Lectures on Discrete Probabilistic Models

Michele Pavon
Dipartimento di Matematica
Università di Padova
via Trieste 63, 35121 Padova
Italy

January 15, 2015

Man can believe the impossible, but man can never believe the improbable.
Oscar Wilde

The human understanding is of its own nature prone to suppose the existence of more order and regularity in the world than it finds.
Sir Francis Bacon

De dolis in huius modi ludis.
Bis luscus[1]

[1]This cryptic inscription was recently found in "loco Arculi" near Milan. It suggests that there might have been scientific work on cheating in gambling that antedates the calculations of Pacioli, Tartaglia and Cardano.

Preface

In the Spring quarter 2007, 2008 and 2009, I had the opportunity to teach a honor, introductory probability course to students of the Scuola Galileiana di Studi Superiori of the University of Padova. These are extremely bright, highly motivated students pursuing a Bachelor Degree in one of the Natural or Engineering Sciences. The students take this thirty-hour course during the second quarter of their first college year, with a math background that includes little more than univariate calculus.

I was therefore confronted with the non trivial task of designing an interesting program that would employ only elementary tools. While preparing my lectures, my main concern was to stimulate the students and to give them a glimpse of *some* profound concepts and recent exciting applications. In order to get the student interested, I tried to spice up my lectures with historical comments and curiosities including extensive evidence of how poor our probabilistic intuition is. I also tried to design a rather original route that would allow me to get *rapidly* to some fundamental principles of statistical mechanics (statics for thermodynamic systems is presented in Chapter 2 after only 12 pages of basic probability). Some connections between statistical physics and the rapidly growing field of Markov Chain Monte Carlo (MCMC) and its generalizations such as the Metropolis algorithm, the Feynman-Kac formula, etc. are then mentioned. Other applications having profound connections with statistical mechanical models and concepts such as the average consensus problem and Google PageRank algorithm are also outlined at the end of the book.

This short book should in no way be considered a complete introduction to probability for which several excellent monographs are available. I mention the evergreen textbook by W.Feller : *An Introduction to Probability Theory and its Applications, Vol.I.*, Second Edition, Wiley, 1957, A. N. Shiryayev, *Probability*, Springer-Verlag, New York, 1984, and the more recent: C. Grin-

stead and J. Snell, *Introduction to Probability*, 2nd edition, AMS, 2006, freely available at http://math.dartmouth.edu/ prob/prob/prob.pdf. The latter books are somewhat ideal for a one to two-semester comprehensive course. The present book, however, might be a suitable textbook for a short probability course for physicists, biologists and other natural and engineering sciences students who are eager to see the modeling and computational power of probability at work in their own fields. It may be also be suitable as secondary reading material for mathematics students with interest in applications. Prerequisites for the course are univariate calculus, and some basic elements of linear algebra and of thermodynamics.

In writing the book, I drew inspiration and examples from the existing literature. Besides the three textbooks mentioned above, the Markov chain part has profited from the excellent book by P. Brémaud: *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer-Verlag, New York, 1999. Part of the statistical mechanics lectures have been inspired by some beautiful, unpublished notes by Francesco Guerra. Other topics are treated here in an original fashion, which I deemed important for students who will most likely eventually pursue research.

I wish to thank my colleagues Paolo Dai Pra, Marco Favretti, Lorenzo Finesso, Stefano Pinzoni and Francesco Ticozzi who read and criticized various portions of the book.

Venice, March 2009 M. P.

Contents

Questionnaire

The following (anonymous) questionnaire was administered to the students on the first day of class. The students were told it was meant to evaluate their probabilistic intuition. They were allotted twenty minutes to complete the test. They were also asked to mark problems they had already been exposed to. Problem 3 was included to show that, although our probabilistic intuition is in many ways rather poor, our (Euclidean) geometric intuition is often quite correct. All of the probabilistic problems were then solved in class after the appropriate background had been developed. They are described in Chapters III and VI.

1. Assume equal probability for the two sexes at each birth.

   (a) A mother has two children. The younger one is a daughter named Giulia. What is the probability that the other child is a girl?

   (b) A mother has two children. The older one is a daughter named Giulia. What is the probability that the other child is a girl?

   (c) A mother has two children. One of them is a daughter. What is the probability that the other child is a girl?

   (d) A mother has two children. One of them is a daughter named Giulia. What is the probability that the other child is a girl?

2. Consider a group of N persons. Disregarding leap years, we assume that the probability of being born in each day of the year is $1/365$. Estimate the smallest N such that the probability of at least two people having the same birthday is greater than $1/2$.

3. - Among all rectangles with a given perimeter, which one has maximal area?

- Among all planar figures with a given perimeter, which one has maximal area?

4. In an urn, there are three coins whose faces are heads-tails, heads-heads and tails-tails. I draw a coin at random and toss it: It results in heads. What is the probability that the other face of the coin is heads?

5. I show you three closed boxes. Only one contains a golden coin and I know which one it is. You choose a box. After that, I open one of the other two boxes that I know is empty. At this point, you are allowed to confirm you first choice or to change it choosing the other closed box.

   (a) you should change your first choice:

   (b) you should confirm your first choice;

   (c) it makes no difference if you confirm or change your choice.

6. The attendants in a theatre cloakroom have lost all the labels for the coats of n men. They therefore decide to return them at random. Let p_n be the probability that at least one of the coats is returned to the owner. As n grows,

   (a) p_n increases;

   (b) p_n decreases;

   (c) p_n does not increase or decrease but it has a limit;

   (d) p_n does not increase or decrease nor does it have a limit.

7. A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green?

Chapter 1

Finite probabilistic models

1.1 Experiments with a finite number of outcomes

We consider first experiments that only have finitely many possible outcomes.

Example 1.1.1

1. Coin tossing has the two possible outcomes H (heads) and T(tails).
2. Rolling a die has the six possible outcomes $1, 2, 3, 4, 5, 6$.

Given an experiment, the set of all possible distinct outcomes

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$$

is called the *sample space*. For coin tossing, $\Omega = \{H, T\}$. For rolling a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. For n coin tosses,

$$\Omega = \{\omega = (a_1, a_2, \ldots, a_n), a_i = H \text{ or } T\}.$$

For a sample space with N elements, let us consider a function

$$p : \Omega \to [0, 1],$$

such that

$$\sum_{i=1}^{N} p(\omega_i) = 1. \tag{1.1}$$

We call p a *probability distribution* and $p(\omega_i)$ the *probability* of outcome ω_i. A subset A of Ω, $A \in \mathcal{P}(\Omega)$[1], is called an *event*. For instance, when rolling a

[1] $\mathcal{P}(\Omega)$ denotes the collection of all subsets of Ω.

die, $A = \{2, 4, 6\}$ is the event "the outcome is even". We define

$$\mathbb{P}(A) := \sum_{\{\omega_i \in A\}} p(\omega_i), \quad \mathbb{P}(\emptyset) = 0. \tag{1.2}$$

Such a map $\mathbb{P}$ is called a *probability measure* on $\mathcal{P}(\Omega)$. It enjoys the following properties.

1. $\mathbb{P}(\Omega) = 1$;
2. Let $A_1, A_2, \ldots, A_m$ be pairwise disjoint subsets of Ω. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} \mathbb{P}(A_i).$$

These two properties imply

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
2. $A \subseteq B \Rightarrow \mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A)$ (hence $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$);
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ (hence $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$).

Exercise 1.1.2 Establish the following relation:

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

Consider $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$. Then the *uniform distribution* is given by $p_u(\omega) = \frac{1}{N}$. It follows that, for any $A \subseteq \Omega$, $\mathbb{P}(A) = \frac{|A|}{N}$. Here $|A|$ denotes the cardinality of A.

Example 1.1.3

- for a fair coin, $p_u(H) = p_u(T) = \frac{1}{2}$;
- for a fair die, $p_u(i) = \frac{1}{6}$, $i = 1, 2, \ldots, 6$;
- for n tosses of a fair coin, there are 2^n distinct outcomes. Hence

$$p_u((a_1, a_2, \ldots, a_n)) = 2^{-n}.$$

A lot of the early work on probability (Cardano) concerned dice rolling motivated by gambling. A short essay on three dice rolling entitled "Sulla scoperta dei dadi" was written around 1620 by Galileo Galilei. Gambling also motivated some of the pioneers (Pacioli and Fontana (nickname Tartaglia)) to study the more challenging *problem of points*. It is one of the problems posed in 1654 by Chevalier De Méré to Pascal which is solved in the famous Pascal-Fermat correspondence: Two players A and B play a series of games where each one has probability $1/2$ of winning (say, they toss a fair coin). Before they complete the series of games they had agreed upon, they have to stop. How should the stakes be divided? Here is Fermat's reasoning. Suppose the two players had agreed that whoever won five games would take all the money. Suppose the series of games is interrupted after seven games of which four have been won by A and three by B. If they had played two more games, one would have won for sure. Of the four equally likely possibilities, three lead to A winning and only one to B winning. Hence, the stakes should be divided according to the ratio $3:1$.

1.2 Random variables

Let us consider a discrete *probability space* $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. From now on, unless otherwise explicitly stated, $|\Omega| = N$. A function

$$X : \Omega \to \mathbb{R}$$

is called a *random variable*.

Example 1.2.1 Consider the sample space $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$, corresponding to two tosses of a coin, with the uniform distribution p_u. We define the random variable

$$X(\omega) = \#\,\text{of heads in}\ \omega = (a_1, a_2).$$

Then X takes values in the set $\mathcal{X} = \{0,1,2\}$.

Example 1.2.2 Let $\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$ be the sample space corresponding to rolling twice a fair die. We can define a random variable by

$$X(\omega) = \text{sum of dots on the two faces.}$$

Then X takes values in $\mathcal{X} = \{2,3,\ldots,12\}$.

Example 1.2.3 Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a probability space. The *indicator function* of $A \subseteq \Omega$ is the random variable

$$\mathbb{1}_A(\omega) := \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A \end{cases}$$

Here $\mathcal{X} = \{0, 1\}$.

Given a random variable X taking values in $\mathcal{X} = \{x_1, \ldots, x_M\}$, we define the *level sets*

$$A_i := \{\omega : X(\omega) = x_i\}, \quad i = 1, \ldots, M.$$

Observation 1.2.4 The collection $A_1, \ldots, A_M$ yields a *partition* of Ω in that these sets are pairwise disjoint and

$$\bigcup_{i=1}^{M} A_i = \Omega.$$

The random variable X admits the "canonical" representation

$$X(\omega) = \sum_{i=1}^{M} x_i \mathbb{1}_{A_i}(\omega). \tag{1.3}$$

Notice that $|\Omega| < \infty \Rightarrow M = |\mathcal{X}| < \infty$. For each random variable X, it is possible to define a probability p_X on $\mathcal{X}$ by

$$p_X(x_i) := \mathbb{P}\{\omega : X(\omega) = x_i\} = \mathbb{P}(A_i), \quad x_i \in \mathcal{X}.$$

$p_X = (p_X(x_1), \ldots, p_X(x_M))$ is called the *probability distribution* of the random variable X. The corresponding probability measure on $\mathcal{P}(\mathcal{X})$ is called the probability law of X (also "push-forward" measure of $\mathbb{P}$ under X). We assume from here on that $x_i \in \mathcal{X} \Rightarrow p_X(x_i) > 0$.

1.3 Expected value

Let X be a random variable on the probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. The *expected value* (also called *expectation* or *mean*) of X is defined by

$$\mathbb{E}X := \sum_{i=1}^{N} X(\omega_i) p(\omega_i) = \sum_{i=1}^{M} x_i \mathbb{P}(A_i). \tag{1.4}$$

It is immediate that the expectation may also be computed as

$$\mathbb{E}X = \sum_{i=1}^{M} x_i p_X(x_i). \tag{1.5}$$

When more than one probability distribution (measure) is defined on Ω ($\mathcal{P}(\Omega)$) and we want to emphasize the dependence of the expected value on p ($\mathbb{P}$), we write $\mathbb{E}_p X$ ($\mathbb{E}_{\mathbb{P}} X$).

Example 1.3.1 Let $X \equiv c$ be constant. Then $EX = c$.

Example 1.3.2 Consider again Example 1.2.1 with $\mathcal{X} = \{0, 1, 2\}$. Then the sets A_i, $i = 1, 2, 3$, are given by

$$A_1 = \{(T, T)\}, \quad A_2 = \{(H, T), (T, H)\}, \quad A_3 = \{(H, H)\}.$$

Since $p(\omega_i) \equiv \frac{1}{4}$, we get

$$\mathbb{P}(A_1) = p_X(0) = \frac{1}{4}, \quad \mathbb{P}(A_2) = p_X(1) = \frac{1}{2}, \quad \mathbb{P}(A_3) = p_X(2) = \frac{1}{4}.^{2}$$

Hence,

$$\mathbb{E}X = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

Example 1.3.3 Consider again Example 1.2.2. We have

$$p_X(2) = p_X(12) = \frac{1}{36},\ p_X(3) = p_X(11) = \frac{2}{36},\ p_X(4) = p_X(10) = \frac{3}{36},$$
$$p_X(5) = p_X(9) = \frac{4}{36},\ p_X(6) = p_X(8) = \frac{5}{36},\ p_X(7) = \frac{6}{36}.$$

Hence

$$\mathbb{E}X = (2{+}12){\cdot}\frac{1}{36}{+}(3{+}11){\cdot}\frac{2}{36}{+}(4{+}10){\cdot}\frac{3}{36}{+}(5{+}9){\cdot}\frac{4}{36}{+}(6{+}8){\cdot}\frac{5}{36}{+}7{\cdot}\frac{6}{36} = 7,$$

as expected for symmetry reasons.

[2]Writing as late as 1754, d'Alembert states that these three values should be assigned equal probability!

Example 1.3.4 Consider now Example 1.2.1 with the following probability

$$p(H,H) = \alpha^2, P(H,T) = p(T,H) = \alpha\cdot(1-\alpha), p(T,T) = (1-\alpha)^2, \quad 0 \leq \alpha \leq 1.$$

Then

$$\mathbb{E}X = 0 \cdot (1-\alpha)^2 + 1 \cdot 2\alpha \cdot (1-\alpha) + 2 \cdot \alpha^2 = 2\alpha.$$

The following proposition collects some fundamental properties of the expected value.

Proposition 1.3.5 Let $\mathcal{V}$ be the vector space of all random variables on the probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$.

1. $\mathbb{E}$ is *linear* on $\mathcal{V}$, namely

$$\mathbb{E}\{\alpha X + \beta Y\} = \alpha\mathbb{E}X + \beta\mathbb{E}Y, \quad \forall \alpha, \beta \in \mathbb{R}, \forall X, Y \in \mathcal{V};$$

2. $X(\omega) \geq 0 \ \forall \omega \Rightarrow \mathbb{E}X \geq 0$. It follows that $X(\omega) \geq Y(\omega) \ \forall \omega \Rightarrow \mathbb{E}X \geq \mathbb{E}Y$;

3. $|\mathbb{E}X| \leq \mathbb{E}|X|$;

4. $\mathbb{E}\mathbb{1}_A = \mathbb{P}(A)$.

5. Let $X, Y \in \mathcal{V}$. Then they satisfy the *Cauchy-Schwarz inequality*

$$(\mathbb{E}|X \cdot Y|)^2 \leq \mathbb{E}\{X^2\} \cdot \mathbb{E}\{Y^2\}. \tag{1.6}$$

Proof . We prove (1.6), the rest being immediate. Consider the representations (1.3) for X and Y

$$X = \sum_{i=1}^{m} x_i \mathbb{1}_{A_i}, \quad Y = \sum_{j=1}^{n} y_j \mathbb{1}_{B_j}.$$

Observing that $i \neq j \Rightarrow \mathbb{1}_{A_i}(\omega) \cdot \mathbb{1}_{A_j}(\omega) \equiv 0$, we get

$$X^2 = \sum_{i=1}^{m} x_i^2 \mathbb{1}_{A_i}, \quad Y^2 = \sum_{j=1}^{n} y_j^2 \mathbb{1}_{B_j}.$$

We then get

$$\mathbb{E}X^2 = \sum_{i=1}^{m} x_i^2 \mathbb{P}(A_i), \quad \mathbb{E}Y^2 = \sum_{j=1}^{n} y_j^2 \mathbb{P}(B_j). \tag{1.7}$$

Let us assume $X(\omega) \not\equiv 0$ and $Y(\omega) \not\equiv 0$, otherwise (1.6) is trivially satisfied. Then the above representations imply $\mathbb{E}\{X^2\} > 0$ and $\mathbb{E}\{Y^2\} > 0$. Define

$$\tilde{X} := \frac{X}{\sqrt{\mathbb{E}\{X^2\}}}, \quad \tilde{Y} := \frac{Y}{\sqrt{\mathbb{E}\{Y^2\}}}.$$

We observe that $\mathbb{E}\{\tilde{X}^2\} = \mathbb{E}\{\tilde{Y}^2\} = 1$. Now, from

$$\left(|\tilde{X}| - |\tilde{Y}|\right)^2 \geq 0,$$

we get

$$2|\tilde{X}\tilde{Y}| \leq \tilde{X}^2 + \tilde{Y}^2.$$

The latter implies

$$2\mathbb{E}|\tilde{X}\tilde{Y}| \leq \mathbb{E}\{\tilde{X}^2\} + \mathbb{E}\{\tilde{Y}^2\} = 2.$$

We conclude that $\mathbb{E}|\tilde{X}\tilde{Y}| \leq 1$ which is equivalent to (1.6). □

By (1.3), $\mathbb{E}$ is the only linear functional on $\mathcal{V}$ satisfying property 4. of Proposition 1.3.5.
Let X be a random variable on $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. The *variance* of X is defined by

$$\mathbb{V}X := \mathbb{E}\{(X - \mathbb{E}X)^2\}.$$

It measures the dispersion of the values of the random variable about its mean. The quantity $\sigma = \sqrt{\mathbb{V}X}$ is called *standard deviation*. The following formula is often useful to compute the variance

$$\mathbb{V}X = \mathbb{E}\{X^2\} - (\mathbb{E}X)^2. \tag{1.8}$$

Theorem 1.3.6 (*Markov's Inequality*) Let X be a nonnegative random variable on $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. Then

$$\mathbb{P}\{\omega : X(\omega) \geq \epsilon\} \leq \frac{1}{\epsilon}\mathbb{E}X, \quad \forall \epsilon > 0. \tag{1.9}$$

Proof . Write

$$X(\omega) = X(\omega)\cdot\mathbb{1}_{\{\omega:X(\omega)\geq\epsilon\}} + X(\omega)\cdot\mathbb{1}_{\{\omega:X(\omega)<\epsilon\}} \geq X(\omega)\cdot\mathbb{1}_{\{\omega:X(\omega)\geq\epsilon\}} \geq \epsilon\mathbb{1}_{\{\omega:X(\omega)\geq\epsilon\}}.$$

Taking expected values, we get

$$\mathbb{E}X \geq \epsilon\mathbb{P}\{\omega : X(\omega) \geq \epsilon\}.$$

□

Corollary 1.3.7 (*Chebyshev's Inequality*) Let X be a random variable on $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. Then

$$\mathbb{P}\{\omega : |X(\omega) - \mathbb{E}X| \geq \epsilon\} \leq \frac{1}{\epsilon^2}\mathbb{V}X, \quad \forall\epsilon > 0. \tag{1.10}$$

Proof . $\mathbb{P}\{\omega : |X(\omega) - \mathbb{E}X| \geq \epsilon\} = \mathbb{P}\{\omega : |X(\omega) - \mathbb{E}X|^2 \geq \epsilon^2\}$. Then use (1.9). □

1.4 Entropy

Consider the function of a real variable $h(x) = -x\ln x$ on $x > 0$. Notice that

$$\lim_{x\searrow 0} h(x) = 0.$$

Hence, we can extend the domain of h to all of $x \geq 0$. Let p be a probability distribution on the sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$. Then, the (Shannon) *entropy* of p is defined by

$$H(p) := \sum_{i=1}^{N} h(p(\omega_i)) = -E\{\ln p\}. \tag{1.11}$$

The base of the logarithm is not important. In Information Theory, base 2 is employed to express entropy in *bits*. In various contexts, entropy represents the degree of uncertainty of p, our ignorance on the state of a system, etc. In physics, it is a measure of how disordered a system is. In order to discuss entropy further, we need a few basic facts on convex functions [28].

1.5 Convex functions

Let V be a vector space and $K \subseteq V$. The set K is *convex* if whenever $x, y \in K$, then $[x, y] \subseteq K$, where the "segment" $[x, y]$ is defined by

$$[x, y] = \{z : z = \lambda x + (1 - \lambda)y,\ 0 \leq \lambda \leq 1\}.$$

Observation 1.5.1 The intersection of convex subsets of V is convex.

Let $K \subseteq V$ be convex, $f : K \to \mathbb{R}$. The function f is said to be *convex* if it satisfies

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y), \quad \forall x, y \in K, \forall \lambda \in [0, 1]. \quad (1.12)$$

The function f is *strictly convex* on K if it satisfies (1.12) with strict inequality for all $x \neq y$ and $\lambda \in (0, 1)$. The function f is called (strictly) *concave* if $-f$ is (strictly) convex.

Example 1.5.2 $f(x) = mx+q$, $f(x) = x^2$, $f(x) = e^{-x}$, $f(x) = |x|$ are convex functions on $\mathbb{R}$. $f(x) = -\ln x$ is convex on $x > 0$.

A convex function defined on the interval $(a, b) \subseteq \mathbb{R}$ is always *continuous* there and *differentiable* at all but at most countably many points. A twice differentiable function is convex on (a, b) if and only if its second derivative is non-negative there. If its second derivative is positive then it is strictly convex, but the converse does not hold (take e.g. $f(x) = x^4$).

Example 1.5.3 Take $f(x) = -\ln x$ on $x > 0$. Since $f''(x) = \frac{1}{x^2}$ is positive on $x > 0$, we conclude that f is strictly convex there.

If a convex function on $(a, b) \subseteq \mathbb{R}$ is differentiable at x_0 and $f'(x_0) = 0$, then x_0 is a *global minimum point* on (a, b). A strictly convex function has at most one global minimum point.

Example 1.5.4 The convex function $f(x) = 0$ for $x \in (0, 1)$ and $f(0) = f(1) = 1$ is discontinuous at the endpoints of $[0, 1]$. All points in $(0, 1)$ are global minima since f' is zero there. The function $f(x) = |x|$ is convex and admits a unique global minimum on $\mathbb{R}$ at $x = 0$ which is the only point of nondifferentiability. The convex function $f(x) = x$ has a unique global minimum on $[a, b]$ at $x = a$ where the right-hand derivative is not zero. It has no global minimum on $(a, b]$. The function $f(x) = (x - 1)^4$ is strictly convex on $\mathbb{R}$ and has the unique global minimum at $x = 1$ where f' vanishes. The strictly convex function $-\ln x$ has no minima on $x > 0$.

Exercise 1.5.5 Let $K \subseteq V$ be convex and $(\lambda_1, \ldots, \lambda_n) \in [0,1]^n$ satisfy $\sum_{i=1}^n \lambda_i = 1$. Show that if $(x_1, \ldots, x_n) \in K^n$, then $\sum_{i=1}^n \lambda_i x_i \in K$.

Theorem 1.5.6 (*Jensen 1906*) Let K be a convex subset of the vector space V, and let $f : K \to \mathbb{R}$. The following properties are equivalent:

1. f is convex;
2. $epi f := \{(x, \alpha) \in K \times \mathbb{R} | f(x) \leq \alpha\}$ is a convex subset of $V \times \mathbb{R}$;
3. For all $n \in \mathbb{N}$, for all $(x_1, \ldots, x_n) \in K^n$, and for all $(\lambda_1, \ldots, \lambda_n) \in [0,1]^n$ such that $\sum_{i=1}^n \lambda_i = 1$ we have the Jensen inequality

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i). \tag{1.13}$$

Proof . 3 $\Rightarrow$ 1 is apparent. Let us prove first that 1 $\Rightarrow$ 3. We know that property 3 holds for $n = 2$. Suppose now it holds for $n-1$, $n \geq 3$. Then, assuming $\lambda_n < 1$ otherwise the inequality is trivially satisfied, we have

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) = f\left((1-\lambda_n)\sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} x_i + \lambda_n x_n\right).$$

Observe now that

$$0 \leq \frac{\lambda_i}{1-\lambda_n} \leq 1, \quad \sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} = 1.$$

By Exercise 1.5.5, the point $\sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} x_i$ belongs to K. By convexity of f, we now get

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq (1-\lambda_n) f\left(\sum_{i=1}^{n-1} \frac{\lambda_i}{1-\lambda_n} x_i\right) + \lambda_n f(x_n).$$

It now suffices to invoke the induction hypothesis.
To prove 2 $\Rightarrow$ 3, notice that all points $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$ lay in *epi* f. Since the latter is a convex set, if $\sum_{i=1}^n \lambda_i = 1$, $\lambda_i \in [0,1]$, we must have

$$\left(\sum_{i=1}^n \lambda_i x_i, \sum_{i=1}^n \lambda_i f(x_i)\right) = \sum_{i=1}^n \lambda_i (x_i, f(x_i)) \in epi f.$$

The latter implies

$$f\left(\sum_{i=1}^{n}\lambda_i x_i\right) \leq \sum_{i=1}^{n}\lambda_i f(x_i).$$

$1 \Rightarrow 2$ is left as an exercise. □

Observation 1.5.7 Along the same lines as in Theorem 1.5.6, one can show that f is strictly convex if and only if (1.13) holds with strict inequality whenever $\sum_{i=1}^{n}\lambda_i x_i \neq x_j$, $\forall j$.

Corollary 1.5.8 Let K be a convex subset of the vector space V and let $f_\alpha : K \to (-\infty, +\infty]$, $\alpha \in I$, be a collection of convex functions. The *upper hull* of the collection is defined by

$$g(x) = \sup\{f_\alpha(x) | \alpha \in I\},\ x \in K.$$

Then g is convex.

Proof . Observe that

$$epi\ g = \cap_\alpha epi\ f_\alpha.$$

Since each $epi\ f_\alpha$ is convex, so is their intersection. □

Let V be a vector space, and let $A \subseteq X$. The *convex hull* of A, written $conA$, is the intersection of all convex subsets of V containing A. The convex hull of $n+1$ *affinely independent*[3] points of a Euclidean space is called an *n-simplex*. For example, a 1-simplex is a line segment, a 2-simplex is a triangle and a 3-simplex is a tetrahedron.

1.6 The simplex of probability distributions

Let $\mathcal{D}(\Omega)$ denote the family of all probability distributions on the sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$. Then $\mathcal{D}(\Omega)$ is an $(N-1)$-simplex whose vertices are the singular distributions $p_i(\omega_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta. The latter distributions correspond to situations where no randomness is present.

[3]The points $x_1, x_2, \ldots, x_{n+1}$ are called affinely independent if every point x in their convex hull admits a *unique* representation as convex combination of the points.

Theorem 1.6.1 The entropy function H is strictly concave on $\mathcal{D}(\Omega)$. Moreover, it satisfies

$$0 \leq H(p) \leq \ln N. \tag{1.14}$$

In particular, $H(p) = 0$ if and only if p is a vertex of the simplex and $H(p) = \ln N$ if and only if $p = p_u$ the center of the simplex, where $p_u(\omega_i) \equiv \frac{1}{N}$.

Proof The function $h(x) = -x \ln x$ is nonnegative on $[0, 1]$ since both x and $-\ln x$ are nonnegative on $(0, 1]$. Hence, $H(p) \geq 0$ on $\mathcal{D}(\Omega)$. Moreover, $h(0) = h(1) = 0$, and, on $(0, 1]$, $h'(x) = -\ln x - 1$ which is positive for $x \in (0, e^{-1})$ and negative for $x \in (e^{-1}, 1]$. The function h is also strictly concave since, on $(0, 1]$, $h''(x) = -\frac{1}{x}$ is negative. It follows that $H(\cdot)$ is strictly concave on $\mathcal{D}(\Omega)$. Indeed, let $p, q \in \mathcal{D}(\Omega)$ and write $p_i := p(\omega_i)$, $q_i := q(\omega_i)$. Then, $H(\lambda p + (1-\lambda)q) = \sum_{i=1}^{N} h(\lambda p_i + (1-\lambda)q_i)$. For $p \neq q$ and $\lambda \in (0, 1)$, strict concavity of h on $[0, 1]$ implies $h(\lambda p_i + (1-\lambda)q_i) > \lambda h(p_i) + (1-\lambda)h(q_i)$, for some i. Hence, $H(\lambda p + (1-\lambda)q) > \lambda H(p) + (1-\lambda)H(q)$.

Moreover, $H(p) = 0$ implies that each $p(\omega_i)$ is either zero or one. Since they must add to one, it follows that $p(\omega_i) = \delta_{ij}$ for some j.

By Jensen inequality, we also have

$$\begin{aligned} H(p) = \sum_{i=1}^{N} h(p(\omega_i)) = N \sum_{i=1}^{N} \frac{1}{N} h(p(\omega_i)) &\leq N h\Big(\sum_{i=1}^{N} \frac{1}{N} p(\omega_i)\Big) \\ &= N h(\frac{1}{N}) = -N \frac{1}{N} \ln(\frac{1}{N}) = \ln N = H(p_u). \end{aligned}$$

Finally, by strict concavity of h, we get $H(p) < H(p_u)$ whenever $p \neq p_u$ (Observation 1.5.7). □

Problems

Problem 1 Establish the following relations:

a. $\mathbb{P}(A) \cdot \mathbb{P}(A^c) \leq \frac{1}{4}$;

b. $\mathbb{P}(B \cap C) \leq \mathbb{P}(B)\mathbb{P}(C) + \frac{1}{4}$.

Problem 2 Consider rolling three dice and adding the number of dots. The number of triples of numbers from 1 to 6 whose sum is 9 is the same as those with sum 10. Is the probability of 9 equal to the probability of 10?

Problem 3 Let the random variable X take values in $\{1, 2, \ldots, N\}$. Establish the following *telescope formula*:

$$\mathbb{E}X = \sum_{k=1}^{M} \mathbb{P}(X \geq k).$$

Problem 4 Prove (1.8).

Problem 5 Prove $1 \Rightarrow 2$ in Theorem 1.5.6.

Chapter 2

Thermodynamic systems: Statics

2.1 A finite Gibbs variational principle

Consider a physical system completely described (at the *mesoscopic* level) by the discrete state space $\mathcal{T} = \{1, 2, \ldots, N\}$. We can think of this mesoscopic description as originating from a microscopic description where the *phase space* Γ has undergone a "coarse graining" through subdivision into small cells. Each of the cells represents a mesoscopic state. For each state i we consider its *energy* E_i. The function $\mathcal{H} : i \mapsto E_i$ is called *Hamiltonian*. The thermodynamic states of the system are given by probability distributions on $\mathcal{T}$, namely by $\mathcal{S} := \mathcal{D}(\mathcal{T})$. On $\mathcal{S}$, we define the *internal energy* as the expected value of the Energy *observable* in state p

$$U(E, p) = \mathbb{E}_p\{\mathcal{H}\} = \sum_i E_i p_i = \langle E, p \rangle, \tag{2.1}$$

where E denotes the N-dimensional vector with components E_i. Let us also introduce the *Gibbs entropy*

$$S(p) = kH(p) = -k \sum_i p_i \ln p_i, \tag{2.2}$$

where k is Boltzmann's constant. By Theorem 1.6.1, S is nonnegative and strictly concave on $\mathcal{S}$. Let $\bar{E}$ be a constant satisfying

$$E_m = \min_i E_i \leq \bar{E} \leq \frac{1}{N} \sum_i E_i. \tag{2.3}$$

We can think of $\bar{E}$ as the energy of the underlying conservative microscopic system (The reason for taking $\frac{1}{N}\sum_i E_i$ as upper bound in (2.3) is explained in Observation 2.2.2 below). We want to study the following *Maximum Entropy* problem:

$$\text{maximize} \quad \{S(p); p \in \mathcal{S}\} \tag{2.4}$$
$$\text{subject to} \quad U(E,p) = \bar{E}. \tag{2.5}$$

This is an (important) instance of a class of maximum entropy problems, see [26] for a recent survey, where entropy is maximized over probability distributions which give the correct expectation of certain observables in accordance with known macroscopic quantities. Another maximum entropy problem is discussed in the dynamical setting in Section 8.4.

In order to solve this constrained optimization problem, we resort to a fundamental, albeit elementary, result. Let Y be a nonempty set and let $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ denote the extended reals. Consider the maximization of $J : Y \to \bar{\mathbb{R}}$ over the nonempty subset M of Y.

Definition 2.1.1 The map $\Lambda : Y \to \bar{\mathbb{R}}$ is called a *Lagrange Functional* for the optimization problem if it is *finite* and *constant* over M.

Lemma 2.1.2 (*Lagrange Lemma*) Let $\Lambda : Y \to \bar{\mathbb{R}}$ be a Lagrange functional and let $y_0 \in M$ maximize $\mathcal{L} = J + \Lambda$ over Y. Then y_0 maximizes J over M.

Proof . For any $y \in M$, we have $J(y_0) + \Lambda(y_0) \geq J(y) + \Lambda(y) = J(y) + \Lambda(y_0)$. Hence $J(y_0) \geq J(y)$. □

Notice that this apparently innocuous result is in fact a lethal (scientific!) weapon in that it is extremely general. Indeed, it does not require any algebraic nor topological structure on Y and the hypotheses on Λ are also minimal. With this tool, we can attack problem (2.4)-(2.5). Let us introduce the positive orthant $\mathbb{R}^N_+ = \{p = (p_1, p_2, \ldots, p_N) | p_i \geq 0, i = 1, \ldots, N\}$. In our setting, $\mathbb{R}^N_+ = Y$, $S = J$ and

$$M = \{p = (p_1, p_2, \ldots, p_N) | p_i \geq 0, i = 1, \ldots, N, \sum_i p_i = 1, U(E,p) = \bar{E}\}.$$

We take

$$\Lambda(p) = \lambda(\bar{E} - U(E,p)) + \mu(\sum_i p_i - 1),$$

where $\lambda, \mu \in \mathbb{R}$, $\lambda \geq 0$, are called *Lagrange multipliers*. Since $\Lambda(p) \equiv 0$ on M, it is a Lagrange Functional for our problem. The *Lagrangian function* is then given by

$$\mathcal{L}(p, \lambda, \mu) := S(p) + \lambda(\bar{E} - U(E, p)) + \mu(\sum_i p_i - 1). \tag{2.6}$$

In the spirit of Lagrange Lemma 2.1.2, we consider the *unconstrained* maximization of $\mathcal{L}(\cdot, \lambda, \mu)$ over $\mathbb{R}^N_+$. Observing that $\lambda\bar{E}$ and μ are constants, it is equivalent to maximizing over $\mathbb{R}^N_+$ the functional

$$I(p) = -k\sum_i p_i \ln p_i - \lambda \sum_i E_i p_i + \mu \sum_i p_i = \sum_i [-k \ln p_i - \lambda E_i + \mu] p_i. \tag{2.7}$$

Observe that

$$I(p) = \sum_i f_i(p_i), \quad f_i(p_i) = [-k \ln p_i - \lambda E_i + \mu] p_i.$$

Hence, the problem is equivalent to maximize each $f_i(p_i)$ over $\mathbb{R}_+$. Observe that $[-k \ln p_i - \lambda E_i + \mu] p_i$ is strictly concave (it is equal to the strictly concave function $kh(p_i)$ plus a *linear function*) on $\mathbb{R}_+$. Hence the vanishing of f_i' is a *sufficient* condition for a maximum point. Setting $f_i'(p_i) = 0$ for each i, we get the optimality condition

$$-k \ln p_i - k - \lambda E_i + \mu = 0.$$

The latter yields

$$p_i^* = \exp[-1 + \frac{\mu}{k} - \frac{\lambda}{k} E_i]. \tag{2.8}$$

Since each $p_i^* \geq 0$, $p^* = (p_1^*, \ldots, p_N^*)$ is the maximum point of $I(p)$. In order for p^* to solve the original constrained problem, it must lie in M. Let us first worry about condition $\sum_i p_i^* = 1$. We can choose μ so that

$$\exp[1 - \frac{\mu}{k}] = \sum_i \exp[-\frac{\lambda}{k} E_i] := Z\left(\frac{\lambda}{k}\right). \tag{2.9}$$

This guarantees that $\sum_i p_i^* = 1$. The letter Z was chosen by Boltzmann to indicate "zuständige Summe" (pertinent sum). Z is called (canonical) *partition function* in Statistical Mechanics. Define the *absolute temperature*

$T := \lambda^{-1}$ for $\lambda > 0$ and $T = +\infty$ for $\lambda = 0$. Define also the so-called *inverse temperature* β as

$$\beta := \frac{\lambda}{k} = \frac{1}{kT}.$$

We can then rewrite (2.8)-(2.9) as

$$p_i^* = Z(\beta)^{-1} \exp[-\beta E_i], \quad Z(\beta) = \sum_i \exp[-\beta E_i]. \tag{2.10}$$

The *Boltzmann distribution* (2.10) (also called *Gibbs distribution*) corresponds to the equilibrium thermodynamical state at the temperature T. Let us introduce the *Free Energy* functional F defined by

$$F(E, p, T) := U(E, p) - TS(p). \tag{2.11}$$

Since S is strictly concave on $\mathcal{S}$ (Theorem 1.6.1) and $U(E, \cdot)$ is linear, it follows that F is strictly convex on the state space $\mathcal{S}$.

Theorem 2.1.3 (*Gibbs' Principle*) The Boltzmann distribution p^* is a minimum point of the free energy F on $\mathcal{S}$.

Proof . Observe that $-TI(p)$, where $I(p)$ is given by (2.7), and F differ by a constant on $\mathcal{S}$. The result then follows from the above argument. □

2.2 The role of temperature

Consider again the Boltzmann distribution

$$p_i^*(\beta) = Z(\beta)^{-1} \exp[-\beta E_i], \quad Z(\beta) = \sum_i \exp[-\beta E_i], \quad \beta = \frac{1}{kT}. \tag{2.12}$$

For $\beta = 0$, equivalently for $T = +\infty$, we get $p^* = p_u$, namely the Boltzmann distribution coincides with the uniform distribution. This agrees with our intuitive idea of temperature. Mathematically, $\beta = \lambda = 0$ implies that no weight is put on the constraint when maximizing the Lagrangian (2.6). Hence, the maximum point in $\mathcal{S}$ simply is the maximum entropy distribution. What happens when T tends to zero (equivalently, when β tends to $+\infty$)? Let

$$A_m := \{i \in \mathcal{T} | \mathcal{H}(i) = E_m\}.$$

This is the set of *minimum energy* states. Consider a state $i \notin A_m$. Since $Z(\beta) \geq \exp[-\beta E_m]$, we get

$$p_i^*(\beta) = Z(\beta)^{-1} \exp[-\beta E_i] \leq \exp[-\beta(E_i - E_m)].$$

The right hand side tends to zero when $\beta \to +\infty$ since $E_i - E_m > 0$. It follows that

$$\lim_{\beta \to +\infty} \mathbb{P}^*(\beta)(A_m) = 1.$$

We conclude that, when T tends to zero, the Boltzmann distribution tends to concentrate itself on the minimum points of the Hamiltonian function.

We now go back to the original maximum entropy problem (2.4)-(2.5). In doing so, we shall investigate the dependence of the absolute temperature T on the internal energy $\bar{E}$. By Lemma 2.1.2, if the Boltzmann distribution satisfies (2.5), it solves the original constrained maximum entropy problem. Thus, we get the condition

$$\mathbb{E}_{p^*(\beta)}(\mathcal{H}) = \sum_i E_i p_i^*(\beta) = \sum_i E_i Z(\beta)^{-1} \exp[-\beta E_i] := G(\beta) = \bar{E}. \quad (2.13)$$

Observe that if E_i are all equal (constant Hamiltonian function), then necessarily $E_i = \bar{E}, \forall i$ and (2.13) is satisfied for any β. We consider now the more interesting case where the E_i are not all equal. We study the dependence of β on $\bar{E}$ as given by (2.13).

Theorem 2.2.1 Assume that $\mathcal{H}$ is not constant. Then the function $G : \beta \mapsto G(\beta)$ in (2.13) is *strictly decreasing* on $\beta \geq 0$, bijectively mapping $[0, +\infty)$ onto $\left(E_m, \frac{1}{N}\sum_i E_i\right]$. Consequently, there does exist a continuously differentiable inverse function $\beta = G^{-1}(\bar{E})$.

Proof . Observe that

$$G(\beta) = -\frac{1}{Z(\beta)}\frac{dZ}{d\beta} = -\frac{d \ln Z(\beta)}{d\beta}.$$

We have

$$\frac{dG}{d\beta} = \sum_i E_i Z(\beta)^{-1} \exp[-\beta E_i] \left[-E_i + G(\beta)\right] = G(\beta)^2 - \sum_i E_i^2 Z(\beta)^{-1} \exp[-\beta E_i].$$

By (1.8), we then get

$$\frac{dG}{d\beta} = \left(\mathbb{E}_{p^*(\beta)}\mathcal{H}\right)^2 - \mathbb{E}_{p^*(\beta)}\left(\mathcal{H}^2\right) = -\mathbb{V}_{p^*(\beta)}\left(\mathcal{H}\right) < 0. \quad (2.14)$$

□

Observation 2.2.2 The above result shows that, given the value of the internal energy

$$\bar{E} \in \left[E_m, \frac{1}{N} \sum_i E_i \right],$$

there exists one and only one value of the absolute temperature $T = 1/k\beta$ such that the corresponding Boltzmann distribution (2.10) satisfies the constraint (2.13), and therefore solves the original maximum entropy problem (2.4)-(2.5) (with the understanding that $T = 0$ is associated to $\bar{E} = E_m$ since $\lim_{\bar{E} \searrow E_m} G^{-1}(\bar{E}) = +\infty$). Moreover, higher values of $\bar{E}$ correspond to higher values of the temperature.

2.3 The travelling salesman problem

Consider the following problem in combinatorial optimization. A travelling salesman has to visit $n+1, n \geq 2$ cities. Any two cities are connected and he knows all distances. He wishes to find the shortest path that originates and ends in his hometown going through all cities only once (Hamiltonian cycle). The number of such paths is the number of permutations of n objects divided by two, namely $\frac{1}{2}n!$[1] (see Subsection 3.1.3) which grows very rapidly. Trying all the permutations and seeing which one is shortest becomes rapidly impractical. Using techniques from *optimal control*, the problem can be solved in time $O(n^2 2^n)$. Although this is exponential, it is still much better than $O(n!)$.

This problem is an important instance of a NP- hard problem. Roughly speaking, these are problems for which no polynomial-time algorithm is known (meaning, no algorithm is known that can complete the computation in an amount of time which is polynomial in the input size). Another important example is the factorization of large semiprimes (products of two prime numbers). For instance, a recent effort which factored a 200-digit number (RSA-200) took eighteen months and used over half a century of computer time! (RSA is an algorithm for public-key cryptography based on the difficulty of factoring integers). A *quantum computer* could solve this problem more efficiently than a classical computer using *Shor's algorithm* to find its factors. The fundamental reason is that quantum computing permits some kind of massive *parallelism* (to get some intuition on that, look up the

[1] $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$.

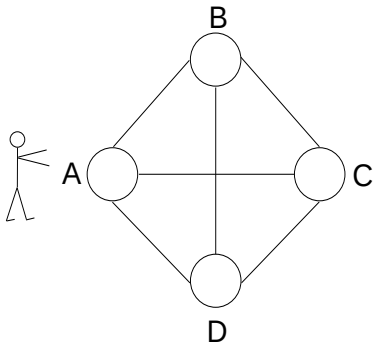

Figure 2.1: If a salesman starts at point A, and if the distances between every pair of points are known, what is the shortest route which visits all points and returns to point A?

solution to Deutsch's problem in Appendix B). This ability would allow a quantum computer to "break" many of the cryptographic systems in use today, in the sense that there would be a polynomial time (in the number of bits of the integer) algorithm for solving the problem.

Now let $N = \frac{1}{2}n!$. For each path $i = 1, \ldots, N$, let E_i denote its length. Let us consider the Boltzmann distribution

$$p_i^*(\beta) = Z(\beta)^{-1} \exp[-\beta E_i], \quad Z(\beta) = \sum_i \exp[-\beta E_i], \quad \beta = \frac{1}{kT}. \qquad (2.15)$$

Notice that the shorter the path, the higher the probability. Hence, if we *sample* from p^*, we have higher chance to get i with shorter length! There is a small problem though: We don't know p^* (if we did, we would know all the path lengths and would have already solved the problem). How can we sample from an unknown distribution? As we shall see in Chapter 10, Subsection 10.4.2, the *Metropolis algorithm* accomplishes precisely that.

Problems

Problem 6 To prove Gibbs' Principle, we have used the fact that the sum of a strictly concave function and a linear function is strictly concave. Prove the following more general result. Let f and g, defined on the convex set K, be convex. Then, if $\alpha > 0$ and $\beta > 0$, $h = \alpha f + \beta g$ is also convex on K. If at least one of f and g is *strictly* convex, so is the function h.

Problem 7 Let $\mathcal{X} = \{1, 2, \ldots, n\}$, and let $\mathcal{D}(\mathcal{X})$ be the simplex of probability distributions on $\mathcal{X}$. Solve the following *Maximum Entropy* problem:

$$\text{maximize} \qquad \{H(p); p \in \mathcal{D}(\mathcal{X})\} \tag{2.16}$$

$$\text{subject to} \qquad \sum_{i=1}^{n} i p_i = \alpha, \tag{2.17}$$

where $0 < \alpha \leq \frac{n+1}{2}$. It si namely the problem of finding the maximum entropy distribution among those with a given expectation.

Problem 8 Consider the travelling salesman problem with 4 cities. Suppose the cities occupy the vertices of a square. Find all the Hamiltonian cycles and their length.

Chapter 3

Uniform probability spaces

3.1 Combinatorics

Let us go back to n coin tosses, where

$$\Omega_n = \{\omega = (a_1, a_2, \ldots, a_n),\ a_i = H \text{ or } T\}, \quad |\Omega_n| = 2^n.$$

The *Law of averages* states that, for n large, the number of heads is "about the same" as the number of tails. In order to see whether this layperson's belief can be turned into a mathematical principle, we pose, as in [4, p. 1], the following question:

Question 3.1.1 *How many of the 2^n sequences of H and T have exactly k heads?*

To answer this question, we need first to develop some basic counting skills. Consider a probability space with the uniform measure $(\Omega, \mathcal{P}(\Omega), \mathbb{P}_u)$. Namely for $A \subseteq \Omega$, $\mathbb{P}_u(A) = \frac{|A|}{N}$. Thus computing the probability of any event reduces to counting its elements. This is often a highly nontrivial task in the realm of *combinatorics*. Combinatorics is a branch of pure mathematics concerned with regrouping and/or ordering a finite set of objects. One wishes to count the different ways. There are two important aspects:

- Is the ordering important? For instance is $(x, y, z) = (z, x, y)$?
- Are repetitions of the same object possible?

3.1.1 Dispositions with repetitions

Consider the task of placing k balls in n cells, where it is allowed to place more than one ball in a cell. There are

$$DR_n^k = n^k$$

such arrangements that are called *dispositions with repetitions* (sometimes, unhappily, permutations with repetitions). Equivalently, one can think of the ordered drawing of k balls from n numbered balls with replacement. For dispositions with repetitions order matters.

Example 3.1.2 In all of the following examples, the possible alternatives are provided by dispositions with repetitions:

- placing 3 balls in 4 cells can be done in $4^3 = 64$ different ways;
- there are 365^{10} different alternatives for the possible birthdays of ten persons none of whom was born on February 29;
- there are 7^k different ways k car accidents could have occurred during a week;
- in an experiment on cosmic rays, there are n^k ways k particles can hit n Geiger counters;
- random numbers: ordering k digits can be done in 10^k ways;
- sex distribution among k individuals (n=2). There are 2^k binary sequences of length k;
- *gene distribution*. Each gene may appear in n different forms $A_1, \ldots, A_n$ (*genotypes*). The descendants of an individual (person, animal, plant) inherit one genotype. When there are k descendants, this can happen in n^k different ways.

3.1.2 Dispositions without repetitions

Consider the task of placing k balls in n cells, where at most one ball is allowed in a cell. The number of possible *dispositions without repetitions* (sometimes permutations without repetitions) is

$$D_n^k = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!},$$

where $0! = 1$ by definition. Equivalently, one can think of the ordered drawing of k balls from n numbered balls without replacement. Order matters.

Example 3.1.3 The number of different three letter words one can form from $\{A, B, C, D, E\}$ without repeating letters is $5 \cdot 4 \cdot 3 = 60$.

Example 3.1.4 Consider Problem 2 in the questionnaire. We have a group on N persons. Assume that the probability of being born in any of the 365 days of the year is the same (we disregard leap years). Find the smallest N such that the probability that at least two individuals in the group have the same birthday is greater than $1/2$.

It is easier to compute the probability p that all N persons have different birthdays. We get

$$p = \frac{D_{365}^N}{DR_{365}^N} = \frac{365 \cdot 364 \cdot 363 \cdots (365 - N + 1)}{365^N} = (1-\frac{1}{365})\cdot(1-\frac{2}{365})\cdot(1-\frac{N-1}{365}). \tag{3.1}$$

For N small, we can ignore all the cross products, getting

$$p \approx 1 - \frac{1 + 2 + \cdots + N - 1}{365} = 1 - \frac{N(N-1)}{2 \cdot 365} = 1 - \frac{N(N-1)}{730}.$$

For $N = 10$, we get $p \approx 0.877$, the correct value being 0.883. For larger N, we get a better approximation using logarithms and the fact that, for small positive x, we have $\log(1 - x) \approx -x$. From (3.1), we then have

$$\log p \approx -\frac{1 + 2 + \cdots + (N-1)}{365} = -\frac{N(N-1)}{730}.$$

For $N = 30$, this approximation yields 0.3037, whereas the correct value is $p = 0.294$. For $N = 22$ the correct value is 0.524 very close to $1/2$. In contrast to our intuition that leads us to choose large $N's$, the answer to the problem is $N = 23$.

3.1.3 Permutations

These are the different ways to order n objects and are therefore called *permutations*. It is a particular case of dispositions without repetitions when $k = n$. Indeed,

$$P_n = D_n^n = \frac{n!}{0!} = n!.$$

Example 3.1.5 In the travelling salesman problem of Section 2.3 with $n+1$ cities, the number of Hamiltonian paths is $\frac{1}{2}n!$. Indeed, each Hamiltonian cycle from and to a given city corresponds to a permutation of the other n cities. Moreover, as we do not distinguish between say $(1,3,4,2,1)$ and $(1,2,4,3,1)$, we need to divide by two.

Example 3.1.6 One evening I'm going to the movies with seven other people one of whom I strongly dislike. If we sit in a row next to each other at random, what is the probability that the unpleasant person sits next to me?

There are 8! different ways we can sit. Let A be the event "I sit next to the unpleasant person". We need to calculate the number of elements of A. These elements can be partitioned in the following two subsets

$$(+,+,\text{me},\text{unpleasant person},+,+,+,+), \tag{3.2}$$

and

$$(+,+,+,\text{unpleasant person},\text{me},+,+,+). \tag{3.3}$$

The number of sequences of the type (3.2) is the number of permutations of seven elements. Similarly for the sequences of type (3.3). We conclude that

$$|A| = 7! + 7! = 2 \cdot 7!.$$

If all arrangements are equally likely, the sought probability is

$$\mathbb{P}(A) = \frac{2 \cdot 7!}{8!} = \frac{1}{4}.$$

3.1.4 Combinations without repetitions

Consider placing k indistinguishable balls in n cells, with at most one ball in each cell. Equivalently, consider drawing simultaneously k balls from n numbered. Here order does not matter. These are called combinations without repetitions. Their number is denoted by C_n^k (or $C(n,k)$). Since k objects may be ordered in $k!$ different ways, we have

$$C_n^k \cdot k! = D_n^k = \frac{n!}{(n-k)!} \Rightarrow C_n^k = \frac{n!}{k!(n-k)!} := \left(\begin{array}{c} n \\ k \end{array} \right).$$

For $k < 0$ or $k > n$, set $\left(\begin{array}{c} n \\ k \end{array} \right) = 0$.

Example 3.1.7 Among the 2^n sequences of outcomes in n coin tosses, there are $\left(\begin{array}{c} n \\ k \end{array} \right)$ that have exactly k heads. For instance, in five coin tosses, there are

$$\left(\begin{array}{c} 5 \\ 2 \end{array} \right) = 10$$

different ways to have exactly two heads.

Observation 3.1.8 In Example 3.1.4, our intuition fails because we do not estimate correctly the number of possible matches

$$\left(\begin{array}{c} 23 \\ 2 \end{array} \right) = 253.$$

The *binomial coefficient* $\left(\begin{array}{c} n \\ k \end{array} \right)$ (read "n choose k") owns its name to the expansion

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^{n-k}\, b^k. \tag{3.4}$$

Notice that the binomial theorem (3.4) implies, taking $a = b = 1$,

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) = 2^n. \tag{3.5}$$

This has an obvious interpretation in terms of n coin tosses in view of Example 3.1.7. Directly from the definition we get

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right), \quad \left(\begin{array}{c} n \\ n \end{array} \right) = \left(\begin{array}{c} n \\ 0 \end{array} \right) = 1.$$

It is not difficult to establish *Pascal's rule*

$$\left(\begin{array}{c} n \\ k \end{array} \right) + \left(\begin{array}{c} n \\ k+1 \end{array} \right) = \left(\begin{array}{c} n+1 \\ k+1 \end{array} \right), \quad 0 \leq k < n. \tag{3.6}$$

It follows from Pascal's identity that the binomial coefficient is always a natural number.

Exercise 3.1.9 Show that for n even $\left(\begin{array}{c} n \\ k \end{array} \right)$ is maximum for $k = n/2$ and for n odd it is maximum for $k = (n-1)/2$ and $(n+1)/2$.

Table 3.1: Rows $0-4$ in Tartaglia's Triangle.

			1					
		1		1				
	1		2		1			
1		3		3		1		
1		4		6		4		1

¿From Pascal's identity (3.6), we get the Tartaglia's Triangle (also named after Pascal, Yang Hui, Omar Khayyám, Pingala, Stiefel,...), see Table 3.1. The triangle's rows are precisely the coefficients which arise in the binomial expansion (3.4). Here are some of the triangle properties.

1. The triangle is symmetric;
2. by (3.5), the sum of the n^{th} row is 2^n;
3. the digits in the n^{th} row compose the number 11^n. For $n = 5$, this amounts to $11^5 = 1 \cdot 10^5 + 5 \cdot 10^4 + 10 \cdot 10^3 + 10 \cdot 10^2 + 5 \cdot 10 + 1 \cdot 10^0$. Similarly for $n > 5$;
4. sums over "diagonals" yield the *Fibonacci numbers*, i.e.

$$\sum_{k=0}^{n} \left(\begin{array}{c} n-k \\ k \end{array} \right) = F(n+1),$$

where the Fibonacci (Leonardo Pisano was posthumously named *filius Bonacci*) numbers $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$ are defined by the recursion

$$F(n+1) = F(n) + F(n-1), \quad F(0) = 0,\ F(1) = 1. \tag{3.7}$$

As observed by Kepler,

$$\lim_n \frac{F(n+1)}{F(n)} = \varphi, \tag{3.8}$$

where $\varphi = \frac{1+\sqrt{5}}{2} \approx 1.618$, called *golden ratio* (*sectio aurea*) is the only positive solution of $x^2 - x - 1 = 0$. This ratio, has fascinated scientists,

artists, musicians, etc. since antiquity as it appears in mathematics, architecture, aesthetics,and so on.[1]

The *tetractys of the decad* was a triangular pattern of points of central importance in Pythagorean philosophy. It contained the *triangular numbers* $(\frac{1}{2}n \cdot (n+1))$ $1, 3, 6, 10$. The *tetrahedral numbers* are the sum of the triangular numbers. Natural, triangular and tetrahedral numbers are found on diagonals of the Tartaglia's triangle. See [13, pp. 107-112] for further historical information on this triangle.

Example 3.1.10 A college employee has to schedule in the finals week ten final exams to take place in one specific classroom. Considering that each working day an exam can be scheduled in the morning or in the afternoon, there are ten available slots. There are three math exams, four physics exams, two chemistry exams and one biology exam. How many different sequences of finals are there? Let us reason as follows: Suppose the employee places first the math exams. This can be done in $\left(\begin{array}{c} 10 \\ 3 \end{array} \right)$ different ways. The employee then proceeds to schedule the physics exams. There are now only seven available slots. Hence, the four physics exams can be scheduled in $\left(\begin{array}{c} 7 \\ 4 \end{array} \right)$ different ways. For the chemistry exams, there are $\left(\begin{array}{c} 3 \\ 2 \end{array} \right)$ choices left. Finally, for the biology exam, there is only $\left(\begin{array}{c} 1 \\ 1 \end{array} \right) = 1$ possibility. We conclude that the different possible exam sequences are

$$\left(\begin{array}{c} 10 \\ 3 \end{array} \right) \times \left(\begin{array}{c} 7 \\ 4 \end{array} \right) \times \left(\begin{array}{c} 3 \\ 2 \end{array} \right) = \frac{10!}{3!7!} \times \frac{7!}{4!3!} \times \frac{3!}{2!1!} = \frac{10!}{3!4!2!}.$$

This is just an instance of a *multinomial coefficient* of the form

$$\frac{n!}{k_1!k_2!\ldots k_r!}, \quad k_i \in \mathbb{N}, \quad \sum_{i=1}^{r} k_i = n, \tag{3.9}$$

which generalizes the binomial coefficient. It represent the number of different ways we can arrange n balls if we can distinguish them by groups.

[1] One of the pioneers of probability was Luca Pacioli who taught mathematics to Leonardo da Vinci. He wrote *De Divina Proportione*, published in 1509, which is mostly devoted to φ (Leonardo did the engraving). Pacioli also gave seminal contributions to the field now known as accounting and wrote a chess treatise *De Ludo Scacchorum* around year 1500 which was only discovered in an aristocratic private library in 2006! The drawings may be due to Leonardo, see Figure 3.1.

Example 3.1.11 Let us solve Problem 6 in the questionnaire. Let p_n the probability that at least one of the coats is returned to the owner, Consider the sample space of permutations of n objects

$$\Omega = \{(a_1, \ldots, a_n),\ 1 \leq a_i \leq n,\ i \neq j \Rightarrow a_i \neq a_j\}.$$

We know that $|\Omega| = n!$. Let $A_i := \{\omega : a_i = i\}$, namely the event "the i^{th} person recovers his coat". Then, denoting by $\mathbb{P}$ the uniform measure on $\mathcal{P}(\Omega)$, we have

$$\begin{aligned} p_n &= \mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_n) \\ &-\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \ldots + \mathbb{P}(A_1 \cap A_2 \cap A_3) \\ &+\mathbb{P}(A_1 \cap A_2 \cap A_4) + \ldots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n). \end{aligned}$$

Observe now that

$$\mathbb{P}(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n},\ \forall i \Rightarrow \sum_{i=1}^{n} \mathbb{P}(A_i) = n \cdot \frac{1}{n} = 1.$$

Next we consider the intersections of two sets

$$i \neq j \Rightarrow \mathbb{P}(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

On the other hand, there are $\left(\begin{array}{c} n \\ 2 \end{array} \right)$ different ways to choose 2 subsets from n. Hence

$$\begin{aligned} &-\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \ldots - \mathbb{P}(A_{n-1} \cap A_n) = -\sum_{1 \leq i < j \leq n} \mathbb{P}(A_i \cap A_j) \\ &= -\frac{n!}{2!(n-2)!} \cdot \frac{1}{n(n-1)} = -\frac{1}{2!}. \end{aligned}$$

Similarly, there are $\left(\begin{array}{c} n \\ 3 \end{array} \right)$ different ways to choose 3 subsets from n. Moreover, each one of these intersections has probability $[n(n-1)(n-2)]^{-1}$. Their total probability is

$$\left(\begin{array}{c} n \\ 3 \end{array} \right) \cdot [n(n-1)(n-2)]^{-1} = \frac{1}{3!}.$$

We conclude that

$$p_n = \mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots + (-1)^{n+1}\frac{1}{n!}.$$

To study the asymptotic behavior, recall first the series expansion

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \Rightarrow e^{-1} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!}.$$

Thus,

$$1-e^{-1} = 1-\sum_{n=0}^{\infty} \frac{(-1)^n}{n!} = -\sum_{n=1}^{\infty} \frac{(-1)^n}{n!} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n!} = 1-\frac{1}{2!}+\frac{1}{3!}-\frac{1}{4!}+\ldots.$$

Hence, as $n \to +\infty$, p_n is not monotone but it admits the limit

$$\lim_{n\to+\infty} p_n = 1 - \frac{1}{e} \approx 0.632121.$$

Contrary to our intuition, p_n is close to the limit for small n. For instance, for $n = 4$, $p_n \approx 0.625$. For $n = 5$, $p_n \approx 0.63333$. For $n = 6$, $p_n \approx 0.63194$. For $n = 7$, $p_n \approx 0.63214$.

3.2 The de Moivre-Stirling formula

To answer Question 3.1.1 effectively, we still lack a tool, namely an efficient way to compute $n!$ when n is large.

Theorem 3.2.1 (de Moivre-Stirling)

$$\lim_{n\to\infty} \frac{n!}{e^{-n}n^{n+\frac{1}{2}}(2\pi)^{\frac{1}{2}}} = 1. \tag{3.10}$$

See [10, p. 50] for a proof. De Moivre discovered that for large n, $n! \sim Ce^{-n}n^{n+\frac{1}{2}}$. Stirling then showed that $C = \sqrt{2\pi}$.

Remark 3.2.2 The quantity

$$n! - e^{-n}n^{n+\frac{1}{2}}(2\pi)^{\frac{1}{2}}$$

tends to infinity as $n \to +\infty$. The *relative error*, however, decreases rapidly. For instance, for $10! = 3,628,800$, the error is 0.8%, for 100! the error is just 0.08%.

We are now ready to answer Question 3.1.1. Consider $2n$ tosses of a coin with n large. We know that there are exactly $\left(\begin{array}{c} 2n \\ n \end{array} \right)$ sequences with exactly n heads. Using Theorem 3.2.1, we get

$$\left(\begin{array}{c} 2n \\ n \end{array} \right) = \frac{(2n)!}{(n!)^2} \approx \frac{e^{-2n}(2n)^{2n}\sqrt{4\pi n}}{e^{-2n}n^{2n}(2\pi n)} = 2^{2n}\frac{1}{\sqrt{\pi n}}. \tag{3.11}$$

Thus, for n large, the fraction of sequences of length $2n$ that have exactly n heads is estimated as

$$\frac{2^{2n}\frac{1}{\sqrt{\pi n}}}{2^{2n}} = \frac{1}{\sqrt{\pi n}}.$$

Thus, for a fair coin, the probability of getting exactly the same number of heads and tails is approximatively $(\pi n)^{-1/2}$. We conclude that, no matter what *Law of averages* may say, such a probability tends to zero as $n \to +\infty$! Moreover, we have already observed (Exercise 3.1.9) that $\left(\begin{array}{c} 2n \\ k \end{array} \right)$ attains its *maximum* at $k = n$. It follows that the probability of getting k heads and $2n-k$ tails in $2n$ tosses of a fair coin tends to zero for each fixed $0 \leq k \leq 2n$ as n tends to infinity. Hence, the probability of getting a $2n$-sequence where the number of heads differs from n by less than a given number M also decreases to zero asymptotically. This might be disconcerting at first. After a moment thought, however, we realize that we were simply asking the wrong question.

Question 3.2.3 *Let $N(n, \epsilon)$ be the number of sequences of length n where the proportion of heads differs from $1/2$ less than ϵ. What is the asymptotic behavior of $N(n, \epsilon) \cdot 2^{-n}$?*

The answer to this question is provided in the next chapter by one of the most fundamental results of all of probability.

Problems

Problem 9 In how many ways can two rooks be placed on the board so that one can capture the other?

Problem 10 We can't stand any more TLA (three letter acronyms). How many different ones are there?

Problem 11 Is it advantageous to bet on at least one six in four rolls of a fair die?

Problem 12 Prove Pascal's identity (3.6).

Problem 13 In 10 coin tosses:

a. How many sequences have exactly two heads in the first three tosses?

b. How many sequences have exactly two heads?

Problem 14

a. Compute

$$\sum_{k=1}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right).$$

b. Use this result to compute the expected value of a random variable X that takes the values $0, 1, \ldots, n$ with probability $p_X(k) = \left(\begin{array}{c} n \\ k \end{array} \right) 2^{-n}$ (binomial distribution).

Problem 15 Use the binomial expansion (3.4) in

$$(1+x)^m (1+x)^{n-m} = (1+x)^n, \quad 1 \leq m < n,$$

to prove the *Vandermonde Identity*

$$\sum_{j=0}^{k} \left(\begin{array}{c} m \\ j \end{array} \right) \left(\begin{array}{c} n-m \\ k-j \end{array} \right) = \left(\begin{array}{c} n \\ k \end{array} \right). \tag{3.12}$$

Deduce, as a particular case,

$$\sum_{j=0}^{m} \left(\begin{array}{c} m \\ j \end{array} \right)^2 = \left(\begin{array}{c} 2m \\ m \end{array} \right). \tag{3.13}$$

Problem 16 Consider again the matching problem, namely Problem 6 in the questionnaire (Example 3.1.11). What is the expected number of matches?

Problem 17 In a parking lot, there are n numbered parking spaces in a row. My parking space is not at either end of the row. I manage to park if at least one of the adjacent spaces is not occupied. I am told that today there are 2 cars in my row. What is the probability that I succeed in parking?

Problem 18 ([10, p.55]) A group of $2N$ boys and $2N$ girls is divided into two groups of equal size. Use the de Moivre-Stirling formula to estimate the probability that each group has the same number of boys and girls.

Figure 3.1: A page of *De ludo scacchorum* by Pacioli with a drawing that some attribute to Leonardo.

Chapter 4

The law of large numbers

4.1 The weak law of large numbers

Theorem 4.1.1 (Weak Law of Large Numbers - Jakob Bernoulli 1713) Recall that $N(n,\epsilon)$ denotes the number of sequences of length n in which the proportion of heads differs from $1/2$ less than ϵ.

$$\lim_{n\to+\infty}\frac{N(n,\epsilon)}{2^n}=1. \tag{4.1}$$

Proof . Let $\Omega_n=\{\omega^{(n)}=(a_1,a_2,\ldots,a_n\},a_i=H\,\text{or}\,T\}$. We define on $(\Omega_n,\mathcal{P}(\Omega_n),\mathbb{P}_n)$, where $\mathbb{P}_n$ is the uniform measure, the following random variables:

$$X_j^n(\omega^{(n)}):=\left\{\begin{array}{ll}1, & \text{if } j^{th} \text{ elementof } \omega^{(n)} \text{ is } H,\\ 0, & \text{if } j^{th} \text{ elementof } \omega^{(n)} \text{ is } T.\end{array}\right. , \quad j=1,2,\ldots,n.$$

We have

$$\mathbb{E}X_j^n=\frac{1}{2},\qquad \mathbb{E}\{X_i^n\cdot X_j^n\}=\frac{1}{4},\quad i\neq j.$$

The second expectation can indeed by computed as $\mathbb{E}\{X_i^n\cdot X_j^n\}=1\cdot\mathbb{P}_n\{\omega^{(n)}:$ element $i=$ element $j=H\}$. We now define the " centered" (zero-mean) random variables

$$\tilde{X}_j^n(\omega^{(n)}):=X_j^n(\omega^{(n)})-\frac{1}{2}.$$

It is easy to verify that

$$\mathbb{E}\{\tilde{X}_i^n\cdot\tilde{X}_j^n\}=\frac{1}{4}\delta_{ij}, \tag{4.2}$$

where δ_{ij} denotes the Kronecker delta. We now define a random variable that counts the number of heads in $\omega^{(n)}$:

$$S^n(\omega^{(n)}) := \sum_{j=1}^{n} X_j^n(\omega^{(n)}).$$

We have

$$S^n - \frac{n}{2} = \sum_{j=1}^{n} \tilde{X}_j^n.$$

Using (4.2), we then get

$$\mathbb{E}\left\{\left(\frac{S^n}{n} - \frac{1}{2}\right)^2\right\} = \frac{1}{n^2}\mathbb{E}\left\{\sum_{i,j=1}^{n} \tilde{X}_i^n \cdot \tilde{X}_j^n\right\} = \frac{1}{n^2} \cdot \frac{n}{4} = \frac{1}{4n}.$$

By Chebyshev's inequality (1.3.7), we now get

$$\mathbb{P}^n\{\omega^{(n)} : \left|\frac{S^n(\omega^{(n)})}{n} - \frac{1}{2}\right| \geq \epsilon\} \leq \frac{1}{\epsilon^2}\mathbb{E}\left\{\left(\frac{S^n}{n} - \frac{1}{2}\right)^2\right\} = \frac{1}{\epsilon^2 4n}.$$

It follows that

$$\lim_{n\to\infty} \mathbb{P}^n\{\omega^{(n)} : \left|\frac{S^n(\omega^{(n)})}{n} - \frac{1}{2}\right| \geq \epsilon\} = 0, \quad \forall \epsilon > 0. \tag{4.3}$$

Since

$$\mathbb{P}^n\{\omega^{(n)} : \left|\frac{S^n(\omega^{(n)})}{n} - \frac{1}{2}\right| \geq \epsilon\} = 1 - \frac{N(n,\epsilon)}{2^n},$$

(4.1) follows. □

4.2 Further applications of combinatorics

4.2.1 A biological application

Exercise 4.2.1 ([10, 54]) n sticks are broken into a longer and a shorter part. The $2n$ pieces are then arranged into pairs from which new sticks are formed. We are interested in the following questions:

1. What is the probability that the parts be reunited in the original order?

2. what is the probability that all longer parts be paired to short pieces?

Motivation comes from biology. When cells are exposed to harmful radiations (such as alpha[1] and beta radiation), some chromosome break. The longer part contains the centrometer. If two longer parts or two shorter parts riunite, the cell dies. We need to compute in how many ways we can form pairs from $2n$ parts. Suppose first that order matters. Then there are

$$\binom{2n}{2} \cdot \binom{2n-2}{2} \cdots \binom{2}{2} = \frac{(2n)!}{2!(2n-2)!} \cdot \frac{(2n-2)!}{2!(2n-4)!} \cdots = \frac{(2n)!}{2^n}.$$

different ways (this just a special case of the multinomial coefficient (3.9)). As we are not interested in the order, we need to divide this quantity by the number of ways we can order n pairs, namely $n!$. We get the number

$$\frac{(2n)!}{2^n \cdot n!}. \tag{4.4}$$

1. The probability they reunite in the original fashion is the reciprocal of (4.4). We get

$$\frac{2^n \cdot n!}{(2n)!} = \frac{1}{1 \cdot 3 \cdot 5 \cdot (2n-1)}.$$

   For $n = 3$, we get $\frac{1}{15}$. For $n = 46$, the number of chromosomes of almost every cell in the human body, this number is excessively small.

2. to get the second probability, we only need to multiply by $n!$ the previous one (think or reordering the shorter parts while the longer are kept fixed). We get, using (3.11) in the last step,

$$\frac{n!}{1 \cdot 3 \cdot 5 \cdot (2n-1)} = \frac{2^n \cdot n! \cdot n!}{(2n)!} = 2^n \cdot \binom{2n}{n}^{-1} \approx 2^{-n}\sqrt{\pi n}.$$

   For $n = 3$, we get $\frac{2}{5}$. For $n = 46$, we get $\approx (1.7) \cdot 10^{-13}$.

[1]Alpha particles, emitted by radioactive nuclei such as uranium or radium, consist of two protons and two neutrons bound together into a particle. It is estimated that chromosome damage from alpha particles is about 100 times greater than that caused by an equivalent amount of other radiation. On the other hand, as radiation interferes with cell division, it can be used for treatment of cancer cells that are among the fastest-dividing in the body.

4.2.2 Boltzmann's loaded dice

Suppose N dice are rolled and we are informed that the total number of spots is $N \cdot 4.5$. We are asked: What proportion of the dice are showing face $i, i = 1, 2, \ldots, 6$? First of all, let us recall that the number of different ways that N dice can fall so that n_i dice show face i is given by the multinomial coefficient (3.9)

$$\frac{N!}{n_1!n_2!\ldots n_6!}, \quad n_i \in \mathbb{N}, \quad \sum_{i=1}^{6} n_i = N. \tag{4.5}$$

We then have a "macrostate" indexed by $(n_1, n_2, \ldots, n_6)$ corresponding to $\frac{N!}{n_1!n_2!\ldots n_6!}$ "microstates", each having probability 6^{-N}. To find the most probable macrostate, we need to maximize the multinomial coefficient (4.5) under the constraint

$$\sum_{i=1}^{6} i \cdot n_i = N \cdot 4.5. \tag{4.6}$$

This procedure will yield the macrostate, among those satisfying (4.6), that *can be realized in more ways*. Let us now use the crude version of Theorem 3.2.1 $N! \approx e^{-N}N^N$. We get

$$\frac{N!}{n_1!n_2!\ldots n_6!} \approx \frac{e^{-N}N^N}{\prod_{i=1}^{6} e^{-n_i} n_i^{n_i}} = \prod_{i=1}^{6} \left(\frac{N}{n_i}\right)^{n_i} = \prod_{i=1}^{6} e^{-n_i \ln\left(\frac{n_i}{N}\right)} =$$
$$e^{-\sum_{i=1}^{6} n_i \ln\left(\frac{n_i}{N}\right)} = e^{NH(p)}, \quad p_i = \frac{n_i}{N}, i = 1, 2, \ldots, 6. \tag{4.7}$$

Thus, for N large, maximizing (4.5) under the constraint (4.6) is almost equivalent to maximizing the entropy

$$H(p) = -\sum_{i=1}^{6} p_i \ln(p_i)$$

under the constraint

$$\sum_{i=1}^{6} i \cdot p_i = 4.5, \tag{4.8}$$

which is a standard maximum entropy problem [2]. By the same variational analysis as in Chapter 2, we get that the solution has the form

$$p_i^* = \frac{e^{\lambda_i}}{\sum_{i=1}^6 e^{\lambda_i}},$$

where the λ_i must be such that

$$\sum_{i=1}^6 i \cdot \frac{e^{\lambda_i}}{\sum_{i=1}^6 e^{\lambda_i}} = 4.5.$$

Hence, the most probable macrostate is $(Np_1^*, Np_2^*, \ldots, Np_6^*)$ and we expect to find $n_i^* = Np_i^*$ dice showing face i. More is true: It can be shown [7, Chapter 13] that, for N large, with probability close to one, other distributions satisfying (4.8) are close to p^*. This fact is sometimes referred to as Entropy Concentration Theorem [16].

4.2.3 Statistical mechanics

Consider a mechanical systems composed of k particles. Let Γ denote the *phase space*, so that there is a bijection between points in Γ and possible states of the system. As in Chapter 2, we subdivide Γ in a large number N of cells ("coarse graining") so that each particle is assigned to one cell. The (thermodynamic) state of the resulting system is given by probability distributions of k particles in N cells.

If the particle are *distinguishable* , the N^k dispositions with repetition are equally likely and one talks of *Maxwell-Boltzmann statistics* (note that more than one particle may wind up in the same cell).

If the particles are *indistinguishable* , there are two types of statistics:

- *Fermi-Dirac statistics* It applies to *fermions* (electrons, protons, neutrons, also called "half-integer spin" particles). These are the elementary particles which constitute ordinary matter. This statistics is based on

[2] As Max Planck once observed, Nature appears to have a "strong preference" for situations of higher entropy

1. it is impossible for two or more particles to be in the same cell (*Pauli Exclusion Principle*[3]);
2. since the particles are indistinguishable, there are $\left(\begin{array}{c} N \\ k \end{array} \right)$ different possibilities, each with probability $\left(\begin{array}{c} N \\ k \end{array} \right)^{-1}$.

- *Bose-Einstein statistics* It applies to *bosons* (photons, pions, also called "integer spin particles"). Bosons are not subject to the Pauli exclusion principle: An unlimited number of particles may occupy the same state at the same time. As the particles are indistinguishable and order does not matter, we have a number of possibilities equal to the number of sequences $(m_1, m_2, \ldots, m_k)$, where the m_i are integers between 1 and N arranged in non decreasing order: $m_1 \leq m_2 \leq \ldots \leq m_k$. We show now that there are precisely

$$CR_N^k := \left(\begin{array}{c} N+k-1 \\ k \end{array} \right)$$

different ways (called *combinations with repetitions*). Indeed, consider the map

$$(m_1, m_2, \ldots, m_k) \to (m_1, m_2 + 1, m_3 + 2, \ldots, m_k + k - 1).$$

The second sequence is *strictly* increasing. Thus it may be viewed as a simple combination of length k of the first $n + k - 1$ natural numbers. Since the map establishes a bijection, we get the desired result. The same reult may be proven by induction via Pascal's formula (3.6).

The Maxwell-Boltzmann statistics represents the "classical" or high-temperature limit of both Fermi-Dirac and Bose-Einstein statistics.

4.3 Infinite sample spaces

There is something very disturbing in (4.3). Everything in the left-hand side depends on n. For each n, we have a *different* sample space. As the Ω_n are

[3]This principle has important implications at the microscopic level. It is also responsible for the large-scale stability of matter: Molecules cannot be pushed arbitrarily close together. The Pauli principle is the reason we do not fall through the floor.

naturally "nested", we are led to consider the space of infinite sequences:

$$\Omega := \{\omega = (a_1, a_2, a_3, \ldots), a_i = H \text{ or } T\}. \tag{4.9}$$

We are immediately faced with a major difficulty. So far, we have only dealt with finite sample spaces. Here, however, the cardinality of Ω is infinite. Let us pause for a moment and see whether we can extend the basic notion of probability space to the case when $|\Omega| = \infty$. Consider first experiments that have *denumerably many* possible outcomes.

Example 4.3.1

1. the number of *alpha* particles emitted by a radioactive substance that reach a certain counter in a certain period (as in a famous experiment by Rutherford *et al.* in 1920));

2. the number of times a web server is accessed per minute;

3. the number of cars that pass through a certain point on a freeway during a given period of time;

4. the number of particles that "scatter" off of a target in a nuclear or high energy physics experiment.

In all of these cases, we can take $\Omega = \mathbb{N}$. For such a sample space, consider a function

$$p : \Omega \to [0, 1],$$

such that

$$\sum_{i=1}^{\infty} p(\omega_i) = 1. \tag{4.10}$$

Call $p(\omega_i)$ the *probability* of outcome ω_i and call a subset $A \subseteq \Omega$ an *event*.

Example 4.3.2 In all of the cases of Example 4.3.1, the natural p is the *Poisson distribution*

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{4.11}$$

where $\lambda > 0$ is the expected number of occurrences during the given time interval. Notice that (4.11) satisfies (4.10). Consider a random variable X

taking the value $k \in \mathbb{N}$ with probability (4.11). We are interested in the expected value and the variance of X.

$$\mathbb{E}X = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} = \lambda \cdot \sum_{j=0}^{\infty} \frac{\lambda^j e^{-\lambda}}{j!} = \lambda.$$

Example 4.3.3 Another example of a probability distribution on $\mathbb{N}$ is provided by the *geometric distribution*

$$p(k) = pq^k, \quad k = 0, 1, 2, \ldots, \tag{4.12}$$

where $q = 1 - p$. The fact that $\sum_k p(k) = 1$ follows immediately from

$$\sum_{k=0}^{\infty} q^k = 1/(1-q) = 1/p.$$

This distribution arises when counting the number of consecutive successes. Suppose we toss a coin with heads=success having probability $q = 1 - p$. Then, the probability of having exactly k consecutive successes before the first failure is $p(k) = pq^k$, see Problem 70.

Define a *probability measure* $\mathbb{P}$ on $\mathcal{P}(\Omega)$ by

$$\mathbb{P}(A) := \sum_{\{\omega_i \in A\}} p(\omega_i), \quad \mathbb{P}(\emptyset) = 0. \tag{4.13}$$

It enjoyes the following properties.

1. $\mathbb{P}(\Omega) = 1$;

2. Let $(A_i)_{i=1}^{\infty}$ be a sequence of *pairwise disjoint* subsets of Ω. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

4.4 Infinite coin tosses

We now return to the infinite tosses of a fair coin, where Ω is given by (4.9). Recall that every number $a \in [0,1)$ admits a unique representation

$$a = \frac{a_1}{2} + \frac{a_2}{2^2} + \frac{a_3}{2^3} + \ldots, \quad a_i = 0, 1. \tag{4.14}$$

Thus, if we write 1 and 0 instead of H and T, respectively, we see that (4.14) sets a bijection between Ω and $[0,1)$. Hence, the cardinality of Ω is $\mathbf{c} = 2^{\aleph_0}$, namely the cardinality of the continuum! Things are worse than that. Not only is the new sample space of infinite cardinality, but, by symmetry reasons, each sequence in Ω should have the same probability. Since $\mathbb{P}(\Omega) = 1$, we must have $\mathbb{P}(a) = 0, \forall \omega \in \Omega$! We conclude that it is *impossible* to construct a probability measure on subsets of Ω starting with the probability p of the "atoms" ω. Before we get depressed, let us remember that any major difficulty in the history of mathematics, physics, etc. has led to great advances. This one represents no exception.

Observation 4.4.1 After all, we are not really interested in the probability of the single number $a \in [0,1)$. We are interested in the probability that a be in a subset of $[0,1)$. Moreover, by symmetry, $\mathbb{P}\{a \in [0, \frac{1}{2})\}$ *must* be $\frac{1}{2}$.

Definition 4.4.2 Let $X \neq \emptyset$ be a set. A collection $\mathcal{F}$ of subsets of X is called a *σ-algebra* (or sigma-algebra) if it satisfies the following properties:

1. $X \in \mathcal{F}$;
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;
3. $A_i \in \mathcal{F}, i = 1, 2, \cdots \Rightarrow (\bigcup_{i=1}^{\infty} A_i) \in \mathcal{F}$.

A collection that satisfies 1 and 2 above, but it is only closed under *finite* unions, is called an *algebra*.

Example 4.4.3 Let $X \neq \emptyset$. Then two (trivial) σ-algebras are $\mathcal{F} = \{\emptyset, X\}$ and $\mathcal{F} = \mathcal{P}(X)$. For $X = \mathbb{R}$ the real numbers, $\mathcal{B}(\mathbb{R})$ denotes the *Borel sets*. This is the smallest (intersection of two σ-algebras is a σ-algebra!) σ-algebra containing all open sets (equivalently, all intervals).

Following the great A. Kolmogorov (1933), we now take a bold step.

Definition 4.4.4 A triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space* if

1. $\Omega \neq \emptyset$;
2. $\mathcal{F}$ is a σ-algebra of subsets of Ω;
3. $\mathbb{P} : \mathcal{F} \to [0, 1]$ is such that
   (a) $\mathbb{P}(\Omega) = 1$;
   (b) $A_i \in \mathcal{F}, i = 1, 2, \ldots$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

.

Remark 4.4.5 Observe that, in the case when the cardinality of Ω is finite or countably infinite, the above definition coincides with the previous ones when $\mathcal{F} = \mathcal{P}(\Omega)$. Nevertheless, even in the case $|\Omega| < \infty$, if may be meaningful to consider a sub-σ-field $\mathcal{F} \subset \mathcal{P}(\Omega)$ of "observable events". For instance, take the case of two indistinguishable dice. Then, every event containing (i, j) must also contain (j, i). There are only 2^{21} such subsets of Ω.[4]

Let us go back to coin tossing and consider the family $\mathcal{F}_0$ of subsets of Ω in (4.9) of the form

$$\{\omega = (a_1, a_2, \ldots) | (a_1, \ldots, a_n) \in A^n\}, \quad \text{where } A^n \in \mathcal{P}(\Omega_n),\ n \geq 1. \quad (4.15)$$

It is easy to verify that $\mathcal{F}_0$ is an algebra. We can define a (finitely additive) probability $\mathbb{P}$ on $\mathcal{F}_0$ by

$$\mathbb{P}(\{\omega = (a_1, a_2, \ldots) | (a_1, \ldots, a_n) \in A^n\}) = \mathbb{P}_n(A^n). \quad (4.16)$$

[4]It was Gerolamo Cardano, who wrote the first book on probability *Liber de Ludo Aleae* sometimes between 1526 and 1563, who realized that the sample space for two die rolls should be taken to be the 36 ordered pairs (i, j) rather than the 21 unordered pairs. Cardano was a mathematician, a physician, an inventor, a chess player. His father, a lawyer and an amateur mathematician, was a friend of Leonardo da Vinci. Cardano published the general solution for cubic and quartic equations. He was first to describe typhoid fever. He researched tuberculosis, asthma, and venereal diseases. Here is an excerpt from *De malo recentiorum medicorum usu libellus*, Venice, 1536, which shows an incredibly modern thinking: "To do nothing with physic is far better than to do too much, and a physician desiring to act rightly should consider a great number of things before setting down prescriptions for the pharmacist to manufacture". He also invented several mechanical devices such as the *Cardan-shaft* used in vehicles to this day.

The extension of $\mathbb{P}$ to a σ-algebra, however, requires one of the most powerful results of measure theory. Let us introduce some notation. When a sequence of sets $\{A_n\}_{n=1}^{\infty}$ satisfies $A_{n+1} \subseteq A_n, \forall n$ and $A = \cap_n A_n$, we write $A_n \downarrow A$.

Theorem 4.4.6 (Carathéodory's Extension Theorem)Let $\mathcal{A} = \sigma(\mathcal{A}_0)$ be the smallest σ-algebra containing an algebra $\mathcal{A}_0$. Let $\mathbb{P}_0$ be a finitely additive probability defined on $\mathcal{A}_0$ satisfying the property

$$A_n \in \mathcal{A}_0, \quad A_n \downarrow \emptyset \Rightarrow \lim_n \mathbb{P}_0(A_n) = 0. \tag{4.17}$$

Then there exists a unique probability $\mathbb{P}$ on $\sigma(\mathcal{A}_0)$ such that $\mathbb{P}_{|_{\mathcal{A}_0}} = \mathbb{P}_0$.

Since (4.16) may be shown to satisfy (4.17) [31, pp.151-152], we get from Carathéodory's extension theorem the following result.

Corollary 4.4.7 Let $\mathcal{F} = \sigma(\mathcal{F}_0)$ be the smallest σ-algebra containing the algebra $\mathcal{F}_0$ of sets (5.6). Then, there exists a unique probability measure $\mathbb{P}$ on $\mathcal{F}$ such that

$$\mathbb{P}(\{\omega = (a_1, a_2, \ldots) | (a_1, \ldots, a_n) \in A^n\}) = \mathbb{P}_n(A^n), \forall A^n \in \mathcal{P}(\Omega_n), \forall n \geq 1.$$

4.5 Random variables in general

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 4.5.1 A function $X : \Omega \to \mathbb{R}$ is called a random variable if

$$\{\omega : X(\omega) \in B\} \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{4.18}$$

Observation 4.5.2 Notice that condition (4.18) ensures that the probability $\mathbb{P}(\{\omega : X(\omega) \in B\})$ has a meaning. It actually suffices to require condition (4.18) when $B = (-\infty, x]$, namely $\{\omega : X(\omega) \leq x\} \in \mathcal{F}, \forall x \in \mathbb{R}$.

Definition 4.5.3 A random variable X is called *simple* if it can be expressed in the form

$$X(\omega) = \sum_{i=1}^{m} x_i \mathbb{1}_{A_i}(\omega), \quad A_i \in \mathcal{F}, x_i \in \mathbb{R}.$$

As before, $A_i = \{\omega : X(\omega) = x_i\}$. The probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega : X(\omega) \in B\}), \quad B \in \mathcal{B}(\mathbb{R}),$$

is called *probability distribution* of X (or push-forward of $\mathbb{P}$ under X). The function

$$F_X(x) := \mathbb{P}\{\omega : X(\omega) \leq x\}$$

is called *(cumulative) distribution function* of X. It can be shown that F enjoys the following properties:

1. F is nondecreasing;

2. $\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to+\infty} F(x) = 1.$

3. F is right-continuous and admits left limit for every $x \in \mathbb{R}$.

When the distribution function F satifies

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{4.19}$$

where f is nonnegative, one says that f is the *density* of the probability distribution. If f is continuous, by the fundamental theorem of calculus, we have $\frac{dF}{dx} = f(x)$. Let X be a random variable with distribution function F_X admitting density f_X. Then

$$\mathbb{P}\{\omega : a \leq X(\omega) \leq b\} = \int_a^b f_X(t)dt.$$

Let $X(\omega) = \sum_{i=1}^m x_i \mathbb{1}_{A_i}(\cdot)$ be a simple random variable. Its expected value is defined by

$$\mathbb{E}X = \sum_{i=1}^m x_i \mathbb{P}(A_i) = \int_\Omega X(\omega)d\mathbb{P}(\omega) = \int_{\mathbb{R}} x d\mathbb{P}_X(x).$$

The expected value may be defined for a large class of random variables using abstract integration theory. But not *all* random variables admit an expected value!

Example 4.5.4 In the so called *St. Petersburg paradox*, you pay a fixed fee to enter, and then a fair coin will be tossed repeatedly until a tail first appears, ending the game. The pot starts at 1 \$ and is doubled every time a head appears. You win whatever is in the pot after the game ends. Thus you win 1 \$ if a tail appears on the first toss, 2 \$ if on the second, etc. What would be a fair price to pay for entering the game? To answer this we need to consider what would be the average payout:

$$\mathbb{E}X = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 4 + \cdots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty.$$

This means that the player should almost surely come out ahead in the long run, no matter how much he pays to enter. You should therefore play the game at *any* price if offered the opportunity. Yet, in published descriptions of the paradox, many people expressed disbelief in the result. Martin [19] quotes Ian Hacking as saying "few of us would pay even 25 \$ to enter such a game" and says most commentators would agree.

4.6 The strong law of large numbers

We are now ready to reformulate the weak law of large numbers. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is as in (4.9), and $\mathcal{F}$ and $\mathbb{P}$ are as in Corollary 4.4.7. Define on Ω the random variables

$$X_j^n(\omega) := \begin{cases} 1, & \text{if } j^{th} \text{ elementof } \omega \text{ is } H, \\ 0, & \text{if } j^{th} \text{ elementof } \omega \text{ is } T. \end{cases}, \quad j = 1, 2, \ldots, n,$$

$$S^n(\omega) := \sum_{j=1}^{n} X_j^n(\omega). \tag{4.20}$$

We can then rewrite (4.3) in the more satisfactory form

$$\lim_{n\to\infty} \mathbb{P}\{\omega : \left| \frac{S^n(\omega)}{n} - \frac{1}{2} \right| \geq \epsilon\} = 0, \quad \forall \epsilon > 0. \tag{4.21}$$

The (bad) intuition underlying the law of averages, however, leads many to believe that something much stronger must be true, namely

$$\lim_{n\to\infty} \frac{S^n(\omega)}{n} = \frac{1}{2}, \quad \forall \omega \in \Omega. \tag{4.22}$$

This is clearly false, just consider $\omega = (T, T, T, T, \ldots)$. The best we can hope is that the sequences for which (4.22) fails form a set of "small probability". Given a probability space, a property is said to hold *almost surely* (a.s.) if the event where it fails has probability zero.

Theorem 4.6.1 (Strong Law of Large Numbers - E. Borel 1909)

$$\lim_{n\to\infty} \frac{S^n(\omega)}{n} = \frac{1}{2}, \quad \text{a.s.} \tag{4.23}$$

Remark 4.6.2 The strong law of large numbers implies the weak law. Consider the random variable X taking the value 1 when the outcome of the coin tossing is H and 0 when it is T. Clearly, $\mathbb{E}X = \frac{1}{2}$. This theorem states that, excepting a set of sequences that has zero probability, the sample average converges to the probabilistic average (the expected value). This is of central importance in statistics.

Problems

Problem 19 Establish (4.2).

Problem 20 Show that also the variance $\mathbb{V}X = \mathbb{E}\{(X - \lambda)^2\}$ of a random variable X with the Poisson distribution (4.10) is equal to λ.

Problem 21 Compute $\mathbb{E}X$ when X has the geometric distribution (4.12).

Problem 22 Let $\mathcal{X} = \{1, 2, \ldots\}$, and let $\mathcal{D}(\mathcal{X})$ be the family of probability distributions on $\mathcal{X}$. Show that the solution to the following generalization of Problem 7:

$$\text{maximize} \qquad \{H(p); p \in \mathcal{D}(\mathcal{X})\} \tag{4.24}$$

$$\text{subject to} \qquad \sum_{i=1}^{\infty} ip_i = \alpha, \tag{4.25}$$

is given by a geometric distribution.

Chapter 5

Binomial distribution. The central limit theorem

5.1 The binomial distribution

Let's return once more to n coin tosses. What happens if the coin is biased? Let p be the probability of heads and $q = 1-p$ the probability of tails. Then, for $\omega \in \Omega_n$, we set

$$p(\omega) = \left(p^{\#of H\, in\, \omega}\right) \cdot \left(q^{\#of T\, in\, \omega}\right). \tag{5.1}$$

For $A \subseteq \Omega_n$, we have $\mathbb{P}(A) = \sum_{\omega \in A} p(\omega)$. Let S^n be defined as in (4.20). Then S^n has the *binomial distribution* $B(n,p)$

$$\mathbb{P}\{\omega : S^n(\omega) = k\} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}. \tag{5.2}$$

This distribution is of central importance both for the theory and for the applications. Of course, we must have

$$\sum_{k=0}^{n} \mathbb{P}\{\omega : S^n(\omega) = k\} = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} = 1.$$

Indeed, by the binomial theorem (3.4), we have

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} = (p + 1 - p)^n = 1.$$

Exercise 5.1.1 Let us compute the expected value of S^n. We get

$$\mathbb{E}S^n = \sum_{k=0}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} = np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1}(1-p)^{n-k}$$
$$= np \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} p^j (1-p)^{n-1-j} = np \cdot (p+1-p)^{n-1} = np.$$

The Poisson distribution (4.11) with parameter $\lambda = np$ can be used as an approximation to $B(n,p)$ of the binomial distribution if n is sufficiently large and p is sufficiently small (say $n \geq 20$ and $p \leq 0.05$).

Example 5.1.2 Consider the n^k dispositions with repetitions (Subsection 3.1.1) for k balls in n cells. What is the probability that a *specific cell* contains exactly $r = 0, 1, \ldots, k$ balls? The r balls may be chosen in $\left(\begin{array}{c} k \\ r \end{array} \right)$ different ways. The other $k-r$ may be placed in the remaining $n-1$ cells in $(n-1)^{k-r}$ different ways. It follows that the probability $p(r)$ that a certain specific cell contains r balls is

$$p(r) = \frac{\left(\begin{array}{c} k \\ r \end{array} \right) \cdot (n-1)^{k-r}}{n^k} = \left(\begin{array}{c} k \\ r \end{array} \right) \left(\frac{1}{n} \right)^r \left(1 - \frac{1}{n} \right)^{k-r}.$$

This is just the binomial distribution $B(k, \frac{1}{n})$.

Example 5.1.3 [10, p.138] Consider the following power supply problem. n workers use intermittently electric power. We are interested in avoiding overloads knowing that we can afford six power units. To simplify the problem, suppose that at each given time each worker has the same probability p of requiring a unit of power. If they work independently, the probability of k workers requiring power simultaneously is precisely $B(k,n,p)$. Suppose that on the average the workers use power 12 minutes per hour so that we can put $p = 0.2$. In this case, the probability of seven or more workers requiring power at the same time is $B(7,10,0.2) + \ldots + B(10,10,0.2) \approx 0.000864$. We should then expect one overload every 1157 minutes, namely one minute in twenty hours.

Example 5.1.4 [10, p.139] We are testing a new serum on cattle. We know that the normal rate of infection of a certain disease is 25%. Suppose n

animals get the serum. How can we evaluate the serum effectiveness? Assume the serum has no effect. Then the probability that k animals remain healthy is given by $B(k, n, 0.75)$. For $k = n = 10$ this probability is 0.056 and for $k = n = 12$ it reduces to 0.032. Thus, if the serum is tested on ten or twelve animals and none catches the infection, this is a strong indication that the serum is effective. Note that, without serum, the probability of out of 17 animals at most one catches the disease is approximativaly 0.05. Thus, it gives *stronger evidence* that the serum is effective if out of seventeen tested animals only one catches the infection than if out of ten all remain healthy! An even stronger evidence of the serum effectiveness is if out of 23 tested animals only two get infected. In statistical *Hypothesis Testing*, one would call *null hypothesis* the fact that the serum has no appreciable effect. The *alternate hypothesis* would then be that the serum has some significant effect. A type 1 error is then rejecting the null hypothesis when it is true. A type 2 error is to accept the null hypothesis when it is false. For instance, if only one of seventeen animal catches the disease and we reject the null hypothesis, we only have a probability of 0.05 of a type 1 error.

Example 5.1.5 Consider an urn containing N_1 white balls and N_2 black balls. A group of r balls is chosen at random. What is the probability $p(k)$ that the group so chosen contains exactly k white balls? Here $0 \leq k \leq min(N_1, r)$. The k white balls can be chosen in $\begin{pmatrix} N_1 \\ k \end{pmatrix}$ different ways. The $r-k$ black balls can be chosen in $\begin{pmatrix} N_2 \\ r-k \end{pmatrix}$ different ways. The r balls can be chosen from the $N_1 + N_2$ balls in $\begin{pmatrix} N_1 + N_2 \\ r \end{pmatrix}$ different ways. Hence

$$p(k) = \frac{\begin{pmatrix} N_1 \\ k \end{pmatrix} \begin{pmatrix} N_2 \\ r-k \end{pmatrix}}{\begin{pmatrix} N_1 + N_2 \\ r \end{pmatrix}}. \tag{5.3}$$

These probabilities form the *hypergeometric distribution* (see a particular case in Example 9.3.2). This distribution is relevant in *quality control*, namely when estimating the percentage of defective items produced from a sample of the whole production. It is also relevant in estimation of the size of an animal population from recapture data, cf. [10, pp.43-47].

5.2 The De Moivre - Laplace Central Limit Theorem

Let us go back once more to coin tossing. Let again $S^n(\omega)$ be the number of heads in the first n tosses as in (4.20). Let $\mathbb{P}$ be as in Corollary 4.4.7. By Theorem 4.6.1, we know that, excepting a set of sequences of zero probability, we have

$$\lim_{n\to\infty} \frac{S^n(\omega)}{n} = \frac{1}{2}.$$

Now suppose we toss the coin $10,000$ times and get $5,213$ heads. Should we suspect that the coin is biased? How large a *fluctuation* is nonsurprising? As observed at the end of Chapter III, for n large, the fraction of sequences of length $2n$ that have exactly n heads for a fair coin is estimated as $\frac{1}{\sqrt{\pi n}}$. It follows that for each $0 \leq l \leq n$,

$$\lim_{n\to\infty} \sum_{k=-l}^{l} \left(\begin{array}{c} 2n \\ n+k \end{array} \right) 2^{-2n} = 0.$$

Actually, recalling that the binomial coefficient $\left(\begin{array}{c} 2n \\ k \end{array} \right)$ attains its maximum at $k = n$, we have that for $m_1 < m_2$ *fixed*

$$\lim_{n\to\infty} \sum_{k=m_1}^{m_2} \left(\begin{array}{c} 2n \\ k \end{array} \right) 2^{-2n} = 0.$$

The fact that we have discussed an even number of tosses is immaterial for the asymptotic behavior of quantities. We conclude that, in the notation of Theorem 4.6.1, we have

$$\lim_{n\to\infty} \mathbb{P}(m_1 \leq S^n \leq m_2) = 0.$$

At this point, one may get the idea that, to get a nonzero limiting probability, we need to allow the size of the range to increase with n. Actually, in view of the estimate $\frac{1}{\sqrt{\pi n}}$ for the probability of n heads in $2n$ tosses, we are lead to consider the following probability

$$\mathbb{P}(\frac{n}{2} - m\sqrt{n} \leq S^n \leq \frac{n}{2} + m\sqrt{n}) = \sum_{|\frac{n}{2}-k| \leq m\sqrt{n}} \left(\begin{array}{c} n \\ k \end{array} \right) 2^{-n}.$$

The following *Central Limit Theorem* implies that the above probability has a nonzero limit as n tends to infinity.

Theorem 5.2.1 (De Moivre (1733))

$$\lim_{n\to\infty} \mathbb{P}\left\{\omega : \frac{S^n - \mathbb{E}S^n}{\sqrt{\mathbb{V}S^n}} \leq x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}} d\xi. \tag{5.4}$$

Observation 5.2.2 This theorem says that asymptotically the cumulative distribution function $F_n(x)$ of the random variables

$$X_n = \frac{S^n - \mathbb{E}S^n}{\sqrt{\mathbb{V}S^n}}$$

approaches the *Gaussian* (normal) distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}} d\xi. \tag{5.5}$$

Namely, $F_n(x) \to \Phi(x), \forall x \in \mathbb{R}$. Notice that the X_n are "normalized", namely $\mathbb{E}X_n = 0$, $\mathbb{V}X_n = 1$ for all n.

Observation 5.2.3 Theorem 5.2.1 answers the above considered question about fluctuations (*normal deviations*). Indeed, recall that $\mathbb{E}S^n = \frac{n}{2}$ and $\mathbb{V}S^n = \frac{n}{4}$ (see the proof of Theorem 4.3 and Exercise 5.1.1). Thus, by Theorem 5.2.1, for large n

$$\mathbb{P}\left\{\omega : \frac{S^n - \mathbb{E}S^n}{\sqrt{\mathbb{V}S^n}} \leq x\right\} = \mathbb{P}\left\{\left(\frac{S^n}{n} - \frac{1}{2}\right) \leq \frac{x}{2\sqrt{n}}\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}} d\xi.$$

This permits to obtain the result for $x > 0$

$$\lim_{n\to+\infty} \mathbb{P}\left\{\left|\frac{S^n}{n} - \frac{1}{2}\right| > \frac{x}{2\sqrt{n}}\right\} = 2\left(1 - \Phi(x)\right) > 0,$$

Notice that the quantity $1 - \Phi(x)$ is very small for relatively small x. For instance, $1 - \Phi(4) = 0.000032$. Tables for the normal distribution $\Phi(x)$ are provided by most textbooks, see e.g. [10, p.167]. If we get $5,213$ heads in $10,000$ tosses, the probability that this be due only to chance is approximatively 0.000064. In fact

$$\frac{5213}{10,000} - \frac{1}{2} > \frac{1}{50} = \frac{4}{2\sqrt{n}}.$$

In this case, we would have strong evidence that the coin is biased.

Observation 5.2.4 What happens when the coin is biased? Consider $\mathcal{F}_0$, the family of subsets of Ω in (4.9) of the form

$$\{\omega = (a_1, a_2, \ldots) | (a_1, \ldots, a_n) \in A^n\}, \quad \text{where } A^n \in \mathcal{P}(\Omega_n),\ n \geq 1. \tag{5.6}$$

We know that $\mathcal{F}_0$ is an algebra. Let $\mathbb{P}$ be the probability measure on $\mathcal{F} = \sigma(\mathcal{F}_0)$ (Theorem 4.4.6) such that

$$\mathbb{P}\{\omega : k \text{ heads in the first n tosses}\} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} = b(k, n, p).$$

Then, the Central Limit Theorem 5.2.1 continues to hold true as shown by Laplace in 1818.

5.3 Random walks

Consider again n coin tossing where the probability of getting H is p. Let a particle start at the origin and move a unit step in direction north-east or south-east each unit of time depending on whether the outcome is H or T, respectively. Figure 5.1 shows eight *symmetric* ($p = 1/2$) random walks.

Example 5.3.1 A random walk models the financial status of a gambler. If the step is sufficiently small, it models effectively the motion of a gas molecule (*physical Brownian motion*).

The path ω is completely specified by $(a_1, a_2, \ldots, a_n)$. Its probability is given by (5.1). The binomial distribution (5.2) describes the probability of the position of the particle after n time units. There are several questions of interest concerning random walks such as:

- number and frequency of zero crossing;
- time needed to reach a certain level;
- time spent over a certain level.

Here is a modern, rather surprising result about random walks .

Theorem 5.3.2 (Arcsine Law - P. Lévy (1939), P. Erdös and M. Kac (1947)) Let $0 < \alpha < 1$. Let $p_n(\alpha)$ be the probability that the fraction of time spent on the positive side after n tosses be less than α. Then

$$\lim_{n\to\infty} p_n(\alpha) = \frac{1}{2\pi} \arcsin(\sqrt{\alpha}) = \frac{1}{\pi} \int_0^\alpha [x(1-x)]^{-\frac{1}{2}}\, dx. \tag{5.7}$$

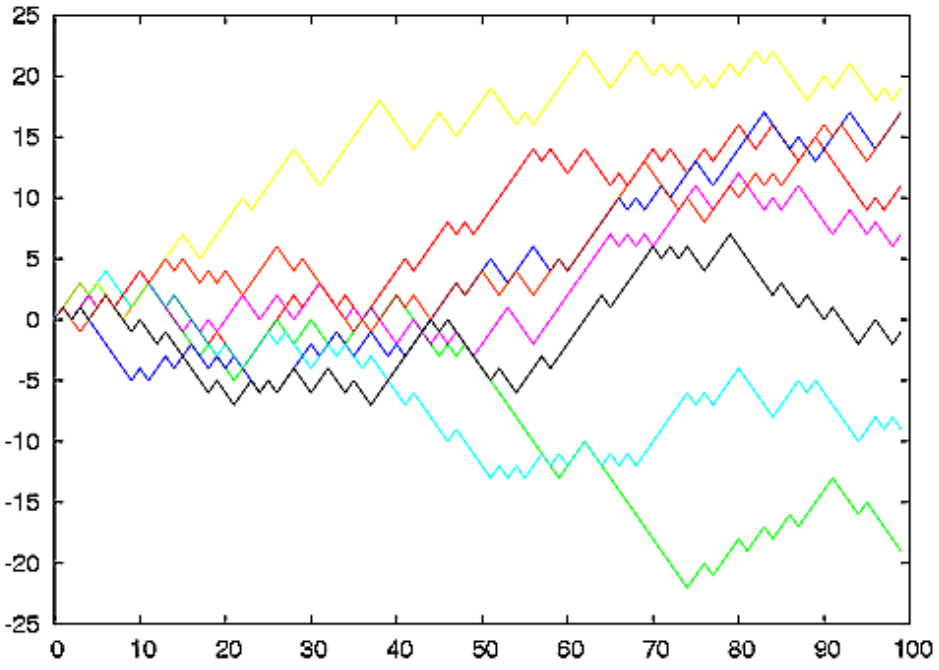

Figure 5.1: Eight random walks in one dimension.

The righ-hand side of (5.7) yields a good approximation even for few tosses such as $n = 20$. Since the integrand in (5.7) is a U shaped curve that tends to infinity for $x \searrow 0$ and $x \nearrow 1$, it is much more likely that the fraction of time spent on the positive side be close to 0 or 1 rather than to $1/2$! More explicitly, the probability that the fraction of time spent on the positive side be in the interval $(0, \epsilon)$ is much larger than say the probability that the fraction of time spent on the positive side be in the interval $(\frac{1}{2}, \frac{1}{2} + \epsilon)$. For instance, for $n \geq 20$, with probability 0.1 the particle spends 99.4% of the time on the same side.

Problems

Problem 23 Show that the variance of S^n $\mathbb{V}S^n = \mathbb{E}\{(S^n - np)^2\}$ in Exercise 5.1.1 is equal to npq.

Problem 24 Compute

$$\sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right).$$

Problem 25 Show that the mode of the binomial distribution $B(n, p)$ is the greatest integer less than or equal to $(n+1)p$. If $m = (n+1)p$ is an integer, then $m-1$ and m are both modes.

Problem 26 Prove the weak law of large numbers (4.21) when S^n has the binomial distribution (5.2).

Problem 27 In a box, there are n balls numbered from 1 to n. Suppose m are successively drawn with replacement. Compute the probability that there are exactly l balls number one when $l = 0, 1, \ldots, m$.

Problem 28 In a school class, there are 8 girls and 12 boys. A group of 8 is chosen at random. What is the probability that there are exactly 4 girls?

Problem 29 Consider the sample space $\Omega = \{\omega = (a_1, a_2), a_i = H \text{ or } T\}$ with the uniform distribution $p_u(\omega) = 1/4, \forall \omega$, corresponding to two coin tosses. Consider also the random variables

$$X_i(\omega) = \begin{cases} 1, & \text{if } a_i = H, \\ 0, & \text{if } a_i = T. \end{cases}$$

Define a new random variable on Ω by

$$\xi(\omega) = \frac{X_1(\omega) + X_2(\omega) - \mathbb{E}\{X_1 + X_2\}}{\sqrt{\mathbb{V}(X_1 + X_2)}}.$$

First, show that $\mathbb{E}\xi = 0$ and $\mathbb{V}(\xi) = 1$. Then, show that $H(p_\xi) > H(p_{X_1}) = H(p_{X_2})$, where $H(p)$ denotes the entropy of the distribution p defined in (1.11).

Chapter 6

Conditional Probability. Independence

6.1 Conditional probability

¿From now on $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ denotes a probability space with $|\Omega| < \infty$. Let $A \subseteq \Omega$ have positive probability $\mathbb{P}(A) > 0$.

Definition 6.1.1 The *conditional probability* of an event B given A (given that A has occurred) is defined by

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{6.1}$$

Observation 6.1.2 For each fixed event A, (6.1) yields a new probability measure on $\mathcal{P}(\Omega)$.

In the case when $\mathbb{P}$ is the uniform measure, we have

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}, \quad \mathbb{P}(A \cap B) = \frac{|A \cap B|}{|\Omega|} \Rightarrow \mathbb{P}(B|A) := \frac{|A \cap B|}{|A|}.$$

Example 6.1.3 Consider Problem 4 in the Questionnaire. Let p denote the probability that the other face be heads. Let A be the event "the shown face is heads" and let B be the event "the two heads coin has been drawn". Then

$$p = \mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

Let us list some basic properties of conditional probability.

1. $\mathbb{P}(A|A) = 1, \quad \mathbb{P}(\emptyset|A) = 0;$
2. $A \subseteq B \Rightarrow \mathbb{P}(B|A) = 1;$
3. $B \subseteq A \Rightarrow \mathbb{P}(B|A) = \frac{\mathbb{P}(B)}{\mathbb{P}(A)}.$

Example 6.1.4 Consider Problem 1 in the Questionnaire.

a. Let F denoted female and M denotes male. Then, $\Omega = \{FF, FM, MF, MM\}$ with the uniform distribution. Let $A = \{FF, MF\}$ and $B = \{FF\}$. We get

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

b. Let $A = \{FF, FM\}$ and $B = \{FF\}$. We get

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

c. Let $A = \{FF, FM, MF\}$ and $B = \{FF\}$. Then

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

d. *Case 1.* Parents have decided to call the first offspring Giulia if she is a female and the second offspring Lucia if both offspring are female. We take as sample space $\Omega = \{GM, MG, GL, MM\}$ with the uniform distribution. Let $A = \{GM, MG, GL\}$ and $B = \{GL\}$. We get, as in case **c.**,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Case 2. Parents choose between the two names Giulia and Lucia at random (e.g. flipping a coin). We take as sample space $\Omega = \{GM, MG, LM, ML, GL, LG, MM\}$ with the distribution:

$$p(GM) = p(MG) = p(LM) = p(ML) = p(GL) = p(LG) = \frac{1}{8},$$
$$p(MM) = \frac{1}{4}.$$

Let $A = \{GM, MG, GL, LG\}$ and $B = \{GL, LG\}$, where M denotes a son. We have $\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(B) = \frac{1}{4}$. Hence

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

6.2 Bayes' rule

From (6.1), we immediately get the *multiplication of probabilities* formula

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A). \tag{6.2}$$

The latter can be generalized by induction to n events $\{A_1, \ldots, A_n\}$ assuming $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0$

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}). \tag{6.3}$$

In the case of $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, formula (6.2) yields *Bayes' rule*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}. \tag{6.4}$$

Consider now a partition (see (1.2.4)) $\mathcal{D} = \{A_1, \ldots, A_m\}$ of Ω with $\mathbb{P}(A_i) > 0, \forall i$. Let $B \subseteq \Omega$. Observing that, since the A_i are pairwise disjoint, the same applies to $(B \cap A_i)$. Hence,

$$B = \bigcup_{i=1}^{m} (B \cap A_i) \Rightarrow \mathbb{P}(B) = \sum_{i=1}^{m} \mathbb{P}(B \cap A_i).$$

From (6.2), we now get the *law of total probability* (law of alternatives) for any event $B \subseteq \Omega$

$$\mathbb{P}(B) = \sum_{i=1}^{m} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \tag{6.5}$$

Example 6.2.1 Consider three boxes A_1, A_2 and A_3 containing $2,040$, 560 and $1,815$ light bulbs, respectively. In box A_1, 5% of the bulbs are defective. In box A_2, 10% of the bulbs are defective. In box A_3, 20% of the bulbs are defective. When a new bulb is needed, a box is chosen at random, and then, in the chosen box, a bulb is chosen at random. What is the probability that a defective bulb is chosen?

Let $A_i, i = 1, 2, 3$ be the event "box A_i is chosen" and let B be the event "a defective bulb is chosen". Then, $\mathbb{P}(A_i) = \frac{1}{3}, i = 1, 2, 3$. Moreover, $\mathbb{P}(B|A_1) = \frac{1}{20}$, $\mathbb{P}(B|A_2) = \frac{1}{10}$ and $\mathbb{P}(B|A_3) = \frac{1}{5}$. By the law of total probability (6.5), we get

$$\mathbb{P}(B) = \sum_{i=1}^{3} \mathbb{P}(B|A_i)\mathbb{P}(A_i) = \frac{1}{3}\left(\frac{1}{20} + \frac{1}{10} + \frac{1}{5}\right) = \frac{7}{60}.$$

Since $\{A, A^c\}$ constitute a partition of Ω, we get the important special case of (6.5)

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c). \tag{6.6}$$

Then, combining (6.4) with (6.5), we get a more general form of Bayes' formula

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_{j=1}^{m} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}. \tag{6.7}$$

By (6.6), the latter has the special case

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)} \tag{6.8}$$

Observation 6.2.2 In statistical applications, the A_i are called *hypotheses*. The $\mathbb{P}(A_i)$ are called *prior* probabilities and the $\mathbb{P}(A_i|B)$ are called *a posteriori* (posterior) probabilities (namely after an experiment has shown that event B has occurred).

Example 6.2.3 An urn contains two coins M_1 and M_2. While the first is a fair coin, M_2 has probability $\frac{1}{3}$ of getting H. Suppose a coin is drawn at random and tossed: We get H. What is the probability that it is the unbiased coin M_1?

We have the sample space $\Omega = \{M_1H, M_1T, M_2H, M_2T\}$. We have the probabilities

$$\mathbb{P}(M_1) = \mathbb{P}(M_2) = \frac{1}{2}, \quad \mathbb{P}(H|M_1) = \mathbb{P}(T|M_2) = \frac{1}{2}, \quad \mathbb{P}(H|M_2) = \frac{1}{3}.$$

By (6.8) we then get

$$\mathbb{P}(M_1|H) = \frac{\mathbb{P}(M_1)\mathbb{P}(H|M_1)}{\mathbb{P}(H|M_1)\mathbb{P}(M_1) + \mathbb{P}(H|M_2)\mathbb{P}(M_2)} = \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{2}} = \frac{3}{5}.$$

Example 6.2.4 Consider Problem 7 in the Questionnaire, due to Tversky and Kahneman 1982. Let A be the event "the cab was Blue" and let B be the event "the witness identifies the cab as Blue". Observe that A^c is the event "the cab was Green". By Bayes' formula (6.8), we get

$$\mathbb{P}(A|B) = \frac{(15/100)(80/100)}{(80/100)(15/100) + (20/100)(85/100)} = \frac{12}{29} \simeq 0.413.$$

In the study of Tversky and Kahneman, most subjects gave probabilities over 0.5, and some gave answers over 0.8. According to Tversky and Kahneman, it is the *representativeness heuristic* that makes people neglect the relevant base rates (only 15% blue cabs). Krosnick showed that when the order of information was reversed, the effects were mitigated.

6.3 Independence

One of the most distinctive concepts of probability is that of independence.

Definition 6.3.1 Two events A and B are called *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \tag{6.9}$$

Observation 6.3.2 Assume $\mathbb{P}(A) > 0$. Then if A and B are independent,

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

That is, the probability of B is unaffected by the fact that A has occurred.

Example 6.3.3 Let us draw at random one card from a French deck. Let A be the event "the card is an ace" and let B be the event " the suit of the card is hearts". We have

$$\mathbb{P}(A \cap B) = \frac{1}{52} = \frac{1}{13} \cdot \frac{1}{4} = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Hence, A and B are independent.

The concept of independence is far from intuitive, as Problem 38 shows.

Example 6.3.4 Let us consider Problem 5 in the questionnaire (the celebrated Monty Hall problem). If you don't change, you win with probability 1/3. If you do change, you win if and only if the box you had first picked was empty which happens with probability 2/3. Alternatively, one can argue as follows. Suppose you have picked box 1. Consider the following three events: A "box 1 contains the coin", B " either box 2 or box 3 contains the coin" and C "at least one of boxes 2 and 3 is empty". Of course, we have

$$\mathbb{P}(A) = \frac{1}{3}, \quad \mathbb{P}(B) = \frac{2}{3}, \quad \mathbb{P}(C) = 1.$$

It follows that

$$\mathbb{P}(A|C) = \frac{1}{3}, \quad \mathbb{P}(B|C) = \frac{2}{3}, \tag{6.10}$$

namely A and C are independent, and B and C are independent. Also notice that my opening of one empty box between 2 and 3 only means that C has occurred, as I know from the start where the coin is. Thus, if you are given the choice between box 1 and boxes $2, 3$ together *before or after the box has been opened*, you should choose the set $\{2, 3\}$ in both cases as B has twice the probability of occurring of A. But this is precisely what you do if you change your initial choice! Since one of boxes 2 and 3 has been opened, the other (closed) box now carries the full probability of $\frac{2}{3}$. Thus, switching doubles your chances of finding the coin. Most people, including many with a scientific education, are unable to find or even accept this solution. This is a powerful illustration of how bad our probabilistic intuition may turn out to be when facing some problems. It may even lead to some sort of hallucinations that overwhelm our rational thinking.

Definition 6.3.5 Given a random variable X, a *median* is any number satisfying

$$\mathbb{P}\{X \geq m\} \geq \frac{1}{2}, \quad \mathbb{P}\{X \leq m\} \geq \frac{1}{2}. \tag{6.11}$$

Example 6.3.6 In the birthday problem (3.1.4), the (unique) median is $m = 23$.

Exercise 6.3.7 Let $X_1, \ldots, X_n$ be mutually independent random variables. Suppose they all have the uniform distribution

$$p_{X_i}(k) = \mathbb{P}\{X_i = k\} = \frac{1}{N}, \quad k = 1, 2, \ldots, N.$$

Let $Y : \omega \to \max\{X_1(\omega), \ldots, X_n(\omega)\}$. Find the distribution and the expected value of Y.

Notice that, by independence of the X_i, we have

$$\mathbb{P}\{Y \leq k\} = \left(\frac{k}{N}\right)^n.$$

It then follows

$$p_Y(k) = \mathbb{P}\{Y = k\} = \mathbb{P}\{Y \leq k\} - \mathbb{P}\{Y \leq k-1\} = \frac{k^n - (k-1)^n}{N^n}.$$

We now find $\mathbb{E}Y$. We have

$$\mathbb{E}Y = \sum_{k=1}^{N} k p_Y(k) = \frac{1}{N^n} \sum_{k=1}^{N} \left[k^{n+1} - k(k-1)^n\right]$$

$$= \frac{1}{N^n} \sum_{k=1}^{N} \left[k^{n+1} - (k-1)^{n+1} - (k-1)^n\right] = \frac{1}{N^n} \left[N^{n+1} - \sum_{k=1}^{N} (k-1)^n\right].$$

For N large,

$$\sum_{k=1}^{N} (k-1)^n \approx \int_0^N x^n dx = \frac{N^{n+1}}{n+1}.$$

We conclude that, if N is large,

$$\mathbb{E}Y \approx N \frac{n}{n+1}. \tag{6.12}$$

To get an idea of a possible application, consider a town with $1,000$ cars whose license plates are numbered $1, 2, \ldots, 1,000$. Suppose we read $n = 10$ license numbers. Then, by (6.12), the expected value of the largest plate number is

$$1,000 \cdot \frac{10}{11} \approx 910.$$

(The median is 934). In statistics, the observed maximum in a sample may be used to estimate the unknown number N.This method was employed during World War II to estimate the enemy production.

Definition 6.3.8 Let $\mathcal{F}_1, \mathcal{F}_2$ be two algebras (Definition 4.4.2). They are called *independent* if all pairs $(A_1, A_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ are independent. It is

not difficult to show that, if A and B are independent, so are the algebras $\mathcal{F}_1 = \{A, A^c, \emptyset, \Omega\}$ and $\mathcal{F}_2 = \{B, B^c, \emptyset, \Omega\}$. Events $A_1, \ldots, A_n$ are called *independent* if, for all $1 \leq i_1 < i_2 < \cdots < i_k \leq n, k = 1, \ldots, n$, they satisfy the condition

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Finally, the algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are called independent if $(A_1, \ldots, A_n), A_i \in \mathcal{F}_i$ are independent.

Observation 6.3.9 Pairwise independence does not imply independence. For instance, take $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ with the uniform distribution. Then $A = \{\omega_1, \omega_2\}$, $B = \{\omega_1, \omega_3\}$ and $C = \{\omega_1, \omega_4\}$ are pairwise independent but

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{4} \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$

Let's go back to n coin tosses with the binomial distribution (5.2). Consider the event $A_k = \{\omega : a_k = H\}$ and let $\mathcal{F}_k = \{A_k, A_k^c, \emptyset, \Omega\}$. We have $\mathbb{P}(A_k) = p, \mathbb{P}(A_k^c) = q$. Moreover, for $k \neq l$, we have

$$\mathbb{P}(A_k \cap A_l) = p^2, \quad \mathbb{P}(A_k \cap A_l^c) = pq, \quad \mathbb{P}(A_k^c \cap A_l^c) = q^2.$$

Thus the $\mathcal{F}_k$ are independent! This is the reason why we talk about "n independent tosses" of a coin. Consider now n probability spaces

$$(\Omega_1, \mathcal{F}_1, \mathbb{P}_1), \ldots (\Omega_n, \mathcal{F}_n, \mathbb{P}_n),$$

with $|\Omega_i| < \infty$. Then we can define a new sample space

$$\Omega = \Omega_1 \times \Omega_2 \cdots \times \Omega_n,$$

where $\omega \in \Omega$, $\omega = (a_1, a_2, \ldots, a_n), a_i \in \Omega_i$. Define also

$$\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \cdots \otimes \mathcal{F}_n$$

to be the algebra generated by sets of the form $A = A_1 \times A_2 \times \cdots \times A_n, A_i \in \mathcal{F}_i$. Define $p(\omega) = p_1(a_1) \cdot p_2(a_2) \cdots p_n(a_n)$ and

$$\mathbb{P}(A) = \sum_{a_1 \in A_1, \ldots, a_n \in A_n} p_1(a_1) \cdot p_2(a_2) \cdots p_n(a_n). \tag{6.13}$$

Then $\mathbb{P}(\Omega) = 1$ and $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space called *direct product* of the spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$. It is easy to verify that the events $B_1 = \{\omega : a_1 \in A_1\}, \ldots, B_n = \{\omega : a_n \in A_n\}$ are independent with respect to $\mathbb{P}$. The same applies to the subalgebras $\mathcal{B}_k := \{B_k : B_k = \{\omega : a_k \in A_k\}, A_k \in \mathcal{F}_k\}$.

We conclude that our standard probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ for n coin tosses

$$\omega = (a_1, \ldots, a_n), a_i = H \text{ or } T, \quad p(\omega) = \left(p^{\#of H\, in\, \omega}\right) \cdot \left(q^{\#of T\, in\, \omega}\right),$$

can be thought of as the direct product of the probability spaces $(\Omega_i, \mathcal{P}(\Omega_i), \mathbb{P}_i)$, where $\Omega_i = \{H, T\}$, $\mathcal{P}(\Omega_i) = \{H, T, \emptyset, \Omega_i\}$ and $\mathbb{P}(\{H\}) = p, \mathbb{P}(\{T\}) = q$. We are now fully justified to say that (6.13) models n *independent coin tosses*.

Exercise 6.3.10 Let $A_1, \ldots, A_n$ be independent. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} \mathbb{P}(A_i^c).$$

If A and B are independent, so are A^c and B^c. Indeed, we have

$$\begin{aligned}\mathbb{P}(A^c \cap B^c) = \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B) = 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)) \\ = 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(B) = (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)).\end{aligned}$$

The same applies to the general case. Hence,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right)^c\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^{n} A_i^c\right) = 1 - \prod_{i=1}^{n} \mathbb{P}(A_i^c).$$

Notice that it follows that the probability that none of the events occur is

$$p = \prod_{i=1}^{n} \left(1 - \mathbb{P}(A_i)\right).$$

6.4 Dependence of random variables

Definition 6.4.1 Let X and Y be two random variables defined on the same probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ and taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. They are called *independent* if they satisfy

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y), \quad \forall x \in \mathcal{X}, \forall y \in \mathcal{Y},$$

where we have used the shorthand notation $\mathbb{P}(X = x)$ to denote $\mathbb{P}(\omega : X(\omega) = x)$. The function p_{XY} defined on $\mathcal{X} \times \mathcal{Y}$ by

$$p_{XY}(x, y) = \mathbb{P}(X = x, Y = y) \tag{6.14}$$

is called *joint probability distribution of X and Y*. Independence is then the property that

$$p_{XY}(x, y) = p_X(x) \cdot p_Y(y),$$

namely p_{XY} factors into the product of the two *marginal distributions* p_X and p_Y. Notice that the following properties hold:

$$p_{XY}(x, y) \geq 0, \quad \sum_{j,k} p_{XY}(x_j, y_k) = 1,$$

$$\sum_k p_{XY}(x_j, y_k) = p_X(x_j), \quad \sum_j p_{XY}(x_j, y_k) = p_Y(y_k).$$

The notion of joint probability distribution readily generalizes to n random variables with similar properties.

Example 6.4.2 Consider n coin tosses with $\omega = (a_1, a_2, \ldots, a_n)$. Let

$$X_i(\omega) = \begin{cases} 1, & a_i = H, \\ 0, & a_i = T \end{cases}$$

Then the X_i are independent.

Definition 6.4.3 Let X and Y be as in the above problem. The *covariance* of X and Y is defined by

$$\mathrm{Cov}\,(X, Y) = \mathbb{E}\{(X - \mathbb{E}X)(Y - \mathbb{E}Y)\} = \mathbb{E}\{XY\} - \mathbb{E}X \cdot \mathbb{E}Y. \tag{6.15}$$

The *correlation coefficient* of X and Y is defined by

$$\rho(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sqrt{\mathbb{V}X \cdot \mathbb{V}Y}}. \tag{6.16}$$

It follows from Problem 40 that two independent random variables are *uncorrelated* ($\mathrm{Cov}\,(X, Y) = 0$). What about the converse?

Example 6.4.4 Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ with the uniform distribution $p_u(\omega_i) = \frac{1}{3}$. Define the random variable $X : \Omega \to \mathbb{R}$ as $X(\omega_1) = 0$, $X(\omega_2) = \frac{\pi}{2}$, and $X(\omega_3) = \pi$. Clearly the probability distribution of X $p_X(x_i)$ is also uniform on $\mathcal{X} = \{0, \frac{\pi}{2}, \pi\}$. Define the random variables on $\mathcal{X}$

$$Y(x) = \sin x, \quad Z(x) = \cos x.$$

Then,

$$\mathbb{E}Y = \sum_{i=1}^{3} Y(x_i)p_X(x_i) = \frac{1}{3} \cdot (0 + 1 + 0) = \frac{1}{3},$$

$$\mathbb{E}Z = \sum_{i=1}^{3} Z(x_i)p_X(x_i) = \frac{1}{3} \cdot (1 + 0 + (-1)) = 0.$$

Hence,

$$\begin{aligned} \mathrm{Cov}\,(Y, Z) &= \mathbb{E}\{(Y - \mathbb{E}Y)(Z - \mathbb{E}Z)\} = \mathbb{E}\{YZ\} \\ &= \sum_{i=1}^{3} Y(x_i)Z(x_i)p_X(x_i) = \frac{1}{3}(0 \cdot 1 + 1 \cdot 0 + 0 \cdot (-1)) = 0. \end{aligned}$$

Thus Y and Z are uncorrelated. They are, however, not independent as

$$\mathbb{P}\{Y = 1, Z = 1\} = 0 \neq \frac{1}{9} = \mathbb{P}\{Y = 1\} \cdot \mathbb{P}\{Z = 1\}.$$

Actually, the dependence between these two random variables is very strong since

$$Y^2 + Z^2 \equiv 1.$$

Thus uncorrelation does not imply independence [1].

Observation 6.4.5 It is easy to verify

$$\mathbb{V}(X + Y) = \mathbb{V}X + \mathbb{V}Y + 2\mathrm{Cov}\,(X, Y).$$

Hence, for independent variables, variances can be added

$$\mathbb{V}(X + Y) = \mathbb{V}X + \mathbb{V}Y. \tag{6.17}$$

For instance, to compute the variance of S^n in Exercise 5.1.1, it suffices to observe that the variance must be n times the variance of the random variable taking the value 1 with probability p. The latter has expected value p and variance $\mathbb{V}X = (1 - p)^2 p + (-p)^2(1 - p) = (1 - p)p$.

[1] It can be shown that uncorrelation implies independence for *jointly Gaussian* random variables [31, p.234].

6.5 Conditional expectation

Let X and Y be two random variables defined on the same probability space and taking values in the finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Definition 6.5.1 The *conditional expectation* of X given that Y has taken the value $y \in \mathcal{Y}$ is the function $g(\cdot)$ defined on $\mathcal{Y}$ by

$$g(y) = \mathbb{E}(X|Y = y) = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x|Y = y). \tag{6.18}$$

The *conditional expectation of X given Y* is the random variable $g(Y)$ which takes the value $\mathbb{E}(X|Y = y)$ with probability $p_Y(y)$. If we have random variables $X, Y_1, \ldots, Y_n$ taking values in $\mathcal{X}$, $\mathcal{Y}_1$, $\ldots$, $\mathcal{Y}_n$, we can define

$$g_n(y_1, \ldots, y_n) = E(X|Y_1 = y_1, \ldots, Y_n = y_n) = \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x|Y_1 = y_1, \ldots, Y_n = y_n).$$

Moreover, $\mathbb{E}(X|Y_1, \ldots, Y_n) := g_n(Y_1, \ldots, Y_n)$. If A is an event, the conditional probability of A given $Y_1, \ldots, Y_n$ is defined by

$$\mathbb{P}(A|Y_1, \ldots, Y_n) := E(\mathbb{1}_A|Y_1, \ldots, Y_n).$$

Observation 6.5.2 If X is independent of $\{Y_1, \ldots, Y_n\}$ (i.e. the corresponding algebras are independent in the sense of Definition 6.3.8), then $\mathbb{P}(X = x|Y_1 = y_1, \ldots, Y_n = y_n) = \mathbb{P}(X = x)$. We then get $\mathbb{E}(X|Y_1 = y_1, \ldots, Y_n = y_n) = EX$. It also follows that

$$\mathbb{E}(X|Y_1, \ldots, Y_n) = \mathbb{E}X. \tag{6.19}$$

Observation 6.5.3 Let f be any function. Then

$$\begin{aligned}\mathbb{E}(X \cdot f(Y)|Y = \bar{y}) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} x \cdot f(y) \cdot \mathbb{P}(X = x|Y = \bar{y}) \\ &= \sum_{x \in \mathcal{X}} x \cdot f(\bar{y}) \cdot \mathbb{P}(X = x|Y = \bar{y}) = f(\bar{y}) \cdot \mathbb{E}(X|Y = \bar{y}).\end{aligned}$$

This implies

$$\mathbb{E}(X \cdot f(Y)|Y) = f(Y) \cdot \mathbb{E}(X|Y). \tag{6.20}$$

In particular, choosing $X \equiv 1$, we get

$$\mathbb{E}(f(Y)|Y) = f(Y). \tag{6.21}$$

Observation 6.5.4 For each n-tuple $(y_1, \dots, y_n) \in (\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n)$, $\mathbb{E}(X|Y_1 = y_1, \dots, Y_n = y_n) := g_n(y_1, \dots, y_n)$ is simply the expected value of X with respect to the conditional probability distribution $\mathbb{P}(X = x|Y_1 = y_1, \dots, Y_n = y_n)$. Hence, it enjoys all the properties of an expectation, see Proposition 1.3.5. This in turn implies that $\mathbb{E}(X|Y_1, \dots, Y_n) = g_n(Y_1, \dots, Y_n)$ enjoys the following properties:

1. $\mathbb{E}(\alpha_1 X_1 + \alpha_2 X_2|Y_1, \dots, Y_n) = \alpha_1 \mathbb{E}(X_1|Y_1, \dots, Y_n) + \alpha_2 \mathbb{E}(X_2|Y_1, \dots, Y_n)$;
2. $X \equiv c \Rightarrow \mathbb{E}(X|Y_1, \dots, Y_n) \equiv c$;
3. $X \geq 0 \Rightarrow \mathbb{E}(X|Y) \geq 0$.

Moreover, by (1.5.6), we have

Proposition 6.5.5 Let φ be convex and let $X, Y_1, \dots, Y_n$ be as above. Then

$$\varphi\left(\mathbb{E}(X|Y_1 = y_1, \dots, Y_n = y_n)\right) \leq \mathbb{E}(\varphi(X)|Y_1 = y_1, \dots, Y_n = y_n). \tag{6.22}$$

This implies

$$\varphi\left(\mathbb{E}(X|Y_1, \dots, Y_n)\right) \leq \mathbb{E}(\varphi(X)|Y_1, \dots, Y_n). \tag{6.23}$$

Another fundamental property is the following.

Theorem 6.5.6 (*Iterated conditioning*)We have

$$\mathbb{E}(\mathbb{E}(X|Y_1, \dots, Y_n)) = \mathbb{E}X. \tag{6.24}$$

Moreover, when $1 \leq i_1 < \cdots < i_k \leq n$, we have

$$\mathbb{E}(\mathbb{E}(X|Y_1, \dots, Y_n)|Y_{i_1}, \dots, Y_{i_k}) = \mathbb{E}(X|Y_{i_1}, \dots, Y_{i_k}). \tag{6.25}$$

Proof . We prove (6.24) with $n = 1$. We have

$$\begin{aligned}\mathbb{E}(\mathbb{E}(X|Y)) &= \sum_{y \in \mathcal{Y}} \mathbb{E}(X|Y = y) p_Y(y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x|Y = y) p_Y(y) \\ &= \sum_{x \in \mathcal{X}} x \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x|Y = y) p_Y(y) = \sum_{x \in \mathcal{X}} x \cdot p_X(x) = \mathbb{E}X.\end{aligned}$$

The case of general n is proved analogously. Proving property (6.25) is left as an exercise. □

Remark 6.5.7 Everything we have done in this section extends without much effort to the case of random variables with denumerable state space provided $E|X| < \infty$, see [5, pp.39-41].

6.6 Estimation

6.6.1 Least squares estimation

Consider two random variables X and Y on the same probability space taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Suppose X represents the phenomenon of interest, but we can only observe Y. If X and Y are not independent, we expect that the value of Y allows us to make inference on the value of X. For any $f : \mathbb{R} \to \mathbb{R}$, the random variable $f(Y)$ is called an *estimator* of X. We say that $f^*(Y)$ is an *optimal least-squares estimator* if it satisfies

$$\mathbb{E}\{(X - f^*(Y))^2\} = \inf_f \mathbb{E}\{(X - f(Y))^2\}.$$

Theorem 6.6.1 The unique optimal least-squares estimator is given by the conditional expectation $f^*(Y) = \mathbb{E}(X|Y)$.

Proof . By (6.20) and (6.24), we have

$$\begin{aligned}\mathbb{E}\left((X - \mathbb{E}(X|Y)) \cdot f(Y)\right) &= \mathbb{E}\left((X \cdot f(Y) - \mathbb{E}(X \cdot f(Y)|Y)\right) \\ &= \mathbb{E}(X \cdot f(Y)) - \mathbb{E}(X \cdot f(Y)) = 0.\end{aligned}$$

Hence, for any estimator $f(Y)$, we have

$$\begin{aligned}\mathbb{E}\{(X - f(Y))^2\} &= \mathbb{E}\{(X - \mathbb{E}(X|Y) + \mathbb{E}(X|Y) - f(Y))^2\} \\ = \mathbb{E}\{(X - \mathbb{E}(X|Y))^2\} &+ 2\mathbb{E}\{(X - E(X|Y)) \cdot (\mathbb{E}(X|Y) - f(Y))\} \\ &+ \mathbb{E}\{(\mathbb{E}(X|Y) - f(Y))^2\} \\ &= \mathbb{E}\{(X - \mathbb{E}(X|Y))^2\} + \mathbb{E}\{(\mathbb{E}(X|Y) - f(Y))^2\}.\end{aligned}$$

Since both terms are nonnegative, we get the result. □

Suppose now we confine ourselves to *linear estimators* of the form

$$f(Y) = a + bY, \quad a, b \in \mathbb{R}.$$

Observe that the function

$$\varphi(a, b) = \mathbb{E}\{(X - (a + bY))^2\} = \mathbb{E}\{(X - bY)^2\} - 2aE\{(X - bY)\} + a^2$$

is quadratic in a. Thus, *for each fixed* b, the infimum is attained at

$$a^*(b) = E\{X - bY\}.$$

Consider now the function

$$\psi(b) := \varphi(a^*(b), b) = \mathbb{E}\{(X - bY)^2\} - \mathbb{E}\{(X - bY)\}^2.$$

A simple calculation yields

$$\psi(b) = b^2 \left[\mathbb{E}\{Y^2\} - (EY)^2\right] - 2b \left[\mathbb{E}\{XY\} - \mathbb{E}X \cdot \mathbb{E}Y\right] + \mathbb{E}\{X^2\} - (\mathbb{E}X)^2.$$

Once more, this is a quadratic function whose minimum is attained at

$$b^* = \frac{\mathbb{E}\{XY\} - \mathbb{E}X \cdot \mathbb{E}Y}{\mathbb{E}\{Y^2\} - (\mathbb{E}Y)^2} = \frac{\mathrm{Cov}\,(X, Y)}{\mathbb{V}Y}.$$

We conclude that the optimal linear least-squares estimator is given by

$$a^* + b^*Y = \mathbb{E}X + \frac{\mathrm{Cov}\,(X, Y)}{\mathbb{V}Y} \left[Y - \mathbb{E}Y\right]. \tag{6.26}$$

The least-squares estimation error is given by

$$\delta^* = \mathbb{E}\{(X - a^* - b^*Y)^2\} = \mathbb{V}X - \frac{\mathrm{Cov}\,(X, Y)^2}{\mathbb{V}Y} = \mathbb{V}X \left[1 - \rho(X, Y)^2\right].$$

Note that $|\rho(X, Y)| = 1 \Rightarrow \delta^* = 0$. By the representation (1.7), this implies that $X = a^* + b^*Y$. In other words, X can be perfectly estimated by a linear estimator. If $\rho(X, Y) = 0$, i.e. the random variables are uncorrelated, then the best linear estimator reduces to a constant $f^*(Y) = \mathbb{E}X$ and the corresponding estimation error is $\delta^* = \mathbb{V}X$.

6.6.2 L^1 estimation

Suppose we are looking instead for a constant a^* that gives the best L^1 approximation of X, namely that solves the problem

$$\inf_{a \in \mathbb{R}} \mathbb{E}|X - a|. \tag{6.27}$$

Proposition 6.6.2 Any median (Definition 6.3.5) for X is a solution of Problem 6.27.

Proof . For $a \in \mathbb{R}$, we have

$$f(a) := \mathbb{E}|X - a| = \sum_{i=1}^{n} |X(\omega_i) - a| p(\omega_i).$$

Clearly, f is a convex function (a linear combination of convex functions with nonnegative coefficients is convex). Consider the two events

$$A_+ := \{\omega_i | X(\omega_i) > a\}, \quad A_- := \{\omega_i | X(\omega_i) < a\}.$$

Then $f(a)$ may be expressed in the form

$$f(a) = \sum_{\omega_i \in A+} (X(\omega_i) - a)p(\omega_i) + \sum_{\omega_i \in A_-} (a - X(\omega_i))p(\omega_i). \tag{6.28}$$

Observe that the sets A_+ and A_- depend on a. Nevertheless, since there are only finitely many ω_i, for h sufficiently small $A_+(a \pm h) = A_+(a)$. Similarly for A_-. Hence we can use expression (6.28) to compute the derivative of f. We get the optimality condition

$$0 = \frac{df}{da} = -\sum_{\omega_i \in A_+} p(\omega_i) + \sum_{\omega_i \in A_-} p(\omega_i).$$

This is equivalent to

$$\mathbb{P}(A_+(a^*)) = \mathbb{P}(A_-(a^*)). \tag{6.29}$$

Consider now $A_0(a^*) = \{\omega_i | X(\omega_i) = a^*\}$. Using (6.29), we get

$$\mathbb{P}\{A_+(a^*) \cup A_0(a^*)\} = \mathbb{P}\{\omega_i | X(\omega_i) \geq a^*\} = 1 - \mathbb{P}\{A_-(a^*)\} = 1 - \mathbb{P}\{A_+(a^*)\}.$$

Let us show that this implies that $\mathbb{P}\{\omega_i | X(\omega_i) \geq a^*\} \geq \frac{1}{2}$. Indeed

$$\begin{aligned}\mathbb{P}\{A_+(a^*) \cup A_0(a^*)\} = 1 - \mathbb{P}\{A_+(a^*)\} &\geq 1 - \mathbb{P}\{A_+(a^*)\} - \mathbb{P}\{A_0(a^*)\} \\ &= 1 - \mathbb{P}\{A_+(a^*) \cup A_0(a^*)\}.\end{aligned}$$

Similarly for the other inequality. We conclude that any median is a solution of Problem 6.27. □

6.6.3 Linear regression

Let us apply Proposition 6.6.2 to l^1 *linear regression* through a point. Suppose we guess a linear dependence between the random variables X and Y. Let $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ be realizations of X and Y, respectively. We seek a straight line $y(x) = a + bx$ through a given point (x_0, y_0). Without

loss of generality, we can take $(x_0, y_0) = (0, 0)$. We are left with the task of finding b such that

$$\sum_{i=1}^{n} |y_i - bx_i|$$

is minimized. Since the $x_i \neq 0$, we can rewrite the above criterion as

$$\sum_{i=1}^{n} |y_i - bx_i| = \sum_{i=1}^{n} |\frac{y_i}{x_i} - b| \cdot |x_i|.$$

Define $\Omega = \{1, 2, \ldots, n\}$, and

$$p(i) = \frac{|x_i|}{\sum_{i=1}^{n} |x_i|}.$$

Then, minimizing $\sum_{i=1}^{n} |y_i - bx_i|$ is equivalent to minimizing

$$\sum_{i=1}^{n} |y_i - bx_i| \cdot \frac{1}{\sum_{i=1}^{n} |x_i|} = \sum_{i=1}^{n} |\frac{y_i}{x_i} - b| \cdot p(i).$$

This is just problem (6.27) for the random variable $\frac{Y}{X}$. We know that medians are solutions.

Example 6.6.3 Let $(x_1, x_2, x_3, x_4) = (1, 2, 3, 4)$ and $(y_1, y_2, y_3, y_4) = (2, 3, 5, 9)$. Here $p(i) = \frac{i}{10}$. We get

$$\left(\frac{y_1}{x_1}, \frac{y_2}{x_2}, \frac{y_3}{x_3}, \frac{y_4}{x_4}\right) = \left(2, \frac{3}{2}, \frac{5}{3}, \frac{9}{4}\right).$$

Ordering these values and associating to them the corresponding probability, we get Table 6.1. Any $\frac{5}{3} \leq m \leq 2$ is a median and therefore solves the problem.

Table 6.1: Distribution of $\frac{Y}{X}$

$\frac{y_i}{x_i}$	$\frac{3}{2}$	$\frac{5}{3}$	2	$\frac{9}{4}$
$p(i)$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{4}{10}$

Problems

Problem 30 Consider the equally likely 4^4 dispositions with repetitions of 4 balls in 4 cells. If we know that the first two balls are in different cells, what is the probability that one of the cells contains exactly three balls?

Problem 31 Suppose the random variable X has the geometric distribution (4.12). Show that for all $k, k_0 \geq 1$, we have

$$\mathbb{P}(X = k + k_0 | X > k_0) = \mathbb{P}(X = k).$$

Problem 32 An epidemic disease is spreading through a country. The country is divides into four states NW, NE, SW and SE having population $1,000,000$, $2,000,000$, $3,000,000$ and $4,000,000$, respectively. It is known that the percentage of infected people in each state is 15%, 7%, 5% and 2%, respectively. A person from this country travels to another country. Find the probability that this person be infected.

Problem 33 There is a country, in the complement of our past and future light cones, called *Schönesland*. The politicians there are divided into two coalitions: The conservative and the liberals. The conservative have 95% of crooks and 5% of idiots. The liberals have 90% of idiots and 10% of crooks. The news report today that one politician was discovered bribing a judge. What is the probability that this politician is a conservative?

Problem 34 In a population that has the same number of male and female, there are 5 male out of 100 that are colorblind and 25 female every 10.000. We choose a person at random from this population. We discover that the person is in fact colorblind. What is the probability that it is a male?

Problem 35 Consider random permutations of the four letters (a, b, c, d). Let A be the event "a precedes b" and B be the event "c precedes d". Prove that A and B are independent.

Problem 36 Suppose a fair coin is tossed twice. Let A be the event "the first toss is head" and B is the event "the two outcomes are equal". Are A and B independent?

Problem 37 Let A and B be independent events. Prove that so are the algebras $\mathcal{F}_1 = \{A, A^c, \emptyset, \Omega\}$ and $\mathcal{F}_2 = \{B, B^c, \emptyset, \Omega\}$.

Problem 38 In a family there are n children. Assume equal probability of birth for the two sexes at each birth. Let A be the event "the family has children of both sexes" and let B be the event "there is at most one daughter". Discuss the independence of events A and B for $n \geq 2$.

Problem 39 Let X_1 and X_2 be two independent Poisson random variables with mean λ_1 and λ_2, respectively. Prove that $Y = X_1 + X_2$ also has a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$.

Problem 40 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $|\Omega| < \infty$. Let X and Y be two indedpendent random variables defined on Ω. Prove the orthogonality relation

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

Problem 41 Let X and Y be independent random variables with state space $\mathcal{X} = \mathcal{Y} = \mathbb{Z}$. Let $Z = X + Y$. Express the probability distribution of Z in terms of p_X and p_Y. Check that your answer agrees with the answer to Problem 39 when the random variables have a Poisson distribution.

Problem 42 Two players play a game flipping a coin several times. At each coin toss, the probabilities of winning are p for player A and $q = 1 - p$ for player B. The capital of A increases by one unit each time A wins, and remains the same each time A loses. To make the game more interesting, the players agree that they will play as many games as the number of cars passing in one minute on a freeway they can observe. Suppose the latter are descibed by a Poisson random variable T with parameter λ. Let $X_i, i \geq 1$ be independent Bernoulli random variables taking the values 1 and 0 with probability p and $q = 1 - p$, respectively. Then, the capital S of player A at the end of the game is given by

$$S(\omega) = X_1(\omega) + X_2(\omega) + \ldots + X_{T(\omega)}(\omega).$$

Assume that T is independent of $X_i, i \geq 1$. Show that S has a Poisson distribution with parameter $p\lambda$.

Problem 43 Consider again the game of Problem 42. Observe that S is the number of wins for player A and $Z := T - S$ is the number of losses for A. Show that S and Z are independent.

Problem 44 Consider the following example from [13, p.143]. Out of 60 people, the number of those that have cancer or not , smoke or not is reported in Table 6.2. We choose a person at random from the group. Consider the following two random variables: $C(\omega) = 1$ if this person has cancer and 0 if not, $S(\omega) = 1$ if this person smokes and 0 if not. Find the joint distribution of C and S and their marginal distributions.

Table 6.2: Cancer and smoking.

	Not smoke	Smoke	Total
Not cancer	40	10	50
Cancer	7	3	10
Totals	47	13	60

Problem 45 Let X and Y be independent, identically distributed random variables with state space $\mathcal{X} = \mathcal{Y} = \mathbb{Z}$. Let $Z = X + Y$. Assuming $\mathbb{E}X = \mathbb{E}Y < \infty$, compute $\mathbb{E}(X|Z)$.

Problem 46 Prove (6.25) when $n = 2$.

Problem 47 Consider the following generalization of Problem 45. Let X_1 and X_2 be independent Poisson random variables with mean λ_1 and λ_2, respectively. Let $Y = X_1 + X_2$. Compute $\mathbb{E}(X_1|Y)$.

Problem 48 Prove that

$$a^* = \sum_{i=1}^{n} f(x_i)p_i.$$

solves the least squares problem

$$\min_{a\in\mathbb{R}} \sum_{i=1}^{n} \left(f(x_i) - a\right)^2 p_i.$$

Problem 49 Suppose we have guessed a linear dependence between X and Y. Find a straight line through $(x_0, y_0) = (0, 0)$ that provides the best l^1 approximation given the realizations $(x_1, \ldots, x_n) = (1, 2, 3, 4, 6)$ and $(y_1, \ldots, y_n) = (1, 4, 7, 9, 11)$.

Chapter 7

Markov chains

7.1 A migration model

Consider a graph with five nodes (representing cities). Suppose there are one million people distributed among the five cities, see Figure 7.1.
¿From year $t = 0$, the population moves every following year according to the following proportions

Table 7.1: Migration proportions

	stay	clokwise migration	counterclockwise migration
fraction	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Let $\pi_i(t), i = 1, 2, 3, 4, 5$ be the fraction of population of city i in year t. Suppose initially all people live in city n.1. Then

$$\pi_i(0) = \begin{cases} 1, & i = 1, \\ 0, & i \neq 1 \end{cases}$$

Let us compute the fractions of population in the following years

$$\pi_i(1) = \begin{cases} \frac{1}{2}, & i = 1, \\ \frac{1}{4}, & i = 2,\ i = 5, \\ 0, & i = 3,\ i = 4. \end{cases} \qquad \pi_i(2) = \begin{cases} \frac{3}{8}, & i = 1, \\ \frac{1}{4}, & i = 2,\ i = 5, \\ \frac{1}{16}, & i = 3,\ i = 4. \end{cases}$$

$$\pi_i(3) = \begin{cases} \frac{20}{64}, & i = 1, \\ \frac{15}{64}, & i = 2,\ i = 5, \\ \frac{7}{64}, & i = 3,\ i = 4. \end{cases} \qquad \pi_i(4) = \begin{cases} \frac{70}{256}, & i = 1, \\ \frac{57}{256}, & i = 2,\ i = 5, \\ \frac{36}{256}, & i = 3,\ i = 4. \end{cases}$$

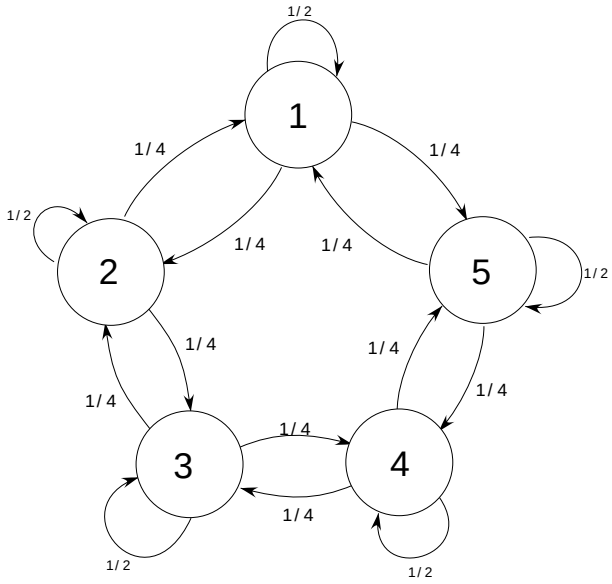

Figure 7.1: Migration among five cities

It is apparent that asymptotically we can expect each city to have the same population of $200,000$. Namely

$$\pi_i(t) \to \frac{1}{5}, \quad \forall i.$$

Observation 7.1.1 We can learn a lot from this simple model. We begin with a few observations

- Notice that $\pi_i \equiv \frac{1}{5}$ is just the uniform distribution on five elements.
- Recall that the uniform distribution is the maximum entropy distribution and that $\pi(0)$ is one of the minimum entropy distribution (Theorem 1.6.1).
- Also notice that π is *invariant* in the sense that

$$\pi(t) = \pi \Rightarrow \pi(t+1) = \pi.$$

- We also observe that there is a compact way to describe the updating mechanism of the migration. Let $\pi(t)$ denote the column vector with

components $\pi_i(t)$. Then

$$\pi(t+1) = P\pi(t), \tag{7.1}$$

where P is the matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix}. \tag{7.2}$$

- The $\pi_i(t+1)$ *do depend* on the $\pi_i(\tau), \tau \leq t$. Recursion (7.1), however, shows that this dependence is rather simple in that the fractions at time $t+1$ can be computed knowing only the fractions at the previous time and the matrix P.

Are we always going to get asymptotically an invariant distribution? Is this going always to be the uniform distribution? Can there be more than one invariant distribution? Is the transition mechanism always increasing the entropy of the distribution? How long does it take to be "close" to the invariant distribution? These are some pertinent questions. Let us consider the same migration example with only one change: City number one is Venice and when people get there they don't want to leave any more. What happens in this case? It is easy to see that the π_i are now updated according to

$$\pi(t+1) = Q^*\pi(t)$$

where Q^* denotes the transpose of the matrix Q given by

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Notice that the rows of Q sum to one (as did those of P), but Q is no more symmetric. Suppose the initial distribution is the uniform $\pi_i(0) \equiv \frac{1}{5}$. We get

$$\pi_i(1) = \begin{cases} \frac{6}{20}, & i = 1, \\ \frac{3}{20}, & i = 2,\ i = 5, \\ \frac{4}{20}, & i = 3,\ i = 4. \end{cases} \qquad \pi_i(2) = \begin{cases} \frac{30}{80}, & i = 1, \\ \frac{10}{80}, & i = 2,\ i = 5, \\ \frac{15}{80}, & i = 3,\ i = 4. \end{cases}$$

A larger and larger fraction is staying in Venice and asymptotically

$$\pi_i = \begin{cases} 1, & i = 1, \\ 0, & i \neq 1 \end{cases}$$

Thus, we have in a sense the inverse of the previous evolution: The maximum entropy distribution goes to one of the minimum entropy ones! This already answers one of our questions: Entropy does not always increase (see Problem 63). What if city number three is Paris, and people don't want to leave from there either? Then the evolution is given by $\pi(t+1) = R^*\pi(t)$, where

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix}. \tag{7.3}$$

Are the people going to wind up in Venice or Paris? If they all start in Venice, they will stay there. Similarly for Paris. And if they start a fraction p in Venice and $q = 1 - p$ in Paris, they will also all not move. Hence the invariant distribution is no longer unique and actually there is a continuum of different invariant distributions. We are beginning to get a glimpse the richness of the migration mechanism.

7.2 The Markov transition mechanism

Often we are interested in modeling the uncertain *evolution* of a phenomenon. The appropriate mathematical model is that of a *stochastic process*.

Definition 7.2.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A family of random variables $\{X(t); t \in T\}$ defined on Ω is called a *stochastic process*. It is called a *continuous-time* process if $T = [a, b], -\infty \leq a < b \leq +\infty$. It is called a *discrete-time* process if $T \subseteq \mathbb{Z}$.

Example 7.2.2 The position of a gas molecule may be described by a continuous time stochastic process $\{X(t); t \geq 0\}$ taking values in $\mathbb{R}^3$. The monthly rain fall may be modeled by a discrete-time process $\{X(t); t \in \mathbb{N}\}$ taking values in $\mathbb{R}_+$. The yearly number of individuals in a population may be described by a discrete-time process $\{X(t); t \in \mathbb{N}\}$ taking values in $\mathbb{N}$.

Definition 7.2.3 Consider the family of random variables $X = \{X(t); t \in \mathbb{N}\}$ all taking values in the *finite* or *countable* set $\mathcal{X}$. The process X is called a *Markov chain* if it satisfies the property

$$\begin{aligned}&\mathbb{P}(X(t+1) = x | X(0) = x_0, X(1) = x_1, \cdots, X(t) = x_t)\\ &= \mathbb{P}(X(t+1) = x | X(t) = x_t), \quad \forall (x_0, x_1, \cdots, x_t, x) \in \mathcal{X}^{t+2}, \quad \forall t \geq 0. \end{aligned} \tag{7.4}$$

Remark 7.2.4 The Markov property if often described in words as " the future of the process does not depend on the (strict) past but only on the present". This is *wrong*. The correct statement is: "The future of the process depends on the past only through the present".

Example 7.2.5 *1. Bernoulli Trials*
Let $X(0), X(1), X(2), \ldots$ be independent Bernoulli trials with

$$\mathbb{P}(X(t) = H) = p, \quad \mathbb{P}(X(t) = T) = q = 1 - p.$$

This sequence is trivially a Markov chain since, by independence,

$$\begin{aligned}\mathbb{P}(X(t+1) = x | X(0) = x_0, X(1) = x_1, \cdots, X(t) = x_t) &= \mathbb{P}(X(t+1 = x)\\ &= \mathbb{P}(X(t+1) = x | X(t) = x_t).\end{aligned}$$

Example 7.2.6 *2. Random walk on $\mathbb{Z}$.*
In the previous example, replace H and T with 1 and -1, respectively. Consider the random variables defined by

$$Y(t) = \sum_{s=0}^{t} X(s).$$

We have $Y(t+1) = Y(t) + X(t+1)$ and

$$\begin{aligned}\mathbb{P}(Y(t+1) = x | Y(0) = y_0, Y(1) = y_1, \cdots, Y(t) = y_t) &= \mathbb{P}(X(t+1) = x - y_t)\\ &= \mathbb{P}(Y(t+1) = x | Y(t) = y_t).\end{aligned}$$

Hence $Y = \{Y(t); t \in \mathbb{N}\}$ is a Markov chain.

These two examples readily generalize to independent random variables $X(t)$ with any discrete distribution.

What is the most effective way to describe a Markov chain? Since $\mathcal{X}$ is countable, we can identify x_i with i and take $\mathcal{X} = \mathbb{N}$. Let us introduce the distribution of $X(t)$

$$\pi_i(t) = \mathbb{P}(X(t) = i).$$

Consider also the *transition probabilities*

$$p_{ij}(t) := \mathbb{P}(X(t+1) = j | X(t) = i). \tag{7.5}$$

In many applications, the transition probabilities p_{ij} do not depend on time. In this case, the Markov chain is called *time-homogeneous*. From now on, we shall deal with this case (unless the opposite is explicitly stated). By the law of total probability (6.5), we get the recursion

$$\pi_j(t+1) = \sum_i p_{ij}\pi_i(t). \tag{7.6}$$

The latter can be expressed more compactly as

$$\pi(t+1) = P^*\pi(t), \tag{7.7}$$

where $\pi(t)^* = (\pi_0(t), \pi_1(t), \pi_2(t), \ldots)$ and the matrix P is given by

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} & \cdots \\ p_{10} & p_{11} & p_{12} & p_{13} & \cdots \\ p_{20} & p_{21} & p_{22} & p_{23} & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

The matrix P, called *matrix of transition probabilities*, *transition matrix* or simply *chain matrix*, enjoys the following properties:

a. $p_{ij} \geq 0, \forall i, \forall j$;

b. $\sum_j p_{ij} = 1, \forall i$.

Property **b.** simply expresses the fact that rows must sum to one since this is the probability of transition to any state starting from i. A matrix satisfying properties **a.** and **b.** is called *stochastic*. Notice that, by the general multiplication of probabilities formula (6.3) and the Markov property (7.4), we have

$$\mathbb{P}(X(0) = i_0, X(1) = i_1, \ldots, X(n) = i_n) = \pi_{i_0}(0) \cdot p_{i_o i_1} \cdots p_{i_{n-2} i_{n-1}} \cdot p_{i_{n-1} i_n}. \tag{7.8}$$

To each transition matrix we can associate its *transition graph* G. This directed graph has the states of $\mathcal{X}$ as nodes and an arc from i to j if and only if $p_{ij} > 0$. The latter probability is then displayed next to the arc, see e.g. Figure 7.1.

Example 7.2.7 *3. Random walk with absorbing barriers*
Take $\mathcal{X} = \{0, 1, \ldots, N\}$ and chain matrix P given by

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ q & 0 & p & 0 & \ldots & 0 \\ 0 & q & 0 & p & \ldots & 0 \\ . & . & . & . & . & \ldots \\ . & . & . & . & . & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{bmatrix}.$$

States 0 and N are *absorbing*: Once the system reaches them, it stays there forever. This Markov chain represents a coin tossing game with two players A and B. At each toss, the probabilities of winning are p for A and $q = 1-p$ for B. $X(t)$ is the capital of A at time t. The latter increases by one unit each time A wins, and decreases by one unit each time A loses. The two absorbing states represent the situation where one of the players' capital has decreased to zero and the game stops. Notice that this example may be viewed as a (realistic) variant of Example 7.2.6.

Consider now the migration model of the previous section where we regard the fraction of population $\pi_i(t)$ as the probability of finding someone in city i at time t. Suppose the chain matrix is given by (7.3). Then cities 1 and 3 are absorbing states, and the chain represents a *cyclical random walk*.

Example 7.2.8 *4. P. and T. Ehrenfest Urn Model*
Consider the following famous *Gedankenexperiment* described by Paul and Tatiana Ehrenfest in 1907: N balls are distributed in two containers A and B. At each time t, a ball is chosen at random and moved from its container to the other. Let $X(t)$ be the number of balls in container A. Suppose that at time t there are l balls in A and $N - l$ in B. Then the probabilities that the system makes a transition to $l - 1$ or to $l + 1$ at time $t + 1$ are given by

$\frac{l}{N}$ and $\frac{N-l}{N}$, respectively. Thus, the chain matrix is

$$P = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ \frac{1}{N} & 0 & 1-\frac{1}{N} & 0 & \dots & \dots & 0 \\ 0 & \frac{2}{N} & 0 & 1-\frac{2}{N} & 0 & \dots & 0 \\ . & . & . & . & . & . & \dots \\ . & . & . & . & . & . & \dots \\ 0 & 0 & 0 & \dots & 1-\frac{1}{N} & 0 & \frac{1}{N} \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

For N large, this is an effective model of molecules diffusing through a membrane while subject to a central force (that is, in the random walk, the particle is subject to an attractive elastic force increasing with the distance from $N/2$). This model of heat exchange played an important role in the dramatic Boltzmann-Zermelo controversy on the *recurrence paradox*. It considerably helped our understanding of how thermodynamic irreversibility can coexist with reversibility at the microscopic level.

Let us introduce the n-step transition probabilities

$$p_{ij}^{(n)} := \mathbb{P}(x(t+n) = j | X(t) = i).$$

We have, in particular

$$p_{ij}^{(2)} := \sum_k p_{ik} p_{kj}.$$

By induction, we get first the *backward equation*

$$p_{ij}^{(n+1)} := \sum_k p_{ik} p_{kj}^{(n)}, \tag{7.9}$$

and then the *Chapman-Kolmogorov equation*

$$p_{ij}^{(m+n)} := \sum_k p_{ik}^{(m)} p_{kj}^{(n)}. \tag{7.10}$$

Let $P^{(n)} = \left(p_{ij}^{(n)}\right)$ be the matrix of n-step transition probabilities. By (7.9), it follows that $P^{(n)} = P^n$, namely the (row times column) product of P with itself n times (this product makes sense also when the state space is infinite). Alternatively, iterate (7.7) to get

$$\pi(t+1) = (P^*)^{t+1} \pi(0). \tag{7.11}$$

We also have

$$\pi_j(t+1) = \sum_{i=0}^{N} p_{ij}^{(t+1)} \pi_i(0), \quad \forall j. \tag{7.12}$$

Comparing (7.11) and (7.12) and taking into account that $\pi(0)$ is arbitrary, it follows that

$$P^{(t)} = P^t. \tag{7.13}$$

Remark 7.2.9 It is now apparent that the initial distribution $\pi(0)$ and the chain matrix P completely determine the Markov chain. Also notice that, since the product of two stochastic matrices is easily seen to be stochastic, P^t is stochastic for all $t \geq 0$.

Besides (7.9), the other important case of (7.10) is the *forward equation*

$$p_{ij}^{(n+1)} := \sum_{k} p_{ik}^{(n)} p_{kj}. \tag{7.14}$$

7.3 Stationary distribution

Definition 7.3.1 A distribution π is called *stationary* for the Markov chain X with chain matrix P if it satisfies

$$\pi = P^* \pi. \tag{7.15}$$

Example 7.3.2 In the migration model of the Section 7.1 with P given by (7.2), the uniform distribution $\pi(i) = 1/5$ is stationary for the chain. In Example 7.2.7, let $\mu_i = \delta_{0i}$ and $\nu_i = \delta_{Ni}$. Then any distribution of the form

$$\bar{p} = \lambda \cdot \mu + (1 - \lambda)\nu, \quad \lambda \in [0, 1]$$

is stationary for the chain.

Example 7.3.3 The distribution

$$\pi_i = \left(\begin{array}{c} N \\ i \end{array} \right) \cdot 2^{-N}$$

may be seen to be stationary for the Ehrenfest urn model of Example 7.2.8. This is just the distribution of i heads in N tosses of a fair coin (5.2).This has a transparent interpretation: Independent of the number of balls that are initially placed in container A, asymptotically the probability of finding i balls in A is the same as if we place the N balls at random with probability $1/2$ of going into either container.

Example 7.3.4 *5. Random walk with reflecting barriers*
Let $\mathcal{X} = \{0, 1, \ldots, N\}$ and let P be given by

$$P = \begin{bmatrix} r_0 & p_0 & 0 & 0 & \ldots & 0 \\ q & 0 & p & 0 & \ldots & 0 \\ 0 & q & 0 & p & \ldots & 0 \\ . & . & . & . & . & \ldots \\ . & . & . & . & . & \ldots \\ 0 & 0 & 0 & \ldots & p_N & r_N \end{bmatrix}.$$

Here $r_0, p_0, p_N, r_N \geq 0$, $r_0 + p_0 = 1$, $r_N + p_N = 1$, $p + q = 1, p > 0, q > 0$. When $p_0 = 1$ and $p_N = 1$, this is a random walk with reflecting barriers at 0 and N. When $p_0 > 0$ and $p_N > 0$ but both $\neq 1$, the barriers are partially reflecting. In example 7.2.8, states 0 and N are reflecting barriers!

Example 7.3.5 *6. A model in queueing theory*
One of the most important fields of application of Markov chains is queueing theory. Consider the following model: In each unit of time, one client is served (unless nobody is waiting). Let $X(t)$ be the number of people queueing at time t ($\mathcal{X} = \mathbb{N}$). Let $\xi(t)$ be the number of persons arriving during period t. We assume that the $\xi(t), t \in \mathbb{N}$ are i.i.d., namely *independent, identically distributed random variables*. Let

$$\mathbb{P}(\xi(t) = i) = q_i, \quad q_i \geq 0, \quad \sum_{i=0}^{\infty} q_i = 1. \tag{7.16}$$

Now observe that if at the beginning of period t there are $i \geq 1 (i = 0)$ persons in line, then at the end of such period there are $i - 1 + \xi(t)$ $(\xi(t))$ persons queueing. Hence

$$X(t+1) = (X(t) - 1)^+ + \xi(t),$$

where $Y^+ := \max(Y, 0)$. The (infinite) transition matrix is then given by

$$P = \begin{bmatrix} q_0 & q_1 & q_2 & q_3 & \ldots \\ q_0 & q_1 & q_2 & q_3 & \ldots \\ 0 & q_0 & q_1 & q_2 & \ldots \\ 0 & 0 & q_0 & q_1 & \ldots \\ . & . & . & . & \ldots \end{bmatrix}.$$

Let us try to develop some intuition on the asymptotic behavior of this process. In each time period, the expected arrival number is

$$\alpha = \sum_{i=0}^{\infty} k \cdot q_k.$$

If $\alpha > 1$, we expect the length of the queue to grow beyond any bound (recall, only one is served in each unit of time). When $\alpha < 1$, the length of the line might tend to some equilibrium value. For the critical case $\alpha = 1$, we expect great instability. This and more general models where the service time is also random (*birth-death process*) are applied in a variety of fields such as traffic flow, advanced telecommunication systems, server queueing, etc.

Example 7.3.6 *7. An inventory model*
Consider the following inventory model: A warehouse can store at most M pieces of a commodity. Let $X(t)$ be the number of pieces in stock at time t. If $X(t) < m$, then the commodity is immediately replenished up to the maximum level M. Suppose the cumulative demand in period t is given by $\xi(t)$ which are i.i.d. with distribution as in (7.16). Then,

$$X(t+1) = \begin{cases} X(t) - \xi(t+1), & m < X(t) \le M, \\ M - \xi(t+1), & X(t) \le m. \end{cases}, \tag{7.17}$$

where $X(t) \in \{M, M-1, \ldots, 1, 0, -1, -2, \ldots\}$.

Notice that equation (7.17) is of the form

$$X(t+1) = f(X(t), \xi(t+1)). \tag{7.18}$$

All processes generated by a recursion (7.18) with the $\xi(t)$ i.i.d. and taking at most countably many values are Markov chains [5, Theorem 2.1]. Example 7.3.5 falls into this class, as well as Example 7.2.6. Not all homogeneous Markov chains, however, admit a "natural" representation of this form.

Example 7.3.7 *8. Success runs*
Consider a sequence of Bernoulli trials for a biased coin $X(t)$. Let us agree that H is called "success" and T is called "failure". Define a sequence of random variables $Y(t), t \ge 1$ with state space $\mathcal{Y} = \{0, 1, \ldots\}$ by

$$Y(t) = \begin{cases} 0, & X(t) = T, \\ k, & X(t) = H, X(t-1) = H, \ldots, X(t-k+1) = H, X(t-k) = T. \end{cases}$$

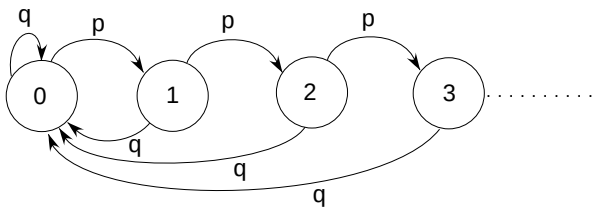

Figure 7.2: Success runs.

Namely, $Y(t)$ is the length of the sequence of successes. It is apparent that only transitions $k \to k+1$ and $k \to 0$ are possible. The chain matrix is

$$P = \begin{bmatrix} q & p & 0 & 0 & 0 & \dots \\ q & 0 & p & 0 & 0 & \dots \\ q & 0 & 0 & p & 0 & \dots \\ . & . & . & . & . & \dots \end{bmatrix}.$$

The theory of runs has a variety of applications such as testing randomness, contagion, imperfections in the production process, intentional mixing or clustering of people, etc. see [10, pp 40-41]. The graph of the chain is depicted in Figure 7.2.

Example 7.3.8 *9. Genetics model*
Consider the following simple haploid model in genetics due to Fisher and Wright. Consider a population of constant size obtained selecting N individuals at random each successive generation. A certain gene with forms A and a has therefore $2N$ representative in each generation. We say that the system is in state $j, j = 0, 1, \dots, 2N$ at time n if A occurs j times in the

n-th population. Assuming random mating, the probability that a gene is of type A in the next generation is then $\frac{j}{2N}$. The probability that we have a transition from j to k individuals with gene A in the following generation is then

$$p_{jk} = \left(\begin{array}{c} 2N \\ k \end{array} \right) \left(\frac{j}{2N} \right)^k \left(1 - \frac{j}{2N} \right)^{N-k},$$

which is just the binomial distribution with $p = \frac{j}{2N}$, cf. Example 5.1.2. States 0 and $2N$ are absorbing. More realistic models take into account the mutation pressure.

Example 7.3.9 *10. Galton-Watson model*
Another example in the class (7.18) is a branching process model. Suppose $\{\xi_n^{(j)}, n \geq 1, j \geq 1\}$ are i.i.d. random variables with integer values. Define

$$X(n+1) = \sum_{j=1}^{X(n)} \xi_{n+1}^{(j)}.$$

By convention, $X(t) = 0$ implies $X(t+1) = 0$. Moreover, $X(0)$ is assumed independent of the $\{\xi_n^{(j)}\}$. Here $X(n)$ represents the number of individuals in the n-th generation of a population (persons, particles, etc.). Individual j of the n-th generation produced $\xi_{n+1}^{(j)}$ descendents. For instance, this model is used in genetics to estimate the survival chances of a mutant gene.

Observation 7.3.10 Suppose the state space is finite $\mathcal{X} = \{0, 1, \ldots, N\}$. Observe that π satisfying (7.15) implies that P has an *eigenvalue* equal to one (the eigenvalues of P and P^* are the same). Moreover, if the uniform distribution is invariant, we have that also *the columns of P sum to one*. A matrix with nonnegative elements such that both rows and columns sum to one is called *doubly stochastic*. Observe finally that if P is symmetric as in (7.2), then for sure it is doubly stochastic and it has the uniform distribution as stationary.

Does every Markov chain posses at least one stationary distribution? When is such a distribution unique? If a unique stationary distribution π exists, does $\pi(t) \to \pi$? These questions will be tackled in Chapter 9.

7.4 Reversibility

What happens if we look at a Markov chain X with time reversed[1]? First of all, it can be shown that the reverse-time process is also Markov. Indeed, the Markov property itself can be formulated in a way that does not priviledge one direction of time. Consider the "past" event

$$A = \{X(0) = x_0, X(1) = x_1, \ldots, X(t-1) = x_{t-1}\},$$

and the "future" event

$$B = \{X(t+1) = x_{t+1}, X(t+2) = x_{t+2}, \ldots\}.$$

We have

$$\mathbb{P}(B|A, X(t) = x_t) = \frac{\mathbb{P}(B \cap A|X(t) = x_t)}{\mathbb{P}(A|X(t) = x_t)}.$$

By the Markov property, we also have

$$\mathbb{P}(B|A, X(t) = x_t) = \mathbb{P}(B|X(t) = x_t).$$

It follows that

$$\frac{\mathbb{P}(B \cap A|X(t) = x_t)}{\mathbb{P}(A|X(t) = x_t)} = \mathbb{P}(B|X(t) = x_t)$$

and

$$\mathbb{P}(A \cap B|x(t) = x_t) = \mathbb{P}(A|X(t) = x_t) \cdot \mathbb{P}(B|X(t) = x_t), \tag{7.19}$$

namely, at each time t, past and future of the process are *conditionally independent* given the present $X(t)$. Notice that (7.19) in turn gives

$$\mathbb{P}(A|B, X(t) = x_t) = \mathbb{P}(A|X(t) = x_t),$$

which is the Markov property for the chain with time reversed. Looking at the process with time reversed is important in the theory of Markov chains, for instance for Monte Carlo simulation and queuing theory. It is also important for conceptual reasons (reversibility in physics[2]). Also observe that in some

[1] We consider, namely, the process $Y(t) = X(-t)$ indexed by the negative integers.

[2] For instance, suppose we know from some historic document that a certain event A (war, migration, draught, etc) occured when there was a lunar eclipse. By integrating *backward in time* the equations of motion of earth and moon we can find when sun, earth and moon were aligned. Hence, we can safely date A.

applications the independent variable might not be time. For instance, $X(k)$ may represent the displacement of the k-th section of a flexible structure. Let us introduce the *reverse-time transition probabilities*

$$q_{ji}(t, \pi(0)) := \mathbb{P}(X(t) = i | X(t+1) = j), \tag{7.20}$$

where we have emphasized the dependence on the initial distribution $\pi(0)$. Also notice that the q_{ji} depend on time even when the p_{ij} don't (time-homogeneous Markov chain). Let us find the relation between the q_{ji} and the p_{ij}. Let $\pi(0)$ be the initial distribution and let $\pi(t)$ be distribution at time t obtained through (7.6). We have

$$\pi_i(t)p_{ij} = \mathbb{P}(X(t) = i, X(t+1) = j) = \pi_j(t+1)q_{ji}(t, \pi(0)). \tag{7.21}$$

It follows that, whenever $\pi_j(t+1) \neq 0$,

$$q_{ji}(t, \pi(0)) = \frac{\pi_i(t)}{\pi_j(t+1)} p_{ij}. \tag{7.22}$$

Observation 7.4.1 Notice that if a distribution π satisfies (7.21) at some time

$$\pi_i p_{ij} = \pi_j q_{ji}(t, \pi(0)),$$

then, summing on both sides with respect to i, we get $\pi_j = \sum_i p_{ij}\pi_i$, it namely follows that π is invariant.

Definition 7.4.2 The Markov chain X with stationary distribution π is said to satisfy the *detailed balance* condition if

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \text{for all } i, j \in \mathcal{X}. \tag{7.23}$$

We record the following important result.

Theorem 7.4.3 Let π be any distribution satisfying (7.23). Then π is invariant for the chain with transition matrix $P = (p_{ij})$.

To understand the detailed balance condition, let us go back to the migration model with P given by (7.2). Suppose we have reached the asymptotic distribution of $1/5$ of population in each city. Nevertheless, people keep moving according to the rules of Table 7.1. Let us pick two cities, say 1 and 2, and

look at their population exchange. The fraction of the total population moving from 1 to 2 is $\frac{1}{4} \cdot \frac{1}{5} = \frac{1}{20}$. Hence each year $50,000$ people move from 1 to 2. But $50,000$ move from 2 to 1. Thus, *disregarding the population exchange with other cities*, the inflow and outflow *between just* 1 *and* 2 is the same. The same applies to any other pair of cities. This is why we talk of *detailed* balance.

A chain satisfying (7.23) is called *reversible* with respect to π. Indeed, if the detailed balance condition (7.23) is satisfied, we get from (7.21)

$$\pi_j q_{ji}(t, \pi) = \pi_i p_{ij} = \pi_j p_{ji}, \quad \text{for all } i, j \in \mathcal{X}. \tag{7.24}$$

Consider now the case when besides satisfying (7.23), we have $\pi_i > 0, \forall i$ (we shall see in Chapter 9 that there exists a large class of Markov chains (*irreducible* chains with finite state space) that satisfy this last condition. Then, we get from (7.24) that $q_{ji}(t, \pi) = q_{ji}(\pi)$ do not depend on time and

$$q_{ji}(\pi) = p_{ji}, \quad \text{for all } i, j \in \mathcal{X}. \tag{7.25}$$

We conclude that if we start with $\pi(0) = \pi$, then

$$\mathbb{P}_\pi(X(t+1) = i | X(t) = j) = \mathbb{P}_\pi(X(t) = i | X(t+1) = j),$$

namely the transition probabilities between two states are the same disregarding the direction of time! It follows that, for $t_1 < t_2 < \cdots < t_m$ and $x_i \in \mathcal{X}$,

$$\begin{aligned} &\mathbb{P}_\pi(X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_n) = x_n) = \\ &\mathbb{P}_\pi(X(t_1) = x_n, X(t_2) = x_{n-1}, \ldots, X(t_n) = x_1). \end{aligned}$$

This explains why the chain is called reversible. Finally, notice that the detailed balance condition may be expressed in the finite state space case as

$$\text{diag}\,(\pi_1, \ldots, \pi_n) P = P^* \text{diag}\,(\pi_1, \ldots, \pi_n),$$

which represents some kind of "symmetry with respect to the invariant distribution".

7.5 Martingales.

7.5.1 Martingales and submartingales

Definition 7.5.1 Consider a discrete time stochastic process $X = \{X(t), t \geq 0\}$ with finite or denumerable state space $\mathcal{X}$. The process $Y = \{Y(t), t \geq 0\}$ is called a *martingale* [3] with respect to X if

1. $\mathbb{E}|Y(t)| < \infty, \forall t \geq 0$;
2. $Y(t)$ is a function of $\{X(s), 0 \leq s \leq t\}, \forall t \geq 0$;
3. $\mathbb{E}(Y(t+1)|X(0), X(1), \ldots, X(t)) = Y(t)$.

It is called a *submartingale* with respect to X when it satisfies 1., 2. and

$$\mathbb{E}(Y(t+1)|X(0), X(1), \ldots, X(t)) \geq Y(t).$$

We can say that a martingale is *conditionally constant* and a submartingale is *conditionally increasing*. The process is called a *supermartingale* when $-Y(t)$ is a submartingale. Finally, notice that the case $Y(t) = X(t)$ is also included.

Example 7.5.2 Consider Example 7.2.6 with $p = q = 1/2$. It is apparent that conditions 1. and 2. above are verified. Moreover, using the independence of the $X(t)$, (6.19), (6.21), and the fact that $\mathbb{E}X(t) = 0, \forall t \geq 0$, we get

$$\begin{aligned}
&\mathbb{E}(Y(t+1)|X(0), X(1), \ldots, X(t)) \\
&= \mathbb{E}(X(t+1)|X(0), X(1), \ldots, X(t)) + \mathbb{E}(Y(t)|X(0), X(1), \ldots, X(t)) \\
&= \mathbb{E}(X(t+1)) + Y(t) = Y(t).
\end{aligned}$$

Thus Y is a martingale with respect to the sequence X. Think of $Y(t)$ as representing the capital of player A at time t. The crucial property of a martingale

$$\mathbb{E}(Y(t+1)|X(0), X(1), \ldots, X(t)) = Y(t)$$

[3] Martingale apparently originates from the southern French town Martigues such as in "chausses à la martinguale". It is nowadays a strap on a horse's harness used to hold down the horse's head. It was already used to indicate a gambling strategy in the eighteenth century: " I took all the gold I found, and playing the martingale, and doubling my stakes continuously, I won every day during the remainder of the carnival", writes the Venetian Giacomo Casanova around 1750.

simply expresses the fact that the game is *fair*. Indeed, the expected capital at time $t+1$, given $X(s), s \leq t$, is the capital at time t. Suppose now that $p > q$. Then

$$\mathbb{E}(Y(t+1)|X(0), X(1), \ldots, X(t)) = \mathbb{E}X(t+1) + Y(t) = p - q + Y(t) > Y(t).$$

Hence, Y is a submartingale with respect to the sequence X.

Example 7.5.3 *Pólya's urn:* An urn contains at time $t = 0$ m white and n black balls. At each following time, a ball is chosen at random from the urn. The ball is then put back into the urn together with another ball of the same colour taken from a large collection available to the experimenter. Hence, the number of balls in the urn grows. Pólya's urn may be viewed as a rough model of contagious diseases, where each occurrence increases the probability of further occurrences. Let $X(t)$ be the number of white balls in the urn at time $t \geq 0$. Let

$$Y(t) = \frac{X(t)}{m+n+t}.$$

Then $Y = \{Y(t), t \geq 0\}$ is a martingale with respect to X.

Proposition 7.5.4 Let $Y = \{Y(t), t \geq 0\}$ be a martingale with respect to $X = \{X(t), t \geq 0\}$. Let φ be a convex function and define $Z(t) := \varphi(Y(t)), t \geq 0$. If $\mathbb{E}|Z(t)| < \infty$, $\forall t \geq 0$, then Z is a submartingale with respect to X.

Proof . By (6.23), we have

$$\begin{aligned}\mathbb{E}(Z(t+1)|X(0), X(1), \ldots, X(t)) &= \mathbb{E}(\varphi(Y(t+1))|X(0), X(1), \ldots, X(t)) \\ &\geq \varphi\left(\mathbb{E}(Y(t+1)|X(0), X(1), \ldots, X(t))\right) = \varphi(Y(t)) = Z(t).\end{aligned}$$

□

7.5.2 Space-time harmonic functions

Consider a Markov chain $X = \{X(t); t \in \mathbb{N}\}$ with state space $\mathcal{X}$ and transition matrix $P = (p_{ij})$. There exists a large class of X-martingales that are constructed as *instantaneous functions* of $X(t)$.

Definition 7.5.5 A function $h : \mathbb{N} \times \mathcal{X} \to \mathbb{R}$ is called *space-time harmonic* if, for every $t \geq 0$ and all $i, j \in \mathcal{X}$, it satisfies the backward equation

$$h(t,i) = \sum_j p_{ij} h(t+1,j). \tag{7.26}$$

Proposition 7.5.6 Let h be space-time harmonic for the Markov chain X. Define the stochastic process $Y = \{Y(t) = h(t, X(t)), t \geq 0\}$. Then, if $E|Y(t)| < \infty, \forall t$, Y is a martingale with respect to X.

Proof . We have

$$\begin{aligned} \mathbb{E}(Y(t+1)|X(0), X(1), \ldots, X(t)) &= \mathbb{E}(h(t+1, X(t+1))|X(0), X(1), \ldots, X(t)) \\ &= \mathbb{E}(h(t+1, X(t+1))|X(t)). \end{aligned}$$

Now observe that for all $i \in \mathcal{X}$

$$\mathbb{E}(h(t+1, X(t+1))|X(t) = i) = \sum_j p_{ij} h(t+1, j) = h(t, i).$$

Thus $\mathbb{E}(h(t+1, X(t+1))|X(t)) = h(t, X(t)) = Y(t)$. Since properties 1. and 2. are also satisfied, we conclude that Y is a martingale with respect to X. □

Problems

Problem 50 Show that the product PQ of two $n \times n$ stochastic matrices P and Q is stochastic.

Problem 51 Let $\{X(t), t \geq 0\}$ be a Markov chain with values in $\mathcal{X}$ and transition matrix P. Let

$$Y(t) = \begin{pmatrix} X(t) \\ X(t+1) \\ X(t+2) \end{pmatrix}.$$

Show that $\{Y(t), t \geq 0\}$ is also a Markov chain with values in $\mathcal{Y} = \mathcal{X}^3$.

Problem 52 Let $\{Y_k\}_{k \geq 1}$ be a sequence of independent random variables all having the geometric distribution (4.12). Define $X(t) := \max(Y_1, \ldots, Y_t)$. Let $X(0)$ be a random variable independent of the $\{Y_k\}_{k \geq 1}$ with values in $\mathbb{N}$. Show that $\{X_k\}_{k \geq 0}$ is a Markov chain and determine its transition matrix.

Problem 53 Consider a two-state Markov chain with transition matrix

$$P = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}.$$

Find the (unique) stationary distribution.

Problem 54 Can the uniform distribution be stationary in Examples 7.2.7 and 7.3.4?

Problem 55 Show that the chain of Problem 52 does not possess a stationary distribution.

Problem 56 Consider Problem 51. Suppose π is a stationary distribution for X. Does Y possess a stationary distribution?

Problem 57 Write the transition matrix of $X(t)$ in Example 7.3.6 when $m = 2$ and $M = 4$.

Problem 58 Prove Theorem 7.4.3.

Problem 59 Show that the Ehrenfest chain 7.2.8 is reversible with respect to the distribution

$$\pi_i = \begin{pmatrix} N \\ i \end{pmatrix} \cdot 2^{-N}.$$

It follows that π is indeed stationary.

Problem 60 Show that the process Y of Example 7.5.3 is an X martingale.

Problem 61 Consider a discrete-time stochastic process $X = \{X(0), X(1), \ldots\}$ with countable state space $\mathcal{X}$ and a random variable Y with finitely many values. Define $Z(t) = \mathbb{E}(Y|X(0), X(1), \ldots, X(t))$. Show that $Z = \{Z(0), Z(1), \ldots\}$ is a martingale with respect to X.

Chapter 8

Thermodynamic systems: Dynamics

8.1 Information divergence

Definition 8.1.1 Let p and q belong to the simplex of probability distributions on $\mathcal{X} = \{1, 2, \ldots, N\}$. We say that the *support* of p is contained in the support of q if $q_i = 0 \Rightarrow p_i = 0$ and write $Supp(p) \subseteq Supp(q)$. The*(Information) Divergence* or *Kullback-Leibler Index* or *Relative Entropy* of q from p is defined to be

$$\mathbb{D}(p\|q) = \begin{cases} \sum_i p(i) \log \frac{p(i)}{q(i)}, & Supp(p) \subseteq Supp(q), \\ +\infty, & Supp(p) \not\subseteq Supp(q). \end{cases}, \tag{8.1}$$

where, as in Chapter 2, $0 \cdot \log 0 = 0$.

Observation 8.1.2 $\mathbb{D}(\cdot, \cdot)$ does not induce a metric since it is not symmetric and, more important, it does not satisfy the triangle inequality.

Proposition 8.1.3 *The divergence defined in (8.1) enjoys the following properties:*

1. $\mathbb{D}(p\|q) \geq 0$*;*
2. $\mathbb{D}(p\|q) = 0$ *if and only if* $p = q$*.*

Proof . Assume $Supp(p) \subseteq Supp(q)$ otherwise the result is trivial. Write

$$\mathbb{D}(p\|q) = \sum_i \left[\frac{p(i)}{q(i)} \log \frac{p(i)}{q(i)}\right] q(i) = \sum_i g\left(\frac{p(i)}{q(i)}\right) q(i),$$

where the function $g(x) = x \log x$ is strictly convex on $x > 0$ (see the proof of Theorem 1.6.1). It now follows from Theorem 1.5.6 that

$$\mathbb{D}(p\|q) = \sum_i g\left(\frac{p(i)}{q(i)}\right) q(i) \geq g\left(\sum_i \frac{p(i)}{q(i)} q(i)\right) = g(1) = 0.$$

This proves both properties 1 and 2 since the inequality is strict unless $p(i) = q(i), \forall i$. □

Property 1. is interpreted in Information Theory in the following sense since Shannon's fundamental work: The average message length is minimized when codes are assigned on the basis of the true probabilities p rather than any other distribution q, see e.g. [7].[1]

8.2 The second law of thermodynamics

Consider again the situation of Chapter 2 where the thermodynamic states of the system are given by probability distributions on the space of mesoscopic states $\mathcal{T} = \{1, 2, \ldots, N\}$. We make the following *assumptions:*

A1. The time evolution of the system is described by a time-homogeneous Markov chain X with state space $\mathcal{T}$;

A2. The Boltzmann distribution

$$\bar{\pi}_i = Z^{-1} \exp\left[-\frac{E_i}{kT}\right], \quad Z = \sum_i \exp\left[-\frac{E_i}{kT}\right], \tag{8.2}$$

is *stationary* for the Markov chain X.

[1] My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". John von Neumann had a better idea, he told me, "You should call it entropy, for two reasons. In the first place your uncertainty function goes by that name in statistical mechanics. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage." Shannon as quoted in M. Tribus, E.C. McIrvine, Energy and information, Scientific American, 224 (September 1971), 178-184.

Of course, the evolution of a physical system occurs in continuous time. A continuous-time Markov chain, however, is essentially a discrete-time Markov chain with a random time scale, see [5, Chapter 8]. The results we present in this chapter have similar counterparts in the continuous time setting. Consider again the free energy introduced in (2.11) $F(E,p,T) = U(E,p) - TS(p)$. Since

$$\ln \bar{\pi}_i = -\ln Z - \frac{E_i}{kT},$$

we have

$$F(E,p,T) = U(E,p) - TS(p) = \sum_i E_i p_i + kT \sum_i p_i \ln p_i$$
$$= -kT \sum_i p_i \ln \bar{\pi}_i - kT \ln Z + kT \sum_i p_i \ln p_i = kT\mathbb{D}(p\|\bar{\pi}) - kT \ln Z.$$

Observe that Z does not depend on p. Hence, minimizing $F(E,\cdot,T)$ or $\mathbb{D}(\cdot\|\bar{\pi})$ over the simplex of thermodynamic states $\mathcal{S}$ is equivalent. By Proposition 8.1.3, however, the unique minimizer of $\mathbb{D}(\cdot\|\bar{\pi})$ on $\mathcal{S}$ is $\bar{\pi}$. Hence, this yields another proof of Gibbs' Principle (see Theorem 2.1.3). We are now ready for the fundamental law of dynamics.

Theorem 8.2.1 (Second Law of Thermodynamics) Let P be the chain matrix, let $\pi(0) \in \mathcal{S}$ and let $\pi(t) := (P^*)^t \pi(0)$. Then,

$$F(E,\pi(t+1),T) \leq F(E,\pi(t),T). \tag{8.3}$$

Proof . We show equivalently that $\mathbb{D}(\pi(t)\|\bar{\pi})$ is nonincreasing. Denote by $P_{ij}(t) = \mathbb{P}_{\pi(0)}(X(t)=i, X(t+1)=j)$ and by $\Pi_{ij} = \mathbb{P}_{\bar{\pi}}(X(t)=i, X(t+1)=j)$. Observe that $Supp(P_{ij})(t) \subseteq Supp(\Pi_{ij})$. Hence, $\mathbb{D}(P_{ij}(t)\|\Pi_{ij})$ is finite. Using $P_{ij}(t) = p_{ij}\pi_i(t)$ and $\Pi_{ij} = p_{ij}\bar{\pi}_i$, we get

$$\mathbb{D}(P_{ij}(t)\|\Pi_{ij}) = \sum_{ij} P_{ij}(t) \ln \frac{P_{ij}(t)}{\Pi_{ij}} = \sum_{ij} p_{ij}\pi_i(t) \ln \frac{p_{ij}\pi_i(t)}{p_{ij}\bar{\pi}_i}$$
$$= \sum_{ij} p_{ij}\pi_i(t) \ln \frac{\pi_i(t)}{\bar{\pi}_i} = \sum_i \left(\sum_j p_{ij} \right) \pi_i(t) \ln \frac{\pi_i(t)}{\bar{\pi}_i} = \mathbb{D}(\pi(t)\|\bar{\pi}). \tag{8.4}$$

By (7.21), we also have

$$\begin{aligned}
&\mathbb{D}(P_{ij}(t)\|\Pi_{ij}) = \sum_{ij} P_{ij}(t) \ln \frac{P_{ij}(t)}{\Pi_{ij}} \\
&= \sum_{ij} q_{ji}(t,\pi(0))\pi_j(t+1) \ln \frac{q_{ji}(t,\pi(0))\pi_j(t+1)}{q_{ji}(\bar{\pi})\bar{\pi}_j} \\
&= \sum_{ij} q_{ji}(t,\pi(0))\pi_j(t+1) \ln \frac{\pi_j(t+1)}{\bar{\pi}_j} + \sum_{ij} q_{ji}(t,\pi(0))\pi_j(t+1) \ln \frac{q_{ji}(t,\pi(0))}{q_{ji}(\bar{\pi})} \\
&= \sum_j \left(\sum_i q_{ji}(t,\pi(0)) \right) \pi_j(t+1) \ln \frac{\pi_j(t+1)}{\bar{\pi}_j} \\
&+ \sum_j \left(\sum_i q_{ji}(t,\pi(0)) \ln \frac{q_{ji}(t,\pi(0))}{q_{ji}(\bar{\pi})} \right) \pi_j(t+1) \\
&= \mathbb{D}(\pi(t+1)\|\bar{\pi}) + \sum_j \mathbb{D}(q_{ji}(t,\pi(0))\|q_{ji}(\bar{\pi}))\pi_j(t+1). \qquad (8.5)
\end{aligned}$$

Comparing (8.4) and (8.5), we get

$$\mathbb{D}(\pi(t+1)\|\bar{\pi}) - \mathbb{D}(\pi(t)\|\bar{\pi}) = -\sum_j \mathbb{D}(q_{ji}(t,\pi(0))\|q_{ji}(\bar{\pi}))\pi_j(t+1). \qquad (8.6)$$

Since both $\mathbb{D}(q_{ji}(t,\pi(0))\|q_{ji}(\bar{\pi})) \geq 0$ and $\pi_j(t+1) \geq 0$, we get (8.3). Notice that (8.6) also implies that its left-hand side does not depend on i. □

8.3 A stronger form of the second law

In the notation of the previous section, assume that there exists $\bar{t}$ such that $\pi_i(t) > 0, \forall i \in \mathcal{X}, \forall t \geq \bar{t}$. This is the case if we assume that, for all $\pi(0)$, $\pi(t) \to \bar{\pi}$ (zeroth law of thermodynamics).

Lemma 8.3.1 The function

$$g(t,j) := \frac{\bar{\pi}_j}{\pi_j(t)}$$

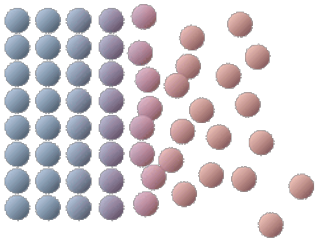

Figure 8.1: A pictorial description of thermodynamic irreversibility: Cold molecules get in contact with a heat source.

is *space-time harmonic* for the chain with initial distribution $\pi(0)$ in the reverse time direction, namely it satisfies

$$g(t+1,j) = \sum_i q_{ji}(t,\pi(0))g(t,i). \tag{8.7}$$

Proof . By (7.21) and the invariance of the Botzmann distribution $\bar{\pi}$, we have

$$\begin{aligned}\sum_i q_{ji}(t,\pi(0))g(t,i) &= \sum_i q_{ji}(t,\pi(0))\frac{\bar{\pi}_i}{\pi_i(t)} = \sum_i \frac{\pi_i(t)}{\pi_j(t+1)}p_{ij}\frac{\bar{\pi}_i}{\pi_i(t)} \\ &= \frac{1}{\pi_j(t+1)}\sum_i \bar{\pi}_i p_{ij} = \frac{\bar{\pi}_j}{\pi_j(t+1)} = g(t+1,j).\end{aligned}$$

□

Lemma 8.3.2 Under the assumptions of the previous lemma, the stochastic process $Y(t) := g(t, X(t))$ is a martingale with respect to X in the reverse time direction.

Proof . First of all, we have

$$\mathbb{E}_{\pi(0)} Y(t) = \sum_i g(t,i)\pi_i(t) = \sum_i \frac{\bar{\pi}_i}{\pi_i(t)}\pi_i(t) = 1, \quad \forall t \geq 0.$$

Second, observe that $Y(t)$ is in fact a function of $X(t)$. Finally, taking Lemma 8.6 into account, a calculation similar to that in the proof of Proposition 7.5.6 yields

$$\mathbb{E}(Y(t)|X(t+1), X(t+2), \ldots) = Y(t+1).$$

□

Theorem 8.3.3 Under the above assumptions, the stochastic process $Z(t) := -\ln Y(t) := -\ln g(t, X(t))$ is a submartingale with respect to X in the reverse time direction.

Proof . Observe that $-\ln$ is a convex function and invoke Proposition 7.5.4. □

The above result is a Markov chain version of [24] which, in turn, was inspired by [9]. Observe that Theorem 8.3.3 gives a stronger, local form of the second law. Indeed, define the *free energy density*

$$\psi(t,i) := E_i + kT \ln \pi_i(t) = -kT \ln g(t,i) - kT \ln Z = kT \ln \frac{\pi_i(t)}{\bar{\pi}_i} - kT \ln Z, \tag{8.8}$$

so that

$$\mathbb{E}_{\pi(0)}(\psi(t, X(t)) = \sum_i \psi(t,i)\pi_i(t) = F(E, \pi(t), T).$$

By Theorem 8.3.3 we now have

$$\mathbb{E}(\psi(s, X(s))|X(t) = i) \geq \psi(t,i), s < t,$$

namely $\psi(t, X(t))$ is conditionally increasing in the reverse time direction. This represents a local specification of the dissipative characteristics of these systems. It will be made more transparent in Section 8.6. Next, we show that Theorem 8.2.1 is indeed a consequence of Theorem 8.3.3.

Corollary 8.3.4 $\mathbb{D}(\pi(t)\|\bar{\pi})$ is nonincreasing.

Proof . It suffices to observe that, by Theorem 8.3.3, we have

$$E(-\ln g(t, X(t)|X(t+1), X(t+2), \ldots) \geq -\ln g(t+1, X(t+1)).$$

Taking expectations, using (6.24) and observing that

$$\mathbb{E}_{\pi(0)}(-\ln g(t, X(t)) = \sum_i \ln \frac{\pi_i(t)}{\bar{\pi}_i} \pi_i(t) = \mathbb{D}(\pi(t)\|\bar{\pi}),$$

we get

$$\mathbb{D}(\pi(t)\|\bar{\pi}) \geq \mathbb{D}(\pi(t+1)\|\bar{\pi}).$$

□

8.4 Schrödinger bridges

Consider a Markov chain $X = \{X(0), X(1), \ldots\}$ with finite state space $\mathcal{X}$. We denote by Π_0^T the joint distribution of $\{X(0), X(1), \ldots, X(T)\}$ on $\mathcal{X}^{T+1}$. Notice that Π_0^T may be thought of as the probability distribution induced by the process X on the space of *trajectories* $(i_0, i_1, \ldots, i_{T-1}, i_T) \in \mathcal{X}^{T+1}$. Suppose that we can estimate the marginal distribution at time T (see Section 8.5), but we find a distribution ρ which differs from $\pi(T) = (P^*)^{T+1}\pi(0)$. Thus, our assumption on the joint distribution of $\{X(0), X(1), \ldots, X(T)\}$ was wrong. It is then natural to consider the following problem. We want to find another Markovian distribution P_0^T on $\mathcal{X}^{T+1}$ which has the observed marginal ρ at time T and is "as close as possible" to the "prior" distribution Π_0^T. To measure how close the distributions are, we use the information divergence. We consider namely the following
Maximum Entropy Problem: Let $\mathcal{D}_{\mathcal{M}}(\rho)$ denote the family of Markovian distributions on $\mathcal{X}^{T+1}$ that have marginal ρ at time T.

$$\text{minimize} \quad \{\mathbb{D}(P_0^T\|\Pi_0^T); P_0^T \in \mathcal{D}_{\mathcal{M}}(\rho)\} \tag{8.9}$$

Observation 8.4.1 At first sight, this problem may appear quite different from Problem (2.4)-(2.5). Observe, however, the following. Maximizing the entropy $S(\cdot)$ is the same as minimizing $\mathbb{D}(\cdot\|p_u)$, where p_u is the uniform

distribution. Moreover, the constraint on the final marginal in Problem 8.9, may be expressed as a *linear* constraint on P_0^T as

$$\sum_{i_0}\sum_{i_1}\cdots\sum_{i_{T-1}} P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T) = \rho(i_T).$$

Thus the two problems have the same form.

In order to facilitate the solution of Problem 8.9, let us introduce the reverse time transition probabilities $q_{ji}(t)$ (7.20) corresponding to P_0^T and $q^\pi_{ji}(t)$ corresponding to Π_0^T. By a reverse-time version of (7.8), we have

$$P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T) = q_{i_1 i_0}(0)\cdot q_{i_2 i_1}(1)\cdot q_{i_T i_{T-1}}(T-1)\cdot \rho_{i_T}, \quad (8.10)$$

$$\Pi_0^T(i_0, i_1, \ldots, i_{T-1}, i_T) = q^\pi_{i_1 i_0}(0)\cdot q^\pi_{i_2 i_1}(1)\cdot q^\pi_{i_T i_{T-1}}(T-1)\cdot \pi_{i_T}(T). \quad (8.11)$$

We then have the following key representation for $\mathbb{D}(P_0^T\|\Pi_0^T)$.

Lemma 8.4.2 In the previous notation, assume that $Supp(P_0^T) \subseteq Supp(\Pi_0^T)$. Let $p(t)$ and $\pi(t)$ denote the marginals of P_0^T and Π_0^T at time t, respectively. Then

$$\mathbb{D}(P_0^T\|\Pi_0^T) = \sum_{k=1}^{T}\sum_{i_k} \mathbb{D}\left(q_{i_k i_{k-1}}(k-1)\|q^\pi_{i_k i_{k-1}}(k-1)\right) p_{i_k}(k) + \mathbb{D}(\rho\|\pi(T)). \quad (8.12)$$

Proof . Using (8.10)-(8.11), we get

$$\mathbb{D}(P_0^T\|\Pi_0^T) = \sum_{i_0}\sum_{i_1}\cdots\sum_{i_T} P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T)\ln\frac{P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T)}{\Pi_0^T(i_0, i_1, \ldots, i_{T-1}, i_T)}$$

$$= \sum_{i_o, i_1, \cdots, i_T} P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T)\left[\sum_{k=1}^{T}\ln\left(\frac{q_{i_k i_{k-1}}(k-1)}{q^\pi_{i_k i_{k-1}}(k-1)}\right) + \ln\left(\frac{\rho_{i_T}}{\pi_{i_T}(T)}\right)\right].$$

Observe now that

$$\sum_{i_o, i_1, \cdots, i_T} P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T)\ln\left(\frac{\rho_{i_T}}{\pi_{i_T}(T)}\right)$$

$$= \sum_{i_T}\sum_{i_o, i_1, \cdots, i_{T-1}} P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T)\ln\left(\frac{\rho_{i_T}}{\pi_{i_T}(T)}\right)$$

$$= \sum_{i_T}\rho_{i_T}\ln\left(\frac{\rho_{i_T}}{\pi_{i_T}(T)}\right) = \mathbb{D}\left(\rho\|\pi(T)\right).$$

Moreover, let $p(i_{k-1}, i_k) = \mathbb{P}(X(k-1) = i_{k-1}, X(k) = i_k)$ be the two times marginal of P_0^T. Observe that $p(i_{k-1}, i_k) = q_{i_k i_{k-1}}(k-1)p_{i_k}(k)$. We then get

$$\begin{aligned}
&\sum_{i_o, i_1, \cdots, i_T} P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T) \sum_{k=1}^{T} \ln\left(\frac{q_{i_k i_{k-1}}(k-1)}{q^{\pi}_{i_k i_{k-1}}(k-1)}\right) \\
&= \sum_{k=1}^{T} \sum_{i_k, i_{k-1}} \sum_{i_j \neq i_k, i_{k-1}} P_0^T(i_0, i_1, \ldots, i_{T-1}, i_T) \ln\left(\frac{q_{i_k i_{k-1}}(k-1)}{q^{\pi}_{i_k i_{k-1}}(k-1)}\right) \\
&= \sum_{k=1}^{T} \sum_{i_k, i_{k-1}} p(i_{k-1}, i_k) \ln\left(\frac{q_{i_k i_{k-1}}(k-1)}{q^{\pi}_{i_k i_{k-1}}(k-1)}\right) \\
&= \sum_{k=1}^{T} \sum_{i_k, i_{k-1}} q_{i_k i_{k-1}}(k-1) p_{i_k}(k) \ln\left(\frac{q_{i_k i_{k-1}}(k-1)}{q^{\pi}_{i_k i_{k-1}}(k-1)}\right) \\
&= \sum_{k=1}^{T} \sum_{i_k} \mathbb{D}\left(q_{i_k i_{k-1}}(k-1) \| q^{\pi}_{i_k i_{k-1}}(k-1)\right) p_{i_k}(k)
\end{aligned}$$

and (8.12) follows. □

Theorem 8.4.3 A solution to Problem 8.9 is given by the distribution $\hat{P}_0^T$ corresponding to the Markov chain with marginal distribution ρ at time T and reverse time transition mechanism equal to that of Π_0^T, namely

$$\hat{q}_{i_k i_{k-1}}(k-1) = q^{\pi}_{i_k i_{k-1}}(k-1), \quad k = 1, 2, \ldots, T. \tag{8.13}$$

Proof . Since both terms in (8.12) are nonnegative, and $\mathbb{D}(\rho \| \pi(T))$ is invariant over $\mathcal{D}_{\mathcal{M}}(\rho)$, the best we can hope for, when minimizing $\mathbb{D}(P_0^T \| \Pi_0^T)$, is to make the first term equal to zero. This is the case if (8.13) holds true. □

Let us compute the forward transition probabilities of $\hat{P}_0^T$. Let $\hat{p}(t)$ and $\hat{p}_{ij}$ denote the marginal at time t and the forward transition probabilities of $\hat{P}_0^T$, respectively. Let p_{ij} be the forward transition probabilities of Π_0^T. By (7.24), we have

$$\hat{p}_i(t)\hat{p}_{ij}(t) = \hat{p}_j(t+1)\hat{q}_{ji}(t), \quad \pi_i(t)p_{ij} = \pi_j(t+1)q^{\pi}_{ji}(t). \tag{8.14}$$

Assume now that $\pi_i(t) > 0, \hat{p}_i(t) > 0, \forall i, 0 \leq t \leq T$. Then (8.14) and (8.13) yield

$$\hat{p}_{ij} = \frac{\varphi(t+1, j)}{\varphi(t, i)} p_{ij}, \quad \varphi(t, i) := \frac{\hat{p}_i(t)}{\pi_i(t)}. \tag{8.15}$$

Observe, moreover, that φ is space-time harmonic with respect to the transition mechanism of Π_0^T since, by (8.15),

$$\sum_j p_{ij}\varphi(t+1,j) = \sum_j \hat{p}_{ij}\varphi(t,i) = \varphi(t,i).$$

Hence, we can say that the optimal solution of Problem 8.9 is obtained from the *a priori* Markov chain Π_0^T through a "multiplicative functional" transformation. The solution to a maximum entropy problem of the form (8.9) is called *Schrödinger bridge*. It is also possible to consider maximum entropy problems of where *both the initial and final marginals are fixed*. Under suitable assumptions, the solution is obtained also in this case through a multiplicative functional transformation of the a priori Markovian evolution, [25].

8.5 Large deviations

Let $X_1, X_2, \ldots$ be i.i.d. random variables with state space $\mathcal{X}$ and common distribution $p(x)$. Suppose we have collected a sample $x_1, x_2, \ldots, x_N$. Then the *empirical distribution* on $\mathcal{X}$ is given by

$$\mu_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{xx^i}.$$

If we are interested in computing the expected value $\sum_{x\in\mathcal{X}} f(x)p(x)$. For N large, we can approximate the latter with

$$\sum_{x\in\mathcal{X}} f(x)\mu_N(x) = \frac{1}{N}\sum_i f(x_i).$$

Consider now N Markov chains $X^i, i = 1, 2, \ldots, N$, all having joint distribution Π_0^T on the (discrete) interval $[0, T]$. These may be thought of as *random vectors* with values in $\mathcal{X}^{T+1}$. Suppose the X^i are independent and suppose that the empirical marginal observed at time T is indeed $\rho \neq \pi(T)$. If N is large (say of the order of Avogadro's number), some form of the law of large numbers applies and we conclude that the systems have evolved in an unlikely way. Of all these unlikely ways, though, which one is the most likely?

In modern probabilistic language, this is a problem of *large deviations of the empirical distribution* μ_N on $\mathcal{X}^{T+1}$ given by

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^i}, \quad N = 1, 2, \dots \tag{8.16}$$

where δ_{X^i} assigns probability one to the trajectory X^i (notice that (8.16) may be viewed as a random variable with values in the simplex of distributions on $\mathcal{X}^{T+1}$). Since $\pi_0^T \notin \mathcal{D}_{\mathcal{M}}(\rho)$, by the law of large numbers, the probability of observing $\mu_N \in \mathcal{D}_{\mathcal{M}}(\rho)$ tends to zero. Nevertheless, for a fixed large N, since we have observed ρ at time T, we know that $\mu_N \in \mathcal{D}_{\mathcal{M}}(\rho)$. We ask, which one is its most probable form? The answer is provided by Sanov' theorem (cf. e.g. [1, Theorem 8.2]) which, loosely speaking, says that the probability of observing $\mu_N \in \mathcal{D}_{\mathcal{M}}(\rho)$ decays as

$$\exp\left[-N \inf\{\mathbb{D}(P_0^T \| \Pi_0^T); P_0^T \in \mathcal{D}_{\mathcal{M}}(\rho)\}\right]. \tag{8.17}$$

Thus, solving the Maximum Entropy Problem 8.9 also solves the large deviation problem! Schrödinger, who clearly formulated the problem in 1931/32 [29, 30] as a problem of large deviations of the empirical distribution[2] solved the problem for Brownian particles (see Example 5.3.1) thanks to his intuition. He was motivated by same striking analogies to quantum mechanics:"Merkwürdige Analogien zur Quantenmechanik, die mir sehr des Hindenkens wert erscheinen". Indeed, when both initial and final marginals are fixed, the Schrödinger bridge from ρ_T to ρ_0 is just the time reversal of the Schrödinger bridge from ρ_0 to ρ_T resembling the time reversibility of quantum mechanics.

8.6 The principle of minimum dissipation

Let us now go back to the setting of Section 8.2. Let Π_0^T and $\bar{\Pi}_0^T$ denote the joint distribution of $\{X(0), X(1), \dots, X(T)\}$ when the initial distribution is $\pi(0)$ and the Boltzmann distribution $\bar{\pi}$, respectively. Let $\mathcal{D}_{\mathcal{M}}(\pi(0))$ denote the family of Markovian distributions on $\mathcal{X}^{T+1}$ that have marginal $\pi(0)$ at

[2]Notice that the probabilistic formalism for such an abstract problem was not available at that time!

time 0. Consider the following
Maximum Entropy Problem:

$$\text{minimize} \quad \{\mathbb{D}(P_0^T \| \bar{\Pi}_0^T); P_0^T \in \mathcal{D}_{\mathcal{M}}(\pi(0))\}. \tag{8.18}$$

Arguing, in the forward direction of time, as in the proof of Lemma 8.4.2, we get another representation of $\mathbb{D}(P_0^T \| \bar{\Pi}_0^T)$:

$$\mathbb{D}(P_0^T \| \bar{\Pi}_0^T) = \sum_{k=0}^{T} \sum_{i_k} \mathbb{D}(p^u_{i_k i_{k+1}}(k) \| p_{i_k i_{k+1}}) p^u_{i_k}(k) + \mathbb{D}(p(0) \| \pi(0)). \tag{8.19}$$

Again, since both terms in the righ-hand side of (8.19) are nonnegative and the second is invariant over $\mathcal{D}_{\mathcal{M}}(\pi(0))$, we have that a solution to the Maximum Entropy Problem 8.18 is provided by the joint distribution of the Markov chain having marginal $\pi(0)$ at time zero and the same (forward) transition mechanism of $\bar{\Pi}_0^T$, namely p_{ij}. But this is precisely the distribution Π_0^T! Thus, we have establish the following variational result.

Theorem 8.6.1 The evolution of the physical system of Section 8.2 is such that it minimizes $\mathbb{D}(P_0^T \| \bar{\Pi}_0^T)$ over all Markovian evolutions having $\pi(0)$ as initial distribution.

Suppose $\pi_i(t) > 0, \forall i, t$. Then, the relation between the corresponding reverse time transition probabilities

$$q_{ji}(t, \pi(0)) = \frac{g(t+1, j)}{g(t, i)} q_{ji}(\bar{\pi}), \quad g(t, i) = \frac{\bar{\pi}_i}{\pi_i(t)} \tag{8.20}$$

is given by the multiplicative transformation induced by the reverse-time space-time harmonic function, see Lemma 8.6.

We finally want to show that Problems (8.9)-(8.18) may be reformulated as *optimal stochastic control* problems (corresponding results for continuous time Markov chains are contained in a Master Thesis by C. M. Pelaggi supervised by F. Guerra in 1988 [27]). Consider a finite Markov chain X with transition matrix P. Suppose that at each time $t > 1$ we are able to modify the transition probabilities of the chain by a suitable control action u. In control problems, it is important to specify what is the information available to the controller to design the control strategy. We assume that at time t, the control action is only based on the knowledge of the state $X(t)$ (*feedback*

or *Markov controls*). It can be showed that the controlled process is still a Markov chain. We denote by p_{ij}^u the transition mechanism of the controlled process X^u which, in general, is not time-homogeneous. In control theory, one seeks to find a control strategy (*policy*) in a suitable class so that the resulting evolution meets certain specifications. In optimal control, the specification is to minimize a given cost function depending on the control and on the state.

Consider a finite Markov chain X. Let Π_{sx}^T be the joint distribution of $\{X(s), X(s+1), \ldots, X(T)\}$ where $X(s) = x \in \mathcal{X}$ with probability one ($\Pi_{sx}^T = \Pi_s^T(\cdot|X(s) = x)$). Let p_{ij} be the (forward) transition probabilities and $\pi(t)$ denote the marginal at time t. For a given feedback control u, we denote by $\pi^u(t)$ the distribution of the Makov chain with δ_{xi} as distribution at time s and transition probabilities p_{ij}^u. Consider the following
Stochastic Control Problem: Find a feedback control u that minimizes

$$J(u) = \sum_{k=s}^{T} \sum_{i_k} \mathbb{D}\left(p_{i_k i_{k+1}}^u(k) \| p_{i_k i_{k+1}}\right) p_{i_k}^u(k) - \sum_{i_T} p_{i_T}^u(T) \ln\left(\frac{\rho_{i_T}}{\pi_{i_T}(T)}\right). \tag{8.21}$$

Notice that $J(\cdot)$ in (8.21) features a "running cost" and a "terminal cost" (a functional of this form is called *Bolza Functional*). To facilitate the solution of this problem, let us introduce the space-time harmonic function (for the uncontrolled dynamics) φ defined by the recursion

$$\varphi(t,i) = \sum_j p_{ij}\varphi(t+1,j), \quad \varphi(T,j) = \frac{\rho_j}{\pi_j(T)}. \tag{8.22}$$

Assume that $\varphi(t,i) > 0, \forall i, \forall\, s \le t \le T$. Observe now that $\varphi(s,x)$ does not depend on u. Hence, Problem 8.21 is equivalent to minimizing $I(u) := J(u) + \ln\varphi(s,x)$. Adding and subtracting $\sum_{k=s}^T \sum_{i_k} \ln\varphi(k,i_k) p_{i_k}^u(k)$ and observing that

$$\begin{aligned}
&\sum_{i_k} \ln\varphi(k,i_k) p_{i_k}^u(k) - \sum_{i_{k+1}} \ln\varphi(k+1,i_{k+1}) p_{i_{k+1}}^u(k+1) \\
&= \sum_{i_k} \ln\varphi(k,i_k) p_{i_k}^u(k) - \sum_{i_{k+1}} \ln\varphi(k+1,i_{k+1}) \sum_{i_k} p_{i_k i_{k+1}}^u(k) p_{i_k}^u(k) \\
&= \sum_{i_k}\sum_{i_{k+1}} p_{i_k i_{k+1}}^u(k) \ln\varphi(k,i_k) p_{i_k}^u(k) - \sum_{i_{k+1}} \ln\varphi(k+1,i_{k+1}) \sum_{i_k} p_{i_k i_{k+1}}^u(k) p_{i_k}^u(k) \\
&= \sum_{i_k}\sum_{i_{k+1}} p_{i_k i_{k+1}}^u(k) \left[\ln\varphi(k,i_k) - \ln\varphi(k+1,i_{k+1})\right] p_{i_k}^u(k),
\end{aligned}$$

we get

$$I(u) = \sum_{k=s}^{T} \sum_{i_k} \mathbb{D}\left(p^u_{i_k i_{k+1}}(k) \| p_{i_k i_{k+1}} \frac{\varphi(k+1, i_{k+1})}{\varphi(k, i_k)} \right) p^u_{i_k}(k).$$

Define

$$p^{u^*}_{ij} = p_{ij} \frac{\varphi(t+1, j)}{\varphi(t, i)}. \tag{8.23}$$

Notice that $p^{u^*}_{ij} \geq 0$ and, by (8.22),

$$\sum_j p^{u^*}_{ij} = \sum_j p_{ij} \frac{\varphi(k+1, j)}{\varphi(k, i)} = \frac{\varphi(k, i)}{\varphi(k, i)} = 1.$$

We conclude that an optimal evolution is provided by the transition probabilities (8.23). It is a simple calculation to verify that for all times $s < t \leq T$ indeed

$$\varphi(t) = \frac{\pi^u(t)}{\pi(t)}.$$

Hence, we have obtained the same solution as in Problem 8.9.
Finally, let us find the stochastic control problem equivalent to Problem 8.18. Let Π^T_s and $\bar{\Pi}^T_s$ denote as before the joint distribution of $\{X(s), X(1), \ldots, X(T)\}$ when the distribution at time $t = 0$ is $\pi(0)$ and the Boltzmann distribution $\bar{\pi}$, respectively. Let $\Pi^{Tx}_s = \Pi^T_s(\cdot | X(T) = x)$ and $\bar{\Pi}^{Tx}_0 = \bar{\Pi}^T_s(\cdot | X(T) = x)$. Consider the reverse-time stochastic control problem of minimizing

$$J_r(u) = \sum_{k=1}^{T} \sum_{i_k} \mathbb{D}\left(q^u_{i_k i_{k-1}}(k-1) \| q_{i_k i_{k-1}}(\bar{\pi}) \right) p^u_{i_k}(k) - \sum_{i_s} p^u_{i_s}(s) \ln \left(\frac{p^u_{i_s}(s)}{\bar{\pi}_{i_s}} \right). \tag{8.24}$$

Assume $\pi_j(t) > 0, \forall j, \forall s \leq t < T$. By Lemma , the function

$$g(t, j) := \frac{\bar{\pi}_j}{\pi_j(t)}$$

is reverse-time space-time harmonic, i.e. $g(t+1, j) = \sum_i q_{ji}(t, \pi(0)) g(t, i)$. Also observe that the problem is equivalent to minimizing $I_r(u) := J_r(u) - \ln g(T, x)$. Adding and subtracting $\sum_{k=s}^{T} \sum_{i_k} \ln g(k, i_k) p^u_{i_k}(k)$, we get

$$I_r(u) = \sum_{k=s}^{T} \sum_{i_k} \mathbb{D}\left(q^u_{i_k i_{k-1}}(k-1) \| q_{i_k i_{k-1}}(\bar{\pi}) \frac{g(k, i_{k+1})}{g(k-1, i_k)} \right) p^u_{i_k}(k).$$

By (8.20),

$$q_{ji}^{u^*}(t) = q_{ji}(\bar{\pi})\frac{g(t+1,j)}{g(t,i)} = q_{ji}(t,\pi(0)).$$

Again, we get the same solution as in Problem 8.18. Since the minimum of J_r is zero, we get the following inequality

$$-\ln\frac{\pi_x(T)}{\bar{\pi}_x} \leq \sum_{k=1}^{T}\sum_{i_k}\mathbb{D}\left(q_{i_k i_{k-1}}^{u}(k-1)\|q_{i_k i_{k-1}}(\bar{\pi})\right)p_{i_k}^{u}(k) - \sum_{i_s}p_{i_s}^{u}(s)\ln\left(\frac{p_{i_s}^{u}(s)}{\bar{\pi}_{i_s}}\right).$$

Recalling that the free energy density and $-kT\ln g(t,j) = -kT\ln\frac{\bar{\pi}_j}{\pi_j(t)}$ differ by a constant (8.8), we see that the above inequality may be interpreted as a *principle of minimum dissipation* .

8.7 The Feynman-Kac formula

Let $\mathcal{X}$ be N dimensional and let $0 \leq t < T$. Let W_{ti} be the distribution on $\mathcal{X}^{T-t+1}$ that assigns equal probability to each sequence of the form $(i, i_{t+1}, \ldots, i_{T-1}, i_T)$. Hence,

$$W_{ti}\left(i_t, i_{t+1}, \ldots, i_{T-1}, i_T\right) = \delta_{ii_t}\cdot\frac{1}{N^{T-t}}.$$

We can then think of W_{ti} as the distribution induced on $\mathcal{X}^{T-t+1}$ by a sequence of i.i.d. random variables (random walk) $\{W(t), W(t+1), \ldots, W(T-1), W(T)\}$ conditioned to start in i at time t ($\frac{1}{N}$ may be viewed as the constant transition probability between any two states for such a process).

Theorem 8.7.1 Let the function h satisfy the backward recursion

$$h(s,i) = \sum_j \frac{1}{N}h(s+1,j)\exp[-V(s+1,j)], \quad h(T,i) = h_T(i), \tag{8.25}$$

where h_T is non negative and not identically zero. Then h admits the following probabilistic representation *(Feynman-Kac formula)*:

$$h(t,i) = \sum_{(i_t,\ldots,i_T)}\left\{h_T(i_T)\exp\left[-\sum_{s=t}^{T-1}V(s+1,i_{s+1})\right]\right\}W_{ti}(i_t,\ldots,i_T). \tag{8.26}$$

Proof . Observe that $h(s,i) > 0, \forall s \in [0,T], \forall i \in \mathcal{X}$. Consider any Markov chain $X = \{X(0), X(1), \ldots\}$ with state space $\mathcal{X}$ and transition probabilities

$$p_{ij}(s) = \frac{1}{N} \frac{h(s+1,j)}{h(s,i)} \exp[-V(s+1,j)], \quad s = 0, 1, \ldots, T-1. \tag{8.27}$$

In view of (8.25), we indeed have $\sum_j p_{ij}(s) = 1, \forall s \in [0, T-1]$. Consider now the identity

$$h(t, X(t)) = h_T(X(T)) \exp\left\{ -\sum_{s=t}^{T-1} [\log h(s+1, X(s+1)) - \log h(s, X(s))] \right\}. \tag{8.28}$$

Let P be the joint distribution of $X(t), X(t+1), \ldots, X(T)$, and let $P_{ti} = P(\cdot | X(t) = i)$. Taking the conditional expectation on both sides of (8.28) given that $X(t) = i$ and using (8.27), we get

$$\begin{aligned}
h(t,i) &= \mathbb{E}(h(t, X(t) | X(t) = i) \\
&= \mathbb{E}_{P_{ti}} \left(h_T(X(T)) \exp\left\{ -\sum_{s=t}^{T-1} [\log h(s+1, X(s+1)) - \log h(s, X(s))] \right\} \right) \\
&= \sum_{(i_t, \ldots, i_T)} h_T(i_T) \exp\left\{ -\sum_{s=t}^{T-1} [\log h(s+1, i_{s+1}) - \log h(s, i_s)] \right\} \delta_{ii_t} p_{i_t i_{t+1}}(t) \cdots p_{i_{T-1} i_T}(T-1) \\
&= \sum_{(i_t, \ldots, i_T)} h_T(i_T) \exp\left\{ \sum_{s=t}^{T-1} \log \left[p_{i_s i_{s+1}}(s) \cdot \frac{h(s, i_s)}{h(s+1, i_{s+1})} \right] \right\} \delta_{ii_t} \\
&= \sum_{(i_t, \ldots, i_T)} h_T(i_T) \exp\left\{ \sum_{s=t}^{T-1} \log \left[\frac{1}{N} \exp[-V(s+1, i_{s+1})] \right] \right\} \delta_{ii_t} \\
&= \sum_{(i_t, \ldots, i_T)} h_T(i_T) \exp\left[-\sum_{s=t}^{T-1} V(s+1, i_{s+1}) \right] \frac{1}{N^{(T-t)}} \delta_{ii_t} \\
&= \sum_{(i_t, \ldots, i_T)} \left\{ h_T(i_T) \exp\left[-\sum_{s=t}^{T-1} V(s+1, i_{s+1}) \right] \right\} W_{ti}(i_t, \ldots, i_T).
\end{aligned}$$

□

Feynman-Kac formulae are related to a *change of probability distribution on path space* (here the space of finite trajectories $(i_t, i_{t+1}, \ldots, i_T)$ in $\mathcal{X}^{T-t+1}$).

Indeed, the same calculations as in the above proof yield

$$P_{ti}(i_t,\ldots,i_T) = \frac{h_T(i_T)}{h(t,i_t)}) \exp\left[-\sum_{s=t}^{T-1} V(s+1,i_{s+1})\right] W_{ti}(i_t,\ldots,i_T),$$

from which formula (8.26) follows by simply summing over all "trajectories" $(i_t, i_{t+1}, \ldots, i_T)$ in $\mathcal{X}^{T-t+1}$. Since W_{ti} is just the uniform distribution on $\mathcal{X}^{T-t+1}$ "pinned" at t, we recognize a strong analogy with the Boltzmann distribution (8.2) where $h(t,i)$ plays the role of the normalizing constant.

Feynman-Kac formulae originated (in the continuous time setting) in quantum and statistical physics. They occur nowadays in sequential Monte Carlo methods (see Section 10.4) that are applied in a variety of fields such as: Statistics, applied probability, Bayesian estimation, directed polymer simulation, genetic and genealogical population models, signal processing, network analysis, etc., see [8].

Problems

Problem 62 Consider a Markov chain $X = \{X(0), X(1), \ldots\}$ with N dimensional state space $\mathcal{X}$ and transition matrix $P = (p_{ij})$. Let μ and ν two probability distributions on $\mathcal{X}$ satisfying $Supp(\nu) \subseteq Supp(\mu)$. Introduce the flow of distributions induced by P

$$\mu(t) := (P^*)^t \mu, \quad \nu(t) := (P^*)^t \nu.$$

Prove that

$$\mathbb{D}(\nu(t+1)\|\mu(t+1)) \leq \mathbb{D}(\nu(t)\|\mu(t)).$$

Observe that, in the case when μ is stationary for the chain ($\mu = P^*\mu$), this represents a generalization of Theorem 8.2.1.

Problem 63 Consider a Markov chain $X = \{X(0), X(1), \ldots\}$ with N dimensional state space $\mathcal{X}$ and doubly stochastic (Observation 7.3.10) transition matrix $P = (p_{ij})$. Let μ be any distribution on $\mathcal{X}$, and consider the corresponding flow $\mu(t) := (P^*)^t \mu$. Prove that the Shannon entropy is non-decreasing along the flow, namely

$$H(\mu(t+1)) \geq H(\mu(t)).$$

Problem 64 Let $Y = \{Y(0), Y(1), \ldots\}$ be a Markov chain with N dimensional state space $\mathcal{X}$. Let π_{ij} be its transition probabilities. Let g satisfy the backward equation

$$g(s,i) = \sum_j \pi_{ij} g(s+1,j) \exp[-V(s+1,j)], \quad g(T,i) = g_T(i), \tag{8.29}$$

where $g_T(i) \geq 0, \forall i \in \mathcal{X}$. Let Π be the joint distribution of $Y(t), Y(t+1), \ldots, Y(T)$, and let $\Pi_{ti} = \Pi(\cdot|Y(t) = i)$. Establish the following extension of the Feynman-Kac formula

$$g(t,i) = \sum_{(i_t,\ldots,i_T)} \left\{ g_T(i_T) \exp\left[-\sum_{s=t}^{T-1} V(s+1, i_{s+1}) \right] \right\} \Pi_{ti}(i_t, \ldots, i_T).$$

Chapter 9

Recurrence and ergodicity

9.1 Communication classes. Closed sets

Consider a Markov chain $X = \{X(0), X(1), \ldots\}$ with state space $\mathcal{X}$.

Definition 9.1.1 We say that state j is *accessible* from i if there exists a time $n \geq 0$ such that

$$p_{ij}^{(n)} = \mathbb{P}(X(t+n) = j | X(t) = i) > 0,$$

where we set

$$p_{ij}^{(0)} = \delta_{ij}.$$

We say that states i and j *communicate* and write $i \sim j$ if i is accessible from j and j is accessible form i (notice that every state i communicates with itself since $p_{ii}^{(0)} = 1$). This is an *equivalence relation* that induces a partition of the states of $\mathcal{X}$ into equivalence classes called *communication classes*. When there is only one communication class the Markov chain is called *irreducible*.

Example 9.1.2 In Example 7.2.7, state 0 is accessible from 1 but not viceversa. There are three communication classes $E_1 = \{0\}, E_2 = \{1, 2, \ldots, N-1\}, E_3 = \{N\}$. The migration model with (7.2), and the processes in Examples 7.2.8, 7.3.4, 7.3.7 are all irreducible chains.

Definition 9.1.3 A state i has *period* $d(i)$ if the following properties are verified:

1. $p_{ii}^{(n)} > 0 \Rightarrow d(i)$ divides n;

2. $d(i)$ is the largest integer such that 1. holds.

If $p_{ii}^{(n)} = 0, \forall n \geq 1$, we set $d(i) = 0$. When $d(i) = 1$, i is called *aperiodic*. It is not difficult to show that if $i \sim j$, then i and j have the same period.

Example 9.1.4 The Ehrenfest model 7.2.8 yields an irreducible chain of period 2.

Definition 9.1.5 A set $C \subseteq \mathcal{X}$ of states is called *closed* if no state outside of C is accessible from any state in C, i.e. $\sum_{j \in C} p_{ij} = 1$, $\forall i \in C$. The smallest closed set containing a set E of states is called the *closure* of E. A single state i that forms a closed set is called *absorbing*.

Observation 9.1.6 Each closed set of states C corresponds to a sub-chain in the following sense. If we delete all rows and columns of P corresponding to states outside of C we get another stochastic matrix.

Observation 9.1.7 Notice that communication classes are not necessarily closed: For instance $E_2 = \{1, 2, \ldots, N-1\}$ in Example 7.2.7 is not closed. It is indeed possible from one communication class to enter another but then it is not possible to come back. Conversely, states in a closed set need not communicate, see Example 9.2.6 below.

Observation 9.1.8 All properties discussed in this section are *topological*, namely they only concern the transition graph.

9.2 Classification of states.

For a a Markov chain $X = \{X(t), t \geq 0\}$ with state space $\mathcal{X} = \mathbb{N}$, we introduce the probability

$$f_{ij}^{(n)} := \mathbb{P}(X(n) = j, X(k) \neq j, k = 1, 2, \ldots, n-1 | X(0) = i).$$

Observing that

$$f_{ij}^{(n)} = p_{ij}^{(n)} - \sum_{k=1}^{n-1} f_{ij}^{(k)} p_{jj}^{(n-k)}, \quad p_{ij}^{(0)} = \delta_{ij},$$

we get the relation

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}. \tag{9.1}$$

Let us define

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)},$$

the probability, starting fron i, of ever reaching the state j and

$$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)},$$

the probability, starting from i, of ever returning to i. It can be shown that if $i \sim j$, $i \neq j$, $f_{ij} = f_{ji} = 1$, cf [31, pp. 536-538].

Definition 9.2.1 A state i is called *recurrent* if $f_{ii} = 1$ and *transient* if $f_{ii} < 1$.

Theorem 9.2.2 1. State i is recurrent if and only if

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty. \tag{9.2}$$

2. If i is recurrent and $i \sim j$, then j is recurrent.

Proof . From (9.1), we get

$$\begin{aligned}\sum_{n=1}^{\infty} p_{ii}^{(n)} &= \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \sum_{n=k}^{\infty} p_{ii}^{(n-k)} \\ &= f_{ii} \sum_{m=0}^{\infty} p_{ii}^{(m)} = f_{ii}\left(1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}\right).\end{aligned}$$

If $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$, it follows that $f_{ii} < 1$, namely i is transient. Suppose now that $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$. In view of (9.1), we have

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{n=1}^{N}\sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{n=k}^{N} p_{ii}^{(n-k)} \leq \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{m=0}^{N} p_{ii}^{(m)}.$$

It follows that

$$f_{ii} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \geq \sum_{k=1}^{N} f_{ii}^{(k)} \geq \frac{\sum_{n=1}^{N} p_{ii}^{(n)}}{\sum_{m=0}^{N} p_{ii}^{(m)}} \to 1, \quad \text{as } N \to \infty,$$

where the last convergence occurs because of the hypothesis. Hence $f_{ii} = 1$ and i is recurrent. To prove 2., assume $i \sim j$. Then there exist t_1 and t_2 such that $p_{ij}^{(t_1)} > 0$ and $p_{ji}^{(t_2)} > 0$. Moreover,

$$p_{ii}^{(n+t_1+t_2)} \geq p_{ij}^{(t_1)} p_{jj}^{(n)} p_{ji}^{(t_2)}.$$

From these two properties it follows that

$$\sum_{n=1}^{\infty} p_{jj}^{(n)} = \infty \Rightarrow \sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty,$$

namely j recurrent implies i recurrent. □

Corollary 9.2.3 Suppose j is a transient state. Then

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty, \quad \forall i.$$

The latter in turn implies $\lim_{n\to\infty} p_{ij}^{(n)} = 0, \forall i$.

Proof .

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=k}^{\infty} p_{jj}^{(n-k)}$$
$$= f_{ij} \sum_{m=0}^{\infty} p_{jj}^{(m)} \leq \sum_{m=0}^{\infty} p_{jj}^{(m)} < \infty,$$

where we have used the fact that the probability f_{ij} is bounded by one. □

Definition 9.2.4 For a recurrent state i the *mean recurrence time* is

$$\mu_i = \sum_{n=1}^{\infty} n \cdot f_{ii}^{(n)}.$$

A recurrent state is called *positive* if $\mu_i < \infty$ and *null* otherwise. A positive recurrent aperiodic state is called *ergodic*.

Theorem 9.2.5 Let j be recurrent aperiodic ($d(j) = 1$). Then

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{f_{ij}}{\mu_j}. \tag{9.3}$$

In particular, if $i \sim j$, we have

$$\lim_{n\to\infty} p_{ij}^{(n)} = (\mu_j)^{-1}, \tag{9.4}$$

where, in the case of a null recurrent state, $(\mu_j)^{-1}$ is set equal to zero. The latter result includes the important particular case

$$\lim_{n\to\infty} p_{jj}^{(n)} = (\mu_j)^{-1}. \tag{9.5}$$

Let j be recurrent of period $k = d(j) > 1$, then

$$\lim_{n\to\infty} p_{jj}^{(nk)} = \frac{k}{\mu_j}. \tag{9.6}$$

The proof is based on a nontrivial result of analysis and is therefore omitted, see [10, Theorems 3,4, Chapter XIII].

Recall that the binomial distribution

$$\pi_i = \begin{pmatrix} N \\ i \end{pmatrix} \cdot 2^{-N}$$

is stationary for the Ehrenfest model 7.2.8. For N large, we can approximate the binomial distribution with a Gaussian distribution. It follows that, asymptotically, we are almost sure to find approximatively one half of balls in each container. For instance, with 10^6 balls, the probability of finding asymptotically more than 505.000 balls in one container is of the order of 10^{-23}. The mean recurrence times of improbable states (such as all balls in one container) are enormous with respect to states where balls are approximatively evenly split between the two containers, see Problem 71.

Example 9.2.6 *11. Finite chain*
Consider a chain with state space $\mathcal{X} = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$ and transition

matrix

$$P = \begin{bmatrix} * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & 0 & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & * & 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & 0 & 0 \end{bmatrix},$$

where $*$ stands for a positive transition probability. First of all, we observe that there are three closed sets $\{0\}, \{1, 2\}, \{3, 4, 5\}$ corresponding to the three stochastic submatrices

$$[*], \quad \begin{bmatrix} 0 & * \\ * & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & * & * \\ 0 & * & * \\ * & 0 & 0 \end{bmatrix}.$$

State $\{0\}$ is absorbing and therefore recurrent. States 1 and 2 are recurrent and periodic of period 2 with mean recurrence time equal to 2. States $3, 4.5$ are ergodic (as we shall see, in a finite chain null recurrent states are impossible). States $6, 7, 8$ are transient since from 6 is is possible to move to the closed sets. From 7 and 8 the chain will eventually pass into 6 and never return.

Theorem 9.2.7 In an irreducible Markov chain, all states belong to the same class: They are either all transient, or all null recurrent, or all positive recurrent. In any case, they all have the same period. The recurrent states can be assigned in a unique way to closed sets $C_1, C_2, \ldots$ which are equivalent classes. Besides the recurrent states, there may be transient states from which the C_k may be reached.

Example 9.2.8 Consider again Example 7.3.5 with $q_0 \in (0, 1)$ and $q_0 + q_1 < 1$. Then the chain is irreducible. If $\alpha = \sum_{i=0}^{\infty} k \cdot q_k > 1$, all states are transient. If $\alpha < 1$, all states are positive recurrent. If $\alpha = 1$, all states are null recurrent see [17, Section 3.5].

Example 9.2.9 Consider again the random walk on $\mathbb{Z}$ of Example 7.2.6, see Figure 9.1: We move one step to the right with probability p and one step

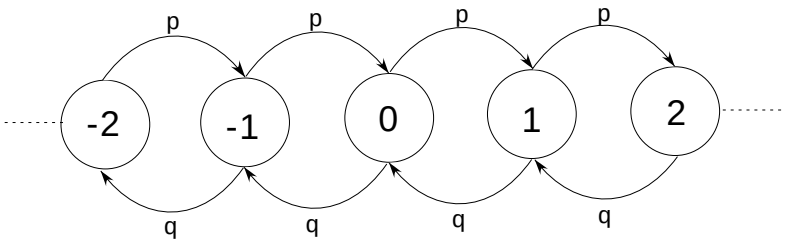

Figure 9.1: Random walk on $\mathbb{Z}$.

to the left with probability $q = 1 - p$. This chain is clearly irreducible. By arguing as in Example 9.3.1 below, one can show that if $p = q = 1/2$, all states are (null) recurrent, if $p \neq q$, all states are transient!

Observation 9.2.10 In an irreducible chain, $p_{jj} > 0$ for some j implies that the chain is not periodic.

Example 9.2.11 Consider a chain with transition matrix

$$P = \begin{bmatrix} q_0 & p_0 & 0 & 0 & 0 & \dots \\ q_1 & 0 & p_1 & 0 & 0 & \dots \\ q_2 & 0 & 0 & p_2 & 0 & \dots \\ . & . & . & . & . & \dots \end{bmatrix}, \quad p_i + q_i = 1, \quad p_i \in (0,1), \quad i = 0, 1, \dots$$

This is simply a generalization of the success runs of Example 7.3.7. Since $p_i \neq 0$ and $q_i \neq 0$, $\forall i$, the chain is irreducible. In view of the above theorem, it suffices to analyze the nature of one state: We pick the 0 state. There is only one way to go in n steps from 0 to 0. Hence,

$$f_{00}^{(1)} = q_0, \quad f_{00}^{(n)} = p_0 \cdot p_1 \cdot \ldots \cdot p_{n-2} \cdot q_{n-1}, n \geq 1.$$

Define $x_0 = 1$ and $x_n = p_0 \cdot p_1 \cdot \ldots \cdot p_{n-1}$. Then

$$f_{00} = \sum_{n=1}^{\infty} f_{00}^{(n)} = \sum_{n=1}^{\infty}(x_{n-1} - x_n) = \lim_{m\nearrow\infty}(1 - x_m) = \lim_{m\nearrow\infty} = 1 - \lim_{m\nearrow\infty}\prod_{i=0}^{m-1} p_i.$$

By a well-known theorem on infinite products [5, Theorem 1.9, p. 422], we have

$$\lim_{m\nearrow\infty}\prod_{i=0}^{m-1} p_i = 0 \Leftrightarrow \sum_{i=0}^{\infty}(1 - p_i) = \infty.$$

We conclude that 0 (with all other states) is recurrent if and only if $\sum_{i=0}^{\infty} q_i = \infty$. For instance, if $q_i = \frac{1}{i+1}, i \geq 1$, all states are recurrent. If instead $q_i = \frac{1}{i^2+1}, i \geq 1$, all states are transient.

Corollary 9.2.12 In a finite Markov chain (chain with finite state space) there are no null recurrent states. A class is recurrent if and only if the chain has no way to leave it. It is impossible that all states be transient. Hence, there is at least one positive recurrence class.

Proof . First of all, observe that it suffices to prove the result for irreducible chains. Suppose the states are all null recurrent. Then, by Theorem 9.2.5, $p_{ij}^{(n)} \to 0, \forall i, \forall j$. The same happens if all states are transient. The rows of P^n, however, sum to one. Hence, its elements cannot all tend to zero. □

Thus, after possible renumbering of the states, the transition matrix of a finite chain has necessarily the structure (see Example 9.2.6):

$$P = \begin{bmatrix} P_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & P_2 & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & P_m & 0 \\ A_1 & A_2 & . & . & . & A_{m+1} \end{bmatrix},$$

While P_k and A_{m+1} are square, $A_1, \ldots, A_m$ need not be. The structure of P^n is similar

$$P^n = \begin{bmatrix} P_1^n & 0 & 0 & 0 & 0 & 0 \\ 0 & P_2^n & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & P_m^n & 0 \\ * & * & . & . & . & A_{m+1}^n \end{bmatrix},$$

Theorem 9.2.13 (Ergodic theorem for Markov chains) Consider a Marhov chain $X = \{X(0), X(1), \ldots\}$ with state space $\mathcal{X}$. Then

1. There exists a stationary distribution if and only if there exists at least one positive recurrent class. In the latter case, all stationary distributions π are such that $\pi(j) = 0$ for all j transient or null recurrent.

2. There exists a unique stationary distribution if and only if there exists a unique positive recurrent class C. In that case, for $j \in C$, we have

$$\pi(j) = \frac{1}{\mu_j}.$$

   Moreover, if $f : \mathcal{X} \to \mathbb{R}$ satisfies

$$\sum_{i \in \mathcal{X}} |f(i)| \pi_i < \infty,$$

   then

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f(X(k)) = \sum_{i \in \mathcal{X}} f(i) \pi_i. \tag{9.7}$$

3. If there are more recurrent classes C_α, let π_α be the distribution on C_α given by

$$\pi_\alpha(j) = \frac{1}{\mu_j}.$$

   Then all stationary distributions are given by "mixtures" of the π_α (the simplex induced by the π_α).

4. The limit

$$\lim_{n \to \infty} (P^*)^n \, \pi(0)$$

   exists independent from $\pi(0)$ if and only if there is a unique positive recurrent, aperiodic class. In that case, every row of P^n tends to π^*.

Corollary 9.2.12 and Theorem 9.2.13 imply the following important fact.

Corollary 9.2.14 A finite Markov chain always has at least one stationary distribution.

The following results complements Theorem 9.2.13.

Theorem 9.2.15 (Perron-Frobenius) Consider a finite Markov chain X with transition matrix P and state space $\mathcal{X} = \{0, 1, \ldots, N\}$. Suppose there exists an integer $m \geq 1$ such that all elements of P^m are strictly positive. Then the limit

$$\lim_{n\to\infty} p_{jk}^{(n)} = \pi_k$$

exists $\forall j, k$, is independent from j, $\pi_k > 0, \forall k = 0, 1 \ldots, N$, $\pi = P^*\pi$ and $\sum_{k=0}^{N} \pi_k = 1$ (namely π is a stationary distribution). Such a chain is called ergodic.

Theorem 9.2.16 (Strong Law of Large Numbers for finite Chains) Consider a finite Markov chain X. Suppose there exists an integer $m \geq 1$ such that all elements of P^m are strictly positive and let π be the stationary distribution. Let $N_n(j)(\omega)$ denote the number of times j appears in $X(0)(\omega), X(1)(\omega), \ldots, X(n-1)(\omega)$. Then

$$\lim_{n\to\infty} \frac{N_n(j)(\omega)}{n} = \pi_j, \quad \text{a.s.} \tag{9.8}$$

9.3 Analysis of examples

We apply the results of the previous section to various examples.

Example 9.3.1 *12. Random walk with reflection at zero*
Consider the following variant of Examples 7.2.6 and 7.3.4. We have a random walk with transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \ldots \\ q & 0 & p & 0 & 0 & \ldots \\ 0 & q & 0 & p & 0 & \ldots \\ . & . & . & . & . & \ldots \end{bmatrix}.$$

with graph as in Figure 9.2

There is a unique communicating class, namely the chain is irreducible. It has period two. Let us compute f_{i1}, namely the probability, starting from i, of reaching 1. We have

$$f_{i1} = qf_{(i-1)1} + pf_{(i+1)1}, \quad i > 1. \tag{9.9}$$

Since

$$q\left(\frac{q}{p}\right)^{i-2} + p\left(\frac{q}{p}\right)^{i} = q\left(\frac{q}{p}\right)^{i-2}\left[1 + \frac{q}{p}\right] = \left(\frac{q}{p}\right)^{i-1},$$

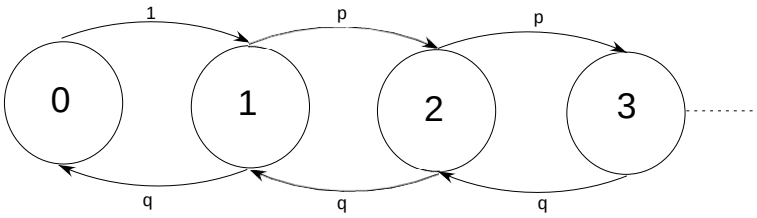

Figure 9.2: Random walk with reflection at zero.

we get that $\left(\frac{q}{p}\right)^{i-1}$ satisfies (9.9). If $p > q$, $f_{i1} = \left(\frac{q}{p}\right)^{i-1} < 1$. Hence, all states are transient and there is no stationary distribution ($p_{ij}^{(n)} \to 0, \forall i, j$). It can be shown [5, p.151] that, starting from state $i \geq 1$ there is the positive probability $1 - (q/p)^i$ of never visiting state 0.

If $q \geq p$, the solution of (9.9) is $f_{i1} = 1, \forall i > 1$ (f_{i1} is a probability and cannot be larger than 1!). Hence, all states are recurrent. We consider now the system of equations originating from equation (7.15) $\pi = P^*\pi$. We get

$$\begin{aligned}
\pi(0) &= q\pi(1), \\
\pi(1) &= \pi(0) + q\pi(2), \\
\pi(2) &= p\pi(1) + q\pi(3), \\
\pi(3) &= p\pi(2) + q\pi(4), \\
&\dots
\end{aligned}$$

Working our way from the top to the bottom we get

$$\begin{aligned}\pi(1) = q\pi(1) + q\pi(2) &\Rightarrow \pi(1) = \frac{q}{p}\pi(2),\\ \pi(2) = q\pi(2) + q\pi(3) &\Rightarrow \pi(2) = \frac{q}{p}\pi(3),\\ &\cdots\end{aligned}$$

Hence,

$$\pi(j) = \frac{p}{q}\pi(j-1), \quad j = 2, 3, \ldots$$

If $p = q$, we get $\pi(1) = \pi(2) = \ldots$ and there is no stationary distribution. In this case, all states are null recurrent. Finally, if $q > p$, condition

$$\sum_{j=0}^{\infty} \pi(j) = 1$$

yields

$$\pi(1)\left[q + 1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2 + \cdots\right] = 1.$$

The latter gives

$$\pi(1) = \frac{q-p}{2q^2}.$$

In conclusion

$$\begin{aligned}\pi(0) &= \frac{q-p}{2q},\\ \pi(j) &= \frac{q-p}{2q^2}\left(\frac{p}{q}\right)^{j-1}, \quad j \geq 1.\end{aligned}$$

In the latter case $q > p$, all states are positive recurrent and the above π is the unique stationary distribution.

Consider Example 7.2.8. This chain is irreducible with a unique positive recurrent class. The unique stationary distribution is the binomial

$$\pi_i = \left(\begin{array}{c} N \\ i \end{array}\right) \cdot 2^{-N}.$$

Moreover, the chain is reversible, see the solution of Problem 59 in Appendix A.
Consider now the random walk with absorbing barriers of Example 7.2.7. States 0 and N are absorbing and recurrent. The other states are transient ($p_{ij}^{(n)} \to 0, \forall i, \forall 1 \leq j \leq N-1$). Define the distributions $\pi_j^1 = \delta_{0j}$ and $\pi_j^2 = \delta_{Nj}$. By Theorem 9.2.13, all stationary distributions are obtained as

$$\pi = \lambda\pi^1 + (1-\lambda)\pi^2.$$

Similar considerations hold for the genetic example 7.3.8.
In the migration model of the Section 7.1 with transition matrix (7.2), the chain is again irreducible with a unique positive recurrent class. The uniform distribution is the only stationary distribution and the chain is reversible.

Example 9.3.2 *13. Laplace*
This example originates with Laplace. N white balls and N black balls are placed in two urns so that each urn contains N balls. Let $X(t)$ denote the number of black balls in the first urn at time t. At each time, one ball is chosen at random in each urn and moved to the other one.Let us first find the transition probabilities. We have

$$p_{jj} = \frac{j}{N}\cdot\frac{N-j}{N} + \frac{N-j}{N}\cdot\frac{j}{N} = \frac{2j(N-j)}{N^2}$$
$$p_{j(j+1)} = \frac{N-j}{N}\cdot\frac{N-j}{N} = \frac{(N-j)^2}{N^2}$$
$$p_{j(j-1)} = \frac{j^2}{N^2}.$$

This chain is very similar to the Eherenfest model (two reflecting barriers, irreducible, unique positive recurrent class) except that it is aperiodic ($p_{jj} > 0$ except for the two reflecting states 0 and N). A stationary distribution π must satisfy

$$\pi_i = \pi_{i+1}p_{(i+1)i} + \pi_i p_{ii} + \pi_{(i-1)}p_{(i-1)i}$$
$$= \pi_{i+1}\frac{(i+1)^2}{N^2} + \pi_i\frac{2i(N-i)}{N^2} + \pi_{i-1}\frac{(N-i+1)^2}{N^2}. \tag{9.10}$$

Let us consider the distribution

$$\pi_i = \frac{\begin{pmatrix} N \\ i \end{pmatrix}\begin{pmatrix} N \\ N-i \end{pmatrix}}{\begin{pmatrix} 2N \\ N \end{pmatrix}} = \frac{\begin{pmatrix} N \\ i \end{pmatrix}^2}{\begin{pmatrix} 2N \\ N \end{pmatrix}} = \frac{(N!)^4}{(i!)^2((N-i)!)^2(2N)!}.$$

It is just the probability of getting i black balls when N are chosen at random from the set of N black and N white balls (see also Problem 18) . In the right-hand side of (9.10), we get

$$\begin{aligned}
&\frac{(N!)^4}{((i+1)!)^2((N-i-1)!)^2(2N)!}\frac{(i+1)^2}{N^2}+\pi_i\frac{2i(N-i)}{N^2}\\
&\qquad+\frac{(N!)^4}{((i-1)!)^2((N-i+1)!)^2(2N)!}\frac{(N-i+1)^2}{N^2}\\
=\frac{(N!)^4(N-i)^2}{(i!)^2((N-i)!)^2(2N!)N^2}&+\pi_i\frac{2i(N-i)}{N^2}+\frac{(N!)^4i^2}{(i!)^2((N-i)!)^2(2N!)N^2}\\
&=\pi_i\left[\frac{(N-i)^2+2i(N-i)+i^2}{N^2}\right]=\pi_i.
\end{aligned}$$

Thus π is invariant. As for Ehrenfest urn model 7.2.8, the mean recurrence times of improbable states such as all white balls in one urn are enormous.

9.4 Absorption analysis

Given a time-homogeneous Markov chain, it is interesting to compute the probability of absorption by a closed set C (recall that C is closed if $\sum_{j\in C} p_{ij} = 1$ for all $i \in C$). It is also relevant to compute the mean time to absorption. Let us illustrate this analysis on the absorbing states of Example 7.2.7. We follow [5, Section 2.3]. Recall that $X(t)$ represents the capital of player A at time t and the transition matrix P is given by

$$P=\begin{bmatrix}
1 & 0 & 0 & 0 & \dots & 0\\
q & 0 & p & 0 & \dots & 0\\
0 & q & 0 & p & \dots & 0\\
. & . & . & . & . & \dots\\
. & . & . & . & . & \dots\\
0 & 0 & 0 & \dots & 0 & 1
\end{bmatrix},\quad p+q=1.$$

Let $i \in [1, N-1]$ and $N-i > 0$ be the initial fortunes of the players A and B. The state space is then $\mathcal{X} = \{0, 1, \dots, N\}$. The state 0 represents the *ruin of gambler* A and the state N represents the ruin of gambler B. Let T be the duration of the game, namely the first time that t at which $X(t) = 0$ or $X(t) = N$. Notice that T is a random variable taking values in the positive

integers. Consider the absorption probability of state N, namely

$$p(i) := \mathbb{P}(X(T) = N | X(0) = i).$$

Then, considering the situation after one step, we get the following relation

$$p(i) = p \cdot p(i+1) + q \cdot p(i-1), \quad p(0) = 0,\ p(N) = 1. \tag{9.11}$$

This is a simple linear homogeneous difference equation with constant coefficients. The solutions are found via the roots of the characteristic equation

$$px^2 - x + q = 0,$$

see Problem 73 below. Suppose first $p \neq q$. Then, there are two simple roots $x_1 = 1$ and $x_2 = q/p$. The general solution of (9.11) is given by

$$p(i) = c_1(1)^i + c_2 \left(\frac{q}{p}\right)^i$$

Using the boundary conditions, we get

$$p(i) = \frac{1 - \left(\frac{q}{p}\right)^i}{1 - \left(\frac{q}{p}\right)^N}. \tag{9.12}$$

Suppose now that $p = q$. Then the general solution is

$$p(i) = c_1(1)^i + c_2 i(1)^i.$$

The boundary conditions imply that $c_1 = 0$ and $c_2 = 1/N$. Hence

$$p(i) = \frac{i}{N}. \tag{9.13}$$

In both cases, it can be shown that the probability that the game lasts forever is zero (one of the two players wins).

Let us now compute the average duration of the game $m(i) = E(T|X(0) = i)$. For $i \in [1, N-1]$, we get the linear, non homogeneous equation

$$m(i) = 1 + p \cdot m(i+1) + q \cdot m(i-1).$$

This because is tossed at least once and because after that the state will be $i+1$ with probability p and $i-1$ with probability q. The boundary conditions are simply $m(0) = 0$ and $m(N) = 0$. All solution are obtained as solutions of the corresponding homogeneous equation (which is the same as before for the probability (9.11)) plus one particular solution of the non homogeneous. Trying $m(i) = \alpha i^2 + \beta i + \gamma$, we get the conditions

$$\alpha(p - q) = 0, \quad \alpha + \beta(p - q) = -1. \tag{9.14}$$

Suppose $p = q$. Then, $\alpha = -1$. The general solution has the form

$$m(i) = c_1 + c_2 i - i^2 + \beta i + \gamma.$$

Imposing the boundary conditions, we get $c_1 + \gamma = 0$ and $(c_2 + \beta)N - N^2 = 0$. Hence, the solution is

$$m(i) = i \cdot (N - i).$$

As expected, the mean duration of a fair game is maximum if the initial capital of the two players is equal ($i = N/2$). In the case when the state space is infinite and there are infinitely many transient states, there is the possibility of never being absorbed by the recurrent class. For a more refined absorption analysis, see [5, Chapter 4].

Example 9.4.1 *9. Genetics: A diploid model*
Diploid cells have two homologous copies of each chromosome, usually one from the mother and one from the father. Nearly all mammals are diploid organisms. Human diploid cells have 46 chromosomes and human haploid gametes (egg and sperm) have 23 chromosomes. Hereditary characters depend on genes which appear in pairs occupaying the same position in the two paired chromosomes. Let us consider the simplest case: Each gene of a particular pair can assume two different forms *called allele* A and a. Three different *genotypes* AA, Aa, aa can be formes. A cell is said to be *homozygous* for a particular gene when identical alleles of the gene are present on both homologous chromosomes such as AA and aa, *heterozygous* otherwise (Aa). Here A is a *dominant*, meaning that *phenotype* manifests itself in AA and Aa and a is a *recessive*, namely phenotype manifests itself only in aa. The riproductive cells (*gametes*) receive only one gene. For instance, organisms of the pure genotype AA or aa produce only one type of gamete. New organisms originate from two gametes of the parents. Each gene has at each time probability $1/2$ to be transmitted and successive trials are independent.

Thus, we can think of the genotypes of n children as the outcome of n independent tosses of two fair coins. The genotypes of the descendants of a pair $Aa \times Aa$ are AA, Aa, aa with probabilities 1/4, 1/2, 1/4, respectively. A pair $AA \times aa$ can have only descendants Aa. Consider the following
Brother-sister mating problem: The direct descendants of a pair mate between them. As there are three genotypes AA, Aa, aa for the parents, there are precisely

$$\left(\begin{array}{c} 3+2-1 \\ 2 \end{array} \right) = \frac{4!}{2!2!} = 6$$

combinations with repetitions (see Section 4.2.3). We have

$$x_0 = AA \times AA,\ x_1 = AA \times Aa,\ x_2 = Aa \times Aa,\ x_3 = Aa \times aa,\ x_4 = aa \times aa,\ x_5 = AA \times aa.$$

With the above rules, the chain matrix is the following

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 1/16 & 1/4 & 1/4 & 1/4 & 1/16 & 1/8 \\ 0 & 0 & 1/4 & 1/2 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Clearly, x_0 and x_4 are absorbent states. Other states are transient. It is interesting to calculate the probability of absorption by states x_0 and x_4 from the various other states. For instance, starting from state x_1, we have

$$f_{10} = \frac{1}{4} f_{00} + \frac{1}{2} + \frac{1}{4} f_{20} = \frac{1}{4} + \frac{1}{2} f_{10} + \frac{1}{4} f_{20} \Rightarrow f_{10} = \frac{1}{2} + \frac{1}{2} f_{20}.$$

One gets

$$(f_{10}, f_{20}, f_{30}, f_{50}) = \left(\frac{3}{4}, \frac{1}{2}, \frac{1}{4}, \frac{1}{2} \right).$$

9.5 Non-Markovian processes

Many evolutions cannot be modeled by a Markov process. For instance, in finance and demography, *moving averages*

$$Y(t) = \frac{1}{N+1} \left(X(s) + X(s+1) + \ldots + X(s+N) \right)$$

are often used to smooth out short-term fluctuations, thus highlighting longer-term trends or cycles. Here, e.g., $s = t - N$ or , in the case when N is even, $s = -\frac{N}{2}$. Suppose the $X(t)$ are independent. Then $Y(t)$ is not a Markov process.

Consider now a Markov chain $X = \{X(0), X(1), \ldots\}$ with state space $\mathcal{X} = \{0, 1, 2\}$ and transition matrix

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Suppose $X(0)$ has the uniform distribution. Let

$$Y(t) = \begin{cases} 1, & X(t) = 2, \\ 0, & X(t) \neq 2 \end{cases}$$

Then, it is easy to argue that

$$\mathbb{P}(Y(t+1) = 1 | Y(t) = 0, Y(t-1) = 0) = 1 \neq \mathbb{P}(Y(t+1) = 1 | Y(t) = 0) = \frac{1}{2}.$$

Hence, $Y(t)$ is *not* a Markov chain. For another example of a non Markovian process concerning Pólya's urn, see Problem 77 below. Given an observable non Markovian process $Y = \{Y(0), Y(1), \ldots\}$, one often seeks to represent Y as $Y(t) = f(t, X(t))$, where $X = \{X(0), X(1), \ldots\}$ is a Markov chain. When this is feasible, one talks about a *hidden Markov model*, see Problem 78.

Problems

Problem 65 Show that indeed $\sim$ of Definition 9.1.1 induces an equivalence relation on $\mathcal{X}$.

Problem 66 Prove that the period is in fact a communication class property.
Hint: Suppose $i \sim j$, and let k and l be such that $p_{ij}^{(k)} > 0$ and $p_{ji}^{(l)} > 0$. Prove first that $d(i)$ divides $k + l$. Second, show that if $n > 0$ is not divisible by $d(i)$, then necessarily $p_{jj}^{(n)} = 0$. Conclude from that that $d(i) \leq d(j)$.

Problem 67 Consider Problem 51. Show that if the chain X is irreducible, so is the chain Y.

Problem 68 Consider the following *oriented random walk on the discrete torus*: Let M and N be natural numbers. Consider a time-homogeneous Markov chain with state space $\mathcal{X} = \{0, \ldots, M-1\} \times \{0, \ldots, N-1\}$ and the following transition mechanism. From state (i, j), it moves to state $(i + 1 \,(\mathrm{mod}\, M), j)$ with probability $\frac{1}{2}$ and to state $(i, j + 1 \,(\mathrm{mod}\, N))$ with probability $\frac{1}{2}$.

a. Show that the chain is irreducible;

b. Show that the chain is aperiodic if and only if M and N are coprime.

Problem 69 Consider Example 7.3.5. Show that the chain is irreducible if and only if $\mathbb{P}(\xi(0) = 0) > 0$ *and* $\mathbb{P}(\xi(0) \geq 2) > 0$.

Problem 70 Consider Example 7.3.7. Compute $P^n = \left(p_{ij}^{(n)}\right)$ and the stationary distribution.

Problem 71 Consider again Example 7.2.8. Suppose $N = 6 \times 10^{23}$ which is approximatively the number of molecules in a mole (*Avogadro's number*). Compute the mean recurrence time μ_N of state N when all balls/molecules are in the first container.

Problem 72 A five state Markov chain has transition matrix

$$P = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

1. Draw the graph of the chain and classify the states;
2. Discuss existence and uniqueness of the stationary distribution.

Problem 73 In the absorption analysis, we used the following fact: Consider the linear, constant-coefficient, homogeneous difference equation

$$\sum_{k=0}^{N} a_k x(n-k) = 0, \quad a_0 \neq 0, \quad a_N \neq 0. \tag{9.15}$$

Suppose the *characteristic equation*

$$\sum_{k=0}^{N} a_k z^{-k} = 0$$

has distinct solutions $\lambda_1, \ldots, \lambda_r, r \leq N$, with respective multiplicities $\nu_1, \ldots, \nu_r$, with $\sum_{i=1}^{r} \nu_i = N$. Then a basis for the N dimensional vector space of solutions is provided by the *signals*

$$\left\{ n^l \lambda_i^n; i = 1, \ldots, r; l = 0, \ldots, \nu_i - 1 \right\}. \tag{9.16}$$

Use this result to prove that each periodic, discrete-time signal x of period N ($x(n+N) = x(n), \forall n \in \mathbb{N}$) admits the *Fouries series representation*

$$x(n) = \sum_{k=0}^{N-1} a_k \varphi_k(n), \quad n \in \mathbb{N},$$

where

$$\varphi_k(n) = e^{ik\frac{2\pi}{N}n}, k = 0, 1, \ldots, N-1,$$

are harmonically related complex exponential (they are all periodic with common period N).

Problem 74 Use the basis (9.16) to write the general solution of the Fibonacci equation (3.7) as

$$F(n) = c_1 \lambda_1^n + c_2 \lambda_2^n.$$

Then use the initial conditions to determine c_1 and c_2. Check that this expression gives the correct Fibonacci numbers for $n = 2$ and $n = 3$. Finally use this expression to prove Kepler's observation (3.8).

Problem 75 Find the mean duration time of the game of Example 7.2.7 considered in Section 9.4 when $p \neq q$.

Problem 76 Let $X(t), t \geq 0$ be a sequence of independent Bernoulli trials. Prove that the process Y defined by

$$Y(t) = \frac{1}{2}\left(X(t) + X(t+1)\right)$$

is not Markovian.

Problem 77 Consider Example 7.5.3 (Pólya's urn). Let $Z(t) = 0$ if at time t a white ball is drawn and $Z(t) = 1$ is a black ball is drawn. Show that Z is not a Markov process.

Problem 78 Let $X = \{X(0), X(1), \ldots\}$ be a Markov chain with finite state space $\mathcal{X}$. Let $f : \mathbb{N} \times \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y}$ is finite. Consider the stochastic process $Y(t) = f(t, X(t))$. Is Y always a Markov chain? Is the compound process $Z = (X, Y)$ taking values in $\mathcal{X} \times \mathcal{Y}$ always a Markov chain? Are the answers to the two above questions the same if both $\mathcal{X}$ and $\mathcal{Y}$ are countably infinite?

Chapter 10

Some modern applications of Markov chains

10.1 The Google PageRank algorithm

During the period 1995-1998 Larry Page and Sergey Brin developed at Stanford University an algorithm design to weight the pages of the World Wide Web which has since been known as *Google's PageRank*. It soon rose to prominence among search engines. Although not all details are available, the algorithm essentially works like this. The ranking of the individual pages is given by the stationary distribution of a Markov chain. Let N be the number of current pages (today, hundreds of billions). Consider the $N \times N$ *connectivity matrix* G defined by: $g_{ij} = 1$ if there is a hyperlink from page i to page j and zero otherwise. Notice that G is huge but very sparse. This link structure is recalculated approximatively once a month. Define another $N \times N$ matrix P as

$$p = (p_{ij})_{i,j=1}^{N}, \quad p_{ij} = p\frac{g_i j}{\sum_j g_{ij}} + (1-p)\frac{1}{N}, \quad 0 < p < 1. \tag{10.1}$$

Observe that $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$. Hence P is stochastic. The corresponding chain is clearly irreducible (the second term in the right-hand side of (10.1) makes so that pages that do not have outgoing links are not absorbing states!). Hence, there is a unique stationary distribution π with $\pi_j > 0, \forall j$ (Theorem 9.2.15). Also notice that P is nearly sparse (it is a rank one modification of a sparse matrix).

In the PageRank algorithm, $p = 0.85$. Ideally, the algorithm should mimic the actual navigation of web users.The stationary distribution, which is also calculated as asymptotic distribution of the chain once a month, provides the ranking. Notice that the latter is independent of the content of the pages and of the specific web search query.

A recent experimental research on real user traffic involving approximatively one hundred thousand users [20] showed the following: The ranking based on the actual frequency with which a site is visited significantly differs from that of the PageRank algorithm. To interpret these findings, the authors analyzed the fundamental assumptions underlying PageRank and found that each one of them is violated by actual user behavior.

Other recent applications of PageRank like algorithms include ranking of doctoral programs or individual scientists production, generating customized reading lists and measuring the impact of the *Blogosphere* on the Web itself.

10.2 Identifying genes in genomic DNA

Markov chains and their generalizations play a central role also in recent attempts to find genes through statistical modeling, see e.g. [34] and references therein. For instance, in *prokaryotes* (bacteria and archaea), genes appear as stretches in an enormously long sequence of four symbols $\mathcal{A} := \{A, C, G, T\}$[1] that alternate with non coding regions. It is widely believed that genes of the same organism can be predicted on the base of statistical modeling. For instance, k-step Markov chains have been used in the modeling. These may be viewed as Markov chains with values in $\mathcal{A}^k$. A popular choice is $k = 5$, in which case the state space $\mathcal{X}$ has cardinality $4^5 = 1,024$. Of course, transition probabilities need to be estimated from known genes. Results are encouraging, but much remains to be done.

10.3 The average consensus problem

Animal aggregations, such as schools of fish, flocks of birds, groups of bees, etc. are believed to use local coordination rules that result in complex intelligent behavior at the group level, see e.g. [22, 11]. This has provided

[1]The four letters are the initials of the bases contained in nucleotides: Adenine, Cytosine, Guanine and Thymine.

inspiration for a whole new exciting field of research in decentralized control dealing with problems of multi-agent systems *consensus* and *flocking*, (see [23] for a recent survey) and *synchronization* [32, 18]. Applications include mobile robots coordination, estimation with distributed sensors and load balancing in computer networking. Typically, in all of these applications, very limited amount of information may be exchanged. The latter may depend on the network topology which varies with the agents position. Design and analysis of decentralized control laws for these systems is usually very hard.

Let us consider a toy version of the coordinated consensus problem. We have N agents whose state at time k is represented by a vector $x(k) \in \mathbb{R}^N$. Their dynamics is simply given by

$$x(k+1) = x(k) + u(k),$$

where $u(k) \in \mathbb{R}^N$ represents the control inputs. The goal is to design a feedback control law $u = Kx$ such that, for any initial condition $x_0 \in \mathbb{R}^N$, the closed loop system $x(k+1) = (I+K)x(k)$ satisfies

$$\lim_{k\to\infty} x(k) = \alpha(x_0)\mathbb{1}, \tag{10.2}$$

where $\mathbb{1}^* = (1, 1, \ldots, 1)$. Notice that $k_{ij} \neq 0$ implies that agent j has to communicate x_j to agent i in order for i to compute its feedback control. Let $\mathcal{G}_K$ be the *communication graph* associatd to K, namely the graph with vertices $\{1, 2, \ldots, N\}$ and arc from j to i if and only if $k_{ij} \neq 0$.

Let $P = I + K$. Clearly, to ensure that $\alpha(x_0)\mathbb{1}$ is an equilibrium point of the closed loop system, we need to impose $K\mathbb{1} = 0$ or, equivalently, $P\mathbb{1} = \mathbb{1}$. If we add the requirement that the elements $p_{ij} \geq 0$, we see that P must be a *stochastic matrix*. Consider the directed graph $\mathcal{G}_K$ with vertices $\{1, 2, \ldots, N\}$ where there is an arc from j to i whenever the element k_{ij} of K is not zero. Given a directed graph $\mathcal{G}$ with vertices $\{1, 2, \ldots, N\}$ that represents the communication network, we say that K is *compatible* with $\mathcal{G}$ if $\mathcal{G}_K \subseteq \mathcal{G}$. The consensus problem is said to be solvable on $\mathcal{G}$ if there exists a feedback matrix K compatible with $\mathcal{G}$ and achieving (10.2). It is always assumed that $\mathcal{G}$ contains all loops (i, i), namely each agent has access to its own state.

The Perron-Frobenius Theorem 9.2.15 ensures convergence under the condition that $\mathcal{G}_K$ be strongly connected (irreducible chain). This happens if and only if there exists an integer $m \geq 1$ such that P^m has only strictly positive elements since the chain is aperiodic. In this case, 1 is an eigenvalue of multiplicity one of P with corresponding eigenvector $\mathbb{1}$. All other eigenvalues λ

of P are in the open unit disc. The *essential spectral radius* of P is defined by

$$\mu(P) = \max\{|\lambda|; \lambda \in \sigma(P) \setminus \{1\}\},$$

it is namely the absolute value of the eigenvalue different from 1 closest to the unit circle[2] It regulates how fast P^n tends to the rank one matrix with all rows equal to the stationary distribution π (it may be seen that the distribution $\pi(t)$ converges to π as μ^t). As the speed of convergence is an essential issue in many applications, it is important to choose K so that the essential spectral radius is as small as possible. In [2], the following problem was solved. Consider a strongly connected graph $\mathcal{G}$ with vertices $\{1, 2, \ldots, N\}$ such that $(j, j) \in \mathcal{G}, \forall j$ and $(i, j) \in \mathcal{G} \Leftrightarrow (j, i) \in \mathcal{G}$. Find a symmetric, stochastic matrix $P = (p_{ij})$ such that $(i, j) \notin \mathcal{G} \Rightarrow p_{ij} = 0$ that has minimum essential spectral radius.

In the average consensus problem, it is often required that consensus is attained at the average of the initial states $\bar{x}(0) = \frac{1}{N}\sum_i x_i(0)$. It can be shown that the baricenter of states is preserved by the dynamics for all initial $x(0)$ if and only if P is doubly stochastic (its columns also sum to one). This is the case if $K\mathbf{1} = 0$ and K is symmetric. The average consensus problem becomes more tractable when the graph $\mathcal{G}$ and the feedback matrix K possess symmetries.

In some applications, such as the random gossip algorithm [3] or when we want to model the possible failure in inter agent communication, it may be sensible to allow P to be random. More explicitly, $k \mapsto P(k)$ is assumed to be a Markov chain taking values in a finite set of stochastic matrices $\{P_1, \ldots, P_m\}$.

10.4 Markov chain Monte Carlo

10.4.1 Monte Carlo methods

Suppose we like to compute the area of a planar figure contained in the square $[0, 1] \times [0, 1]$. We know that points in $[0, 1]$ are in one to one correspondence with infinite binary sequences. Hence, if we toss two coins a large number of times (say $10,000$) we get a miniscule square in $[0, 1] \times [0, 1]$ that we can consider as a point which may or may not be contained in our figure. If we

[2] $-\log \mu$ is called the *mixing rate* of the chain. The quantity $1 - \mu(P)$ is called the *spectral gap*.

can repeat the experiment many times, we can approximate the area by the fraction of points that fall in the figure. This is the essence of integration by simulation which becomes extremely effective in higher dimensions (namely, when the independent variable lies in $\mathbb{R}^p$, $p \geq 2$). Simulation has come a long way since French naturalist Buffon computed the value of π by dropping a needle on a card table where equispaced parallel lines had been drawn. An excellent introductory source is [13], which has many examples and indication of dedicated computer programs.

Monte Carlo methods are computational algorithms that rely on repeated random sampling. Monte Carlo simulation treats deterministic problems by first finding a probabilistic analog. Previous methods of simulation and statistical sampling generally did the opposite, namely used simulation to test a previously understood deterministic problem. Perhaps the most famous early use was by Enrico Fermi in 1930, when he employed a random method to calculate the properties of the newly-discovered neutron. This method was then extensively used in the forties by scientists working on the Manhattan Project at the Los Alamos National Laboratory[3].

Monte Carlo algorithms are the most prominent class of *Randomized algorithms*. These are algorithms that make random choices during their execution. They may give different answers for the same input at different runs and the results may be incorrect. Given the enormously increased computing power, their use and range of applications has considerably grown in recent times, see e.g. [21]. For a recent survey of applications to Systems and Control Theory see [33].

10.4.2 The Metropolis algorithm

The *Markov chain Monte Carlo* (MCMC) method seeks to construct a finite, irreducible, aperiodic Markov chain whose unique stationary distribution is the distribution π we like to sample from. A natural objection to this program is: How can it be easier to construct such a Markov chain rather than construct directly a random variable with distribution π? In many applications of the statistical mechanics models, the probability distribution is only known up to a normalizing factor, the partition function (see Chapter

[3]The name "Monte Carlo" was popularized by Ulam, Fermi, von Neumann, and Metropolis. It apparently originated from the famous casino in Monaco where Ulam's uncle would borrow money to gamble.

2), which is difficult to compute. In other applications, one seeks to compute expectations of a functional of a large dimensional random vector. In both situations, it is convenient to employ MCMC. Nowadays applications of MCMC include simulating noisy images, textures, protein structures, approximate counting for polymer models, Bayesian statistics[4] and scientific computing. A typical MCMC algorithm will actually find a Markov chain that is *reversible* with the distribution π. This because the simplest way to guarantee that π is stationary for a certain transition matrix is to impose the detailed balance condition (7.23), cf. (Theorem 7.4.3).

The *Metropolis algorithm* is the foundation stone of the Markov chain Monte Carlo. We illustrate this algorithm on the travelling salesman problem of Section 2.3 (for a beautiful introduction to Markov chains and randomized algorithms see [15]). Recall that we would like to sample from the (unknown) Boltzmann distribution

$$\bar{\pi}_i(\beta) = Z(\beta)^{-1} \exp[-\beta E_i], \quad Z(\beta) = \sum_i \exp[-\beta E_i], \quad \beta = \frac{1}{kT}, \qquad (10.3)$$

where E_i is the length of path $i, i = 1, \ldots, \frac{1}{2}n!$, where n is the number of cities to be visited. The structure of the algorithm is the following: We try to run a Markov chain $X(t+1) = f(X(t), \xi(t))$, where the $\xi(t)$ are i.i.d., hoping that it will have the desired distribution $\bar{\pi}$ as stationary. Concretely, we proceed as follows. Start from any admissible path/permutation:

1. Let $i \in \{1, 2, \ldots, \frac{1}{2}n!\}$ be the present path;

2. Chose two cities at random i_r and i_s along the path i and *exchange* their order. This yields a new path j (recall that any two cities are connected). More explicitly, if

$$i = (i_1, \ldots, i_{r-1} i_r, i_{r+1}, \ldots, i_{s-1}, i_s, i_{s+1}, \ldots, i_n)$$

then

$$j = (i_1, \ldots, i_{r-1}, i_s, i_{s-1}, \ldots, i_{r+1}, i_r, i_{s+1} \ldots, i_n).$$

[4]It has been observed that the reason the Bayesian approach has become more popular in statistics is MCMC!

3. Accept path j with probability

$$\alpha = \min\left(\frac{\bar{\pi}_j}{\bar{\pi}_i}, 1\right) = \min\left(\exp\left(\frac{E_i - E_j}{T}\right), 1\right)^{5}.$$

Thus, if $E_i \geq E_j$, namely path i is longer than path j, always accept path j. If, instead, $E_i < E_j$, *accept path j with probability* $\exp\frac{E_i - E_j}{T}$. This strategy has the potential to overcome paths that represent local minima!

If Q is a symmetric transition matrix of an irreducible chain, consider the transition probabilities

$$p_{ij} = \begin{cases} q_{ij} \min\left(\exp\left(\frac{E_i - E_j}{T}\right), 1\right), & i \neq j, \\ 1 - \sum_{k,k\neq i} q_{ik} \min\left(\exp\left(\frac{E_i - E_j}{T}\right), 1\right), & i = j. \end{cases}$$

We have that this Metropolis chain is reversible with respect to the Boltzmann distribution $\bar{\pi}$ (10.3).Indeed, for $i \neq j$,

$$\bar{\pi}_i p_{ij} = q_{ij}\bar{\pi}_i \min\left(\exp\left(\frac{E_i - E_j}{T}\right), 1\right) = q_{ij} \min\left(\bar{\pi}_i, \bar{\pi}_j\right) = \bar{\pi}_j p_{ji},$$

since $q_{ij} = q_{ji}$.

10.4.3 Simulated annealing

In metallurgy, annealing is a heat treatment used to change strength and hardness of materials. The metal is first heated (usually to glowing) for a while and then *slowly cooled* so that it settles in a minimum energy state. The latter is the hardest, corresponding to a microstructure with larger crystals. This can be used to improve the Metropolis algorithm described above. Indeed, in the latter, if T is large, we get a fast convergence, but not necessarily to a global minimum. When T is small, on the other hand, we eventually do approach a global minimum, but convergence may be incredibly slow as there is the possibility of getting trapped for a long time in local minima. Hence, we must run the Metropolis algorithm first with a large temperature T_1 up to time N_1. Then use a lower temperature T_2 up to time N_2, etc. If the sequences $T_i \searrow 0$ and $N_i \nearrow \infty$ are carefully chosen, namely temperatures are slowly lowered compared to the N_i, convergence to a global minimum occurs in a reasonable time, see [15, Chapter 13] for details.

[5] To do this, generate a random variable U with uniform distribution on $[0, 1]$. If $U \leq \alpha$, accept the new path. If $U > \alpha$, stay with old path.

10.5 Distribution of epithelial cells

A *simple epithelium* is a tissue composed of a single layer of cells. Since the early microscopic studies of animal tissues (Schwann, 1847), it had been observed that cells are organized in monolayer epithelia as an array of (irregular) polygonal forms, with a majority of hexagonal shapes. Historically, the "cobblestone paving" appearance of simple epithelia had been believed to reflect optimal cell packing. Indeed, many biological systems form certain geometric configurations that minimize surface energy or maximize space filling. Examples are provided by honeycombs, insect retinal cells and compressed soap bubbles. The pattern in proliferating epithelia is, however, much more irregular. The fact that the *average number* of cell sides exponentially approaches six had been known for some time to be a consequence of the following fact: Each cell division produces two new vertices and three new edges. Nevertheless, this observation does not imply a prevalence (or even existence) of exagons, nor does it provide any further information on the distribution of the sides number.

This long standing mystery was recently solved in [12]. Contrary to the optimal packing hypothesis, it was shown there that mitosis (cell division) suffices to explain the distribution of polygonal shapes found in monolayer epithelia. The model, provided by a simple Markov chain, where a cell's state is its number of sides, originated as follows. Experimental observation of the growth of the wing primordium of the fruitfly *Drosophila melanogaster* during four days of larval development showed no large-scale sorting or migration within this epithelium. The only significant cellular movements occurred during mitosis when polygonal cells rounded up and divided into two daughter polygons. Further experimental work led to the hypothesis that cell division should be sufficient to account for the cell shape distribution in simple proliferating epithelia. Let $S(t)$ denote the number of sides/vertices of the cell at time t. Since triangular cells are never observed, it was assumed that each of the two daughter cells receives at least two vertices (tricellular junctions) from the parent cell. The remaining $S(t) - 4$ parental junctions are distributed between the two daughters according to fair coin tossing. Hence, the first daughter receives all but two junctions according to the binomial distribution $B(S(t) - 4, \frac{1}{2})$. This originates a first stochastic matrix P whose non zero entries are normalized numbers of the Tartaglia's triangle in Figure 3.1. There is, however, another effect to take into account. On mitosis, a cell adds one side to each of two neighboring cells. If there are N cells in

the epithelium, after one mitosis cycle producing $2N$ cells, $2N$ new sides are added. Hence, on the average, each cell gains one side (this may be thought of as a *mean field approximation*). Thus we have another stochastic matrix S which is just a shift matrix, namely $s_{ij} = 1$ if $j = i+1$ and zero otherwise. The two matrices are then given by

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & \dots \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \dots \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} & 0 & \dots \\ \frac{1}{16} & \frac{4}{16} & \frac{6}{16} & \frac{4}{16} & \frac{1}{16} & \dots \\ . & . & . & . & . & \dots \end{bmatrix}, \quad S = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 1 & \dots \\ . & . & . & . & . & . & \dots \end{bmatrix},$$

where $i, j = 4, 5, 6, \dots$. The transition matrix of the chain is then $Q = PS$ given by

$$Q = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \dots \\ 0 & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} & 0 & \dots \\ 0 & \frac{1}{16} & \frac{4}{16} & \frac{6}{16} & \frac{4}{16} & \frac{1}{16} & \dots \\ . & . & . & . & . & . & \dots \end{bmatrix}. \tag{10.4}$$

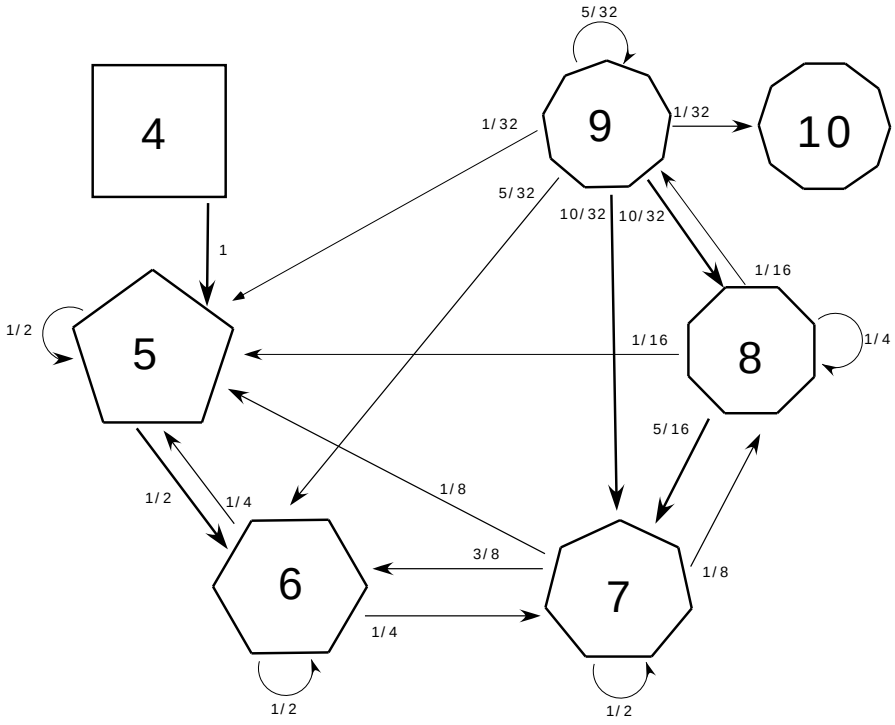

Figure 10.1: Portion of chain graph for cell shapes in monolayer epithelia. Transition from states with equilibrium probability less than 10^{-4} are not depicted.

The most relevant part of the chain graph is depicted in Figure 10.1. The state 4 is transient, all the other states are recurrent. The stationary distribution of cell shape π satisfying $\pi = Q^T\pi$, with Q as in (10.4) has $\pi_4 = 0, \pi_j > 0, \forall j > 4$. More explicitly, we have $\pi_5 \sim 0.289$, $\pi_6 \sim 0.464$, $\pi_7 \sim 0.208$, $\pi_8 \sim 0.0359$, $\pi_9 \sim 0.0028$, other states having equilibrium probability less than 10^{-4}. In the developing *Drosophila* wing, the actual polygon distribution closely matched the predicted distribution after only eight generations. Only the small percentage of four sided cells was not predicted from the model. This effect is believed to be due to the mean field approximation. Empirical evidence from frogs, fruit flies, and hydra suggests that the stationary distribution is exhibited by almost all multicellular animals, see Figure 10.2.

comparison.jpg

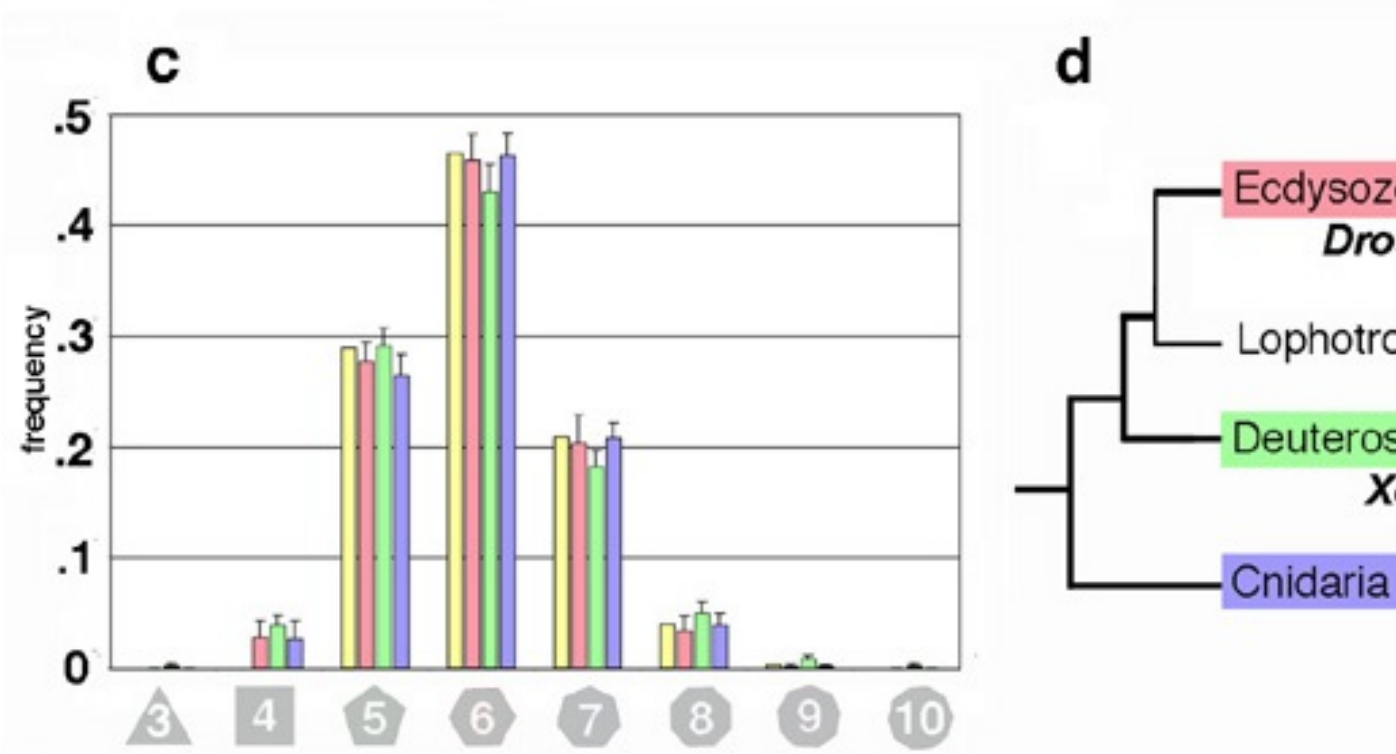

Figure 10.2: Comparison of theoretical-experimental distribution of polygonal shapes in monolayer epithelia.

Problems

Problem 79 In the average consensus problem outlined in Section 10.3, show that the baricenter of states $\frac{1}{N}\sum_i x_i(0)$ is preserved for all $x(0)$ by the evolution if and only if P is doubly stochastic.

Bibliography

[1] R. Azencott, *Grandes déviations et applications*, Ecole d'Eté de Probabilitès de Saint Flour VIII, Lecture Notes in Mathematics 774, Springer-Verlag, 1980.

[2] S. Boyd, P. Diaconis, and L. Xiao, Fastest mixing Markov chain on a graph, *SIAM Review* 46, 667-689, 2004.

[3] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, Randomized gossip algorithms, *IEEE Trans. Inf. Theory* 52, 2508-2530, 2006.

[4] L. Breiman, *Probability*, Addison-Wesley, Reading, Mass., 1968.

[5] P. Brémaud, *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer-Verlag, New York, 1999.

[6] K. L. Chung, *Elementary Probability Theory*, Springer, New York, Fourth Edition, 2003.

[7] T. Cover and J. Thomas, *Elements of information theory*, Second Edition, Wiley, 2006.

[8] P. Del Moral, *Feynman-Kac Formulae*, Springer-Verlag, 2004.

[9] J. L. Doob, A Markov chain theorem, *Probability & Statistics* (The H. Cramér Volume), Wiley, 1959, pp. 50-57.

[10] W. Feller, *An Introduction to Probability Theory and its Applications, Vol.I.* Second Edition Wiley, 1957.

[11] G. Flierl, D. Grunbaum, S. Levin, and D. Olson, From individuals to aggregations: the interplay between behavior and physics, *J. Theoretical Biology* 196, 397-454, 1999.

[12] M. Gibson, A. Patel, R. Nagpal and N. Perrimon, The emergence of geometric order in proliferating metazoan epithelia, *Nature* 442 , 1038-1041, 2006.

[13] C. Grinstead and J. Snell, *Introduction to Probability*, 2nd edition, Americal Mathematical Society, 2006, freely available at http://math.dartmouth.edu/ prob/prob/prob.pdf

[14] F. Guerra, Notes for a Statistical Mechanics course, University of Rome "La Sapienza", 1987 (in Italian).

[15] O. Häggström, *Finite Markov Chains and Algorithmic Applications*, London Mathematical Society Student Texts 52, Cambridge University Press, 2002.

[16] E.T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, Sept. 1982.

[17] S. Karlin, *A first course in stochastic processes*, Academic Press, 1966.

[18] M. Marodi, F. d'Ovidio, and T. Vicsek, Synchronization of oscillators with long range interaction: Phase transition and anomalous finite size effects, *Physical Review E* 66, 011109, 2002.

[19] R. Martin, The St. Petersburg Paradox, *The Stanford Encyclopedia of Philosophy* (Fall 2004 Edition), Ed. Edward N. Zalta. Stanford, California: Stanford University.

[20] M. Meiss, F. Menczer, S. Fortunato, A. Flammini, A. Vespignani: Ranking Web Sites with Real User Traffic. Proc. WSDM 2008, also at http://www.informatics.indiana.edu/fil/Papers/click.pdf

[21] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*, Cambridge University Press, 2005.

[22] A. Okubo, Dynamical aspects of animal grouping: swarms, schools, flocks, and herds, *Advances in Biophysics* 22, 1-94, 1986.

[23] R. Olfati-Saber, J. A. Fax, and R. M. Murray, Consensus and Cooperation in Networked Multi-Agent Systems, *Proceedings of the IEEE*, 95, 215-233, 2007.

[24] M.Pavon, Stochastic control and nonequilibrium thermodynamical systems, *Appl. Math. and Optimiz.* **19** (1989), 187-202.

[25] M. Pavon and F. Ticozzi, Discrete-time classical and quantum Markovian evolutions: Maximum entropy problems on path space, *J. Math. Phys.*, **51**, 042104-042125 (2010) doi:10.1063/1.3372725. Preprint arXiv:math-ph/0811.0933v2.

[26] M. Pavon and A. Ferrante, On the geometry of maximum entropy problems, *SIAM Review*, **55-3**, 2013, 415-439, preprint arXiv: math/1112.5529v3 [math.OC].

[27] C. M. Pelaggi, Principi variazionali stocastici per funzionali della densità e per la funzione importanza, "Laurea" (Master) Thesis, University of Rome "La Sapienza", 1988.

[28] T. Rockafellar, *Convex analysis*. Princeton University Press, 1970.

[29] E. Schrödinger, Über die Umkehrung der Naturgesetze, *Sitzungsberichte der Preuss Akad. Wissen. Berlin, Phys. Math. Klasse* (1931), 144-153.

[30] E. Schrödinger, Sur la théorie relativiste de l'électron et l'interpretation de la mécanique quantique, *Ann. Inst. H. Poincaré* **2**, 269 (1932).

[31] A. N. Shiryayev, *Probability*, Springer-Verlag, New York, 1984.

[32] S. H. Strogatz, From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators, *Physics D: Nonlinear Phenomena* 143, 1-20, 2000.

[33] R. Tempo and H. Ishii, Monte Carlo and Las Vegas randomized algorithms for systems and control, *European J. of Control* 13 (2007), 189-203.

[34] M. Vidyasagar, S. S. Mande, Ch. V. Siva Kumar Reddy and V. V. Raja Rao, The 4M (Mixed Memory Markov Models) algorithm for finding genes in prokaryotic genomes, *IEEE Trans. Aut. Control* 53 (2008), 26-37.

Appendix A

Answers to problems

A.1 Chapter 1

Problem 1

a. The maximum point of the function $f(x) = x - x^2$ occurs at $x = 1/2$. Hence

$$\mathbb{P}(A) \cdot \mathbb{P}(A^c) = \mathbb{P}(A) - \mathbb{P}(A)^2 \leq \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4};$$

b. using the previous result, and the monotonicity of probability, we get

$$\begin{aligned}\mathbb{P}(B \cap C) = \mathbb{P}(B \cap C)\left[\mathbb{P}(B) + \mathbb{P}(B^c)\right] &\leq \mathbb{P}(B \cap C)\left[\mathbb{P}(B) + \mathbb{P}((B \cap C)^c)\right] \\ &\leq \mathbb{P}(B \cap C)\mathbb{P}(B) + \frac{1}{4} \leq \mathbb{P}(B)\mathbb{P}(C) + \frac{1}{4}.\end{aligned}$$

Problem 3

$$\begin{aligned}&\mathbb{E}X = \sum_{k=1}^{M} k\mathbb{P}(X = k) = \\ &1 \cdot \mathbb{P}(X = 1) + 2 \cdot \mathbb{P}(X = 2) + 3 \cdot \mathbb{P}(X = 3) + \cdots + M \cdot \mathbb{P}(X = M) \\ &= (\mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \cdots + \mathbb{P}(X = M)) \\ &+ (\mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \cdots + \mathbb{P}(X = M)) + \cdots + \mathbb{P}(X = M) \\ &= \sum_{k=1}^{M} \mathbb{P}(X \geq k).\end{aligned}$$

Problem 5 We show directly that $1 \Leftrightarrow 2$. Suppose f is convex. Let (x, α) and (y, β) be two points in *epi* f. Then, for $\lambda \in [0, 1]$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \leq \lambda\alpha + (1-\lambda)\beta.$$

It follows that

$$\lambda(x, \alpha) + (1-\lambda)(y, \beta) = (\lambda x + (1-\lambda)y, \lambda\alpha + (1-\lambda)\beta) \in epi f.$$

Viceversa, if $epi f$ is convex, then for $x, y \in K$

$$\lambda(x, f(x)) + (1-\lambda)(y, f(y)) = (\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)) \in epi f.$$

This implies that $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$.

A.2 Chapter 2

Problem 7 We follow the same argument as in Section 2.1. The *Lagrangian function* is given by

$$\mathcal{L}(p, \lambda, \mu) := H(p) + \lambda(\alpha - \sum_{i=1}^{n} i p_i) + \mu(\sum_{i=1}^{n} p_i - 1).$$

Equivalently, we can maximize over $\mathbb{R}^N_+$ the functional

$$I(p) = -\sum_i p_i \ln p_i - \lambda \sum_{i=1}^{n} i p_i + \mu \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} [-\ln p_i - \lambda i + \mu] p_i.$$

Arguing as above, we get the optimality condition

$$-\ln p_i - 1 - \lambda i + \mu = 0.$$

The latter yields

$$p_i^* = \exp[-1 + \mu - \lambda i] > 0.$$

Next, we choose μ so that $\sum_i p_i^* = 1$, namely

$$\exp[1 - \mu] = \sum_{i=1}^{n} \exp[-\lambda i]. \tag{A.1}$$

Let $q := \exp[-\lambda]$. Then the optimal distribution has the form

$$p_i^* = cq^i.$$

If $q < 1$,

$$p_i^* = \frac{1-q}{1-q^n} q^{i-1},$$

since

$$(1-q)\sum_{i=1}^{n} q^{i-1} = 1 - q^n.$$

Arguing as in Theorem 2.2.1, one can see that for each $0 < \alpha < \frac{n+1}{2}$ there is a unique λ such that the corresponding distribution satisfies the expectation constraint. For $\alpha = \frac{n+1}{2}$, we get as solution the uniform distribution ($\lambda = 0$)

$$p_i^* = \frac{1}{n}.$$

A.3 Chapter 3

Problem 9 I can place the first rook on 64 different squares and then the second on 14, for each choice of the first. This exausts all the posssibilities since a rook can capture another if and only if it can be captured. Hence, the answer is $64 \times 14 = 896$.

Problem 10 There are 26 letters in the English alphabet. The total number of TLA is $26^3 = 17,576$ (dispositions without repetitions of 3 balls in 26 cells).

Problem 11 Let us consider $\Omega = \{\omega = (a_1, a_2, a_3, a_4), a_i \in \{1, 2, 3, 4, 5, 6\}\}$. We have $|\Omega| = 6^4 = 1296$. There are $5^4 = 625$ sequences ω in which six does not appear. The probability of at least one six is therefore $1 - \frac{625}{1296} = \frac{671}{1296} \simeq 0.5177$. Hence, it is advantageous to bet. Chevalier De Méré knew that (probably form experience). He deduced that it would favorable to bet on at least one double six in 24 rolls. We have, however, $1 - \left(\frac{35}{36}\right)^{24} \simeq 0.4914$. When he started losing, be blamed it on the silly mathematics until Pascal in 1654 explained to him how to compute chances.

Problem 13

a. There are exactly $\left(\begin{array}{c} 3 \\ 2 \end{array} \right) = 3$ ways of getting 2 heads in the first three tosses. Hence, there are $3 \times 2^7 = 384$ sequences of length 10 that have two heads in the first three tosses;

b. There are $\left(\begin{array}{c} 10 \\ 2 \end{array} \right) = 45$ such sequences.

Problem 14

a. Differentiating

$$(1+x)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) x^k$$

we get

$$n(1+x)^{n-1} = \sum_{k=1}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) x^{k-1} = \sum_{k=0}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) x^{k-1}.$$

Evaluating such expression at $x = 1$, we get

$$n2^{n-1} = \sum_{k=0}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right).$$

b. Multiplying the latter inequality on both sides by 2^{-n}, we get

$$\frac{n}{2} = \sum_{k=0}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) 2^{-n} = \mathbb{E}\{X\}.$$

Problem 15 Recall first that

$$\left(\begin{array}{c} n \\ n-k \end{array} \right) = \left(\begin{array}{c} n \\ k \end{array} \right).$$

Hence (3.13) is indeed a particular case of (3.12). To prove (3.12), one may use induction or, as suggested, the binomial expansion in

$$(1+x)^n (1+x)^m = (1+x)^{n+m}.$$

We give instead a different proof which is more combinatorial in spirit. Consider $n+m$ coin tosses. We know that there are

$$\left(\begin{array}{c} n+m \\ r \end{array} \right)$$

different ways to get exactly r heads. Let $0 \leq k \leq r$. If we get k heads in the first n tosses, and $r-k$ in the last m tosses, we would have a total of exactly r in the $n+m$ tosses. The first result can occur in $\left(\begin{array}{c} n \\ k \end{array} \right)$ different ways, the second in $\left(\begin{array}{c} m \\ r-k \end{array} \right)$ different ways. Hence, there are

$$\left(\begin{array}{c} n \\ k \end{array} \right) \times \left(\begin{array}{c} m \\ r-k \end{array} \right)$$

different ways to obtain k heads in the first n tosses and $r-k$ in the last m. As k varies between 0 and r, we get all the different ways to obtain r heads in $n+m$ tosses. Namely

$$\sum_{k=0}^{r} \left(\begin{array}{c} n \\ k \end{array} \right) \left(\begin{array}{c} m \\ r-k \end{array} \right) = \left(\begin{array}{c} n+m \\ r \end{array} \right).$$

Problem 16 As in Example 3.1.11, consider the sample space of permutations of n objects

$$\Omega = \{(a_1, \ldots, a_n),\ 1 \leq a_i \leq n,\ i \neq j \Rightarrow a_i \neq a_j\}.$$

Let $N(\omega)$ denote the number of matches in $\omega \in \Omega$. Then, we have

$$N(\omega) = \sum_{i=1}^{n} \mathbb{1}_{A_i}(\omega),$$

where, as in Example 3.1.11, A_i is the event "the i^{th} person recovers his coat". Hence

$$\mathbb{E}N = \sum_{i=1}^{n} \mathbb{E}\mathbb{1}_{A_i} = \sum_{i=1}^{n} 1\mathbb{P}(A_i) = n \times \frac{1}{n} = 1.$$

Hence, the expected number of matches is exactly one, independently of $n \geq 1$! In the words of K. L. Chung [6, p.173]: "this is neat, but it must be considered as a numerical accident".

Problem 17 A possible way to answer this question is the following. Introduce the events $A_1 := \{(\ldots, \text{empty}, \text{my space}, \ldots)\}$ and $A_2 := \{(\ldots, \text{my space}, \text{empty}, \ldots)\}$. We need to compute $\mathbb{P}(A_1 \cup A_2)$, where $\mathbb{P}$ is the uniform probability measure. We get

$$\mathbb{P}(A_1) = \mathbb{P}(A_2) = \frac{\left(\begin{array}{c} n-2 \\ 2 \end{array} \right)}{\left(\begin{array}{c} n-1 \\ 2 \end{array} \right)} = \frac{(n-2)!(n-3)!}{(n-4)!(n-1)!} = \frac{n-3}{n-1}.$$

Moreover,

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}\{(\ldots, \text{empty}, \text{my space}, \text{empty}, \ldots)\} = \frac{\left(\begin{array}{c} n-3 \\ 2 \end{array} \right)}{\left(\begin{array}{c} n-1 \\ 2 \end{array} \right)}$$

$$= \frac{(n-3)!(n-3)!}{(n-5)!(n-1)!} = \frac{(n-3)(n-4)}{(n-1)(n-2)}.$$

We then get

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) = \frac{n-3}{n-1} + \frac{n-3}{n-1} - \frac{(n-3)(n-4)}{(n-1)(n-2)}$$

$$= \frac{n-3}{n-1}\left(2 - \frac{n-4}{n-2}\right) = \frac{(n-3)n}{(n-1)(n-2)}.$$

For $n = 4$, we get $\mathbb{P}(A_1 \cup A_2) = \frac{2}{3}$. For $n = 5$, $\mathbb{P}(A_1 \cup A_2) = \frac{5}{6}$. For $n = 6$, $\mathbb{P}(A_1 \cup A_2) = \frac{9}{10}$.

There is, however, a much faster way to solve this problem. Let $B := (A_1 \cup A_2)^c$. Then $B = \{(\ldots, \text{car}, \text{my space}, \text{car}, \ldots)\}$ represents the event "I cannot park". We have

$$\mathbb{P}(B) = \left(\begin{array}{c} n-1 \\ 2 \end{array} \right)^{-1} = \frac{2!(n-3)!}{(n-1)!}.$$

Hence,

$$\mathbb{P}(A_1 \cup A_2) = 1 - \mathbb{P}(B) = 1 - \frac{2!(n-3)!}{(n-1)!} = 1 - \frac{2}{(n-1)(n-2)} = \frac{(n-3)n}{(n-1)(n-2)}.$$

Problem 18 There are $\begin{pmatrix} 4N \\ 2N \end{pmatrix}$ different ways to choose a group of $2N$ persons from $4N$ persons. Of these, those with exactly N boys and N girls are $\begin{pmatrix} 2N \\ N \end{pmatrix}^2$. Infact, there are $\begin{pmatrix} 2N \\ N \end{pmatrix}$ ways to choose N boys out of the $2N$ e as many for the girls. We conclude that the sought probability is

$$\frac{\begin{pmatrix} 2N \\ N \end{pmatrix}^2}{\begin{pmatrix} 4N \\ 2N \end{pmatrix}} = \frac{[(2N)!]^4}{(N!)^4(4N)!} \simeq \sqrt{\frac{2}{N\pi}},$$

where the last estimate is obtained via the de Moivre-Stirling formula (3.2.1).

A.4 Chapter 4

Problem 19 If $i \neq j$, we get

$$\mathbb{E}\{\tilde{X}_i^n \cdot \tilde{X}_j^n\} = \mathbb{E}\{X_i^n \cdot X_j^n\} - \frac{1}{2}\mathbb{E}X_i^n - \frac{1}{2}\mathbb{E}X_j^n + \frac{1}{4} = \frac{1}{4} - \frac{1}{4} - \frac{1}{4} + \frac{1}{4} = 0.$$

For $i = j$, it suffices to observe that $\left(\tilde{X}_i^n\right)^2 \equiv \frac{1}{4}$.

Problem 21

$$\begin{aligned}\mathbb{E}X &= \sum_{k=0}^{\infty} kpq^k = \sum_{k=1}^{\infty} kpq^k = q\sum_{k=1}^{\infty} [(k-1)+1]\, pq^{k-1} \\ &= q\sum_{j=0}^{\infty} jpq^j + qp\sum_{j=0}^{\infty} q^j = q\mathbb{E}X + \frac{pq}{1-q} = q\mathbb{E}X + q.\end{aligned}$$

It follows that $(1-q)\mathbb{E}X = q$ which yields

$$\mathbb{E}X = \frac{q}{p}.$$

A.5 Chapter 5

Problem 24 Let X be a random variable with binomial distribution $B(n, \frac{1}{2})$. Using

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

we get

$$\sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right) = 2^n \left(\sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right) 2^{-n} \right) = 2^n \left(VX + (EX)^2 \right).$$

Since $\mathbb{E}X = \frac{n}{2}$ and $\mathbb{V}X = \frac{n}{4}$, we finally get

$$\sum_{k=0}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array} \right) = 2^n \left(\frac{n}{4} + \frac{n^2}{4} \right) = 2^{n-2} n(1+n).$$

Problem 27 This problem is equivalent to that of Exercise 5.1.2. We get a particular case of the binomial distribution with $p = 1/n$:

$$p_l = \left(\begin{array}{c} m \\ l \end{array} \right) \frac{1}{n^l} \left(1 - \frac{1}{n} \right)^{m-l}.$$

Problem 28 We can choose the four girls in $\left(\begin{array}{c} 8 \\ 4 \end{array} \right)$ different ways. Similarly, the boys four boys can be chosen in $\left(\begin{array}{c} 12 \\ 4 \end{array} \right)$ different ways. The total number of ways we can choose eight people from twenty is $\left(\begin{array}{c} 20 \\ 8 \end{array} \right)$. Hence, the sought probability p is one of the hypergeometric distribution

$$\frac{\left(\begin{array}{c} 8 \\ 4 \end{array} \right) \left(\begin{array}{c} 12 \\ 4 \end{array} \right)}{\left(\begin{array}{c} 20 \\ 8 \end{array} \right)} = \frac{(12!)^2 8!}{(4!)^3 20!}.$$

Problem 29 $\mathbb{E}\xi = 0$ is apparent. It follows that $\mathbb{V}(\xi) = \mathbb{E}(\xi)^2 - (\mathbb{E}\xi)^2 = \mathbb{E}(\xi)^2$ is given by

$$\mathbb{V}(\xi) = \mathbb{E}(\xi)^2 = \frac{\mathbb{E}\left\{ \left((X_1 + X_2) - \mathbb{E}\{X_1 + X_2\} \right)^2 \right\}}{\mathbb{V}(X_1 + X_2)} = \frac{\mathbb{V}(X_1 + X_2)}{\mathbb{V}(X_1 + X_2)} = 1.$$

Using base 2 logarithms, we get the entropies:

$$H(p_{X_1}) = H(p_{X_2}) = -\left[\frac{1}{2}\cdot\log\frac{1}{2} + \frac{1}{2}\cdot\log\frac{1}{2}\right] = 1.$$

To compute $H(p+\xi)$, notice that $\xi(\omega) = \sqrt{2}\left[X_1(\omega) + X_2(\omega) - 1\right]$, takes the values $\sqrt{2}$, 0 e $-\sqrt{2}$ with probability 1/4, 1/2 e 1/4, respectively. We then get

$$H(p_\xi) = -\left[\frac{1}{4}\cdot\log\frac{1}{4} + \frac{1}{2}\cdot\log\frac{1}{2} + \frac{1}{4}\cdot\log\frac{1}{4}\right] = \frac{3}{2}.$$

We conclude that

$$H(p_{X_1}) = H(p_{X_2}) = 1 < \frac{3}{2} < 1 + \frac{1}{2}\cdot\log\pi = H(p_G),$$

where p_G is the Gaussian density with first moment equal to zero and second moment equal to one appearing in the De Moivre - Laplace Theorem. Finally, observe that the entropy of the normalized random variables

$$\tilde{X}_i := \frac{X_i - \mathbb{E}X_i}{\sqrt{\mathbb{V}(X_i)}}, \quad i = 1, 2$$

is the same as for the non normalized.

A.6 Chapter 6

Problem 30 Let A be the event "the first two balls are in different cells" and B the event "one cell contains exactly three balls". Then

$$\mathbb{P}(A) = \frac{4\times 3\times 4^2}{4^4} = \frac{3}{4}, \quad \mathbb{P}(A\cap B) = \frac{4\times 3\times 2}{4^4} = \frac{3}{32}.$$

We get

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)} = \frac{3}{32}\times\frac{4}{3} = \frac{1}{8}.$$

Of course, we get more rapidly the same result observing that we have two possibilities for the third ball and only one for the last one $\mathbb{P}(B|A) = 2/(4\cdot 4) = 1/8$.

Problem 32 Let $A_i, i = 1, 2, 3, 4$ be the events "the person comes from region NW, NE, SW and SE, respectively. Let B be the event "the person is infected'. Then, by the law of total probability (6.5)

$$\mathbb{P}(B) = \frac{1}{10}\frac{15}{100} + \frac{2}{10}\frac{7}{100} + \frac{3}{10}\frac{5}{100} + \frac{4}{10}\frac{2}{100} = \frac{52}{1,000}.$$

Problem 34 Let A be the event "the person is male" and B the event "the person is colorblind". Then A^c is the event "the person is female". Let's use (6.8) and get

$$\mathbb{P}(A|B) = \frac{(1/2)(1/20)}{(1/20)(1/2) + (1/400)(1/2)} = \frac{20}{21}.$$

Problem 38 Since $\mathbb{P}(A^c) = 2 \cdot 2^{-n}$, we have $\mathbb{P}(A) = \frac{2^n-2}{2^n}$. B occurs if all children are male or there is exactly one daughter (the latter can happen in n different ways). These two cases are mutually exclusive. Hence, $\mathbb{P}(B) = \frac{1+n}{2^n}$. The event $A \cap B$ is "there is exactly one daughter". It has probability $\mathbb{P}(A \cap B) = \frac{n}{2^n}$. We conclude that A and B are independent if and only if it holds

$$\frac{n}{2^n} = \frac{2^n - 2}{2^n} \cdot \frac{1+n}{2^n}.$$

This equation becomes $n = 2^{n-1} - 1$. The graphs of $f(x) = x + 1$ and $g(x) = 2^{x-1}$ only intersect once on $x \geq 0$ for $x = 3$, see Figure A.1. We conclude that the two events are independent for $n = 3$ and dependent for any other $n \geq 2, n \neq 3$!.

Problem 39 See the solution of Problem 41 below.

Problem 40 Observe first that $\mathbb{E}\{(X - \mathbb{E}X)(Y - \mathbb{E}Y)\} = \mathbb{E}XY - \mathbb{E}X \cdot \mathbb{E}Y$. Let $\{x_1, \ldots, x_m\}$ be the values taken by X and let $A_i := \{\omega : X(\omega) = x_i\}$, $i = 1, \ldots, m$. Similarly let $\{y_1, \ldots, y_n\}$ be the values taken by Y and let $B_j := \{\omega : Y(\omega) = y_j\}$, $j = 1, \ldots, n$. We get

$$\begin{aligned}\mathbb{E}XY = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j \mathbb{P}(A_i \cap B_j) &= \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j \mathbb{P}(A_i) \cdot \mathbb{P}(B_j) \\ &= \sum_{i=1}^{m} x_i \mathbb{P}(A_i) \sum_{j=1}^{n} y_j \mathbb{P}(B_j) = \mathbb{E}X \cdot \mathbb{E}Y,\end{aligned}$$

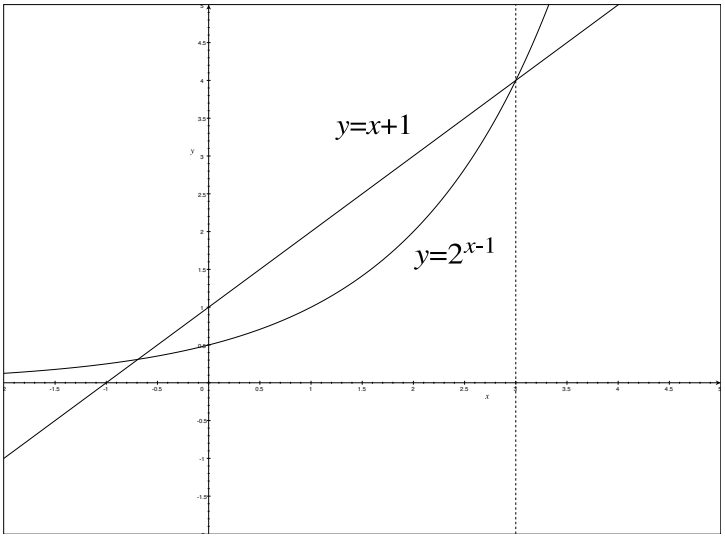

Figure A.1: Graphs of f and g.

where we have used the independence of X and Y.

Problem 41

$$p_Z(k) = \sum_{j=-\infty}^{\infty} p_X(j) \cdot p_Y(k-j) = (p_X * p_Y)(k).$$

Namely, the distribution of Z is just the *convolution* of the distributions of X and Y. In the case when X and Y have Poisson distribution with parameter λ_1 and λ_2, respectively, we get

$$\begin{aligned} p_Z(k) &= \sum_{j=0}^{k} p_X(j) \cdot p_Y(k-j) = \sum_{j=0}^{k} \frac{\lambda_1^j e^{-\lambda_1}}{j!} \cdot \frac{\lambda_2^{k-j} e^{-\lambda_2}}{(k-j)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} \left(\begin{array}{c} k \\ j \end{array} \right) \lambda_1^j \lambda_2^{k-j} = \frac{e^{-(\lambda_1+\lambda_2)}}{k!} (\lambda_1+\lambda_2)^k. \end{aligned} \tag{A.2}$$

We conclude that Z is also Poisson with parameter $\lambda_1 + \lambda_2$.

Problem 42

$$\mathbb{P}(S=k)=\mathbb{P}(X_1+X_2+\cdots X_T=k)=\mathbb{P}\left(\bigcup_{n=k}^{\infty}\{X_1+X_2+\cdots X_n=k,T=n\}\right)$$

$$=\sum_{n=k}^{\infty}\mathbb{P}\left(X_1+X_2+\cdots X_n=k,T=n\right)=\sum_{n=k}^{\infty}\mathbb{P}\left(X_1+X_2+\cdots X_n=k\right)\mathbb{P}\left(T=n\right)$$

$$=\sum_{n=k}^{\infty}\left(\begin{array}{c}n\\k\end{array}\right)p^kq^{n-k}\frac{\lambda^ne^{-\lambda}}{n!}=\frac{(p\lambda)^ke^{-\lambda}}{k!}\sum_{n=k}^{\infty}\frac{(q\lambda)^{n-k}}{(n-k)!}$$

$$=\frac{(p\lambda)^ke^{-\lambda}}{k!}\sum_{j=0}^{\infty}\frac{(q\lambda)^j}{j!}=\frac{(p\lambda)^ke^{-\lambda}}{k!}e^{q\lambda}=\frac{(p\lambda)^ke^{-p\lambda}}{k!}.$$

Problem 43 Recall that S has a Poisson distribution with parameter $p\lambda$ (Problem 42). Let us define $Y_i:=1-X_i, i\geq 1$. Then clearly the Y_i are also i.i.d. Bernoulli trials with parameter $q=1-p$. Moreover, we have

$$Z(\omega)=Y_1(\omega)+Y_2(\omega)+\ldots+Y_{T(\omega)}(\omega).$$

It follows that also Z has a Poisson distribution with parameter $q\lambda$. Now let $j,k\geq 0$. Then

$$\mathbb{P}(S=k,T-S=j)=\mathbb{P}(S=k,T=k+j)=\mathbb{P}(X_1+\cdots+X_{k+j}=k,T=k+j)$$

$$=\mathbb{P}(X_1+\cdots+X_{k+j}=k)\mathbb{P}(T=k+j)=\left(\begin{array}{c}(k+j)\\k\end{array}\right)p^kq^l\frac{\lambda^{k+j}e^{-\lambda}}{(k+j)!}=\frac{(p\lambda)^ke^{-p\lambda}}{k!}\frac{(q\lambda)^je^{-q\lambda}}{j!}.$$

We conclude that $\mathbb{P}(S=k,T-S=j)=\mathbb{P}(S=k)\mathbb{P}(T-S=j)$, namely S and $Z=T-S$ are independent.

Problem 45 By symmetry, $\mathbb{E}(X|Z)=\mathbb{E}(Y|Z)$. Hence,

$$2\mathbb{E}(X|Z)=\mathbb{E}(X+Y|Z)=Z=X+Y.$$

We conclude that $\mathbb{E}(X|Z)=\frac{1}{2}(X+Y)$.

Problem 47 Using (A.2), we get, for $k\geq j$,

$$\mathbb{P}(X_1=j|Y=k)=\frac{\mathbb{P}(X_1=j,X_1+X_2=k)}{\mathbb{P}(X_1+X_2=k)}=\frac{\mathbb{P}(X_1=j,X_2=k-j)}{\mathbb{P}(X_1+X_2=k)}$$

$$=\frac{\frac{\lambda_1^j}{j!}e^{-\lambda_1}\frac{\lambda_2^{k-j}}{(k-j)!}e^{-\lambda_2}}{\frac{(\lambda_1+\lambda_2)^k}{k!}e^{-(\lambda_1+\lambda_2)}}=\left(\begin{array}{c}k\\j\end{array}\right)\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^j\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{k-j}.$$

For $0 \leq j \leq k$, we get a binomial distribution (5.2) with

$$p = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Since we know that the expected value of such a binomial distribution is kp, we get

$$\mathbb{E}(X_1|Y = k) = k\frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

We conclude that

$$\mathbb{E}(X_1|Y) = \frac{\lambda_1}{\lambda_1 + \lambda_2}Y = \frac{\lambda_1}{\lambda_1 + \lambda_2}(X_1 + X_2).$$

Problem 49 We follow the same procedure as in Example 6.6.3. We seek b minimizing $\sum_{i=1}^{5} |y_i - bx_i|$. Define the probabilities

$$p(\{i\}) = \frac{|x_i|}{\sum_{i=1}^{5} |x_i|}, \quad i = 1, 2, 3, 4, 5.$$

Moreover, we compute

$$(y_1/x_1, y_2/x_2, y_3/x_3, y_4/x_4, y_5/x_5) = (1, 2, 7/3, 9/4, 11/6).$$

Ordering these values, we get the table

Table A.1: Distribution of $\frac{Y}{X}$

$\frac{y_i}{x_i}$	1	$\frac{11}{6}$	2	$\frac{9}{4}$	$\frac{7}{3}$
$p(i)$	$\frac{1}{16}$	$\frac{6}{16}$	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{3}{16}$

We know that medians are solutions. We conclude that the unique solution is $b^* = 2$.

A.7 Chapter 7

Problem 59 Recall that the detailed balance condition

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \forall i, j$$

implies that π is invariant (Theorem 7.4.3). We now check that the binomial distribution

$$\pi_i = \left(\begin{array}{c} N \\ i \end{array} \right) \cdot 2^{-N}$$

satisfies the detailed balance condition. Since in the transition matrix of Example 7.2.8 $p_{ij} = 0 \Leftrightarrow p_{ji} = 0$, we only need to verify the condition when $p_{ij} \neq 0$. This happens if and only if i and j differ by one. We can take $i = j - 1$. Then, we have

$$\left(\begin{array}{c} N \\ j-1 \end{array} \right) \cdot 2^{-N} \left(1 - \frac{j-1}{N} \right) = \frac{(N-1)!}{(j-1)!(N-j+1)!} \cdot 2^{-N} = \left(\begin{array}{c} N \\ j \end{array} \right) \cdot 2^{-N} \frac{j}{N},$$

which is the detailed balance condition.

Problem 60 Conditions 1. and 2. of Definition 7.5.1 are clearly satisfied (each $Y(t)$ only takes finitely many values). Consider next

$$\begin{aligned} \mathbb{E}(Y(t+1)|X(0) = x_0, X(1) = x_1, \cdots, X(t) = x_t) &= \mathbb{E}(Y(t+1)|X(t) = x_t) \\ &= \frac{1}{m+n+t+1} \mathbb{E}(X(t+1)|X(t) = x_t) \\ &= \frac{1}{m+n+t+1} \left(x_t \cdot \frac{m+n+t-x_t}{m+n+t} + (x_t+1) \cdot \frac{x_t}{m+n+t} \right) \\ &= \frac{x_t}{m+n+t} = \mathbb{E}(Y(t)|X(t) = x_t). \end{aligned}$$

We conclude that $\mathbb{E}(Y(t+1)|X(t)) = \mathbb{E}(Y(t)|X(t)) = Y(t)$. Hence, Y is an X martingale.

A.8 Chapter 9

Problem 66 Following the hint, suppose $i \sim j$, and let k and l be such that $p_{ij}^{(k)} > 0$ and $p_{ji}^{(l)} > 0$. It follows that $p_{ii}^{(k+l)} \geq p_{ij}^{(k)} \cdot p_{ji}^{(l)} > 0$. Hence, $d(i)$ divides $k+l$. Suppose now that $d(i)$ does not divide $n > 0$. It follows that $k+l+n$ is also not divisible by $d(i)$ and $p_{ii}^{(k+l+n)} = 0$. Since $p_{ii}^{(k+l+n)} \geq p_{ij}^{(k)} \cdot p_{jj}^{(n)} \cdot p_{ji}^{(l)} > 0$, it follows that $p_{jj}^{(n)} = 0$. We conclude from that $p_{jj}^{(n)} > 0$ implies that $d(i)|n$. The latter implies $d(i) \leq d(j)$. By symmetry, $d(j) \leq d(i)$ must also hold and the two periods are equal.

Problem 70 We get

$$P^2 = \begin{bmatrix} p & pq & q^2 & 0 & 0 & \dots \\ p & pq & 0 & q^2 & 0 & \dots \\ p & pq & 0 & 0 & q^2 & \dots \\ . & . & . & . & . & \dots \end{bmatrix}, \quad P^3 = \begin{bmatrix} p & pq & pq^2 & q^3 & 0 & 0 & \dots \\ p & pq & pq^2 & 0 & q^3 & 0 & \dots \\ p & pq & pq^2 & 0 & 0 & q^3 & \dots \\ . & . & . & . & . & . & \dots \end{bmatrix}.$$

By induction, one easily gets

$$P^n = \begin{bmatrix} p & pq & pq^2 & \dots & pq^{n-1} & q^n & 0 & 0 & \dots \\ p & pq & pq^2 & \dots & pq^{n-1} & 0 & q^n & 0 & \dots \\ p & pq & pq^2 & \dots & pq^{n-1} & 0 & 0 & q^n & \dots \\ . & . & . & . & . & . & . & . & \dots \end{bmatrix}.$$

There is a unique positive recurrent, aperiodic class. By Theorem 9.2.13, there is a unique stationary distribution π. Moreover, every row of P converges to π^*. Hence, $\pi_j = pq^j, j = 0, 1, 2, \dots$ which is a geometric distribution. (The invariance relation $\pi = P^*\pi$ may be readily checked observing that $\sum_{k=0}^{\infty} q^k = 1/(1-q) = 1/p$).

Problem 71 Observe that $\pi_N = 2^{-N}$. Since there is a unique positive recurrent class (of period 2), it follows from Theorem 9.2.13, point 2., that

$$\mu_N = 2^N = 2^{(6\times 10^{23})}.$$

To get an idea of such mean recurrence time, suppose that the time unit is a second and that N be "only" 20.000. Then $\mu_N \simeq 2 \times 10^{6000}$ *years*!. The state $N/2$, however, has a mean recurrence time $\mu_{(N/2)}$ which is less than 6 minutes!

Problem 72 1. We see immediately from the graph of the chain that there are two positive recurrent classes of states $C_1 = \{0, 2\}$ and $C_2 = \{3, 4\}$ (see Figure A.2). The closed sets are C_1, C_2 e $\{0, 1, 2\}$. State $\{1\}$ is an equivalence class by itself. Since from 1 it is possible to go (in one step) into the class C_1 and never come back, it follows that 1 is transient. Other states are necessarily all positive recurrent, since in a finite chain there are no null recurrent states.

2. Since there are two positive recurrent classes, the stationary distribution (whose existence is always guaranteed for a finite chain) is not unique. Let

$$\pi_j^1 = \begin{cases} 1/2, & j = 0, 2 \\ 0, & \text{otherwise.} \end{cases}, \quad \pi_j^2 = \begin{cases} 1/2, & j = 3, 4 \\ 0, & \text{otherwise.} \end{cases}$$

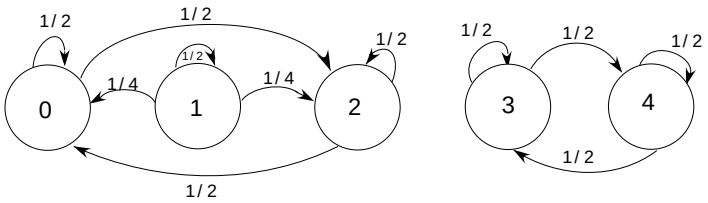

Figure A.2:

Then both π^1 and π^2 are stationary. By the ergodic theorem 9.2.13, point 3., it follows that all other stationary distributions π are obtained as a convex combination of π^1 and π^2:

$$\pi = \lambda\pi^1 + (1-\lambda)\pi^2.$$

Problem 75 Consider the non homogeneous equation

$$1 + p \cdot m(i+1) - m(i) + q \cdot m(i-1) = -1.$$

When $p \neq q$, conditions (9.14) yield

$$\alpha = 0, \quad \beta = \frac{1}{1-2p}.$$

Particular solutions are

$$m_p(i) = \frac{1}{1-2p} i + \gamma.$$

Since the corresponding homogeneous equation is (9.11), the general solution of the non homogeneous equation is

$$m(i) = c_1 + c_2 \left(\frac{q}{p}\right)^i + \frac{i}{1-2p} + \gamma.$$

Imposing the boundary conditions $m(0) = m(N) = 0$, we get

$$m(i) = \frac{1}{1-2p}\left[i - \frac{N\left(1 - \left(\frac{q}{p}\right)^i\right)}{1 - \left(\frac{q}{p}\right)^N}\right].$$

For instance, with $p = 3/4$, $N = 4$ and $i = 1$, we get the average duration of the game $m(1) = 3.4$.

Appendix B

Deutsch's problem

B.1 Qubit

In a classical computer, the basic unit of information storage is the *bit* (binary digit) taking a value of either 0 or 1. The state of a transistor in a processor, the magnetization of a surface in a hard disk and the presence of current in a cable can all be used to represent bits in the same computer. In quantum computing, the unit of quantum information is the *qubit* (quantum bit). That information is described by a state vector in a two-level quantum system whose mathematical description is a two-dimensional vector space over $\mathbb{C}$. The two computational basis states are conventionally written as $|0\rangle$ and $|1\rangle$. A pure qubit state is any *superposition* of these two orthonormal states:

$$\alpha|0\rangle + \beta|1\rangle,$$

where α and β are complex numbers satisfying

$$|\alpha|^2 + |\beta|^2 = 1.$$

Thus, the qubit can take infinitely many values. When we measure this qubit in the standard basis, the probability of outcome $|0\rangle$ is $|\alpha|^2$ and the probability of outcome $|1\rangle$ is $|\beta|^2$. Given the computational basis, we can identify states with unit vectors in $\mathbb{C}^2$ trough the map

$$\alpha|0\rangle + \beta|1\rangle \mapsto \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

More precisely, states correspond to equivalence classes of unit vectors, where two vectors are equivalent if they differ by multiplication by a complex number of modulus one. Similarly, states of a two qubit system can be identified with (equivalence classes of) unit vectors in $\mathbb{C}^4$. Any two-level quantum system can be used to represent a qubit. Several physical implementations which approximate two-level systems to various degrees have been realized. These include photon polarization, electron spin[1], nuclear spin (addressed through NMR), energy states (ground and first excited states) of an atom, electron localization in quantum dots pairs, etc.

A *quantum logic gate* is a transformation on a system with a small number of qubits. Quantum logic gates are reversible, unlike many classical logic gates. Quantum logic gates are represented by *unitary*[2] matrices . The most common quantum gates operate on spaces of one or two qubits. Quantum logic gates are there represented by 2×2 or by 4×4 unitary matrices, respectively. For example, the *Hadamard gate* H operates on a single qubit. It is represented by the Hadamard matrix

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Observe that H, besides being unitary, is Hermitian and, therefore, *involutory*, namely $H^2 = I$. A set of *universal quantum gates* is any set of gates to which any operation possible on a quantum computer can be reduced, that is, any other unitary operation can be expressed as a finite sequence of gates from the set. One simple set of two-qubit universal quantum gates is the Hadamard gate (H), a certain phase rotation gate, and the C-NOT (controlled NOT) gate operating on two qubits. The C-NOT gate flips the second qubit (called *target* qubit) if and only if the first qubit (called *control* qubit) is 1. Let us denote the four basis vectors by

$$|0\rangle \otimes |0\rangle, \quad |0\rangle \otimes |1\rangle, \quad |1\rangle \otimes |0\rangle, \quad |1\rangle \otimes |1\rangle.$$

Then the C-NOT operates on the basis vectors according to

$$|x_1\rangle \otimes |x_2\rangle \mapsto |x_1\rangle \otimes |x_1 \oplus x_2\rangle,$$

[1]The electron spin has the dimension of an angular momentum, but it only takes two values $\pm 1/2$ along any direction.

[2]An $n \times n$ matrix A with complex entries is called unitary if $A^*A = I_n$, where $*$ denotes transposition plus conjugation.

where $\oplus$ denotes the sum mod two. It has the matrix representation

$$CNOT = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

B.2 Deutsch's algorithm

In Deutsch's problem, there is an unknown function $f : \{0,1\} \to \{0,1\}$. We wish to learn if f is *constant* ($f(0) = f(1)$) or *balanced* ($f(0) \neq f(1)$). Any classical deterministic algorithm requires *two* evaluations of f to answer this question. Consider now a two qubit in the state $|x\rangle \otimes |y\rangle = |0\rangle \otimes |1\rangle$. We apply *Hadamard transform* H to each qubit. We get,

$$H|x\rangle = H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) = |x'\rangle, \quad H|y\rangle = H|1\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) = |y'\rangle.$$

Consider now the quantum implementation U_f of the function f which maps $|x'\rangle \otimes |y'\rangle$ to $|x'\rangle \otimes |f(x') \oplus y'\rangle$.[3] Explicitly, we have

$$U_f|x'\rangle \otimes |y'\rangle = \begin{cases} |x'\rangle \otimes |y'\rangle, & \text{if } f(0) = f(1) = 0, \\ -|x'\rangle \otimes |y'\rangle, & \text{if } f(0) = f(1) = 1 \\ |y'\rangle \otimes |y'\rangle, & \text{if } f(0) = 0, = f(1) = 1 \\ -|y'\rangle \otimes |y'\rangle, & \text{if } f(0) = 1, = f(1) = 0 \end{cases}$$

Apply now Hadamard transform once more. Since $H|x'\rangle \otimes |y'\rangle = |0\rangle \otimes |1\rangle$, we get $|0\rangle \otimes |1\rangle$ if $f(0) = f(1) = 0$ and $-|0\rangle \otimes |1\rangle$ if $f(0) = f(1) = 1$. Notice that in both cases, the first qubit is $|0\rangle$. Similarly, since $H|y'\rangle \otimes |y'\rangle = |1\rangle \otimes |1\rangle$, we get $|1\rangle \otimes |1\rangle$ if $f(0) = 0, f(1) = 1$ and $-|1\rangle \otimes |1\rangle$ if $f(0) = 1, f(1) = 0$ and the first qubit is always $|1\rangle$. Hence, reading out the first qubit, we learn whether f is constant or balanced. We conclude that a quantum computer could answer the question about f after just *one* evaluation!

[3]Notice that it is necessary to have U_f acting on *two* qubits so that this represents a *reversible* quantum gate.