{X_2, X_1} \mu_{X_1} = P_{\leftarrow}\{\underline{X}\} \quad (16.28)$$

where the proof of the last equalities relies on repeated use of the identities  $q_{X_{t-1}, X_t} \mu_{X_t} = p_{X_t, X_{t-1}} \mu_{X_{t-1}}$ . The probability of the sequence in the reverse process is given by

$$\frac{1}{N} \log P_{\leftarrow}\{\underline{X}\} = \frac{1}{N} \sum_{t=2}^N \log p_{X_{t-1}, X_t} + \frac{1}{N} \log \mu_{X_t} \quad (16.29)$$

$$= \sum_{s, s'} \frac{k_{s, s'}}{N} \log p_{s', s} + \frac{1}{N} \log \mu_{X_t} \quad (16.30)$$

where  $k_{s, s'}$  is the number of transitions from  $s'$  to  $s$  in the sequence  $\underline{X}$ . As  $N \rightarrow \infty$ , the fraction  $k_{s, s'}/N$  of transitions  $s' \rightarrow s$  converges to the probability  $p_{s, s'} \mu_{s'}$ . Therefore

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_{\leftarrow}\{\underline{X}\} = \sum_{s, s'} p_{s, s'} \mu_{s'} \log p_{s', s} = \sum_{s, s'} p_{s, s'} \mu_{s'} \log q_{s, s'} \quad (16.31)$$

where the proof of the last equality is left as an exercise. Therefore, for large  $N$

$$P_{\leftarrow}\{\underline{X}\} \simeq P_{\rightarrow}\{\underline{X}\} e^{-N\Sigma} \quad (16.32)$$

where

$$\Sigma \equiv D_{KL}[P_{\rightarrow} \| P_{\leftarrow}] = \sum_{s, s'} p_{s, s'} \mu_{s'} \log \frac{p_{s', s}}{p_{s, s'}} = \sum_{s, s'} p_{s, s'} \mu_{s'} \log \frac{p_{s, s'}}{q_{s, s'}} \quad (16.33)$$

is called the *entropy production*. As long as  $q_{s, s'} \neq p_{s, s'}$ , the probability of the forward process is exponentially (in  $N$ ) more likely than the backward one, because  $D_{KL}[P_{\rightarrow} \| P_{\leftarrow}] > 0$ . Hence given the transition matrix  $p_{s, s'}$ , we can detect the arrow of time because the two transition probabilities  $p_{s, s'}$  and  $q_{s, s'}$  are different and they define two distinguishable stochastic processes. Furthermore, notice that the Kullback-Leibler divergence is symmetric in this case, i.e.  $D_{KL}[P_{\leftarrow} \| P_{\rightarrow}] = D_{KL}[P_{\rightarrow} \| P_{\leftarrow}]$ . This reflects the mirror symmetry of the directions of the time arrow: the forward arrow of time under the reverse process is as unlikely as the backward arrow under the forward Markov chain.

If, instead, the Markov chain is reversible, i.e.  $q_{s, s'} = p_{s, s'}$ , then there is no way in which the arrow of time can be detected.

The entropy production is a measure of how much the forward process is more likely than the reversed one, which is expressed in Eq. (16.33) as the difference between the logarithms of the forward and the backward transition probabilities. Indeed irreversibility is related to the existence of a *probability current*, whereby these two terms do not cancel each other and the net probability flow is non-zero.

**Exercise 16.13**

Show that a Markov chain with two states is always reversible. Irreversibility requires at least three states and a probability current that either runs clockwise or counter-clockwise through the states.

**16.4 Data compression and coding theory**

Data compression deals with the problem of optimally representing messages. We refer to Chapter 5 of COVER for a detailed discussion. This is a short summary of the main ideas. The relation between information theory and coding was already hinted at in the introduction. As discussed there, the typical setting is the one where Alice and Bob need to communicate using a binary channel. Then Alice will *encode* her messages to Bob in a string of bits, transmit this string over the channel, and Bob will read it and *decode* it to get the original message. A message  $\underline{X} = (X_1, \dots, X_n)$  is a sequence of symbols  $X_i \in \chi$  drawn from an alphabet  $\chi$ . The simplest example is a text (e.g. a book) which is a sequence of ASCII characters (letters, numbers, spaces, punctuation, etc). But you can likewise think of images, e.g. digital pictures of paintings, as sequences of RGB values for each pixel. Ultimately, each message is stored in digital devices in the form of sequences of zeros and ones, so there is a function  $C(\underline{X})$  that associates to each message  $\underline{X}$  a string  $C(\underline{X})$  of bits. Coding theory deals with the problem of finding ways of representing the data as efficiently as possible, i.e. with the minimal number of bits. Each bit can be thought of as the answer to a yes/no question, so efficient coding, i.e. the problem of optimally<sup>14</sup> representing information, coincides with the problem of eliciting information in an optimal manner, that we already discussed.

Coding theory enters into play, for example, when you use a data compression algorithm (e.g. gzip) on your computer that transforms a text file written in ASCII code into a file that occupies less space on the hard disk of your computer. Compression is possible because messages contain regularities. For example, if the character “q” is always followed by “u” in a text, a code that translates “q” and “u” by different sequences of bits (called codewords) is less efficient than one that codes the pair “qu” directly. Indeed, what the compression program does when you invoke it, is to scan the file you want to compress in search of regularities, i.e. of patterns that occur very frequently. Formally we shall consider messages as being generated as random draws from a probability distribution. Then the knowledge of the probability distri-

<sup>14</sup>In the sense of most parsimoniously.



bution is what makes optimal compression possible. This is why probability theory, coding theory and information theory are so intimately connected.<sup>15</sup>

The main result in coding theory, due to Shannon, makes this connection explicit in the simple case of messages generated as i.i.d. draws from a distribution  $p(x)$  with  $x \in \chi$ . We already discussed Shannon's theorem when we introduced information theory. Let us briefly recall it. Shannon theorem is a consequence (or restatement) of the Asymptotic Equipartition Property. The latter says that, almost surely a message  $\underline{X} = (X_1, \dots, X_n)$  composed of characters drawn independently from the same distribution  $p(x)$  belongs to the set  $A_n$  of typical sequences, which contains  $|A_n| \sim 2^{nH[X]}$  elements. If we label all messages  $\underline{X} \in A_n$  with an integer  $C(\underline{X})$  we can take the binary representation of  $C(\underline{X})$  as the code.<sup>16</sup> Then, almost surely for  $n \rightarrow \infty$ ,  $C(\underline{X}) \leq 2^{nH[X]}$  which means that at most  $H[X]$  bits per character are needed to transmit a message.

This strategy, however, is not very practical because the calculation of  $C(\underline{X})$  requires ranking all messages which are exponentially many in  $n$ . This is practically unfeasible. Also if a message is composed of two parts  $\underline{X} = (\underline{X}_1, \underline{X}_2)$  the code of  $\underline{X}$  is not easily related to those of its parts. For messages where  $X_i$  are drawn as i.i.d. variables from the same distribution, it may be more practical to consider codes such that

$$C(\underline{X}) = (c(X_1), \dots, c(X_n))$$

that are sequences of *codewords*  $c(x)$  each of which corresponds to a character  $x \in \chi$ . So the key question is, how should the function  $c(x)$  be chosen?

We already encountered examples of codes in the introduction, for the

---

<sup>15</sup>A theatre play, such as Othello, is an example of a message, because it is a sequence of letters. It is definitely true that any understanding of the production of Shakespeare has to do with a better understanding of the regularities that one can find in his works. Yet, thinking of his works as being generated as a random draw from a probability distribution seems somewhat extreme, and it is at best an approximation. The simplest such approximation is to think of each letter as being drawn independently at random from a probability distribution. The fact that letters from 'a' to 'z' do not occur with the same probability allows a certain degree of compression of Othello. Furthermore, one realises that certain words (e.g. 'the' or 'and') occur much more frequently than others (e.g. 'Iago') and some (e.g. 'yqat') never occur. This leads to better approximations of the generative process, which affords further compression. Furthermore, the occurrence of words depends on the occurrence of other words in the same act or even in other acts. The more regularities one detects the better one can compress Othello. Note that some of these features are generic of English texts some are generic of Shakespeare's production and some are specific of Othello.

<sup>16</sup>A good way of labelling messages is by their rank in probability, from the most probable to the least probable. You can check that in this way at most  $H[X]$  bits per character are needed to transmit a message.

case where  $\chi = \{a, b, c, d\}$  has four elements, reported on the right.<sup>17</sup> This should allow one to translate each sequence of bits, such as

0010110101001 ...

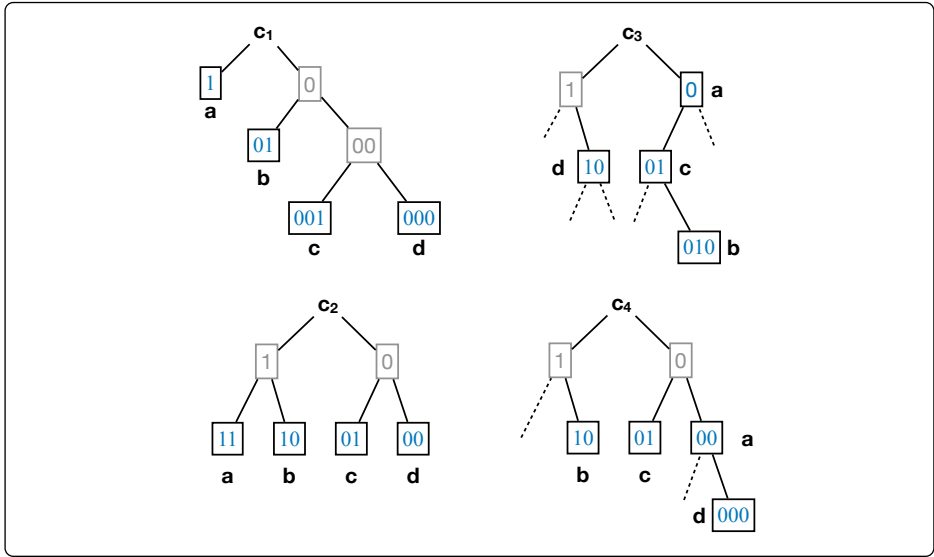
into a sequence of characters in  $\chi$ . A minimal requirement for codes is that they be *uniquely decodable*. This means that any sequence of bits that is produced by translating a sequence of characters should be decodable in a unique manner. This does not happen if there are two or more sequences of characters  $X$  that correspond to the same sequence of bits. The three codes  $c_1, c_2$  and  $c_3$  satisfy this property. For example,  $c_1$  would translate that sequence into *cbabbc* ... whereas  $c_2$  will give *dbaccd* .... In both cases, the translated sequence can be computed as we scan the sequence of bits from left to right. Codes that have this property are called *instantaneous* codes, because they allow to instantaneously translate bit-strings into messages. The key property that makes a code an instantaneous code is that no codeword is the prefix of another codeword, i.e. no codeword coincides with the leftmost part of another codeword.

This is not true for  $c_3$  for which  $c_3(a)$  is a prefix of  $c_3(b)$  and  $c_3(c)$ , for example. In this case it is not possible to figure out what the translation of the leftmost bits is unless one considers also the bits that come after. For example, according to  $c_3$ , the first 0 in the sequence above could correspond to *a* or to the beginning of the codewords for *b* or *c*. However the latter two options should be discarded because the second bit is a 0, which is not compatible with either a *b* or a *c*. If the first character is an *a* the second can be a *b* or a *c*. Yet it cannot be a *b* because otherwise the bits that follow (11 ...) do not correspond to a decodable sequence ( $c_3$  has no codewords that starts with 11). So the first characters should be *accddd* ... but the next characters depend on what the following characters are. Hence  $c_3$  is not an instantaneous code. Finally code  $c_4$  is not uniquely decodable. For example the bit string 000000 could either be *aaa* or *dd*.

We shall focus on instantaneous codes only. Each code admits a representation as a tree, as shown in Figure 39. For instantanous codes, the codewords correspond to the leaves of the tree (the terminal nodes) and the

<sup>17</sup>Four examples of codes:

$\chi$	$c_1(x)$	$c_2(x)$	$c_3(x)$	$c_4(x)$
<i>a</i>	1	11	0	00
<i>b</i>	01	10	010	01
<i>c</i>	001	01	01	10
<i>d</i>	000	00	10	000



**Figure 39.** Representation of the codes  $c_1, c_2, c_3$  and  $c_4$  as trees.

length  $\ell(x) = |c(x)|$  of each codeword (i.e. the number of bits) corresponds to the distance of the corresponding node from the root (which is the top most node). For instantaneous codes, the lengths  $\ell(x)$  satisfy Kraft's inequality

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1. \quad (16.34)$$

This is very easily proven.<sup>18</sup>

With some more effort one can show (see COVER) that for any set of lengths  $\mathcal{L} = \{\ell_1, \ell_2, \dots, \ell_{|\mathcal{X}|}\}$  that satisfy Kraft's inequality Eq. (16.34), i.e. such that  $\sum_i 2^{-\ell_i} \leq 1$ , there is at least one instantaneous code  $c(x)$  such that the lengths  $|c(x)|$  match exactly the  $\ell_i$ 's.

### Exercise 16.14

There is more than one code that corresponds to the same lengths. Count the number of codes which have the same lengths as the codes  $c_1$  and  $c_2$ .

<sup>18</sup>Proof: let  $\bar{\ell} = \max_{x \in \mathcal{X}} \ell(x)$ . Then continue the tree to all nodes at distance  $\bar{\ell}$  from the root. For each word  $x$ , this results in  $2^{\bar{\ell} - \ell(x)}$  nodes at distance  $\bar{\ell}$  down the codeword corresponding to  $x$ . The number of these nodes is  $\sum_{x \in \mathcal{X}} 2^{\bar{\ell} - \ell(x)}$ . This number has to be smaller than the total number of nodes at distance  $\bar{\ell}$  from the root, which is  $2^{\bar{\ell}}$ . This leads to Eq. (16.34).

Among all instantaneous codes, we want to find those that make the expected length of the bit-string it produces as short as possible, when characters are drawn from a distribution  $P\{X = x\} = p_x$ . The two results above imply that it is enough to find a set  $\mathcal{L}$  of lengths that satisfy Kraft's inequality and we're guaranteed that an instantaneous code with those lengths exists. So it is enough to solve the problem

$$\min_{\ell \in \mathcal{L}} \mathbb{E}[\ell(X)] \quad (16.35)$$

over all sets  $\mathcal{L} = \{\ell(x) : \sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1\}$  of lengths that satisfy Kraft's inequality. Introducing this constraint with a Lagrange multiplier, leads to the problem<sup>19</sup>

$$\min_{\ell \in \mathcal{L}, \lambda} \left[ \sum_{x \in \mathcal{X}} p_x \ell(x) - \lambda \left( \sum_{x \in \mathcal{X}} 2^{-\ell(x)} - 1 \right) \right]. \quad (16.36)$$

What makes this problem complicated is that  $\ell(x)$  must be an integer variable. If we neglect this problem and minimise over real values of  $\ell(x)$ , then we're going to obtain a lower bound. The latter problem is simple and is solved by setting to zero the first order derivative of the objective function in Eq. (16.36). This yields  $\ell(x) = -\log_2 p_x$  and

$$\min_{\ell \in \mathcal{L}} \mathbb{E}[\ell(X)] \geq H[X] = - \sum_{x \in \mathcal{X}} p_x \log_2 p_x. \quad (16.37)$$

If you take the smallest integer  $\ell(x)$  which is larger than  $-\log_2 p_x$ , then you can get better estimate of the minimal expected length. The smallest integer larger than  $-\log_2 p_x$  is smaller than  $-\log_2 p_x + 1$ . Therefore the expected length must be smaller than  $H[X] + 1$ . Taken together these results show that for any  $X$  there is an instantaneous code that allows to represent  $X$  with an expected number of bits that is bounded by

$$H[X] \leq \min_{\ell \in \mathcal{L}} \mathbb{E}[\ell(X)] \leq H[X] + 1. \quad (16.38)$$

This result can be improved by invoking *block coding*. This means that, in sending a message  $\underline{X} = (X_1, \dots, X_n)$  with  $n \gg 1$ , instead of using codes that translate each  $X_i$  separately, we can look for the instantaneous codes

<sup>19</sup>Note the sign of the  $\lambda$  term. The most efficient codes are those which have shorter code-words, so those for which the left hand side of Eq. (16.34) is as large as possible, i.e. for which Eq. (16.34) is satisfied as an equality.

that translate a pair  $X_i, X_{i+1}$  of successive variables, or a subsequence  $\underline{X}_i^{(m)} = (X_{i+1}, \dots, X_{i+m})$  of  $m$  successive characters. The same argument that we have applied above implies that

$$H[\underline{X}_i^{(m)}] \leq \min_{\ell \in \mathcal{L}} \mathbb{E} [\ell(\underline{X}_i^{(m)})] \leq H[\underline{X}_i^{(m)}] + 1.$$

However  $H[\underline{X}_i^{(m)}] = mH[X]$  which means that block coding can achieve a compression that satisfies

$$H[X] \leq \frac{1}{m} \min_{\ell \in \mathcal{L}} \mathbb{E} [\ell(\underline{X}_i^{(m)})] \leq H[X] + \frac{1}{m}. \quad (16.39)$$

This result, for  $m \rightarrow \infty$ , coincides with Shannon's bound that ensures that at most  $H[X]$  bits per character need to be exchanged by Alice and Bob in order to communicate messages generated from the distribution  $p_x$ .

This derivation also tells us how optimal codes should look like. Indeed the equation  $\ell(x) = -\log_2 p_x$  tells us that short codewords should be assigned to most probable characters. The Huffman coding algorithm, for example, is based on the idea of iteratively assigning bits to the least probable values of  $x$ , by grouping them together.<sup>20</sup> We refer to COVER for a detailed discussion of this and other algorithms.

### Exercise 16.15

Check that  $c_1$  and  $c_2$  satisfy Kraft's inequality as an equality whereas  $c_3$  does not satisfy it. What about  $c_4$ ? Can you find an instantaneous code for which Kraft's inequality is not satisfied as an equality?

Data compression is only the simplest of the problems discussed in coding theory. A different class of problems have to do with the fact that most daily life communication channels are affected by noise. The string of bits in output is not equal to the one in input, because some bits may be turned from 0 to 1 or viceversa. Communication over noisy channels requires *error correcting codes*, i.e. codes with a built in redundancy that can help recover the original

<sup>20</sup>**Huffman codes:** Huffman coding algorithm reconstruct the tree from the bottom, starting from a partition of the set  $\chi$  of words into singleton sets  $\{x\}$  with an associated probability  $p_x$ . At every step, the algorithm generates a new partition from the old one by merging the two sets  $\mathcal{S}$  and  $\mathcal{S}'$  with the smallest probability, assigning to the new set  $\mathcal{S} \cup \mathcal{S}'$  the sum of the probabilities  $p_{\mathcal{S} \cup \mathcal{S}'} = p_{\mathcal{S}} + p_{\mathcal{S}'}$ . At the same time, the algorithm assigns bits 0 and 1 to the edges joining the nodes corresponding to  $\mathcal{S}$  and  $\mathcal{S}'$  to  $\mathcal{S} \cup \mathcal{S}'$ . The algorithm ends when the partition formed by the single set  $\chi$  is reached, i.e. when all words are merged in the same set. The codeword of  $x$  is given by the sequence of bits associated to all the merging of sets  $\mathcal{S}$  that contain  $x$ , starting from the root  $\chi$ , down to the set  $\{x\}$ .

message, even if that was corrupted by noise. This is a fascinating subject which we will not discuss, however. Yet again, the solution has to do with understanding what the typical messages that need to be transmitted are and how typically they would be corrupted by noise. This allows to get precise bounds, again in terms of entropies, on the amount of redundancy that needs to be embedded in messages, in order to achieve an error free communication.

If you understood the main gist of the arguments discussed above, then you may consider pondering on the following questions:

1. What do you expect the sequence of bits of an optimally compressed sequence  $X_1, \dots, X_n$  should look like? What is the probability that a (randomly chosen) bit is equal to one? What is the difference of this sequence from a sequence of random i.i.d. bits?
2. In all our discussion we have assumed a binary alphabet for the codes. Yet the same results can be derived for codes in an alphabet with three different characters (e.g. 0, 1 and 2), or the 26 characters of the English alphabet. How would this change the results, e.g. Eq. (16.34) and Eq. (16.39)?
3. Languages (e.g. English, French, Chinese, etc) might be thought of as the codes that we use to communicate. A text is a representation of something (an object, a concept, an idea, etc) that is coded as a sequence of characters. Yet, if you look at texts as coded messages, the coding looks rather inefficient. For example, you may delete a certain fraction of characters from a text but still be able to reconstruct the entire text or grasp the gist of the text. The most frequent words in a text (e.g. “the”, “and”, “this”, etc) do not carry any meaning<sup>21</sup> and the least frequent words are very informative on the content of the text. There is a lot of (apparently useless) redundancy in language. Why did humans converged to such inefficient ways of communicating?

---

<sup>21</sup>George Zipf found that for a text like the Holy Bible, the frequency with which the  $r^{\text{th}}$  most frequent word occurs is roughly inversely proportional to  $r$ . This is true for many texts (but not for phone directories) and for texts written in different languages. This implies that the number of words that occur  $k$  times is proportional to  $1/k^2$ , or that the number of occurrences of words used  $k$  times is inversely proportional to  $k$ . This is reminiscent of the Asymptotic Equipartition Property, that states that the number of typical sequences is inversely proportional to their probability. Is this a coincidence or does it hints to the fact that our language has evolved so that text shares some statistical properties with typical sequences?

**Exercise 16.16**

Let a text be generated by first drawing a subject  $Z \in \mathcal{Z}$  and then a message  $\underline{X} = (X_1, \dots, X_n)$  of  $n$  characters  $X_i \in \mathcal{X}$  drawn independently from a distribution  $p(x|z) = P(X_i = x|Z = z)$ . There are two possible strategies: *A*) use the same code irrespective of the subject, and *B*) first code the subject  $Z$  and then code the text  $\underline{X}$  depending on the subject (two way code). Note that code *B* represents each text  $\underline{X}$  optimally, at the expense of the extra cost of coding  $Z$ , whereas texts  $\underline{X}$  are never coded optimally with strategy *A*, with an over expenditure of bits that should grow with  $n$ . Show that, irrespective of this, the two way code *B* is never the best one.





# Chapter 17

## Large deviation theory

“It is just more likely, that is all. It is a good guess. And we always try to guess the most likely explanation, keeping in the back of the mind the fact that if it does not work we must discuss the other possibilities.” (R.P. Feynmann, 1965)

Having discussed typical events, let us discuss a-typical events.<sup>1</sup> There are two reasons (at least) why a-typical events may be of interest. First we may be interested in *rare* events that involve fluctuations of quantities that are larger than what one typically expects. For example, the credit rating of an insurance company is based on its estimated default probability. This occurs if an unexpectedly large number of contracts in its portfolio demand claims that exceed the equity<sup>2</sup>  $A$  of the insurance company. The claims  $X_i$  from contracts  $i = 1, \dots, n$  can be modeled as random variables and the default corresponds to the event

$$D = \{S_n \geq A\}, \quad S_n = \sum_{i=1}^n X_i.$$

If  $n \gg 1$ , which is the case in this example, we know that as long  $\mathbb{E}[S_n] < A$  this even does not typically occur. So default  $D$  is an a-typical event.

Communications engineers face a similar problem: they need to calculate safe buffer and bandwidth sizes for network traffic which arises from a population of many users. This entails estimating the probability of traffic overflow, making sure that these will be very rare events. In both cases, we

---

<sup>1</sup>There are several textbooks devoted to Large Deviation Theory, as e.g. [29].

<sup>2</sup>The equity is a measure of the value of the company, and it equals the amount of money that would result if all of the assets of the company were liquidated and all debts were paid off.

want to estimate how small is the probability of the large deviation and how do we expect it to occur.

In a stylised picture, biological evolution occurs through random mutations. Most of them have neutral or deleterious effects, and the accumulation of such deleterious mutations generally decreases the reproduction probability — the fitness — of descendants. Yet some rare mutation bring advantages that increase the fitness of individuals carrying them, whose descendants will reproduce faster. So the fitness of the population as a whole does not decrease, because evolution is propelled by rare events.

More in general, when we study a phenomenon we might represent our current state of knowledge with a distribution  $Q(\omega)$  defined on the sample space of all possible realisations  $\omega \in \Omega$  of that phenomenon. You may think of  $\omega$  as a complete description of that phenomenon and of  $Q$  as the distribution encoding all known (experimental) facts. The distribution  $Q$  is the *theory* that allows us to predict the value  $\mu_Q = \mathbb{E}_Q[X]$  of a quantity  $X(\omega)$ . Clearly, we're interested in predictions of the theory  $Q$  going beyond the range of events that have been used to derive it.

This prediction can be tested in a repeated series of independent experiments  $\underline{X} = (X_1, \dots, X_n)$  and, if  $\mathbb{E}_Q[X]$  is finite, we expect that  $S_n/n \cong \mathbb{E}_Q[X]$  for  $n$  large. If this expectation is confirmed by the experiment, then the experiment brings no new information. But if  $S_n/n$  is very different from  $\mathbb{E}_Q[X]$ , then the experimental result calls for a revised theory  $P$  that can accommodate all existing knowledge and the new observation. In this case, the experiment is an a-typical event because the theory  $Q$  is wrong.<sup>3</sup> How should we revise the theory  $Q \rightarrow P$  in order to incorporate the new information? And how much did we learn?

The study of rare (a-typical) events is the domain of *Large Deviation Theory*. Let us start by formalising the main questions and concepts in the case of sequences  $\underline{X} = (X_1, \dots, X_n)$  of i.i.d. random variables. Let us assume that the variance  $\mathbb{V}[X_i] = \sigma^2 < \infty$  is finite, so that both the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) hold. Then, for large  $n$ , the mean  $S_n/n$  will be very close to  $\mu = \mathbb{E}[X]$  (LLN) and the sum  $S_n$  is well approximated by  $S_n \simeq n\mu + \sigma\sqrt{n}\zeta$  where  $\zeta$  is a Gaussian random variable with zero mean and unit variance (CLT). This is what we *typically* expect.

<sup>3</sup>This logic is routinely applied in statistics, when we want to test an hypothesis. Then  $Q(\omega)$  stands for the distribution that we expect if a certain hypothesis  $H_0$  is satisfied. A practical example is that of subjects that receive a treatment for a certain disease. Then one wants to rule out the *null hypothesis*  $H_0$  that the treatment is completely ineffective, on the basis of a sample  $\underline{X}$  of measurement of a quantity  $X$  that is known to be relevant. In hypothesis testing, we take  $Q(x)$  as the distribution that  $X$  would follow in untreated patients. In this case, if the treatment is effective then the sample  $\underline{X}$  is a-typical.

Yet it may happen to observe *large deviations*,<sup>4</sup> i.e. events such that, for some  $\epsilon > 0$ ,

$$A_n(\bar{x}) = \left\{ \underline{X} : \left| \frac{1}{n} \sum_{i=1}^n X_i - \bar{x} \right| < \epsilon \right\} \quad (17.1)$$

with  $\bar{x} \neq \mu$ . These are clearly *a-typical* events that we expect to occur with a vanishingly small probability, as  $n \rightarrow \infty$ .

The questions that we shall focus on are:

1. what is the probability  $P\{A_n(\bar{x})\}$  of the large deviation? More specifically, since  $P\{A_n(\bar{x})\} \rightarrow 0$  as  $n \rightarrow \infty$ , we shall be interested in the leading behaviour of  $P\{A_n(\bar{x})\}$  with  $n$ .
2. Conditional on the fact that  $A_n(\bar{x})$  occurs, what is the distribution of the  $X_i$ ? In other words, how are large deviations typically realised?

The answers to these questions depend on the distribution from which the sample  $\underline{X}$  is drawn. We shall discuss separately the different cases.

## 17.1 Large deviations for i.i.d. variables with finite support

Consider<sup>5</sup> a sequence of  $n$  i.i.d. random variables  $\underline{X} = (X_1, \dots, X_n)$  drawn from a distribution  $Q(x)$  over a finite alphabet  $x \in \mathcal{X}$  (i.e.  $|\mathcal{X}| < +\infty$ ). The probability of a sample  $\underline{X}$  is given by<sup>6</sup>

$$P\{\underline{X}\} = \prod_{i=1}^n Q(X_i) = \prod_{x \in \mathcal{X}} Q(x)^{nP_{\underline{X}}(x)} = e^{-n\mathcal{H}[P_{\underline{X}}] - nD_{KL}[P_{\underline{X}} \| Q]}. \quad (17.2)$$

where

$$P_{\underline{X}}(x) = \frac{1}{n} |\{i : X_i = x\}| \quad (17.3)$$

<sup>4</sup>NY Times reports on Dec. 11, 2021 that Kentucky “was hit by four tornadoes [...] including one that stayed on the ground for more than 200 miles.” The Governor of Kentucky said “This has been the most devastating tornado event in our state’s history, [...] The level of devastation is unlike anything I have ever seen.” This is a very unlikely event according to the distribution of past events. Scientists suspect that this suggests that the distribution of severity of these events has changed because of climate change.

<sup>5</sup>This part is discussed in COVER, Chapter 11.

<sup>6</sup>Remember that

$$\mathcal{H}[P] = - \sum_{x \in \mathcal{X}} P(x) \log P(x)$$

is the entropy as a functional of  $P(x)$ .

is the empirical distribution,<sup>7</sup> which is the fraction of points in the sample that are equal to  $x$ . In particular, the probability of a sample  $P\{\underline{X}\}$  only depends on its type  $P_{\underline{X}}$ . The event  $A_n$  also can be defined in terms of types, as a subset in the space of distributions<sup>8</sup>  $\mathcal{P}$  or of types  $\mathcal{P}_n \subseteq \mathcal{P}$  of samples of  $n$  points. More precisely, the event defined in Eq. (17.1) can be rewritten as  $A_n = \{P_{\underline{X}} \in \mathcal{A}_n \subseteq \mathcal{P}_n\}$  where

$$\mathcal{A}_n = \{P \in \mathcal{P}_n : |\mathbb{E}_P[X] - \bar{x}| < \epsilon\}, \quad \mathbb{E}_P[X] = \sum_{x \in \mathcal{X}} P(x)x \quad (17.4)$$

is a subset of the space of distributions defined on  $\mathcal{X}$ .

The probability that an event  $A_n$  occurs can be written as

$$P\{A_n\} = \sum_{\underline{X} \in A_n} P\{\underline{X}\} = \sum_{\underline{X} \in A_n} e^{-n\mathcal{H}[P_{\underline{X}}] - nD_{KL}[P_{\underline{X}}\|Q]} \quad (17.5)$$

$$= \sum_{P \in \mathcal{A}_n} \sum_{\underline{X}: P_{\underline{X}}=P} e^{-n\mathcal{H}[P] - nD_{KL}[P\|Q]} \quad (17.6)$$

$$= \sum_{P \in \mathcal{A}_n} e^{-n\mathcal{H}[P] - nD_{KL}[P\|Q]} \left| \left\{ \underline{X} : P_{\underline{X}} = P \right\} \right| \quad (17.7)$$

$$\sim \sum_{P \in \mathcal{A}_n} e^{-nD_{KL}[P\|Q]} \quad (17.8)$$

where we used the fact that, by Eq. (17.2)  $P\{\underline{X}\}$  only depends on  $P_{\underline{X}}$  in the first equation, and the fact that the number  $\left| \left\{ \underline{X} : P_{\underline{X}} = P \right\} \right|$  of samples with  $P_{\underline{X}} = P$  is  $\sim e^{n\mathcal{H}[P]}$ , by the Asymptotic Equipartition Property.<sup>9</sup>

If  $|\bar{x} - \mathbb{E}_Q[X]| < \epsilon$  then the event  $A_n$  is typical, which means that there is at least one distribution  $P \in \mathcal{A}_n$  that is very close to  $Q$ , and that asymptotically converges to it. Therefore for these distributions  $D_{KL}[P\|Q] \rightarrow 0$  as  $n \rightarrow \infty$  and, as a consequence,  $P\{A_n\} \rightarrow 1$ . If  $\bar{x}$  is significantly different from  $\mathbb{E}_Q[X]$

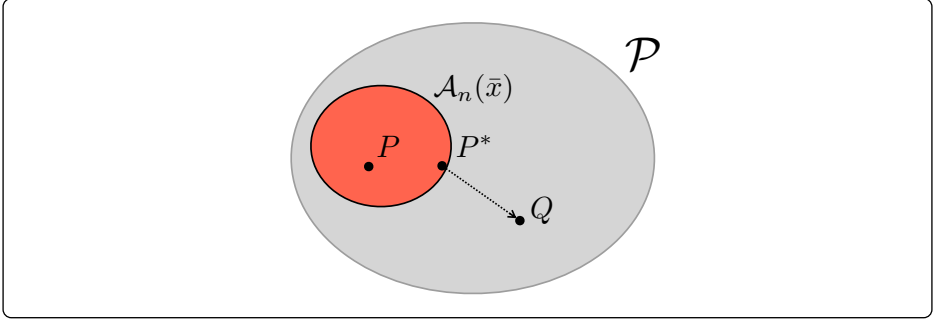
<sup>7</sup> $P_{\underline{X}}(x)$  is called the type of  $\underline{X}$ . We refer to COVER, Chapter 11 for a detailed discussion of types.

<sup>8</sup>The space of distributions is defined as

$$\mathcal{P} = \left\{ P : \mathcal{X} \rightarrow \mathbb{R}, P(x) \geq 0, \sum_{x \in \mathcal{X}} P(x) = 1 \right\}.$$

The set  $\mathcal{P}_n$  of types is a subset of  $\mathcal{P}$  of distributions where, for all  $x \in \mathcal{X}$ ,  $p(x) = k_x/n$  with  $k_x = 0, 1, \dots, n$  and  $\sum_{x \in \mathcal{X}} k_x = n$ .  $\mathcal{P}_n$  is a discrete set of points in  $\mathcal{P}$ . For each  $x \in \mathcal{X}$ ,  $k_x$  can take  $n+1$  values, so the number of points in  $\mathcal{P}_n$  can be at most  $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$ . As  $n$  increases, the number of points in  $\mathcal{P}_n$  becomes denser and denser, so that each  $P \in \mathcal{P}$  can be approximated to arbitrary precision by a  $P \in \mathcal{P}_n$  if  $n$  is sufficiently large.

<sup>9</sup>Let us remind that  $a_n \sim e^{cn}$ , where  $c$  is a constant, means that  $\frac{1}{n} \log a_n \rightarrow c$  as  $n \rightarrow \infty$ .



**Figure 40.** Sketch of the minimisation problem in large deviation theory. We note in passing that the relative entropy  $D_{KL}[P\|Q] \geq D_{KL}[P\|P^*] + D_{KL}[P^*\|Q]$  satisfies the opposite of the triangle inequality (see COVER Theorem 11.6.1).

then  $Q$  is “far” from any  $P \in \mathcal{A}_n$ . Then  $A_n$  is an  $a$ -typical event and its probability vanishes as  $n \rightarrow \infty$ . Every type  $P \in \mathcal{A}_n$  contributes with a term which is exponentially small in  $n$ , with a coefficient that is proportional to  $D_{KL}[P\|Q]$ . Then for  $n$  large, we expect that the sum will be dominated by the type

$$P^* = \arg \min_{P \in \mathcal{A}_n} D_{KL}[P\|Q] \quad (17.9)$$

that is “closest” to  $Q$ , in terms of  $D_{KL}$  divergence. Indeed, taking only the term  $P = P^*$  in the sum over  $\mathcal{A}_n$  in Eq. (17.8), one gets  $P\{A_n\} \geq e^{-nD_{KL}[P^*\|Q]}$ . On the other hand,  $e^{-nD_{KL}[P\|Q]} \leq e^{-nD_{KL}[P^*\|Q]}$  that provides an upper bound

$$P\{A_n\} \leq e^{-nD_{KL}[P^*\|Q]} |\mathcal{A}_n| \quad (17.10)$$

$$\leq (1+n)^{|\mathcal{X}|} e^{-nD_{KL}[P^*\|Q]} \quad (17.11)$$

where we used the fact that the number  $|\mathcal{A}_n|$  of types  $P \in \mathcal{A}_n$  is upper bounded by the total number of types  $|\mathcal{P}_n|$ , which is less than  $(n+1)^{|\mathcal{X}|}$ . This means that  $P\{A_n\}$  decays exponentially with a rate which is equal to  $D_{KL}[P^*\|Q]$ . This is the content of *Sanov’s theorem*, i.e.

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n\} = -D_{KL}[P^*\|Q]. \quad (17.12)$$

Summarising, the leading order in the behaviour of the probability of an  $a$ -typical event  $A_n$  when  $n \rightarrow \infty$ , is given by  $P\{A_n\} \sim e^{-nD_{KL}[P^*\|Q]}$  where  $P^*$  is the solution of Eq. (17.9).

Let us illustrate this for the case

$$\mathcal{A}_n = \left\{ P : \sum_{x \in \mathcal{X}} P(x)f(x) \geq \bar{f} \right\}$$

that corresponds to event  $A_n$  where the average of  $f(X)$  over a sample  $\underline{X}$  of points drawn independently from  $Q(x)$ , is larger than  $\bar{f}$ . If

$$\mathbb{E}_Q[f(X)] \equiv \sum_{x \in \mathcal{X}} Q(x)f(x) \geq \bar{f}$$

then  $Q \in \mathcal{A}_n$  and the event is typical. The interesting case is when  $\mathbb{E}_Q[f(X)] < \bar{f}$  because then  $A_n$  is an atypical event where the sample average

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \geq \bar{f}$$

does not satisfy the law of large numbers. In order to compute  $P\{A_n\}$  we should first solve the problem Eq. (17.9). This is done introducing Lagrange multipliers and solving the problem

$$\min_{P, \beta, \lambda} \left[ D_{KL}[P \| Q] + \beta \left( \sum_{x \in \mathcal{X}} P(x)f(x) - f_0 \right) + \lambda \left( \sum_{x \in \mathcal{X}} P(x) - 1 \right) \right],$$

where  $f_0 \geq \bar{f}$  has to be chosen so as to satisfy Eq. (17.9). Equating the derivative of the objective function in this minimisation problem to zero, shows that the solution has the form

$$P_\beta(x) = \frac{Q(x)e^{-\beta f(x)}}{Z(\beta)} \quad (17.13)$$

where

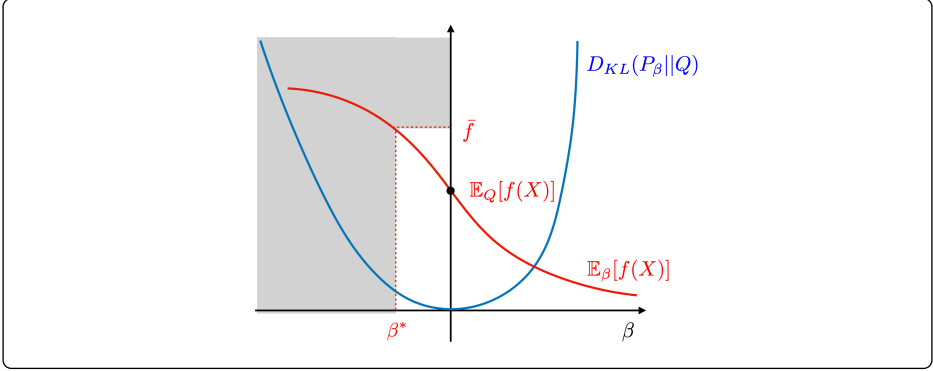
$$Z(\beta) = \mathbb{E}_Q[e^{-\beta f(X)}] = \sum_{x \in \mathcal{X}} Q(x)e^{-\beta f(x)} \quad (17.14)$$

is the normalisation constant.<sup>10</sup> The parameter  $\beta$  has to be fixed so that

$$\mathbb{E}_\beta[f(X)] = \sum_{x \in \mathcal{X}} P_\beta(x)f(x) = -\frac{d}{d\beta} \log Z(\beta) \quad (17.15)$$

where we used  $\mathbb{E}_\beta[\dots]$  for expectations over the distribution  $P_\beta$ . Notice that when  $\beta = 0$  then  $P_\beta(x) = Q(x)$  is the original distribution. For this reason, the curve  $\mathbb{E}_\beta[f(X)]$  takes the value  $\mathbb{E}_Q[f(X)]$  for  $\beta = 0$ . In other words, the point  $\beta = 0$  corresponds to typical events, where the law of large numbers holds. Varying  $\beta$  one “explores” rare events with large fluctuations of the

<sup>10</sup>  $Z(\beta)$  is often called the *partition function*. Note that the derivatives of  $\log Z(\beta)$  is closely related to the cumulant generating function of  $X$ ,  $\phi(h) = \log Z(-h)$ . We use this property to relate the derivatives of  $\log Z(\beta)$  to the cumulants of  $X$  under the distribution  $P_\beta$ .



**Figure 41.** The shaded region corresponds to the event  $\mathcal{A}_n$ .

sample mean of  $f$ . In particular,  $\mathbb{E}_\beta[f(X)]$  is a decreasing function of  $\beta$  (see Figure 41), because

$$\frac{d\mathbb{E}_\beta[f(X)]}{d\beta} = -\left\{\mathbb{E}_\beta[f^2(X)] - \mathbb{E}_\beta[f(X)]^2\right\} = -\mathbb{V}_\beta[f(X)] \leq 0.$$

So the event  $\mathcal{A}_n$  corresponds to all those  $\beta$  for which  $\mathbb{E}_\beta[f(X)] \geq \bar{f}$ , i.e. to the region  $\beta \leq \beta^*$  where  $\beta^*$  is such that  $\mathbb{E}_{\beta^*}[f(X)] = \bar{f}$ .

Among all the distributions  $P_\beta$  with  $\beta \leq \beta^*$  we should choose that one with the smallest  $D_{KL}[\cdot \| Q]$ . Now

$$D_{KL}[P_\beta \| Q] = -\beta \mathbb{E}_\beta[f(X)] - \log Z(\beta)$$

and

$$\frac{dD_{KL}[P_\beta \| Q]}{d\beta} = \beta \mathbb{V}_\beta[f(X)]$$

has the same sign of  $\beta$ . Therefore,  $D_{KL}[P_\beta \| Q]$  has a minimum at  $\beta = 0$  and its minimum for  $\beta \leq \beta^* \leq 0$  is attained at  $\beta^*$ . Summarizing,

$$P\{\mathcal{A}_n\} \sim e^{-nD_{KL}[P^* \| Q]}, \quad D_{KL}[P^* \| Q] = -\beta^* \bar{f} - \log Z(\beta^*)$$

where  $\beta^*$  satisfies  $\mathbb{E}_{\beta^*}[f(X)] = \bar{f}$  and  $P^* = P_{\beta^*}$ .

What is the meaning of the distribution  $P_{\beta^*}$ ? In order to address this question, let us compute the marginal distribution of the first  $m$  variables  $\underline{X} = (X_1, \dots, X_m)$

$$P(\underline{X} | \mathcal{A}_n(\bar{x})) = P\{X_1 = x_1, \dots, X_m = x_m | \mathcal{A}_n(\bar{x})\}$$

when  $n \rightarrow \infty$  with  $m$  finite, conditional on the occurrence of the large deviation  $A_n(\bar{x})$ . We observe that<sup>11</sup>

$$P(\underline{X}|A_n(\bar{x})) = \frac{P\{X_1 = x_1, \dots, X_m = x_m\} \cap A_{n-m}(\bar{x}')}{P\{A_n(\bar{x})\}} \quad (17.16)$$

$$= \frac{Q(x_1) \cdots Q(x_m) P\{A_{n-m}(\bar{x}')\}}{P\{A_n(\bar{x})\}} \quad (17.17)$$

$$\simeq P_{\beta^*}(x_1) \cdots P_{\beta^*}(x_m) \quad (17.18)$$

where in the first equality, the event  $A_{n-m}(\bar{x}')$  is the event that the  $n - m$  variables  $(X_{m+1}, \dots, X_n)$  sum up to

$$\sum_{i=m+1}^n X_i = n\bar{x} - \sum_{i=1}^m x_i \equiv (n - m)\bar{x}'. \quad (17.19)$$

In Eq. (17.17) we use the fact that the variables  $X_i$  are independent and they are drawn from  $Q$ . Finally, Eq. (17.18) holds because<sup>12</sup>

$$\frac{P\{A_{n-m}(\bar{x}')\}}{P\{A_n(\bar{x})\}} \simeq e^{-(n-m)D_{KL}[P_{\beta'}\|Q] + nD_{KL}[P_{\beta}\|Q]} \quad (17.20)$$

$$\begin{aligned} &\simeq e^{(n-m)\beta'\bar{x}' - n\beta\bar{x} - n[\log Z(\beta') - \log Z(\beta)] - m \log Z(\beta')} \\ &\simeq \frac{1}{Z(\beta^*)^m} e^{-\beta^* \sum_{i=1}^m x_i} \end{aligned} \quad (17.21)$$

for  $n \rightarrow \infty$ . Eq. (17.17) shows that in the limit  $n \rightarrow \infty$  the joint distribution of  $\underline{X}$  coincides with the distribution of  $m$  variables  $X_1, \dots, X_m$  which are drawn independently from the same distribution  $P_{\beta^*}(x)$ . In loose words, *the large deviation is realised as a typical sample of independently drawn variables from a distribution  $P_{\beta^*}(x)$ , which is different from  $Q$ .*

<sup>11</sup>We use the previous results with  $f(x) = x$  for simplicity.

<sup>12</sup>Here we use the shorthand  $\beta = \beta^*(\bar{x})$  and  $\beta' = \beta^*(\bar{x}')$ . The second line follows from the fact that

$$D_{KL}[P^*\|Q] = -\beta\bar{x} - \log Z(\beta).$$

In the first term of the exponent we use Eq. (17.19) so that

$$(n - m)\beta'\bar{x}' - n\beta\bar{x} = n(\beta' - \beta)\bar{x} - \sum_{i=1}^m x_i$$

The first term cancels with

$$\log Z(\beta') - \log Z(\beta) \simeq -(\beta' - \beta)\bar{x} + \dots$$

that is obtained expanding  $\log Z(\beta')$  around  $\beta$  (note that  $\beta - \beta' \sim \bar{x} - \bar{x}'$  is of order  $1/n$ ) using  $\bar{x} = -\frac{d}{d\beta} \log Z(\beta)$ .



**Exercise 17.1**

Compute the function

$$I(\bar{x}) = - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n(\bar{x})\}$$

for a Poisson distribution with mean  $\mathbb{E}_Q[X] = \lambda$ . Show that the solution satisfies the relation  $I(\bar{x}) = \lambda f(\bar{x}/\lambda)$ . Show that the same relation should hold for any infinitely divisible distribution.

There are in principle many other ways in which a sample that satisfies  $A_n(\bar{x})$  could be realised. Any other distribution  $P \in \mathcal{A}_n(\bar{x})$  such that  $\mathbb{E}_P[X] = \bar{x}$  would generate samples that satisfies  $A_n(\bar{x})$ , typically. However, the probability to generate samples with type  $P_{\bar{X}} = P$  is  $e^{-nD_{KL}[P||Q]}$ , which is exponentially smaller (in  $n$ ) than the probability of typical samples generated as i.i.d. draws from  $P^*$  in Eq. (17.9). The distribution that is most likely to be observed is the “closest” to  $Q$  in terms of the KL divergence.<sup>13</sup>

## 17.2 Large deviations for i.i.d. continuous variables with thin tails

The same solution can be derived<sup>14</sup> by a direct calculation for the cases where  $X_i \in \mathbb{R}$  are continuous i.i.d. random variables whose common pdf  $q(x)$  decays at least exponentially fast.<sup>15</sup> We refer to this case by saying that  $q(x)$  has *thin tails*. The case of *fat tails*, where  $q(x)$  decays slower than an exponential, will be discussed later.

Let  $A_n(\bar{x})$  be the event that the mean falls in an interval  $[\bar{x}, \bar{x} + d\bar{x})$  for an infinitesimal  $d\bar{x}$ . Then  $P\{A_n(\bar{x})\} = p_n(\bar{x})d\bar{x}$  where  $p_n(\bar{x})$  is the pdf of  $\bar{x}$ . This

<sup>13</sup>Remember that the type  $P_{\bar{X}}$  of a random sample of i.i.d. draws from a distribution is not random at all when  $n \rightarrow \infty$ , by the Glivenko-Cantelli theorem.

<sup>14</sup>This derivation can be found also in the appendix of [30].

<sup>15</sup>I.e. distributions such that for some  $\lambda, K > 0$

$$\lim_{x \rightarrow \pm\infty} q(x)e^{\lambda|x|} \leq K.$$

can be computed using the integral representation of the delta function<sup>16</sup>

$$p_n(\bar{x}) = n \int_{-\infty}^{\infty} \prod_{i=1}^n dx_i q(x_i) \delta\left(\sum_i x_i - n\bar{x}\right) \quad (17.22)$$

$$= n \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{ikn\bar{x}} \left[ \int dx q(x) e^{-ikx} \right]^n \quad (17.23)$$

$$= n \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{ng(ik)} \quad (17.24)$$

where the function  $g(\beta)$  is defined as.

$$g(\beta) = \beta\bar{x} + \log \int dx q(x) e^{-\beta x}$$

The integral in Eq. (17.24) can be evaluated by the saddle point method. This entails looking at the stationary point of  $g(\beta)$  and expanding around it. The maximum of  $g(\beta)$  is attained at  $\beta^*(\bar{x})$  that satisfies the equation  $g'(\beta) = 0$ , i.e.

$$\bar{x} = \frac{1}{Z(\beta)} \int dx x q(x) e^{-\beta x}, \quad Z(\beta) = \int dx q(x) e^{-\beta x} = \mathbb{E}_Q [e^{-\beta X}] \quad (17.25)$$

Then one can perform the integral in Eq. (17.24) by substituting

$$g(ik) = g(\beta^*) + \frac{g''(\beta^*)}{2} (ik - \beta^*)^2 + O(ik - \beta^*)^3$$

Upon changing variables to  $y = \sqrt{ng''(\beta)}(k + i\beta^*)$  one can check that higher order terms in the expansion of  $g$  beyond the second one are small for  $n$  large and can be neglected. Therefore one can compute the Gaussian integral with the result

$$p_n(\bar{x}) \simeq \sqrt{\frac{n}{2\pi g''(\beta^*)}} e^{ng(\beta^*)} \sim e^{ng(\beta^*)} \quad (17.26)$$

where the leading order behavior in  $n$  is retained in the last equation.

<sup>16</sup>The Dirac's  $\delta(x)$  function is defined as that (generalized) function such that for any function  $f(x)$

$$\int_{-\infty}^{\infty} dx f(x) \delta(x - x_0) = f(x_0)$$

In particular with  $f(x) = 1$  this shows that  $\delta(x - x_0)$  is a pdf whose mass is concentrated in  $x_0$ . With  $f(x) = e^{ikx}$  the relation above shows that the Fourier transform of  $\delta(x)$  is 1. Hence

$$\delta(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{-ikx}.$$

Also note that  $\delta(ax) = \delta(x)/a$ .

**Exercise 17.2**

There are other ways in which a large deviation  $\bar{x}$  can be realised. Imagine that a large deviation  $\bar{x} = \mathbb{E}_Q[X] + a$  is observed, with  $a > 0$ . The “explanation” of large deviation theory is that the event  $A_n(\bar{x})$  occurs because  $X_i$  are actually not drawn from  $q(x)$  but from  $p_\beta(x)$  of Eq. (17.27), with  $\beta$  determined by the condition  $\bar{x} = \mathbb{E}_\beta[X]$ . A different explanation is that, instead, the  $X_i$  are drawn i.i.d. from a “shifted” distribution  $p_a(x) = q(x - a)$ . Show, for the specific example of exponential random variables,  $q(x) = e^{-x}$  for  $x \geq 0$  and  $q(x) = 0$  for  $x < 0$ , that the “shifted” distribution hypothesis is much less plausible than the one offered by large deviation theory.

There are few things to observe in this result:

1. The form of Eq. (17.25) that fixes  $\beta^*$  is of the form  $\bar{x} = \mathbb{E}_\beta[X]$  where the expectation is taken on the modified distribution

$$p_\beta(x) = \frac{q(x)e^{-\beta x}}{Z(\beta)} \quad (17.27)$$

This is not a coincidence, as we’re going to see.

2. The second derivative of  $g$  is positive as it is the variance of a random variable  $X$  with pdf  $p_\beta(x)$

$$g''(\beta) = \mathbb{E}_\beta[X^2] - \mathbb{E}_\beta[X]^2 = \mathbb{V}_\beta[X]$$

3. The marginal joint distribution of a finite number  $m$  of variables, say  $\underline{X} = (X_1, \dots, X_m)$  conditional on the occurrence of  $A_n(\bar{x})$ , defined as

$$\begin{aligned} p(\underline{X}|A_n(\bar{x})) dx_1 \cdots dx_m \\ = P\{X_1 \in [x_1, x_1 + dx_1), \dots, X_m \in [x_m, x_m + dx_m)|A_n(\bar{x})\} \end{aligned}$$

can be estimated as before, and

$$\lim_{n \rightarrow \infty} p(x_1, \dots, x_m | A_n(\bar{x})) = p_\beta(x_1) \cdots p_\beta(x_m)$$

This shows that the large deviation is realised as an independent draw of variables from the distribution  $p_\beta(x)$ .

4. The expression of the rate of exponential decay of the probability  $P\{A_n(\bar{x})\}$  can be written as

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n(\bar{x})\} = g(\beta^*) = -D_{KL}[p_{\beta^*} \| q]$$

as shown by a direct calculation. This is the same result as Sanov's theorem Eq. (17.12). The fact that  $P\{A_n(\bar{x})\}$  is related to a relative entropy is not accidental, as we discussed earlier.

### 17.3 Large deviations and the Legendre transform

The function

$$I(\bar{x}) = - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n(\bar{x})\} \quad (17.28)$$

is called the Cramer's function or the large deviation (rate) function. As shown above,  $I(\bar{x}) = D_{KL}[P_{\beta^*(\bar{x})} \| Q]$  is a relative entropy. Rephrasing the steps we did above, the practical recipe to compute the Cramer's function is condensed in the following steps:<sup>17</sup>

1. Compute the cumulant generating function

$$\phi(h) = \log \int dx q(x) e^{hx} = \log \mathbb{E}_Q [e^{hX}]$$

2. Take a derivative of  $\phi$  and compute

$$\bar{x}(h) = \frac{d\phi}{dh} \quad (17.29)$$

3. invert this function and compute  $h(\bar{x})$

4. compute

$$I(\bar{x}) = \bar{x}h(\bar{x}) - \phi[h(\bar{x})]$$

The variables  $h$  and  $\bar{x}$  are called *conjugate variables*. Notice that the function  $\phi(h)$  has to be concave, i.e. its second derivative must be positive. This is always true in the present case, because  $\phi''(h) = \mathbb{V}_\beta[X] > 0$  is given by the variance of  $X$  on the distribution  $P_\beta$  (with  $\beta = -h$ ). Indeed the steps above

---

<sup>17</sup>In the derivation above we had

$$I(\bar{x}) = -g(\beta^*), \quad h = -\beta$$

and  $\phi(h) = \log Z(\beta)$ . The reason for this change of notation will become clear in what follows.

“map” a concave function  $\phi(h)$  into another concave function  $I(\bar{x})$ , because you can easily check that  $I''(\bar{x}) = 1/\phi''(h) > 0$ .

As a general remark, note that the function  $I(\bar{x})$  contains (and it has to be consistent with) both the law of large numbers and the central limit theorem. The first implies that  $I(\bar{x}) = 0$  when  $\bar{x} = \mathbb{E}_Q[X]$ . The second implies that the pdf of  $\bar{x}$  is well approximated by a Gaussian for  $\bar{x} \simeq \mathbb{E}_Q[X]$ , i.e.

$$p_n(\bar{x}) \simeq \sqrt{\frac{n}{2\pi\mathbb{V}_Q[X]}} e^{-\frac{n(\bar{x}-\mathbb{E}_Q[X])^2}{2\mathbb{V}_Q[X]}}.$$

Therefore,  $I(\bar{x}) \simeq \frac{(\bar{x}-\mathbb{E}_Q[X])^2}{2\mathbb{V}_Q[X]} + \dots$  is well approximated by a quadratic function for  $\bar{x} \simeq \mathbb{E}_Q[X]$ . This can indeed be checked explicitly, because the second derivative of  $I(\bar{x})$  for  $\bar{x} = \mathbb{E}_Q[X]$  is the inverse of the second derivative of the cumulant generating function  $\phi(h)$  for  $h = 0$ , which is the variance  $\mathbb{V}_Q[X]$ .

### Exercise 17.3

Compute the Cramer function  $I(\bar{x})$  for the exponential distribution  $p(x) = e^{-x}$ ,  $x \geq 0$ .

The mathematics described here is that of Legendre transforms.<sup>18</sup> This mathematical construction does not arise accidentally. Consider the following constrained optimisation problem

$$I(\bar{x}) = \min_{P: x(P)=\bar{x}} U(P) \quad (17.30)$$

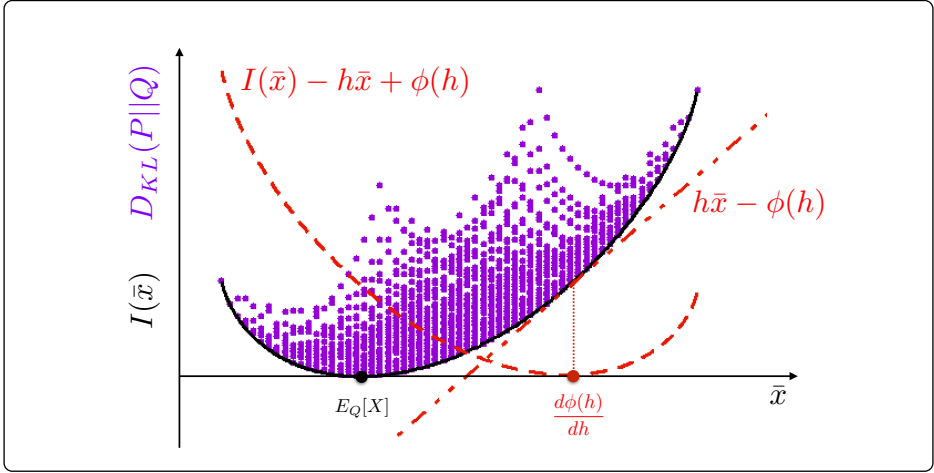
where  $P \in \mathbb{R}^d$  is a  $d$ -dimensional vector and the function  $U(P)$  is concave.<sup>19</sup> In the case of large deviations for distributions with finite support,  $P$  is a distribution,  $U(P) = D_{KL}[P\|Q]$  and  $x(P) = \sum_x P(x)x$  is a linear function of  $P$  (an expected value).  $P$  identifies a point in the  $(x, U)$  plane, with  $x = x(P)$ , and the solution of the problem lies on the boundary in the  $(x, U)$  plane between points that can be achieved for some value of  $P$  and points that cannot be achieved. This boundary is the function  $I(\bar{x})$  that we want to characterise (see Figure 42).

<sup>18</sup>A warmly suggested reading on the Legendre transform, which discusses its geometric interpretation and gives much intuition on its nature, can be found in [31].

<sup>19</sup>I.e.

$$U(\lambda P_1 + (1 - \lambda)P_0) \leq \lambda U(P_1) + (1 - \lambda)U(P_0).$$

for  $\lambda \in [0, 1]$ .



**Figure 42.** Construction of the large deviation function for  $Q(x)$  defined for  $x \in \chi = \{1, 2, 3, 4\}$  and  $Q(1) = 2Q(2) = 4Q(3) = 4Q(4) = 1/2$  and  $n = 20$  points. The red curves show the construction implied by the Legendre transform for  $h = 1$ .

With the introduction of Lagrange multipliers, we transform the problem in Eq. (17.30) into<sup>20</sup>

$$I(\bar{x}) = \min_P \max_h \{U(P) - h[x(P) - \bar{x}]\} \quad (17.31)$$

$$= \max_h \{\bar{x}h - \phi(h)\} \quad (17.32)$$

$$\phi(h) = \max_P \{hx(P) - U(P)\}. \quad (17.33)$$

In this way we relate the original optimisation problem Eq. (17.30) to a *dual* problem Eq. (17.33).

In order to understand the meaning of  $h$ , consider the same problem, but for a value  $\bar{x} + d\bar{x}$  of the constraint. Then if  $P^*(\bar{x})$  is the point where the extreme is achieved in the original problem,

$$I(\bar{x} + d\bar{x}) = U(P^*(\bar{x} + d\bar{x})) = U(P^*(\bar{x})) + \nabla_P U \cdot \delta P^* + \dots$$

where  $\delta P^* = P^*(\bar{x} + d\bar{x}) - P^*(\bar{x})$ . The first order conditions of the optimisation in Eq. (17.31) on  $P$  imply that  $\nabla_P U = h \nabla_P x$ . Hence the equation above reads  $I(\bar{x} + d\bar{x}) = I(\bar{x}) + h \nabla_P x \delta P^* + \dots$ . The equation  $x(P(\bar{x})) = \bar{x}$ , on the other

<sup>20</sup>The fact that the optimisation over  $h$  is a maximisation derives from the fact that it is the solution of the optimisation of a concave function  $hx(P) - U(P)$ . As  $I(\bar{x})$  inherits its concavity from  $U(P)$ ,  $\phi(h)$  inherits its convexity from  $hx(P) - U(P)$ .

hand, implies that  $\nabla_P x \delta P^* = d\bar{x}$ . These, taken together, show that

$$h = \frac{dI}{d\bar{x}}$$

is the slope of the tangent of the curve that is the locus of the set of solutions of the optimisation in the  $(\bar{x}, U)$  plane. This set can equivalently be described by the coordinate  $h$ . Indeed, because of the concavity of  $U(P)$ , the function  $h(\bar{x})$  is an increasing function. Furthermore, this description is totally equivalent to the one in terms of  $\bar{x}$ . If we let  $P(h)$  be the solution of the problem in Eq. (17.33), then one has

$$\begin{aligned} \phi(h + dh) &= x(P(h + dh))(h + dh) - U(P(h + dh)) \\ &= \phi(h) + \bar{x}(h)dh + [h\nabla_P x - \nabla_P U] \delta P + \dots \end{aligned}$$

The term in braces vanishes because of the first order conditions of the problem in Eq. (17.33). Therefore one concludes that

$$\bar{x} = \frac{d\phi}{dh}.$$

Indeed the relation between  $I(\bar{x})$  and  $\phi(h)$  is completely symmetric, i.e.

$$I(\bar{x}) + \phi(h) = \bar{x}h,$$

so  $I$  is the Legendre transform of  $\phi$  and  $\phi$  is the Legendre transform of  $I$ . Indeed, notice that Eq. (17.33) can be rewritten as

$$\phi(h) = \max_{\bar{x}} \left[ h\bar{x} - \min_{P: x(P)=\bar{x}} U(P) \right] = \max_{\bar{x}} [h\bar{x} - I(\bar{x})]. \quad (17.34)$$

The Legendre transform is not a mere change of variables. Rather it is a mapping of the solution  $(\bar{x}, I)$  of a constrained optimisation problem Eq. (17.30) into the solution  $(h, \phi)$  of a dual unconstrained optimisation problem (Eq. (17.33)). The Legendre transform provides a precise prescription for identifying the *conjugate* variable  $h$  that should be used in the transformed problem.<sup>21</sup>

<sup>21</sup>The Legendre transform is the bread and butter of statistical mechanics. As we shall see, the thermodynamics of an isolated system is described by distributions of maximal entropy, which is called the *microcanonical ensemble*. In an isolated system the energy  $E$  is a constant of the motion and hence it is fixed, as well as the volume  $V$  and the number of particles. This problem can be related to the description of a system in equilibrium with its environment (the *heat bath*) removing the constraint on  $E$ . In this description, which is the *canonical ensemble*, the new variable is the temperature  $T$  and the objective function is the free energy  $F = \langle E \rangle - TS$ . Likewise, the constraint on fixed volume  $V$  can be removed with a Legendre transform that maps the problem in one where the pressure  $P$  is fixed, and the constraint on  $N$  can be removed introducing the chemical potential  $\mu$ . As an Exercise, identify in each of these cases what are the variables  $\bar{x}$  and  $h$  and what are the functions  $I(\bar{x})$  and  $\phi(h)$ .

Let us illustrate the properties of  $I(\bar{x})$  for sums  $S_n = \sum_{k=1}^n X_k$  of binary variables that take values  $X_k = \pm 1$  with equal probability. Then, both the recipe above and a direct calculation using Stirling's approximation of the binomial coefficient, show that

$$I(\bar{x}) = \frac{1 - \bar{x}}{2} \ln(1 - \bar{x}) + \frac{1 + \bar{x}}{2} \ln(1 + \bar{x})$$

which is just the relative entropy  $D_{KL}[P_\beta \| Q]$  between the distribution  $P_\beta = (\frac{1-\bar{x}}{2}, \frac{1+\bar{x}}{2})$  and the uniform distribution  $Q = (1/2, 1/2)$ , as it should.

#### Exercise 17.4

Compute  $I(\bar{x})$  in both ways for the case of binary variables discussed in the text.

The expansion for  $|\bar{x}| \ll 1$  yields  $I(\bar{x}) \simeq \frac{1}{2}\bar{x}^2 + O(\bar{x}^4)$  for  $\bar{x} \ll 1$ , which is consistent with the law of large number and the central limit theorem for  $|\bar{x}| \sim 1/\sqrt{n} \ll 1$ . For larger values of  $\bar{x}$  the function  $I(\bar{x})$  provides much more informations on the large deviation properties of the mean  $S_n/n$ . Note that  $I(\bar{x})$  is defined only for  $\bar{x} \in [-1, 1]$ . Indeed also  $|S_n/n| \leq 1$  by definition in this case. Next note that  $I(\pm 1) = \ln 2$ , and indeed the probability that  $S_n = \pm n$  is exactly  $2^{-n}$ .

## 17.4 How much do we learn?\*

Let us go back to our discussion<sup>22</sup> where the distribution  $Q$  encodes our current state of knowledge, i.e. our *theory*. The theory  $Q$  predicts that an observable  $X$  should take a value  $\approx \mathbb{E}_Q[X]$ . When we perform an experiment and measure  $X$ , the measurement may be consistent with this prediction or not. In the latter case we need to revise our theory  $Q$  and replace it by  $P_\beta$ , depending on the observed value  $\bar{x}$  of  $X$ . How much do we learn?

The uncertainty is reduced from  $\mathcal{H}[Q]$  to  $\mathcal{H}[P_\beta]$ . Hence the acquired information is

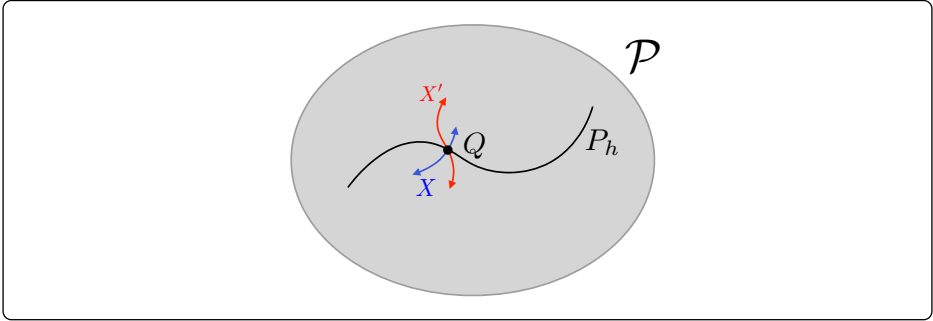
$$-\Delta\mathcal{H} = \mathcal{H}[Q] - \mathcal{H}[P_\beta] \quad (17.35)$$

$$= I(\bar{x}) + \mathbb{E}_Q[(e^{hX - \phi(h)} - 1) \log Q], \quad (17.36)$$

where the second line results from a trite calculation using the results in previous sections. The first term  $I(\bar{x}) = D_{KL}[P_\beta \| Q]$  quantifies how surprising the

<sup>22</sup>This section is a side remark, and it should be taken as a digression for curious students.





**Figure 43.** Probing the space of distributions around  $Q$ . Each experiments  $X$  explores the space along a different trajectory  $P_h$ .

result of the experiment is. The second instead, has the form of a covariance<sup>23</sup> between  $e^{hX - \phi(h)}$  and  $\log Q$ . Hence it depends on what observable  $X$  has been probed in the experiment. This allows us to ask, given  $Q$ , what quantity  $X$  should be probed in order for the experiment to be as informative as possible? Yet  $\Delta\mathcal{H}$  also depends on  $h$ , i.e. on the observed value  $\bar{x}$  of  $X$ . One way to address this question is to “explore the neighbourhood” of  $Q$ , searching for “directions”  $X$  where the reduction in uncertainty  $\Delta\mathcal{H}$  increases faster. Hence we expand  $\Delta\mathcal{H}$  for small values of  $h$  and, after some work, we find<sup>24</sup>

$$\begin{aligned} \Delta\mathcal{H} &\simeq -h \operatorname{Cov}_Q(X, \log Q) \\ &\quad - \frac{1}{2}h^2 \left\{ \mathbb{V}_Q[X] + \operatorname{Cov}_Q \left[ (X - \mathbb{E}_Q[X])^2, \log Q \right] \right\} + O(h^3) \end{aligned}$$

which is an interesting result. The leading linear term implies that the largest change in  $\Delta\mathcal{H}$  occurs when  $X = \log Q$ , which is the  $X$  that maximises the covariance with  $\log Q$ . Note indeed that, by the Asymptotic Equipartition Property, the value of  $-\log Q \approx \mathcal{H}[Q]$  permits to identify the set of typical outcomes.

The choice  $X = \log Q$  explores the space of distributions along the curve of parametric distributions<sup>25</sup>

$$P_h(x) = \frac{1}{\mathbb{E}_Q[Q^h]} Q^{1+h}(x).$$

<sup>23</sup>Note that  $\mathbb{E}_Q[e^{hX - \phi(h)}] = 1$ .

<sup>24</sup>We remind that the covariance is defined as

$$\operatorname{Cov}_Q(X, Y) = \mathbb{E}_Q[(X - \mathbb{E}_Q[X])(Y - \mathbb{E}_Q[Y])]$$

where the index specifies that the expectation is taken with respect to  $Q$ .

<sup>25</sup>In a statistical mechanics analogy, as we shall see  $Q$  takes the form  $Q(x) = \frac{1}{Z} e^{-E(x)/T}$ , where  $T$  is the temperature. Then also  $P_h(x)$  has the same form, with  $T' = T/(1+h)$ . In

The change  $\Delta\mathcal{H}$  can however be either positive or negative, depending on whether  $h < 0$  or  $h > 0$ . In order to make sure that the measurement reduces the uncertainty on the system, the measured quantity  $X$  should be such that  $\text{Cov}_Q(X, \log Q) = 0$ , so that the linear term vanishes.

The first term of order  $h^2$  is  $I(\bar{x}) \simeq \frac{1}{2}h^2\mathbb{V}_Q[X]$ , which suggests that the most potentially surprising experiments, are those that probe quantities with large fluctuations. This is indeed a well established recipe in experimental design.

## 17.5 Weakly correlated variables: phase transitions and the Gartner-Ellis theorem

The results we have derived so far for large deviations extend to the case where the random variables  $X_i$  are weakly dependent. How weak the dependence can be will be clarified below.<sup>26</sup>

Consider the following situation: we have a sample  $X_1, \dots, X_n$  drawn i.i.d. from a distribution, but we're not sure what the distribution is. With probability  $a$  the sample comes from the distribution  $P$  and with probability  $1 - a$  it comes from the distribution  $Q$ . Both  $P$  and  $Q$  have either finite support or thin tails. What is the probability  $P\{A_n(\bar{x})\}$  in this case? Clearly

$$\mathbb{E}[X] = a\mathbb{E}_P[X] + (1 - a)\mathbb{E}_Q[X],$$

where  $\mathbb{E}_P[\dots]$  and  $\mathbb{E}_Q[\dots]$  stand for expectations on the distributions  $P$  and  $Q$ , respectively. Do we expect that the law of large numbers

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow a\mathbb{E}_P[X] + (1 - a)\mathbb{E}_Q[X]$$

holds?

---

addition

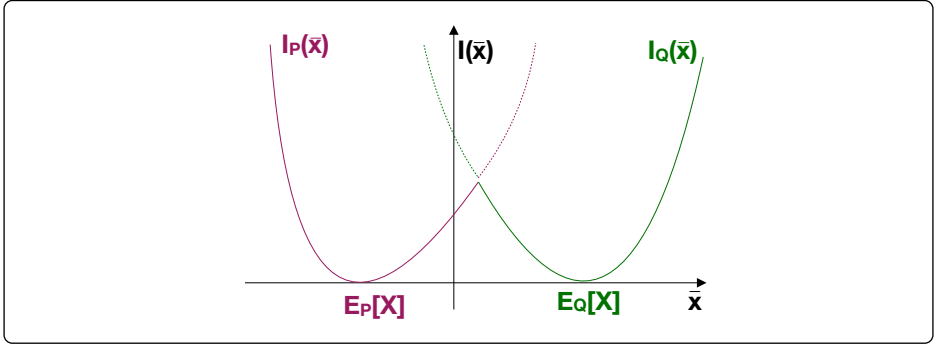
$$\begin{aligned} \Delta\mathcal{H} = & -\frac{h}{T^2}\mathbb{V}_Q[E] \\ & -\frac{h^2}{2T^2}\left[\mathbb{V}_Q[E] + \frac{1}{T}\mathbb{E}_Q[(E - \mathbb{E}_Q[E])^3]\right] + \dots \end{aligned}$$

and the coefficient of the linear term in  $h$  is the specific heat.

<sup>26</sup>To give an idea, one example where the theory applies is when random variables interact only “locally”. This means that for each  $X_i$  there is a finite subset  $\partial_i \subset \{1, \dots, n\}$  of indices such that, conditional on the values of the variables  $X_j$  for  $j \in \partial_i$ ,  $X_i$  is independent of all the other variables  $k \notin \partial_i$ , i.e.

$$P\{X_i|X_j, \forall j \neq i\} = P\{X_i|X_j, \forall j \in \partial_i\}.$$

A Markov process, where  $X_i$  only depends on  $X_{i-1}$  and  $X_{i+1}$  (i.e.  $\partial_i = \{i-1, i+1\}$ ), is a sequence of weakly dependent random variables.



**Figure 44.** The construction of the Cramer function  $I(\bar{x})$  for the example discussed in the text.

The answer can be found by a direct calculation:

$$P\{A_n(\bar{x})\} = aP\{A_n(\bar{x})|P\} + (1-a)P\{A_n(\bar{x})|Q\} \quad (17.37)$$

$$\sim ae^{-nI_P(\bar{x})} + (1-a)e^{-nI_Q(\bar{x})} \quad (17.38)$$

where  $P\{A_n(\bar{x})|W\}$  is the probability of the large deviation, conditional on the assumption that the variables  $X_i$  are drawn i.i.d. from the distribution  $W = P$  or  $Q$ , and

$$I_W(\bar{x}) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n(\bar{x})|W\} = \min_{P \in \mathcal{A}_n(\bar{x})} D_{KL}[P\|W].$$

It is now clear that

$$I(\bar{x}) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n(\bar{x})\} = \min[I_P(\bar{x}), I_Q(\bar{x})]. \quad (17.39)$$

Notice that:

- The curve  $I(\bar{x})$  touches the  $\bar{x}$  axis in *two* points  $\bar{x} = \mathbb{E}_P[X]$  and  $\bar{x} = \mathbb{E}_Q[X]$ . This means that, typically we expect that the sample mean converges to either  $\mathbb{E}_P[X]$  or to  $\mathbb{E}_Q[X]$ , but *not* to  $\mathbb{E}[X]$ . This violation of the law of large numbers occurs because the variables  $X_1, \dots, X_n$  are *not* independent. Indeed, knowledge of a subset  $k$  of the  $X_i$  allows us to infer whether the right distribution is  $P$  or  $Q$ , and hence informs us on the values of the remaining  $n - k$ .
- The curve  $I(\bar{x})$  is not convex. Locally it is convex, apart from the point  $\bar{x}_c$  where  $I_P(\bar{x}_c) = I_Q(\bar{x}_c)$ , where it has a cusp.

- The derivative  $h$  of  $I(\bar{x})$  is no longer a continuous function of  $\bar{x}$ . Rather it has a jump at the point  $\bar{x}_c$ , i.e.  $\lim_{\bar{x} \rightarrow \bar{x}_c^\pm} I'(\bar{x}) = h_\pm$ .
- Following the geometric construction of the function  $\phi(h)$ , one finds that the function  $\phi(h)$  is not single valued in the interval  $h \in [h_+, h_-]$  and that it is not continuous.

**The Maxwell construction and the Gärtner-Ellis theorem.** The fact that  $I(\bar{x})$  derived above is non convex makes the recipe based on the Legendre transform, that we discussed for i.i.d. variables inapplicable. The *Gärtner-Ellis theorem* describes what happens if we apply this recipe anyhow. Suppose that the function

$$\bar{\phi}(h) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E} [e^{h(X_1 + \dots + X_n)}] \quad (17.40)$$

exists and is finite, for  $h$  in a neighbourhood of the origin. Then the convex hull  $\bar{I}(\bar{x})$  of the large deviation function is given by the Legendre transform of  $\bar{\phi}(h)$ .

### Exercise 17.5

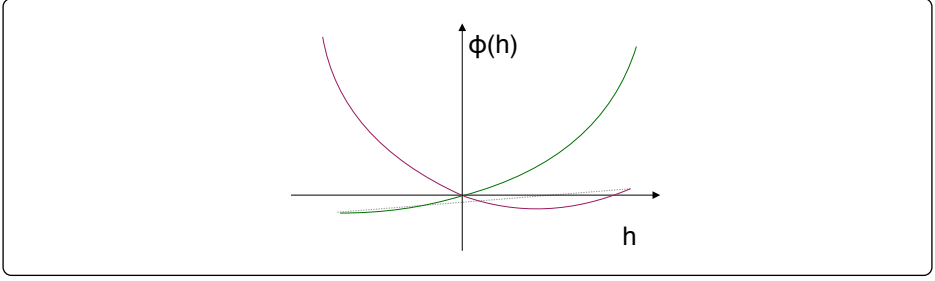
Let  $\underline{X} = (X_1, \dots, X_n)$  where  $X_i = Y_0 Y_i$ , with  $Y_0 = \pm 1$  with equal probability, and  $Y_i \in \{0, 1\}$  are i.i.d. random variables with  $P\{Y_i = 1\} = p = 1 - P\{Y_i = 0\}$ , and they are all independent of  $Y_0$ . Compute the large deviation function for the random variables  $X_i$ , i.e.

$$I(\bar{x}) = - \lim_{n \rightarrow \infty} \frac{1}{n} \log P \left\{ \sum_{i=1}^n X_i \in [\bar{x}, \bar{x} + \epsilon) \right\}$$

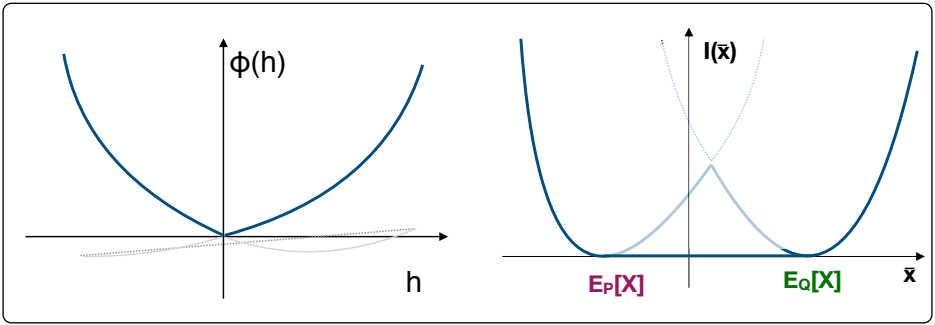
for some  $\epsilon > 0$ . Compute the function  $\bar{I}(\bar{x})$  by Gärtner-Ellis theorem, i.e. as the Legendre transform of  $\bar{\phi}$ . What is the posterior distribution that the true distribution is  $P$ , given  $A_n(\bar{x})$ ?

Let us see how this works for the problem we discussed above, of a sequence  $\underline{X}$  of variables which is drawn i.i.d. from either  $P$  or  $Q$ . It is easy to see that  $\mathbb{E} [e^{h(X_1 + \dots + X_n)} | W] = e^{n\phi_W(h)}$ , where  $\phi_W(h)$  is drawn in Figure 45 for  $W = P$  or  $Q$ . Then

$$\begin{aligned} \bar{\phi}(h) &= \lim_{n \rightarrow \infty} \frac{1}{n} \log [a e^{n\phi_P(h)} + (1-a) e^{n\phi_Q(h)}] \\ &= \max [\phi_P(h), \phi_Q(h)] \end{aligned} \quad (17.41)$$



**Figure 45.** The functions  $\phi_P$  and  $\phi_Q$  for the example discussed in the text.



**Figure 46.** The Gärtner-Ellis theorem applied to the problem of a sequence  $\underline{X}$  drawn i.i.d. from either  $P$  or  $Q$ .

as shown in Figure 46 (left). Notice that  $\bar{\phi}(h)$  has a cusp — i.e. a discontinuity in its first derivative — for  $h = 0$ . The derivative of  $\bar{\phi}(h)$  as  $h \rightarrow 0^+$  equals  $\mathbb{E}_Q[X]$  whereas when  $h \rightarrow 0^-$  one finds  $\bar{\phi}'(h) = \mathbb{E}_P[X]$ .

The Legendre transform  $\bar{I}(\bar{x})$  of  $\bar{\phi}(h)$  is shown in Figure 46 (right). This function  $\bar{I}(\bar{x})$  is identical to  $I(\bar{x})$ , except for the part in the interval  $\bar{x} \in [\mathbb{E}_P[X], \mathbb{E}_Q[X]]$ , where  $I(\bar{x})$  is replaced by a straight line.

The Gärtner-Ellis theorem provides the solution to a different yet related problem, which is the case where an unknown fraction of the variables are drawn from  $P$  and the rest from  $Q$ . Specifically, let  $X_i$  be drawn from  $P$  if  $i \leq \nu n$  and from  $Q$  if  $i > \nu n$ , with  $\nu \in [0, 1]$  which is unknown.

Again we consider the event  $A_n(\bar{x})$ , i.e. that the mean of a sample  $X_1, \dots, X_n$  of points obtained in this way equals  $\bar{x}$ , and we want to compute the probability of  $A_n(\bar{x})$ . The probability of finding a large deviation with a sample mean

equal to  $\bar{x}$  is

$$\begin{aligned} P\{A_n(\bar{x})\} &= \int_0^1 d\nu \int d\bar{x}_P \int d\bar{x}_Q P\{A_{\nu n}(\bar{x}_P)|P\} P\{A_{(1-\nu)n}(\bar{x}_Q)|Q\} \\ &\quad \delta(\bar{x} - \nu\bar{x}_P - (1-\nu)\bar{x}_Q) \\ &\sim \int_0^1 d\nu \int d\bar{x}_P \int d\bar{x}_Q e^{-n[\nu I_P(\bar{x}_P) + (1-\nu)I_Q(\bar{x}_Q)]} \\ &\quad \delta(\bar{x} - \nu\bar{x}_P - (1-\nu)\bar{x}_Q) \end{aligned}$$

where we assume a uniform prior on  $\nu$ . For all values of  $\bar{x} \in [\mathbb{E}_P[X], \mathbb{E}_Q[X]]$  this multiple integral is dominated by the values  $\bar{x}_P = \mathbb{E}_P[X]$  and  $\bar{x}_Q = \mathbb{E}_Q[X]$ , and  $\nu$  such that  $\bar{x} = \nu\mathbb{E}_P[X] + (1-\nu)\mathbb{E}_Q[X]$ , because then  $I_P(\bar{x}_P) = I_Q(\bar{x}_Q) = 0$ , and one finds that

$$I_\nu(\bar{x}) = -\lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n^{(\nu)}(\bar{x})\} = 0 \quad \forall \bar{x} \in [\mathbb{E}_P[X], \mathbb{E}_Q[X]].$$

Put differently, for every  $\bar{x} \in [\mathbb{E}_P[X], \mathbb{E}_Q[X]]$  it is possible to find a value

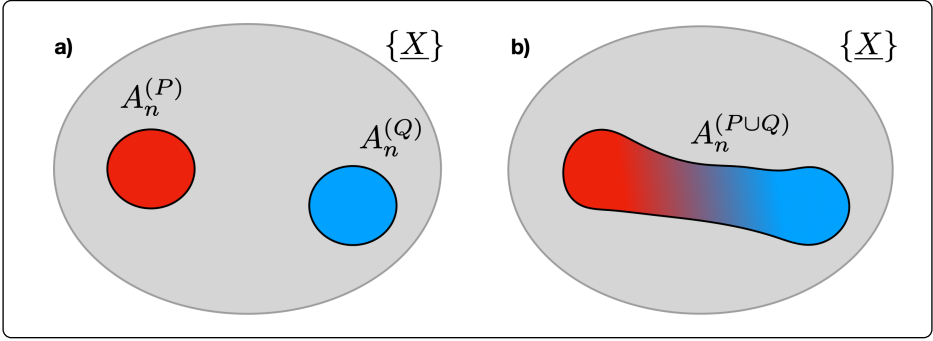
$$\nu = \frac{\mathbb{E}_Q[X] - \bar{x}}{\mathbb{E}_Q[X] - \mathbb{E}_P[X]} \in [0, 1] \quad (17.42)$$

such that the above construction allows us to realise the large deviation  $\bar{x}$  as a typical event (i.e. with  $I_\nu(\bar{x}) = 0$ ).

As we're going to discuss (see footnote 17) the replacement of the non-concave part of  $I(\bar{x})$  with a straight line is conceptually identical to the Maxwell's construction in thermodynamics. In physics this construction relates the thermodynamics of homogenous but unstable states to that of inhomogeneous states, which are a mixture of two homogeneous states. Here, it relates the (large deviation) properties of a system which is either in one *pure* state ( $P$ ) or in another ( $Q$ ), to one which is a *mixture*  $P_\nu = \nu P + (1-\nu)Q$  of the two states. Mathematically, the first case is described by the Cramer function  $I(\bar{x})$  while the mixture is described by its *convex hull*  $\tilde{I}(\bar{x})$ , defined in Eq. (17.39), which is the Legendre transform of  $\tilde{\phi}(h)$  in Eq. (17.41).

### Exercise 17.6

Consider yet a different problem where each of the variables  $X_i$  is drawn from  $P$ , with probability  $\nu$ , or from  $Q$  with probability  $1 - \nu$ . What is the large deviation function  $I(\bar{x})$  in this case when  $\nu$  is known and when  $\nu$  is unknown?



**Figure 47.** Pictorial representation of the space of typical samples  $\underline{X}$  drawn i.i.d. from either  $P$  or  $Q$  (a) or from mixtures  $\nu P + (1 - \nu)Q$  (b).

Notice the difference in the structure of the typical set in the different cases. When the sample is drawn from one of the distributions but we do not know which one, the typical set is the union of two disjoint sets, the typical set of samples generated from  $P$  and of those generated from  $Q$ . When instead each point may be generated from either  $P$  or  $Q$  with unknown probabilities, then the typical set extends to the union of typical sets of all mixtures  $\nu P + (1 - \nu)Q$  for all  $\nu \in [0, 1]$ . This will be an important point when we will discuss statistical inference, which deals with finding those models  $Q$  such that a given data set  $\underline{X}$  may be considered a typical draw.

### 17.5.1 Large deviations for Markov Chains

A further example of a sequence of weakly dependent random variables is given by Markov Chains. Let us recall that a Markov Chain  $Z_0, Z_1, \dots, Z_t, \dots$  is a sequence of random variables that take values in a discrete set  $\mathcal{S}$ , and which is defined by a transition matrix

$$p_{s,s'} = P\{Z_t = s | Z_{t-1} = s'\}, \quad s, s' \in \mathcal{S}. \quad (17.43)$$

We restrict our attention to irreducible Markov Chains for which the distribution  $p\{Z_t = s\}$  converges, as  $t \rightarrow \infty$ , to the unique invariant measure  $\mu_s$  which satisfies the equation  $\mu_s = \sum_{s'} p_{s,s'} \mu_{s'}$ .

For an observable  $X_t$  with a distribution  $P\{X_t = x | Z_t = s\} = q(x|s)$  that depends only on the state  $Z_t$  at time  $t$ , we expect that its time average between times  $\tau + 1$  and  $\tau + N$  converges as  $N \rightarrow \infty$  to the expected value of  $X_t$  on  $\mu_s$ , for  $\tau \rightarrow \infty$ , i.e.

$$\lim_{\tau \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=\tau+1}^{\tau+N} X_t \rightarrow \mathbb{E}_\mu[X_t] \equiv \sum_{x,s} x q(x|s) \mu_s.$$

What is the probability to observe instead a value  $\bar{x}$  different from  $\mathbb{E}_\mu[X_t]$ ? In order to apply Eq. (17.40) we need to compute the expected value

$$\mathbb{E} \left[ e^{h \sum_t X_t} \right] = \sum_{s_\tau, s_{\tau+1}, \dots, s_{\tau+N}} \prod_{t=\tau+1}^{\tau+N} p_{s_t, s_{t-1}} \mathbb{E} \left[ e^{h X_t} | s_t \right] P_0(s_\tau) \quad (17.44)$$

$$= \sum_{s_\tau, s_{\tau+N}} \{ \hat{U}^N \}_{s_{\tau+N}, s_\tau} P_0(s_\tau), \quad (17.45)$$

where  $\{ \hat{U}^N \}_{s_{\tau+N}, s_\tau}$  is the  $s_{\tau+N}, s_\tau$  element of the  $N^{\text{th}}$  power of the matrix  $U_{s,s'} = \mathbb{E} \left[ e^{h X_t} | s \right] p_{s,s'}$ . In the repeated matrix multiplication, the dominant component is the one corresponding to the largest eigenvalue of  $\hat{U}$ , corresponding to the right eigenvector

$$\lambda v_s = \sum_{s'} U_{s,s'} v_{s'} = \sum_{s'} \mathbb{E} \left[ e^{h X_t} | s \right] p_{s,s'} v_{s'} \quad (17.46)$$

which leads to  $\mathbb{E} \left[ e^{h \sum_t X_t} \right] \sim \lambda^N$ . Note that, by virtue of the Perron-Frobenius theorem,  $\lambda$  and all components of  $v_s$  are positive, because, if the chain is irreducible, the matrix  $\hat{U}^N$  has all strictly positive elements for  $N$  large enough. Hence the limit in Eq. (17.40) leads to  $\tilde{\phi}(h) = \log \lambda$ .

Summarising, the recipe of large deviations for a Markov Chain is *i)* compute the matrix  $\hat{U}$ , *ii)* compute its largest eigenvalue  $\lambda$  as a function of  $h$ , *iii)* compute the rate function  $\bar{I}(\bar{x})$  from the Legendre transform of  $\tilde{\phi}(h) = \log \lambda$ . The distribution of  $Z_t$  conditional on the large deviation is given by the normalised right eigenvector

$$P\{Z_t = s | A_n(\bar{x})\} = \frac{v_s}{\sum_{s'} v_{s'}}$$

(which implicitly depends on  $h$ , which is the solution of  $\frac{d\tilde{\phi}}{dh} = \bar{x}$ ). Note that when  $h \rightarrow 0$ , this distribution reverts back to the invariant measure  $\mu_s$ .

## 17.6 Large deviations for fat tailed distributions

The Cramer function  $I(\bar{x})$  has the property that it is positive and it vanishes for  $\bar{x} = \mathbb{E}[X]$ , which corresponds to the point  $h = 0$ . The machinery above works if  $\phi(h)$  exists at least for  $h$  in an open neighbourhood of the origin. This requires that the pdf of  $X$  decays at least as an exponential for  $|x| \rightarrow \infty$ . What happens if this is not true?



We shall call fat tailed distribution any distribution  $Q(x)$  for which

$$\lim_{|x| \rightarrow \infty} \frac{1}{|x|} \log Q(x) = 0 \quad (17.47)$$

for  $x \rightarrow +\infty$  or  $x \rightarrow -\infty$ , or both. In this limit,  $e^{hx}Q(x)$  diverges for at least one value of  $h$  in the neighbourhood of  $h = 0$  as  $x \rightarrow \pm\infty$ .

For simplicity, we focus on the right tail of the pdf, and assume that  $Q(x)$  vanishes at least exponentially fast as  $x \rightarrow -\infty$ . This includes stretched exponential distributions  $Q(x) \sim e^{-ax^\alpha}$  with  $\alpha < 1$  and power law distributions  $Q(x) \sim Ax^{-\gamma}$  for  $x \gg 1$ . Again we focus on the event

$$A_n(\bar{x}) = \left\{ \underline{X} : \left| \frac{1}{n} \sum_{i=1}^n X_i - \bar{x} \right| < \epsilon \right\}$$

for some arbitrarily small  $\epsilon > 0$  and our goal is to compute the Cramer's function  $I(\bar{x})$  in Eq. (17.28). For  $h \leq 0$  we can follow the recipe outlined in the previous sections because  $\mathbb{E}[e^{hX}]$ , and hence  $\phi(h)$ , is finite. This allows us to define the Cramer function  $I(\bar{x})$  for all  $\bar{x} \leq \mathbb{E}_Q[X]$ , which is expected to vanish as  $\bar{x} \rightarrow \mathbb{E}_Q[X]$  with a quadratic behaviour  $I(\bar{x}) \simeq \frac{1}{2\mathbb{V}_Q[X]} (\bar{x} - \mathbb{E}_Q[X])^2 + \dots$  for  $\bar{x} \lesssim \mathbb{E}_Q[X]$ .

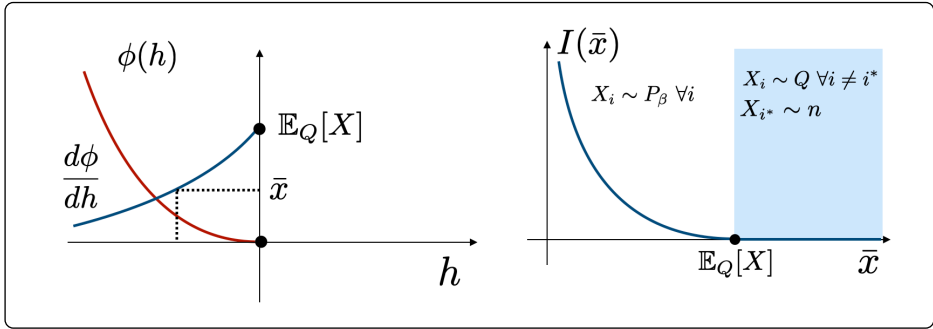
However, for  $h > 0$  this recipe does not work because the integral that defines  $\phi(h)$  diverges. In order to explore the behaviour of  $I(\bar{x})$  for  $\bar{x} > \mathbb{E}_Q[X]$ , let us consider the event

$$\begin{aligned} \tilde{A}_n(\bar{x}) = \bigcup_{i^*=1}^n \left\{ \underline{X} : \left| \frac{1}{n-1} \sum_{i \neq i^*} X_i - \mathbb{E}_Q[X] \right| < \epsilon, X_{i^*} = x_n^* \right\} \\ x_n^* = n\bar{x} - (n-1)\mathbb{E}_Q[X] \end{aligned} \quad (17.48)$$

In words,  $\tilde{A}_n(\bar{x})$  describes a large deviation event where the mean  $\frac{1}{n} \sum_i X_i = \bar{x}$  deviates from the expected value  $\mathbb{E}_Q[X]$ , but all the excess of the mean is concentrated on only one variable  $X_{i^*} = x_n^*$ , which is proportional to  $n$ , whereas all the other variables are “typical”, i.e.  $X_i \approx \mathbb{E}_Q[X]$ . The probability of this event is

$$P\{\tilde{A}_n(\bar{x})\} \geq (1 - \epsilon)nQ(n\bar{x} - (n-1)\mathbb{E}_Q[X])$$

where the factor  $1 - \epsilon$  comes from the fact that the  $n - 1$  variables  $i \neq i^*$  take typical values, the factor  $n$  accounts for the fact that  $i^*$  can take  $n$  values, and the last factor is the probability of  $X_{i^*} = x_n^*$ .



**Figure 48.** Large deviations for a (right) fat tailed distributions: Sketch of the Legendre transform construction for  $h \leq 0$  (left) and the resulting Cramer function (right).

The event  $\tilde{A}_n(\bar{x})$  is only one way in which the large deviation can occur, therefore  $\tilde{A}_n(\bar{x}) \subseteq A_n(\bar{x})$ . As a consequence  $P\{A_n(\bar{x})\} \geq P\{\tilde{A}_n(\bar{x})\}$  and

$$I(\bar{x}) = - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{A_n(\bar{x})\} \quad (17.49)$$

$$\leq - \lim_{n \rightarrow \infty} \frac{1}{n} \log P\{\tilde{A}_n(\bar{x})\} = 0 \quad (17.50)$$

where the last equality is a consequence of Eq. (17.47). Therefore, for all  $\bar{x} \geq \mathbb{E}_Q[X]$  the Cramer function vanishes,  $I(\bar{x}) = 0$ .

In loose words, “democratic” ways to realise large deviations, where  $\bar{x}$  is obtained as the average of i.i.d. draws from a modified distribution, are not typical. For fat tailed distributions, large deviations typically concentrate on a single variable  $X_{i^*}$  which is responsible for the whole excess of the mean  $\bar{x}$ . The symmetry between the variables, which are identically distributed *a priori*, is *broken spontaneously*, because one of them takes an *extensive* value (i.e. a value proportional to  $n$ ). *Spontaneously* refers to the fact that, *a priori*, any variable  $X_{i^*}$  can carry the excess deviation.

The fact that  $I(\bar{x}) = 0$  for all  $\bar{x} \geq \mathbb{E}_Q[X]$  implies that  $I(\bar{x})$  has a singularity at  $\bar{x} = \mathbb{E}_Q[X]$  in the second derivative. This is the analogue of a second order phase transition in statistical physics,<sup>27</sup> that generally occur when a symmetry of the system is spontaneously broken,<sup>28</sup> precisely as in the current situation where the *a priori* (permutation) symmetry between the variables  $X_i$  is broken.

<sup>27</sup>In thermodynamics, the order of a transition is defined as the order of the derivative of the thermodynamic potential that develops a singularity at the critical point. As we shall see,  $I(\bar{x})$  is related to the entropy in statistical mechanics.

<sup>28</sup>The typical example is the spontaneous magnetisation of metals when the temperature is decreased below the Curie temperature.

This phenomenon is similar to the typical behaviour of sums of i.i.d. random variables with a pdf  $p(x)$  that decays slower than  $|x|^{-2}$ . As we have seen, in that case averages are dominated by few variables which are of the same order of the whole sum.<sup>29</sup> Yet in that case the expected value of  $X$  does not exist so large deviations cannot be defined.

---

<sup>29</sup>In the special case where  $X$  are Cauchy variables  $p(x) = \pi^{-1}(1 + x^2)^{-1}$ , you can check that the  $\sum_i X_i/n$  is itself a Cauchy variable. Therefore the probability of a large deviation

$$P\{A_n(\bar{x})\} = \frac{1}{\pi} \frac{1}{1 + \bar{x}^2}$$

does not decay exponentially with  $n$ . Actually it does not decay at all.



# Chapter 18

## States of knowledge

Now that we have a quantitative notion of information, we can address the problem of finding distributions that are consistent with a given state of knowledge. Just like Socrates has been claimed to say that

The only true wisdom is in knowing you know nothing

it seems the only state of knowledge we can precisely identify is the one where we “know nothing”. If lack of information can be measured by the entropy, the state where we know nothing corresponds to a probability distribution of maximal entropy. In addition, as we shall see, large deviation theory allows us to be precise in understanding how new information can be incorporated in our current state of knowledge (i.e. in probability distributions). This “becomes a methodology for a very general type of scientific reasoning”, as claimed by E. T. Jaynes [32]. We shall discuss this general approach and then, statistical mechanics as one of its particular applications.

### 18.1 Maximum entropy

Consider the case of a discrete random variable  $X \in \chi$  drawn from a finite set  $\chi$ . The state of maximal ignorance corresponds to a distribution  $p(x) = P\{X = x\}$  of maximal entropy<sup>1</sup>

$$p(x) = \frac{1}{|\chi|}. \quad (18.1)$$

Indeed, in order to dispel uncertainty the number of binary questions we need to ask is as large as possible, i.e.  $H[X] = \log_2 |\chi|$ . In this state, we’re also

---

<sup>1</sup>You can show this by studying the maximisation of  $\mathcal{H}[p]$  with the normalisation constraint  $\sum_x p(x) = 1$ .

maximally uncertain on what is the best way to ask questions.<sup>2</sup> The state of maximal ignorance is also such that the distribution of  $X$  is invariant under any permutation of the possible values  $x \in \chi$ . This is consistent with a state of knowledge where we don't know anything that can distinguish event  $\{X = x\}$  from event  $\{X = x'\}$ .

Now assume that we know that

$$\mathbb{E}[F(X)] = \sum_{x \in \chi} p(x)F(x) = f \quad (18.2)$$

for a function<sup>3</sup>  $F(X)$ . Then the distribution that encodes this and only this information, is given by the one that maximises the entropy, subject to these constraints. This implies that we have to solve the problem:

$$\max_{p, \lambda, \nu} \left\{ - \sum_{x \in \chi} p(x) \log p(x) + \lambda \left[ \sum_{x \in \chi} p(x)F(x) - f \right] + \nu \left[ \sum_{x \in \chi} p(x) - 1 \right] \right\}.$$

The solution is

$$p_\lambda(x) = \frac{1}{Z(\lambda)} e^{\lambda F(x)} \quad (18.3)$$

where  $Z(\lambda)$  ensures normalisation, and the value of  $\lambda$  should be adjusted in such a way that Eq. (18.2) is satisfied, i.e.

$$\mathbb{E}[F(X)] = \frac{d \log Z}{d \lambda} = f. \quad (18.4)$$

---

<sup>2</sup>In this case, the optimal way to elicit information is to ask questions that split the number of possible alternatives in half each time. If  $|\chi| = 2^H$ , there are  $\binom{2^H}{2^{H-1}}$  ways to choose how to make the first question,  $\binom{2^{H-1}}{2^{H-2}}$  ways to pose the second and so on. In total there are

$$\mathcal{N} = \prod_{k=0}^{H-1} \binom{2^{H-k}}{2^{H-k-1}}^{2^k}$$

ways to ask the  $H$  questions. Which of these ways one chooses to ask questions is irrelevant. If  $p(x)$  were not independent of  $x$ , some of these ways would be better than others. In a state of maximal ignorance there is no clue of how to pose questions in a smart way.

<sup>3</sup>We expect that  $\mathbb{E}[F(X)] = f$  based on theoretical grounds, or this knowledge may come from the fact that, in a series of  $N \gg 1$  independent experiments where we measure the variables  $Y_i = F(X_i)$  for  $i = 1, \dots, N$ , we observe that

$$\frac{1}{N} \sum_{i=1}^N F(X_i) \simeq f,$$

and that we expect the Law of Large Numbers to hold.

Note that the solution to this problem is unique. The way to show this is to observe that  $\lambda$  is the solution of a convex optimisation problem. Indeed Eq. (18.4) corresponds to the first order condition of the maximisation of the entropy as a function of  $\lambda$

$$\Sigma(\lambda) = \mathcal{H}[p_\lambda] = \log Z(\lambda) - \lambda \mathbb{E}[F(X)] .$$

where  $\mathbb{E}[F(X)]$  is a function of  $\lambda$ . Note that

$$\frac{d\Sigma}{d\lambda} = -\lambda \frac{d\mathbb{E}[F(X)]}{d\lambda} = -\lambda \mathbb{V}[F(X)]$$

has the opposite sign of  $\lambda$ , where  $\mathbb{V}[F(X)] \geq 0$  is the variance of  $F(X)$  under the distribution  $p_\lambda$ . So  $\Sigma(\lambda)$  has a unique maximum at  $\lambda = 0$ , because it increases for  $\lambda < 0$  and it decreases for  $\lambda > 0$ .

Yet it is important to stress that the entropy

$$S(f) = \max_{p: \mathbb{E}[F(X)] = f} \mathcal{H}[p] \quad (18.5)$$

is a function of  $f$ , which is the independent variable. The variables  $f$  and  $\lambda$  are conjugate under the Legendre transform that maps the problem Eq. (18.5) into the conjugate problem<sup>4</sup>

$$\psi(\lambda) = \min_p [-\mathcal{H}[p] - \lambda \mathbb{E}[F]] \quad (18.6)$$

The solution of Eq. (18.5) is given by  $S(f) = \log Z(\lambda) - \lambda f$ , where  $\lambda = \lambda(f)$  is given by the solution of Eq. (18.4), whereas the solution of Eq. (18.6) is given by

$$\psi(\lambda) = \min_f [-S(f) - \lambda f] = -\log Z(\lambda). \quad (18.7)$$

The function  $\psi$  is not an entropy.<sup>5</sup> It is called a *free energy*.

---

<sup>4</sup>This follows from

$$\begin{aligned} S(f) &= \min_{\lambda} \max_p \{ \mathcal{H}[p] + \lambda (\mathbb{E}[F] - f) \} \\ &= \min_{\lambda} \left\{ -\lambda f - \min_p [-\mathcal{H}[p] - \lambda \mathbb{E}[F]] \right\} \\ &= \min_{\lambda} \{ -\lambda f - \psi(\lambda) \} \end{aligned}$$

<sup>5</sup>Note that  $-\psi(\lambda)$  is the cumulant generating function of the random variable  $F(X)$ .

Summarising, the maximisation of the entropy at a fixed value of  $f = \mathbb{E}[F]$  corresponds to the minimisation of the free energy  $\psi$  at a fixed value of the conjugate parameter  $\lambda$ . Because of this

$$\lambda(f) = -\frac{dS}{df} \quad \text{and} \quad f(\lambda) = -\frac{d\psi}{d\lambda} \quad (18.8)$$

and the functions  $S$  and  $\psi$  stand in the relation  $S + \psi = -\lambda f$ .

### Exercise 18.1

The construction discussed in this section is identical to the one we have followed in large deviation theory. for  $Q(x) = 1/|\mathcal{X}|$ . What is the relation between the parameters  $h, \bar{x}$  and  $\lambda, f$ , and between the functions  $I, \phi$  and  $S, \psi$ ?

This construction generalises in a straightforward manner to the case where  $F(X) = (F_1(X), \dots, F_K(X))$  is a vector of  $K$  observables and  $f = (f_1, \dots, f_K)$  is a vector of measurements. The solution of the maximisation of the entropy is again given by Eq. (18.3) with  $\lambda = (\lambda_1, \dots, \lambda_K)$  being a vector of parameters, fixed by eqs. (18.4), where the derivative is replaced by the gradient, and  $\lambda F(x) = \sum_k \lambda_k F_k(x)$  is given by the dot product.

There are several ways to see that Eq. (18.3) is the correct choice that encodes only the information that  $\mathbb{E}[F(X)] = f$  in the probability of  $X$ , as discussed in [33]. Let us discuss one of them. Imagine the situation where you have a sample of  $n \gg 1$  values of  $X$ , that you think are drawn from a distribution  $p(x)$ . Then the analogous of Eq. (18.2) is

$$\bar{f} = \frac{1}{n} \sum_{i=1}^n F(X_i) = \sum_{x \in \mathcal{X}} P_{\underline{X}}(x) F(x). \quad (18.9)$$

where  $P_{\underline{X}}(x)$  is the fraction of times that the outcome  $x$  occurs in the sample  $\underline{X} = (X_1, \dots, X_n)$ . The number of samples  $\underline{X}$  that correspond to a given  $P_{\underline{X}} = P$  is

$$\left| \{ \underline{X} : P_{\underline{X}} = P \} \right| = \frac{n!}{\prod_x [nP(x)]!} \simeq e^{n\mathcal{H}[P]}$$

where the second relation is a trite application of Stirling's formula. Then it is clear that, among all the possible distributions  $P$  that are consistent with Eq. (18.9) those for which  $\mathcal{H}[P]$  is maximal correspond to an overwhelmingly larger number of samples. So the probability that the observed sample is not one of these, is negligibly small as  $n \rightarrow \infty$ .



Distributions of maximal entropy are special because the probability of a sample  $\underline{X} = (X_1, \dots, X_n)$

$$p^{(k)}(\underline{X}) = \frac{1}{Z^n} \exp \left\{ \sum_k \lambda_k \sum_{i=1}^n F_k(X_i) \right\}$$

depends on the data *only* through the empirical averages

$$\hat{f}_k(\underline{X}) = \frac{1}{n} \sum_{i=1}^n F_k(X_i)$$

of  $F_k$ . Therefore these averages contain all the information that is needed to identify the parameters  $\lambda$  of the distribution  $p^{(k)}$ . All other information in the sample is uninformative noise. This is why the empirical averages  $\hat{f}_k$  are called *sufficient statistics*. This should not be surprising. Indeed, the distribution  $p^{(k)}$  has been derived precisely as the one that encodes the state of knowledge in which the values of  $F$ , and only these, are known.

### 18.1.1 Generalised thermodynamics

**Equilibrium:** the principle of maximum entropy can also be applied to a system composed of two or more parts, of which we know the value of an aggregate quantity. More precisely, let  $X_{1+2} = (X_1, X_2)$  be the variables that specify the state of the combined system, where  $X_i$  are the variables of subsystem  $i$  (with  $i = 1$  or  $2$ ). These can be the coordinates that specify microscopic states in physical systems, but we shall deal with them as (vectors of) random variables in the general case. Let  $F(X)$  be an additive quantity.<sup>6</sup>  $F(X_{1+2}) = F(X_1) + F(X_2)$ . Then the state of maximal entropy where this quantity takes a specific expected value

$$f_{1+2} = \mathbb{E} [F(X_{1+2})] \quad (18.10)$$

is given by the maximum entropy distribution

$$p_{1+2}(x) = \frac{1}{Z(\lambda)} e^{\lambda[F(x_1)+F(x_2)]} = p_1(x_1)p_2(x_2). \quad (18.11)$$

where the variables  $X_1$  and  $X_2$  are independent, which is indeed consistent with a maximum entropy state. In addition, the distribution of the states of

<sup>6</sup>In physics, additive quantities are proportional to the size of the system and they are called *extensive*. Examples include the entropy, the volume, the energy and the number of particles. Variables that are independent of the system's size — such as the temperature, the pressure and the particle density — are called *intensive*.

the subsystems are also maximum entropy states  $p_i(x_i) = \frac{1}{Z_i(\lambda_i)} e^{\lambda_i F(X_i)}$ . This is again consistent with the principle of maximum entropy. Furthermore, the conjugate variable takes the same value of  $\lambda_i = \lambda$ . This is again a consequence of maximum entropy. Indeed the entropy  $H[X_{1+2}] = S_{1+2}(f_{1+2})$  is related to the entropy of the subsystems  $H[X_i] = S_i(f_i)$  by the relation

$$S_{1+2}(f_{1+2}) = \max [S_1(f_1) + S_2(f_{1+2} - f_1)] \quad (18.12)$$

The first order condition of this maximisation problem requires that  $f_1$  be such that

$$\frac{d}{df_1} [S_1(f_1) + S_2(f_{1+2} - f_1)] = \frac{dS_1}{df_1} - \frac{dS_2}{df_2} \Big|_{f_2=f_{1+2}-f_1} = 0.$$

This, in view of Eq. (18.8) applied to each subsystem, implies

$$\lambda_1 = \lambda_2 = \lambda \quad (18.13)$$

In words, the maximum entropy principle is associated to a notion of *equilibrium* where each of the parts has the same value of the conjugate variables  $\lambda_i$ . In physics, conjugate variables of extensive variables are called *intensive*, meaning that they are independent of system size. This is because thermodynamic potentials — i.e. the functions  $S$  and  $\psi$  — are themselves extensive, so the conjugate variable to an extensive variable cannot be extensive. In a maximum entropy equilibrium all the intensive variables take the same value in each part of the subsystem. In other words, equilibrium states are homogeneous. This is called the *zeroth law* in thermodynamics. This generalises to systems composed of many parts  $X_\ell$ ,  $\ell = 1, \dots, L$  in a straightforward manner.

**The first law of thermodynamics:** consider now a different problem where the observables  $F_k(X)$  change slightly, i.e.  $F_k \rightarrow F_k + \delta F_k$  and the measurement also changes  $f_k \rightarrow f_k + \delta f_k$ . This transformation involves an arbitrary (infinitesimal) change of both the “internal” parameters  $F_k$  and of the “external” variables  $f_k$ , and it can be regarded as a generalised infinitesimal “thermodynamic” transformation. The new system is described by new parameters  $\lambda'_k = \lambda_k + \delta \lambda_k$ , which are again given by the solution of eqs. (18.4).

The change in the entropy, to leading order, can be written as<sup>7</sup>

$$\delta\mathcal{H} = \mathcal{H}[p_{\lambda+\delta\lambda}] - \mathcal{H}[p_{\lambda}] \simeq \sum_{k=1}^K \lambda_k \delta Q_k \quad (18.14)$$

where

$$\delta Q_k = -\delta f_k + \mathbb{E}[\delta F_k(X)] \quad (18.15)$$

is a generalised “heat”, that is composed of two parts. The first is due to the action  $\delta f_k$  of the external variables on the system and the second is the change of the internal observables. Put differently, the change  $\delta f_k$  of  $f_k$  in any transformation between maximum entropy states is given by two terms, one is the “work”  $\mathbb{E}[\delta F_k(X)]$  done on the system and the other is due to the change  $\delta Q_k$  in the information content. Eq. (18.15) is the analog of the *first law of thermodynamics* in physics.

### 18.1.2 Maximum entropy learning\*

Maximal entropy — sometimes called *maxent* — provides a procedure to learn theories from data. Imagine we’re interested to acquire knowledge about an unknown quantity  $X$ , that we know takes values in a finite set  $X \in \chi$ . Our goal is to learn the distribution  $p(x) = P\{X = x\}$  and to reduce our uncertainty about  $X$ . If we’re in a state of total ignorance about  $X$  then our starting point is the maximum entropy distribution  $p^{(0)}(x) = 1/|\chi|$ . Imagine that we make an experiment and measure<sup>8</sup> the observable  $\mathbb{E}[Y_1] = \mathbb{E}[f_1(X)]$ . If the value  $f_1 = \mathbb{E}[Y_1]$  that we obtain is consistent with the theory, i.e. if

$$f_1 = \sum_{x \in \chi} p^{(0)}(x) F_1(x)$$

---

<sup>7</sup>The entropy at the maximum is given by

$$\mathcal{H}[p_{\lambda}] = -\lambda f + \log Z(\lambda)$$

where  $\lambda f = \sum_k \lambda_k F_k$  stands for the scalar product. The change in the first term is given by  $\delta(\lambda f) = \delta\lambda f + \lambda \delta f$ . The change in the second term instead is given by  $\delta \log Z = \delta\lambda \mathbb{E}[F] + \lambda \mathbb{E}[\delta F]$ , where expected values are taken with respect to  $p_{\lambda}$ , and hence  $\mathbb{E}[F] = f$  so that the terms proportional to  $\delta\lambda$  cancel.

<sup>8</sup>For example, we can take a sample  $\underline{Y}_1 = (Y_1^{(1)}, \dots, Y_1^{(N)})$  and estimate

$$\mathbb{E}[Y] \simeq \frac{1}{N} \sum_{i=1}^N Y_1^{(i)},$$

if  $N$  is very large.

then the experiment confirms the theory. If it does not, then, in order to include this observation, the theory has to be modified as

$$p^{(1)}(x) = \frac{1}{Z^{(1)}(\lambda_1)} e^{\lambda_1 F_1(x)},$$

where  $\lambda_1$  has to be fixed so that  $\sum_{x \in \mathcal{X}} p^{(1)}(x) F_1(x) = \frac{\partial \log Z^{(1)}}{\partial \lambda_1} = f_1$ . This procedure can be repeated by performing further experiments on other observables  $Y_k = F_k(X)$ , for  $k = 2, 3, \dots$ . At each step, if the prediction of the current theory  $p^{(k-1)}$  does not match the outcome  $f_k$  of the experiment, i.e. if  $f_k \neq \sum_{x \in \mathcal{X}} p^{(k-1)}(x) F_k(x)$ , then the theory has to be refined  $p^{(k-1)} \mapsto p^{(k)}$  with the procedure given above. In this way the theory  $p^{(k)}$  encodes, at each step, all the knowledge that has been accumulated in past experiments. Notice that if  $\lambda_k = 0$  then  $p^{(k)} = p^{(k-1)}$ .

The entropy  $\mathcal{H}[p^{(k)}]$  is clearly a non-increasing function of  $k$ , so it generally decreases in the process of refining the theory.<sup>9</sup> The difference  $H[p^{(k-1)}] - H[p^{(k)}]$  is the amount of information that is learned in the  $k^{\text{th}}$  step.

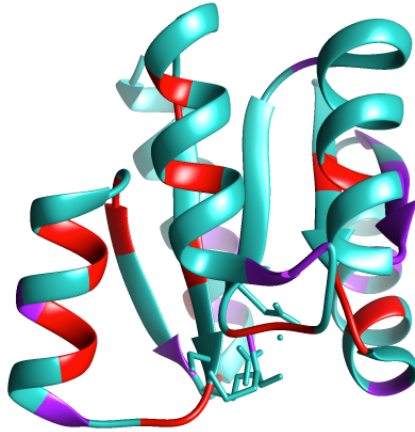
There are different ways in which the principle of maximal entropy enters statistical inference. For example, one should be aware that each statistical method which are based on the covariance of the data — as e.g. principal component analysis or K-means clustering — implicitly assume that the data follows Gaussian statistics. Indeed, the conclusions drawn from these methods would be exactly the same if the data were drawn from a Gaussian distribution that reproduces the empirical covariance. All information contained on higher order statistics (e.g. three point correlations) is lost.

In other situations maximum entropy distributions are assumed precisely because one intends to focus on specific properties. For example, in the problem of the reconstruction of the three dimensional structure of proteins from their sequence, one can assume that the stability of the structure depends on the presence of contacts between amino acids that attract each other. These are amino-acids which are close in space even if they are far apart along the sequence. Because of their relevance for the stability of the three dimensional structure, these amino-acid pairs should be conserved by evolution, or rather they should co-evolve. This means that a mutation on one of them should be accompanied by a compensatory mutation on the other.

In a data set of many sequences of proteins with the same structure, this

---

<sup>9</sup>Remember our discussion on the mutual information: the knowledge of a random variable  $Y$  decreases our uncertainty on  $X$  *a priori*, but *a posteriori* there may be values of  $Y$  such that the entropy of  $X$  is actually larger. Why is this not the case in the situation we're discussing here?



**Figure 49.** The three dimensional structure of a protein.

reflects in the distribution of pairwise correlations between amino-acids suggesting that contacts can be identified by fitting a model of pairwise interacting amino-acids on the data. For more information on this, see [34].

### 18.1.3 Continuous variables

It seems natural to generalise the discussion above to continuous variables  $X$  with pdf  $p(x)$ , by adopting the differential entropy  $h[X]$  instead of  $H[X]$  and replacing partial with functional derivatives. So, for example, the distribution of maximal (differential) entropy for  $X \in [0, \infty)$  with  $E[X] = \mu$  is the exponential  $p(x) = \mu^{-1}e^{-x/\mu}$  and the distribution of maximal entropy for  $X \in \mathbb{R}$  with  $E[X] = \mu$  and  $V[X] = \sigma^2$  is the Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The main problem with this approach is that re-parametrisation invariance is lost. Imagine two observers that want to make inference on the same system and measure the same quantity  $\phi$ . Yet the first observer represent the observables  $\phi(x)$  as a function of  $x$  and the second as a function of  $y$ , where  $y = f(x)$ , with  $f(x)$  a strictly increasing function of  $x$ . Hence, the second observer represents the same quantity with a different function  $\tilde{\phi}(y) = \phi(f^{-1}(y))$ . On

the basis of the same data  $\underline{\phi} = (\phi_1, \dots, \phi_n)$  and the same measurement

$$\bar{\phi} = \frac{1}{n} \sum_{i=1}^n \phi_i$$

their states of knowledge would be encoded in the two distributions

$$p(x) = \frac{1}{Z} e^{\theta \phi(x)}, \quad \tilde{p}(y) = \frac{1}{\tilde{Z}} e^{\tilde{\theta} \bar{\phi}(y)}$$

respectively, where we assume that the two distributions are normalisable (i.e.  $Z, \tilde{Z} < +\infty$ ). Yet these correspond to two different states of knowledge. Indeed, by a change of variable, the pdf  $\tilde{p}(y)$  for the second observer would correspond to

$$\tilde{p}(x) = \tilde{p}(f(x)) \frac{df(x)}{dx} = \frac{1}{\tilde{Z}} e^{\tilde{\theta} \phi(x)} \frac{df(x)}{dx}$$

which is different from  $p(x)$ . Indeed it is not even a maximum (differential) entropy distribution.<sup>10</sup> Indeed, the two observers maximise two different functions  $h[X]$  and  $h[Y]$  subject to the same constraint. It is no wonder that their states of knowledge are different. The problem is that for continuous variables the differential entropy does not provide a way to encode a state of complete ignorance, rather it allows us only to quantify changes in our state of knowledge. The issue of how to represent, from first principles, a state of ignorance for continuous variables, corresponds to the problem of choosing the non-informative prior in Bayesian statistics that is discussed in [8]. The bottom line is that, when possible, symmetries of the problem can be used to determine the prior. In order to give a flavour of the argument, imagine we want to find the pdf  $p_0(x)$  that encodes the state of complete ignorance for a random variable  $X \in \mathbb{R}$ . We shall call this a *prior* because this pdf represents what is known on  $X$  *before* we make any measurement. Imagine two observers, one that measures the variable  $X$  and the other that measures  $Y = X + a$ , with  $a \in \mathbb{R}$  a constant. Because of translation invariance, the state of knowledge of the two observers must be the same, they both have no clue of what the value of  $X$  (or  $Y$ ) is, i.e. they should both use the same prior  $p_0$ . They must also assign the same probability  $p_0(x)dx = p_0(y)dy$  to the same intervals of  $X$ . This means that  $p_0(x) = p_0(x + a)$  for all values of  $a$ , which means that

$$p_0(x) = c$$

<sup>10</sup>For discrete variables  $X$  this problem does not arise. Both observers assign the same probabilities to corresponding values of  $X$  and  $Y$ , because  $f$  is a bijection between discrete values.

must be a constant. The problem is that in order for this pdf to be normalisable one should have  $c \rightarrow 0$ , i.e.  $p_0(x)$  is an *improper prior*.

### Exercise 18.2

Using the same argument, show that the prior that encodes a state of complete ignorance on a positive real random variable  $X > 0$  is  $p_0(x) = c/x$ . This is again an improper prior.

In order to understand the origin of the problem, let's go back to the discrete case. There, the state of complete ignorance is the one which is further away from the state of complete knowledge  $X = x$ , in terms of the minimal number of binary questions that need to be asked to determine  $X$ . If  $X$  is continuous, it is clear that the minimal number of binary questions should be infinite. This tallies with the fact that, when symmetries can be used, one finds improper (i.e. non normalisable) priors, i.e. priors for which  $h[X] = +\infty$ .

Even if it is disturbing, the fact that  $p_0$  is not normalisable, does not prevent us from using it in learning. Imagine indeed that we collect a sample  $\underline{\phi}$  of  $N$  independent observations of the variable  $\phi(X)$ , and we observe that

$$\frac{1}{N} \sum_{i=1}^N \phi(X_i) = \bar{\phi}.$$

Then we can use the machinery of large deviation theory to incorporate this information in the state of knowledge  $p_0$ . Formally, the updated state of knowledge now would read

$$p(x|\bar{\phi}) = \frac{p_0(x)e^{\lambda\phi(x)}}{Z(\lambda)}, \quad Z(\lambda) = \int_{-\infty}^{\infty} dx p_0(x) e^{\lambda\phi(x)}. \quad (18.16)$$

If we substitute  $p_0(x) = c$ , the constant  $c$  cancels in both the numerator and the denominator. So the fact that  $p_0(x)$  is improper, does not prevent  $p(x|\bar{\phi})$  to be a proper pdf, provided that  $Z(\lambda) < \infty$ .<sup>11</sup>

<sup>11</sup>A limiting procedure that could be applied is to limit the values of  $X$  to the interval  $[-1/(2c), 1/(2c)]$ , do the calculation, and then let  $c \rightarrow 0$ . This is an example of a *regularisation*, a technique used to remove singularities from a problem. The prior  $p_0$  should be invariant under affine transformation  $X' = a + bX$  for all  $a \in \mathbb{R}$  and all  $b > 0$ . This suggests that location and scale of a random variable  $X$  both need improper priors and both introduce a singularity that needs to be regularised. An interesting question, which is open to the best of my knowledge, is: are these the only (primitive) singularities or can there be other ones?

Yet there's another problem with Eq. (18.3). Take the example where our current state of knowledge  $p^{(0)}$  implies that  $X$  is a Gaussian random variable with mean  $\mu$  and variance  $\sigma^2$ . On the basis of this, you would predict that  $S = E[X^3]$  should take the value  $S = \mu^3 + 3\mu\sigma^2$ . Imagine you observe that  $S$  is significantly different from this value. What should you conclude?

If you try to incorporate this information in the distribution, you end with a distribution

$$p^{(1)}(x) = \frac{1}{Z} e^{\lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3}$$

that cannot be normalised, so the recipe of maximum entropy fails.

There is a way to accommodate the observation  $S \neq \mu^3 + 3\mu\sigma^2$  that requires a minimal modification of the distribution  $p^{(0)}$ . Take

$$\tilde{p}^{(1)}(x) = \epsilon \delta(x - \Lambda) + (1 - \epsilon) p^{(0)}(x)$$

then, a trite calculation leads to

$$\mathbb{E}[X] = \mu + \epsilon(\Lambda - \mu) \quad (18.17)$$

$$\mathbb{V}[X] = \sigma^2 + \epsilon[\sigma^2 + (1 - \epsilon)\mu(\mu - 2\Lambda)] \quad (18.18)$$

$$\mathbb{E}[X^3] = \mu^3 + 3\mu\sigma^2 + \epsilon[\Lambda^3 - 3\mu\sigma^2 - 3\mu^3] . \quad (18.19)$$

If we take

$$\Lambda = (S - \mu^3 - 3\mu\sigma^2)^{1/3} \epsilon^{-1/3}$$

then in the limit  $\epsilon \rightarrow 0$  we recover all the three observed moments. At the same time, in this limit,  $p^{(1)} \rightarrow p^{(0)}$  which is the original distribution. Formally this is correct, but what does it mean?

The fact that  $h[\tilde{p}^{(1)}] = h[p^{(0)}]$  implies that the observation on  $S$  does not dispel any uncertainty on  $X$ .<sup>12</sup> The distribution  $p^{(1)}$  can be realised by a sample of  $n \sim \epsilon^{-1}$  observations of  $S_i = X_i^3$ , in  $n - 1$  of which,  $X_i$  is a typical draw from  $p^{(0)}$ , and one of them takes value  $X_{i^*} = \Lambda \sim n^{1/3}$  which is very large. All this is reminiscent of the discussion we had concerning large deviations of fat tailed distributions.

Indeed the pdf of  $S$ , behaves asymptotically as

$$P\{S \in [s, s + ds]\} \sim e^{-c|s|^{2/3}} ds, \quad |s| \rightarrow \infty .$$

<sup>12</sup>This discussion suggests that statistical analysis should be carried out on variables whose distribution has thin tails. For example, gene expression is measured in experiments based on PCR (Polymerase Chain Reaction), which is a method by which a weak signal is amplified in a multiplicative process. As a result, the outcome of PCR is a concentration (of mRNA) which has a very broad distribution. For this reason, it is customary to base statistical analysis on the logarithm of the concentration and to discuss gene activation or suppression in terms of fold-increase or decrease of the measured concentration.



Therefore  $S$  has a fat tailed distribution. As we have seen in the previous chapter, we expect that large deviations (or unexpected events) of samples drawn from such distributions occur in a peculiar manner, where one of the points in the sample attains an anomalously large (or small) value, whereas all the others take typical values. In this situation, the observation on  $S$  cannot change the state of knowledge on the variable  $X$ .

This indicates what type of observables will bring new information, in the sense that unexpected events allow us to update our state of knowledge on  $X$ , and what observables do not. This suggests that it is useless to sample observables which have a fat tailed distribution under the current state of knowledge, if our goal is to test a theory  $p^{(0)}$ .

#### 18.1.4 What can we learn?

Remember our discussion on complex systems that maximise a complex function  $U(\underline{s}, \bar{s})$  over a set of variables  $\vec{s} = (\underline{s}, \bar{s})$  which are known only in part, because  $\bar{s}$  are *unknown unknowns*. We concluded that the probability to observe a certain value  $\underline{s}$  is given by

$$P\{\underline{s}^* = \underline{s}\} = \frac{1}{Z(\beta)} e^{\beta u_{\underline{s}}},$$

where  $u_{\underline{s}} = \mathbb{E}[U(\underline{s}, \bar{s}) | \underline{s}]$  is the known part of the function that is optimised and  $\beta > 0$  depends on the optimisation over unknown variables.

If we do not know the function  $u_{\underline{s}}$ , can we use the procedure outlined above to learn it? In other words, can the function  $u_{\underline{s}}$  be learned from a series of experiments?

Let  $p^{(0)}(\underline{s})$  be the distribution that encodes the current state of knowledge about the system. For a quantity  $q_{\underline{s}}$ , it is possible to compute its distribution

$$p^{(0)}(q) = \sum_{\underline{s}} p^{(0)}(\underline{s}) \delta(q - q_{\underline{s}})$$

Imagine running an experiment where the value  $q_{\text{exp}}$  is measured. In particular, for a complex system, we can assume that  $\underline{s}$  is a high dimensional vector of weakly dependent variables. So that the distribution of  $q$  should be sharply peaked around its expected value  $\mathbb{E}^{(0)}[q] = \sum_{\underline{s}} p^{(0)}(\underline{s}) q_{\underline{s}}$ , and hence  $q_{\text{exp}} \approx \mathbb{E}^{(0)}[q]$ .

If  $q_{\text{exp}} \approx \mathbb{E}^{(0)}[q]$  within experimental errors, then the state of knowledge  $p^{(0)}$  does not need to be updated. Otherwise it has to be revised.<sup>13</sup> In the latter

<sup>13</sup>There is a long tradition of experiments designed to test our state of knowledge in physics. For example, until 1964, we expected that the laws of Nature should be invariant under

case, the standard recipe to update  $p^{(0)}$  is given by Large Deviation Theory. This maintains that the new distribution should be such that  $\mathbb{E}^{(1)}[q] = q_{\text{exp}}$ , without assuming anything else. More precisely, the amount of information that the measurement gives on the state  $\underline{s}$  is given by the mutual information  $I(\underline{s}, q) = D_{KL}[p^{(1)} \| p^{(0)}]$ . Hence,  $p^{(1)}$  should be the distribution with  $\mathbb{E}^{(1)}[q] = q_{\text{exp}}$  for which  $D_{KL}[p^{(1)} \| p^{(0)}]$  is minimal. The distribution that satisfies this requirement is

$$p^{(1)}(\underline{s}) = \frac{1}{Z(g)} p^{(0)}(\underline{s}) e^{g q_{\underline{s}}}, \quad Z(g) = \int dq p^{(0)}(q) e^{g q} \quad (18.20)$$

where  $g$  is adjusted in such a way to satisfy  $\mathbb{E}^{(1)}[q] = q_{\text{exp}}$ . This process can be continued with additional measures of different observables  $q'_s, q''_s, \dots$ , and, in principle, it leads to infer

$$\beta u_{\underline{s}} = \log p^{(0)}(\underline{s}) + g q_{\underline{s}} + g' q'_s + g'' q''_s + \dots \quad (18.21)$$

to the desired accuracy from a series of experiments.

This recipe, however, only works for quantities which have a distribution which falls off faster than exponential as  $q \rightarrow \pm\infty$ . If  $-\log p^{(0)}(q) \simeq c|q|^\gamma$  for  $|q| \rightarrow \infty$  with  $\gamma < 1$ , then the integral defining  $Z(g)$  in Eq. (18.20) is not defined. There is no well defined way to incorporate an observation  $q_{\text{exp}} \neq \mathbb{E}[q]$  in the current state of knowledge in this case. This clearly applies to  $u_{\underline{s}}$  itself. The only models  $u_{\underline{s}}$  that can be learned are those for which the density of states

$$\mathcal{N}(u) du = |\{\underline{s} : u_{\underline{s}} \in [u, u + du)\}|$$

has thin tails, i.e. decays like or faster than an exponential as  $u \rightarrow \infty$ . In this sense, systems where  $\mathcal{N}(u)$  have an exponential behaviour with  $u$  are special, because they separates the region of learnable systems — those for which  $\mathcal{N}(u)$  has thin tails — from unlearnable ones — those where  $u_{\underline{s}}$  has a fat tailed distribution. Interestingly, these are the systems that are best at learning according to [35, 36].

---

time reversal  $T$ . The *CPT* theorem states that the laws of Nature should be invariant under the combined transformation *CPT*, where *C* stands for charge conjugation and *P* for parity transformations. The discovery of the violation of the *CP* symmetry in experiments on the decays of neutral kaons, changed our state of knowledge in particle physics.

## Chapter 19

# Statistical mechanics

Statistical mechanics describes the macroscopic behaviour of systems of many particles. Let us briefly recall the standard approach to statistical mechanics, following ref. [37] — to which we refer as LANDAU— or [38] — to which we refer as KARDAR. A *configuration*  $\underline{X} = (X_1, \dots, X_n)$  of a physical systems is a vector of the  $n$  coordinates  $X_i$  of the particles, where  $n \approx 6 \cdot 10^{23}$  is of the order of Avogadro's number.<sup>1</sup> The coordinates  $X_i$  satisfy Newton's law of classical mechanics, that defines a trajectory  $\underline{X}(t)$  of the configuration in *phase space*  $\Gamma$ . These provide a complete microscopic description of the system. Statistical mechanics aims at deriving a statistical description from mechanics, in which the information on the microscopic configuration  $\underline{X}$  is lost. The idea of statistical mechanics is that time averages can be replaced by statistical averages on an ensemble of systems with a distribution  $p(\underline{X})$  on  $\Gamma$ . This means that, for any observable  $O(\underline{X})$ ,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_{t_0}^{t_0+T} dt O(\underline{X}(t)) = \int_{\Gamma} d\underline{X} p(\underline{X}) O(\underline{X}). \quad (19.1)$$

The objective of statistical mechanics is to compute  $p(\underline{X})$ , so that the values of observables can be computed theoretically.

For an isolated systems that is not subject to external forces, the energy  $E(\underline{X})$  is a constant of the motion. So the dynamics  $\underline{X}$  only spans the manifold of  $\Gamma$  with a fixed value of  $E(\underline{X}) = U$ . In addition, Liouville's theorem<sup>2</sup> ensures that the probability  $p(\underline{X})$  is also a constant of the motion. Under the *ergodic*

---

<sup>1</sup>In classical mechanics the coordinates  $X_i = (q_i, p_i)$  of a particle specify its position  $q_i$  and its momentum  $p_i$ , each of which is a  $d$  dimensional vector.

<sup>2</sup>The Liouville theorem concerns the evolution of the probability distribution  $p(\underline{X}, t)$  under Hamiltonian dynamics.

*hypothesis*, that affirms that all states  $\underline{X}$  with the same energy  $E(\underline{X}) = U$  are visited,<sup>3</sup> one concludes that for an isolated system whose initial energy is in a narrow interval  $E(\underline{X}) \in [U, U + \Delta)$  around  $U$

$$p(\underline{X}) = \begin{cases} \frac{1}{\Delta\Gamma(U)} & \text{if } E(\underline{X}) \in [U, U + \Delta) \\ 0 & \text{otherwise} \end{cases} \quad (19.2)$$

where  $\Delta\Gamma(U)$  is the volume of  $\Gamma$  such that  $E(\underline{X}) \in [U, U + \Delta)$ . Eq. (19.2) is called the *microcanonical ensemble*. There is no proof that the ergodic hypothesis is true in general, and there are several counter-examples. In practice, however, for systems of many degrees of freedom the ergodic hypothesis typically holds because the set of initial conditions for which Eq. (19.1) fails vanishes as  $n \rightarrow \infty$ . These arguments are discussed in detail by Baldwin et al. [56], which is a recommended reading.

For a subsystem of a larger isolated system, instead, one can argue that

$$p(\underline{X}) = \frac{1}{Z(\beta)} e^{-\beta E(\underline{X})}, \quad Z(\beta) = \int_{\Gamma} d\underline{X} e^{-\beta E(\underline{X})}, \quad (19.3)$$

where  $\beta$  is the inverse temperature.<sup>4</sup> Eq. (19.3) is called the *canonical ensemble*.

It combines Hamilton's equations

$$\dot{p}_i = -\frac{\partial E}{\partial q_i} \quad \dot{q}_i = \frac{\partial E}{\partial p_i}$$

with the continuity equation in phase space  $\frac{\partial p}{\partial t} + \nabla_{\underline{X}}(p\dot{\underline{X}}) = 0$ , that states that the change in  $p(\underline{X}, t)$  in any volume  $d\underline{X}$  is due to trajectories  $\underline{X}(t)$  entering or leaving  $d\underline{X}$ . As a result, the Liouville theorem states that

$$\frac{dp}{dt} = \frac{\partial p}{\partial t} + [\nabla_{\underline{X}} p] \dot{\underline{X}} = 0.$$

<sup>3</sup>The ergodic hypothesis has the same flavour of irreducibility in the case of Markov chains. In an irreducible chain, every state can be reached from any other state by a sequence of transitions with positive probability. In Markov chains irreducibility ensures the uniqueness of the invariant distribution, i.e. it ensures that the asymptotic distribution is the same, irrespective of the initial conditions. Likewise, the ergodic hypothesis ensures that all states  $\underline{X}$  with energy  $U$  can be reached from any other state  $\underline{X}'$  with the same energy.

<sup>4</sup>One way to reach this conclusion is the one offered by LANDAU: if  $p(\underline{X}(t))$  is a constant of the motion, then it has to be a function of the constants of the motion. For a system at rest, this means that  $p(\underline{X}) = f(E(\underline{X}))$  must be a function of the energy. The function  $f(\cdot)$  can be identified by requiring that if  $\underline{X}_1$  and  $\underline{X}_2$  are two subsystems of a larger system in equilibrium, and if the interaction between particles are short ranged, then  $\underline{X}_1$  and  $\underline{X}_2$  should be independent, i.e.  $p_{1+2}(\underline{X}_1, \underline{X}_2) = p_1(\underline{X}_1)p_2(\underline{X}_2)$ . At the same time, the energy of the combined system is additive, i.e.  $E_{1+2}(\underline{X}_1, \underline{X}_2) = E_1(\underline{X}_1) + E_2(\underline{X}_2)$ . This implies that  $f(E) = e^{a+bE}$  should have an exponential form, as in Eq. (19.3). Note that the assumption of independence of  $\underline{X}_1$  and  $\underline{X}_2$  implies that  $H[\underline{X}_1, \underline{X}_2]$  is maximal.

Both Eqs. (19.2) and (19.3) have the form of a maximum entropy distribution. Indeed Boltzmann spent considerable effort to show that the distribution  $p(x, t)$  of the coordinates of a single particle in a classical fluid, satisfies an equation — later named after him — that admits  $-\mathcal{H}[p]$  as a Lyapunov function.<sup>5</sup> In other words, the entropy cannot decrease, i.e.

$$\frac{d\mathcal{H}[p]}{dt} \geq 0. \quad (19.4)$$

This is called the *Boltzmann H-theorem*. This result relies on the so-called *molecular chaos hypothesis* that states that when two particles of the fluid collide, we can assume that their velocities are independent random variables.<sup>6</sup> This is not true, strictly speaking, because the same two molecules collide many times with each other, so in principle the velocity of one of the particles depends on the exchanges of momentum it had with previous particles, including the other one. Yet this hypothesis makes a lot of sense, because between two consecutive collisions between the same two particles, both of them collide with so many other particles that the memory of past encounters is “lost in collisions”.

The molecular chaos hypothesis has nothing to do with physics. It is a purely statistical hypothesis, that however has remarkable consequences. Indeed, the laws of motion of classical mechanics are invariant for time reversal whereas Eq. (19.4) states that the entropy is not. The  $H$  theorem is also in contradiction with Poincaré recurrence theorem that states that an Hamiltonian system that starts in a given state will return to it, or arbitrarily close to it, after a sufficiently long time. Apparently also time reversal invariance is “lost in collisions”. How is this possible?

Loosely speaking, this is because “sufficiently long” means an astronomically long time. The time a system spends in equilibrium states is astronomically longer than that spent in non-equilibrium states, because the latter are astronomically more numerous than the latter. This is a statement of the same nature as the Asymptotic Equipartition Property. Hence, a system prepared in a non-equilibrium state will soon relax to equilibrium, but it is practically impossible to observe an equilibrium state that will evolve into a non-equilibrium one. Furthermore, the molecular chaos hypothesis assumes that particles loose memory of their previous encounters. Understandably,

<sup>5</sup>A Lyapunov function is a function that decreases on all the trajectories of the dynamics.

<sup>6</sup>Without this assumption, the equation for the distribution of the coordinates of a single particle would depend on the joint distribution of the coordinates of two particles. The latter, in turn, satisfies an equation that involves the joint distribution of even more particles. This is the so-called BBGKY hierarchy of equations (see KARDAR). The molecular chaos hypothesis closes this hierarchy, by assuming independence.

a system that loses memory will converge to a state of maximal ignorance, i.e. a state of maximum entropy. Note, furthermore, that the molecular chaos hypothesis itself entails maximum entropy at the molecular scale, by assuming independence of the momenta of two colliding particles.

In this derivation, the appearance of the entropy as the functional whose maximisation describes equilibrium states looks like a coincidence. In hindsight it makes a lot of sense:<sup>7</sup> the equilibrium state of macroscopic systems can be described even if all microscopic details are ignored completely, which is exactly what the maximum entropy principle implies. The fact that equilibrium states of a system are described by a state of maximal ignorance means that they can't be distinguished. Non-equilibrium states can be distinguished because there are many ways of driving a system out of equilibrium. The information on how a system is driven out of equilibrium, is precisely the information which is lost when the system relaxes back to equilibrium — a state of maximal ignorance.

Ultimately, the macroscopic behaviour arises from the interplay of two key quantities: the probability  $p(\underline{X})$  of configurations, or its logarithm which is proportional to the *energy*, and the number  $W(E)$  of configurations with the same probability (or with the same energy), which is the *entropy*  $S(E) = \log W(E)$ . The tradeoff between energy and entropy has its roots in typical behaviour, and it is of the same nature of the one that relates the probability of typical sequences to their number in the Asymptotic Equipartition Property.

Before continuing, it is worth to remark that the same system can be described at three different levels:

**Configurations.** Classical mechanics describes the system at the level of configurations  $\underline{X}$ , which is the vector of coordinates and momenta for all particles.

**States.** The same system can be described in terms of the single particle probability distribution  $p(x)$ . For example, the Boltzmann equation (see KARDAR) is based on the distribution  $p(x)$  of the coordinates of single particles.

**Thermodynamic variables.** The macroscopic description of the equilibrium of the system is described by thermodynamic variables. Some of

---

<sup>7</sup>From what we have learned so far, the entropy measures exactly how much information is “lost in collisions”, i.e. how much the uncertainty on a system whose microscopic state is initially described by a state  $p_0(\underline{X})$  increases in time. It is worth remembering, at this point, that the exact knowledge of the state of a system is theoretically impossible, because an infinite number of bits would be needed to specify exactly the position of even a single particle.

these are *extensive*, like the internal energy, the entropy and the volume, in the sense that they are proportional to the number  $n$  of particles. Some are *intensive*, such as the temperature the pressure or the chemical potential. Some of these have a mechanical origin, like the energy, in the sense that they are functions of the coordinates  $\underline{X}$  of the system. Others have a purely statistical origin, such as the entropy, in the sense that they depend on the distribution  $p(\underline{X})$ .

We have seen these different levels of description when we described the properties of sequences  $\underline{X}$  of many i.i.d. random variables. As a consequence of the Asymptotic Equipartition Property, we have seen that all typical sequences correspond to the same type  $P_{\underline{X}}$  (which corresponds to a state) and that averages  $\frac{1}{n} \sum_i f(X_i)$ , such as e.g. the energy, become deterministic (i.e. non-random) quantities, much like thermodynamic variables. The limit  $n \rightarrow \infty$  corresponds to the thermodynamic limit, when the number of particles in the system and its volume both diverge. A macroscopic physical system with a number of particles which is of the order of Avogadro's number  $n \simeq 6 \cdot 10^{23}$ , is very close to this limit.

The coordinates of the particles are not independent random variables, because of the presence of interactions. Yet these interactions are local, which means that each particle interacts with only a finite number of other particles. Particles which are sufficiently far apart are in practice independent. So the vector  $\underline{X}$  can be considered as a sequence of weakly interacting particles, for which the description of the Asymptotic Equipartition Property applies.<sup>8</sup>

## 19.1 Statistical mechanics as maximum entropy inference

The attempt to derive the macroscopic behaviour — i.e. thermodynamics — from the laws of classical mechanics relies on the ergodic and the molecular chaos hypotheses, and it arrives at the maximum entropy principle. Yet neither of these hypotheses is strictly necessary. The Hamiltonian, which is the energy of the system as a function of its coordinates, contains all information on its dynamics. The energy itself is a constant of the motion. Therefore, the

---

<sup>8</sup>It is important to note that interaction between particles (e.g. collisions in a gas) are essential for the system to reach an homogeneous equilibrium. If particles did not interact, each would follow its trajectory with its own conserved quantities (e.g. momentum and kinetic energy). We can consider the coordinates of particles as random variables precisely because in between two observations at different point in time, the same particle has undergone so many interactions with other particles, that its state is totally unpredictable.

maximum entropy principle can be invoked at the outset to predict the state of the system: the best description of a system is the one that does not make unnecessary assumptions. This leads to Eq. (19.2) for an isolated system at energy  $E$  and to Eq. (19.3) for a system in contact with a larger system with which it can exchange energy.<sup>9</sup> This approach to statistical mechanics was proposed by Jaynes [32], to which we refer for more details.

As we have seen, the microcanonical and the canonical distributions are direct consequences of the maximum entropy principle. The second law of thermodynamics is implicit with it, whereas the zeroth (Eq. (18.13)) and the first (Eq. (18.15)) law of thermodynamics, as we have seen, are also a direct consequence of it. Let us review them here. The equilibrium entropy of the system with internal energy  $U = \mathbb{E}[E(\underline{X})]$  is defined as

$$S(U) = \max_{p: \mathbb{E}[E]=U} \mathcal{H}[p]. \quad (19.5)$$

The inverse temperature  $\beta = \frac{1}{k_B T}$  is defined as<sup>10</sup>

$$\beta = \frac{dS}{dU}$$

The zeroth law says that the temperature of systems in thermal equilibrium must be the same. For a system composed of two parts, the entropy satisfies

$$S(U) = \max_{U_1} [S_1(U_1) + S_2(U - U_1)] \quad (19.6)$$

where  $S_1$  and  $S_2$  are the entropies of subsystems 1 and 2, each of which is the solution of the maximisation of the entropy on the respective subsystem, with  $\mathbb{E}[E_i(\underline{X}_i)] = U_i$  ( $i = 1, 2$ ). The first order condition of the maximisation in Eq. (19.6) yields

$$\left. \frac{dS_1}{dU_1} - \frac{dS_2}{dU_2} \right|_{U_2=U-U_1} = \beta_1 - \beta_2 = 0$$

which implies that the temperatures in the two subsystems must be the same, in equilibrium. If the system is slightly out of equilibrium and  $\beta_1 \neq \beta_2$ , then we expect the equilibrium will be restored by means of an exchange in energy between the two parts of the system. If the energy of system 1 increases by

<sup>9</sup>Indeed, the energy  $E(\underline{X})$  is a sufficient statistics of Eqs. (19.2), (19.3).

<sup>10</sup>With respect to the notation used in the derivation of Eq. (18.15), here  $f = U$  is the internal energy,  $\beta = -\lambda$  is the inverse temperature and  $\beta F(\beta) = \phi(\lambda)$  defines the free energy  $F(\beta)$ .



$dU_1$  then that of system 2 decreases by the same amount, because the energy of the combined system is conserved. Hence the increase in  $S$  is given by

$$dS = (\beta_1 - \beta_2)dU_1 \geq 0.$$

This states that energy (i.e. heat) will pass from hotter to colder bodies, and not vice-versa.<sup>11</sup> This is Clausius statement of the *second law of thermodynamics*.

In a thermodynamic transformation where the energy levels  $E(\underline{X})$  change by  $\delta E(\underline{X})$  and the internal energy  $U$  changes by  $dU$ , the change in entropy  $S(U)$  is given by the same argument leading to Eq. (18.15), i.e.  $dS = \beta (dU - \mathbb{E}[\delta E(\underline{X})])$ . This can be rewritten as

$$dU = \delta Q + \delta W. \quad (19.7)$$

which is the *first law of thermodynamics*. In Eq. (19.7)  $\delta Q = dS/\beta$  is the heat supplied to the system, whereas  $\delta W = \mathbb{E}[\delta E(\underline{X})]$  is the work done on the system. Note that  $dU$  is an exact differential (i.e. it is the differential of a state variable), whereas  $\delta Q$  and  $\delta W$  are not.

For an isolated system of  $n$  particles in a finite volume  $V$ , the entropy  $S(U, V, n)$  is a function of  $U$ ,  $V$  and  $n$ . The thermodynamics description of systems in thermal equilibrium at temperature  $1/\beta$  is obtained from the Legendre transform, and it is given by the *free energy*

$$F(\beta, V, n) = \frac{1}{\beta} \min_p [\beta \mathbb{E}[E] - \mathcal{H}[p]] = -\frac{1}{\beta} \log Z(\beta, V, n). \quad (19.8)$$

Hence, in practice, the equilibrium of a system at (inverse) temperature  $\beta$  is derived from the partition function  $Z$  in Eq. (19.3) using Eq. (19.8) to compute the free energy.

The same recipe of Legendre transform can be applied to obtain a description where the volume or the number of particles are allowed to change. The first step is to identify the conjugate variable and the second to find the corresponding thermodynamic potential. For example, for systems that can freely expand in their environment (i.e.  $V$  is not fixed), at temperature  $1/\beta$ , the conjugate variable that replaces  $V$  is the pressure

$$P = -\frac{\partial F}{\partial V}$$

---

<sup>11</sup>If  $\beta_1 > \beta_2$ , i.e. if 1 is colder than 2, then  $dU_1 \geq 0$ , which means that heat will flow from 2 to 1.

and the thermodynamic potential is  $G(\beta, P, n) = F + PV$ . Likewise, the conjugate variable to  $n$  is the chemical potential  $\mu = \frac{\partial F}{\partial n}$  and the potential<sup>12</sup> is  $\Omega(\beta, V, \mu) = F - \mu n$ . We refer to LANDAU or KARDAR for a detailed discussion.

In the rest of this Chapter, we shall focus on applying the recipe of statistical mechanics in few interesting cases, where we shall put to use what we have learned.

## 19.2 The classical ideal gas

In order to illustrate the concepts discussed so far, let us consider a gas of  $n$  non-interacting particles of mass  $m$  in a box of volume  $V$  at temperature  $1/\beta$ . The Hamiltonian, in terms of the canonical coordinates  $(\mathbf{q}, \mathbf{p})$ , is given by

$$E(\mathbf{q}, \mathbf{p}) = \sum_{i=1}^n \frac{p_i^2}{2m}$$

where  $\mathbf{q} = (q_1, \dots, q_n)$  is the vector of positions of the particles and  $\mathbf{p}$  the vector of momenta. In  $D$  dimensions these are  $nD$  dimensional vectors. The canonical partition function is obtained integrating over all coordinates, i.e.

$$Z(\beta, V, n) = \int d\mathbf{q} d\mathbf{p} e^{-\beta E(\mathbf{q}, \mathbf{p})} \quad (19.9)$$

$$= V^n \left[ \int_{-\infty}^{\infty} d\mathbf{p} e^{-\frac{\beta \mathbf{p}^2}{2m}} \right]^{nD} \quad (19.10)$$

$$= V^n \left( \frac{2\pi m}{\beta} \right)^{\frac{nD}{2}}. \quad (19.11)$$

Hence the free energy is given by

$$F = -\frac{1}{\beta} \log Z = -\frac{n}{\beta} \log \left[ V \left( \frac{2\pi m}{\beta} \right)^{\frac{D}{2}} \right]. \quad (19.12)$$

All thermodynamics quantities of interest, like the energy

$$U = \frac{\partial}{\partial \beta}(\beta F) = \frac{nD}{2\beta}$$

---

<sup>12</sup>This is called the *grand potential* and it corresponds to a distribution  $p$  that is called the *grand-canonical ensemble*.

or the pressure

$$P = -\frac{\partial F}{\partial V} = \frac{n}{\beta V}$$

are obtained from  $F$ .

A disturbing fact of this result is that the free energy of the ideal gas is not extensive. Indeed  $F/n$  diverges in the thermodynamic limit  $n \rightarrow \infty$  with  $V/n$  finite (see Eq. (19.12)). The root of the problem lies in the entropy

$$S = \beta(U - F) = n \log V + \frac{nD}{2} \log \left( \frac{2\pi em}{\beta} \right) \quad (19.13)$$

which is non-extensive, because of the first term. This fact, known as the *Gibbs paradox* is not a mistake. The entropy is exactly what it should be, i.e.

$$S = -n \int d\vec{q} d\vec{p} \rho(\vec{q}, \vec{p}) \log \rho(\vec{q}, \vec{p}) \quad (19.14)$$

where  $\rho(\vec{q}, \vec{p})$  is the single particle probability density function

$$\rho(\vec{q}, \vec{p}) = \frac{1}{V} \left( \frac{\beta}{2\pi m} \right)^{D/2} e^{-\beta \frac{p^2}{2m}}.$$

The origin of the “paradox” lies in the distribution of positions, which is uniform in the volume occupied by the gas. The fact that there is no paradox, is illustrated by comparing the state of a gas of  $n$  particles in a volume  $V$  at temperature  $1/\beta$  with that of the same system which is divided in two equal parts of volume  $V/2$ , each of which contains  $n/2$  particles. The free energy of the split system is smaller than that of the original system by an amount  $n \log 2$ . This is precisely the number of bits needed to specify in which part each of the molecules is confined when the gas is divided in two parts. The calculation leading to Eq. (19.12) correctly accounts for this loss of information. Ultimately, the “paradox” arises because the particles are distinguishable and they carry their own identity as they travel around the system. Note that there is no Gibbs paradox if instead of a gas one considers a solid, where each atom is localised in its specific location. It is only when particles are allowed to exchange their rôles by physics that problems with extensivity arise.

There are two ways to recover an extensive free energy. The first is to consider indistinguishable particles. For example, you can check that in a Bose gas the free energy is extensive. The second is to remember the discussion we had on generating functions for labelled objects Eq. (7.19). There we learned that when counting distinguishable (i.e. labelled) objects we need to modify our mathematical counting device, in order to preserve the composition

property. After all, the partition function is nothing but a generating function, so that discussion suggests that we should use a modified partition function, dividing it by  $n!$ . In this way the free energy

$$F = -\frac{1}{\beta} \log \frac{Z}{n!} = -\frac{n}{\beta} \log \left[ \frac{V}{n} \left( \frac{2\pi m}{\beta} \right)^{\frac{D}{2}} \right] - \frac{n}{\beta}$$

regains its extensive character. Yet, this mathematical artifice spoils the interpretation of the entropy in information theoretic terms.

### Exercise 19.1

Note that  $\vec{p}$  and  $\vec{q}$  are continuous variables and Eq. (19.14) is a differential entropy. Does introducing a finite precision in our measurement of the positions and momenta, e.g. as suggested by Heisenberg uncertainty principle, fixes Gibbs paradox?

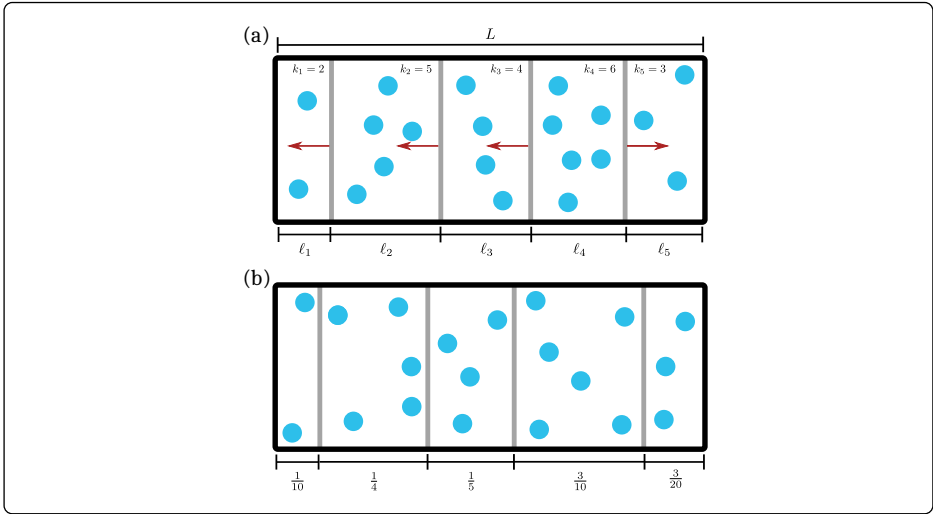
### Exercise 19.2

In a Bose gas, particles occupy single particle states with energy  $\hbar k^2 / (2m)$  with each component of  $\vec{k}$  taking discrete values  $2\pi\ell / L$ , and  $V = L^3$ . Compute the free energy of the ideal Bose gas and check that it is extensive.

Summarising, non-extensivity derives from the assumption of distinguishability of classical mechanics. Quantum mechanics shows that this assumption is wrong: particles are indistinguishable. Yet one could envisage a system of classical distinguishable particles and one could ask whether the non-extensive part of the free energy could have some physical effect, and if so, can it be used? In other words, can the information on which particle is which be used to do work?

## 19.2.1 The Slilárd information engine\*

Far from providing an answer, let us see how (some) information can be turned into work in a simple case. Let us consider the same system, but in a box that is partitioned into  $q$  different compartments, by  $q - 1$  vertical walls. Let  $\ell_a$  be the longitudinal length of compartment  $a = 1, \dots, q$ . We assume the box to have unit area in the perpendicular direction, so  $\ell_a$  is also the volume of partition  $a$ . We denote by  $X_i$  the partition to which particle  $i$  belongs. The



**Figure 50.** An ideal gas in a box divided by walls. If walls can move horizontally the gas in different partitions will expand or contract so as to reach equilibrium.

probability to find particle  $i$  in box  $a$  is

$$p_a = \frac{\ell_a}{L}, \quad L = \sum_{a=1}^q \ell_a$$

and since all particles are independent, the probability of a configuration  $\underline{X} = (X_1, \dots, X_n)$  is given by

$$P(\underline{X}) = \prod_{i=1}^n p_{X_i} = \prod_{a=1}^q p_a^{n_a} \quad (19.15)$$

where  $n_a = |\{i : X_i = a\}|$  is the number of particles in partition  $a$  and  $\sum_{a=1}^q n_a = n$ . Each of the sub-systems is an ideal gas of  $n_a$  particles in thermal equilibrium at temperature  $1/\beta$ . We can derive the Hamiltonian of the system in terms of the coordinates  $\underline{X}$ , by equating Eq. (19.15) with the Boltzmann equation  $P(\underline{X}) = \frac{1}{Z} e^{-\beta E(\underline{X})}$ . This gives

$$E(\underline{X}) = -\frac{1}{\beta} \sum_{a=1}^q n_a \log p_a = \sum_{a=1}^q \epsilon_a n_a, \quad \epsilon_a = -\frac{1}{\beta} \log \frac{\ell_a}{L}, \quad (19.16)$$

with  $Z(\beta) = 1$ . This is equivalent to a system of  $n$  particles distributed on  $q$  energy levels  $\epsilon_a$ . Let us now imagine that each wall is allowed to move

freely in the longitudinal direction. The pressure in each partition is given by  $P_a = \frac{n_a}{\beta \ell_a}$ , which is also the force acting on the walls. The wall between partition  $a$  and  $a + 1$  will experience a force equal to  $P_a - P_{a+1}$  and it will move accordingly, until the new positions  $\hat{\ell}_a$  are such that the pressure in each compartment is the same

$$\hat{P}_a = \frac{n_a}{\beta \hat{\ell}_a} = P$$

where  $P = \frac{n}{\beta L}$ . This condition implies that the new positions  $\hat{\ell}_a$  should be such that

$$\hat{P}_a = \frac{\hat{\ell}_a}{L} = \frac{n_a}{n}$$

is exactly equal to the probabilities  $p_a$  that make the configuration  $\underline{X}$  as likely as possible (these will be called the *maximum likelihood* parameters in the next chapter).

### Exercise 19.3

The temperature of Bernoulli trials: let  $\underline{X} = (X_1, \dots, X_n)$  be a sequence of Bernoulli trials, where  $X_i = 0, 1$ . We wish to interpret this as a system of  $n$  independent particles in a two state system. Each particle has energy  $\epsilon(X)$  depending on which state  $X = 0, 1$  it is in, so the energy is given by

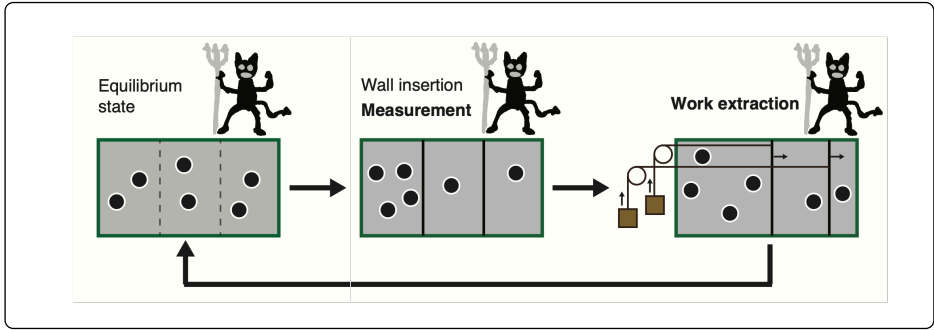
$$E(\underline{X}) = \sum_{i=1}^n \epsilon(X_i).$$

Take  $\epsilon(0) = 1$  and  $\epsilon(1)$  such that  $\mathbb{E}[E(\underline{X})] = 0$ . With this choice, all effects of the randomness should be ascribed to the temperature, which is the only free parameter. By equating  $p(\underline{X})$  to  $\frac{1}{Z} e^{-\beta E(\underline{X})}$  show that the (inverse) temperature is given by  $\beta = p \log \frac{p}{1-p}$ . Comment the result.

The work done in this transformation is

$$W = \sum_{a=1}^q \int_{\ell_a}^{\hat{\ell}_a} d\ell P_a(\ell) = \sum_{a=1}^q \int_{\ell_a}^{\hat{\ell}_a} d\ell \frac{n_a}{\beta \ell} = \frac{n}{\beta} D_{KL}[\hat{p} \| p], \quad (19.17)$$

which is related to the Kullback-Leibler divergence between the final and the initial state. So  $\beta W$  is related to the change in the information content of



**Figure 51.** The Szilard engine.

the system. The key point is that this work can be used *if the number  $n_a$  of particles in each partition is known*.<sup>13</sup>

#### Exercise 19.4

Show that  $\beta W = nI(\mathbf{n}, q)$  can also be written as the mutual information between the  $x$ -position  $q$  of one of the particles and the vector  $\mathbf{n} = (n_1, \dots, n_q)$ .

Eq. (19.17) equates the work that can be done by a quasi-static transformation to a Kullback-Leibler divergence. This relation is true in general. Indeed the work done in a quasi-static transformation at constant temperature equals the difference in the free energy, because by the first law of thermodynamics,  $\Delta W = \Delta E[E] - \Delta Q = \Delta E[E] - T\Delta S = \Delta F$ . This relation is true if the initial and final states, as well as all intermediate states, are equilibrium states. It also holds whatever is the distribution  $p(\underline{X})$  of the initial state, provided that the entropy of the final state is given by Shannon's formula  $S = k_B H[\underline{X}]$ , with  $\Delta F = k_B T D_{KL}[p \| p_{\text{eq}}]$ , as shown by Esposito and Van den Broeck [40], where  $p_{\text{eq}}$  is the equilibrium distribution.

Finally note that, because of Eq. (19.16), the work is precisely given by its

<sup>13</sup>If the number  $n_a$  of particles in each partition is known, it is possible to anticipate in which direction the walls will move and to exploit the movements of the walls to perform work (e.g. by lifting a weight attached to the walls). Szilárd used this to devise a cyclic transformation of an ideal system in which the knowledge of  $\mathbf{n}$  makes it possible to extract work from a system at finite temperature  $1/\beta$ . Szilárd proposed this as a simple manifestation of Maxwell's demon idea, i.e. that knowledge of the microscopic state of a system makes it possible to (apparently) violate the second law of thermodynamics. Note that the amount of work  $W = nI(\mathbf{n}, q)/\beta$  can be measured in bits. For more details, see [39].

loss in internal energy, i.e.

$$W = E(\underline{X}) - E'(\underline{X})$$

due to the change  $\delta\epsilon_a$  in the energy levels.

### 19.3 The Ising model

The Ising model is the workhorse of statistical mechanics. It describes a magnetic system where each atom is characterised by a magnetic moment — a *spin* for short — that can either point up or down. Hence  $\underline{X} = (X_1, \dots, X_n)$  is a vector of variables  $X_i = \pm 1$  that take only two values. The Hamiltonian is defined as

$$E(\underline{X}) = -h \sum_{i=1}^n X_i - J \sum_{\langle i,j \rangle} X_i X_j$$

The first term describes the influence of an external magnetic field  $h$  that promotes an alignment of the spins in the direction of the sign of  $h$  (remember that states of minimal energy are more likely). In the second, the sum runs on all pairs  $\langle i, j \rangle$  of interacting spins and promotes states where spins are aligned (for  $J > 0$ ). In real physical systems, a spin  $i$  interacts only with spins of atoms that are nearby in space. Here we consider the *mean field* version of the model, where the sum on  $\langle i, j \rangle$  is replaced by a sum over all pairs, but with an intensity reduced by a factor  $n$ , i.e.

$$E(\underline{X}) = -h \sum_{i=1}^n X_i - \frac{J}{n} \sum_{i < j} X_i X_j. \quad (19.18)$$

The factor  $1/n$  ensures that the energy of the *ground state*<sup>14</sup>

$$\min_{\underline{X}} E(\underline{X}) = -|h|n - \frac{J}{2}(n-1) \propto n$$

is extensive.

In order to compute the partition function, we use the fact that

$$\sum_{i < j} X_i X_j = \frac{1}{2} \left( \sum_i X_i \right)^2 - \frac{1}{2} \sum_i X_i^2 = \frac{1}{2} \left( \sum_i X_i \right)^2 - \frac{n}{2}.$$

<sup>14</sup>This is obtained by aligning all spins with  $h$ , i.e.  $X_i = \text{sign } h$  for all  $i$ .



Therefore, neglecting the last term  $n/2$ , that only contributes a constant, and with  $M = \sum_i X_i$

$$Z(\beta) = \sum_{\underline{X}} e^{-\beta E(\underline{X})} \quad (19.19)$$

$$= \sum_{M=-n}^n \binom{n}{\frac{n+M}{2}} e^{\beta h M + \frac{\beta J}{2n} M^2} \quad (19.20)$$

$$\simeq \frac{n}{2} \int_{-1}^1 dm e^{-n\beta g(m, \beta, h)} \quad (19.21)$$

where we changed variables from  $M$  to  $m = M/n$ . The function  $g$  is given by

$$g(m, \beta, h) = -\frac{1}{\beta} \mathcal{H}[m] - hm - \frac{J}{2} m^2, \quad (19.22)$$

where

$$\begin{aligned} \mathcal{H}[m] &= \frac{1}{n} \log \left( \binom{n}{\frac{1+m}{2}n} \right) \\ &\simeq -\frac{1+m}{2} \log \frac{1+m}{2} - \frac{1-m}{2} \log \frac{1-m}{2} \end{aligned} \quad (19.23)$$

is the entropy of a random variable  $X = \pm 1$ , with  $P\{X = +1\} = \frac{1+m}{2}$ , and the last expression is a trite application of Stirling's formula. Eq. (19.21) can be evaluated by the saddle point method, so the free energy per particle is given by<sup>15</sup>

$$f(\beta, h) = -\frac{1}{\beta} \lim_{n \rightarrow \infty} \frac{1}{n} \log Z(\beta) = g(m^*, \beta, h) \quad (19.24)$$

where  $m^*(\beta, h)$  is the solution of the equation  $\frac{\partial g}{\partial m} = 0$ . The equation for  $m^*$  can be put in the form<sup>16</sup>

$$m^* = \tanh(\beta h + \beta J m^*). \quad (19.25)$$

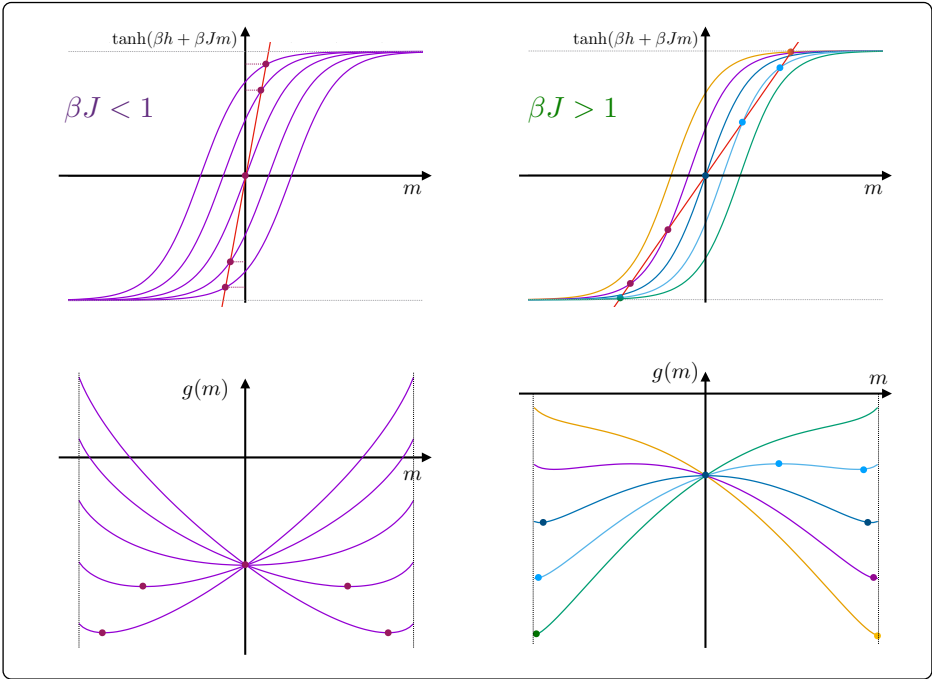
<sup>15</sup>Note that  $m^*$  is the conjugate variable to  $h$ , because

$$\frac{df}{dh} = \frac{\partial g}{\partial h} + \frac{\partial g}{\partial m} \frac{dm^*}{dh} = -m^*$$

because  $\frac{\partial g}{\partial m} = 0$  when  $m = m^*$ .

<sup>16</sup>Here we use the relation

$$\operatorname{arc} \tanh m = \frac{1}{2} \log \frac{1+m}{1-m}$$



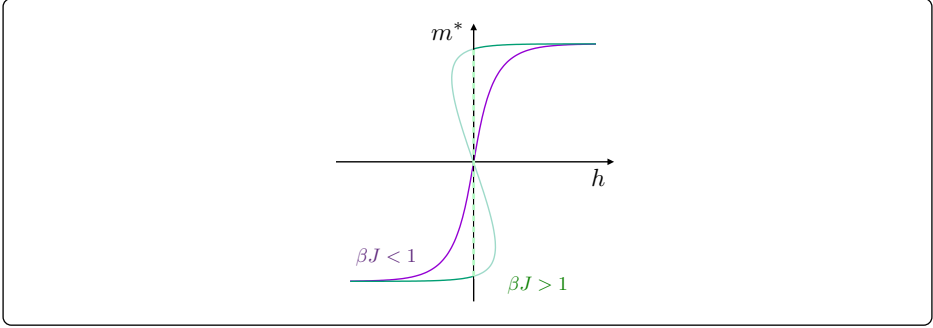
**Figure 52.** Graphical solution of the saddle point equations for the magnetisation of the mean field Ising model. The top graphs show the solutions of Eq. (19.25) and the bottom graphs the corresponding values of  $g$  for high temperatures ( $\beta J < 1$ , left) and low temperatures ( $\beta J > 1$ , right).

The behaviour of  $m^*$  as a function of  $h$  can be analysed by plotting the left and the right hand side of Eq. (19.25), as in Figure 52.

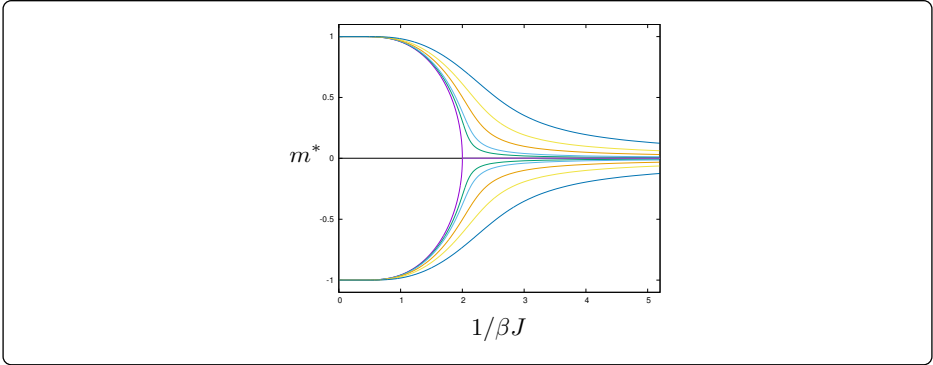
For  $\beta J < 1$  the solution to Eq. (19.25) is unique for all values of  $h$ , and it continuously increases from  $-1$  to  $1$ . For  $\beta J > 1$ , instead, there is an interval of  $h$  around the origin where Eq. (19.25) has three solutions. The plot (see Figure 53) of all three solutions, as a function of  $h$ , exhibits an s-shaped behaviour around the origin. Inspection of the function  $g$  reveals that one of them is a maximum of  $g$ , that corresponds to an unstable state. The other two are minima. Of these, the one that has the same sign of  $h$  attains a lower minimum. This corresponds to the equilibrium state. The other minimum is called a meta-stable state.

Therefore the correct solution “jumps” between the negative solution and the positive one, as soon as  $h$  crosses zero. The resulting value of  $m^*$  as a function of  $h$  is shown in Figure 53.

As a function of temperature, as shown in Figure 54, we observe that for  $h \neq 0$  the magnetisation varies between  $m^* = \pm 1$  (depending on whether



**Figure 53.** The solutions of Eq. (19.25) as a function of  $h$  for  $\beta J = \frac{1}{2}$  and  $\beta J = 2$ . In the latter case, also the unstable and metastable solutions are shown. The dashed vertical line connects the negative and the positive branch of the equilibrium solution, that corresponds to the global minimum of  $g$ .



**Figure 54.** The solutions of Eq. (19.25) as a function of  $1/\beta J$  for different values of  $h$ .

$h > 0$  or  $h < 0$ ) for  $\beta \rightarrow \infty$ , and  $m^* = 0$  for  $\beta = 0$ . For  $h = 0$  instead, we observe a singular behavior. For  $\beta J < 1$  the magnetisation vanishes ( $m^* = 0$ ), whereas for  $\beta J > 1$  the magnetisation splits in two branches of opposite sign. At  $h = 0$  the distribution of  $M$  is symmetric for changes  $M \rightarrow -M$ . This symmetry (which is explicitly broken when  $h \neq 0$ ) is spontaneously broken. For  $\beta J > 0$ , in spite of the fact that  $\mathbb{E}[M] = 0$ , in each realisation of the Ising model we find that the magnetisation takes value  $M = \pm nm^*$ .

In the region  $\beta J > 0$ , the distribution  $p(\underline{X})$  is “divided” into two *pure states* that correspond to the phases with opposite magnetisations. The distribution of the magnetisation per spin  $m = M/n$  in the two states is approximately a Gaussian

$$p_{\pm}(m) \simeq \sqrt{\frac{n}{2\pi\mathbb{V}[m]}} e^{-n \frac{(m \mp m^*)^2}{2\mathbb{V}[m]}}$$

where  $\mathbb{V}[m]$  is the inverse of the second partial derivative of  $g$  with respect to  $m$ , computed at  $m^*$ . Any state with a magnetisation  $\bar{m} \in [-m^*, m^*]$  (on the vertical dashed line of Figure 53) can be obtained as a mixture

$$\bar{p}(m) = \frac{1 + \bar{m}/m^*}{2} p_+(m) + \frac{1 - \bar{m}/m^*}{2} p_-(m).$$

This corresponds to a situation where the two phases *coexist*. In physical systems with short range interactions this is realised by having a fraction  $\frac{1+\bar{m}/m^*}{2}$  in the phase with magnetisation  $m^*$  and the rest in the opposite phase.<sup>17</sup>

The singularity at the phase transition  $\beta J = 1$  becomes evident if one studies the behaviour of the *susceptibility*, which quantifies the change in the observable  $m^*$  if the parameter  $h$  is varied

$$\chi = \frac{dm^*}{dh} \tag{19.26}$$

$$= (1 - m^{*2}) \left[ \beta + \beta J \frac{dm^*}{dh} \right] \tag{19.27}$$

$$= \beta \frac{1 - m^{*2}}{1 - \beta J (1 - m^{*2})}. \tag{19.28}$$

For  $h = 0$  and  $\beta J < 1$ , the magnetisation vanishes. Hence the susceptibility

$$\chi = \frac{\beta}{1 - \beta J}, \quad h = 0, \quad \beta J < 1$$

diverges as  $\beta J \rightarrow 1^-$ .

<sup>17</sup>The thermodynamic phase a physical system with short range correlations, corresponds to a system of weakly dependent variables which can be described in terms of a single particle distribution function, as if each particle's coordinate were drawn independently from a distribution  $P$ . The situations where more than one phase coexists, is then analogous to the case of a sample  $\underline{X}$  of i.i.d. draws from either  $P$  or  $Q$ , which we discussed earlier in the context of large deviations. As we saw, the large deviation function  $I(\bar{x})$  is non-convex in that case. In thermodynamics, non-convex thermodynamic potentials are un-physical. For example, the van der Waals theory of liquids, predicts a non-convex potential. This contrasts with thermodynamic stability because it results in a non-monotonic relation between pressure and volume. The *Maxwell construction* remedies to this problem, by drawing an horizontal line that cuts the non-monotonic part of the  $P - V$  curve in such a way that the areas above and below the line in the  $P - V$  plot are equal. This condition identifies a quasi-static cycle that can be performed exerting no work, which means that the two states at the extremes of the cut have the same free energy. Mathematically this is equivalent to the construction leading to  $\bar{I}(\bar{x})$ . Physically, the states on the horizontal line in the Maxwell construction, are mixture of the two phases, where a fraction of the system is in one phase and the rest is in the other. In physics, these are the thermodynamically stable states and they can be realised because it is possible to grow bubbles of one phase into the other. The energetic cost of the mixed states is of the order of the interface between the two phases, which is negligible with respect to the bulk energy, which is proportional to the volume.

**Exercise 19.5**

Using Eq. (19.25), prove that the same behaviour  $\chi \sim 1/|\beta J - 1|$  also attains as  $\beta J \rightarrow 1^+$ .

The singular behaviour of thermodynamic quantities at a second order phase transition point is traditionally characterised in terms of the *critical exponents* which describe the singular behaviour of thermodynamic quantities close to a phase transition. For example, the divergence of the susceptibility as the temperature approaches the critical point is usually described as  $\chi \sim |\beta - \beta_c|^{-\gamma}$ , where  $\beta_c$  is the critical value of the (inverse) temperature and  $\gamma$  is an exponent. Hence we found that  $\gamma = 1$  for the mean field Ising model (with  $\beta_c = 1/J$ ).

**Exercise 19.6**

The exponent  $\tilde{\beta}$  is defined by the behaviour of  $m^* \sim (\beta J - 1)^{\tilde{\beta}}$  for  $\beta J \rightarrow 1^+$  at  $h = 0$  and the exponent  $\delta$  by the behaviour  $m^* \sim h^{1/\delta}$  when  $h \rightarrow 0$  with  $\beta J = 1$ . Using the expansion of Eq. (19.25), find  $\tilde{\beta} = 1/2$  and  $\delta = 3$ .

Note that, the large deviation function  $I(\bar{x})$  of the magnetisation  $\bar{x} = \frac{1}{n} \sum_i X_i$  can be computed directly for the mean field Ising model, and it reads

$$I(\bar{x}) = \beta f(\beta, h) - \mathcal{H}(\bar{x}) - \beta h \bar{x} - \frac{\beta J}{2} \bar{x}^2 \quad (19.29)$$

where  $\mathcal{H}(\bar{x})$  is the function in Eq. (19.23). You can check that for  $\beta J > 1$  this function is not convex and therefore its Legendre transform has a singularity when the conjugate parameter equals  $-h$ .

## 19.4 The Random Energy Model

The Ising model describes a magnetic system where all atoms interact in the same way. There are other systems where, because of impurities of different types (called generically *disorder*), the interaction can be of either sign and they can involve more than two spins. As a way to model these situations, you can consider an Ising model where each of the interactions  $J_{i,j}$  is drawn at random from some distribution.<sup>18</sup> As a result of this, the energy  $E(\underline{X})$  itself, for a fixed configuration  $\underline{X}$ , becomes a random variable. The Random

<sup>18</sup>These models are called *spin glasses*.

energy Model (REM) was proposed by Derrida in 1981 [25] as an extreme realisation of these systems, where each energy  $E(\underline{X})$  is drawn from a Gaussian distribution

$$p(E) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{E^2}{2n}},$$

independently, for each configuration<sup>19</sup>  $\underline{X}$ . With  $\underline{X} = (X_1, \dots, X_n)$  and  $X_i = \pm 1$ , the partition function is a sum of  $N = 2^n$  random variables. One may expect that, because of the law of large numbers,

$$Z(\beta) = \sum_{\underline{X}} e^{-\beta E(\underline{X})} \simeq 2^n \mathbb{E} [e^{-\beta E}] = 2^n e^{\frac{n}{2}\beta^2}. \quad (19.30)$$

This immediately yields the free energy

$$F(\beta) = -\frac{1}{\beta} \log Z(\beta) \simeq -\frac{n}{2} \log 2 - \frac{n}{2} \beta^2, \quad (19.31)$$

the internal energy

$$U(\beta) = -\frac{\partial}{\partial \beta} \log Z(\beta) \simeq -n\beta, \quad (19.32)$$

and the entropy

$$S(\beta) = \beta(U - F) = n \log 2 - \frac{n}{2} \beta^2. \quad (19.33)$$

The problem with this solution is that, for  $\beta > \beta_c = \sqrt{2 \log 2}$  the entropy becomes negative. This is not possible, because the entropy of a discrete variable  $\underline{X}$  must be non-negative. In order to understand what is going wrong, let us compute the minimal energy (which is called the *ground state energy*)  $E_0 = \min_{\underline{X}} E(\underline{X})$ . This is the minimum of  $N = 2^n$  i.i.d. random variables, and  $N$  is really very large. Let us write  $E(\underline{X}) = -\sqrt{n} Y(\underline{X})$ , so that  $Y(\underline{X})$  are  $N$  i.i.d. Gaussian random variables with mean zero and variance one. Then

$$E_0 = \min_{\underline{X}} E(\underline{X}) = -\sqrt{n} \max_{\underline{X}} Y(\underline{X}) \simeq -\sqrt{n} \sqrt{2 \log N}$$

where we have used the expression of the coefficient  $a_N \simeq \sqrt{2 \log N}$  that we have computed for the maxima of Gaussian random variables. This shows that the minimal value that the energy can take is  $E_0 = -n\sqrt{2 \log 2}$ . There

<sup>19</sup>The reason why the variance of  $E$  is taken to be proportional to  $n$  is because this ensures that the thermodynamic quantities are extensive, i.e. proportional to  $n$ , as we shall see. The variance of  $E$  is also the specific heat at infinite temperature, which has to be extensive.

are no states with energy lower than that. This means that Eq. (19.32) has to be modified as

$$U(\beta) = -n \min \left( \beta, \sqrt{2 \log 2} \right).$$

Notice that the change in  $U$  occurs precisely at the value  $\beta_c$  where the entropy vanishes. For  $\beta \geq \beta_c$  the Gibbs distribution  $p(\underline{X}) = \frac{1}{Z(\beta)} e^{-\beta E(\underline{X})}$  is concentrated on very few states with energy close to  $E_0$ . Hence  $S \simeq 0$  for  $\beta \geq \beta_c$ , i.e.

$$S(\beta) = \max \left[ n \log 2 - \frac{n}{2} \beta^2, 0 \right].$$

The problem with the calculation leading to Eq. (19.31) is that the law of large numbers only holds when the number of terms that contribute to the sum is large. This occurs only for  $\beta < \beta_c$ . For  $\beta > \beta_c$  the partition function is *not* self-averaging. The assumption Eq. (19.30) goes under the name of the *annealing approximation*. It makes the calculation easy, as in this case, but often wrong, specially at low temperatures. In general, the self-averaging quantities are the extensive ones, like the free energy. This means that rather than taking the expected value of  $Z(\beta)$  one has to take the so-called *quenched average*

$$F(\beta) = -\frac{1}{\beta} \mathbb{E} [\log Z(\beta)]$$

which is much harder computationally.<sup>20</sup>

### 19.4.1 A gas of weakly interacting particles and the Grand Canonical ensemble

The phenomenon of concentration of large deviations is a simple realisation of a second order phase transition. This is made more apparent by translating the problem into that of the statistical mechanics of an interacting gas problem. This problem was first discussed in a paper of Bialas et al. [41]. Later the same

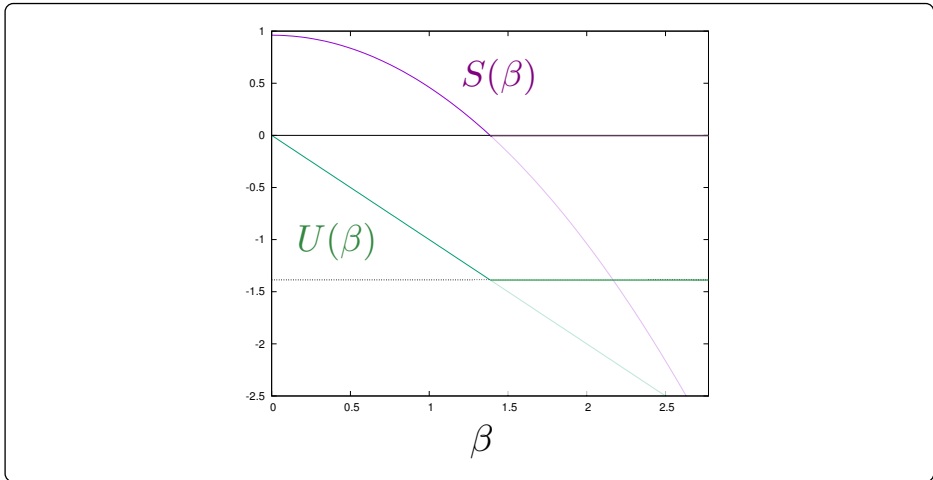
<sup>20</sup>The trick that is often used to deal with  $\mathbb{E} [\log Z(\beta)]$  is to write

$$\log Z = \lim_{r \rightarrow 0} \frac{Z^r - 1}{r}$$

which has the advantage that one needs to take the expected values of powers of  $Z$ . For integer  $r$ ,

$$Z^r(\beta) = \sum_{\underline{X}_1} e^{-\beta E(\underline{X}_1)} \dots \sum_{\underline{X}_r} e^{-\beta E(\underline{X}_r)}$$

is the partition function of  $r$  replicas of the same system. This is why the approach that uses this trick to compute  $\mathbb{E} [\log Z]$  is called the *replica method*. This method is based on computing  $\mathbb{E} [Z^r]$  for integer  $r$ , then to interpret the result for real values of  $r$  (by analytic continuation) and finally to take the limit as  $r \rightarrow 0$ .



**Figure 55.** The entropy and the internal energy of the REM model.

phenomenon was discussed in a much broader set of particle models, known as *zero-range processes* (see e.g. Evans et al. [42]).

Consider  $N$  particles distributed in  $n$  boxes (or states) and let the Hamiltonian be given by

$$E(\underline{X}) = \sum_{i=1}^n \log(1 + X_i)$$

where  $X_i$  is the number of particles in box  $i = 1, \dots, n$ , and  $\sum_i X_i = N$ . This is a gas of particles with weak attractive on-site interaction.<sup>21</sup> We shall consider the equilibrium distribution of the particles at temperature  $1/\beta$ . You can consider the boxes arranged on a  $d$ -dimensional lattice, with particles jumping from box to box, with any dynamics that obeys detailed balance with the Hamiltonian above.<sup>22</sup> The probability of a configuration  $\underline{X}$  is given by

$$P\{\underline{X}\} = \frac{1}{Z(\beta, N)} e^{-\beta E(\underline{X})} \delta_{\sum_i X_i, N} = \frac{1}{Z(\beta, N)} \prod_{i=1}^n (1 + X_i)^{-\beta} \delta_{\sum_i X_i, N},$$

<sup>21</sup>In order to see why the interaction is an attractive one, take two sites with  $X_1$  and  $X_2$  particles. The configurations where all particles are moved to the same site has always a lower energy, because  $\log(1 + X_1) + \log(1 + X_2) \geq \log(1 + X_1 + X_2)$ .

<sup>22</sup>Detailed balance for a system at fixed temperature means that the probability  $w(\underline{X} \rightarrow \underline{X}')$  of a transition from state  $\underline{X}$  to state  $\underline{X}'$  satisfies

$$\frac{w(\underline{X} \rightarrow \underline{X}')}{w(\underline{X}' \rightarrow \underline{X})} = \frac{P(\underline{X}')}{P(\underline{X})} = e^{-\beta[E(\underline{X}') - E(\underline{X})]}.$$

(see the chapter on Markov chains).



where the delta function imposes the constraint of particle number conservation, and the canonical partition function  $Z(\beta, N)$  is obtained as usual summing the Boltzmann factor  $e^{-\beta E(X)}$  over all states with  $N$  particles.<sup>23</sup>

A simpler way to study the system is to use the Grand Canonical ensemble instead of the Canonical one. The Grand Canonical ensemble describes a system where the number of particles is allowed to fluctuate, as if the system were in contact with a larger system at the same temperature, with which it can exchange particles. The variable  $N$  is replaced by its conjugate variable, which is the chemical potential  $\mu$ . This entails introducing a statistical weight  $e^{-\beta\mu}$  for each particle. The Grand Canonical partition function is given by

$$\mathcal{Z}(\beta, \mu) = \sum_{N=0}^{\infty} e^{-\beta\mu N} Z(\beta, N) = \left[ \sum_{x=0}^{\infty} (1+x)^{-\beta} e^{-\beta\mu} \right]^n.$$

The thermodynamic potential in this ensemble is called the *grand potential*

$$\Omega(\beta, \mu) = -\frac{1}{\beta} \log \mathcal{Z}(\beta, \mu).$$

The expected number of particles in the system is obtained as

$$\mathbb{E}[N] = \frac{\partial}{\partial \mu} \Omega(\beta, \mu)$$

From which one defines the density  $\rho = \mathbb{E}[N]/n$ . In the thermodynamic limit ( $n \rightarrow \infty$ ) the grand canonical ensemble's description is equivalent to the description of the canonical ensemble because the variance of the density<sup>24</sup>  $\rho = N/n$  is proportional to  $1/n$ . Hence, the density  $\rho$  converges to a constant value  $\rho = \mathbb{E}[N]/n$ , which is a function of  $\mu$ . Hence, adjusting the chemical potential  $\mu$  it is possible to change the density  $\rho$  of particles in the sub-system. This implies that the gas is in a state where the number of particles on each site has a distribution

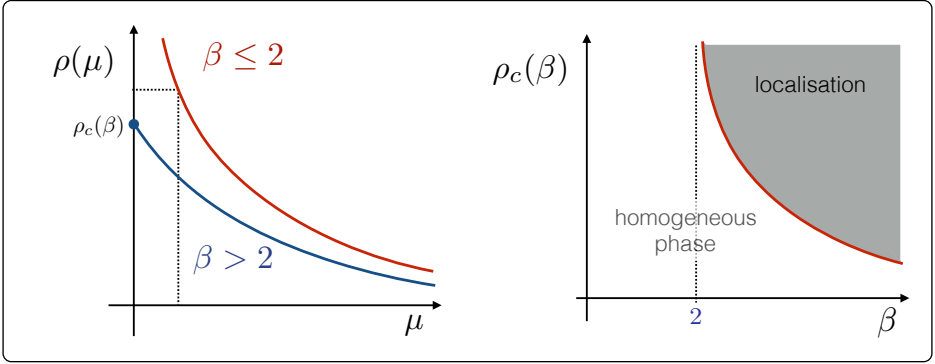
$$P\{X_i = x\} = A(1+x)^{-\beta} e^{-\beta\mu x}, \quad x = 0, 1, \dots \quad (19.34)$$

where  $A(\beta, \mu)$  is a normalisation constant.

The emphasis is different but the machinery and the concepts are exactly the same as the ones used for discussing large deviations or states of maximal

<sup>23</sup>Notice that this corresponds exactly to studying large deviations from a distribution  $Q(x) = Q_0(1+x)^{-\beta}$  where  $\mathbb{E}[X] = \rho = N/n$ . The Cramer function  $I(\rho) = S(\rho_0) - S(\rho)$  is related to the decrease in entropy with respect to the typical case where  $\rho_0 = \mathbb{E}_Q[X]$ , and  $S(\rho_0) = \mathcal{H}[Q]$  is the entropy of the distribution  $Q$ .

<sup>24</sup>which is obtained from the second derivative of  $\Omega$  with respect to  $\mu$ .



**Figure 56.** A weakly interacting gas. Left: density as a function of chemical potential. Right: the phase diagram.

entropy. The grand canonical trick is biasing *a priori* probabilities (with  $\mu = 0$ ) on the distribution of particles in each box in such a way as to recover states with a given density as large deviations, i.e. as typical outcomes under the biased distribution.

However, the trick only works as long as  $\mathcal{Z}(\beta, \mu)$  is well defined. In our case this corresponds to  $\mu \geq 0$ , because  $\mathcal{Z}$  is undefined for  $\mu < 0$ . In our case, the density, as a function of  $\mu$  and  $\beta$ , is given by

$$\rho(\beta, \mu) = A \sum_{x=0}^{\infty} x(1+x)^{-\beta} e^{-\beta\mu x}. \quad (19.35)$$

When  $\beta \leq 2$  the density diverges in the limit  $\mu \rightarrow 0^+$ . Therefore, for every value of  $N/n$  it is possible to find a value of  $\mu$  such that  $\rho(\beta, \mu) = N/n$ .

For  $\beta > 2$ , instead, the limit

$$\lim_{\mu \rightarrow 0} \rho(\mu) = \rho_c(\beta) < +\infty$$

is finite. All states with a density of particles smaller than  $\rho_c$  can be described by finding the value of  $\mu > 0$  such that  $\rho(\mu)$  equals  $N/n$ . In order to achieve states with a density  $\rho > \rho(0)$  the symmetry between the different boxes has to be broken. The most likely state for a gas with  $N/n > \rho_c$  is given by a situation where all sites  $i \neq i^*$  but one have  $\rho_c$  particles on average, with  $X_i$  distributed independently according to Eq. (19.34) with  $\mu = 0$ , and the remaining one ( $i^*$ ) gathers all the excess  $N - (n-1)\rho_c$  particles. These are the states of maximal entropy, hence these are those that are expected to be typically observed. In summary, as we increase the density of particles, if  $\beta > 2$ , the system crosses

the critical point  $\rho_c(\beta)$  beyond which a finite fraction of the particles “localises” on a single site.

### Exercise 19.7

Plot the phase diagram in the  $(\rho, T)$  plane, solving numerically the equation for  $\rho(\mu)$ . Perform a numerical simulations with a Metropolis algorithms where moves are just hopping of single particles from one box to another. Verify numerically the phase transition.

One may wonder what is the relation between this phenomenon and Bose-Einstein condensation (BEC). Without entering into many details, BEC is a phase transition where a finite fraction of the particles of a quantum ideal gas of bosons condensates in the state with zero momentum, where the single particle wave function is completely delocalised. Also in that case, the analysis can be carried out in the grand canonical ensemble, with the introduction of the chemical potential  $\mu > 0$  that should be fixed so that the density equals  $\rho = \frac{n}{V}$ . In  $d > 2$ , however, there is a maximal density of particles that can be accommodated in states with non-zero momentum, which is

$$\rho_c(d) = A \int_0^\infty dz \frac{z^{d/2-1}}{e^z - 1} \quad (19.36)$$

where  $A$  is a constant. When the density  $\rho > \rho_c$  exceeds this threshold, a finite fraction of the particles have to “condensate” in the state of zero momentum. We refer to other sources [38] for the derivation of this result. Here we observe that the critical density in our case can be written as<sup>25</sup>

$$\rho_c(\beta) = A \sum_{x=0}^{\infty} (1+x)^{-\beta} x \quad (19.37)$$

$$= A \sum_{x=0}^{\infty} (1+x)^{-\beta+1} - 1 \quad (19.38)$$

$$= \frac{A}{\Gamma(\beta-1)} \int_0^\infty dz \frac{z^{\beta-2}}{e^z - 1} - 1. \quad (19.39)$$

<sup>25</sup>Here we used the identity

$$(1+x)^{-a} = \frac{1}{\Gamma(a)} \int_0^\infty dz z^{a-1} e^{-(1+x)z}$$

in the last step, in order to sum the series on  $x$ .

The condition for the existence of the BEC (i.e.  $\rho_c(d) < +\infty$ ) is then very similar to the one that guarantees the existence of a localised phase in the gas we're studying (i.e.  $\rho_c(\beta) < +\infty$ ). In both cases, the power of  $z$  in the integral should be larger than zero, to prevent the divergence of the integral when  $z \rightarrow 0$ . Therefore, at least at a formal level, the localisation transition discussed here is similar to the BEC phase transition.

## 19.5 A teaser in stochastic thermodynamics\*

The coordinates  $X$  of a physical system define its state.  $X$  provides also all the information that enter the laws of motion of the system, which define how  $X$  evolves over time. In other words, the dynamics of a physical system is such that conditional to the present state  $X_t$  the future  $(X_{t+1}, X_{t+2}, \dots)$  is independent of the past  $(\dots, X_{t-2}, X_{t-1})$ . In technical terms,  $X_t$  is a Markov process. The dynamics satisfies general laws of thermodynamics, which we will now derive in the simplest case where  $X_t$  evolves as a Markov chain.

Consider a Markov chain  $\underline{X} = (X_0, X_1, \dots, X_N)$  defined on a finite state space  $X_t \in \mathcal{S}$  with transition matrix  $\hat{p}^{(t)}$  with elements

$$p_{s,s'}^{(t)} = P\{X_t = s | X_{t-1} = s'\} \quad t = 1, \dots, N.$$

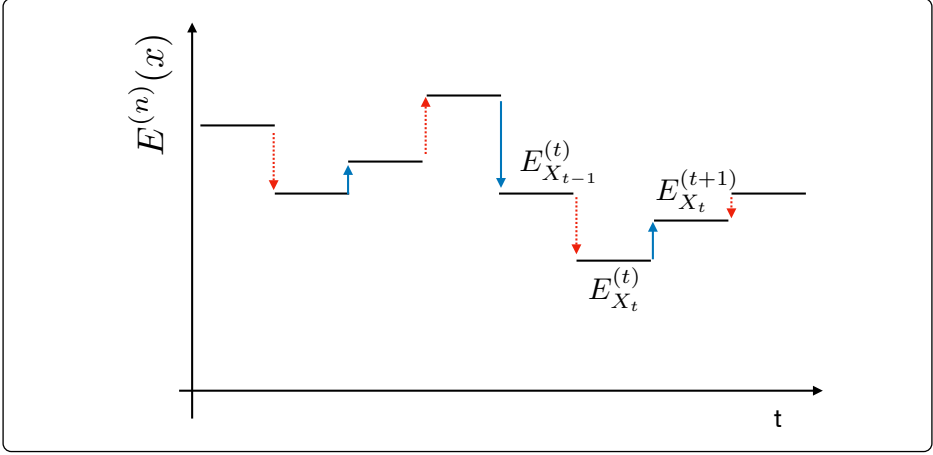
We explicitly allow for a time dependence in the transition matrix because we want to describe general situations in which the system under study can be manipulated. We assume that all states  $s \in \mathcal{S}$  are ergodic for each of the transition matrices  $\hat{p}^{(t)}$ . Hence each transition matrix  $\hat{p}^{(t)}$  admits an invariant measure  $\mu_s^{(t)} = \sum_{s'} p_{s,s'}^{(t)} \mu_{s'}^{(t)}$ . For simplicity, we consider a situation where the system is unperturbed until  $t = 0$ , so that the transition probability is  $\hat{p}^{(1)}$  for all  $t \leq 0$ . Hence we can assume that the distribution of  $X_0$  is  $P\{X_0 = s\} = \mu_s^{(1)}$ .

Following Hack et al. [43], we define the energy function at time  $t$  as

$$E_s^{(t)} = -\log \mu_s^{(t)},$$

and  $E_s^{(0)} = E_s^{(1)}$ . The reason for this choice is that it is consistent with equilibrium statistical mechanics. Indeed, a physical system with this energy function will converge over time to an equilibrium state  $\mu_s^{(t)} = \frac{1}{Z} e^{-\beta E_s^{(t)}}$  (with  $Z = \beta = 1$ ) which coincides with the invariant measure associated with the Markov chain with transition matrix  $\hat{p}^{(t)}$ .

Over time, the energy will change either because the state changes or because the energy function itself changes. Therefore one can define work  $W$



**Figure 57.** The energy changes either because work is done on the system (blue full arrows) or because the state  $X_t$  changes and heat is released (dotted red arrows).

and heat  $Q$  for the transformation  $\underline{X}$  as

$$W(\underline{X}) = \sum_{n=0}^{N-1} [E_{X_n}^{(n+1)} - E_{X_n}^{(n)}] \quad (19.40)$$

$$Q(\underline{X}) = \sum_{n=1}^N [E_{X_n}^{(n)} - E_{X_{n-1}}^{(n)}] . \quad (19.41)$$

Work  $W$  is defined as the change of energy levels due to some applied force, when the state  $X_t$  is fixed. Note that  $W$  is work done on the system, because  $W > 0$  when the energy increases. Heat  $Q$  is the change in energy due to the variation of the state  $X_t$ , on a fixed energy landscape. Their sum

$$W(\underline{X}) + Q(\underline{X}) = E_{X_N}^{(N)} - E_{X_0}^{(0)} = \Delta E(\underline{X}) \quad (19.42)$$

is the variation of the energy during the transformation. This is the first law of thermodynamics. When  $\Delta E = 0$  work done on the system is transformed into heat.

In order to derive the second law of thermodynamics, let us recall that the reverse Markov chain is defined as

$$q_{s,s'}^{(t)} \equiv P\{X_{t-1} = s | X_t = s', t\} = \frac{P_{s',s}^{(t)} \mu_s^{(t)}}{\mu_{s'}^{(t)}} .$$

Now note that the probability of this trajectory in the reverse process is

$$Q_{\leftarrow}(X_N, \dots, X_0) \equiv q_{X_0, X_1}^{(1)} q_{X_1, X_2}^{(2)} \cdots q_{X_{N-1}, X_N}^{(N)} \mu_{X_N}^{(N)} \quad (19.43)$$

$$\begin{aligned} &= p_{X_1, X_0}^{(1)} \frac{\mu_{X_0}^{(1)}}{\mu_{X_1}^{(1)}} p_{X_2, X_1}^{(2)} \frac{\mu_{X_1}^{(2)}}{\mu_{X_2}^{(2)}} \cdots p_{X_N, X_{N-1}}^{(N)} \frac{\mu_{X_{N-1}}^{(N)}}{\mu_{X_N}^{(N)}} \mu_{X_N}^{(N)} \\ &= p_{X_N, X_{N-1}}^{(N)} \cdots p_{X_2, X_1}^{(2)} p_{X_1, X_0}^{(1)} \mu_{X_0}^{(1)} \prod_{n=1}^{N-1} \frac{\mu_{X_n}^{(n)}}{\mu_{X_n}^{(n+1)}} \\ &= P_{\rightarrow}(X_0, \dots, X_N) e^{-W(X)} \end{aligned} \quad (19.44)$$

which coincides with Eq. (16.27), i.e. when  $\hat{p}^{(t)} = \hat{p}^{(1)}$  for all  $t$  then  $W = 0$  and  $Q_{\leftarrow}(X_N, \dots, X_0) = P_{\rightarrow}(X_0, \dots, X_N)$ .

Summing over all values of  $X_0, \dots, X_N$  one obtains Jarzinski equality  $\mathbb{E}[e^{-W}] = 1$ . A consequence of Eq. (19.44) is that

$$\mathbb{E}[W] = \mathbb{E}\left[\log \frac{P_{\rightarrow}(X_0, \dots, X_N)}{Q_{\leftarrow}(X_N, \dots, X_0)}\right] = D_{KL}[P_{\rightarrow} \| Q_{\leftarrow}] \geq 0 \quad (19.45)$$

which is a generalised second law of thermodynamics, that states that no work can be extracted from a thermodynamic transformation between states with the same free energy.<sup>26</sup>

It is worth to contrast Eq. (19.45) with Eq. (16.33), which defined the entropy production  $\Sigma \equiv D_{KL}[P_{\rightarrow} \| P_{\leftarrow}]$ . While Eq. (19.45) compares the direct and the inverse process in a time varying set-up, Eq. (16.33) compares the two arrows of time in the same process with time independent transition probabilities, thereby providing a measure of irreversibility.

<sup>26</sup>Our choice of  $Z = \beta = 1$  implies that the free energy  $F = -\beta^{-1} \log Z = 0$  at all times.

## Chapter 20

# Statistical inference

Essentially, all models are wrong, some are useful. (G.P.E. Box)<sup>1</sup>

G.P.E. Box was a statistician not a physicist. A physicist would probably think that Newton's law is the right model for phenomena described by classical mechanics (e.g. bodies falling from a height). Yet, even there, every experiment is subject to effects that cannot be controlled by the experimenter. Evidently these effects are not in the simple model

$$m\ddot{h} = -g$$

that describes the trajectory of the falling body.<sup>2</sup> No matter how much care we take, when we take measures many times, we'll always get slightly different numbers. That's why we take experimental averages.

In most cases, all the statistics a physicist needs does not go beyond mean and variance, because experiments can be repeated many times and the control of experimental conditions can be improved.<sup>3</sup> As one moves away from physics, one faces phenomena that are not ruled by known fundamental laws, with experiments that cannot be done or repeated. All one has is a series of

---

<sup>1</sup>This sentence is attributed to George P.E. Box, quoting his paper in Journal of the American Statistical Association [44]. The paper does not contain such a sentence, yet it clearly discusses it at length, together with an interesting discussion of the scientific method and some notes on the remarkable life of R.A. Fisher, one of the founding fathers of statistics. The paper is a recommended reading.

<sup>2</sup>Here  $h(t)$  would be the height of the body,  $m$  its mass and  $g$  the acceleration due to the gravitational force.

<sup>3</sup>In particle physics, there is a convention that, in order for an experimental result to be considered a discovery, it should be statistically validated to a confidence of 99.99994%, which corresponds to an interval of five standard deviations around the average of a Gaussian distribution.

observations and the question is what can we learn from these. Learning is not only describing, it is also *generalising*, i.e. predicting data which has not been seen yet. The trade-off between accuracy, i.e. how well you describe the data you have already seen, and generalisation, i.e. how well you predict the outcome of experiments not yet done, is a key issue in statistics.

Here we shall discuss a general class of problems that is usually encountered in classical statistics and inference. We're given a sample  $\underline{X} = \{X_1, \dots, X_n\}$ . We suppose  $X_i$  are i.i.d. draws from an (unknown) distribution<sup>4</sup>  $Q(x)$ . Our goal is that of inferring  $Q$  or to say something about it. The first framework is that of *hypothesis testing*, where we have to choose between two alternative possibilities for  $Q$ . This is a decision problem. How can we make this decision in the best possible way? How can we do it in order that the number of samples needed to make the correct decision with a given accuracy is minimal? This does not apply only to statistics. Any alarm system needs to sense the environment and test whether the particular conditions under which a specific response is needed occur. Taking this decision optimally and as quickly as possible (i.e. with the least number of samples) may determine life or death.<sup>5</sup>

Next we move to parameter estimation. In this case, we know (or we assume) that the distribution that has generated the data belongs to a parametric family of distributions, but we do not know the value of the parameter(s). Maximising the likelihood is the simplest recipe but, strictly speaking, it is not the right answer. Indeed it produces an estimate of the parameter that depends on the data, and that varies as the data accumulates. Indeed, the right way to think about the problem is that we should encode our state of knowledge on the parameters into a distribution. Before we see the data, this is the *prior* distribution and after we see the data we can update our state of knowledge using Bayes rule. In this way the likelihood can be used to compute the *posterior* distribution. But how do we choose a prior distributions that correspond to a given state of ignorance on the parameters, and which prior corresponds to a complete state of ignorance on the parameters of a given model?

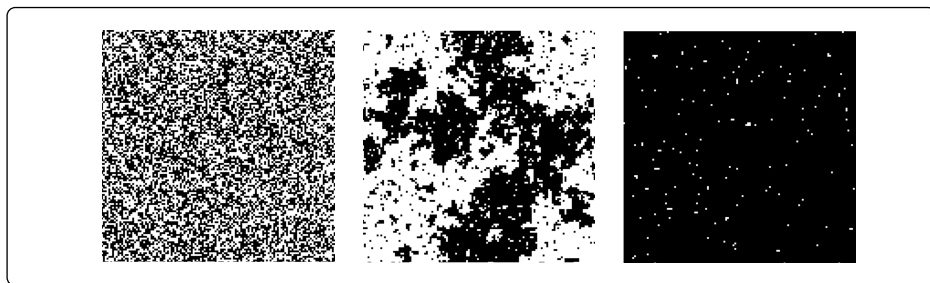
At any rate, we expect that as the data accumulates, prior knowledge becomes less and less relevant. Indeed, as we shall see, maximum likeli-

---

<sup>4</sup>In other words, the problem is to find a distribution  $Q$  such that  $\underline{X}$  can be considered as a typical sample drawn from  $Q$ . Remember that there are  $\sim e^{nH[P_{\underline{X}}]}$  typical samples, where  $P_{\underline{X}}$  is the type of the sample, and for each of these, the probability to be drawn from  $Q$  is  $e^{\sum_x P_{\underline{X}}(x) \log Q(x)}$ , so the probability that  $\underline{X}$  is a typical sample is  $\sim e^{-nD_{KL}[P_{\underline{X}}||Q]}$ .

<sup>5</sup>For example, an organism has to decide whether to switch on or not a genetic program, or to switch metabolism from one state to another. This depends on the estimate that the organism computes of the concentrations of different nutrients and toxins in the environment.





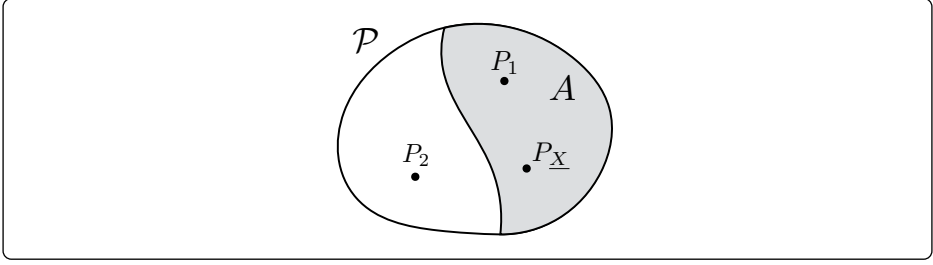
**Figure 58.** Three snapshots of a 2-D Ising model at different temperatures. What is your uncertainty on the temperature in the three cases?

hood estimates converge to asymptotic values, or equivalently, the posterior distribution becomes more and more sharply peaked around the maximum likelihood estimates. The speed of convergence, i.e. the width of the posterior distribution, is controlled by the Fisher Information. The (inverse of the) Fisher Information quantifies the uncertainty in parameter estimation or how much the estimates of parameters change when new data is added. When the Fisher Information is large, the estimated parameters do not change much if new data is added. This means that the model estimated on past data also describes well yet unseen data. This means that it generalises well.

Statistical inference is the inverse problem to statistical mechanics. While the first deals with understanding which model best describes a dataset, the latter studies which behaviour — i.e. which type of data — is generated by a given model. The Fisher Information has its counterpart in the susceptibility in statistical mechanics, because it quantifies how observables change when the parameters change. This gives a special significance to those points where the susceptibility becomes very large (or diverge in the thermodynamic limit). These are associated to critical behaviour at continuous phase transitions in statistical mechanics. In statistical inference, we expect that these same “critical” models to be good at generalising.

Finally we discuss model selection, which is the situation where we have to choose between several parametric models, with different level of complexity and detail. Hypothesis testing is not the right approach here, as noted by Akaike [45], because all hypotheses are wrong to start with. Several recipes have been suggested (such as AIC, BIC, MDS, etc) and it is important to see where they come from and how they are related.

This is an outline of the results discussed in chapter 11 of COVER plus other subjects. We discuss the main ideas and refer to textbooks for the derivations. In this chapter we consider a sample  $\underline{X} = \{X_1, \dots, X_n\}$ , where



**Figure 59.** A sketch of the space of probability distributions and hypothesis testing.

$X_i$  can be considered as the outcomes of  $n$  independent experiments, run under the same conditions. Therefore we think of  $X_i$  as i.i.d. draws from an (unknown) distribution  $Q(x)$ .

This chapter concentrates on the regime of classical statistics, where the dimensionality of the data and of the parameters of the models are finite and the number of samples diverge ( $n \rightarrow \infty$ ). In this regime we can rely on the asymptotic results that we have discussed thus far. We shall briefly comment at the end about high dimensional statistical inference, a regime in which the dimensionality of the data or of model's parameter is of the same order or larger than the number of data points.

## 20.1 Hypothesis testing

A simple example of hypothesis testing is the case where we have two alternative hypotheses on the unknown distribution  $Q$ :

$$H_1 : Q = P_1 \quad H_2 : Q = P_2,$$

and we want to decide which one is most appropriate for the data  $\underline{X}$ . We restrict attention to the case where  $X_i \in \chi$  takes values in a finite set  $\chi$  (with  $n \gg |\chi|$ ).

The way to design a statistical test is to define an acceptance region  $A$  for  $H_1$ , such that if  $\underline{X} \in A$  then  $H_1$  is accepted and  $H_2$  is rejected and vice-versa.<sup>6</sup> Of course  $P_1 \in A$  and  $P_2 \notin A$ , because when  $n \gg 1$  we expect that  $P_{\underline{X}} \approx P_1$  if  $H_1$  is true. There are many possible ways to define  $A$ . What is the best way to choose  $A$ ?

In order to address this issue, let us introduce the error probabilities

$$\alpha = P_1(\bar{A}) \equiv \sum_{\underline{X} \notin A} P_1(\underline{X}) \quad \beta = P_2(A) \equiv \sum_{\underline{X} \in A} P_2(\underline{X}). \quad (20.1)$$

<sup>6</sup> $A$  can either be defined in the space of samples  $\underline{X}$  or in the space  $\mathcal{P}$  of types  $P_{\underline{X}}$ .

These are the probabilities to reject the hypotheses  $H_1$  or  $H_2$ , if they are correct, i.e. if  $\underline{X}$  were actually drawn from  $P_1$  or  $P_2$ , respectively. The optimal choice for  $A$  is the one that makes  $\alpha$  and  $\beta$  as small as possible.

The answer to this problem is given by the *Neyman-Pearson lemma*, that states that the optimal acceptance region is defined in terms of thresholds on likelihood ratios. Indeed, if

$$A = \left\{ \underline{X} : \frac{P_1(\underline{X})}{P_2(\underline{X})} \geq T \right\} \quad (20.2)$$

where  $T > 0$  is an arbitrary threshold, then there is no other acceptance region  $B$  such that  $\tilde{\alpha} < \alpha$  and  $\tilde{\beta} \leq \beta$ , with

$$\tilde{\alpha} = P_1(\tilde{B}) \quad \tilde{\beta} = P_2(B)$$

and  $\alpha, \beta$  given by Eq. (20.1) and  $A$  given in Eq. (20.2)<sup>7</sup> The significance of this result is particularly transparent if one takes logarithms. Then the acceptance region reads

$$A = \left\{ \underline{X} : D_{KL}[P_{\underline{X}} \| P_1] \leq D_{KL}[P_{\underline{X}} \| P_2] - \frac{1}{n} \log T \right\}.$$

For  $T = 1$ , the hypothesis which has to be accepted is the one closest to the data in terms of relative entropy.

Let us now compute the error probabilities. Take  $\beta$  for example and let's focus on the case  $T = 1$  for simplicity. The event that a sample  $\underline{X}$  generated as i.i.d. draws from  $P_2$  lands in  $A$  is clearly a large deviation. Hence  $\beta$  can be computed using Sanov's theorem Eq. (17.12). This tells us that

$$\beta \sim e^{-nD_{KL}[P^* \| P_2]}$$

where

$$P^* = \arg \min_{P \in A} D_{KL}[P \| P_2].$$

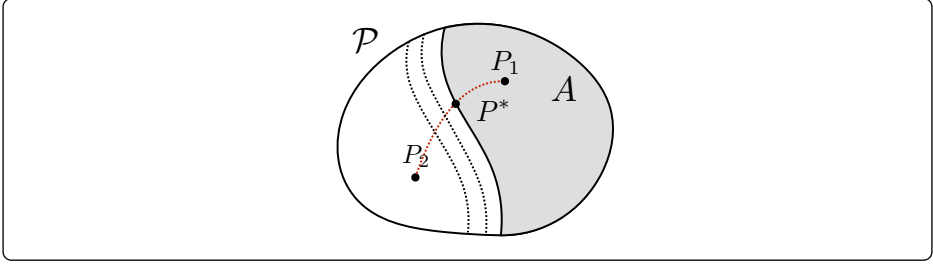
The constrain that sequences  $\underline{X}$  having types  $P_{\underline{X}} = P$  belong to  $A$  can be imposed with a Lagrange multiplier in the optimisation problem

$$\min_{P \in A} D_{KL}[P \| P_2] = \min_{P \in A, \lambda} [D_{KL}[P \| P_2] + \lambda [D_{KL}[P \| P_1] - D_{KL}[P \| P_2]]].$$

<sup>7</sup>The proof of this statement relies on the inequality

$$[\phi_A(\underline{X}) - \phi_B(\underline{X})] [P_1(\underline{X}) - TP_2(\underline{X})] \geq 0$$

where, for any set  $S$ ,  $\phi_S(\underline{X}) = 1$  if  $\underline{X} \in S$  and  $\phi_S(\underline{X}) = 0$  otherwise. This inequality is easily proven considering the different cases, e.g. if  $\underline{X} \in A$  and  $\underline{X} \notin B$ , the first factor is one and the second is positive, because  $P_1(\underline{X}) \geq TP_2(\underline{X})$  for  $\underline{X} \in A$ . Taking the sum of this inequality on all  $\underline{X}$ , one finds  $\alpha + T\beta \leq \tilde{\alpha} + T\tilde{\beta}$ .



**Figure 60.** As  $\lambda$  varies, the distribution  $P_\lambda$  connects the distributions  $P_1 = P_{\lambda=1}$  and the distribution  $P_2 = P_{\lambda=0}$ . The different values of  $\lambda$  correspond to different choices of  $T$  for the acceptance region.

The solution of this constrained minimisation problem is given by

$$P_\lambda(x) = \frac{P_1^\lambda(x)P_2^{1-\lambda}(x)}{\sum_{x'} P_1^\lambda(x')P_2^{1-\lambda}(x')} \quad (20.3)$$

where  $\lambda$  should be fixed so that  $D_{KL}[P_\lambda \| P_1] = D_{KL}[P_\lambda \| P_2]$  (for  $T = 1$ ). The calculation for  $\alpha$  can be done in the same manner, and the solution turns out to be the same as Eq. (20.3), i.e.

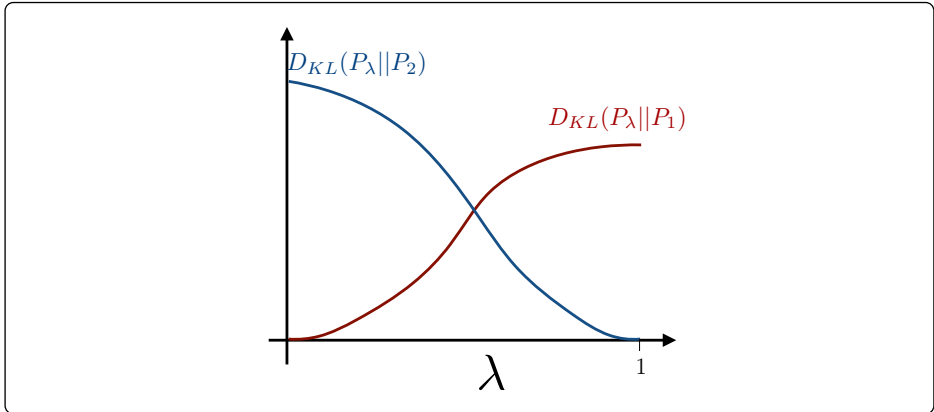
$$\alpha \sim e^{-nD_{KL}[P_\lambda \| P_1]}, \quad \beta \sim e^{-nD_{KL}[P_\lambda \| P_2]}$$

where  $\lambda$  is such that  $D_{KL}[P_\lambda \| P_1] = D_{KL}[P_\lambda \| P_2]$ .

### Exercise 20.1

Show that, for a given  $T$ , the value of  $\lambda$  that optimises  $D_{KL}[P_\lambda \| P_1]$  is the same as the one that determines  $\beta$ .

Graphically, the space  $\mathcal{P}$  of all probability measures on  $\mathcal{X}$ , is divided in two parts: the acceptance region  $A$  for  $H_1$ , with  $P_1 \in A$ , and the acceptance region  $\bar{A}$  for  $H_2$ , with  $P_2 \in \bar{A}$ . In the determination of  $\beta$  we look at the distribution  $P \in A$  that is closest to  $P_2$  whereas  $\alpha$  entails looking for the distribution  $P \in \bar{A}$  that is closest to  $P_1$ . The general solution, of both problems is given by  $P_\lambda$ : as  $\lambda$  varies in  $[0, 1]$ , the point  $P_\lambda$  traces a path in  $\mathcal{P}$  that goes from  $P_2$  (for  $\lambda = 0$ ) to  $P_1$  (for  $\lambda = 1$ ). This path intersects the boundary of  $A$  in a single point, that corresponds to the solution of both problems. Notice that as  $T$  varies from 0 to  $\infty$ ,  $A$  shrinks from the entire space  $\mathcal{P}$  excluding a small neighbourhood of  $P_2$ , for  $T = 0^+$ , to a small neighbourhood of  $P_1$ , for  $T \gg 1$ . Correspondingly the parameter  $\lambda$  changes from 0, for  $T = 0^+$ , to 1 for  $T \gg 1$ .



**Figure 61.** A sketch of the space of probability distributions and parameter estimation.

The next question to be addressed is how to choose  $T$ . One recipe is that of fixing  $\alpha^* = \epsilon$  to a preassigned value independent of  $n$ . Then the smallest error  $\beta^*$  that can be achieved is given by

$$\beta^* \sim e^{-nD_{KL}[P_1\|P_2]}$$

for large  $n$ . This result is known as *Stein's Lemma*.

A different way to choose  $\lambda$  comes from a Bayesian approach. We assume that hypotheses  $H_1$  and  $H_2$  have prior probabilities  $\pi_1$  and  $\pi_2$ . Then it is natural to require that the posterior error

$$P_e = \pi_1\alpha + \pi_2\beta$$

should be as small as possible. By Sanov's theorem

$$P_e \simeq \pi_1 e^{-nD_{KL}[P_\lambda\|P_1]} + \pi_2 e^{-nD_{KL}[P_\lambda\|P_2]} \sim e^{-n \min\{D_{KL}[P_\lambda\|P_1], D_{KL}[P_\lambda\|P_2]\}}$$

for  $n$  large.

### Exercise 20.2

This analysis is reminiscent of the one we discussed in the context of the Gartner-Ellis theorem. Reformulate the problem and the results in terms of large deviations.

Therefore, the optimal value of  $\lambda$  is the one for which

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log P_e = \min\{D_{KL}[P_\lambda\|P_1], D_{KL}[P_\lambda\|P_2]\}$$

is the largest possible, which is indeed the point  $\lambda^*$  where  $D_{KL}[P_{\lambda^*} \| P_1] = D_{KL}[P_{\lambda^*} \| P_2]$ . You just need to sketch a plot of the two distances as  $\lambda$  varies in  $[0, 1]$ , to convince yourself of this. The value

$$C(P_1, P_2) = \max_{\lambda \in [0,1]} \min\{D_{KL}[P_\lambda \| P_1], D_{KL}[P_\lambda \| P_2]\}$$

is called the Chernoff bound and it provides the minimal *a posteriori* error in Bayesian hypothesis testing. Notice that prior distributions, for  $n$  large, do not play any role. Put differently, the Chernoff bound yields the minimal number of samples  $(-\log P_e)/C(P_1, P_2)$  that are needed to reach a decision with a given level of confidence  $P_e$ .

### Exercise 20.3

Let  $P_1(X = 1) = P_1(X = 0) = 1/2$  and  $P_2(X = 1) = p = 1 - P_1(X = 0)$ . Compute the minimal number of points needed to distinguish between  $P_1$  and  $P_2$  on the basis of a sample  $\underline{X}$  of independent observations of  $X_i = 0, 1$ , with a precision of 1%. Plot the result as a function of  $p \in [0, 1]$ .

### Exercise 20.4

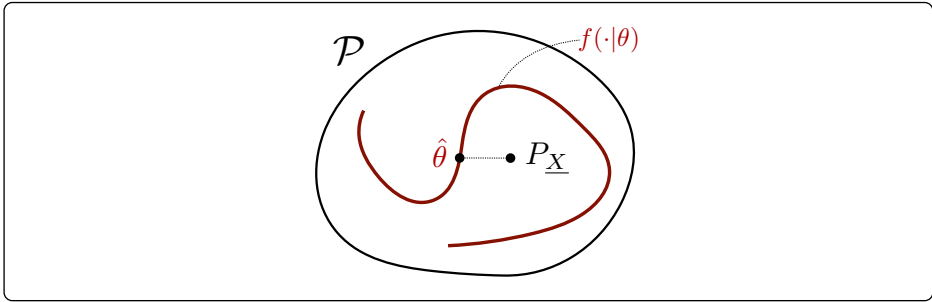
Show that

$$C(P_1, P_2) = - \min_{\lambda \in [0,1]} \log \left[ \sum_x P_1^\lambda(x) P_2^{1-\lambda}(x) \right].$$

## 20.2 Parameter estimation and the Fisher Information

Let us now consider the case where the distribution  $Q$  has the parametric form  $f(x|\theta)$ . In other words, we either know or assume that  $\underline{X}$  can be considered as i.i.d. draws from  $f(x|\theta_0)$  for an unknown parameter  $\theta_0$ . The question then becomes that of estimating the value  $\theta_0$  on the basis of a sample  $\underline{X} = (X_1, \dots, X_n)$  of  $n$  data points. Here  $\theta$  can be a vector of parameters, though most of the arguments can be discussed for a single parameter.

The simplest idea is that of finding the parameter  $\hat{\theta}$  that maximises the likelihood  $P\{\underline{X}|\theta\} = \prod_{i=1}^N f(X_i|\theta)$ . This is the maximum likelihood estimator (MLE). Of course the parameter that maximises  $P\{\underline{X}|\theta\}$  also maximises



**Figure 62.** The volume of indistinguishable distributions in parameter space is an ellipsoid.

$\log P\{\underline{X}|\theta\}$ , therefore

$$\hat{\theta}(\underline{X}) = \arg \max_{\theta} \sum_{i=1}^N \log f(X_i|\theta) \quad (20.4)$$

$$= \arg \min_{\theta} D_{KL}[P_{\underline{X}}|f(\cdot|\theta)] \quad (20.5)$$

where we used again types and got rid of constants that do not depend on  $\theta$ . Note that this is consistent with Large Deviation Theory and Sanov's theorem. You can visualise the result in the space of distributions in the following manner: the parametric family  $f(x|\theta)$  identifies a manifold in this space and the MLE identifies that point on the manifold that is closest, in terms of KL divergence, to the type.

Imagine that  $\underline{X}$  is drawn i.i.d. from  $f(x|\theta_0)$ , so there is one point on the manifold,  $\theta_0$  which is the true value of  $\theta$ . For every draw of  $\underline{X}$  the MLE  $\hat{\theta}$  will take a different value, so  $\hat{\theta}$  is a random variable. What is its distribution?

A more complete description of  $\theta$  can be obtained using Bayes rule. If  $p_0(\theta)$  is the distribution encoding all prior<sup>8</sup> knowledge on  $\theta$ , then the distribution of  $\theta$ , after we see the data (the posterior), is given by

$$p(\theta|\underline{X}) = \frac{1}{P\{\underline{X}\}} \prod_{i=1}^n f(X_i|\theta)p_0(\theta), \quad P\{\underline{X}\} = \int d\theta \prod_{i=1}^n f(X_i|\theta)p_0(\theta). \quad (20.6)$$

When  $n$  is very large, we expect this distribution to be well approximated by a Gaussian. The argument is the following: define

$$\mathcal{L}(\theta, \underline{X}) = \frac{1}{n} \sum_{i=1}^n \log f(X_i|\theta) = \int dx P_{\underline{X}}(x) \log f(x|\theta).$$

<sup>8</sup>I.e. before seeing the sample.

This is expected to be finite as  $n \rightarrow \infty$  by the law of large numbers.<sup>9</sup> Then by definition,

$$p(\theta|\underline{X}) = \frac{1}{P\{\underline{X}\}} e^{n\mathcal{L}(\theta, \underline{X})} p_0(\theta)$$

so the distribution is sharply peaked around the MLE  $\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta, \underline{X})$ :

$$\mathcal{L}(\theta|\underline{X}) = \mathcal{L}(\hat{\theta}|\underline{X}) - \frac{\Gamma}{2}(\theta - \hat{\theta})^2 + \dots$$

with  $\Gamma(\hat{\theta}) = -\partial_{\theta}^2 \mathcal{L}|_{\theta=\hat{\theta}}$ . This also implies that the integral that yields  $P\{\underline{X}\}$  can be computed by the saddle point method

$$P\{\underline{X}\} \simeq e^{n\mathcal{L}(\hat{\theta}|\underline{X})} \sqrt{\frac{2\pi}{n\Gamma(\hat{\theta})}} p_0(\hat{\theta})$$

so that

$$p(\theta|\underline{X}) \simeq \sqrt{\frac{n\Gamma(\hat{\theta})}{2\pi}} e^{-\frac{n\Gamma(\hat{\theta})}{2}(\theta - \hat{\theta})^2} \frac{p_0(\theta)}{p_0(\hat{\theta})}$$

Notice that, for  $n \gg 1$ ,  $p(\theta|\underline{X})$  is well approximated by a Gaussian because it is very small outside the interval  $|\theta - \hat{\theta}| \sim 1/\sqrt{n}$ , and  $p_0(\theta) \approx p_0(\hat{\theta})$  for  $\theta$  in this interval.

Therefore the prior plays no role in the limit  $n \rightarrow \infty$ . The variance of  $\theta$  is  $1/(\Gamma(\hat{\theta})n)$ , i.e. the typical error that we expect on the MLE is

$$|\hat{\theta} - \theta| \sim 1/\sqrt{n\Gamma(\hat{\theta})}.$$

When  $\theta$  is a vector,  $\Gamma$  is replaced by (minus) the Hessian<sup>10</sup> of  $\mathcal{L}$  at  $\hat{\theta}$ , and

$$p(\theta|\underline{X}) \simeq \left(\frac{n}{2\pi}\right)^{d/2} \sqrt{\det \Gamma(\hat{\theta})} e^{-\frac{n}{2} \sum_{\alpha, \beta} (\theta_{\alpha} - \hat{\theta}_{\alpha}) \Gamma_{\alpha, \beta}(\hat{\theta}) (\theta_{\beta} - \hat{\theta}_{\beta})} \frac{p_0(\theta)}{p_0(\hat{\theta})}.$$

<sup>9</sup>The law of large numbers implies that  $\mathcal{L}(\theta, \underline{X})$  converges to

$$\mathbb{E}[\log f(X|\theta)] = -H[\theta_0] - D_{KL}(\theta_0||\theta)$$

as  $n \rightarrow \infty$ , where we used shorthands for the entropy of  $f(x|\theta_0)$  and for the  $D_{KL}$  between  $f(x|\theta_0)$  and  $f(x|\theta)$ . This is maximal for  $\theta = \hat{\theta} = \theta_0$ , which shows that when  $n \rightarrow \infty$  maximum likelihood estimates are *consistent*, i.e. they converge to the true value.

<sup>10</sup>The Hessian is the matrix of second derivatives. In this case,

$$\Gamma_{\alpha, \beta}(\theta) = -\frac{\partial^2}{\partial \theta_{\alpha} \partial \theta_{\beta}} \mathcal{L}(\theta).$$



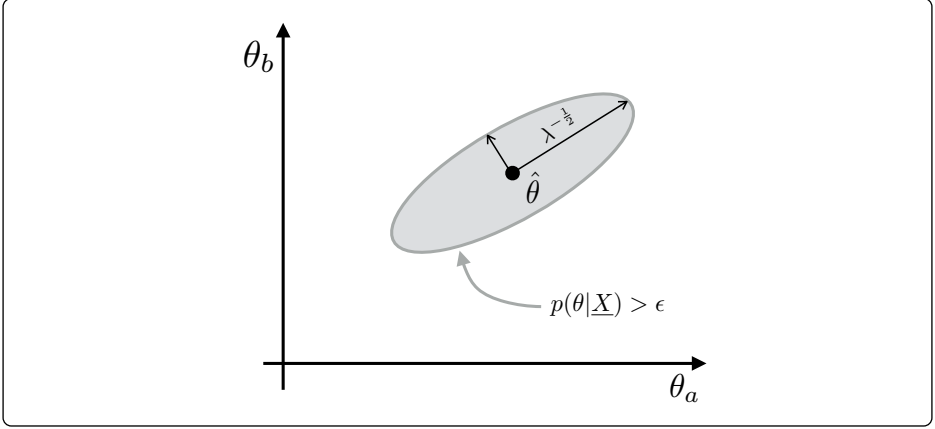


Figure 63. ???

The inverse of the Hessian yields the covariance matrix of the parameters:

$$\mathbb{E} \left[ (\theta_\alpha - \hat{\theta}_\alpha)(\theta_\beta - \hat{\theta}_\beta) \right] \simeq \frac{\Gamma_{\alpha,\beta}^{-1}}{n}.$$

The log-posterior  $\log p(\theta|\underline{X})$  around  $\hat{\theta}$  is well approximated by a quadratic function. So the region where  $p(\theta|\underline{X}) > \epsilon$  are well approximated by ellipsoids whose principal axes are aligned to the eigenvectors of  $\hat{\Gamma}(\hat{\theta})$ . The length of the axes is proportional to  $1/\sqrt{\lambda}$ , where  $\lambda$  is the corresponding eigenvalue of  $\Gamma$ . Large eigenvalues  $\lambda$  correspond to directions where the posterior probability decreases steeply, and hence the uncertainty of the parameters  $\theta$  along these directions are small. Small eigenvalues instead correspond to directions along which  $p(\theta|\underline{X})$  is flatter, and the errors on  $\theta$  may be large. These flat directions have been called *sloppy modes* by J. Sethna and collaborators [46], who have found that they appear in many cases where models with many parameters had been used to fit experimental data. In some cases, the error along sloppy modes, which is  $1/\sqrt{\lambda n}$  can be quite large. This situation occurs, when the model is very complex, i.e. it depends on too many parameters. This general situation is called *over-fitting* and is an indication that the model is not appropriate or that the dataset is not large enough.

The matrix  $\Gamma$  depends generally not only on the maximum likelihood point  $\hat{\theta}$  where the expansion is done, but also on the data  $\underline{X}$  itself. There is a case where this is not true, which is worth recalling. These are models in the *exponential family*:

$$f(x|\theta) = q(x)e^{\theta\tau(x)+\phi(\theta)} \quad (20.7)$$

You can check that the Hessian of  $\mathcal{L}$  w.r.t.  $\theta$  does not depend on  $\underline{X}$ , and the negative of the Hessian becomes<sup>11</sup>

$$\Gamma_{\alpha,\beta} = J_{\alpha,\beta}(\theta) = - \int dx f(x|\theta) \frac{\partial^2}{\partial \theta_\alpha \partial \theta_\beta} \log f(x|\theta).$$

This is the *Fisher information*.

As we have seen, exponential families describe the typical way in which large deviation are realised and they arise from a constrained maximum entropy (or minimal  $D_{KL}$ ) principle. For these models there is a sharp separation between *relevant* variables and irrelevant ones. Indeed, you can directly check that all that matters for computing the maximum likelihood parameter  $\hat{\theta}$  is the average<sup>12</sup>

$$\bar{\tau}(\underline{X}) = \frac{1}{n} \sum_{i=1}^n \tau(X_i)$$

of the random variable  $\tau(X)$ . This is called a *sufficient statistics*. All other information contained in the sample, besides the value of  $\bar{\tau}$ , is irrelevant. This is shown by the fact that the distribution of the sample conditional on the event  $\bar{\tau}(\underline{X}) = t$  is independent of  $\theta$ . Indeed,

$$\begin{aligned} P(\underline{X}|\theta, \bar{\tau}(\underline{X}) = t) &= \frac{P(\underline{X}, \bar{\tau}(\underline{X}) = t|\theta)}{P(\bar{\tau}(\underline{X}) = t|\theta)} \\ &= \frac{1}{Q(t)} \prod_{i=1}^n q(X_i), \quad Q(t) = \sum_{\underline{X}} \prod_{i=1}^n q(X_i) \delta_{\bar{\tau}(\underline{X}), t}. \end{aligned}$$

This result is a special case of the *Fisher-Neyman factorisation theorem*, that states that  $T(\underline{X})$  is a sufficient statistics if and only if  $P(\underline{X}|\theta) = Q(\underline{X})g(\theta, T(\underline{X}))$ , for some functions  $Q$  and  $g$ .

<sup>11</sup>When  $f(x|\theta)$  is given by Eq. (20.7) the Fisher information matrix is

$$J_{\alpha,\beta}(\theta) = - \frac{\partial^2 \phi(\theta)}{\partial \theta_\alpha \partial \theta_\beta}.$$

<sup>12</sup>Indeed the first order conditions of the maximisation of the likelihood yield

$$\bar{\tau}(\underline{X}) = - \frac{\partial \phi}{\partial \theta} = \mathbb{E}[\tau(X)]$$

where the expected value is taken on  $f(x|\theta)$  of Eq. (20.7).

### 20.2.1 The Data Processing Inequality and Sufficient statistics

Consider the typical inference problem:<sup>13</sup> imagine we have a theory that depends on an unknown parameter  $\theta$ . In order to gain information about  $\theta$ , we run a series of independent experiments each of which returns a measure  $X$  of a given observable. We think of  $X$  as being drawn from a distribution<sup>14</sup>  $f(x|\theta)$  that depends on  $\theta$ . Let  $\underline{X}$  be the sample obtained from the series of independent experiments. In order to “extract” the information on  $\theta$  from the sample, we form a particular combination  $T(\underline{X})$  of the variables in the sample. For example  $T(\underline{X})$  can be an estimator  $\hat{\theta}(\underline{X})$  of  $\theta$ . Yet  $\underline{X}$  is generated from  $\theta$  and  $T(\underline{X})$  from  $\underline{X}$ , i.e. in terms of Markov chains  $\theta \rightarrow \underline{X} \rightarrow T(\underline{X})$ . This means that, conditional on  $\underline{X}$ ,  $T$  and  $\theta$  are independent.

It is intuitive that  $T(\underline{X})$  cannot provide more information on  $\theta$  than the information that  $\underline{X}$  contains about  $\theta$ . I.e.

$$I(T(\underline{X}), \theta) \leq I(\underline{X}, \theta). \quad (20.8)$$

This is a particular instance of the data processing inequality which we discussed in Eq. (16.19) (see also COVER 2.8).<sup>15</sup>

$T(\underline{X})$  is called *sufficient statistics* for  $\theta$  if Eq. (20.8) holds as an equality, i.e. if  $I(T(\underline{X}), \theta) = I(\underline{X}, \theta)$ . Equivalently,  $T(\underline{X})$  is called *sufficient statistics* for  $\theta$  if  $\theta \rightarrow T(\underline{X}) \rightarrow \underline{X}$ , i.e. if conditional on  $T$ ,  $\theta$  and  $X$  are independent. This states precisely that, when  $T(\underline{X})$  is known,  $\underline{X}$  does not contain any information about  $\theta$ . Not all distributions  $f(x|\theta)$  admit a sufficient statistics for  $\theta$ . In general it is not true that the information on the generative model can be condensed into few empirical averages.

As a simple example, in the case of Bernoulli trials, the number of success in  $n$  trials is a sufficient statistics for  $p$ . Indeed, the probability  $P\{\underline{X}|k\}$  of any string  $\underline{X}$  of  $n$  trials with  $k$  successes has the same probability, which

<sup>13</sup>This part is discussed in COVER Section 2.8 and 2.9.

<sup>14</sup>Note that  $f(x|\theta)$  is derived from the theory itself we want to test. Yet the only part of the theory that is assumed to be unknown is the value of  $\theta$ . The rest is assumed to be true. This is not dissimilar from how the mass of the Higg’s boson was measured at CERN.

<sup>15</sup>As a reminder, the data processing inequality states that if there are three variables  $X$ ,  $Y$  and  $Z$  and  $X$  and  $Z$  are conditionally independent, given  $Y$ , then

$$I[X; Z] \leq I[X; Y]. \quad (20.9)$$

Conditional independence on  $Y$ , which can be denoted as  $X \rightarrow Y \rightarrow Z$ , means that the joint distribution of  $X, Y, Z$  has the form

$$p(x, y, z) = p(x|y)p(z|y)p(y)$$

so that  $p(x, z|y) = p(x|y)p(z|y)$ .

is independent of  $p$ . For any prior distribution  $\rho(p)$  of  $p$ , it is easy to find that  $I(\underline{X}, p) = I(k, p)$  by a direct calculation. Notice that the entropy of the distribution of  $\underline{X}$  is

$$H[\underline{X}] = nH(p), \quad H(p) = -p \log p - (1 - p) \log(1 - p)$$

proportional to  $n$ , whereas the entropy of  $k$  is bounded above by  $\log(n + 1)$ .<sup>16</sup> Therefore most of the information in the sample  $\underline{X}$  is irrelevant, and only a tiny fraction contains relevant information on the generative model (in this case). This fact is general, indeed, from Eq. (20.32) it is possible to compute the number of bits gained on  $\theta$ , because the mutual information  $I(\underline{X}, \theta)$  is obtained by

$$I(\underline{X}, \theta) = \mathbb{E} [D_{KL}(p(\theta|\underline{X})\|p_0(\theta))] \quad (20.10)$$

where the expected value is taken over the distribution  $p(\underline{X}) = \int d\theta p(\underline{X}|\theta)p_0(\theta)$ . For large  $n$ , within the saddle point approximation, we have that

$$D_{KL}(p(\underline{X}|\theta)\|p_0(\theta)) \simeq \frac{d}{2} \log \frac{n}{2\pi e} + \log \sqrt{\det \hat{\Gamma}(\hat{\theta})} - \log p_0(\hat{\theta}). \quad (20.11)$$

The first two terms count the number of bits needed to know a Gaussian vector to precision  $\sim 1/\sqrt{n}$  whereas the last term counts how surprising the outcome  $\hat{\theta}$  is on the basis of prior information. Only  $(\log n)/2$  bits per parameter are learned because, given the data, each  $\theta$  can be estimated to a precision  $n^{-1/2}$ .

### Exercise 20.5

What is a sufficient statistics for the Poisson distribution?

Notice that *i)* as already mentioned, the number of “useful” bits (those that are informative on  $\theta$ ) is very small compared to the total number of bits, *ii)* the leading contribution  $\frac{d}{2} \log n$  only depends on the number of parameters and it does not depend on the model used, as long as it has  $d$  parameters, finally *iii)* we learn these many bits irrespective of whether the model is right or wrong, i.e. whether the data  $\underline{X}$  are generated from  $f(x|\theta)$ , for some  $\theta_0$ , or not.

The second and last terms depend on  $\underline{X}$ . The second accounts for the uncertainty  $\delta\theta$  on the parameters and it is large when the posterior distribution on  $\theta$  is sharply peaked around its maximum  $\hat{\theta}$ . The second term is small when

<sup>16</sup>This is because  $k$  can only take  $n + 1$  values. Indeed, for  $n$  large,

$$H(k) \simeq \log[2\pi enp(1 - p)]/2.$$

the posterior distribution of  $\theta$  is broad. This is a signature of overfitting and it suggests that the modeller didn't do a good job in choosing the model. The last term informs us that we learn more if the parameters *a posteriori* turn out to attain values  $\hat{\theta}$  that are very unlikely *a priori*. Yet, a reasonable modeller would not choose a prior such that values  $\hat{\theta}$  for which the statistical errors are small are very unlikely. This suggests that the sum of the last two terms should compensate each other, in a principled approaches to statistical inference. This is the case, as we shall see.

### 20.2.2 The Fisher Information

How much information does a datapoint  $X$  carries on the unknown parameter  $\theta$ ? If  $f(X|\theta)$  varies sharply with  $\theta$ , then  $X$  provides a lot of information on the likely values of  $\theta$ , because a small deviation in  $\theta$  will result in a large deviation of  $f(X|\theta)$ . Conversely, if  $f(X|\theta)$  as a function of  $\theta$  is flat, an observation  $X$  will not make it possible to identify  $\theta$  with high precision. In order to turn this observation into a quantitative measure,<sup>17</sup> we need to consider how the information content, which is related to  $-\log f(X|\theta)$ , changes with  $\theta$ . This leads us to consider the *score*, which is a random variable defined as

$$S(\omega) = \frac{\partial}{\partial \theta} \log f(X(\omega), \theta),$$

where  $X$  is a random variable with distribution  $f(x|\theta)$ . Notice that

$$\mathbb{E}[S] = \int dx f(x|\theta) \frac{\partial}{\partial \theta} \log f(X(\omega), \theta) = \frac{\partial}{\partial \theta} \int dx f(x|\theta) = 0.$$

So the expected value of the score is not a good candidate to measure how sharply  $f(X|\theta)$  varies with  $\theta$ , i.e. how much information an observation  $X$  provides on  $\theta$ . The next natural candidate is the variance of the score, which is the *Fisher Information*. This is defined as

$$J(\theta) \equiv \mathbb{V}[S] = \mathbb{E}[S^2] \quad (20.12)$$

$$= \int dx f(x|\theta) \left[ \frac{\partial}{\partial \theta} \log f(X(\omega), \theta) \right]^2 \quad (20.13)$$

$$= - \int dx f(x|\theta) \frac{\partial^2}{\partial \theta^2} \log f(X(\omega), \theta) \quad (20.14)$$

where the last equality is derived using integration by parts (we assume that boundary terms can be neglected). The idea underlying the Fisher information

<sup>17</sup>Let us first discuss the scalar case and then the case where  $\theta \in \mathbb{R}^d$  is a  $d$ -dimensional parameter.

is that variations of the coding cost provide information on the distribution that generate a random variable, so the broader the distribution of coding costs the more informative is a variable  $X$  on the (parameters of the) distribution from which it is drawn.

### Exercise 20.6

Compute the score and the Fisher information for the binary distribution

$$f(x|\theta) = \theta^x(1 - \theta)^{1-x}, \quad x = 0, 1$$

and for the Poisson distribution

$$f(x|\theta) = \frac{\theta^x}{x!} e^{-\theta}, \quad \theta \in \mathbb{N}.$$

The same definition applies to a sample  $\underline{X}$  of  $n$  i.i.d. variables. Then the score  $S(\underline{X}) = \sum_i S(X_i)$  is just the sum of the scores, because  $f(\underline{X}|\theta) = \prod_i f(X_i|\theta)$ . Since  $S(X_i)$  are i.i.d. random variables, the Fisher information for a sample  $\underline{X}$  is  $nJ(\theta)$ .

For a model  $f(x|\theta)$  that depends on  $d$  parameters  $\theta = (\theta_1, \dots, \theta_d)$ , the score becomes a random vector with components

$$S_\alpha(\omega) = \frac{\partial}{\partial \theta_\alpha} \log f(X(\omega), \theta),$$

and the Fisher Information is the covariance of the scores, with elements:

$$J_{\alpha,\beta}(\theta) = \mathbb{E} [S_\alpha S_\beta] \quad (20.15)$$

$$= \int dx f(x|\theta) \left[ \frac{\partial}{\partial \theta_\alpha} \log f(x, \theta) \right] \left[ \frac{\partial}{\partial \theta_\beta} \log f(x, \theta) \right] \quad (20.16)$$

$$= - \int dx f(x|\theta) \frac{\partial^2}{\partial \theta_\alpha \partial \theta_\beta} \log f(x, \theta) \quad (20.17)$$

### Exercise 20.7

Compute the scores and the Fisher information matrix for the Gaussian distribution

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x-\theta_1)^2}{2\theta_2}}.$$

### 20.2.3 The Cramer-Rao bound

The significance of the Fisher information for parameter estimation is made precise by the Cramer-Rao bound. Let  $T(\underline{X})$  be an estimator of the parameter  $\theta$ . An estimator is unbiased if  $\mathbb{E}[T] = \theta$ . It is consistent if  $T(\underline{X}) \rightarrow \theta$  as  $n \rightarrow \infty$ , in probability. So for example, if  $f(x|\theta)$  is a Gaussian<sup>18</sup> with mean  $\theta$  and unit variance, the sample mean  $\bar{x} = \frac{1}{n} \sum_i X_i$  is an unbiased estimator, but also  $X_1$  is an unbiased estimator. However while  $\bar{x}$  is consistent, by the Law of Large Numbers,  $X_1$  is not. A measure of the quality of an unbiased estimator is given by its variance  $\mathbb{V}[T]$ , that characterises the statistical error of the estimate  $T(\underline{X})$ . The Cramer-Rao bound states that, given an unbiased estimator  $T(\underline{X})$  of  $\theta$ , the variance of  $T$  satisfies<sup>19</sup>

$$\mathbb{V}[T] \geq \frac{1}{nJ(\theta)}. \quad (20.18)$$

This is remarkable, because even without knowing the estimator, one can give a lower bound to its variance. Loosely speaking,  $J(\theta)$  quantifies the maximal amount of information each observation carries on the parameter  $\theta$ . Estimators that saturate the Cramer-Rao bound are called efficient.

The generalisation of Cramer-Rao bound to the case where  $\theta$  is a  $d$ -dimensional vector of parameters, states that the covariance matrix  $\hat{C}$ , whose elements are  $\mathbb{E}[(\theta_a - \mathbb{E}[\theta_a])(\theta_b - \mathbb{E}[\theta_b])]$ , satisfies

$$\hat{C} - \frac{1}{n}\hat{J}^{-1} \geq 0$$

in the sense that  $\hat{C} - \frac{1}{n}\hat{J}^{-1}$  is a non-negative definite matrix.

<sup>18</sup>An estimator for the parameter  $\theta$  can be built observing that the expected value of a function  $g(x)$  is in general a function of  $\theta$ , i.e.  $\mathbb{E}[g(X)] = G(\theta)$ . On the other hand, by the law of large numbers, we expect that  $\bar{g}(\underline{X}) = \frac{1}{n} \sum_i g(X_i) \rightarrow G(\theta)$ , as  $n \rightarrow \infty$ . Inverting this relation it is possible to obtain a consistent estimator  $\hat{\theta} = G^{-1}(\bar{g}(\underline{X}))$  of  $\theta$ . A consistent estimator is one that converges to the true value, when  $n \rightarrow \infty$ . Hence its variance vanishes as  $n \rightarrow \infty$ .

<sup>19</sup>The proof of Eq. (20.18) is straightforward. It relies on the Cauchy-Schwartz inequality  $\text{Cov}(S, T) \leq \sqrt{\mathbb{V}[T]\mathbb{V}[S]}$  and the observation that  $\text{Cov}(S, T) = \mathbb{E}[ST]$ , and

$$\begin{aligned} \mathbb{E}[ST] &= \int dx f \left[ \frac{\partial}{\partial \theta} \log f \right] T \\ &= \frac{\partial}{\partial \theta} \int dx f T = \frac{\partial}{\partial \theta} \mathbb{E}[T] \\ &= \frac{\partial}{\partial \theta} \theta = 1 \end{aligned}$$

**Exercise 20.8**

Show that if  $f(x|\theta)$  is an exponential family as in Eq. (20.7), then it admits estimators for which the Cramer-Rao bound holds as an equality.

**20.2.4 Distinguishability of distributions and Fisher information**

Further insight on the meaning of the Fisher Information is given by looking at the problem of parameter inference in the following way: let  $\hat{\theta}$  be the maximum likelihood estimate of  $\theta$  for a given sample  $\underline{X}$ . This means that  $\hat{\theta}$  maximises the log likelihood per point

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i|\theta) \quad (20.19)$$

Now imagine to consider a different sample  $\underline{X}'$  that is very similar to  $\underline{X}$  and let  $\hat{\theta}(\underline{X}')$  be the maximum likelihood estimate of  $\theta$  for  $\underline{X}'$ . If the samples are similar, we expect the maximum likelihood estimates to be close to each other, i.e.  $|\delta\theta| \ll 1$  where  $\delta\theta = \hat{\theta}(\underline{X}') - \hat{\theta}(\underline{X})$ . Can these two samples be distinguished? Can one say that they come from different distributions? Stein's lemma provides a quantitative answer to this question, for a given error threshold  $\epsilon$ . Indeed, you can think of a test between two hypotheses

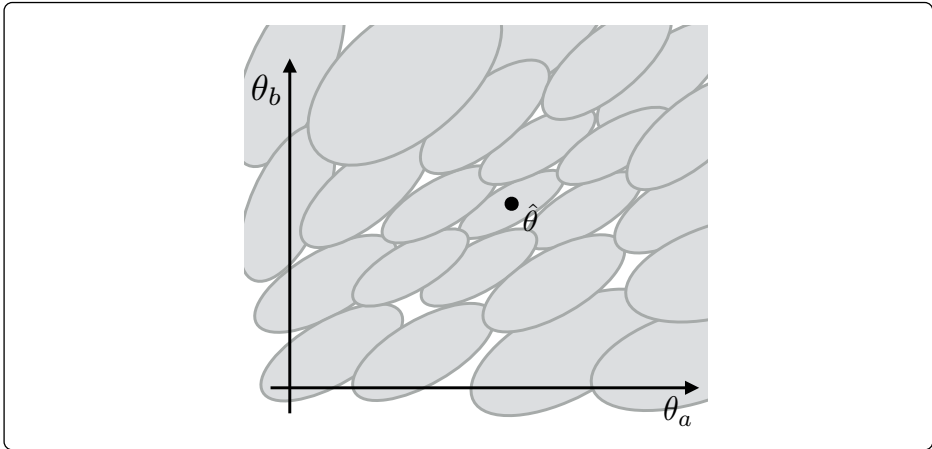
$$H_1 : Q(x) = P_1(x) = f(x|\theta), \quad H_2 : Q(x) = P_2(x) = f(x, |\theta^* + \delta\theta).$$

If we set  $\alpha = \epsilon$ , then  $\beta = e^{-nD_{KL}(f(x|\hat{\theta})||f(x|\hat{\theta} + \delta\theta))}$ . The two distributions cannot be distinguished if  $\beta \gg \epsilon$ , on the basis of a sample of  $n$  points, at a confidence level  $\epsilon$ .

This means that there is a region of distributions around the point  $\hat{\theta}$  that cannot be distinguished from  $f(x|\hat{\theta})$ . A measure of the size of this region is given by that  $\delta\theta$  for which  $\beta = \epsilon$ . This condition implies

$$\begin{aligned} -\frac{1}{n} \log \beta &= D_{KL}(f(x|\hat{\theta})||f(x|\hat{\theta} + \delta\theta)) \\ &\simeq -\frac{1}{2} \sum_{a,b} \delta\theta_a \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta_a} \log f(X|\theta) \right) \left( \frac{\partial}{\partial \theta_b} \log f(X|\theta) \right) \right] \delta\theta_b + \dots \\ &= \frac{1}{2} \delta\theta \hat{J}(\hat{\theta}) \delta\theta + \dots \end{aligned}$$





**Figure 64.** The Fisher information introduces a metric in the space of parameters. This insight is developed using differential geometry, in what is called *Information Geometry*.

where the second equality comes from the expansion of  $\log f(x|\hat{\theta}+\delta\theta)$  in small  $\delta\theta$  to second order. There are two non-trivial aspects worth remarking here. First, the linear terms in the expansion of  $D_{KL}$  vanishes. Second, although  $D_{KL}$  is not symmetric in its arguments, it is symmetric for small “distances”. So the size of the region where distributions cannot be distinguished is given by

$$\delta\theta^T \hat{J}(\hat{\theta}) \delta\theta \leq \frac{2}{n} |\log \epsilon|. \quad (20.20)$$

The intuition behind this equation is the same as that of the Cramer-Rao bound. This result also offers an insight on the nature of the mapping between the space of samples  $\underline{X}$  and the space of the parameters  $\theta$  of the models. Where the Fisher Information is large, the discriminative power of the model is larger, because the size of the region around  $\theta$  where models cannot be distinguished is smaller. Pictorially, the result above allows us to discretize the region of parameters in cells of indistinguishable models, a discretisation that becomes finer and finer as  $n$  increases. For a model with  $K$  parameters, the argument above generalises in a straightforward manner. The condition above identifies cells of elliptic shape, whose main axis are proportional to the inverse of the square root of the eigenvalues of the matrix  $J(\theta)$  and their directions are given by the eigenvectors. The volume of such cells is proportional to  $1/\sqrt{\det J(\theta)}$ . Therefore, in a region of unit volume, there are a number of distinguishable distribution which is proportional to  $\sqrt{\det J(\theta)}$ . Within a maximum entropy approach, each of these models should be *a priori* equiprobable. This means

that the non-informative prior for  $\theta$  should be

$$p_J(\theta) = \frac{1}{c_f} \sqrt{\det J(\theta)}.$$

This prior is called *Jeffrey's prior*. The normalisation constant

$$c_f = \int d\theta \sqrt{\det J(\theta)} \quad (20.21)$$

gives an estimate of the number of distinguishable models, when it is finite. The number  $c_f$  is an intrinsic degeneracy (or uncertainty) on the possible model that has generated the data, that is lifted when we observe the data. So the logarithm of  $c_f$  quantifies the information in bits that we learn about the model, and it can be taken as a measure of the intrinsic complexity of the model. A very complex model induces a fine resolution on the sample space, and it makes it possible to distinguish many samples. A sample  $\underline{X}$  can be explained well by the model with  $\theta \approx \hat{\theta}(\underline{X})$ , but it's very unlikely (or atypical) for  $\theta$  that is significantly different from  $\hat{\theta}(\underline{X})$ . On the contrary, a simple model induces a coarser resolution on the space of samples. Even very different samples can be described by the same model. In the extreme case of a model that assigns the same probability  $f(x) = p$  to all outcomes  $x$ , no samples can be distinguished, because this model assigns the same probability to all samples. This suggests a relation between maximal entropy and minimal complexity, which is reminiscent of the basic idea in coding theory: compressed representations (i.e. reduction of entropy) are achieved by exploiting the structure or patterns in the data (i.e. their complexity).

### Exercise 20.9

Verify that  $D_{KL}(p|q)$  is symmetric under exchange of the arguments if  $q$  and  $p$  are close, i.e. if  $q = p + \delta p$  with  $\delta p \ll 1$ .

### Exercise 20.10

Compute the Fisher Information for i) a binary variable  $P\{X = 1\} = p = 1 - P\{X = 0\}$ , ii) an exponential variable with mean  $\mu$ , and iii) a Gaussian with mean  $\mu$  and variance  $\sigma^2$ . Can you compute the distribution  $p_J$  for these examples? What is their *intrinsic complexity*  $c_f$ ?

## 20.2.5 Exponential families

Consider the case

$$f(x, \theta) = e^{\theta_0 + \sum_{k=1}^s \theta_k g_k(x)}$$

where the distribution depends on a vector  $\theta = (\theta_1, \dots, \theta_d)$  of parameters and

$$\theta_0 = -\log \sum_x e^{\sum_k \theta_k g_k(x)}$$

is a normalization constant. This distribution is called an exponential family, and as we have seen it has several nice properties: it is the maximum entropy distribution consistent with the constraints  $\mathbb{E}[g_k(x)] = \bar{g}_k$  for  $k = 1, \dots, d$ . As already mentioned,  $g_k(x)$  are sufficient statistics.<sup>20</sup> This is the distribution that is typically considered in statistical mechanics, where the operators  $g_k$  are the terms that define the energy. The parameters  $\theta_k$  correspond to the conjugate variables (the temperature, the chemical potential, the magnetic field, etc ...) of the observables. Then  $\theta_0$  is proportional to the corresponding thermodynamic potential.

Notice that  $\frac{\partial}{\partial \theta_k} \theta_0 = -E_{\bar{\theta}}[g_k(X)]$ . Taking a further derivative, it is easy to compute the Fisher information and to show that

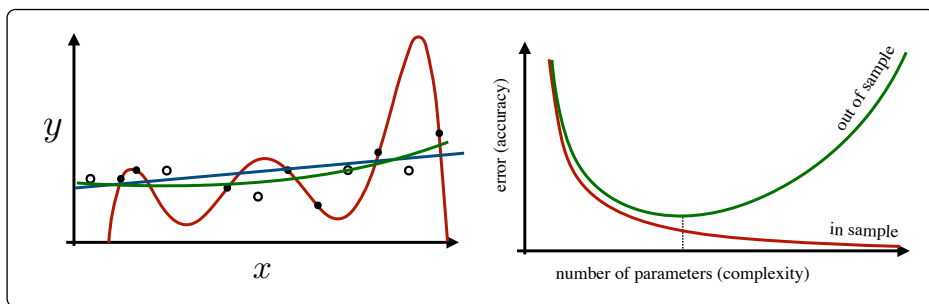
$$J_{k,\ell}(\vec{\theta}) = \mathbb{E}[g_k(X)g_\ell(X)] - \mathbb{E}[g_k(X)]\mathbb{E}[g_\ell(X)] \quad (20.22)$$

$$= -\frac{\partial^2 \theta_0}{\partial \theta_k \partial \theta_\ell} \quad (20.23)$$

$$= \frac{\partial \mathbb{E}[g_k(X)]}{\partial \theta_\ell} = \frac{\partial \mathbb{E}[g_\ell(X)]}{\partial \theta_k}. \quad (20.24)$$

The first relation tells us that the Fisher Information is the covariance matrix of the observables. The last equation tells us that it is also the matrix of susceptibilities. This is natural. In physical systems, a susceptibility tells us how much the behaviour of a system changes when we change some parameter (e.g. the temperature). Inference corresponds to the inverse problem where the behaviour of the system is known and is given by the data, whereas the parameters are the quantities one aims at computing. A model that describes well the data is one that *generalises well*, i.e. a model whose parameters  $\theta$  do not change much if the data changes a little (e.g. if a new data point is added). The best models are those with a large Fisher Information, i.e. with a large susceptibility. In physics, models with a large susceptibility are those that describe systems close to a phase transition, that exhibit anomalously large fluctuations (see Eq. (20.22)). Therefore the theory of critical phenomena in physics plays a particular role in inference problems: when a model is appropriately chosen to describe a data set, it is likely that inference will return models that are close to critical points [47].

<sup>20</sup>The sample averages of  $g_k(x)$  are sufficient statistics for  $\theta_k$ .



**Figure 65.** Left: the same data can be fitted by different models of different complexity, such as polynomial of higher degrees. Right: while the error on the training set (the sample used to fit the model) decreases, that on yet unseen (out of sample) data first decreases and then increases.

## 20.3 Model selection

Speech is silver, silence is golden.

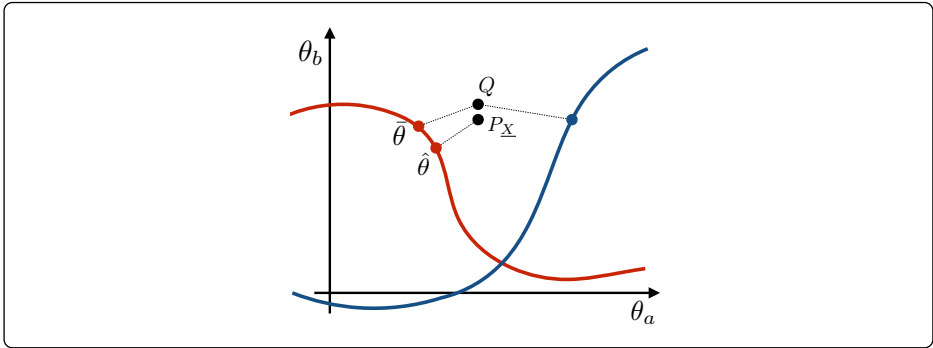
Imagine that you're not sure about the model  $f(x|\theta)$  that describes your data. In this situation, you may be uncertain among a number  $M$  of models  $f_m(x|\theta_m)$  each depending on a different number  $d_m$  of parameters  $\theta_m = (\theta_{1,m}, \dots, \theta_{d_m,m})$  ( $m = 1, \dots, M$ ). Some models may be very complex and give a detailed description of the data, others may be more parsimonious and provide a rougher description of the data. Loosely speaking, different models provide a description of the data with different degree of detail. Which model should we listen to?

The typical example is that of points  $(X_i, Y_i)$  on the plane, that may be described by different models of the form

$$Y_i = a_0 + a_1X_i + a_2X_i^2 + \dots + a_{d-2}X_i^{d-2} + \xi_i$$

where  $\xi_i$  are i.i.d. random variables from a Gaussian distribution with zero mean and variance  $\sigma^2$ , so  $\theta = (a_0, \dots, a_{d-2}, \sigma)$ . The number  $d$  of parameters can be taken as a measure of the complexity of the model, with more complex models containing simpler ones as special cases.

For a sample of  $n$  points, the model with  $d = n + 1$  parameters provides a perfect fit, because it can exactly interpolate between the points. Yet, it does not describes appropriately other data that may become available (out of sample data). This means that a very complex model does not “generalise” well. At the same time, if a new point is added and we estimate again the model, we expect that the parameters will change considerably, and the Fisher



**Figure 66.** Model selection and the AIC. Two models are shown in the space of distributions.

Information is expected to be small. Conversely, a linear fit ( $d = 3$ ) provides a less accurate description, but this will be more robust with respect to the addition of further points. So there is a trade-off between accuracy and generalisability in statistical inference, which is what we want to address here. In particular, if the data were really generated by a high order polynomial, as long as the size  $n$  of the sample is not large enough, it might well be that a linear fit might be the best option, because a fit with a high order polynomial may return coefficients that are very far from the true ones. Yet, as the number of points becomes large, we expect more and more features of the true model to surface, and then we expect the “appropriate” models to have higher and higher complexity.

So it is clear that the best model depends on a trade-off between accuracy, quantified by the log-likelihood, and complexity. But how can we quantify complexity?

### 20.3.1 Akaike Information Criterium (AIC)

Let us consider a sample  $\underline{X}$  that is generated from an unknown distribution  $Q$ . The above discussion implies that, in order to find the best model that describes  $\underline{X}$ , we cannot rely only on the likelihood. The most complex model (i.e. the one with more parameters) will also be the most likely one. However this model will not be efficient in generalisation, i.e. in describing data which have not been used to estimate the model’s parameters. Akaike proposed a method to correct the maximum likelihood, by adding a *complexity penalty*, in order to score different models.

The starting point of the AIC is that the appropriate quantity to score the validity of a model  $f(x|\theta)$  is the distance

$$D_{KL}(Q\|\hat{\theta}) = S(Q|Q) - S(Q|\hat{\theta}) \quad (20.25)$$

between the true distribution and the maximum likelihood distribution,  $f(x|\hat{\theta})$ . Here we used the shorthand<sup>21</sup>  $S(Q|P) = \int dx Q(x) \log P(x)$  for the likelihood of  $P$  with respect to  $Q$ . The maximum likelihood estimator  $\hat{\theta} = \arg \max_{\theta} S(P_{\underline{X}}|\theta)$  does not coincide with the point

$$\bar{\theta} = \arg \max_{\theta} S(Q|\theta)$$

where  $D_{KL}(Q\|\theta)$  is minimal. Yet these two points are close, when  $n \gg 1$ , because  $P_{\underline{X}} \rightarrow Q$  when  $n \rightarrow \infty$ , and hence  $\hat{\theta} \rightarrow \bar{\theta}$ , because the two optimisation problems that define  $\hat{\theta}$  and  $\bar{\theta}$  coincide asymptotically.

Then, we can estimate  $S(Q|\hat{\theta})$  by expanding  $S(Q|\bar{\theta})$  around its maximum  $\bar{\theta}$

$$S(Q|\hat{\theta}) = S(Q|\bar{\theta}) - \frac{1}{2} \sum_{a,b} (\hat{\theta}_a - \bar{\theta}_a) \mathbb{E}_Q \left[ \frac{\partial S_b}{\partial \theta_a} \right] (\hat{\theta}_b - \bar{\theta}_b) + \dots \quad (20.26)$$

where we used the definition of the score,  $S_a = \frac{\partial}{\partial \theta_a} \log f(X|\theta)$ . To leading order, we can replace the expected value over  $Q$  in the quadratic form, with the expected value over  $f(x|\bar{\theta})$ , therefore

$$\mathbb{E}_Q \left[ \frac{\partial S_b}{\partial \theta_a} \right] \approx \int dx f(x|\bar{\theta}) \frac{\partial^2}{\partial \theta_a \partial \theta_b} \log f(x|\bar{\theta}) = -\mathbb{E} [S_a S_b] = -J_{a,b}$$

where we used the fact that  $\mathbb{E} [S_a] = 0$ . Therefore

$$S(Q|\hat{\theta}) \simeq S(Q|\bar{\theta}) + \frac{1}{2} \sum_{a,b} (\hat{\theta}_a - \bar{\theta}_a) J_{a,b}(\bar{\theta}) (\hat{\theta}_b - \bar{\theta}_b) + \dots \quad (20.27)$$

$$\begin{aligned} S(Q|\hat{\theta}) &\simeq S(Q|\bar{\theta}) + \frac{1}{2} \sum_{a,b} J_{a,b}(\bar{\theta}) \mathbb{E} \left[ (\hat{\theta}_a - \bar{\theta}_a) (\hat{\theta}_b - \bar{\theta}_b) \right] + \dots \\ &\simeq S(Q|\bar{\theta}) - \frac{1}{2n} \sum_{a,b} J_{a,b} J_{a,b}^{-1} + \dots = S(Q|\bar{\theta}) - \frac{d}{2n} + \dots \end{aligned} \quad (20.28)$$

In Eq. (20.28) we approximate the quadratic form with its expected value over the distribution  $\bar{\theta}$ , and use the fact that  $\hat{\theta}$  is a Gaussian random variable with expected value  $\bar{\theta}$  and covariance  $\hat{J}^{-1}/n$ .

<sup>21</sup>  $-S(Q|P) = H[Q] + D_{KL}(Q\|P)$  is called the *cross entropy*. We also use the simplified notation  $D_{KL}(Q\|\theta)$  for the Kullback-Leibler divergence between  $Q$  and  $f(x|\theta)$ .

We use this expression in Eq. (20.25) and we replace  $Q$  with  $P_{\underline{X}}$  in the first argument of  $S(\cdot|\cdot)$  of the resulting expression, i.e.<sup>22</sup>

$$D_{KL}(Q\|\hat{\theta}) \simeq S(Q|Q) - S(Q|\bar{\theta}) + \frac{d}{2n} + \dots \quad (20.29)$$

$$\simeq S(P_{\underline{X}}|Q) - S(P_{\underline{X}}|\bar{\theta}) + \frac{d}{2n} + \dots \quad (20.30)$$

As a function of  $\theta$ ,  $S(P_{\underline{X}}|\theta)$  is maximal at  $\hat{\theta}$ . We then expand around  $\hat{\theta}$  and estimate the leading term as in Eqs. (20.26)–(20.28), which leads to

$$S(P_{\underline{X}}|\bar{\theta}) \simeq S(P_{\underline{X}}|\hat{\theta}) - \frac{d}{2n} + \dots$$

Taken together, this leads to

$$D_{KL}(Q\|\hat{\theta}) \simeq S(P_{\underline{X}}|Q) - S(P_{\underline{X}}|\hat{\theta}) + \frac{d}{n} + \dots \quad (20.31)$$

When the last expression is used to compare different models, the first term  $S(P_{\underline{X}}|Q)$  is the same for all models, whereas the other two can be computed from the data and the knowledge of each model. This means that the likelihood per data point of each model has to be penalised by a term  $-d/n$  which only depends on the number of parameters of each model.

For some unknown reason, Akaike defined his complexity with a factor  $-2n$ , i.e. he defines

$$AIC = 2d - 2 \sum_{i=1}^n \log f(X_i|\hat{\theta})$$

as the score that should be used to compare different models, suggesting that the model with the lowest AIC should be preferred.

The way in which AIC should be thought is as an estimate of the expected value of the Kullback-Leibler divergence between the true model and the model with maximum likelihood parameters.

### 20.3.2 Bayesian Information Criterion (BIC)

A different way to address the question of how to penalise models for their complexity comes from a direct Bayesian approach. Before seeing the data,

<sup>22</sup>The distance between  $Q$  and  $P_{\underline{X}}$  vanishes as  $n \rightarrow \infty$ . Likewise  $\hat{\theta}$  is expected to converge to  $\bar{\theta}$  as  $n \rightarrow \infty$ . Yet the distance between  $Q$  or  $P_{\underline{X}}$  and the distribution  $f(\cdot|\theta)$  does not vanish. The reason why the substitution  $Q \rightarrow P_{\underline{X}}$  cannot be done directly in Eq. (20.25) is that  $\hat{\theta}$  also depends on  $\underline{X}$ , whereas  $\bar{\theta}$  does not.

you may have some prior information on which model is most likely. This is encoded in the prior probability  $P_0(m)$ ,  $m = 1, \dots, M$ , where  $M$  is the number of models. If you have no prior information at all, maximum entropy suggests that you should take  $P_0(m) = 1/M$ , which is what I will assume.

When you see the data  $\underline{X}$ , this allows you to revise your prior estimate of the models, and to compute the *posterior* probability  $P(m|\underline{X})$  using Bayes rule

$$P(m|\underline{X}) = \frac{P(\underline{X}|m)P_0(m)}{\sum_{m'=1}^M P(\underline{X}|m')P_0(m')} \quad (20.32)$$

where the likelihood of model  $m$

$$P(\underline{X}|m) = \int d\theta_m f_m(\underline{X}|\theta_m) p_{0,m}(\theta_m)$$

is computed by averaging the likelihood of each model over the prior distribution  $p_{0,m}(\theta)$  of the parameters. In order to estimate  $P(\underline{X}|m)$ , for  $n$  large, we observe that the integrand is of the form  $f(\underline{X}|\theta_m) = e^{n\mathcal{L}(\theta_m)}$  (see Eq. (20.19)) and hence we can resort to the saddle point method. Hence we find the maximum  $\hat{\theta}_m$  of  $\mathcal{L}(\theta_m)$  and expand around it. To second order

$$\mathcal{L}(\theta_m) = \mathcal{L}(\hat{\theta}_m) - \frac{1}{2} \sum_{j,k} (\theta_{j,m} - \hat{\theta}_{j,m}) J_{j,k}(\hat{\theta}_m) (\theta_{k,m} - \hat{\theta}_{k,m}) + \dots$$

Upon changing variables to  $z_j = \sqrt{n}(\theta_{j,m} - \hat{\theta}_{j,m})$ , the integral can be done by Gaussian integration, with the result

$$P(\underline{X}|m) \simeq e^{n\mathcal{L}(\hat{\theta}_m) - \frac{d_m}{2} \log n - C_m} \quad (20.33)$$

where

$$C_m = \frac{1}{2} \log \det J(\hat{\theta}_m) - \frac{d_m}{2} \log(2\pi) - \log p_{0,m}(\theta_m)$$

is a constant (see [48] for a detailed discussion).

In the simplest case where the models are *a priori* equally probable ( $P_0(m) = 1/M$ ), the most probable model is the most likely one, i.e. the one that has the largest likelihood  $P(\underline{X}|m)$ . Yet this is not only given by the value of the likelihood at the maximum  $\hat{\theta}_m$ , but it is also penalised by terms (the second and third in the exponent of Eq. (20.33)) that account for the complexity of the model  $m$ .

Notice that while the likelihood term in the exponent is proportional to  $n$ , the second is proportional to  $\log n$  and the third is a constant. Therefore, for  $n$



very large, the term that dominates is the likelihood, but for smaller values of  $n$  the second term becomes important, and for even smaller values of  $n$  even the third term becomes important. Notice in particular, that when  $n$  is not very large, even the choice of the prior becomes relevant. In these situations, choosing a prior such as  $p_J$  that introduces no bias in the space of samples, becomes important.

Selecting the best model based on the first two terms is called Bayesian Information Criterium (BIC) whereas the last term is usually associated to the Minimal Description Length (MDL) [48]. Both these two criteria are more severe in penalising models than AIC.

A final remark. At a formal level, for  $n$  very large, the denominator in Eq. (20.32), which is nothing but the probability of  $\underline{X}$  (called *evidence* in inference), looks like a partition function

$$P\{\underline{X}\} \simeq \sum_{m=1}^M e^{-nF_m}, \quad F_m = -\mathcal{L}(\hat{\theta}_m) + \frac{K_m}{2n} \log n + \frac{C_m}{n} \quad (20.34)$$

where  $n$  plays the role of the inverse temperature and, as we have seen,  $F_m$  can be regarded as the free energy of the model  $m$ . Imagine a situation such as that described at the beginning of this section, where model  $m$  includes as special cases simpler models. Passing from model  $m$  to model  $m+1$  entails *switching on* a coefficient  $a_{m-1}$  that was set to zero in model  $m$ . This is equivalent to breaking a symmetry between models with  $a_{m-1} > 0$  and models with  $a_{m-1} < 0$ . As we have seen, as  $n$  increases (i.e. as the temperature decreases), the complexity of the model that dominates the sum in Eq. (20.34) typically increases. This is a common phenomenon in physics: as the temperature decreases, the state of matter passes through phases of decreasing degrees of symmetry.

**An illustrative case: two states Dirichelet's model.** Consider a repeated experiment where there are two possible outcomes  $X = 0, 1$  and assume there are  $n$  independent observations,  $k$  with  $X = 1$  and  $n - k$  with  $X = 0$ . There are two possible models: in the first the two states are equiprobable, i.e.  $p_0(X = 1) = p_0(X = 0) = 1/2$ . In the second, the states have different probabilities  $p_1(X = 1) = p = 1 - p_1(X = 0)$ . These correspond to different models that we can identify with different partitions of “states”  $X$ , according to their probabilities. So the first case corresponds to a model  $\mathcal{M}_0 = [(\{0, 1\}, 1/2)]$  where the two states are symmetric because they have the same probability, whereas the second to a model  $\mathcal{M}_1 = [(\{0\}, 1-p), (\{1\}, p)]$ . Clearly  $P\{\underline{X}|\mathcal{M}_0\} = 2^{-n}$  whereas for  $\mathcal{M}_1$  the likelihood  $P\{\underline{X}|\mathcal{M}_1\}$  can be obtained by integrating

the likelihood over the prior distribution of the parameter  $p$ , for which we take a Dirichelet form<sup>23</sup>

$$P_0(p) = \frac{\Gamma(a)^2}{\Gamma(2a)} p^{a-1} (1-p)^{a-1}.$$

The probability of the data  $\underline{X}$  given model  $\mathcal{M}_1$ , is obtained averaging the likelihood  $p(\underline{X}|p, \mathcal{M}_1) = p^k (1-p)^{n-k}$  over the prior  $P_0(p)$  and one obtains:

$$P\{\underline{X}|\mathcal{M}_1\} = \frac{\Gamma(2a)\Gamma(k+a)\Gamma(n-k+a)}{\Gamma(a)^2\Gamma(n+2a)}. \quad (20.35)$$

In order to compare the two models, we invoke Bayes rule and compute the posterior probability

$$P(\mathcal{M}_i|\underline{X}) = \frac{P(\underline{X}|\mathcal{M}_i)P_0(\mathcal{M}_i)}{\sum_j P(\underline{X}|\mathcal{M}_j)P_0(\mathcal{M}_j)} = \frac{P(\underline{X}|\mathcal{M}_i)P_0(\mathcal{M}_i)}{P(\underline{X})}$$

where  $P_0(\mathcal{M}_i)$  is the prior probability of model  $i$ . For the sake of simplicity, we're going to assume that all models are *a priori* equally likely.<sup>24</sup> So the most probable model is the one with the highest likelihood  $P\{\underline{X}|\mathcal{M}\}$ . In the present case, it is easy to check that, for  $n \gg 1$ , in the representative case of a uniform prior ( $a = 1$ ) we have that as long as

$$\left| \frac{k}{n} - \frac{1}{2} \right| < \sqrt{\frac{\log(2n/\pi)}{8n}}$$

the symmetric model  $\mathcal{M}_0$  should be preferred.

This argument extends in a straightforward way to the general case where the outcome  $X$  can take more than two values or states. The argument above suggests that, in the general case, for each pair of states  $X = s$  and  $X = s'$  their probability should be the same, unless they occur in the data a sufficiently different number of times. If  $k_s \approx k_{s'}$  instead, they should be assigned the

<sup>23</sup>This choice is convenient because the posterior distribution over model  $\mathcal{M}_1$ , which is obtained by Bayes rule,

$$P(p|\hat{s}, \mathcal{M}_1) = \frac{\Gamma(k+a)\Gamma(n-k+a)}{\Gamma(n+2a)} p^{k+a-1} (1-p)^{n-k+a-1}.$$

keeps the form of a Dirichelet's distribution. Priors with these property are called *conjugate priors*.

<sup>24</sup>By Occam's razor, one would be tempted to prefer simpler models, i.e. those with fewer parameters. Yet Occam's razor already arises from the integration over the parameters implied by Bayes rule, without the need to introduce it *ad hoc*.

same probability, i.e. the symmetry between states  $s$  and  $s'$  should not be broken.

Given the set  $\mathcal{S}$  of states  $s$  that are seen (with multiplicity  $k_s > 0$ ), then a generic model  $\mathcal{M} = [\mathcal{Q}, \vec{\mu}]$  is one where different states are divided into a partition

$$\mathcal{Q} = (Q_1, Q_2, \dots, Q_N), \quad \bigcup_{q=1}^N Q_q = \mathcal{S}$$

of a number  $N$  of disjoint sets, and each state in the  $q^{\text{th}}$  subset of the partition ( $s \in Q_q$ ) has the same probability  $\mu_q$ . If  $m_q = |Q_q|$  is the number of states in subset  $Q_q$ , then  $\mu_q$  satisfies the normalisation

$$\sum_q m_q \mu_q = 1. \quad (20.36)$$

Any possible partition corresponds to a different model, including the one where each state is in the same subset ( $s \in Q_1, \forall s$ ), and the one where each state is in a different subset ( $s \in Q_s, \forall s$ ). Hence, each partition  $\mathcal{Q}$  identifies a different model  $\mathcal{M}$ , and it is possible to carry out Bayesian model selection on the set of all these models. We refer to Haimovici and Marsili [57] for a detailed discussion. In brief, what one finds is that the most likely models are those that group states that are observed a similar number of times in the same subset of the partition. It is clear that, as  $n$  increases, unless some symmetry implies that some states should have the same probability, the degeneracies between states will be lifted because all states will be observed a sufficiently different number of times. This series of successive *symmetry breaking* events as  $n$  increases, is equivalent to phase transitions in physics, as the temperature decreases, according to the analogy discussed above.

### 20.3.3 Minimum Description Length

Conclusions very similar to those of Bayesian Model Selection, can be drawn from a seemingly very different perspective, that of Minimum Description Length. The problem is the following: Alice chooses a value of  $\theta$ , draws  $n$  i.i.d. samples from  $f(x|\theta)$  and send them over to Bob. Before receiving  $\underline{X}$ , he has to make enough space available on his hard drive. Bob does not know  $\theta$  but he knows that the sample  $\underline{X}$  is drawn from  $f(x|\theta)$ . How much space should he reserve?

If Bob knew the distribution  $P(\underline{X})$  he could store efficiently the data in  $-\log P(\underline{X})$  bits. If he could see the data before deciding how much space to reserve, then he could compute the MLE  $\hat{\theta}$  and use  $P(\underline{X}) = f(\underline{X}|\hat{\theta}(\underline{X}))$

for coding the data, so he would need  $-n\mathcal{L}(\hat{\theta})$  bits. However he has to take this decision before seeing the data. There are several equivalent ways to see the problem: one is to ask Alice to send the parameters  $\theta$  before making the decision. Then Bob should set aside enough space to store  $\theta$  besides the space  $-n\mathcal{L}(\theta)$  needed to store  $\underline{X}$ . But then Bob would need to know what is the distribution  $p_0(\theta)$  from which Alice has drawn  $\theta$ , which looks like the problem of choosing a prior.

A different solution is given by assuming that Bob wants to avoid at all cost to end up in a situation where he would not have enough space. To play it safe, he will assume that, whatever  $P(\underline{X})$  he chooses to encode the data, Alice, knowing it, will choose the worst possible sample  $\underline{X}$ . Knowing this, Bob will choose the  $P(\underline{X})$  that minimises the amount of disk space he has to reserve. This problem can be formalised by introducing the regret of  $P$  for  $\underline{X}$

$$\mathcal{R}(\underline{X}, P) = -\log P(\underline{X}) + \log f(\underline{X}|\hat{\theta})$$

which is the difference between the number of bits used by Bob, if he adopts the code  $P(\underline{X})$  and the minimal possible coding cost  $-\log f(\underline{X}|\hat{\theta})$ . Then the MDL code is

$$\bar{P} = \arg \min_P \max_{\underline{X}} \mathcal{R}(\underline{X}, P).$$

The solution to this minimax problem turns out to be surprisingly simple:

$$\bar{P}(\underline{X}) = \frac{f(\underline{X}|\hat{\theta}(\underline{X}))}{\sum_{\underline{X}'} f(\underline{X}'|\hat{\theta}(\underline{X}'))}$$

which is called the *normalised maximum likelihood*. The number of bits needed by Bob are

$$-\log \bar{P}(\underline{X}) = -\log f(\underline{X}|\hat{\theta}(\underline{X})) + \log \sum_{\underline{X}'} f(\underline{X}'|\hat{\theta}(\underline{X}')) \quad (20.37)$$

The second term must be equal to the additional disk space Bob would need to store the parameters  $\hat{\theta}$ . The best model should be the one that allows for the most concise description, i.e. with the minimal value of  $-\log \bar{P}(\underline{X})$ .

In order to estimate the second term of Eq. (20.37), let us assume that  $f(x|\theta)$  belongs to an exponential family (so the Hessian of the likelihood is given by the Fisher Information) and consider the integral

$$\int d\theta \sqrt{\det J(\theta)} f(\underline{X}|\theta) \simeq \left( \frac{2\pi}{n} \right)^{d/2} f(\underline{X}|\hat{\theta}(\underline{X}))$$

where we used saddle point integration. Now summing over  $\underline{X}$  and taking the logarithm one finds

$$\log \sum_{\underline{X}'} f(\underline{X}' | \hat{\theta}(\underline{X}')) \simeq \frac{d}{2} \log \frac{n}{2\pi} + \log \int d\theta \sqrt{\det J(\theta)}.$$

The first term comes because each of the  $d$  parameters in  $\hat{\theta}$  is known to a precision  $1/\sqrt{n}$ , which requires  $(\log n)/2$  bits. In addition the parameters are not independent, which is what the second term accounts for. So the last term in the expression above encodes the *intrinsic complexity* of the model  $f(x|\theta)$ .

Codes in MDL are efficient in a very precise manner:  $P(\underline{X})$  provides a generative model for samples generated as i.i.d. draws from  $f(x|\theta)$  from some unknown  $\theta$ . This means that the code-length

$$\ell = -\frac{1}{n} \sum_{\underline{X}} \bar{P}(\underline{X}) \log \bar{P}(\underline{X})$$

achieved by MDL is the smallest possible. In order to check this idea, one can study the large deviations of the code-length. This is discussed in Cubero et al. [58] that finds that MDL codes sit precisely at a phase transition in terms of the code-length. There are only distributions that encode samples with a higher code-length, attempts to achieve a lower coding cost triggers a localisation phase transition like the ones we discussed for fat tailed distributions.

## 20.4 The high dimensional limit and beyond

Yet in truth there is no form that is with or without features; he is cut off from all eyes that look for features. With features that are featureless he bears a featured body, and the features of living beings with their featured bodies are likewise.

(the Immeasurable Meanings Sutra, foreword to the Lotus Sutra)

All the discussion up to now has focused on the limit  $n \rightarrow \infty$  when the range of variability  $|\chi|$  of the data points  $X$  were kept fixed.

There are cases, which are of considerable current research interest, where the data is very high dimensional. Examples range from recording of neural activity and gene expression data to time series in economics and finance. It is not rare that each point  $X_i$  consists of a point in a  $d$ -dimensional space, where the dimension can range in the thousands, and that the sample size  $n$  consists of few hundreds of data points.

Describing these data necessarily requires models that depend on many parameters, at least as many as the number  $d$  of variables. Things are made worse by the fact that, in the end, what one would like to estimate are the interactions among the variables that are responsible for the observed behaviour. Yet, if the number of variables is  $d$ , the number of possible pairwise interaction grows with  $d^2$ . The situation is even worse as in many of these systems we have no reason to believe that pairwise interactions are the relevant ones. The number of three body interactions grows in number as  $d^3$  etc ... and the total number of possible interactions among  $d$  variables is  $2^d - 1$ . Even Big Data is not big enough.

This is clearly a situation where the saddle point approximation used in the previous section becomes questionable and all the results we discussed so far cannot be applied. There are two different ways of approaching statistical inference in these situations. The first invokes *regularisation* schemes that inhibit large fluctuations of inferred parameters by constraining them. In practice this entail introducing priors on parameters. For example,  $L_2$  regularisation correspond in maximising an objective function that is a linear combination of the log-likelihood and the sum of squares of the parameters. This implicitly corresponds to assuming a Gaussian prior distribution on parameters.

A different approach is that of resorting to *dimensional reduction* schemes, such as principal component analysis (PCA) or data clustering. PCA aims at identifying directions in the  $d$  dimensional space along which the data exhibit a significant variation. In its simplest form, these directions correspond to the eigenvectors of the largest eigenvalues of the covariance matrix.

Data clustering aims at mapping each point  $X$  of the sample to a discrete variable  $s = 1, \dots, S$ , which is the label of the cluster to which point  $X$  belongs. The mapping  $X \rightarrow s$  aims at grouping similar points in the same cluster, where similarity is defined in terms of a distance between points  $X_i$  in the sample. For the same data, there is a large choice of data clustering methods, based on different distances and algorithms. Which of these methods should one choose?

Any regularisation, dimensional reduction or data clustering approach implicitly entail some assumptions on the data. For example,  $L_2$  regularisation implies a Gaussian prior, as mentioned above, and PCA implicitly assumes a Gaussian generative model because the method is based on pairwise statistics. These may be strong hypotheses in the high dimensional regime, which may be uncontrollable or arbitrary specially in case where the generative model of the data is completely unknown. Furthermore, statistical inference in the high dimensional limit is strongly affected by computational limits, that in many

cases limit the choices to algorithms. Algorithms that require computational times that scale as the square of the dimensionality or of the number of data points are often already unaffordable.<sup>25</sup>

The theory discussed in these lecture notes may provide a critical analysis of what assumptions we are projecting on the data when using one or the other method of data analysis.

## 20.5 Beyond statistical inference: learning and intelligence

All the approaches discussed thus far do not address the fundamental issue of what learning actually is. All the problems we discussed define learning at the outset, providing a solution in terms of an optimisation problem. Learning is a fundamental feature of living beings. A fundamental aspect of it is that it must be possible to identify “interesting” patterns in the data before understanding why they are interesting. Furthermore it must be possible to do so on the basis of very little data. This is possible because uninteresting data (noise) is described by maximum entropy distribution and, as such, it is detectable. This leads to a notion of learning intended as “making sense of data that make sense” that has been developed on the basis of a notion of *relevance* (see e.g. [36]). This goes well beyond the material discussed in this lecture notes, but it makes sense to discuss how far the landscape of concepts we have discussed thus far may bring us in addressing fundamental issues in cognition. Indeed David Marr [55] has argued that the conceptual underpinnings of cognitive functions are independent of whether they are implemented in-silico or in a biological brain. If these are true, these functions should be based on principles of information theory and statistics.

It must be said at the outset that one of the main hurdles in this venture is that a precise definition of concepts such as awareness, intelligence or consciousness is still lacking.<sup>26</sup> While these concepts may be hard to define in general, it may be easier to define them within the limited scope of some simple models. Therefore, following Marr, one can address these questions studying simple artificial neural network models trained on complex data, by dissecting their internal states.

---

<sup>25</sup>Further computational issues arise when the inference problem involves non-convex optimisation problems. In these cases, inference may turn out to be a computationally hard problem. Issues of this type arise in the high noise regime of signal detection problems, see e.g. [30].

<sup>26</sup>Until recently, we relied on Turing’s test as an operational definition of intelligence. The advent of large language models has shown all the limits of this definition.

For example, an absolute, quantitative notion of relevance can distinguish systems that “know” from those that “do not know”. Those that do not know, like systems in statistical mechanics, have an internal state consistent with the maximum entropy principle. The internal state of systems that know should instead be described by states of maximal relevance. If the relevance of the internal state of a system can be measured, then it can also be measured by the system itself, leading to a very rudimentary notion of *awareness* as “knowing to know”. Most importantly, an information theoretic notion of relevance would allow a system to know that it knows irrespective of what it knows, just like the entropy measures information content irrespective of what that information is about. What type of architectures would support this function in a neural network? And what architecture would support an infinite recursion of “knowing of knowing of ... knowing to know”, which may approach a primitive notion of consciousness?

As for intelligence, it has been argued that intelligent behaviour relies on “extreme generalisation [intended as] the ability to handle entirely new tasks that only share abstract commonalities with previously encountered situations, applicable to any task and domain within a wide scope” [52]. This suggests that the ability of abstraction is a prerequisite for intelligence. Finding “abstract commonalities” requires a representation that may encompass a wide variety of tasks and which, therefore, should be independent of any task. In other words, an intelligence spanning an unbounded scope of tasks should navigate a universal map with a metric defined in terms of “abstract commonalities”. But how does this abstract, universal representation come about? This is a question which has been much debated in the context of language. Chomsky has convincingly shown that languages share a common structure — the so-called *universal grammars* — that entails the capacity of infinite recursion [53] thus making it possible to generate an infinite variety of sentences with a finite vocabulary.<sup>27</sup> The fact that this capacity emerges in children without exposure to much data (spoken language) has led to the hypothesis that universal grammars need to be biologically hardwired, an hypothesis that is not widely accepted [54]. Yet, such universal representations could emerge spontaneously in deep cortical areas which integrate input inputs from a broad set of sources, across all sensory modalities, not just from spoken language. While this hypothesis is hard to test in the context of language, it is much easier to test it within simple machine learning models. When these are trained on complex data of increasing variety one should

---

<sup>27</sup>The actual form of language as it is spoken or written derives from this universal grammar through a series of transformations that encodes abstract semantic structures as well as grammatical rules.



expect the internal representations of these models to converge to universal distributions.

I hope that the material presented in these lecture notes can help addressing deep questions about cognition, learning, and intelligence, in the simplest possible models, anchoring approaches on principled theoretical frameworks rooted in the fundamental laws of information and probability.



# Chapter 21

## Exercises for the second part

These exercises are of different degree of complexity, some are open ended, but given the material discussed in the lectures, you should be able to tackle them.

1. Imagine that  $O$  is friend with  $A$ , and  $A$  has  $n$  friends  $B_1, B_2, \dots, B_n$ , each of which is friend with  $C$  (who is not a friend of  $A$  and  $O$ ). You can draw a graph where persons are the nodes and links are friendship. Imagine that  $O$  is positive for a virus that can be transmitted to friends with probability  $p$ . Compute the probability that  $C$  gets infected. Compute it in the case  $p = 1/2$  and  $n = 4$ .
2. Let  $Z_n = \max\{X_1, \dots, X_n\}$  be the maximum of  $n$  independent and identically distributed random variables  $X_i \geq 0$  with pdf  $p(x) = \gamma x^{\gamma-1} e^{-x^\gamma}$ . Find sequences  $a_n$  and  $b_n$  such that the variable  $Y_n$  defined by  $Z_n = a_n + b_n Y_n$  of has a non-degenerate distribution as  $n \rightarrow \infty$ . Find  $P\{Y < x\}$ .
3. Let  $X$  be a non-negative integer random variable with expected value  $\lambda$ . What is your best estimate of its second moment? Imagine that you get a sample of  $N \gg 1$  i.i.d. observations  $X_i$  of  $X$  and that the sample mean is close to  $\lambda$ . Yet the sample second moment is twice as small as this best prediction. What do you conclude? What if instead you find that the sample second moment is twice as large as what you expect?
4. Consider the variable

$$Y = \alpha X + Z$$

where  $X$  and  $Z$  are independent Gaussian variables with mean zero and unit variance. Compute the mutual information  $I(X, Y)$ .

5. Compute the large deviation function for random variables with distribution  $p(x) = e^{-x-e^{-x}}$ ,  $x \in \mathbb{R}$ .
6. Consider the *accelerated* random walk  $S_{n+1} = S_n + nX_n$  where  $X_n$  are i.i.d. random variables that take values  $\pm 1$  with the same probability and  $S_0 = 0$ . Find the value of  $a$  for which the variable  $Z_n = n^{-a}S_n$  admits a limiting distribution. Using this show that the probability that the accelerated random walk returns to a neighbourhood  $S_n \in [-K, K]$  of the origin infinitely often is zero ( $K$  is a finite positive integer).
7. Winning in desperate situations. Imagine you are a coach and, in the game you are playing there are  $n$  rounds left and your team is under by  $n\gamma > 0$  points. The final score of the match will be

$$\sum_{i=1}^n X_i - n\gamma$$

in your favour, where  $X_i$  is the score difference in round  $i$  ( $X_i > 0$  is in your favor,  $X_i < 0$  if your opponent scores).

Assume that  $X_i$  is i.i.d. (very unrealistic, but...) drawn in each round from the probability

$$Q_q(x) = \begin{cases} 0 & \text{w.p. } 1 - 2q \\ +1 & \text{w.p. } q(1 - \eta q) \\ -1 & \text{w.p. } q(1 + \eta q) \end{cases}$$

with  $\eta > 0$ . You can choose the parameter  $q \in [0, 1/2]$ . Informally, you can decide how much to attack or defend, but if you decide to attack (large  $q$ ) then your opponent can score more easily ( $\eta > 0$ ).

In the end, you are interested in events  $E = \{\sum_i X_i \geq n\gamma\}$  in which you win or draw the match. How would you find the best “tactics”  $q^*$ ?

8. Consider a machinery that can undergo mis-functions at random times. The waiting time distribution (pdf) for mis-function events is  $p(\tau) = e^{-\tau}$  (i.e. mis-functions is a Poisson process). The machinery breaks down completely when  $n$  consecutive mis-function events occur. Estimate the probability that the time to breakdown is larger than  $n\bar{\tau}$ , with  $\bar{\tau} \in \mathbb{R}^+$ , and  $n$  is very large.
9. Compute the large deviation function  $I_m(\bar{x})$  for random variables  $X_i \geq 0$  with distribution  $p(x) = x^{m-1}e^{-x}/\Gamma(m)$  and for binomial random

variables  $P(X_i = k) = \binom{m}{k} p^k (1-p)^{m-k}$ . Check that, in both cases, the solution has the scaling form  $I_m(\bar{x}) = m I_1(\bar{x}/m)$ . Why is this so?

10. Consider a random walk  $S_n = X_1 + X_2 + \dots + X_n$  where the steps  $X_i$  are i.i.d. random variables with pdf  $p(x) = \frac{1}{4} e^{-\sqrt{|x|}}$  for  $x \in (-\infty, \infty)$ . Consider the limit of  $Z(t) = \sqrt{dt} S_{n=t/dt}$  when  $dt \rightarrow 0$ . Is the random curve so obtained continuous? Consider now random walks which attain the value  $S_N = vN$  and show that a continuous time limit can now be achieved when  $N = T/dt \rightarrow \infty$ , for the function  $\tilde{Z}(t) = dt S_{n=t/dt}$  when  $dt \rightarrow 0$  ( $t \leq T$ ). Draw a typical realisation of  $\tilde{Z}(t)$  for  $t \in [0, T]$ . What would this graph look like if instead  $X_i$  were i.i.d. Gaussian variables with mean zero and unit variance?
11. Consider the large deviations of sums of uniform random variables (i.e.  $p(x) = 1$  for  $0 \leq x \leq 1$  and  $p(x) = 0$  otherwise). Estimate the behavior of the Cramer function  $I(\bar{x})$  for *i*)  $\bar{x} \approx 1/2$ , *ii*)  $\bar{x} \approx 0$  and *iii*)  $\bar{x} \approx 1$ .
12. Let there be  $n + 1$  boxes labeled  $\omega = 0, 1, \dots, n$ , with  $n$  even. One of the boxes contains a prize, the others are empty. The probability that the prize is in  $\omega = 0$  is  $p$  whereas the probability that it is in any other box is  $(1 - p)/n$ . You have the options to open the box  $\omega = 0$  or to open simultaneously all boxes  $\omega > n/2$ . Show that the option that gives you more information on where the prize is not always the most convenient one. Show that if  $n > n^*(p)$  the second option is more informative than the first, and find  $n^*(p)$ . Show that for  $p = 1/(n + 1)$  the second option is always more informative. Show that, under one of two options, the uncertainty on where the prize is can increase, if  $n > \bar{n}(p)$ , and find  $\bar{n}(p)$ .
13. How many bits do you need to specify a Gaussian random variable with mean  $\mu$  and variance  $\sigma^2$  to a precision of  $n$  bits? Notice that the result depends on  $\sigma$  but not on  $\mu$ . Yet the binary representation of a Gaussian variable with  $\mu = 10^4$  up to precision  $\Delta$ , is very likely to contain more bits than a Gaussian variable with  $\mu = 10$ . Can you explain this apparent paradox?
14. Compute the differential entropy for a multi-dimensional Gaussian with mean  $\vec{\mu}$  and covariance  $\text{Cov}[X_i, X_j] = A_{i,j}$ .
15. Show that, in the case of Bernoulli trials, the number of successes is a sufficient statistics for the probability  $p$  of success. What is a sufficient statistics for the Poisson distribution?

16. Blackwell-Rao estimator: let  $\hat{X} = (X_1, \dots, X_n)$  be a sample of i.i.d. draws from a Poisson distribution with parameter  $\lambda$ . We want to estimate the probability  $e^{-\lambda}$  that  $X_{n+1} = 0$ . A very rough unbiased estimator is  $\delta = 1$  if  $X_1 = 0$  and  $\delta = 0$  otherwise. This is not a consistent estimator but it is unbiased. In order to get a better estimator, Blackwell-Rao theorem suggests to look for a sufficient statistics  $T$  for the unknown parameter and to consider the estimator  $\delta_{BR}(T) = \mathbb{E}[\delta|T]$ . In the present case, a sufficient statistics for  $\lambda$  is  $\sum_{i=1}^n X_i$ . Compute the Blackwell-Rao improved estimator  $\delta_{BR}$  and show that it is consistent (in fact, since  $T$  is complete and  $\delta$  is unbiased, Lehmann-Scheffé theorem implies that  $\delta_{BR}$  is the unique minimum variance unbiased estimator).
17. Among  $n$  objects at most one of them may be lighter or heavier. Given a balance find an upper bound to the number of weighings necessary for finding the lighter or heavier object, if it exists. For  $n = 12$  what would be the optimal first weighing?
18. Let  $p(s) = P\{S = s\}$  be the probability distribution of a discrete random variable  $S \in \mathcal{S}$  with  $|\mathcal{S}| < +\infty$ . Mixing this distribution with that of a deterministic variable taking value  $s_0 \in \mathcal{S}$  yields the distribution

$$p_\alpha(s) = (1 - \alpha)p(s) + \alpha\delta_{s,s_0}, \quad \alpha \in [0, 1]$$

Show that the entropy of the new variable  $S'$  described by this distribution is given by

$$H[S'] = (1 - \alpha)[H[S] - h(p(s_0))] + h(q),$$

$$h(x) = -x \log x - (1 - x) \log(1 - x)$$

where  $q = \alpha + (1 - \alpha)p(s_0)$  and  $H[S]$  is the entropy of the original variable  $S$  (measured in nats). Show that, for sufficiently small  $\alpha$ ,  $H[S']$  increases if  $p(s_0) < e^{-H[S]}$ .

19. Drawing with and without replacement. An urn contains  $r$  red,  $w$  white, and  $b$  black balls. Which has higher entropy, drawing  $k \geq 2$  balls from the urn with replacement or without replacement? Set it up and show why.
20. Let  $X, Y \geq 0$  be two random variables with joint distribution density  $p(x, y) = Ax^\theta e^{-x-xy}$ . Compute the normalization constant and find the interval of  $\theta$  where this is a well defined probability density. Compute the mutual information between  $X$  and  $Y$ .

21. Mutual information and copulas: given two random variables  $X, Y$  with joint pdf  $p(x, y)$  and distribution

$$P(x, y) = \int^x dx' \int^y dy' p(x' y')$$

The marginal distributions are

$$P_x(x) = \int^x dx' \int dy' p(x' y'), \quad P_y(y) = \int dx' \int^y dy' p(x' y').$$

the copula function is defined by the identity<sup>1</sup>

$$P(x, y) = C(P_x(x), P_y(y))$$

The idea is that the transformation  $(X, Y) \rightarrow (U = P_x(X), V = P_y(Y))$  maps the marginal densities to uniform ones so the distribution  $C(U, V)$  contains information on the statistical dependence of  $X$  and  $Y$  that is independent of the marginal distributions. Show that

$$I(X, Y) = D_{KL}(C, Q)$$

where  $Q(u, v) = uv$  is the uniform distribution in  $[0, 1]^2$ . Discuss the result. Generalize the result to  $n > 1$  random variables  $X_1, \dots, X_n$ .

22. Let  $\mathbf{X} = (X_1, \dots, X_N)$  be a vector of random variables whose marginal distributions  $p(x_i)$  are all Gaussian with zero mean and unit variances. Prove that

$$I(X_1, \dots, X_N) \geq -\frac{1}{2} \log \det \hat{C}$$

where  $\hat{C}$  is the covariance matrix with elements  $c_{i,j} = \mathbb{E}[X_i, X_j]$ . Notice that any random variable  $\tilde{X}_i$  can be transformed into a Gaussian variable  $X_i = \phi(\tilde{X}_i)$  by a suitable transformation. Then these bounds can be used to determine instances where variables have non-trivial statistical dependencies. For more information, see [51].

23. Let  $p_{\Sigma}(\mathbf{x})$  be the  $N$  dimensional multivariate Gaussian distribution with zero average, unit variance and correlation matrix  $\Sigma$ . Show that

$$D_{KL}(p_{\Sigma'} \| p_{\Sigma}) = \frac{1}{2} \left[ \log \frac{\det \Sigma}{\det \Sigma'} + \text{Tr}(\Sigma^{-1} \Sigma') - N \right]$$

<sup>1</sup>See R. B. Nelsen, An Introduction to Copulas, Lecture Notes in Statistics (New York: Springer, 1999) or some other standard text for an introduction to copulas.

24. Consider a random rectangular box in  $d$  dimensions. Each side  $X_i$  is an i.i.d. random variable with uniform distribution in  $[0, 1]$  ( $i = 1, \dots, d$ ). What is the expected value of the volume of the random box? What is the side  $\ell$  of the hypercube in  $d$  dimensions that has the same volume of the random box, in the limit  $d \rightarrow \infty$ ? How big can the radius  $r_d$  of an hyper-sphere that can be contained in the random box be, for  $d \gg 1$ ?
25. Optimal binning: imagine you have a sample  $\underline{X} = (X_1, \dots, X_N)$  of  $N$  observations of a random variable  $X$  with unknown pdf  $p(x)$  with support on  $[0, 1]$ . In order to estimate  $p(x)$ , you divide the interval in  $m$  bins of size  $1/m$ . What is the optimal number of bins  $m$  if you want to minimise the relative uncertainty on the point  $(x, p(x))$  of the graph of the pdf? (*hint*: having  $m$  small gives a lot of precision on the estimate of  $p$  but a poor resolution on  $x$ , large  $m$  gives high resolution of  $x$  but large errors in  $p$ .)
26. Let  $X_i$  be the  $x$  coordinate of the  $i^{\text{th}}$  particle of an ideal gas at temperature  $T$  in a cubic box of size one. Now imagine to set a wall at position  $\ell \in [0, 1]$  perpendicular to the  $x$  direction and let  $Y$  be the number of particles to the left of it. Depending on  $Y$  the wall will experience unequal pressures from the left and from the right. If  $Y$  is known the force that results from this difference in pressure can be used in order to perform work.<sup>2</sup>

Show that the expected value of the work done by the system in the isothermal expansion of the gas<sup>3</sup> is equal to

$$\frac{\mathbb{E}[W]}{K_B T} = NI(X_1, Y).$$

Show that, in this case

$$\sum_{i=1}^N I(X_i, Y) \leq I(\underline{X}, Y) = E \left[ \log \frac{p(\underline{X}|Y)}{p(\underline{X})} \right]$$

so that  $\mathbb{E}[W] \leq K_B T I(\underline{X}, Y)$ , which generalises the second law of thermodynamics to cases where information on the microscopic state of a system is available.

<sup>2</sup>This relation between information and work in thermodynamics has been first epitomised in Maxwell's demon, a creature that observing the velocities of particles in a gas can open and close a small gate and create free energy differences that can be used to perform work.

<sup>3</sup>The work differential is given by  $dW = PdV$  where  $P = nK_B T/V$  is the pressure of an ideal gas and  $V$  is its volume.



When  $N = 1$  the work extracted is precisely equal to  $K_B T$  times the information that the measurement  $Y$  gives on the microscopic state  $\underline{X}$  of the gas. When  $N > 1$  not all the information  $I(\underline{X}, Y)$  extracted from the measurement can be used to perform work.

27. Consider two random variables  $X, Y = 0, \pm 1$ , with  $\mu(X = \pm 1) = \mu(Y = \pm 1) = \mu$ ,  $\mu(X = 0) = \mu(Y = 0) = 1 - 2\mu$  and  $E[XY] = 0$ . Find the distribution  $p(X, Y)$  with a given mutual information  $I(X, Y) = I_0$ . How does the distribution with maximal  $I(X, Y)$  looks like?
28. *Chernoff bound*: in hypothesis testing, let  $X \in \{0, 1\}$  be a random variable that has distribution  $P\{X = 1|H_1\} = p$  and  $P\{X = 0|H_1\} = 1 - p$  under hypothesis  $H_1$  and  $P\{X = 1|H_2\} = q$  and  $P\{X = 0|H_2\} = 1 - q$  under hypothesis  $H_2$ . Compute  $\lambda^*(p, q)$  and check that this satisfies the symmetry  $\lambda^*(p, q) = 1 - \lambda^*(q, p)$ . Why is this so?
29. Imagine that vaccines are being developed to contrast an ongoing pandemics. Let's assume that a vaccine is efficient if a subsequent blood test reveals the presence of antibodies with high probability  $p$ . If instead the vaccine is not efficient, antibodies are detected with a baseline probability  $q$ . Can you help the health authorities to decide how many people should be tested in the trial phase, in order to conclude that the vaccine is efficient with high confidence?
30. *Bias in the MLE estimate of the entropy*. Let  $X_1, \dots, X_n$  be  $n$  i.i.d. draws from a distribution  $p_x$  for  $x, X_i \in \chi$  in a finite set. The Maximum Likelihood Estimate (MLE) of  $p_x$  is  $\hat{p}_x = k_x/n$  where  $k_x$  is the number of  $X_i = x$ . Show that the MLE estimate of the entropy

$$\hat{H} = - \sum_{x \in \chi} \hat{p}_x \log \hat{p}_x$$

is a biased estimator of the entropy

$$H[p] = - \sum_{x \in \chi} p_x \log p_x$$

in the sense that

$$\begin{aligned} \mathbb{E}[\hat{H}] - H[p] &\simeq -\frac{|\chi| - 1}{2n} + \frac{1}{12n^2} \left[ 1 + \sum_{x \in \chi} \frac{5}{p_x} \right] \\ &+ \frac{1}{12n^3} \left[ \sum_{x \in \chi} \frac{9}{p_x^2} - \sum_{x \in \chi} \frac{5}{p_x} \right] + O(n^{-4}) \end{aligned}$$

*Hint:* use repeatedly the formula

$$\log k = \int_0^\infty \frac{du}{u} [e^{-u} - e^{-ku}].$$

31. Differential entropy estimates: let  $x_1, \dots, x_N$  be a sample of  $N$  i.i.d. draws from a pdf  $p(x)$ . Show that an estimate of the differential entropy can be given by

$$h[p] = - \int dx p(x) \log p(x) \approx \frac{1}{N-1} \sum_{i=1}^{N-1} \log(x_{i+1} - x_i) + \log N.$$

32. Compute the differential entropy of the random variable  $X \geq e$  with pdf  $p(x) = 1/[x(\ln x)^2]$  for  $x \geq e$  and  $p(x) = 0$  for  $x < e$ .
33. *Coarse graining.* Let  $p(x)$  and  $q(x)$  be two distributions for the discrete random variable  $X \in \chi$ , where  $|\chi| < +\infty$  takes a finite number of values. Let  $Z = f(X)$  be a random variable that takes values on a set  $\mathcal{Z}$ , with  $|\mathcal{Z}| \leq |\chi|$ . The transformation  $f : \chi \rightarrow \mathcal{Z}$  generates a representation of  $X$  that eliminates some details, because different values of  $X$  can be mapped into the same value of  $Z$ . In this sense it is a coarse graining. Let  $\tilde{p}(z) = \sum_{x:f(x)=z} p(x)$  and  $\tilde{q}(z) = \sum_{x:f(x)=z} q(x)$  be the distributions of  $Z$ . Show that  $D_{KL}(\tilde{p} \parallel \tilde{q}) \leq D_{KL}(p \parallel q)$ . In loose words, the two distributions  $p$  and  $q$  approach each other under coarse graining.
34. Compute the Cramer function for a sequence  $\underline{X}$  of i.i.d. Gaussian random variables with unit variance and mean which, for all of them, is either  $\mathbb{E}[X] = \mu$  with some probability  $\nu$ , or  $-\mu$  with probability  $1 - \nu$ . Compute the function

$$\bar{\phi}(h) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E} [e^{h(X_1 + \dots + X_n)}]$$

and its Legendre transform  $\bar{I}(\bar{x})$ . What is the distribution density  $p(x|\bar{x})$  of the variables, conditional to the value of  $\bar{x}$ ?

35. Use the identity

$$e^{\frac{\beta J}{2n} M^2} = \sqrt{\frac{\beta J n}{2\pi}} \int_{-\infty}^{\infty} dm e^{-\frac{\beta J n}{2} m^2 + \beta J m M}$$

with  $M = \sum_i X_i$  and compute the partition function

$$Z(\beta) = \sum_{\underline{X}} e^{-\beta E(\underline{X})}$$

for the mean field Ising model. Recover the equation of state Eq. (19.25).

36. Compute Jeffrey's prior for the Poisson and the Gaussian distribution.
37. Compute Jeffrey's prior for the binary distribution  $P(X = 1) = p = 1 - P(X = 0)$ . Show that the unconditional distribution of the sufficient statistics  $T(\underline{X}) = X_1 + \dots + X_n$  for a sample  $\underline{X}$  of  $n$  i.i.d. draws, under Jeffrey's prior, obeys an arc-sine law for  $n \gg 1$ . What is the distribution of  $T$  instead under Laplace's prior  $p_0(p) = 1$  for  $p \in [0, 1]$ ?
38. Let  $\epsilon \in \mathbb{R}$  be a random variable with distribution

$$f(x|\theta) = \frac{1}{\sigma} e^{g\left(\frac{x-\mu}{\sigma}\right)} \quad (21.1)$$

where  $\theta = (\mu, \sigma)$  and  $e^{g(z)}$  is a pdf such that

$$\int_{-\infty}^{\infty} dz e^{g(z)} z = 0, \quad \int_{-\infty}^{\infty} dz e^{g(z)} z^2 = 1.$$

Check that the expected values of the scores vanish. Show that the Fisher Information is given by

$$\begin{aligned} J_{\mu, \mu} &= \frac{1}{\sigma^2} \mathbb{E} \left[ (g'(Z))^2 \right], & J_{\mu, \sigma} &= \frac{1}{\sigma^2} \mathbb{E} \left[ Z (g'(Z))^2 \right], \\ J_{\sigma, \sigma} &= \frac{1}{\sigma^2} \left\{ \mathbb{E} \left[ (Z g'(Z))^2 \right] - 1 \right\} \end{aligned}$$

and therefore Jeffrey's prior is proportional to  $p_J(\mu, \sigma) \propto 1/\sigma^2$ . Show also that, if  $g(z|\vartheta)$  depends on other parameters  $\vartheta \in \mathbb{R}^d$ , then  $J_{\mu, \vartheta_a} \propto \sigma^{-1}$  and  $J_{\vartheta_a, \vartheta_b}$  is independent of  $\mu$  and  $\sigma$ . Conclude from this that Jeffrey's prior is independent of  $\mu$  and proportional to  $\sigma^{-2}$  for all distributions of the form (21.1). Discuss the result in terms of dimensional analysis.

39. Let  $p(\mu, \sigma|\underline{X})$  be the posterior distributions of the parameters of a Gaussian

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

given a sample  $\underline{X} = (X_1, \dots, X_n)$  of  $n$  i.i.d. observations.

Show that assuming the uninformative prior  $p_0(\mu, \sigma) = c/\sigma$ , the posterior  $p(\mu, \sigma | \underline{X})$  is also improper for  $n = 1$ .

Show that for  $n \geq 2$  provided  $X_1 \neq X_2$ , the posterior is instead a proper probability density.

Show that the same is true under Jeffrey's prior.<sup>4</sup>

40. *Model selection in the limit of uninformative priors.*

Let  $\underline{X} = (X_1, \dots, X_n)$  be a sample of  $n$  points that we know have a Gaussian distribution. In order to decide whether the correct distribution has mean  $\mathbb{E}[X] = 0$  and variance  $\sigma^2$  or a mean  $\mathbb{E}[X] = \mu$  and variance  $\sigma^2$ , we can assume that  $\mu$  has a prior distribution

$$p_0(\mu) = \sqrt{\frac{\epsilon}{2\pi}} e^{-\frac{\epsilon}{2}\mu^2}$$

and that  $\sigma$  has a prior  $p_0(\sigma)$  in both models.

Show that in the limit  $\epsilon \rightarrow 0$  of total *a priori* ignorance about  $\mu$ , the model with  $\mathbb{E}[X] = 0$  will always be selected, in a Bayesian model selection scheme. Interpret the result.

*Hint:* it is possible to show that, when  $p_0(\sigma) = c/\sigma$ , to leading order in  $\epsilon$  the model with  $\mu \neq 0$  is more likely than that with  $\mu = 0$  when

$$\epsilon > \frac{n}{x^2} \left( 1 - \frac{\bar{x}^2}{x^2} \right)^{\frac{n}{2}}$$

where  $\bar{x} = \frac{1}{n} \sum_i X_i$  and  $\overline{x^2} = \frac{1}{n} \sum_i X_i^2$ . In loose words, only if the initial uncertainty on  $\mu$  is finite it is possible to conclude that  $\mu \neq 0$ .

41. What is the probability that democracy works in a random population?

Consider a population of  $N$  individuals with preferences over three choices,  $A, B$  and  $C$ . Let the preference ranking over the alternatives

<sup>4</sup>The intuition about this result is the following. A state of complete ignorance about  $\mu$  is one where an infinite number of bits would be needed to specify  $\mu$  to any precision  $\Delta$ , because  $|\mu|$  can be arbitrarily large. Similarly, an infinite number of bits would be needed to specify  $\sigma$  to any precision  $\Delta$ . When we see two data point, we can estimate the scale of both  $\mu \approx (X_1 + X_2)/2$  and  $\sigma \approx X_1 - X_2$ . This removes the divergencies and yields a state of knowledge that is a finite number of bits away from the knowledge of  $\mu$  and  $\sigma$  to a finite precision. One may conjecture that those associated to scale and location are the only *primitive* divergencies in the state of knowledge about a real random variable  $x$ . Then two points should be sufficient to make the posterior derived from Jeffrey's prior finite, whatever is the distribution  $f(x|\theta)$ , no matter how many parameters it depends on.

be random and independent for each individual. Consider pairwise majority voting among the alternatives, e.g. if the number of individuals which prefer  $A$  to  $B$  is larger than  $N/2$  that the majority prefers  $A$  to  $B$ . In the limit  $N \rightarrow \infty$ , find the probability that pairwise majority voting is transitive, i.e. that if the majority prefers  $A$  to  $B$  and  $B$  to  $C$ , then the majority also prefers  $A$  to  $C$ .

42. Let  $\mathbf{s} = (s_1, \dots, s_n)$  be a string of  $n$  bits ( $s_i = 0, 1$ ). Consider the distribution

$$p(\mathbf{s}) = \frac{1}{Z} e^{-gE(\mathbf{s})}, \quad E(\mathbf{s}) = \max\{i : s_i = 1\}$$

In words, the function  $E$  returns the largest index  $k$  such that  $s_k = 1$  (and  $E(\mathbf{s}) = 0$  if  $s_i = 0$  for all  $i$ ). Notice that any  $\mathbf{s}$  with  $s_k = 1$  and  $s_\ell = 0$  for all  $\ell > k$  has probability  $p(\mathbf{s}) = e^{-gk}/Z$ , where  $Z$  is a normalisation constant. Compute the partition function  $Z$ , the expected value of  $E$  and its variance. Show that in the limit  $n \rightarrow \infty$  this model features a phase transition.

43. Let  $\hat{x} = (x_1, \dots, x_n)$  be a sample of variables drawn from an exponential distribution  $p(x) = \theta e^{-\theta x}$ . How big do you expect  $n$  should be in order to know the parameter  $\theta > 0$  to a precision  $\Delta$ ? *i)* give a heuristic argument for an estimate of the asymptotic behavior of  $n$  with  $\Delta$  for  $\Delta \ll 1$ , *ii)* describe the calculation that you would do to prove this result *iii)* do the calculation. [There are different ways to find the solution. If yours needs a prior on  $\theta$ , use  $p_0(\theta) = \alpha e^{-\alpha\theta}$ ]
44. The Z-channel: let  $X, Y \in \{0, 1\}$  be two binary random variables. Let  $P\{Y = 0|X = 0\} = 1$  and  $P\{Y = 1|X = 1\} = 1 - \epsilon$ . This is called the Z-channel in channel coding. You can think of  $X$  as being the input of a noisy communication channel that gives  $Y$  as output. Hence when the input is  $X = 0$  it is transmitted without error, whereas when  $X = 1$  the input may be corrupted by noise. Show that when the distribution of  $X$  is such that  $I(X, Y)$  is maximal,  $X = 1$  should be more probable than  $X = 0$ .



# Bibliography

- [1] E. Wigner, *The unreasonable effectiveness of mathematics in the natural sciences*, *Commun. Pure Appl. Math.* **13** (1960) 1.
- [2] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, U.K. (2003).
- [3] W. Feller, *An introduction to probability theory and its applications. Vol. I*, Third edition, John Wiley & Sons Inc., New York (1968).
- [4] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley (2005) [DOI: [10.1002/047174882x](https://doi.org/10.1002/047174882x)].
- [5] B.V. Gnedenko, *Theory of probability*, Mir Publishers, Moskow, Russia (1969).
- [6] E. Marinari and G. Parisi, *Trattatello di probabilità - teoria delle probabilità per fisici, computational scientists and computer scientists*, <http://www.formulas.it/formulog/wp-content/uploads/2013/01/trattatello.pdf>.
- [7] D. Kahneman and A. Tversky, *Prospect theory: An analysis of decision under risk*, in *Handbook of the fundamentals of financial decision making: Part I*, World Scientific (2013), pp. 99–127.
- [8] E. Jaynes, *Prior Probabilities*, *IEEE Trans. Syst. Sci. Cybernetics* **4** (1968) 227.
- [9] B. De Finetti, *Sul significato soggettivo della probabilita*, *Fund. Math.* **17** (1931) 298, <https://eudml.org/doc/212523>.
- [10] S. Vosoughi, D. Roy and S. Aral, *The spread of true and false news online*, *Science* **359** (2018) 1146.
- [11] *The book Numerical Recipes*, 3rd ed., <http://numerical.recipes>.

- [12] P. Flajolet and A. Odlyzko, *Singularity Analysis of Generating Functions*, *SIAM J. Discrete Math.* **3** (1990) 216.
- [13] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, U.K. (2009).
- [14] G.E. Andrews and K. Eriksson, *Integer partitions*, Cambridge University Press, Cambridge, U.K. (2004).
- [15] W. Feller, *On the kolmogorov–smirnov limit theorems for empirical distributions*, in *Selected Papers I*, Springer (2015), pp. 735–749.
- [16] M.E.J. Newman, *The Structure and Function of Complex Networks*, *SIAM Rev.* **45** (2003) 167.
- [17] T. Baravi, O. Feinerman and O. Raz, *Echo chambers in the Ising model and implications on the mean magnetization*, *J. Stat. Mech.* **2022** (2022) 043402.
- [18] J.M. Honig, *Critical Phenomena*, in J.M. Honig ed., *Thermodynamics*, 4th ed., Elsevier (2014), p. 375–405 [DOI: 10.1016/b978-0-12-416705-6.00007-5].
- [19] M. Tribus and E.C. McIrvine, *Energy and information*, *Sci. Am.* **225** (1971) 179, <https://www.jstor.org/stable/e24923110>.
- [20] C.W. Gardiner, *Handbook of stochastic methods for physics, chemistry and the natural sciences*, vol. 13 of *Springer Series in Synergetics*, third ed., Springer-Verlag, Berlin (2004).
- [21] P. Erdos and A. Rényi, *On cantor series with convergent  $\sum 1/q_n$* , *Ann. Univ. Sci. Budapest. Eötvös. Sect. Math* **2** (1959) 93.
- [22] X. Gabaix, *The granular origins of aggregate fluctuations*, *Econometrica* **79** (2011) 733.
- [23] J. Bouchaud, *Levy flights and related topics in physics*, *Lect. Notes Phys.* **450** (1995).
- [24] L.M. Brown, *Renormalization: From Lorentz to Landau (and beyond)*, Springer-Verlag (1993).
- [25] B. Derrida, *Random-energy model: An exactly solvable model of disordered systems*, *Phys. Rev. B* **24** (1981) 2613.



- [26] M. Marsili, *The peculiar statistical mechanics of Optimal Learning Machines*, [arXiv: 1904.09144](#) [DOI: [10.1088/1742-5468/ab3aed](#)].
- [27] W. Bialek, *Biophysics: searching for principles*, Princeton University Press (2012).
- [28] R.W. Yeung, *Information theory and network coding*, Springer Science & Business Media (2008).
- [29] R.S. Ellis, *Entropy, large deviations, and statistical mechanics*, Springer Verlag, New York, Berlin (1985).
- [30] M. Mezard and A. Montanari, *Information, physics, and computation*, Oxford University Press (2009).
- [31] R.K.P. Zia, E.F. Redish and S.R. McKay, *Making sense of the Legendre transform*, *Am. J. Phys.* **77** (2009) 614.
- [32] E.T. Jaynes, *Information Theory and Statistical Mechanics*, *Phys. Rev.* **106** (1957) 620.
- [33] Y. Tikhochinsky, N.Z. Tishby and R.D. Levine, *Alternative approach to maximum-entropy inference*, *Phys. Rev. A* **30** (1984) 2638.
- [34] F. Morcos et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families*, *Proc. Nat. Acad. Sci.* **108** (2011) <bbl:err:pages>.
- [35] R.J. Cubero et al., *Statistical Criticality arises in Most Informative Representations*, *J. Stat. Mech.* **1906** (2019) 063402 [[arXiv: 1808.00249](#)].
- [36] M. Marsili and Y. Roudi, *Quantifying Relevance in Learning and Inference*, [arXiv: 2202.00339](#) [DOI: [10.1016/j.physrep.2022.03.001](#)].
- [37] L.D. Landau, E.M. Lifšic, E.M. Lifshitz and L. Pitaevskii, *Statistical physics: theory of the condensed state*, vol. 9, Butterworth-Heinemann (1980).
- [38] M. Kardar, *Statistical physics of particles*, Cambridge University Press, Cambridge, U.K. (2007).
- [39] J. Song et al., *Optimal work extraction and mutual information in a generalized Szilárd engine*, *Phys. Rev. E* **103** (2021) 052121.
- [40] C.V. den Broeck and M. Esposito, *Second law and Landauer principle far from equilibrium*, *EPL* **95** (2011) 40004.

- [41] P. Bialas, Z. Burda and D. Johnston, *Phase diagram of the mean field model of simplicial gravity*, *Nucl. Phys. B* **542** (1999) 413 [[arXiv: gr-qc/9808011](#)].
- [42] M.R. Evans, S.N. Majumdar and R.K.P. Zia, *Factorized steady states in mass transport models on an arbitrary graph*, *J. Phys. A* **39** (2006) 4859.
- [43] P. Hack, S. Gottwald and D.A. Braun, *Jarzynski's Equality and Crooks' Fluctuation Theorem for General Markov Chains with Application to Decision-Making Systems*, *Entropy* **24** (2022) 1731.
- [44] G.E.P. Box, *Science and Statistics*, *J. Am. Statist. Assoc.* **71** (1976) 791.
- [45] H. Akaike, *A new look at the statistical model identification*, *IEEE Trans. Automatic Control* **19** (1974) 716.
- [46] B.B. Machta, R. Chachra, M.K. Transtrum and J.P. Sethna, *Parameter Space Compression Underlies Emergent Theories and Predictive Models*, [arXiv: 1303.6738](#) [DOI: [10.1126/science.1238723](#)].
- [47] I. Mastromatteo and M. Marsili, *On the criticality of inferred models*, *J. Stat. Mech.* **1110** (2011) P10012 [[arXiv: 1102.1624](#)].
- [48] I.J. Myung, V. Balasubramanian and M.A. Pitt, *Counting probability distributions: Differential geometry and model selection*, *Proc. Nat. Acad. Sci.* **97** (2000) 11170.
- [49] C. de Mulatier and M. Marsili, *Bayesian Inference of Minimally Complex Models with Interactions of Arbitrary Order*, [arXiv: 2008.00520](#).
- [50] M. Marsili, I. Mastromatteo and Y. Roudi, *On sampling and modeling complex systems*, *J. Stat. Mech.* **1309** (2013) P09003 [[arXiv: 1301.3622](#)].
- [51] D.V. Foster and P. Grassberger, *Lower bounds on mutual information*, *Phys. Rev. E* **83** (2011) 010101.
- [52] F. Chollet, *On the Measure of Intelligence*, [arXiv: 1911.01547](#).
- [53] N. Chomsky, *Aspects of the Theory of Syntax*, The MIT Press, Cambridge, U.S.A. (1965), <https://www.jstor.org/stable/j.ctt17kk81z>.
- [54] M. Tomasello, *Constructing a language: A usage-based theory of language acquisition*, Harvard University Press (2005), <https://www.hup.harvard.edu/books/9780674017641>.

- [55] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman and Company, San Francisco (1982).
- [56] M. Baldovin, G. Gradenigo, A. Vulpiani and N. Zanghì, *On the foundations of statistical mechanics*, *Phys. Rept.* **1132** (2025) 1.
- [57] A. Haimovici and M. Marsili, A. Haimovici and M. Marsili, *Criticality of mostly informative samples: A Bayesian model selection approach*, *J. Stat. Mech.* **1510** (2015) P10013.
- [58] R.J. Cubero, M. Marsili and Y. Roudi, *Minimum Description Length Codes Are Critical*, *Entropy* **20** (2018) 755.



# Index

Bertrand paradox, [16](#)

Chance, [11](#)

Classical probability, [31](#)

Combinations, [32](#)

Combinatorics, [32](#)

de Finetti, Bruno, [19](#)

Jaynes, Edwin T., [xiii](#), [17](#)

Kolmogorov's axioms, [5](#)

Ordered samples, [32](#)

Permutations, [32](#)

Plausible reasoning, [20](#)

Probability, [11](#)

Product rule, [24](#)

Randomness, [11](#)

Stirling's approximation, [34](#)

Sum rule, [26](#)





# Lectures on probability, information and large scale behaviour

*Matteo Marsili*

This book stems from lecture notes prepared by the author over decades of teaching. It is divided into two parts. The first part focuses on classical probability, with the aim of helping students develop a solid intuition about random phenomena. It guides them in translating real-world problems into probabilistic language and equips them with the tools needed to derive quantitative insights. The second part investigates the typical behaviours that emerge in asymptotic regimes - such as the law of large numbers, limit theorems, and large deviations - and elucidates their connection to information theory. This latter part offers a unifying perspective, based on principles of information theory and statistical mechanics, for understanding large-scale phenomena like phase transitions that appear across disciplines such as statistical physics, inference, coding theory, and computer science.