MONOGRAPH SERIES
INFORMATION TECHNOLOGIES: RESEARCH
AND THEIR INTERDISCIPLINARY APPLICATIONS

1

ŁUKASZ DĘBOWSKI

INFORMATION THEORY AND STATISTICS

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES

Warsaw, 2013

Publication issued as a part of the project:
'Information technologies: research and their interdisciplinary applications',
Objective 4.1 of Human Capital Operational Programme.
Agreement number UDA-POKL.04.01.01-00-051/10-00.

Publication is co-financed by European Union from resources of European Social Fund.

Project leader: Institute of Computer Science, Polish Academy of Sciences

Project partners: System Research Institute, Polish Academy of Sciences, Nałęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences

Editors-in-chief: Olgierd Hryniewicz
Jan Mielniczuk
Wojciech Penczek
Jacek Waniewski

Reviewer: Brunon Kamiński

Łukasz Dębowski
Institute of Computer Science, Polish Academy of Sciences
Lukasz.Debowski@ipipan.waw.pl
http://www.ipipan.waw.pl/staff/l.debowski/

Publication is distributed free of charge

© Copyright by Łukasz Dębowski
© Copyright by Institute of Computer Science, Polish Academy of Sciences, 2013

ISBN 978-83-63159-06-1
e-ISBN 978-83-63159-07-8

Cover design: Waldemar Słonina

Table of Contents

Preface 5

1 Probabilistic preliminaries 7

2 Entropy and information 15

3 Source coding 27

4 Stationary processes 39

5 Ergodic processes 47

6 Lempel-Ziv code 57

7 Gaussian processes 63

8 Sufficient statistics 71

9 Estimators 81

10 Bayesian inference 93

11 EM algorithm and K-means 101

12 Maximum entropy 109

13 Kolmogorov complexity 117

14 Prefix-free complexity 125

15 Random sequences 135

Solutions of selected exercises 147

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’. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.’

Claude Shannon (1916–2001)

Preface

The object of information theory is the amount of information contained in the data, which is the length of the shortest description of the data given which the data may be decoded. In parallel, the interest of statistics and machine learning lies in the analysis and interpretation of data. Since certain data analysis methods can be used for data compression, the domains of information theory and statistics are intertwined. Concurring with this view, this monograph offers a uniform introduction into basic concepts and facts of information theory and statistics.

The present book is intended for adepts and scholars of computer science and applied mathematics, rather than of engineering. The monograph covers an original selection of problems from the interface of information theory, statistics, physics, and theoretical computer science, which are connected by the question of quantifying information. Problems of lossy compression and information transmission, traditionally discussed in information theory textbooks, have been omitted, as of less importance for the addressed audience. Instead of this, some new material is discussed, such as excess entropy and computational optimality of Bayesian inference.

The book is divided into three parts with the following contents:

Shannon information theory: The first part is a basic course in information theory. Chapter 1 introduces necessary probabilistic notions and facts. In Chapter 2, we present the concept of entropy and mutual information. In Chapter 3, we introduce the problem of source coding and prefix-free codes. The following Chapter 4 and Chapter 5 concern stationary and ergodic stochastic processes, respectively. In Chapter 4 we also discuss the concept of excess entropy, for the first time in terms of a monograph. Chapters 4 and 5 form background to present the Lempel-Ziv code, which is done in Chapter 6. The discussion of Shannon information theory is concluded in Chapter 7, where we discuss the entropy and mutual information for Gaussian processes.

Mathematical statistics: The second part of the book is a brief course in mathematical statistics. In Chapter 8, we discuss sufficient statistics and introduce basic statistical models, namely exponential families. In Chapter 9, we define the maximum likelihood estimator and present its core properties. Chapter 10 concerns Bayesian inference and the problem how to choose a reasonable prior. Chapter 11 discusses the expectation-maximization algorithm. The course is concluded in Chapter 12, where we exhibit the maximum entropy principle and its origin in physics.

Algorithmic information theory: The third part is an exposition of the field of Kolmogorov complexity. In Chapter 13, we introduce plain Kolmogorov complexity and a few related facts such as the information-theoretic Gödel theorem. Chapter 14 is devoted to prefix-free Kolmogorov complexity. We exhibit the links of prefix-free complexity with algorithmic probability and its analogies to entropy such as symmetry of mutual information. The final Chapter 15 treats on Martin-Löf random sequences. Besides exhibiting various characterizations of random sequences, we discuss computational optimality of Bayesian inference for Martin-Löf random parameters. This book is the first monograph to treat on this interesting topic, which links algorithmic information theory with classical problems of mathematical statistics.

A preliminary version of this book has been used as a textbook for a monograph lecture addressed to students of applied mathematics. For this reason each chapter is concluded with a selection of exercises, whereas solutions of chosen exercises are given at the end of the book.

Preparing this book, I have drawn from various sources. The most important book sources have been: Cover and Thomas (2006), Billingsley (1979), Breiman (1992), Brockwell and Davis (1987), Grenander and Szegő (1958), van der Vaart (1998), Keener (2010), Barndorff-Nielsen (1978), Bishop (2006), Grünwald (2007), Li and Vitányi (2008), and Chaitin (1987). The chapters on algorithmic information theory owe much also to a few journal articles: Chaitin (1975a), Vovk and V'yugin (1993), Vovk and V'yugin (1994), V'yugin (2007), Takahashi (2008). All those books and papers can be recommended as complementary reading. At the reader's leisure, I also recommend a prank article by Knuth (1984).

Last but not least, I thank Jan Mielniczuk, Jacek Koronacki, Anna Zalewska, and Brunon Kamiński, who were the first readers of this book, provided me with many helpful comments regarding its composition, and helped me to correct typos.

1

Probabilistic preliminaries

Probability and processes. Expectation. Conditional expectation. Borel-Cantelli lemma. Markov inequality. Monotone and dominated convergence theorems. Martingales. Levy law.

Formally, probability is a normalized, additive, and continuous function of events defined on a suitable domain.

Definition 1.1 (probability space). Probability space $(\Omega, \mathcal{J}, P)$ *is a triple where* Ω *is a certain set (called the* event space*),* $\mathcal{J} \subset 2^{\Omega}$ *is a* σ-field*, and* P *is a* probability measure *on* $(\Omega, \mathcal{J})$*. The* σ*-field* $\mathcal{J}$ *is an algebra of subsets of* Ω *which satisfies*

- $\Omega \in \mathcal{J}$*,*
- $A \in \mathcal{J}$ *implies* $A^c \in \mathcal{J}$*, where* $A^c := \Omega \setminus A$*,*
- $A_1, A_2, A_3, ... \in \mathcal{J}$ *implies* $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{J}$*.*

The elements of $\mathcal{J}$ *are called* events*, whereas the elements of* Ω *are called* elementary events*. Probability measure* $P : \mathcal{J} \to [0, 1]$ *is a normalized measure, i.e., a function of events that satisfies*

- $P(\Omega) = 1$*,*
- $P(A) \geq 0$ *for* $A \in \mathcal{J}$*,*
- $P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} P(A_n)$ *for pairwise disjoint events* $A_1, A_2, A_3, ... \in \mathcal{J}$*,*

If the event space Ω is a finite set, we may define probability measure P by setting the values of $P(\{\omega\})$ for all elementary events $\omega \in \Omega$ and we may put $\mathcal{J} = 2^{\Omega}$.

Example 1.1 (cubic die). The elementary outcomes of a cubic die are $\Omega = \{1, 2, 3, 4, 5, 6\}$. Assuming that the outcomes of the die are equiprobable, we have $P(\{n\}) = 1/6$ so that $P(\Omega) = 1$.

For an uncountably infinite Ω, we may encounter many different measures such that $P(\{\omega\}) = 0$ for all elementary events $\omega \in \Omega$. In that case $\mathcal{J}$ need not be necessarily equal 2^{Ω}.

Example 1.2 (Lebesgue measure on $[0, 1]$*).* Let $\Omega = [0, 1]$ be the unit interval. We define $\mathcal{J}$ as the intersection of all σ-fields that contain all subintervals $[a, b]$, $a, b \in [0, 1]$. Next, the *Lebesgue measure* on $\mathcal{J}$ is defined as the unique measure P such that $P([a, b]) = b - a$. In particular, we obtain $P(\{\omega\}) = 0$ and $P(\Omega) = 1$.

The Lebesgue measure is a generalization of the concept of length from the set of sections to a larger family of sets $\mathcal{J}$. It turns out, however, that $\mathcal{J} \neq 2^{\Omega}$ and the Lebesgue measure cannot be uniquely extended to the family of all subsets of Ω (Billingsley, 1979, Section 3).

The measure-theoretic approach puts different applications of probability on a uniform basis. Using the same framework we can discuss both finite and infinite probability spaces. Let us notice for instance that an uncountably infinite event space arises also when elementary events are infinite sequences of some discrete values. In that case it is convenient to introduce the concept of a random variable.

Definition 1.2 (random variable). *Function $X : \Omega \to \mathbb{X}$ is called a* discrete random variable *if set $\mathbb{X}$ is countable and events*

$$(X = x) := \{\omega \in \Omega : X(\omega) = x\}$$

belong to the σ-field $\mathcal{J}$ for each $x \in \mathbb{X}$. Analogously, function $Y : \Omega \to \mathbb{R}$ is called a real random variable *if events $(Y \leq r) := \{\omega \in \Omega : Y(\omega) \leq r\}$ belong to the σ-field $\mathcal{J}$ for each $r \in \mathbb{R}$.*

Example 1.3 (stochastic process). An infinite sequence of random variables is called a *stochastic process*. For instance, let

$$\Omega = \left\{\omega = (\omega_i)_{i=-\infty}^{\infty} : \omega_i \in \mathbb{X}, i \in \mathbb{Z}\right\}$$

and let $\mathcal{J}$ contain all cylinder sets $\{\omega \in \Omega : \omega_i = s\}$, where i varies over integers and $s \in \mathbb{X}$. Then we may define discrete variables $X_i : \Omega \to \mathbb{X}$ as $X_i(\omega) := \omega_i$.

Let us write $\mathbf{1}\{\phi\} = 1$ if proposition ϕ is true and $\mathbf{1}\{\phi\} = 0$ if proposition ϕ is false. The characteristic function of a set A is defined as

$$I_A(\omega) := \mathbf{1}\{\omega \in A\}. \tag{1.1}$$

The supremum $\sup_{a \in A} a$ is defined as the least real number r such that $r \geq a$ for all $a \in A$. On the other hand, infimum $\inf_{a \in A} a$ is the largest real number r such that $r \leq a$ for all $a \in A$. Having these concepts we may define the expectation.

Definition 1.3 (expectation). *Let P be a probability measure. For a discrete random variable $X \geq 0$, the* expectation *(integral, or average) is defined as*

$$\int X \, \mathrm{d}P := \sum_{x : P(X=x) > 0} P(X = x) \cdot x.$$

For a real random variable $X \geq 0$, we define

$$\int X \, \mathrm{d}P := \sup_{Y \leq X} \int Y \, \mathrm{d}P,$$

where the supremum is taken over all discrete variables Y that satisfy $Y \leq X$. Integrals over subsets are defined as

$$\int_A X \,\mathrm{d}P := \int X I_A \,\mathrm{d}P.$$

For random variables that assume negative values, we put

$$\int X \,\mathrm{d}P := \int_{X>0} X \,\mathrm{d}P - \int_{X<0} (-X) \,\mathrm{d}P,$$

unless both terms are infinite. A more frequent notation for the expectation is

$$\boldsymbol{E} X \equiv \boldsymbol{E}_P X \equiv \int X \,\mathrm{d}P,$$

where we suppress the index P in $\boldsymbol{E}_P X$ for probability measure P.

In the following we shall define conditional expectation and conditional probability with respect to a real random variable. Two necessary prerequisites are a definition of a signed measure and the Lebesgue-Radon-Nikodym theorem. The signed measure is a certain generalization of probability measure.

Definition 1.4 (measure). *The pair $(\Omega, \mathcal{J})$ where Ω is a certain set and $\mathcal{J} \subset 2^\Omega$ is a σ-field is called a* measurable space. *Function $\mu : \mathcal{J} \to \mathbb{R}$ is called a* finite signed measure *if*

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n) \text{ for pairwise disjoint sets } A_1, A_2, A_3, ... \in \mathcal{J}.$$

If additionally $\mu(A) \geq 0$ for all $A \in \mathcal{J}$ then μ is called a finite measure.

Definition 1.5 (mutually singular and absolutely continuous measures). *Let $\mu, \nu : \mathcal{J} \to \mathbb{R}$ be two finite measures. Measures μ and ν are called* mutually singular *if $\mu(A) = \nu(A^c) = 0$ for a certain set $A \in \mathcal{J}$. This fact is written as $\nu \perp \mu$. In contrast, measure ν is called* absolutely continuous *with respect to μ if $\mu(A) = 0$ implies $\nu(A) = 0$ for any set $A \in \mathcal{J}$. This fact is written as $\nu \ll \mu$.*

The integral with respect to a finite measure is defined in the same way as the integral with respect to a probability measure. Moreover, for a σ-field $\mathcal{G}$, we say that a real function $f : \Omega \to \mathbb{R}$ is *$\mathcal{G}$-measurable* if $\{\omega \in \Omega : f(\omega) \leq r\} \in \mathcal{G}$. Let us recall that $\mathcal{J}$-measurable functions have been called random variables in context of probability theory.

Theorem 1.1 (Lebesgue-Radon-Nikodym theorem). *Let $\mu, \nu : \mathcal{J} \to [0, \infty)$ be two finite measures. There exist two unique finite measures $\nu_\perp$ and $\nu_\ll$ such that*

$$\nu = \nu_\perp + \nu_\ll$$

where $\nu_{\perp} \perp \mu$ *and* $\nu_{\ll} \ll \mu$. *Moreover there exists a* $\mathcal{J}$*-measurable function* $f : \Omega \to \mathbb{R}$, *called the Radon-Nikodym derivative of* $\nu_{\ll}$ *with respect to* μ, *such that*

$$\nu_{\ll}(A) = \int_A f \,\mathrm{d}\mu$$

for all $A \in \mathcal{J}$. *Function* f *is given uniquely up to a set of measure* 0, *i.e., if* f, g *are two Radon-Nikodym derivatives then* $\mu(\{\omega \in \Omega : f(\omega) \neq g(\omega)\}) = 0$.

Theorem 1.2 (Hahn-Jordan decomposition). *Let* $\nu : \mathcal{J} \to \mathbb{R}$ *be a finite signed measure. There exist two unique finite measures* ν^+ *and* ν^- *such that*

$$\nu = \nu^+ - \nu^-$$

where $\nu^+ \perp \nu^-$.

Proofs of these theorems can be found in Billingsley (1979, Section 32).

Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{J}$. With respect to this σ-field, we define the concept of conditional expectation.

Definition 1.6 (conditional expectation). Conditional expectation *of a real random variable* X *given* σ*-field* $\mathcal{G}$ *is a* $\mathcal{G}$*-measurable function* $\boldsymbol{E}\,[X|\mathcal{G}]$ *such that*

$$\int_B \boldsymbol{E}\,[X|\mathcal{G}]\,\mathrm{d}P = \int_B X\,\mathrm{d}P \tag{1.2}$$

for any event $B \in \mathcal{G}$.

Theorem 1.3. *Conditional expectation* $\boldsymbol{E}\,[X|\mathcal{G}]$ *exists if* $\boldsymbol{E}X$ *exists and is unique up to sets of probability* 0.

Proof. Define a function $\nu : \mathcal{G} \to \mathbb{R}$ as

$$\nu(B) = \int_B X\,\mathrm{d}P.$$

Function ν is a finite signed measure if $\mathbf{E}\,X = \int X\,\mathrm{d}P$ exists. Let ν^+ and ν^- provide the Hahn-Jordan decomposition of measure ν. These measures are absolutely continuous with respect to P. Let f^+ and f^- be their Radon-Nikodym derivatives. If we put $\mathbf{E}\,[X|\mathcal{G}] = f^+ - f^-$ then we obtain (1.2). The uniqueness of conditional expectation follows by uniqueness of Hahn-Jordan decomposition and the uniqueness of Radon-Nikodym derivatives.

Having conditional expectation, we can define conditional probability with respect to a σ-field.

Definition 1.7 (conditional probability). Conditional probability *of event* A *given* σ*-field* $\mathcal{G}$ *is defined as random variable*

$$P(A|\mathcal{G}) = \boldsymbol{E}\,[I_A|\mathcal{G}]\,.$$

Consecutively, we define conditional probability of an event A given a real random variable (or a stochastic process) Y as $P(A|Y) = P(A|\mathcal{G})$ where $\mathcal{G}$ is the intersection of all σ-fields with respect to which Y is measurable.

Next, we shall be interested in limits of sequences and events. The upper limit of a sequence is defined as

$$\limsup_{n\to\infty} a_n := \lim_{n\to\infty} \sup_{m\geq n} a_m$$

and the lower limit of a sequence is defined as

$$\liminf_{n\to\infty} a_n := \lim_{n\to\infty} \inf_{m\geq n} a_m.$$

These limits exist for any sequence but they may be different. For example, $\limsup_{n\to\infty}(-1)^n = 1$ and $\liminf_{n\to\infty}(-1)^n = -1$. We have $\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = a$ if and only if there exists $\lim_{n\to\infty} a_n = a$.

Analogously we define the upper and the lower limit of a sequence of events. Recall the definition (1.1). We put

$$\limsup_{n\to\infty} A_n := B, \text{ where } I_B = \limsup_{n\to\infty} I_{A_n}$$

and

$$\liminf_{n\to\infty} A_n := B, \text{ where } I_B = \liminf_{n\to\infty} I_{A_n}.$$

Equivalently, we have

$$\limsup_{n\to\infty} A_n = \{\omega : \omega \in A_m \text{ for infinitely many } m\}$$

and

$$\liminf_{n\to\infty} A_n = \{\omega : \omega \in A_m \text{ for all but finitely many } m\}.$$

For a random proposition $\varPhi$, we say that $\varPhi$ holds with probability 1 if $P(\{\omega : \varPhi(\omega) \text{ is true}\}) = 1$. For proving that some events hold with probability 1, the following proposition is particularly useful.

Theorem 1.4 (Borel-Cantelli lemma). *If $\sum_{m=1}^{\infty} P(A_m) < \infty$ for a family of events $A_1, A_2, A_3, \ldots$ then*

$$P\Bigl(\limsup_{n\to\infty} A_n\Bigr) = 0.$$

Proof. Notice that $\sum_{m=1}^{\infty} P(A_m) < \infty$ implies

$$\lim_{m\to\infty} \sum_{k=m}^{\infty} P(A_k) = 0.$$

Hence we obtain

$$\begin{aligned} &P(\{\omega : \omega \in A_m \text{ for infinitely many } m\}) \\ &\quad = P(\{\omega : \forall_{m\geq 1}\exists_{k\geq m}\omega \in A_k\}) \\ &\quad = P\Big(\bigcap_{m=1}^{\infty}\bigcup_{k=m}^{\infty} A_k\Big) \leq \inf_{m\geq 1} P\Big(\bigcup_{k=m}^{\infty} A_k\Big) \leq \inf_{m\geq 1}\sum_{k=m}^{\infty} P(A_k) = 0. \end{aligned}$$

Another handy fact in probability theory is Markov inequality.

Theorem 1.5 (Markov inequality). *Let $\epsilon > 0$ be a fixed constant and let Y be a random variable such that $Y \geq 0$. We have*

$$P\big(Y \geq \epsilon\big) \leq \frac{\boldsymbol{E}Y}{\epsilon}.$$

Proof. Consider random variable $Z = Y/\epsilon$. We have

$$P\big(Y \geq \epsilon\big) = P\big(Z \geq 1\big) = \int_{Z\geq 1} \mathrm{d}P \leq \int_{Z\geq 1} Z\,\mathrm{d}P \leq \int Z\,\mathrm{d}P = \frac{\mathbf{E}Y}{\epsilon}.$$

Finally, let us recall a few results concerning sequences of random variables. If we have a sequence of random variables converging to a limit, we can ask whether the expectation of the limit equals the limit of expectations. In general it is not true but there are three important results stating when the order of expectation and the limit may be switched.

Theorem 1.6 (monotone convergence theorem). *Let $(X_i)_{i=1}^{\infty}$ be a sequence of nonnegative ($X_i \geq 0$), nondecreasing ($X_{i+1} \geq X_i$) real random variables. Then function $X = \lim_{n\to\infty} X_n$ is also a real random variable and*

$$\lim_{n\to\infty}\int X_n\,\mathrm{d}P = \int X\,\mathrm{d}P.$$

Theorem 1.7 (Fatou lemma). *Let $(X_i)_{i=1}^{\infty}$ be a sequence of nonnegative ($X_i \geq 0$) real random variables. Then function $X = \liminf_{n\to\infty} X_n$ is also a real random variable and*

$$\liminf_{n\to\infty}\int X_n\,\mathrm{d}P \geq \int X\,\mathrm{d}P.$$

Theorem 1.8 (Lebesgue dominated convergence theorem). *Let $(X_i)_{i=1}^{\infty}$ be a sequence of real random variables which are dominated by an integrable real random variable Y, i.e., $|X| \leq Y$ and $\int Y\,\mathrm{d}P < \infty$. If there exists limit $X = \lim_{n\to\infty} X_n$ then*

$$\lim_{n\to\infty}\int X_n\,\mathrm{d}P = \int X\,\mathrm{d}P.$$

Proofs of these results can be found in Billingsley (1979, Section 16).

An important instance of a stochastic process is a martingale.

Definition 1.8 (filtration and martingale). *A sequence of σ-fields $(\mathcal{G}_i)_{i=1}^{\infty}$ is called a* filtration *if $\mathcal{G}_{n+1} \supset \mathcal{G}_n$. A sequence of random variables $(X_i)_{i=1}^{\infty}$ is called a* martingale *with respect to filtration $(\mathcal{G}_i)_{i=1}^{\infty}$ if each X_n is $\mathcal{G}_n$-measurable, $\boldsymbol{E}\,|X_n| < \infty$, and $\boldsymbol{E}\,[X_{n+1}|\mathcal{G}_n] = X_n$.*

A particular example of a martingale is a collection of conditional probabilities with respect to a rising sequence of σ-fields.

Example 1.4. Let $(X_i)_{i=1}^{\infty}$ be an arbitrary sequence of random variables and let $\mathcal{G}_n$ be the intersection of all σ-fields with respect to which $(X_1, ..., X_n)$ is measurable. Then $(\mathcal{G}_i)_{i=1}^{\infty}$ is a filtration and the sequence of conditional probabilities $(P(A|\mathcal{G}_i))_{i=1}^{\infty}$ is a martingale with respect to this filtration.

A fundamental property of martingales is that they converge with probability 1. Here we will state this result in a particular case.

Theorem 1.9 (Levy law). *Let $(X_i)_{i=1}^{\infty}$ and $\mathcal{G}_n$ be as in Example 1.4. Define $\mathcal{G}$ as the intersection of all σ-fields which contain all $\mathcal{G}_n$. Equality*

$$\lim_{n\to\infty} P\big(A|\mathcal{G}_n\big) = P\big(A|\mathcal{G}\big)$$

holds with probability 1.

A proof of this theorem can be found in Billingsley (1979, Section 35).

Exercises

1. *(Monty Hall paradox)* A participant of the "Let's Make A Deal" quiz hosted by Monty Hall is exposed to three closed doors. Behind one of the doors there is an expensive car, behind two other doors there are two goats. Monty Hall asks the participant to choose a door. It is known that there is a goat behind one of the not selected doors. This door is opened and the goat is shown. Now the participant is asked to choose one of the remaining two doors. He will get what is behind it. Should he choose the same door as before or the other one?
2. Prove that if f is a measurable function so is $|f|$. Is the converse true?
3. Prove that

$$\begin{aligned} \mathbf{E}\,[\mathbf{E}\,[X|\mathcal{G}]] &= \mathbf{E}\,X, \qquad (1.3)\\ \mathbf{E}\,[X|\mathcal{J}] &= X \text{ holds with probability } 1,\\ \mathbf{E}\,[X|\,\{\emptyset, \varOmega\}] &= \mathbf{E}\,X. \end{aligned}$$

4. Define $P(A|B) = P(A \cap B)/P(B)$ for $P(B) > 0$. Show that, for a discrete random variable X, we have $P(A|X)(\omega) = P(A|X = x)$ if $X(\omega) = x$.
5. Show that the upper limit $\limsup_{n\to\infty} a_n$ and the lower limit $\liminf_{n\to\infty} a_n$ exist for any sequence $(a_n)_{n=1}^{\infty}$.

6. Prove that

$$P\big(\liminf_{n\to\infty} A_n\big) \leq \liminf_{n\to\infty} P\big(A_n\big),$$
$$P\big(\limsup_{n\to\infty} A_n\big) \geq \limsup_{n\to\infty} P\big(A_n\big).$$

7. Prove the second Borel-Cantelli lemma: If $\sum_{m=1}^{\infty} P(A_m) = \infty$ for a family of independent events $A_1, A_2, A_3, ...$ then

$$P\big(\limsup_{n\to\infty} A_n\big) = 1.$$

8. Let X be a random variable with expectation $\mathbf{E}\, X = \mu$ and variance $\mathbf{E}\, [X-\mu]^2 = \sigma^2$. Prove the Chebyshev inequality: For any real number $k > 0$,

$$P\big(|X-\mu| \geq k\sigma\big) \leq \frac{1}{k^2}.$$

9. Let $(X_i)_{i=1}^{\infty}$ be independent identically distributed random variables (see Definition 2.2) with finite expectation $\mathbf{E}\, X_i = \mu$ and finite variance $\mathbf{E}\, [X_i - \mu]^2 = \sigma^2$. Using Markov inequality, prove the weak law of large numbers

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0$$

for any $\epsilon > 0$.

10. Let $(X_i)_{i=1}^{\infty}$ be independent identically distributed random variables with finite expectation $\mathbf{E}\, X_i = \mu$, finite variance $\mathbf{E}\, [X_i - \mu]^2 = \sigma^2$, and finite fourth moment $\mathbf{E}\, [X_i - \mu]^4 = \mu_4$. Using Markov inequality and Borel-Cantelli lemma, demonstrate the strong law of large numbers

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1. \tag{1.4}$$

11. Give an example of a sequence of random variables $(X_i)_{i=1}^{\infty}$ such that $\lim_{n\to\infty} \int X_n \,\mathrm{d}P \neq \int (\lim_{n\to\infty} X_n)\,\mathrm{d}P$.
12. Prove the result from Example 1.4.

2

Entropy and information

Entropy. Kullback-Leibler divergence. Mutual information.

Entropy of a random variable on a probability space is the fundamental concept of information theory developed by Claude Shannon (1916–2001) in papers (Shannon, 1948, 1951). Basic definitions in probability, such as probability spaces, random variables, and expectations, have been reviewed in Chapter 1. In information theory the following random variables play an important role.

Definition 2.1 (probability as a random variable). *Let X and Y be discrete variables and A be an event on a probability space $(\mathit{\Omega}, \mathcal{J}, P)$. We define $P(X)$ as a discrete random variable such that*

$$P\big(X\big)(\omega) = P\big(X = x\big) \iff X(\omega) = x.$$

Analogously we define $P(X|Y)$ and $P(X|A)$ as

$$\begin{aligned} P\big(X|Y\big)(\omega) = P\big(X = x|Y = y\big) &\iff X(\omega) = x \text{ and } Y(\omega) = y, \\ P\big(X|A\big)(\omega) = P\big(X = x|A\big) &\iff X(\omega) = x, \end{aligned}$$

where the conditional probability is $P(B|A) = P(B \cap A)/P(A)$ for $P(A) > 0$.

We write $P(Y) = P(X_1, X_2, ..., X_n)$ for $Y = (X_1, X_2, ..., X_n)$. (The same convention will be adopted for other functions of random variables.) Given this concept we may easily define independent variables.

Definition 2.2 (independence). *We say that random variables $X_1, X_2, ..., X_n$ are independent if*

$$P\big(X_1, X_2, ..., X_n\big) = \prod_{i=1}^{n} P\big(X_i\big).$$

Analogously, we say that random variables $X_1, X_2, X_3, ...$ are independent if $X_1, X_2, ..., X_n$ are independent for any n.

Example 2.1. Let $\mathit{\Omega} = [0, 1]$ be the unit section and let P be the Lebesgue measure. Define real random variable $Y(\omega) = \omega$. If we consider its binary expansion $Y = \sum_{i=1}^{\infty} 2^{-i} Z_i$, where $Z_i : \mathit{\Omega} \to \{0, 1\}$, then $P(Z_1, Z_2, ..., Z_n) = 2^{-n} = \prod_{i=1}^{n} P(Z_i)$. Consequently, variables $Z_1, Z_2, Z_3, ...$ are independent.

Another concept that we need is the expectation $\mathbf{E}\, X$ of a random variable X (see Definition 1.3). One of fundamental properties of the expectation is its additivity.

Theorem 2.1. *If $\boldsymbol{E}X + \boldsymbol{E}Y$ exists then $X + Y$ is defined with probability 1, $\boldsymbol{E}(X+Y)$ exists and*

$$\boldsymbol{E}\big(X+Y\big) = \boldsymbol{E}X + \boldsymbol{E}Y.$$

Now we will introduce the main concept of information theory, which is the entropy. Some interpretation of this quantity is the average uncertainty carried by a random variable or a tuple of random variables, regardless of their particular values. We expect that uncertainty adds for probabilistically independent sources. Thus entropy $H(X)$ is a functional of random variable $P(X)$ which is additive for independent random variables. Formally, for $P(X,Y) = P(X)P(Y)$, we postulate $H(X,Y) = H(X) + H(Y)$. Because $\log(xy) = \log x + \log y$ for the logarithm function, the following definition comes as a very natural idea.

Definition 2.3 (entropy). *The* entropy *of a discrete variable X is defined as*

$$H\big(X\big) := \boldsymbol{E}\big[-\log P\big(X\big)\big]. \qquad (2.1)$$

Traditionally, it is assumed that log *is the logarithm to the base* 2*.*

Because $\log P(X) \le 0$, we put the minus sign in the definition (2.1) so that entropy be positive. Equivalently, we have

$$H\big(X\big) = -\sum_{x:P(X=x)>0} P\big(X=x\big)\log P\big(X=x\big),$$

Indeed, we can verify that for $P(X,Y) = P(X)P(Y)$,

$$\begin{aligned} H\big(X,Y\big) &= \mathbf{E}\left[-\log P\big(X,Y\big)\right] = \mathbf{E}\left[-\log P\big(X\big) - \log P\big(X\big)\right] \\ &= \mathbf{E}\left[-\log P\big(X\big)\right] + \mathbf{E}\left[-\log P\big(X\big)\right] = H\big(X\big) + H\big(Y\big). \end{aligned}$$

Example 2.2. Let $P(X=0) = 1/3$ and $P(X=1) = 2/3$. Then

$$H\big(X\big) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = \log 3 - 2/3 = 0.918....$$

We obtain the same value for $P(X=0) = 2/3$ and $P(X=1) = 1/3$ because entropy depends on distribution $P(X)$ rather than on particular values of X. On the other hand, for $P(X=0) = 1/2$ and $P(X=1) = 1/2$, we have

$$H\big(X\big) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2 = 1.$$

The plot of function $H(X)$ for a binary variable (cf., Figure 1) shows that $H(X)$ attains maximum 1 when the variable values are equiprobable whilst $H(X)$ attains minimum 0 when the probability is concentrated on a single value.

What is the range of function $H(X)$ in general? Because function $f(p) = -p\log p$ is strictly positive for $p \in (0,1)$ and equals 0 for $p = 1$, it can be easily seen that:

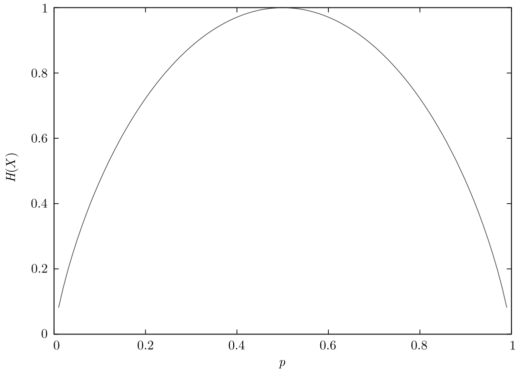

Fig. 1. Entropy $H(X) = -p \log p - (1-p)\log(1-p)$ for $P(X=0) = p$ and $P(X = 1) = 1 - p$.

Theorem 2.2. *$H(X) \geq 0$, whereas $H(X) = 0$ if and only if X assumes only a single value.*

This fact agrees intuitively with the idea that constants carry no uncertainty. On the other hand, assume that X takes values $x \in \{1, 2, ..., n\}$ with equal probabilities $P(X = x) = 1/n$. Then we have

$$H(X) = -\sum_{x=1}^{n} \frac{1}{n} \log \frac{1}{n} = \sum_{x=1}^{n} \frac{1}{n} \log n = \log n.$$

As we will see, $\log n$ is the maximal value of $H(X)$ given X assumes values in $\{1, 2, ..., n\}$. That fact agrees with the intuition that the highest uncertainty occurs for uniformly distributed variables. The simplest proof of this property goes via Kullback-Leibler divergence and Jensen inequality, which are objects of their own interest.

First, it is convenient to enhance the notation and introduce discrete probability distributions.

Definition 2.4. *A discrete probability distribution is a function $p : \mathbb{X} \to [0, 1]$ defined on a countable set $\mathbb{X}$ such that $p(x) \geq 0$ and $\sum_x p(x) = 1$.*

For example we may put $p(x) = P(X = x)$. Thus we can define the entropy of a probability distribution.

Definition 2.5 (entropy revisited). *The entropy of a discrete probability distribution is denoted as*

$$H(p) := -\sum_{x:p(x)>0} p(x) \log p(x).$$

For two distributions we define a similar function.

Definition 2.6 (KL divergence). Kullback-Leibler divergence*, or* relative entropy *of probability distributions p and q is defined as*

$$D(p||q) := \sum_{x:p(x)>0} p(x) \log \frac{p(x)}{q(x)} = -\sum_{x:p(x)>0} p(x) \log q(x) - H(p).$$

We will show that $D(p||q) \geq 0$. For this purpose, we will consider a special class of functions.

Definition 2.7 (convex and concave functions). *A real function $f : \mathbb{R} \to \mathbb{R}$ is* convex *if*

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

for $p_i \geq 0$, $i = 1, 2$, and $p_1 + p_2 = 1$. Moreover, f is called strictly convex *if*

$$p_1 f(x_1) + p_2 f(x_2) > f(p_1 x_1 + p_2 x_2)$$

for $p_i > 0$, $i = 1, 2$, and $p_1 + p_2 = 1$. We say that function f is concave *if $-f$ is* convex*, whereas f is* strictly concave *if $-f$ is* strictly convex*.*

Example 2.3. If function f has a positive second derivative then it is strictly convex. Hence functions $h(x) = -\log x$ and $g(x) = x^2$ are strictly convex.

The expectation of a convex function is greater than the function of the expected argument.

Theorem 2.3 (Jensen inequality). *If f is a convex function and p is a discrete probability distribution over real values then*

$$\sum_{x:p(x)>0} p(x) f(x) \geq f\Big(\sum_{x:p(x)>0} p(x) \cdot x \Big).$$

Moreover, if f is strictly convex then

$$\sum_{x:p(x)>0} p(x) f(x) = f\Big(\sum_{x:p(x)>0} p(x) \cdot x \Big)$$

holds if and only if distribution p is concentrated on a single value.

The proof proceeds by an easy induction on the number of values that p assumes.

Now we prove the requested proposition.

Theorem 2.4. *We have*

$$D(p||q) \geq 0,$$

where the equality holds if and only if $p = q$.

Proof. By the Jensen inequality for $f(y) = -\log y$, we have

$$\begin{aligned} D(p||q) &= -\sum_{x:p(x)>0} p(x) \log \frac{q(x)}{p(x)} \geq -\log\Big(\sum_{x:p(x)>0} p(x) \frac{q(x)}{p(x)}\Big) \\ &= -\log\Big(\sum_{x:p(x)>0} q(x)\Big) \geq -\log 1 = 0, \end{aligned}$$

with equality if and only if $p = q$.

Given this fact we can show when entropy is maximized.

Theorem 2.5. *Let X assume values in $\{1, 2, ..., n\}$. We have $H(X) \leq \log n$, whereas $H(X) = \log n$ if and only if $P(X = x) = 1/n$.*

Remark: If the range of variable X is infinite then entropy $H(X)$ may be infinite.

Proof. Let $p(x) = P(X = x)$ and $q(x) = 1/n$. Then

$$0 \leq D(p||q) = \sum_{x:p(x)>0} p(x) \log \frac{p(x)}{1/n} = \log n - H\big(X\big),$$

where the equality occurs if and only if $p = q$.

The next important question is what is the behavior of entropy under conditioning. The intuition is that given additional information, the uncertainty should decrease. So should entropy. There are, however, two distinct ways of defining conditional entropy.

Definition 2.8 (conditional entropy). Conditional entropy *of a discrete variable X given event A is*

$$H\big(X|A\big) := H(p) \textit{ for } p(x) = P\big(X = x|A\big).$$

Conditional entropy *of X given a discrete variable Y is defined as*

$$H\big(X|Y\big) := \sum_{y:P(Y=y)>0} P\big(Y = y\big) H\big(X|Y = y\big).$$

Both $H(X|A)$ and $H(X|Y)$ are nonnegative.

Theorem 2.6. *$H(X|Y) = 0$ holds if and only if $X = f(Y)$ for a certain function f except for a set of probability 0.*

Proof. Observe that $H(X|Y) = 0$ if and only if $H(X|Y = y) = 0$ for all y such that $P(Y = y) > 0$. This holds if and only if given $(Y = y)$ with $P(Y = y) > 0$, variable X is concentrated on a single value. Denoting this value as $f(y)$, we obtain $X = f(Y)$, except for the union of those sets $(Y = y)$ which have probability 0.

Let us note that inequality $H(X|A) \leq H(X)$ need not hold.

Example 2.4. Let $P(X = 0|A) = P(X = 1|A) = 1/2$, whereas $P(X = 0|A^c) = 1$ and $P(X = 1|A^c) = 0$. Assuming $P(A) = 1/2$, we have $P(X = 0) = (1/2) \cdot (1/2) + (1/2) = 3/4$ and $P(X = 0) = (1/2) \cdot (1/2) = 1/4$ so

$$H(X) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = \log 4 - \frac{3}{4}\log 3 = 0.811....$$

On the other hand, we have $H(X|A) = \log 2 = 1$.

Despite that fact, it is true that $H(X|Y) \le H(X)$ holds in general. Thus entropy decreases given additional information on average. Before we prove it, let us observe:

Theorem 2.7. *We have*

$$H(X|Y) = \boldsymbol{E}\left[-\log P(X|Y)\right].$$

Proof. Observe

$$\begin{aligned} H(X|Y) &= \sum_{y:P(Y=y)>0} P(Y=y)H(X|Y=y) \\ &= -\sum_{x,y:P(X=x,Y=y)>0} P(Y=y)P(X=x|Y=y)\log P(X=x|Y=y) \\ &= -\sum_{x,y:P(X=x,Y=y)>0} P(X=x,Y=y)\log P(X=x|Y=y) \\ &= \mathbf{E}\left[-\log P(X|Y)\right]. \end{aligned}$$

Because $P(Y)P(X|Y) = P(X,Y)$, by Theorem 2.7 we obtain

$$H(Y) + H(X|Y) = H(X,Y).$$

Hence

$$H(X,Y) \ge H(Y).$$

To show that $H(X)$ is greater than $H(X|Y)$, it is convenient to introduce another important concept.

Definition 2.9 (mutual information). Mutual information *between discrete variables X and Y is defined as*

$$I(X;Y) := \boldsymbol{E}\left[\log\frac{P(X,Y)}{P(X)P(Y)}\right].$$

Let us observe that $I(X;X) = H(X)$. Hence entropy is sometimes called *self-information.*

Mutual information is nonnegative because it is a special instance of Kullback-Leibler divergence.

Theorem 2.8. *We have*

$$I\big(X;Y\big) \geq 0,$$

where the equality holds if and only if X and Y are independent.

Proof. Let $p(x,y) = P(X = x, Y = y)$ and $q(x,y) = P(X = x)P(Y = y)$. Then we have

$$I\big(X;Y\big) = \sum_{(x,y):p(x,y)>0} p(x,y) \log \frac{p(x,y)}{q(x,y)} = D(p||q) \geq 0$$

with the equality exactly for $p = q$.

By the definition of mutual information and by Theorem 2.7,

$$\begin{aligned} H\big(X,Y\big) + I\big(X;Y\big) &= H\big(X\big) + H\big(Y\big), \\ H\big(X|Y\big) + I\big(X;Y\big) &= H\big(X\big). \end{aligned} \tag{2.2}$$

Hence by Theorem 2.8, we have

$$\begin{aligned} H\big(X\big) + H\big(Y\big) &\geq H\big(X,Y\big), \\ H\big(X\big) &\geq H\big(X|Y\big),\ I\big(X;Y\big). \end{aligned}$$

Moreover, we have $H(X|Y) = H(Y)$ if X and Y are independent, which also agrees with intuition.

In a similar fashion as for entropy, we may introduce conditional mutual information.

Definition 2.10 (conditional mutual information). Conditional mutual information *between discrete variables X and Y given event A is*

$$\begin{aligned} I\big(X;Y|A\big) := D(p||q) \text{ for } p(x,y) = P\big(X = x, Y = y|A\big) \\ \text{and } q(x,y) = P\big(X = x|A\big)P(Y = y|A). \end{aligned}$$

Conditional mutual information *between discrete variables X and Y given variable Z is defined as*

$$I\big(X;Y|Z\big) := \sum_{z:P(Z=z)>0} P\big(Z = z\big) I\big(X;Y|Z = z\big).$$

Both $I(X;Y|A)$ and $I(X;Y|Z)$ are nonnegative. As in the case of conditional entropy, the following proposition is true:

Theorem 2.9. *We have*

$$I\big(X;Y|Z\big) := \boldsymbol{E}\left[\log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right].$$

The notion of conditional information is useful when analyzing conditional independence.

Definition 2.11 (conditional independence). *Variables $X_1, X_2, ..., X_n$ are* conditionally independent *given Z if*

$$P\big(X_1, X_2, ..., X_n | Z\big) = \prod_{i=1}^{n} P\big(X_i | Z\big).$$

Analogously, we say that variables $X_1, X_2, X_3, ...$ are conditionally independent given Z if $X_1, X_2, ..., X_n$ are conditionally independent given Z for any n.

Example 2.5. Let $Y = f(Z)$ be a function of variable Z, whereas X be an arbitrary variable. Variables X and Y are conditionally independent given Z. Indeed, we have

$$\begin{aligned} P\big(X = x, Y = y | Z = z\big) &= P\big(X = x | Z = z\big)\mathbf{1}\{y = f(z)\} \\ &= P\big(X = x | Z = z\big)P\big(Y = y | Z = z\big). \end{aligned}$$

Example 2.6. Let variables X, Y, and Z be independent assuming with equal probability values 0 and 1. Variables $U = X + Z$ and $W = Y + Z$ are conditionally independent given Z. Indeed, we have

$$\begin{aligned} &P\big(U = u, W = w | Z = z\big) = P\big(X = u - z, Y = w - z\big) \\ &= P\big(X = u - z\big)P\big(Y = w - z\big) = P\big(U = u | Z = z\big)P\big(W = w | Z = z\big). \end{aligned}$$

It can be checked, however, that U and V are not independent.

Definition 2.12 (Markov chain). *A stochastic process $(X_i)_{i=-\infty}^{\infty}$ is called a* Markov chain *if*

$$P\big(X_i | X_{i-1}, X_{i-2}, ..., X_{i-n}\big) = P\big(X_i | X_{i-1}\big)$$

holds for any $i \in \mathbb{Z}$ and $n \in \mathbb{N}$.

Example 2.7. For a Markov chain $(X_i)_{i=-\infty}^{\infty}$, variables X_i and X_k are conditionally independent given X_j if $i \le j \le k$. Indeed, after simple calculations we obtain $P(X_k | X_i, X_j) = P(X_k | X_j)$, and hence

$$P\big(X_i, X_k | X_j\big) = P\big(X_i | X_j\big)P\big(X_k | X_i, X_j\big) = P\big(X_i | X_j\big)P\big(X_k | X_j\big).$$

As in the case of plain mutual information the following fact is true:

Theorem 2.10. *We have*

$$I\big(X; Y | Z\big) \ge 0,$$

where the equality holds if and only if X and Y are conditionally independent given Z.

Of particular interest is this generalization of formula (2.2):

Theorem 2.11. *We have*

$$I(X;Y|Z) + I(X;Z) = I(X;Y,Z).$$

Remark: Hence, variables X and (Y,Z) are independent if and only if X and Z are independent and X and Y are independent given Z.

Proof.

$$\begin{aligned} I(X;Y|Z) + I(X;Z) &= \mathbf{E}\left[\log\frac{P(X,Y,Z)P(Z)}{P(X,Z)P(Y,Z)}\right] + \mathbf{E}\left[\log\frac{P(X,Z)}{P(X)P(Z)}\right] \\ &= \mathbf{E}\left[\log\frac{P(X,Y,Z)}{P(X)P(Y,Z)}\right] = I(X;Y,Z). \end{aligned}$$

Finally, one can ask whether conditional entropy and mutual information may be expressed by entropies of tuples of variables. The answer is positive if the entropies are finite.

Theorem 2.12. *If entropies $H(X)$, $H(Y)$, and $H(Z)$ are finite, we observe these identities:*

$$\begin{aligned} H(X|Y) &= H(X,Y) - H(Y), \\ I(X;Y) &= H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y), \\ I(X;Y|Z) &= H(X|Z) + H(Y|Z) - H(X,Y|Z) \\ &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z), \end{aligned}$$

where all terms are finite and nonnegative.

The proof is left as an easy exercise.

Exercises

1. *(Entropy)* We are tossing a coin until the first tail is obtained. The outcome of the experiment is the number of tossings. Compute the entropy.
2. We are drawing balls from an urn. The outcome of the experiment is a sequence of drawn balls. Is the entropy higher for drawing with replacement or for drawing without replacement?
3. We have two random variables X and Y with disjoint sets of values. Let Z take values $P(Z=0)=p$ and $P(Z=1)=1-p$ and be independent from X and Y. Compute the entropy of variable

$$U = \begin{cases} X, & \text{if } Z = 0, \\ Y, & \text{if } Z = 1. \end{cases}$$

4. For a function g show that

$$H(g(X)) \le H(X).$$

5. Show that function $d(X,Y) = H(X|Y)+H(Y|X)$ is a distance (a metric), i.e., it satisfies:
   - $d(X,Y) = 0$ if there is a bijection between X and Y;
   - $d(X,Y) = d(Y,X)$;
   - $d(X,Z) \leq d(X,Y) + d(Y,Z)$.
6. *(Venn diagrams)* The dependence between entropy, conditional entropy and mutual information can be depicted by Venn diagrams. The diagram for two variables is given in Figure 2, whereas the diagram for three variables is presented in Figure 3.

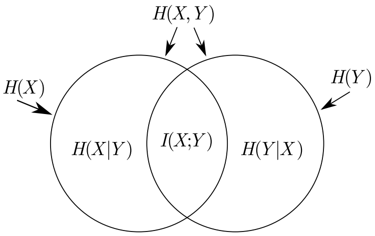

Fig. 2. Venn diagram for two random variables.

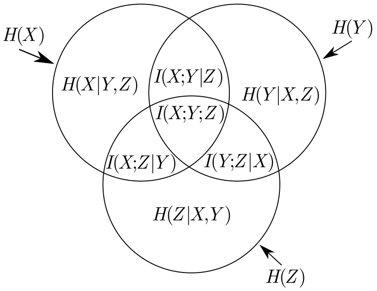

Fig. 3. Venn diagram for three random variables.

Quantity $I(X;Y;Z)$, appearing in Figure 3, is called *triple information.* It can be defined as

$$I\big(X;Y;Z\big) := I\big(X;Y\big) - I\big(X;Y|Z\big).$$

Show that $I(X;Y;Z) = I(Y;X;Z) = I(Y;Z;X)$. Construct also variables X, Y, Z such that $I(X;Y;Z) > 0$ and $I(X;Y;Z) < 0$.

7. Let A be an event. Show that

$$\big|I\big(X;Y\big) - P\big(A\big)I\big(X;Y|A\big) - P\big(A^c\big)I\big(X;Y|A^c\big)\big| \leq 1.$$

8. *(Data-processing inequality)* Let X and Z be conditionally independent given Y. Show that
$$I\big(X;Y\big) \geq I\big(X;Z\big).$$
9. For a function g show that
$$I\big(X;g(Y)\big) \leq I\big(X;Y\big).$$
10. For a Markov chain $(X_i)_{i=-\infty}^{\infty}$ prove that
$$I\big(X_i;X_k\big) \leq I\big(X_i;X_j\big) \text{ for } i \leq j \leq k.$$
11. *(Chain rules)* Prove the chain rule
$$H\big(X_1,...,X_n\big) = H\big(X_1\big) + \sum_{i=2}^{n} H\big(X_i|X_1,...X_{i-1}\big).$$
12. Let variables $X_1, X_2, ..., X_n$ be independent and conditionally independent given Z. Prove that
$$I\big((X_1,X_2,...,X_n);Z\big) = \sum_{i=1}^{n} I\big(X_i;Z\big).$$
13. *(Fano inequality)* Let random variable $X : \Omega \to \mathbb{X}$ be approximated by a random variable $\hat{X}$, which is a function of another variable Y. Denote the probability of error $p_e = P(X \neq \hat{X})$ and the entropy $H(p_e) = -p_e \log p_e - (1-p_e)\log(1-p_e)$. Show that
$$H\big(p_e\big) + p_e \log \operatorname{card} \mathbb{X} \geq H\big(X|\hat{X}\big) \geq H\big(X|Y\big).$$
14. Let X and $\hat{X}$ be two independent random variables with distributions $p(x) = P(X = x)$ and $r(x) = P(\hat{X} = x)$. Show that
$$P\big(X = \hat{X}\big) \geq 2^{-H(p)-D(p||r)}.$$
15. *(Infinite entropy)* Consider a random variable X taking values in natural numbers (without zero) whose distribution is
$$P\big(X = n\big) = \frac{C}{n(\log n)^{\beta}}, \quad n \geq 1,$$
where $\beta \in (1,2]$. Show that entropy $H(X)$ is infinite.
16. Consider another variable X taking values in natural numbers. Show that $H(X) < \infty$ if $\mathbf{E}\,X < \infty$.
17. *(Stirling approximation)* Show that
$$\left(\frac{n}{e}\right)^n \leq n! \leq n\left(\frac{n}{e}\right)^n$$
Hint: Use bounds $\int_0^n \ln x \,\mathrm{d}x \leq \ln n! \leq \int_1^n \ln x \,\mathrm{d}x + \ln n$.

18. Use the Stirling approximation to show that for $p \in [0,1]$ and $k_n = \lceil np \rceil$ we have
$$\lim_{n\to\infty} \frac{1}{n} \log \binom{n}{k_n} = H(p),$$
where $H(p) = -p\log p - (1-p)\log(1-p)$,
19. *(Bregman divergence)* Let ϕ be a differentiable and strictly convex function of vector $x = (x_1, x_2, ..., x_k)$. Bregman divergence is defined as
$$d_\phi(x,y) = \phi(x) - \phi(y) - \sum_i (x_i - y_i)\frac{\partial\phi(y)}{\partial y_i}.$$
Show that for $\phi(p) = -H(p) = \sum_i p_i \log p_i$, Bregman divergence equals Kullback-Leibler divergence, i.e., $d_\phi(p,q) = D(p||q)$.
20. Show that $d_\phi(x,y) \geq 0$ and equality holds if and only if $x = y$.
21. Bregman information of a random variable X is defined as $I_\phi(X) = \mathbf{E}\, d_\phi(X, \mathbf{E}\, X)$. Show that $I_\phi(X) = \mathbf{E}\,\phi(X) - \phi(\mathbf{E}\, X)$.
22. What is the Bregman divergence for $\phi(x) = \sum_i x_i^2$? What is Bregman information $I_\phi(X)$ in that case?
23. *(Generalized Pythagoras theorem)* Define $\operatorname{argmin}_{x\in S} f(x)$ as the argument $x \in S$ for which the function f attains the minimal value. Let S be a convex set of points, let $x_1 \in S$ and let $x_2 = \operatorname{argmin}_{x\in S} d_\phi(x, x_3)$. Show that
$$d_\phi(x_1,x_2) + d_\phi(x_2,x_3) \leq d_\phi(x_1,x_3).$$
Hint: Let $x_\lambda = \lambda x_1 + (1-\lambda)x_2$. Show that
$$0 \leq \left.\frac{\partial d_\phi(x_\lambda, x_3)}{\partial\lambda}\right|_{\lambda=0} = d_\phi(x_1,x_3) - d_\phi(x_1,x_2) - d_\phi(x_2,x_3).$$

3

Source coding

Uniquely decodable codes. Kraft inequality. Shannon-Fano code. Huffman code.

In the previous chapter we have shown that entropy satisfies many intuitive identities. Now we will work on the links between entropy and coding. First we need to introduce some basic concepts in coding. The link with entropy will arise when we seek for optimal codes.

Definition 3.1 (injection). *Function f is called an* injection *if $x \neq y$ implies $f(x) \neq f(y)$.*

In coding theory we consider injections that map elements of a countable set $\mathbb{X}$ into strings over a countable set $\mathbb{Y}$. The set of these strings is denoted as $\mathbb{Y}^+ = \bigcup_{n=1}^{\infty} \mathbb{Y}^n$. Sometimes we also consider set $\mathbb{Y}^* = \{\lambda\} \cup \mathbb{Y}^+$ where λ is the empty string. Sets $\mathbb{X}$ and $\mathbb{Y}$ are called alphabets.

Definition 3.2 (code). *Any injection $B : \mathbb{X} \to \mathbb{Y}^*$ will be called a* code.

In this chapter we consider mostly binary codes, i.e., codes for which $\mathbb{Y} = \{0, 1\}^*$. On the other hand, the alphabet $\mathbb{X}$ may consist of letters, digits or other symbols.

Example 3.1. An example of a code:

symbol x:	code word $B(x)$:
a	0
b	1
c	10
d	11

The original purpose of coding is to transmit some representations of strings written with symbols from an alphabet $\mathbb{X}$ through a communication channel which passes only strings written with symbols from a smaller alphabet $\mathbb{Y}$. Thus the idea of a particularly good coding is that we should be able to reconstruct coded symbols from the concatenation of their codes. Formally speaking, the following property is desired.

Definition 3.3 (uniquely decodable code). *Code $B : \mathbb{X} \to \mathbb{Y}^*$ is called* uniquely decodable *if the code extension*

$$B^* : \mathbb{X}^* \ni (x_1, ..., x_n) \mapsto B(x_1)...B(x_n) \in \mathbb{Y}^*$$

is also an injection.

Example 3.2. The code given in Example 3.1 is not uniquely decodable because $B(ba) = 10 = B(c)$.

Example 3.3. However, this code is uniquely decodable:

symbol x:	code word $B(x)$:
a	0c
b	1c
c	10c
d	11c

The above code is a special case of a more general construction called a comma-separated code.

Definition 3.4 (comma-separated code). *Let* $c \notin \mathbb{Y}$*. Code* $B : \mathbb{X} \to (\mathbb{Y} \cup \{c\})^*$ *is called* comma-separated *if for each* $x \in \mathbb{X}$ *there exists a string* $w \in \mathbb{Y}^*$ *such that* $B(x) = wc$*. Symbol* c *is called the comma.*

Theorem 3.1. *Each comma-separated code is uniquely decodable.*

Proof. For a comma-separated code B, let us decompose $B(x) = \phi(x)c$. We first observe that $B(x_1)...B(x_n) = B(y_1)...B(y_m)$ holds only if $n = m$ (the same number of c's on both sides of equality) and $\phi(x_i) = \phi(y_i)$ for $i = 1, ..., n$. Next, we observe that function ϕ is a code. Hence string $B(x_1)...B(x_n)$ may be only the image of $(x_1, ..., x_n)$ under the mapping B^*. This means that code B is uniquely decodable.

Another recipe for producing a uniquely decodable code is to restrict the length of code words.

Definition 3.5 (fixed-length code). *Let* n *be a fixed natural number. Code* $B : \mathbb{X} \to \mathbb{Y}^n$ *is called a* fixed-length code.

Example 3.4. An example of a fixed-length code:

symbol x:	code word $B(x)$:
a	00
b	01
c	10
d	11

Theorem 3.2. *Each fixed-length code is uniquely decodable.*

Proof. Consider a fixed-length code B. We observe that $B(x_1)...B(x_n) = B(y_1)...B(y_m)$ holds only if $n = m$ (the same length of strings on both sides of equality) and $B(x_i) = B(y_i)$ for $i = 1, ..., n$. Because B is an injection, string $B(x_1)...B(x_n)$ may be only the image of $(x_1, ..., x_n)$ under the mapping B^*. Hence, code B is uniquely decodable.

In the second turn we may ask what is the shortest code to encode a given set of symbols, where the symbols appear with given probabilities. Let $|w|$ denote the length of a string $w \in \mathbb{Y}^*$, measured in the number in symbols. For a random variable $X : \Omega \to \mathbb{X}$, we will be interested in the expected code length

$$\mathbf{E}\,\left|B(X)\right| = \sum_{x \in \mathbb{X}} P(X = x)\,\left|B(x)\right|.$$

Example 3.5. Consider the following distribution and a code:

symbol x:	$P(X = x)$:	code word $B(x)$:
a	1/2	0C
b	1/6	1C
c	1/6	10C
d	1/6	11C

We have $\mathbf{E}\,|B(X)| = 2 \cdot \frac{1}{2} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} = 2\frac{1}{3}$.

Specifically, we are interested in codes that minimize the expected code length for a given probability distribution. In this regard, both comma-separated codes and fixed-length codes have advantages and drawbacks. If certain symbols appear more often than others then comma-separated codes allow to code them as shorter strings and thus to spare space. On the other hand, if all symbols are equiprobable then a fixed-length code without a comma occupies less space than the same code with a comma.

In general, there arises a lower bound for the expected code length which holds for any uniquely decodable code. Namely, $\mathbf{E}\,|B(X)|$ cannot be less than a multiplicity of entropy $H(X)$. This link is particularly easy for binary codes. The first step is to observe the following inequality.

Theorem 3.3 (Kraft inequality). *For any uniquely decodable code $B : \mathbb{X} \to \{0,1\}^*$ we have inequality*

$$\sum_{x \in \mathbb{X}} 2^{-|B(x)|} \le 1. \tag{3.1}$$

Proof. (By Brockway McMillan.) Consider an arbitrary L. Let $a(m,n,L)$ denote the number of sequences $(x_1, ..., x_n)$ such that $|B(x_i)| \le L$ and the length of $B^*(x_1, ..., x_n)$ equals m. We have

$$\left(\sum_{x:|B(x)| \le L} 2^{-|B(x)|} \right)^n = \sum_{m=1}^{nL} a\big(m,n,L\big) \cdot 2^{-m}.$$

Because the code is uniquely decodable, we have $a(m,n,L) \le 2^m$. Therefore

$$\sum_{x:|B(x)| \le L} 2^{-|B(x)|} \le \big(nL\big)^{1/n} \xrightarrow{n \to \infty} 1.$$

Letting $L \to \infty$, we obtain (3.1).

Hence we obtain the main theorem, which links coding and entropy.

Theorem 3.4 (source coding inequality). *For any uniquely decodable code $B : \mathbb{X} \to \{0,1\}^*$, the expected length of the code satisfies inequality*

$$\boldsymbol{E}\left|B(X)\right| \geq H(X), \tag{3.2}$$

where $H(X)$ is the entropy of X.

Proof. Introduce probability distributions $p(x) = P(X = x)$ and

$$r(x) = \frac{2^{-|B(x)|}}{\sum_{y\in\mathbb{X}} 2^{-|B(y)|}}.$$

We have

$$\begin{aligned}
\mathbf{E}\left|B(X)\right| - H(X) &= -\sum_{x\in\mathbb{X}} p(x)\log\Big(r(x)\sum_{y\in\mathbb{X}} 2^{-|B(y)|}\Big) + \sum_{x:p(x)>0} p(x)\log p(x) \\
&= \sum_{x:p(x)>0} p(x)\log\frac{p(x)}{r(x)} - \log\Big(\sum_{x\in\mathbb{X}} 2^{-|B(x)|}\Big) \\
&= D(p||r) - \log\Big(\sum_{x\in\mathbb{X}} 2^{-|B(x)|}\Big).
\end{aligned}$$

That difference is nonnegative by nonnegativity of Kullback-Leibler divergence, Theorem 2.4, and Kraft inequality, Theorem 3.3.

One can ask whether there exist codes for which the source coding inequality is close to equality. In fact, there exists a simple class of codes which is sufficient to encode a given set of symbols optimally. There are two mirror-like definitions.

Definition 3.6 (prefix-free code). *A code B is called* prefix-free *if no code word $B(x)$ is a prefix of another code word $B(y)$, i.e., it is not true that $B(y) = B(x)u$ for $x \neq y$ and a string $u \in \mathbb{Y}^*$.*

Definition 3.7 (suffix-free code). *A code B is called* suffix-free *if no code word $B(x)$ is a suffix of another code word $B(y)$, i.e., it is not true that $B(y) = uB(x)$ for $x \neq y$ and a string $u \in \mathbb{Y}^*$.*

Example 3.6. Codes in Examples 3.3 and 3.4 are prefix-free. Moreover, the code in Example 3.4 is also suffix-free.

Example 3.7. A code which is prefix-free but not suffix-free:

symbol x:	code word $B(x)$:
a	10
b	0
c	110
d	111

Example 3.8. A code which is suffix-free but not prefix-free:

symbol x:	code word $B(x)$:
a	01
b	0
c	011
d	111

Theorem 3.5. *Any prefix-free or suffix-free code is uniquely decodable.*

Proof. Without loss of generality we shall restrict ourselves to prefix-free codes. The proof for suffix-free codes is mirror-like. Let B be a prefix-free code and assume that $B(x_1)...B(x_n) = B(y_1)...B(y_m)$. By the prefix-free property the initial segments $B(x_1)$ and $B(y_1)$ must match exactly and $x_1 = y_1$. The analogous argument applied by induction yields $x_i = y_i$ for $i = 2, .., n$ and $n = m$. Thus code B is uniquely decodable.

For prefix-free codes there exists a theorem converse to the Kraft inequality.

Theorem 3.6. *If function $l : \mathbb{X} \to \mathbb{N}$ satisfies inequality*

$$\sum_{x \in \mathbb{X}} 2^{-l(x)} \le 1 \tag{3.3}$$

then we may construct a prefix-free code $B : \mathbb{X} \to \{0,1\}^*$ such that $|B(x)| = l(x)$.

Proof. (By Nicholas J. Pippenger.) Let u be the k-th element of set $\{0,1\}^l$ enumerated in the lexicographic order. We define section $s(u) := [k2^{-l}, (k+1)2^{-l})$ as the set of all real numbers whose binary expansions begin with string $0.u$. We observe that code B is prefix-free if and only if sections $s(B(x))$ and $s(B(y))$ are disjoint for $x \ne y$.

Because the code domain $\mathbb{X}$ is countable, we may assume without loss of generality that $\mathbb{X} = \{1, 2, ..., n\}$ or $\mathbb{X} = \mathbb{N}$. Then we define B by iteration as follows. First, we denote sets of sections $s(B(y))$ excluded before the x-th iteration as $N(1) := \emptyset$ and $N(x) := \bigcup_{y=1}^{x-1} s(B(y))$ for $x > 1$. Next, we define $B(x) := u$, where u is the *first* element of set $\{0,1\}^{l(x)}$ in the lexicographic order such that sets $s(u)$ and $N(x)$ are disjoint. It is obvious that B defined in this way is prefix-free and satisfies $|B(x)| = l(x)$, as long as strings u with the requested property exist.

Now we will show that strings u with the requested property exist if inequality (3.3) is satisfied. The proof of existence rests on this fact, which can be shown easily by induction: Set $[0,1) \setminus N(x)$ can be represented as a sum of finitely many sections $[k2^{-l}, (k+1)2^{-l})$ of *different* l, which appear in $[0,1)$ in order of decreasing l. Let 2^{-m} be the length of the largest available of these sections. By the mentioned fact, we have

$$2^{-m+1} > 1 - \sum_{y=1}^{x-1} 2^{-l(y)} \ge 2^{-m}. \tag{3.4}$$

The requested string u exists if and only if $2^{-l(x)} \leq 2^{-m}$. In view of (3.4), the latter condition holds if and only if

$$1 - \sum_{y=1}^{x-1} 2^{-l(y)} \geq 2^{-l(x)}.$$

But this condition is satisfied by (3.3).

We observe that Kraft inequality (3.1) may be satisfied with equality.

Definition 3.8 (complete code). *A code $B : \mathbb{X} \to \{0,1\}^*$ is called* complete *if*

$$\sum_{x \in \mathbb{X}} 2^{-|B(x)|} = 1.$$

Example 3.9. A code which is prefix-free, suffix-free, and complete.

symbol x:	code word $B(x)$:
a	00
b	01
c	10
d	11

Example 3.10. Another code which is prefix-free, suffix-free, and complete.

symbol x:	code word $B(x)$:
a	01
b	000
c	100
d	110
e	111
f	0010
g	0011
h	1010
i	1011

Source: Gillman and Rivest (1995).

Now we return to the question whether equality in the source coding inequality may be approximately achieved.

Definition 3.9 (Shannon-Fano code). *A prefix-free code $B : \mathbb{X} \to \{0,1\}^*$ is called a* Shannon-Fano code *if*

$$|B(x)| = \lceil -\log P(X = x) \rceil .$$

Theorem 3.7. *Shannon-Fano codes exist for any distribution and satisfy*

$$H(X) \leq \boldsymbol{E} \left| B(X) \right| \leq H(X) + 1. \tag{3.5}$$

Proof. We have

$$\sum_{x\in\mathbb{X}} 2^{-\lceil -\log P(X=x)\rceil} \le \sum_{x\in\mathbb{X}} 2^{\log P(X=x)} \le 1.$$

Hence Shannon-Fano codes exist by Theorem 3.6. Inequality (3.5) follows by

$$-\log P\big(X=x\big) \le \big|B\big(x\big)\big| \le -\log P\big(X=x\big)+1.$$

In spite of inequality (3.5), Shannon-Fano code is not necessarily the shortest possible code.

Example 3.11. Consider the following distribution and codes:

symbol x:	$P(X=x)$:	code word $B(x)$:	code word $C(x)$
a	$1-2^{-5}$	0	0
b	2^{-6}	100000	10
c	2^{-6}	100001	11

Code B is a Shannon-Fano code, whereas code C is another code. We have $H(X)=0.231...$, $\mathbf{E}\,|B(X)|=1.15625$, and $\mathbf{E}\,|C(X)|=1.03125$. For no symbol code C is worse than code B, whereas for less probable symbols code C is much better.

A code that minimizes the expected length $\mathbf{E}\,|B(X)|$ is known under the name of Huffman code. To introduce this code we need first to uncover a relationship between prefix-free codes and binary trees.

Definition 3.10 (binary tree). *A* binary tree *is a directed acyclic connected graph where each node has at most two children nodes (left and/or right one) and at most one parent node. The node which has no parents is called the* root node. *The nodes which have no children are called* leaf nodes. *We assume that links to the left children are labeled with* 0*'s whereas links to the right children are labeled with* 1*'s. Moreover, some nodes may be labeled with some symbols as well.*

Definition 3.11 (path). *We say that a binary tree contains a* path $w\in\{0,1\}^*$ *if there is a sequence of links starting from the root node and labeled with the consecutive symbols of* w*. We say that the path is ended with symbol* $a\in\mathbb{X}$ *if the last link of the sequence ends in a node labeled with symbol* a*.*

Definition 3.12 (code tree). *The* code tree *for a code* $B:\mathbb{X}\to\{0,1\}^*$ *is a labeled binary tree which contains a path* w *if and only if* $B(a)=w$ *for some* $a\in\mathbb{X}$*, and exactly in that case we require that path* w *is ended with symbol* a*.*

Example 3.12. Consider codes:

symbol x:	code word $B(x)$:	code word $C(x)$:
a	0	00
b	1	01
c	10	10
d	11	110
e	00	111

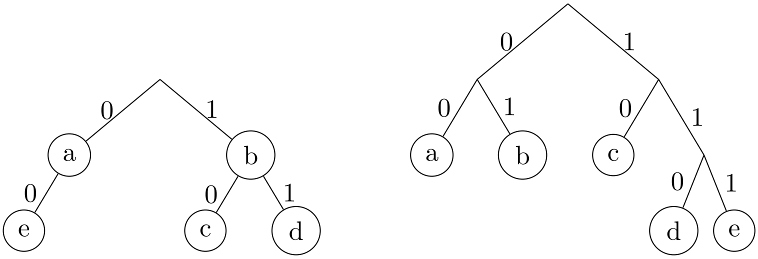

Fig. 4. The code trees for the codes from Example 3.12.

The code trees for these codes are depicted in Figure 4.

It is easy to observe the following fact.

Theorem 3.8. *There is a one-to-one correspondence between the binary codes and code trees. Moreover, a code is prefix-free if and only if the leaf nodes are the only nodes labeled.*

In the next step, we will add some weights to the code trees, which stem from the distribution of symbols.

Definition 3.13 (weighted code tree). *The* weighted code tree *for a prefix code $B : \mathbb{X} \to \{0,1\}^*$ and a probability distribution $p : \mathbb{X} \to [0,1]$ is the code tree for code B where the nodes are enhanced with the following weights: (1) for a leaf node with symbol a, we add weight $p(a)$, (2) to other (internal) nodes, we ascribe weights equal to the sum of weights of their children.*

Example 3.13. Consider this distribution and the code C from Example 3.12:

symbol x:	$p(x)$:	code word $C(x)$:
a	0.2	00
b	0.3	01
c	0.1	10
d	0.2	110
e	0.2	111

The weighted code tree is depicted in Figure 5.

Now we can describe the Huffman code.

Definition 3.14 (Huffman code). *The* Huffman code *for a probability distribution $p : \mathbb{X} \to [0,1]$ is a code whose weighted code tree is constructed by the following algorithm:*

1. *Create a leaf node for each symbol and add them to a list.*
2. *While there is more than one node in the list:*
   (a) *Remove two nodes of the lowest weight from the list.*

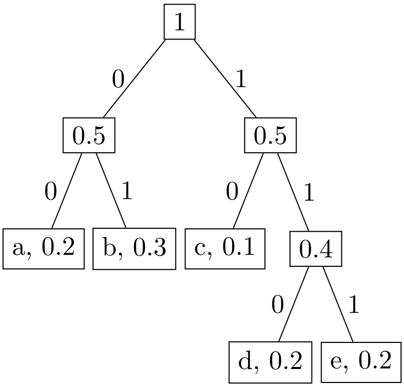

Fig. 5. The weighted code tree for Example 3.13.

(b) Create a new internal node with these two nodes as children and with weight equal to the sum of the two nodes' weights.
(c) Add the new node to the list.

3. The remaining node is the root node and the tree is complete.

Example 3.14. The Huffman code for the distribution from Example 3.13 is:

symbol x:	$p(x)$:	Huffman code $B(x)$:
a	0.2	00
b	0.3	10
c	0.1	110
d	0.2	111
e	0.2	01

The corresponding Huffman code tree is depicted in Figure 6.

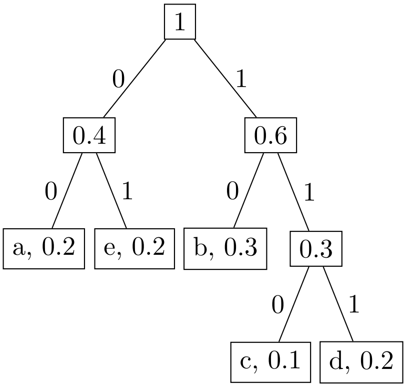

Fig. 6. The Huffman code tree for Example 3.14.

It can be proved that no code fares better than the Huffman code if the probability distribution is known.

Theorem 3.9. *For any probability distribution $p(x) = P(X = x)$, the Huffman code achieves the minimum expected length $\boldsymbol{E}\,|B(X)|$.*

Proof. A code B will be called optimal if $\mathbf{E}\,|B(X)|$ achieves the minimum for a given distribution $p(x) = P(X = x)$. We will use the this fact:

Consider the two symbols x and y with the smallest probabilities. Then there is an optimal code tree C such that these two symbols are sibling leaves in the lowest level of C's code tree.

To prove this fact, observe the following. Every internal node in a code tree for an optimal code must have two children. (Surely, if some internal node had only a single child, we might discard this node.) Then let B be an optimal code and let symbols a and b be two siblings at the maximal depth of B's code tree. Assume without loss of generality that $p(x) \le p(y)$ and $p(a) \le p(b)$. We have $p(x) \le p(a)$, $p(y) \le p(b)$, $|B(a)| \ge |B(x)|$, and $|B(b)| \ge |B(y)|$. Now let C's code tree differ from the B's code tree by switching $a \leftrightarrow x$ and $b \leftrightarrow y$. Then we obtain

$$\begin{aligned}
&\mathbf{E}\,\big|C(X)\big| - \mathbf{E}\,\big|B(X)\big| \\
&\qquad = -p(x)\,|B(x)| - p(a)\,|B(a)| + p(a)\,|B(x)| + p(x)\,|B(a)| \\
&\qquad\quad - p(y)\,|B(y)| - p(b)\,|B(b)| + p(b)\,|B(y)| + p(y)\,|B(b)| \\
&\qquad = (p(a) - p(x))(|B(x)| - |B(a)|) \\
&\qquad\quad + (p(b) - p(y))(|B(y)| - |B(b)|) \le 0.
\end{aligned}$$

Hence code C is also optimal.

Now we will proceed by induction on the number of symbols in the alphabet $\mathbb{X}$. If $\mathbb{X}$ contains only two symbols, then Huffman code is optimal. In the second step, we assume that Huffman code is optimal for $n-1$ symbols and we prove its optimality for n symbols. Let C be an optimal code for n symbols. Without loss of generality we may assume that symbols x and y having the smallest probabilities occupy two sibling leaves in the lowest level of C's code tree. Then from the weighted code tree of C we construct a code C' for $n-1$ symbols by removing nodes with symbols x and y and ascribing a symbol z to its parent node. Hence we have

$$\mathbf{E}\,|C'(X')| = \mathbf{E}\,|C(X)| - p(x) - p(y),$$

where variable $X' = z$ if $X \in \{x, y\}$ and $X' = X$ otherwise. On the other hand, let B' be the Huffman code for X' and let B be the code constructed from B' by adding leaves with symbols x and y to the node with symbol z. By construction, code B is the Huffman code for X. We have

$$\mathbf{E}\,\big|B'(X')\big| = \mathbf{E}\,\big|B(X)\big| - p(x) - p(y).$$

Because $\mathbf{E}\,|B'(X')| \le \mathbf{E}\,|C'(X')|$ by optimality of Huffman code B', we obtain $\mathbf{E}\,|B(X)| \le \mathbf{E}\,|C(X)|$. Hence Huffman code B is also optimal.

Exercises

1. *(Prefix-free codes)* Which of the following codes are prefix-free?
   (a) $\{0, 01, 1\}$,
   (b) $\{01, 101, 11\}$,
   (c) $\{0, 10, 110\}$,
   (d) $\{00, 010, 110, 11\}$.
2. We say that a prefix-free code is maximal if no superset of the code words is prefix-free. Show that a maximal prefix-free code $B : \mathbb{X} \to \{0,1\}^*$ is complete if $\mathbb{X}$ is finite, whereas it need not be so if $\mathbb{X}$ is infinite.
3. *(D-ary codes)* Show that for any uniquely decodable code $B : \mathbb{X} \to \{0, 1, ..., D-1\}^*$,

   $$\sum_{x \in \mathbb{X}} D^{-|B(x)|} \leq 1.$$

   Using this inequality, show further that $\mathbf{E}\,|B(X)| \geq H(X)/\log D$.
4. Let a uniquely decodable code $B : \mathbb{X} \to \{0, 1, ..., D-1\}^*$ satisfy

   $$\sum_{x \in \mathbb{X}} D^{-|B(x)|} < 1.$$

   Show that there exist arbitrarily long sequences in $\{0, 1, ..., D-1\}^*$ which cannot be decoded into sequences of codewords.
5. *(Huffman codes)* Find the Huffman codes for the following distributions:

   (a)

symbol x:	$P(X = x)$:
a	1/12
b	1/6
c	1/4
d	1/3
e	1/6

   (b)

symbol x:	$P(X = x)$:
a	1/11
b	3/11
c	2/11
d	1/11
e	3/11
f	1/11

   (c)

symbol x:	$P(X = x)$:
a	1/7
b	1/2
c	1/6
d	3/42
e	5/42

6. Which of these codes cannot be Huffman codes for any probability distribution?
   (a) $\{001, 01, 1\}$,

(b) $\{01, 10, 11\}$,
(c) $\{0, 10, 11\}$,
(d) $\{00, 010, 10, 11\}$.

7. Consider a random variable X with distribution

symbol x:	$P(X = x)$:
a	1/3
b	1/3
c	4/15
d	1/15

Show that there are two Huffman codes for this distribution: one has lengths $(1, 2, 3, 3)$ and the other has lengths $(2, 2, 2, 2)$. Use this result to demonstrate that the length of some Huffman code can be greater for some symbol than the length of the Shannon-Fano code.

8. We say that X has a dyadic distribution if for each x there exists an integer k such that $P(x) = 2^{-k}$. Show that the length of the Huffman code for a dyadic distribution is the same as the length of the Shannon-Fano code and satisfies $\mathbf{E}\ |B(X)| = H(X)$.

9. *(Elias omega code)* Binary expansion is an example of a code for natural numbers which is not prefix-free. We can correct this code to make it prefix-free by preceding the binary expansion with a recursive representation of its length. In this way we obtain the Elias omega code.
The algorithm for the Elias omega encoding is as follows:
(a) Put 0 at the end of the code.
(b) If the coded number is 1, stop. Otherwise, write the binary representation of the coded number before the code.
(c) Repeat the previous step with the coded number equal to the number of digits written in the previous step minus 1.
In this way we obtain the following correspondence:

number n:	code word $B(n)$:
1	0
2	10 0
3	11 0
4	10 100 0
5	10 101 0
6	10 110 0
7	10 111 0
8	11 1000 0

Find the algorithm for decoding the Elias omega code.

4

Stationary processes

Stationary processes. Markov models. Hidden Markov models. Entropy rate. Entropy rate as the limiting compression rate. Excess entropy.

An infinite collection of random variables on a probability space is called a stochastic process. In this chapter we will study stationary processes, whose probability distributions are invariant under shifting. For the sake of ergodic processes that will be discussed in the next chapter, it pays off to start with a more abstract setup and to begin with probability measures that are invariant under an arbitrary transformation.

Definition 4.1 (invariant measure). *Consider a probability space $(\Omega, \mathcal{J}, P)$ and an invertible operation $T : \Omega \to \Omega$ such that $T^{-1}A \in \mathcal{J}$ for $A \in \mathcal{J}$. Measure P is called T-*invariant *and T is called P-*preserving *if*

$$P\big(T^{-1}A\big) = P\big(A\big)$$

for any event $A \in \mathcal{J}$.

Definition 4.2 (dynamical system). *A dynamical system $(\Omega, \mathcal{J}, P, T)$ is a quadruple that consists of a probability space $(\Omega, \mathcal{J}, P)$ and a P-preserving operation T.*

To check whether a given measure P is T-invariant, we often do not need to check whether $P(T^{-1}A) = P(A)$ for all events $A \in \mathcal{J}$. Usually it suffices to check this condition only for $A \in \mathcal{A}$ where $\mathcal{A}$ is some reasonable subset of $\mathcal{J}$.

Definition 4.3 (generated σ-field). *We say that subset $\mathcal{J} \subset 2^\Omega$ is a σ-*field generated *by subset $\mathcal{A} \subset 2^\Omega$ if it is the intersection of all σ-fields that contain $\mathcal{A}$. We denote this fact as $\mathcal{J} = \sigma(\mathcal{A})$.*

Definition 4.4 (π-system). *A subset $\mathcal{A} \subset 2^\Omega$ is called a π-*system *if for all $A, B \in \mathcal{A}$ we have $A \cap B \in \mathcal{A}$.*

Theorem 4.1 (π-λ theorem). *Let $\mathcal{A}$ be a π-system and let P and P' be two probability measures on $\mathcal{J} = \sigma(\mathcal{A})$. If $P'(A) = P(A)$ for all $A \in \mathcal{A}$ then $P' = P$.*

The proof of this theorem can be found in Billingsley (1979, Section 2).

Now we can see that $P'(A) := P(T^{-1}A)$ is also a probability measure. Thus if we have $P(T^{-1}A) = P(A)$ for all $A \in \mathcal{A}$, where $\mathcal{A}$ is a π-system, then $P(T^{-1}A) = P(A)$ holds also for all $A \in \sigma(\mathcal{A})$.

Example 4.1 (rotation). Let $\Omega = (0,1]$ and let $\mathcal{J}$ be generated by sections $(a,b]$, $a,b \in (0,1]$. The intersection of two sections is a section so the set of all sections is a π-system that generates $\mathcal{J}$. Let P be the Lebesgue measure on $\mathcal{J}$, defined by $P((a,b]) = b-a$, and define $T(\omega) = (\omega + r) \mod 1$ for an $r \in (0,1]$. We have $P(T^{-1}(a,b]) = P((a,b])$. Hence P is T-invariant.

A stationary process is a concept tightly related to an invariant measure.

Definition 4.5 (stationary process). *A stochastic process $(X_i)_{i=-\infty}^{\infty}$, where $X_i : \Omega \to \mathbb{X}$ are discrete random variables, is called* stationary *if there exists a distribution of blocks $p : \mathbb{X}^* \to [0,1]$ such that*

$$P\big(X_{i+1} = x_1, ..., X_{i+n} = x_n\big) = p(x_1...x_n) \tag{4.1}$$

for any $i \in \mathbb{Z}$ and $n \in \mathbb{N}$.

Example 4.2 (IID process). If variables X_i are independent and have identical distribution $P(X_i = x) = p(x)$ then $(X_i)_{i=-\infty}^{\infty}$ is stationary.

It is easy to see that, if we have an invariant measure, we can produce a stationary process given an arbitrary random variable.

Example 4.3. Let measure P be T-invariant and let $X_0 : \Omega \to \mathbb{X}$ be a random variable on $(\Omega, \mathcal{J}, P)$. Define random variables $X_i(\omega) = X_0(T^i\omega)$. For

$$\begin{aligned} A &= \big(X_{i+1} = x_1, ..., X_{i+n} = x_n\big) \\ &= \big\{\omega : X_0(T^{i+1}\omega) = x_1, ..., X_0(T^{i+n}\omega) = x_n\big\}, \end{aligned}$$

we have

$$\begin{aligned} T^{-1}A &= \big\{T^{-1}\omega : X_0(T^{i+1}\omega) = x_1, ..., X_0(T^{i+n}\omega) = x_n\big\} \\ &= \big\{\omega : X_0(T^{i+2}\omega) = x_1, ..., X_0(T^{i+n+1}\omega) = x_n\big\} \\ &= \big(X_{i+2} = x_1, ..., X_{i+n+1} = x_n\big). \end{aligned}$$

Hence, because $P(T^{-1}A) = P(A)$, process $(X_i)_{i=-\infty}^{\infty}$ is stationary.

An interesting question is whether given a distribution of blocks, we can construct a probability space with the requested stationary process and an invariant measure. The answer is positive if the distribution of blocks satisfies a consistency criterion, cf. Billingsley (1979, Section 36).

Theorem 4.2 (process theorem). *Let distribution of blocks $p : \mathbb{X}^* \to [0,1]$ satisfy conditions*

$$\sum_{x\in\mathbb{X}} p(xw) = p(w) = \sum_{x\in\mathbb{X}} p(wx) \tag{4.2}$$

and $p(\lambda) = 1$, where λ is the empty word. Let event space

$$\Omega = \big\{\omega = (\omega_i)_{i=-\infty}^{\infty} : \omega_i \in \mathbb{X}\big\}$$

consist of infinite sequences and introduce random variables $X_i(\omega) = \omega_i$. Let $\mathcal{J}$ be the σ-field generated by all cylinder sets $(X_i = s) = \{\omega \in \Omega : X_i(\omega) = s\}$. Then there exists a unique probability measure P on $\mathcal{J}$ that satisfies (4.1).

If there are some other random variables considered in an application, then we have to consider a larger probability space, whose construction proceeds analogously. With appropriate modifications, Theorem 4.2 may be generalized to nonstationary, real-valued, and real-time processes.

Theorem 4.3. *Let $(\Omega, \mathcal{J}, P)$ be the probability space constructed in Theorem 4.2. Measure P is T-invariant for the operation*

$$\left(T\omega\right)_i := \omega_{i+1},$$

called shift. Moreover, we have $X_i(\omega) = X_0(T^i\omega)$.

Proof. By the π-λ theorem it suffices to prove $P(T^{-1}A) = P(A)$ for $A = (X_{i+1} = x_1, ..., X_{i+n} = x_n)$. But $X_i(\omega) = X_0(T^i\omega)$. Hence $T^{-1}A = (X_{i+2} = x_1, ..., X_{i+n+1} = x_n)$, as shown in Example 4.3. In consequence, we obtain $P(T^{-1}A) = P(A)$ by stationarity.

Definition 4.6. *The triple $(\Omega, \mathcal{J}, P)$ constructed in Theorem 4.2 and the quadruple $(\Omega, \mathcal{J}, P, T)$ constructed in Theorem 4.3 will be called the* probability space *and the* dynamical system generated *by a stationary process $(X_i)_{i=-\infty}^{\infty}$ (with a given block distribution $p : \mathbb{X}^* \to [0, 1]$).*

Now, once we have the process theorem, let us investigate a few further examples of stationary processes. The first example are Markov processes.

Theorem 4.4. *A Markov chain $(X_i)_{i=-\infty}^{\infty}$ is stationary if and only if it has marginal distribution $P(X_i = k) = \pi_k$ and transition probabilities $P(X_{i+1} = l|X_i = k) = p_{kl}$ which satisfy*

$$\pi_l = \sum_k \pi_k p_{kl}. \tag{4.3}$$

Proof. If $(X_i)_{i=-\infty}^{\infty}$ is stationary then the marginal distribution $P(X_i = k)$ and transition probabilities $P(X_{i+1} = l|X_i = k)$ may not depend on i. We also have

$$P(X_{i+1} = k_1, ..., X_{i+n} = k_n) = p(k_1...k_n) := \pi_{k_1} p_{k_1 k_2} ... p_{k_{n-1} k_n}$$

Function $p(k_1...k_n)$ satisfies (4.2). Hence we obtain (4.3). On the other hand, if $P(X_i = k) = \pi_k$ and $P(X_{i+1} = l|X_i = k) = p_{kl}$ hold with (4.3) then function $p(k_1...k_n)$ satisfies (4.2) and the process is stationary.

If the process variables assume a finite number of values, relationship (4.3) can be written in the vector notation as

$$\begin{pmatrix} \pi_1 \; \pi_2 \dots \pi_n \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} = \begin{pmatrix} \pi_1 \; \pi_2 \dots \pi_n \end{pmatrix}.$$

Matrix (p_{kl}) is called the transition matrix. For a given transition matrix the stationary distribution may not exist or there may be many stationary distributions.

Example 4.4. Let variables X_i assume values in natural numbers and let $P(X_{i+1} = k + 1|X_i = k) = 1$. Then the process $(X_i)_{i=1}^{\infty}$ is not stationary. Indeed, assume that there is a stationary distribution $P(X_i = k) = \pi_k$. Then by (4.3) we obtain $\pi_{k+1} = \pi_k$ for any k. Such distribution does not exist if there are infinitely many k.

Example 4.5. For the transition matrix

$$\begin{pmatrix} p_{11} \, p_{12} \\ p_{21} \, p_{22} \end{pmatrix} = \begin{pmatrix} 1 \, 0 \\ 0 \, 1 \end{pmatrix}$$

any vector of form

$$\begin{pmatrix} \pi_1 \, \pi_2 \end{pmatrix} = \begin{pmatrix} a \; 1 - a \end{pmatrix}, \qquad\qquad a \in [0, 1],$$

is the solution of equation (4.3).

Once we have a stationary process we can construct another stationary process as a function of that process.

Theorem 4.5. *If a process $(X_i)_{i=-\infty}^{\infty}$ is stationary then the process $(Y_i)_{i=-\infty}^{\infty}$ where $Y_k = f(T^k(X_i)_{i=-\infty}^{\infty})$ is also stationary.*

In particular, we have the following case:

Definition 4.7. *Process $(Y_i)_{i=-\infty}^{\infty}$ is called a* hidden Markov chain *if $Y_i = f(X_i)$ where $(X_i)_{i=-\infty}^{\infty}$ is a Markov chain. (If variables X_i assume finitely many values then $(Y_i)_{i=-\infty}^{\infty}$ is also called a* finite-state source*.)*

Now let us consider information theoretic properties of stationary processes. We will be interested in the information carried by strings of consecutive variables, called blocks.

Definition 4.8 (block). Blocks *of variables are written as $X_k^l = (X_i)_{k \le i \le l}$.*

The entropy of a block drawn from a stationary process depends only on the block length.

Definition 4.9 (block entropy). *The entropy of the block of n variables drawn from a stationary process will be denoted as*

$$H(n) := H\big(X_1^n\big) = H\big(X_1, ..., X_n\big) = H\big(X_{i+1}, ..., X_{i+n}\big).$$

For convenience, we also put $H(0) = 0$.

Theorem 4.6. *Let Δ be the difference operator, $\Delta F(n) := F(n) - F(n-1)$. Block entropy satisfies*

$$\Delta H(n) = H\big(X_n|X_1^{n-1}\big),$$
$$\Delta^2 H(n) = -I\big(X_1; X_n|X_2^{n-1}\big),$$

where $H(X_1|X_1^0) := H(X_1)$ and $I(X_1; X_2|X_2^1) := I(X_1; X_2)$.

Remark: Hence, for any stationary process, block entropy $H(n)$ is nonnegative ($H(n) \geq 0$), nondecreasing ($\varDelta H(n) \geq 0$) and concave ($\varDelta^2 H(n) \leq 0$).

Proof. We have

$$\begin{aligned} H\big(X_n|X_1^{n-1}\big) &= H\big(X_1^n\big) - H(X_1^{n-1}) \\ &= H(n) - H(n-1) = \varDelta H(n), \\ -I\big(X_1; X_n|X_2^{n-1}\big) &= H\big(X_1^n\big) - H\big(X_1^{n-1}\big) - H\big(X_2^n\big) + H\big(X_2^{n-1}\big) \\ &= H(n) - 2H(n-1) + H(n-2) = \varDelta^2 H(n). \end{aligned}$$

In the following we will introduce a certain limiting quantity.

Definition 4.10 (entropy rate). *The* entropy rate *of a stationary process will be defined as*

$$h = \lim_{n\to\infty} \varDelta H(n) = H\big(1\big) + \sum_{n=2}^{\infty} \varDelta^2 H(n). \tag{4.4}$$

By the previous theorem, we have $0 \leq h \leq H(1)$.

Example 4.6. Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary Markov chain with marginal distribution $P(X_i = k) = \pi_k$ and transition probabilities $P(X_{i+1} = l|X_i = k) = p_{kl}$. We have $\varDelta H(n) = H(X_n|X_1^{n-1}) = H(X_n|X_{n-1})$, so

$$h = -\sum_{kl} \pi_k p_{kl} \log p_{kl}.$$

The name “entropy rate” is motivated by the following identity.

Theorem 4.7. *Entropy rate satisfies equality*

$$h = \lim_{n\to\infty} \frac{H(n)}{n}.$$

Proof. Difference $\varDelta H(\cdot)$ is nonincreasing. Hence block entropy $H(n) = H(m) + \sum_{k=m+1}^{n} \varDelta H(k)$ satisfies inequalities

$$H(m) + (n-m) \cdot \varDelta H(n) \leq H(n) \leq H(m) + (n-m) \cdot \varDelta H(m). \tag{4.5}$$

Putting $m = 0$ in the left inequality in (4.5), we obtain

$$\varDelta H(n) \leq H(n)/n \tag{4.6}$$

Putting $m = n-1$ in the right inequality in (4.5), we hence obtain $H(n) \leq H(n-1)+\varDelta H(n-1) \leq H(n-1)+H(n-1)/(n-1)$. Thus $H(n)/n \leq H(n-1)/(n-1)$. Because function $H(n)/n$ is nonincreasing, the limit $h' := \lim_{n\to\infty} H(n)/n$ exists. By (4.6), we have $h' \geq h$. Now we will prove the converse. Putting $n = 2m$ in the right inequality in (4.5) and dividing both sides by m we obtain $2h' \leq h'+h$ in the limit. Hence $h' \leq h$.

Entropy rate equals the minimal compression rate with which the data typical for a stationary process can be compressed.

Theorem 4.8. *For a stationary process* $(X_i)_{i=-\infty}^{\infty}$*, let* L_n *denote the minimal expected compression rate of a uniquely decodable code* $B_n : \mathbb{X}^n \to \{0,1\}^*$ *for the block of* n *variables. That is,*

$$L_n := \inf_{B_n} \frac{1}{n} \boldsymbol{E} \left| B_n\bigl(X_1, ..., X_n\bigr) \right| .$$

We claim that $\lim_{n\to\infty} L_n = h$*.*

Proof. Assume that B_n is the Shannon-Fano code for the block $(X_1, ..., X_n)$. Then $H(X_1^n) \le nL_n \le \mathbf{E}\,|B_n(X_1, ..., X_n)| \le H(X_1^n) + 1$. Hence the claim follows.

There is another interesting limiting quantity, which is called excess entropy.

Definition 4.11 (excess entropy). *Excess entropy of a stationary process is defined as*

$$E := \lim_{n\to\infty} E(n), \tag{4.7}$$

where

$$E(n) = I\bigl(X_{-n+1}^0; X_1^n\bigr)$$

is the mutual information between adjacent blocks of length n*.*

For any process, excess entropy E is a definite value from range $[0,\infty]$ because $E(\cdot)$ is nondecreasing.

The name “excess entropy” is motivated by the following theorem:

Theorem 4.9. *For the functions*

$$\begin{aligned} \bar{E}(n) &= H(n) - n\Delta H(n), \\ \tilde{E}(n) &= H(n) - nh \end{aligned}$$

we have inequalities

$$\begin{gathered} 0 \le \bar{E}(n) \le E(n) \le \bar{E}(2n) \le E, \\ E(n) \le \tilde{E}(n) \le E, \end{gathered}$$

whereas $\bar{E}(\cdot)$ *and* $\tilde{E}(\cdot)$ *are nondecreasing. Hence*

$$E = \lim_{n\to\infty} E(n) = \lim_{n\to\infty} \bar{E}(n) = \lim_{n\to\infty} \tilde{E}(n).$$

Proof. We have

$$\begin{aligned}
E(n) &= H\bigl(X_1^n\bigr) - H\bigl(X_1^n|X_{-n+1}^0\bigr) = \sum_{i=1}^n \bigl[H(X_i|X_1^{i-1}) - H(X_i|X_{-n+1}^{i-1})\bigr] \\
&= \sum_{i=1}^n \bigl[\Delta H(i) - \Delta H(i+n)\bigr] = -\sum_{i=1}^n \sum_{j=1}^n \Delta^2 H(i+j) \\
&= -\sum_{k=2}^n (k-1)\Delta^2 H(k) - \sum_{k=n+1}^{2n} (2n-k+1)\Delta^2 H(k), \\
\bar{E}(n) &= \sum_{i=1}^n [\Delta H(i) - \Delta H(n)] = -\sum_{i=1}^n \sum_{j=i+1}^n \Delta^2 H(j) \\
&= -\sum_{k=2}^n (k-1)\Delta^2 H(k).
\end{aligned}$$

Each of quantities $\bar{E}(n)$, $E(n)$, $\bar{E}(2n)$ is a sum of nonnegative terms $-\Delta^2 H(k)$ multiplied by nonnegative factors. The factors rise for the consecutive quantities $\bar{E}(n)$, $E(n)$, $\bar{E}(2n)$. Hence we have $0 \le \bar{E}(n) \le E(n) \le \bar{E}(2n)$. By a similar argument, $\bar{E}(n) \le \bar{E}(n+1)$, so $\bar{E}(\cdot)$ is nondecreasing. Hence there exists limit $\lim_{n\to\infty} \bar{E}(n)$ and it equals $\lim_{n\to\infty} E(n)$.

For $m > n$ define

$$\begin{aligned}
\tilde{E}(n;m) = H(n) - n\Delta H(m) &= \bar{E}(m) + n\bigl[\Delta H(n) - \Delta H(m)\bigr] \\
&= -\sum_{k=2}^n (k-1)\Delta^2 H(k) - \sum_{k=n+1}^{m} n\Delta^2 H(k). \qquad (4.8)
\end{aligned}$$

By (4.8) we have $\tilde{E}(n;m) \le \tilde{E}(n;m+1)$. Hence there exists limit $\lim_{m\to\infty} \tilde{E}(n;m)$ and it equals $\tilde{E}(n)$. Comparing the coefficients at $-\Delta^2 H(k)$, we infer that $E(n) \le \tilde{E}(n;m) \le \bar{E}(m)$ for $m \ge 2n$. By a similar argument, $\tilde{E}(n;m) \le \tilde{E}(n+1;m)$. For $m \to \infty$ these two sets of inequalities imply $E(n) \le \tilde{E}(n) \le E$ and $\tilde{E}(n) \le \tilde{E}(n+1)$. Hence $\lim_{n\to\infty} \bar{E}(n) = \lim_{n\to\infty} \tilde{E}(n)$ and $\tilde{E}(\cdot)$ is nondecreasing.

Exercises

1. *(Markov processes)* Find the stationary distributions (π_i) for transition matrices

$$(p_{ij}) = \begin{pmatrix} 1/10 & 9/10 \\ 1/12 & 11/12 \end{pmatrix},$$

$$(p_{ij}) = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix},$$

$$(p_{ij}) = \begin{pmatrix} 3/8 & 5/8 \\ 7/10 & 3/10 \end{pmatrix},$$

$$(p_{ij}) = \begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 2/3 & 0 \\ 1/6 & 1/2 & 1/3 \end{pmatrix}.$$

2. *(Subadditive functions)* A function f is called subadditive $f(n+m) \leq f(n) + f(m)$. Assuming $f(0) = 0$, show that if $\Delta^2 f(n) \leq 0$ then the function is subadditive.
3. Let sequence $(a_n)_{n\in\mathbb{N}}$ be nonnegative. Consider four conditions:
   (a) Sequence $(a_n)_{n\in\mathbb{N}}$ has decreasing increments if $(a_n - a_{n-1})_{n\in\mathbb{N}}$ is decreasing.
   (b) Sequence $(a_n)_{n\in\mathbb{N}}$ has decreasing nths if $(n^{-1}a_n)_{n\in\mathbb{N}}$ is decreasing.
   (c) Sequence $(a_n)_{n\in\mathbb{N}}$ is subadditive if $a_{n+m} \leq a_n + a_m$.
   (d) Sequence $(a_n)_{n\in\mathbb{N}}$ has descending nths if $\limsup_{n\to\infty} n^{-1}a_n = \inf_{n\in\mathbb{N}} n^{-1}a_n$.

   Show that (a) $\implies$ (b) $\implies$ (c) $\implies$ (d) but the converse is not true. Moreover if (a) holds then $(a_n)_{n\in\mathbb{N}}$ increases and $\lim_{n\to\infty}(a_n - a_{n-1}) = \lim_{n\to\infty} n^{-1}a_n$
4. *(Concave functions)* Assuming $f(0) = 0$, show that if $\Delta^2 f(n) \leq 0$ then $p_1 f(n) + p_2 f(m) \leq f(p_! n + p_2 m)$ for $p_i \geq 0$, $p_1 + p_2 = 1$.
5. *(Entropy rate)* Show that a stationary process is a sequence of independent identically distributed variables if and only if $h = H(1)$.
6. Let $(Y_i)_{i=-\infty}^{\infty}$ be a stationary process whose variables assume k distinct values. Show that $(Y_i)_{i=-\infty}^{\infty}$ is a sequence of independent uniformly distributed variables if and only if $h = \log k$.
7. For a stationary process show that

$$\lim_{n\to\infty} \frac{H\big(X_1^n | X_{-k}^0\big)}{n} = \lim_{k\to\infty} \frac{H\big(X_1^n | X_{-k}^0\big)}{n} = h.$$

8. *(Excess entropy)* Show that excess entropy is finite for a hidden Markov chain $(Y_i)_{i=-\infty}^{\infty}$ where $Y_i = f(X_i)$ and variables of the Markov chain X_i assume a finite number of values.
9. For a stationary process and $k \leq n/2$, show that

$$I\big(X_1^{k-1}; X_k^n\big) \leq I\big(X_1^k; X_{k+1}^n\big).$$

10. *(Randomly stopped sequences)* Let $(Y_i)_{i=-\infty}^{\infty}$ be a stationary process. Moreover, let S be a random variable assuming values in natural numbers, where events $(S = n)$ and $(Y_1^n = y_1^n)$ are independent. Show that $H(Y_1^S) \leq H(S) + H(Y_1^{\lceil \mathbf{E}\, S \rceil})$.
11. Let $(Y_i)_{i=-\infty}^{\infty}$ be a stationary process with entropy rate h. Moreover, let S be a random variable assuming values in natural numbers, where events $(S = n)$ and $(Y_{n+1}^m = y_{n+1}^m)$ are independent. Show that $H(Y_1^S) \geq H(S|(Y_i)_{i=1}^{\infty}) + h\mathbf{E}\, S$, where $H(S|(Y_i)_{i=1}^{\infty}) = \lim_{n\to\infty} H(S|Y_1^n)$.

5

Ergodic processes

Ergodic systems. Ergodic theorem. Ergodic Markov processes. Shannon-McMillan-Breiman theorem.

In this chapter, we will discuss the ergodic theorem, which generalizes the strong law of large numbers to the case of stationary ergodic processes. To begin with, let us note that the probability space generated by a stationary process supports many interesting events, whose probability is well defined. Some of them belong to the so called invariant algebra.

Definition 5.1 (invariant algebra). *Let $(\Omega, \mathcal{J}, P, T)$ be a dynamical system. The set of events which are invariant with respect to operation T,*

$$\mathcal{I} := \left\{ A \in \mathcal{J} : A = T^{-1}A \right\},$$

will be called the invariant algebra*.*

It may be insightful to substantiate this concept by a few examples.

Example 5.1. Let $(\Omega, \mathcal{J}, P, T)$ be the dynamical system generated by a stationary process $(X_i)_{i=-\infty}^{\infty}$, where $X_i : \Omega \to \{0, 1\}$. The operation T results in shifting variables X_i, i.e., $X_i(\omega) = X_0(T^i\omega)$ for all $i \in \mathbb{Z}$. Thus the following events belong to the invariant algebra $\mathcal{I}$:

$$\left\{ \omega : X_i(\omega) = 1 \text{ for all } i \in \mathbb{Z} \right\} = \bigcap_{i=-\infty}^{\infty} (X_i = 1), \tag{5.1}$$

$$\left\{ \omega : X_i(\omega) = 1 \text{ for infinitely many } i \geq 1 \right\} = \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} \left(X_j = 1 \right), \tag{5.2}$$

$$\left\{ \omega : \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) = a \right\} = \bigcap_{p=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \left(\left| \frac{1}{n} \sum_{k=1}^{n} X_k - a \right| \leq \frac{1}{p} \right). \tag{5.3}$$

If the process $(X_i)_{i=-\infty}^{\infty}$ is a sequence of independent identically distributed variables then the probability of the mentioned events is 0 or 1. In the case of event (5.1) direct evaluation yields probability 0 for $P(X_i = 1) < 1$ and 1 for $P(X_i = 1) = 1$. The probability of event (5.2) is also 0 or 1 by the Kolmogorov zero-one law. Finally, if the strong law of large numbers holds, the probability of (5.3) is 1 if and only if $a = \mathbf{E}\, X_i$, where $\mathbf{E}\, X_i$ is the expectation of random variable X_i.

Following our intuition for independent variables, we may think that a stationary process is well-behaved if the probability of invariant events is 0 or 1. Formally, this condition is known as ergodicity.

Definition 5.2 (ergodicity). *A dynamical system* $(\Omega, \mathcal{J}, P, T)$ *is called* ergodic *if any event from the invariant algebra has probability* 0 *or* 1, *i.e.,*

$$A \in \mathcal{I} \implies P(A) \in \{0, 1\} .$$

Analogously, we call a stationary process $(X_i)_{i=-\infty}^{\infty}$ ergodic *if the dynamical system generated by this process is ergodic.*

Now we will present the ergodic theorem, which gives a more intuitive characterization of ergodic systems. In 1931, Georg David Birkhoff (1884–1944) showed the following fact:

Theorem 5.1 (individual ergodic theorem). *Let* $(\Omega, \mathcal{J}, P, T)$ *be a dynamical system and define stationary process* $X_i(\omega) := X_0(T^i\omega)$ *for a real random variable* X_0 *on the probability space* $(\Omega, \mathcal{J}, P)$. *The dynamical system is ergodic if and only if for any real random variable* X_0 *where* $\boldsymbol{E}\,|X_0| < \infty$ *equality*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \boldsymbol{E} X_0 \tag{5.4}$$

holds with probability 1.

The concept of a dynamical system has its roots in statistical mechanics. Historically, it was imagined that variables X_k represent a certain time evolution of a state function X_0 subject to iteration of operation T, which describes the dynamics of a physical system. Thus, the dynamical system is ergodic if and only if the time average of *any* state function of the system equals its expectation.

In information theory, we often use this property of ergodic processes in the following way. Namely, for a stationary ergodic process $(X_i)_{i=-\infty}^{\infty}$, with probability 1 we have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left[-\log P(X_k|X_{k-m}^{k-1}) \right] &= \mathbf{E} \left[-\log P(X_0|X_{-m}^{-1}) \right] \\ &= H\big(X_0|X_{-m}^{-1}\big). \end{aligned}$$

This equality holds since $P(X_k|X_{k-m}^{k-1})$ is a random variable on the probability space generated by process $(X_i)_{i=-\infty}^{\infty}$.

The proof of the ergodic theorem consists of two parts. The "if" part is very easy and will be presented first.

Proof of Theorem 5.1($\Leftarrow$): Assume that (5.4) holds for any random variable X_0. Let $I_A(\omega) = \mathbf{1}\{\omega \in A\}$ be the indicator function of an event A. If A belongs to the invariant algebra, we have $\omega \in A \iff \omega \in T^{-i}A \iff T^i\omega \in A$ so $I_A = I_A \circ T^i$. Hence

$$I_A = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I_A \circ T^i = \mathbf{E}\, I_A = P\big(A\big)$$

holds with probability 1. Because $I_A \in \{0, 1\}$ with probability 1, we obtain $P(A) \in \{0, 1\}$. Thus the dynamical system is ergodic. □

Now we shall present a more involved proof of the “only if” part by Adriano Garsia. First, we will demonstrate an auxiliary fact called the maximal ergodic theorem.

Lemma 5.1 (maximal ergodic theorem). *Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary process where $\boldsymbol{E}\,|X_1| < \infty$. Define $S_n = \sum_{k=1}^{n} X_k$ and $M_n = \max(0, S_1, S_2, ..., S_n)$. We have*

$$\int_{M_n>0} X_1 \, \mathrm{d}P \geq 0.$$

Proof. For each $k \leq n$ we have $M_n \circ T \geq S_k \circ T$. Hence

$$X_1 + M_n \circ T \geq X_1 + S_k \circ T = S_{k+1}.$$

Let us write it as

$$X_1 \geq S_{k+1} - M_n \circ T, \quad k = 1, ..., n.$$

But we also have

$$X_1 \geq S_1 - M_n \circ T.$$

Both inequalities yield $X_1 \geq \max(S_1, S_2, ..., S_n) - M_n \circ T$. Hence

$$\begin{aligned} \int_{M_n>0} X_1 \, \mathrm{d}P &\geq \int_{M_n>0} \big[M_n - M_n \circ T\big] \, \mathrm{d}P \\ &= \int_{M_n>0} M_n \, \mathrm{d}P - \int_{M_n>0} M_n \circ T \, \mathrm{d}P \\ &\geq \int M_n \, \mathrm{d}P - \int M_n \circ T \, \mathrm{d}P = 0. \end{aligned}$$

In the next step, we will prove the ergodic theorem proper.

Proof of Theorem 5.1($\Rightarrow$): Assume that the dynamical system is ergodic. Without loss of generality, let us assume $\mathbf{E}\, X_1 = 0$. Statement (5.4) can be derived applying the proof below to process $(X_i - \mathbf{E}\, X_i)_{i=-\infty}^{\infty}$. For a fixed ϵ denote the event

$$G = \Big(\limsup_{n\to\infty} S_n/n > \epsilon\Big).$$

We define random variable

$$X_i^*(\omega) = (X_i(\omega) - \epsilon)\mathbf{1}\{\omega \in G\}$$

and using X_i^* we define S_n^* and M_n^* as in the statement of the maximal ergodic theorem.

By the maximal ergodic theorem, we have

$$\int_{M_n^*>0} X_1^* \, \mathrm{d}P \geq 0.$$

The remaining part of the proof is not so difficult. Events

$$F_n = (M_n^* > 0) = \big(\max_{1\le k\le n} S_k^* > 0\big)$$

converge to

$$F = \big(\sup_{k\ge1} S_k^* > 0\big) = \big(\sup_{k\ge1} S_k^*/k > 0\big) = \big(\sup_{k\ge1} S_k/k > \epsilon\big) \cap G = G.$$

Inequality $\mathbf{E}\,|X_1^*| \le \mathbf{E}\,|X_1| + \epsilon < \infty$ allows to use the Lebesgue dominated convergence theorem (Theorem 1.8), which yields

$$\lim_{n\to\infty} \int_{F_n} X_1^* \,\mathrm{d}P = \int_F X_1^* \,\mathrm{d}P.$$

Hence

$$\int_G X_1^* \,\mathrm{d}P \ge 0.$$

But $G \in \mathcal{I}$ so $\int_G X_1 \,\mathrm{d}P = 0$, regardless whether $P(G) = 0$ or $P(G) = 1$. Hence

$$\int_G X_1^* \,\mathrm{d}P = \int_G X_1 \,\mathrm{d}P - \epsilon P\big(G\big) = -\epsilon P\big(G\big),$$

and thus $P(G) = 0$. Parameter ϵ was chosen arbitrarily so we have $\limsup_{n\to\infty} S_n/n \le 0$ with probability 1. Applying the analogous reasoning to process $(-X_i)_{i=1}^\infty$, we also obtain $\liminf_{n\to\infty} S_n/n \ge 0$. Hence $\lim_{n\to\infty} S_n/n = 0$. □

The ergodic theorem may be generalized to the case of stationary nonergodic processes. In that case, the right hand side of (5.4) equals conditional expectation $\mathbf{E}\,[X_1|\mathcal{I}]$ (see Definition 1.6). The interested reader is referred to Breiman (1992, Section 6.5).

Before we give some examples of ergodic processes, it is insightful to present an instance of a process which is not ergodic.

Example 5.2 (nonergodic process). Let $(U_i)_{i=-\infty}^\infty$ and $(W_i)_{i=-\infty}^\infty$ be independent stationary ergodic processes having different distributions and let an independent variable Z have distribution $P(Z = 0) = p \in (0, 1)$ and $P(Z = 1) = 1 - p$. We will show that process $(X_i)_{i=-\infty}^\infty$, where

$$X_i = \mathbf{1}\{Z = 0\}U_i + \mathbf{1}\{Z = 1\}W_i,$$

is stationary but not ergodic. Assume that $P(U_1^p = w) \ne P(W_1^p = w)$. Then

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n \mathbf{1}\{X_k^{k+p-1} = w\} \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n \Big(\mathbf{1}\{Z = 0\}\mathbf{1}\big\{U_k^{k+p-1} = w\big\} + \mathbf{1}\{Z = 1\}\mathbf{1}\big\{W_k^{k+p-1} = w\big\}\Big) \\
&= \mathbf{1}\{Z = 0\}P\big(U_1^p = w\big) + \mathbf{1}\{Z = 1\}P\big(W_1^p = w\big), \qquad (5.5)
\end{aligned}$$

which is not constant. Hence $(X_i)_{i=-\infty}^\infty$ is not ergodic. The expectation of (5.5) equals, however, $P(X_1^p = w)$.

A stationary process is ergodic if there is no such random "switch" like the variable Z in the above example. To substantiate this claim, let us present a proposition that specifies which Markov chains are ergodic. The proof can be found in Breiman (1992, Theorem 7.16).

Theorem 5.2. *Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary Markov chain, where $P(X_{i+1} = l|X_i = k) = p_{kl}$, $P(X_i = k) = \pi_k$, and the variables assume values from a countable set. The following conditions are equivalent:*

1. *Process $(X_i)_{i=-\infty}^{\infty}$ is ergodic.*
2. *There are no two disjoint closed sets of states; a set A of states is called closed if $\sum_{l\in A} p_{kl} = 1$ for each $k \in A$.*
3. *For a given transition matrix (p_{kl}) there exists a unique stationary distribution.*

Proving 2. $\implies$ 1. is more difficult than the converse. In fact, the converse statement is quite easy. For suppose that there are two disjoint closed sets of states A and B. Then we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{X_k \in A\} = \mathbf{1}\{X_1 \in A\},$$

which is not constant. Thus the process is not ergodic. That is, 1. $\implies$ 2.

Now we may present a few examples of ergodic processes.

Example 5.3. A sequence of independent identically distributed random variables is an ergodic process.

Remark: Hence the strong law of large numbers (1.4) is a corollary of the ergodic theorem (Theorem 5.1).

Example 5.4. Consider a stationary Markov chain $(X_i)_{i=-\infty}^{\infty}$ where $P(X_{n+1} = j|X_n = i) = p_{ij}$ and $P(X_1 = i) = \pi_i$. It is ergodic for

$$\begin{pmatrix}\pi_1 \; \pi_2\end{pmatrix} = \begin{pmatrix}1/2 \; 1/2\end{pmatrix}, \qquad \begin{pmatrix}p_{11} & p_{12} \\ p_{21} & p_{22}\end{pmatrix} = \begin{pmatrix}0 & 1 \\ 1 & 0\end{pmatrix},$$

and nonergodic for

$$\begin{pmatrix}\pi_1 \; \pi_2\end{pmatrix} = \begin{pmatrix}a \; 1-a\end{pmatrix}, \qquad \begin{pmatrix}p_{11} & p_{12} \\ p_{21} & p_{22}\end{pmatrix} = \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}.$$

Ergodicity of some hidden Markov chains follows from this proposition.

Theorem 5.3. *If $(X_i)_{i=-\infty}^{\infty}$ is an ergodic process then process $(Y_i)_{i=-\infty}^{\infty}$ where $Y_k = g(T^k(X_i)_{i=-\infty}^{\infty})$ is also ergodic.*

In particular a hidden Markov chain $Y_k = g(X_k)$ is ergodic, too.

Proof. By the ergodic theorem, the process $(X_i)_{i=-\infty}^{\infty}$ is ergodic if and only if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Z_k = \mathbf{E}\, Z_0$$

for any process $Z_k = f(T^k(X_i)_{i=-\infty}^{\infty})$. In particular, we may take $Z_k = h(T^k(Y_i)_{i=-\infty}^{\infty})$ for an arbitrary function h. Hence process $(Y_i)_{i=-\infty}^{\infty}$ is ergodic by the ergodic theorem.

An important corollary of the ergodic theorem is the Shannon-McMillan-Breiman theorem which states that the blocks in a stationary process are asymptotically equidistributed.

Theorem 5.4 (Shannon-McMillan-Breiman theorem). *Let $P(X_1^n)$ be the random variable that assumes value $P(X_1^n = x_1^n)$ if and only if $X_1^n = x_1^n$. For a stationary ergodic process $(X_i)_{i=-\infty}^{\infty}$, where X_i assume finitely many values, equality*

$$-\lim_{n\to\infty} \frac{1}{n} \log P\big(X_1^n\big) = h \tag{5.6}$$

holds with probability 1, *where h is the entropy rate of $(X_i)_{i=-\infty}^{\infty}$.*

Theorem 5.4 has been generalized by Chung to variables that assume countably infinitely many values, but the proof is more involved than we wish to give here.

In order to demonstrate the Shannon-McMillan-Breiman theorem, we introduce two quantities

$$\begin{aligned} P^k\big(X_1^n\big) &= P\big(X_1^k\big) \prod_{i=k+1}^{n} P\big(X_i|X_{i-k}^{i-1}\big), \quad k \ge 1 \\ h_\infty &= \mathbf{E}\,\left[-\log P(X_1|X_{-\infty}^{0})\right] \end{aligned}$$

and we prove the following three auxiliary results.

Lemma 5.2. *For a stationary ergodic process $(X_i)_{i=-\infty}^{\infty}$, equalities*

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{P^k(X_1^n)}{P(X_1^n)} \le 0, \tag{5.7}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^{0})} \le 0 \tag{5.8}$$

hold with probability 1.

Proof. Let A denote the support set of $P(X_1^n)$. We have

$$\mathbf{E}\left[\frac{P^k(X_1^n)}{P(X_1^n)}\right] = \sum_{x_1^n \in A} P(X_1^n = x_1^n) \frac{P^k(X_1^n = x_1^n)}{P(X_1^n = x_1^n)} \le 1. \tag{5.9}$$

By Markov inequality (Theorem 1.5) and (5.9), we have

$$P\left(\frac{P^k(X_1^n)}{P(X_1^n)} \geq n^2\right) \leq \frac{1}{n^2}$$

or

$$P\left(\frac{1}{n}\log\frac{P^k(X_1^n)}{P(X_1^n)} \geq \frac{2}{n}\log n\right) \leq \frac{1}{n^2}$$

Noting that $\sum_{n=1}^{\infty} 1/n^2 < \infty$, we infer from the Borel-Cantelli lemma (Theorem 1.4) that the event

$$\frac{1}{n}\log\frac{P^k(X_1^n)}{P(X_1^n)} \geq \frac{2}{n}\log n$$

occurs only finitely often with probability one. Hence we obtain (5.7).

To infer (5.8), we use a similar reasoning. First, let $B(X_{-\infty}^0)$ denote the support set of $P(X_1^n|X_{-\infty}^0)$. Using conditional expectation (see Definition 1.6), we obtain

$$\begin{aligned}\mathbf{E}\left[\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\right] &= \mathbf{E}\left[\mathbf{E}\left[\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\middle|X_{-\infty}^0\right]\right] \\ &= \mathbf{E}\left[\sum_{x_1^n\in B(X_{-\infty}^0)} P\big(X_1^n = x_1^n|X_{-\infty}^0\big)\frac{P(X_1^n = x_1^n)}{P(X_1^n = x_1^n|X_{-\infty}^0)}\right] \leq 1. \qquad (5.10)\end{aligned}$$

Applying the same arguments using Markov's inequality to (5.10), we obtain (5.8).

Lemma 5.3. *For a stationary ergodic process* $(X_i)_{i=-\infty}^{\infty}$*, equalities*

$$-\lim_{n\to\infty}\frac{1}{n}\log P^k\big(X_1^n|X_{-\infty}^0\big) = H\big(X_1|X_{-k+1}^0\big), \quad k \geq 1, \qquad (5.11)$$

$$-\lim_{n\to\infty}\frac{1}{n}\log P\big(X_1^n|X_{-\infty}^0\big) = h_\infty \qquad (5.12)$$

hold with probability 1.

Proof. Functions of ergodic processes are ergodic processes. Hence by the ergodic theorem, we obtain

$$\begin{aligned}-\lim_{n\to\infty}\frac{1}{n}\log P^k\big(X_1^n|X_{-\infty}^0\big) &= -\lim_{n\to\infty}\frac{1}{n}\log P\big(X_1^k\big) - \lim_{n\to\infty}\frac{1}{n}\sum_{i=k+1}^{n}\log P\big(X_i|X_{i-k}^{i-1}\big) \\ &= 0 + H\big(X_1|X_{-k+1}^0\big), \\ -\lim_{n\to\infty}\frac{1}{n}\log P\big(X_1^n|X_{-\infty}^0\big) &= -\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log P\big(X_i|X_{-\infty}^{i-1}\big) \\ &= h_\infty.\end{aligned}$$

Lemma 5.4. *For a stationary process $(X_i)_{i=-\infty}^{\infty}$, where X_i assume finitely many values, we have* $\lim_{k\to\infty} H(X_1|X_{-k}^0) = h = h_\infty$.

Proof. In the previous chapter we have defined $h := \lim_{k\to\infty} H(X_1|X_{-k}^0)$. Now it remains to show that $h = h_\infty$. By the Levy law for conditional probabilities (Theorem 1.9) we have

$$\lim_{k\to\infty} P\big(X_1 = x_1|X_{-k}^0\big) = P\big(X_1 = x_1|X_{-\infty}^0\big)$$

with probability 1. Since the alphabet is finite and function $p \log p$ is bounded, the dominated convergence theorem (Theorem 1.8) allows to interchange the expectation and the limit. This yields

$$\lim_{k\to\infty} H\big(X_1|X_{-k}^0\big) = \lim_{k\to\infty} \mathbf{E}\left[-\sum_x P(X = x_1|X_{-k}^0) \log P(X = x_1|X_{-k}^0)\right]$$
$$= \mathbf{E}\left[-\sum_x P(X = x_1|X_{-\infty}^0) \log P(X = x_1|X_{-\infty}^0)\right] = h_\infty.$$

Now the proof of the main proposition follows.

Proof of Theorem 5.4: On the one hand, from (5.7) and (5.11), we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P(X_1^n)} \le H\big(X_1|X_{-k}^0\big).$$

On the other hand, from (5.8) and (5.12), we obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{P(X_1^n)} \ge h_\infty.$$

Since $\lim_{k\to\infty} H(X_1|X_{-k}^0) = h = h_\infty$, this yields (5.6). □

Exercises

1. *(Invariant algebra)* Show that the invariant algebra is a σ-field.
2. *(Weakly stationary processes)* A stochastic process $(X_i)_{i=-\infty}^{\infty}$ is called *weakly stationary* if $\mathbf{E}\, X_i = \mu$ and $\mathbf{E}\,(X_i - \mu)(X_j - \mu) = \sigma(|i-j|)$ for $i, j \in \mathbb{Z}$. Moreover, the process is called *weakly ergodic* if

   $$\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_i - \mu\right]^2 = 0.$$

   Show that the process is weakly ergodic if and only if

   $$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma(|i-j|) = 0.$$

   Prove also that the latter condition is satisfied if $\lim_{k\to\infty} \sigma(k) = 0$.

3. Compute $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k$ for a stationary nonergodic Markov chain, where X_i assume a finite number of values.
4. *(Gambling and information theory)* Following Cover and Thomas (2006), let us consider this curious problem. Bookmakers at horse races return x_k dollars for 1 dollar paid for a horse k if it wins and they pay nothing if it loses. In each race, there is exactly one horse that wins. Obviously our gain depends on the stakes chosen by the bookmakers and the probabilities of horses' winnings. It turns out that the optimal strategy of betting does not depend on bookmakers' stakes. Show that this holds for the following scheme: Let W_n be our capital after the n-th race. Out of W_{n-1} dollars that we have before the n-th race we bet $b_k W_{n-1}$ dollars for the horse $k = 1, ..., q$, where $b_k \geq 0$ and $\sum_{k=1}^{q} b_k = 1$. Moreover, let K_i be the horse that wins in the i-th race. We assume that $(K_i)_{i=-\infty}^{\infty}$ forms a stationary ergodic process with $P(K_i = k) = p_k$. Find the optimal b_k and show that they do not depend on x_k.

6

Lempel-Ziv code

The problem of universal compression. The definition of Lempel-Ziv code. Universality of the Lempel-Ziv code.

From Theorem 4.8, we know that for any stationary process $(X_i)_{i=-\infty}^{\infty}$ there exists a sequence of uniquely decodable codes $B_n : \mathbb{X}^n \to \{0,1\}^*$ that achieve the compression rate equal to the entropy rate. That is

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left|B_n\left(X_1^n\right)\right| = h.$$

In the proposition above, we may take B_n equal to the Shannon-Fano code for the probability distribution of block X_1^n. To compute the Shannon-Fano code we need however to know the probability distribution of the block. Such a situation is unlikely in practical applications of data compression, where we have no prior information about the probability distribution of blocks. Fortunately, as an important corollary of the ergodic theorem, there exist universal codes whose compression rates tend to the entropy rate for any stationary process.

Definition 6.1 (L^1-universal code). *A uniquely decodable code $B : \mathbb{X}^* \to \{0,1\}^*$ is called* L^1-universal *if for any stationary process (not necessarily ergodic) we have*

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{E}\left|B\left(X_1^n\right)\right| = h.$$

Definition 6.2 (universal code with probability 1). *On the other hand, a uniquely decodable code $B : \mathbb{X}^* \to \{0,1\}^*$ is called* universal with probability 1 *if for any stationary ergodic process inequality*

$$\limsup_{n\to\infty} \frac{\left|B\left(X_1^n\right)\right|}{n} \le h.$$

holds with probability 1*.*

The second condition is stronger.

Theorem 6.1. *Let code B be universal with probability* 1*. If there exists a constant K such that*

$$\left|B\left(x_1^n\right)\right| \le Kn$$

for each string x_1^n then code B is L^1-universal.

We omit the proof of this theorem because it relies on the ergodic decomposition for stationary processes, the proof of which is quite difficult (cf. Kallenberg, 1997, Theorem 9.12).

The problem of universal compression falls under the scope of statistics. Indeed, the interest of statisticians lies in identifying parameters of a stochastic process basing on the data typical for that process. Entropy rate of an ergodic process is an example of such a parameter. When we have a universal code, we may estimate the entropy rate as the compression rate yielded by the code.

Now we will present the oldest known universal code, called the Lempel-Ziv (LZ) code. The code was derived by Abraham Lempel (1936–) and Jacob Ziv (1931–) in 1977 (Ziv and Lempel, 1977). The LZ code is partly implemented in the Unix programs `gzip` and `compress`. It is worth noting, however, that neither of those programs yields a universal code because the buffer length imposed in both `gzip` and `compress` is limited.

Definition 6.3 (LZ code). *For simplicity of the algorithm description we assume that the compressed data are binary sequences, that is* $\mathbb{X} = \{0, 1\}$*. The Lempel-Ziv compression algorithm is as follows.*

1. *The compressed sequence is parsed into a sequence of shortest phrases that have not appeared before (except for the last phrase). For example, the sequence* 001010010011100... *is split into phrases* $0, 01, 010, 0100, 1, 11, 00, ...$.
2. *In the following, each phrase is described using a binary index of the longest prefix that appeared earlier and a single bit that follows that prefix. For the considered sequence, this representation is as follows:* $(0,0)(1,1)(10,0)(11,0)(0,1)(101,1)(1,0)$.

Let C_n be the number of phrases in the compressed block X_1^n. If we know C_n, we need $\log C_n$ bits to identify the prefix index for each phrase and 1 bit to describe the following bit. Thus the LZ code uses $|B(X_1^n)| = C_n \left[\log C_n + O(1)\right]$ bits in total to describe the whole sequence.

Now we will prove that LZ code is universal. A splitting of a sequence into distinct phrases will be called a *distinct parsing* of the sequence. Universality of the LZ code follows from this proposition.

Theorem 6.2. *Let* $(X_i)_{i=-\infty}^{\infty}$ *be a stationary ergodic process and let* C_n *be the number of phrases in a distinct parsing of block* $(X_1, X_2, ..., X_n)$*. With probability* 1 *we have*

$$\limsup_{n\to\infty} \frac{C_n \left[\log C_n + O(1)\right]}{n} \le h. \tag{6.1}$$

Remark: Hence the LZ code is universal with probability 1. Moreover, inequality $C_n[\log C_n + O(1)] \le Kn$ follows from Lemma 6.1, discussed below. Hence by Theorem 6.1 the LZ code is also L^1-universal.

For the derivation of Theorem 6.2 we need a couple of auxiliary statements.

Lemma 6.1. *The number of phrases C_n in any distinct parsing of block $(X_1, X_2, ..., X_n)$ satisfies inequality*

$$\lim_{n\to\infty} \frac{C_n \log n}{n} \le 1.$$

Proof. Let $n_k = \sum_{j=1}^{k} j2^j = (k-1)2^{k+1} + 2$ be the sum of lengths of distinct phrases that are not longer than k. The number of phrases C_n in a distinct parsing will be maximal if the phrases are as short as possible. For $n_k \le n < n_{k+1}$ this happens if we take all phrases of length $\le k$ and $\delta/(k+1)$ phrases of length $k+1$, where $\delta = n - n_k$. Then

$$C_n \le \sum_{j=1}^{k} 2^j + \frac{\delta}{k+1} = 2^{k+1} - 2 + \frac{\delta}{k+1} \le \frac{n_k}{k-1} + \frac{\delta}{k+1} \le \frac{n}{k-1}.$$

In the following we will provide a bound for k given n. We have $n \ge n_k = (k-1)2^{k+1} + 2 \ge 2^k$, so

$$k \le \log n.$$

Moreover $n < n_{k+1} = k2^{k+2} + 2 \le (\log n + 2)2^{k+2}$. Hence

$$k + 2 > \log \frac{n}{\log n + 2}.$$

Further transformations yield $k - 1 > \log n - \log(\log n + 2) - 3$. Hence we obtain the claim.

For the next lemma we need some new notation. Let P^k denote the measure of the k-th order Markov approximation of the process $(X_i)_{i=-\infty}^{\infty}$. That is

$$P^k\big(X_{-k+1}^n | X_{-k+1}^0\big) := \prod_{i=1}^{n} P\big(X_i | X_{i-k}^{i-1}\big).$$

Moreover, assume that sequence $(X_1, X_2, ..., X_n)$ is parsed into C_n distinct phrases $(Y_1, Y_2, ..., Y_{C_n})$. Let W_i denote the k bits preceding Y_i. Next, let C_n^{lw} denote the number of phrases Y_i that have length l and context $W_i = w$.

Lemma 6.2 (Ziv inequality). *We have inequality*

$$-\log P^k\big(X_1, X_2, ..., X_n | W_1\big) \ge \sum_{l,w} C_n^{lw} \log C_n^{lw}.$$

Proof. Observe that

$$-\log P^k\big(X_1, X_2, ..., X_n | W_1\big) = -\sum_{j=1}^{C_n} \log P\big(Y_j | W_j\big)$$

$$= -\sum_{l,w} C_n^{lw} \cdot \frac{1}{C_n^{lw}} \sum_{j:|Y_j|=l, W_j=w} \log P^k\big(Y_j|W_j\big)$$
$$\geq -\sum_{l,w} C_n^{lw} \log\Big(\frac{1}{C_n^{lw}} \sum_{j:|Y_j|=l, W_j=w} P^k\big(Y_j|W_j\big)\Big),$$

where the inequality follows from the Jensen inequality because the logarithm function is concave. Because the phrases Y_j under the sum are distinct, we have $\sum_{j:|Y_j|=l,W_j=w} P^k(Y_j|W_j) \leq 1$. Hence the claim follows.

Lemma 6.3. *Let L be a nonnegative random variable taking values in integers and having expectation $\boldsymbol{E}L$. Then entropy $H(L)$ is bounded by inequality*

$$H\big(L\big) \leq \big(\boldsymbol{E}L + 1\big) \log\big(\boldsymbol{E}L + 1\big) - \boldsymbol{E}L \log \boldsymbol{E}L.$$

The proof of this lemma will be discussed in Chapter 12 as an exercise.

Now we can prove the main theorem.

Proof of Theorem 6.2: Let L and W be random variables such that

$$P\big(L = l, W = w\big) = \frac{C_n^{lw}}{C_n}.$$

The expectation of L is

$$\mathbf{E}\,L = \sum_{l,w} \frac{lC_n^{lw}}{C_n} = \frac{n}{C_n}.$$

Hence by Lemma 6.3, we obtain

$$\begin{aligned}
H\big(L\big) &\leq \big(\mathbf{E}\,L + 1\big) \log\big(\mathbf{E}\,L + 1\big) - \mathbf{E}\,L \log \mathbf{E}\,L \\
&= \Big(\frac{n}{C_n} + 1\Big) \log\Big(\frac{n}{C_n} + 1\Big) - \frac{n}{C_n} \log\frac{n}{C_n} \\
&= \Big(\frac{n}{C_n} + 1\Big) \log\big(C_n + n\big) + \Big(\frac{n}{C_n} + 1\Big) \log\frac{1}{n} - \frac{n}{C_n} \log\frac{n}{C_n} \\
&= \Big(\frac{n}{C_n} + 1\Big) \log\big(C_n + n\big) - \frac{n}{C_n} \log n + \log\frac{1}{C_n} + \log n - \log n \\
&= \Big(\frac{n}{C_n} + 1\Big) \log\big(C_n + n\big) - \Big(\frac{n}{C_n} + 1\Big) \log n + \log\frac{n}{C_n} \\
&= \log\frac{n}{C_n} + \Big(\frac{n}{C_n} + 1\Big) \log\Big(\frac{C_n}{n} + 1\Big).
\end{aligned}$$

On the other hand, $H(W) \leq k$, so

$$H\big(L, W\big) \leq H\big(L\big) + H\big(W\big) \leq \log\frac{n}{C_n} + \Big(\frac{n}{C_n} + 1\Big) \log\Big(\frac{C_n}{n} + 1\Big) + k.$$

Then by Lemma 6.1, we have

$$\begin{aligned}
&\lim_{n\to\infty}\frac{C_n}{n}H\big(L,W\big)\\
&=\lim_{n\to\infty}\left(-\frac{C_n}{n}\log\frac{C_n}{n}+\Big(\frac{C_n}{n}+1\Big)\log\Big(\frac{C_n}{n}+1\Big)+\frac{C_n}{n}k\right)=0.
\end{aligned}$$

Now using Lemma 6.1 again, the Ziv inequality, and the ergodic theorem, we obtain

$$\begin{aligned}
\limsup_{n\to\infty}\frac{C_n\big[\log C_n+O(1)\big]}{n}&=\limsup_{n\to\infty}\Big(\frac{C_n\log C_n}{n}-\frac{C_n}{n}H\big(L,W\big)\Big)\\
&=\limsup_{n\to\infty}\frac{1}{n}\Big(C_n\log C_n+C_n\sum_{l,w}\frac{C_n^{lw}}{C_n}\log\frac{C_n^{lw}}{C_n}\Big)\\
&=\limsup_{n\to\infty}\frac{1}{n}\sum_{l,w}C_n^{lw}\log C_n^{lw}\\
&\le-\lim_{n\to\infty}\frac{1}{n}\log P^k\big(X_1^n|X_{-k+1}^0\big)\\
&=-\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log P\big(X_i|X_{i-k}^{i-1}\big)\\
&=\mathbf{E}\,\big[-\log P\big(X_i|X_{i-k}^{i-1}\big)\big]=H\big(X_i|X_{i-k}^{i-1}\big).
\end{aligned}$$

with probability 1. This inequality holds for any k. Considering $k\to\infty$, we obtain (6.1). □

The LZ code is very simple but its convergence to the entropy rate is not very fast. There are codes which compress particular sources better, such as grammar-based codes for natural language (Kieffer and Yang, 2000). Usually the better the compression, the harder is a code to compute. The limit of efficient compression is set by the Kolmogorov complexity but, as we will learn in Chapter 13, the Kolmogorov complexity itself is not computable.

Exercises

1. The LZ algorithm, as described in this chapter, may have problems with parsing the last phrase of a compressed sequence. Propose a modification of the algorithm which solves this problem and argue that a so modified code is universal.
2. Find the LZ parsings for the sequences:
   (a) 0101010101010101010...,
   (b) 1001000100001000001...,
   (c) 001001001001001001...,
   (d) 1011001100011000011....

3. Find the sequences corresponding to these LZ parsings:
   (a) $(0,1)(1,0)(1,1)(0,0)(10,1)(11,1)...$,
   (b) $(0,0)(0,1)(10,1)(11,0)(100,1)(0,1)...$,
   (c) $(0,1)(0,0)(10,1)(11,0)(100,1)(11,1)...$,
   (d) $(0,0)(0,1)(10,1)(10,0)(11,1)(100,0)....$
   Check whether these parsings are produced by the LZ algorithm.
4. Consider the constant sequence 00000000....
   (a) Produce the LZ parsing for this sequence.
   (b) Show that the number of bits per symbol for prefixes of that sequence tends to zero with the increasing length.
5. Produce a sequence for which the number of phrases in the LZ parsing grows as fast as possible.
6. Produce a sequence for which the number of phrases in the LZ parsing grows as slow as possible.

7

Gaussian processes

Differential entropy and Kullback-Leibler divergence. Stationary Gaussian processes. Autocorrelation and partial autocorrelation. Information measures for Gaussian processes. Innovation variance and generalized variance.

In this chapter we will discuss entropy of continuous rather than discrete probability distributions. Let us recall that probability density ρ of a (continuous) real random variable X is a function such that

$$P(X \in A) = \int_A \rho(x)\,\mathrm{d}x. \tag{7.1}$$

For a probability density we may define entropy in the following way.

Definition 7.1 (differential entropy). *The* (differential) entropy *of probability density ρ is defined as*

$$H(\rho) := -\int \rho(x)\ln\rho(x)\,\mathrm{d}x.$$

The natural logarithm (ln) rather than the binary logarithm (log) is used for convenience. Unlike the discrete case, the differential entropy may be negative and infinite.

Example 7.1. Consider the density of the Gauss distribution:

$$\rho(x) = \left[\frac{1}{2\pi\sigma^2}\right]^{1/2}\exp\left[-\frac{x^2}{2\sigma^2}\right].$$

We obtain

$$H(\rho) = -\ln\left[\frac{1}{2\pi\sigma^2}\right]^{1/2} + \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}\ln(2\pi e\sigma^2),$$

which is negative for $\sigma < 1/\sqrt{2\pi e}$.

For a real random variable X with density (7.1) we will define the differential entropy as

$$H\big(X\big) := H(\rho).$$

In contrast to differential entropy, the continuous analogue of Kullback-Leibler divergence is nonnegative like in the discrete case.

Definition 7.2 (KL divergence). *The* Kullback-Leibler divergence *for probability densities ρ and ρ^* is defined as*

$$D(\rho||\rho^*) := \int \rho(x) \ln \frac{\rho(x)}{\rho^*(x)} \, \mathrm{d}x.$$

Theorem 7.1. *We have*

$$D(\rho||\rho^*) \geq 0$$

with equality if and only if $\rho = \rho^*$ except for a set of measure 0.

Proof. Observe that $\ln x \geq 1 - 1/x$ with equality if and only if $x = 1$. Hence

$$\begin{aligned} D(\rho||\rho^*) &= \int \rho(x) \ln \frac{\rho(x)}{\rho^*(x)} \, \mathrm{d}x \\ &\geq \int \rho(x) \left[1 - \frac{\rho^*(x)}{\rho(x)} \right] \mathrm{d}x = \int \rho(x) \, \mathrm{d}x - \int \rho^*(x) \, \mathrm{d}x = 0 \end{aligned}$$

with equality if and only if $\rho = \rho^*$ except for a set of measure 0.

As we recall from Chapter 2, mutual information for discrete random variables is equal to the Kullback-Leibler divergence between the joint distribution and the product of marginal distributions. The same property holds for mutual information between real random variables X and Y if we define it as

$$\begin{aligned} I(X;Y) &:= H(X) + H(Y) - H(X,Y) \\ &= \int \rho_{XY}(x,y) \log \frac{\rho_{XY}(x,y)}{\rho_X(x)\rho_Y(y)} \, \mathrm{d}x \, \mathrm{d}y, \end{aligned} \tag{7.2}$$

where ρ_{XY} is the density of the random vector (X, Y) and ρ_X, ρ_y are the marginal densities of it. We will adopt (7.2) as the definition of mutual information for real variables. Then, by Theorem 7.1, we have $I(X;Y) \geq 0$ with equality if and only if X and Y are independent, exactly as in the discrete case.

For completeness, we will also define the conditional entropy and conditional mutual information as

$$\begin{aligned} H(X|Y) &:= H(X,Y) - H(Y), \\ I(X;Y|Z) &:= H(X|Z) + H(Y|Z) - H(X,Y|Z). \end{aligned}$$

Quantity $H(X|Y)$ need not be nonnegative, whereas $I(X;Y|Z)$ is.

For a pair of real random variables X and Y let us also denote the covariance $\mathrm{Cov}(X,Y) := \mathbf{E}\,[XY] - \mathbf{E}\,X \mathbf{E}\,Y$, the norm $||X|| := \sqrt{\mathrm{Cov}(X,X)}$ and the correlation $\mathrm{Corr}(X,Y) := \mathrm{Cov}(X,Y)/\,||X|| \cdot ||Y||$. In the following, we will consider the class of processes whose every distribution is Gaussian (normal) centered around 0.

Definition 7.3 (zero-mean Gaussian process). *A* zero mean Gaussian process $(X_i)_{i=1}^{\infty}$ *is a process such that the densities of vector* X_1^n *take form*

$$\rho(x_1^n) = \frac{1}{\left[(2\pi)^n \det \Gamma(n)\right]^{1/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i \left[\Gamma(n)^{-1}\right]_{ij} x_j\right], \qquad (7.3)$$

where $\Gamma(n)$ *is the autocovariance matrix.*

The following proposition states when the Gaussian process with a given autocovariance exists, cf. Brockwell and Davis (1987, Proposition 1.6.2).

Theorem 7.2. *The Gaussian process with densities (7.3) exists if and only if for each* n *matrix* $\Gamma(n)$ *is symmetric and nonnegative definite, i.e., if* $\Gamma(n)_{ij} = \Gamma(n)_{ji}$ *and*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i \Gamma(n)_{ij} a_j \geq 0 \qquad (7.4)$$

for every vector a_1^n*.*

Condition (7.4) is equivalent to

$$\mathbf{E}\left[\sum_{i=1}^{n} a_i X_i\right]^2 \geq 0. \qquad (7.5)$$

For a Gaussian process, quantity $\mathbf{E}\left[X_i - \sum_{j\in K} a_j X_j\right]^2$ is a quadratic function of coefficients a_j with a single minimum denoted $a_j = \phi_{ij}^K$ (Brockwell and Davis, 1987, Theorem 2.3.1). It can proved that the so called innovation

$$X_i - P_K X_i,$$

where $P_K X_i := \sum_{j\in K} \phi_{ij}^K X_j$, is independent of X_j, $j \in K$ (Brockwell and Davis, 1987, Proposition 1.6.6). $P_K X_i$ is called the best linear predictor of X_i given $(X_j)_{j\in K}$.

For a stationary Gaussian process, we have $\Gamma(n)_{ij} = \operatorname{Cov}(X_i, X_j) = \gamma(i-j)$, where $\gamma(\cdot)$ is called the autocovariance function. We further define autocorrelation function (ACF) $\rho(n)$, best linear predictor coefficients ϕ_{ni}, partial autocorrelation function (PACF) $\alpha(n)$, and relative innovation variance v_n as

$$\begin{aligned} \rho(n) &:= \operatorname{Corr}\left(X_{n+1}, X_1\right) = \gamma(n)/\gamma(0), && (7.6)\\ \phi_{ni} &:= \phi_{n+1,n+1-i}^{\{1,\dots,n\}}, \\ \alpha(n) &:= \operatorname{Corr}\left(X_{n+1} - P_{\{2,\dots,n\}} X_{n+1}, X_1 - P_{\{2,\dots,n\}} X_1\right), && (7.7)\\ v_n &:= \frac{\left|\left|X_{n+1} - P_{\{1,\dots,n\}} X_{n+1}\right|\right|^2}{\left|\left|X_{n+1}\right|\right|^2}. \end{aligned}$$

Given either $\rho(n)$ or $\alpha(n)$, the other quantities may be computed using the Durbin-Levinson algorithm (Durbin, 1960; Brockwell and Davis, 1987, Proposition 5.2.1).

Theorem 7.3 (Durbin-Levinson algorithm). *Some algorithm for computing $\alpha(n)$ given $\rho(n)$ is as follows. First, we set $v_0 = 1$ and then for $n = 1, 2, ...$ we iterate*

$$\alpha(n) = \left[\rho(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\rho(n-j)\right] \Big/ v_{n-1}, \tag{7.8}$$

$$\phi_{nj} = \phi_{n-1,j} - \alpha(n)\phi_{n-1,n-j}, \quad j \in \{1, ..., n-1\}, \tag{7.9}$$

$$\phi_{nn} = \alpha(n), \tag{7.10}$$

$$v_n = \left[1 - \alpha(n)^2\right] v_{n-1}. \tag{7.11}$$

Theorem 7.2 about the existence of a Gaussian process has its simpler analogue for the partial autocorrelation function, cf. Ramsey (1974); Schur (1917).

Theorem 7.4. *A stationary Gaussian process with densities (7.3) exists if and only if $\alpha(\cdot)$ computed from $\rho(\cdot) = \gamma(\cdot)/\gamma(0)$ using formulae (7.8)–(7.11) satisfies two conditions:*

1. *$|\alpha(m)| \le 1$ for all $m \ge 1$,*
2. *if $|\alpha(k)| = 1$ for some $k > 0$, then $\alpha(m)$ is not determined for $m > k$.*

Proof. Condition (7.4) is equivalent to (7.5). Using induction on n, we will show that this is equivalent to the requested conditions 1. and 2. on the PACF up to $m = n$. Assume that (7.5) holds for some n and any a_i whereas conditions 1. and 2. are satisfied up to $m = n$. Then

$$\begin{aligned} \mathbf{E}\left[\sum_{i=1}^{n+1} a_i X_i\right]^2 &= \mathbf{E}\left[\sum_{i=1}^{n} b_i X_i\right]^2 + a_{n+1}^2 \mathbf{E}\left[X_{n+1} - P_{\{1,\dots,n\}} X_{n+1}\right]^2 \\ &= \mathbf{E}\left[\sum_{i=1}^{n} b_i X_i\right]^2 + a_{n+1}^2 v_n. \end{aligned}$$

Hence the sum is nonnegative for any a_{n+1} if and only if $v_n \ge 0$. According to (7.11), this is equivalent to conditions 1. and 2. on the PACF up to $m = n + 1$.

A similar theorem can be formulated also for nonstationary processes (Dégerine and Lambert-Lacroix, 2003).

Partial autocorrelation function is not only useful to check whether a given stationary process exists but it can be also used to efficiently compute the block entropies of the process. Let us inspect the information-theoretic properties of a Gaussian process. We will denote the block entropy

$$H(n) := H(X_1^n) = H(\rho),$$

where ρ is given by (7.3). As proved by Cover and Thomas (2006, Theorem 8.4.1) we have:

Theorem 7.5. *For a Gaussian process with densities (7.3),*

$$H(n) = \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log \det \mathit{\Gamma}(n). \tag{7.12}$$

A straightforward corollary of the above theorem is this proposition:

Theorem 7.6. *For a pair of Gaussian random variables X and Y,*

$$I\big(X;Y\big) = -\frac{1}{2} \log \left[1 - \operatorname{Corr}(X,Y)^2\right]. \tag{7.13}$$

Proof. We have

$$\begin{aligned} I\big(X;Y\big) &= H\big(X\big) + H\big(Y\big) - H\big(X,Y\big) \\ &= \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log \operatorname{Var} X + \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log \operatorname{Var} Y \\ &\quad - \log(2\pi e) - \frac{1}{2}\log\left[(\operatorname{Var} X)(\operatorname{Var} Y) - \operatorname{Cov}(X,Y)^2\right] \\ &= -\frac{1}{2}\log\left[1 - \operatorname{Corr}(X,Y)^2\right]. \end{aligned}$$

As in the case of discrete-valued processes, we have Theorem 4.6. Now we are in a position to show that the second difference of the block entropy is a simple function of the partial autocorrelation.

Theorem 7.7. *For a stationary Gaussian process $(X_i)_{i=-\infty}^{\infty}$,*

$$I\big(X_1;X_n\big) = -\frac{1}{2}\log\left[1 - \rho(n-1)^2\right], \tag{7.14}$$

$$I\big(X_1;X_n|X_{2:n-1}\big) = -\mathit{\Delta}^2 H(n) = -\frac{1}{2}\log\left[1 - \alpha(n-1)^2\right]. \tag{7.15}$$

Proof. Equality (7.14) follows directly from (7.6) and (7.13). To demonstrate equality (7.15) let us notice that innovation $X_i - P_K X_i$ is independent of X_j, $j \in K$, whereas $P_K X_i$ is a function of X_j, $j \in K$. Since entropy $H(X_i + f((X_j)_{j\in K})|(X_j)_{j\in K})$ equals entropy $H(X_i|(X_j)_{j\in K})$ for any measurable function f then

$$H\big(X_i|(X_j)_{j\in K}\big) = H\big(X_i - P_K X_i|(X_j)_{j\in K}\big) = H\big(X_i - P_K X_i\big).$$

Hence

$$\begin{aligned} I\big(X_1;X_n|X_2^{n-1}\big) &= H\big(X_1|X_2^{n-1}\big) + H\big(X_n|X_2^{n-1}\big) - H\big(X_1,X_n|X_2^{n-1}\big) \\ &= H\big(X_1 - P_{\{2,\ldots n-1\}}X_1\big) + H\big(X_n - P_{\{2,\ldots n-1\}}X_n\big) \\ &\quad - H\big(X_1 - P_{\{2,\ldots n-1\}}X_1, X_n - P_{\{2,\ldots n-1\}}\big) \\ &= I\left(X_1 - P_{\{2,\ldots n-1\}}X_1; X_n - P_{\{2,\ldots n-1\}}\right). \end{aligned}$$

Because the innovations also have Gaussian distribution then (7.15) follows from (7.7) and (7.13).

By formulae (7.15) and (7.12) it follows that Durbin-Levinson algorithm is an efficient algorithm to compute the determinant of the autocovariance matrix. Indeed (7.15) and (7.8) imply

$$\begin{aligned} \Delta H(n) - H(1) &= \sum_{k=2}^{n} \Delta^2 H(k) = \frac{1}{2} \log v_{n-1} \le 0, \\ H(n) - nH(1) &= \sum_{k=1}^{n} [\Delta H(k) - H(1)] = \frac{1}{2} \sum_{k=1}^{n} \log v_{k-1} \le 0. \end{aligned}$$

On the other hand, from (7.12) we obtain

$$H(n) - nH(1) = \frac{1}{2} \left[\log \det \Gamma(n) - n \log \gamma(0)\right].$$

Hence $\det \Gamma(n) = \gamma(0)^n \prod_{k=1}^{n} v_{k-1} = \gamma(0)^n \prod_{k=1}^{n-1} \left[1 - |\alpha(k)|^2\right]^{n-k}$.

Moreover from Theorem 7.4 we obtain this statement:

Theorem 7.8. *For any concave real function $H(\cdot)$ such that $H(0) = 0$ there exists a stationary Gaussian process such that $H(\cdot)$ is its block entropy function.*

For a stationary Gaussian process let us define entropy rate h and excess entropy E via formulae (4.4) and (4.7). Because the analogues of Theorems 4.7 and 4.9 also hold in the Gaussian case, these two quantities can be related to the following two concepts researched in time series theory (Brockwell and Davis, 1987; McLeod, 1998):

Definition 7.4 (innovation variance). *The innovation variance of a stationary Gaussian process is defined as*

$$\sigma^2 := \lim_{n\to\infty} \left(\det \Gamma(n)\right)^{1/n} = \frac{1}{2\pi e} \exp(2h).$$

Definition 7.5 (generalized variance). *The generalized variance of a stationary Gaussian process is defined as*

$$g = \lim_{n\to\infty} \frac{\det \Gamma(n)}{\sigma^{2n}} = \exp(2E).$$

There exist alternative formulae for the innovation variance and the generalized variance in terms of spectral density.

Theorem 7.9 (Herglotz theorem). *For any autocovariance function γ, there exists a unique finite measure F on $I = (-\pi, \pi]$ such that*

$$\gamma(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(ik\omega)\, \mathrm{d}F(\omega). \tag{7.16}$$

Moreover, for any finite measure F, function γ given by (7.16) is an autocovariance function of a certain stationary process (cf. Brockwell and Davis, 1987, Theorem 4.3.1).

Measure F is called the spectral measure of the process. By the Lebesgue-Radon-Nikodym theorem (Theorem 1.1), any measure F on $I = (-\pi, \pi]$ can be decomposed as

$$F = F_{\perp} + F_{\ll},$$

where measure $F_{\perp}$ is mutually singular with the Lebesgue measure m on $I = (-\pi, \pi]$ and $F_{\ll}$ is absolutely continuous with respect to m. According to the second part of the Lebesgue-Radon-Nikodym theorem, there exists a measurable function f such that

$$F_{\ll}(A) = \int_A f(\omega)\,\mathrm{d}\omega.$$

For the spectral measure F, function f is called spectral density.

Theorem 7.10 (Kolmogorov formula). *Let f be the spectral density. The innovation variance equals*

$$\sigma^2 = \begin{cases} \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\omega)\,\mathrm{d}\omega\right], & \int_{-\pi}^{\pi} |\log f(\omega)|\,\mathrm{d}\omega < \infty, \\ 0, & \textit{else} \end{cases}$$

(cf. Grenander and Szegő, 1958, Section 5.2).

Theorem 7.11. *Suppose that $F_{\perp} = 0$, spectral density satisfies $f(\omega) \in [a, b] \subset (0, \infty)$ for all $\omega \in (-\pi, \pi]$ and there exists the derivative f' which satisfies Lipschitz condition $|f'(\omega_1) - f'(\omega_2)| < \text{const}\,|\omega_1 - \omega_2|^{\alpha}$ for $0 < \alpha < 1$. Then the generalized variance equals*

$$g = \exp\left[\frac{1}{\pi}\int_{|z|<1} \left|\frac{d}{dz}\log\psi(z)\right|^2 \mathrm{d}(\operatorname{Re} z)\,\mathrm{d}(\operatorname{Im} z)\right],$$

where function ψ is given by

$$\psi(z) = \exp\left[\frac{1}{2}h(z)\right], \quad h(z) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i\omega} + z}{e^{i\omega} - z}\log f(\omega)\,\mathrm{d}\omega \tag{7.17}$$

for $|z| < 1$ (cf. Grenander and Szegő, 1958, Sections 1.1, 1.13, 1.14, 5.5).

Exercises

1. Derive the Durbin-Levinson algorithm (7.8)–(7.11).
2. Derive the Yule-Walker equations

$$\sum_{i=1}^{n} \phi_{ni}\rho(k - i) = \rho(k), \quad n \geq 1, \quad k \in \{1, ..., n\}. \tag{7.18}$$

3. We say that a Gaussian process $(X_i)_{i=-\infty}^{\infty}$ has a moving average representation MA(∞) if
$$X_i = \sum_{k=0}^{\infty} \psi_k Z_{i-k},$$
where $\sum_{k=0}^{\infty} |\psi_k| < \infty$ and $(Z_i)_{i=-\infty}^{\infty}$ is a Gaussian white noise, i.e., $\mathbf{E}\, Z_i = 0$ and $\mathbf{E}\, [Z_i Z_j] = \mathbf{1}\{i = j\}$. Compute the autocorrelation function of an MA(∞) process.
4. Let $X_i = aZ_i + bZ_{i-1}$, where $(Z_i)_{i=-\infty}^{\infty}$ is a Gaussian white noise. Compute the ACF and PACF.
5. Process $(X_i)_{i=-\infty}^{\infty}$ is called exchangeable if vectors of variables $(X_{i_1}, ..., X_{i_n})$ have a distribution that depends only on n. By definition, all exchangeable processes are stationary. Gaussian exchangeable processes have autocorrelation function
$$\rho(n) = \begin{cases} 1, & n = 0, \\ p, & n \neq 0, \end{cases}$$
where $0 \leq p \leq 1$. Show that in that case we have
$$\alpha(n) = \phi_{ni} = \frac{p}{(n-1)p + 1}.$$
6. Show that for $d \in (-\infty, 1/2)$ there exists a weakly stationary process, called ARIMA$(0, d, 0)$, that has parameters
$$\rho(n) = \prod_{i=1}^{n} \frac{i + d - 1}{i - d}, \tag{7.19}$$
$$\phi_{nk} = -\binom{n}{k} \frac{(k - d - 1)!(n - d - k)!}{(-d - 1)!(n - d)!}, \tag{7.20}$$
$$\alpha(n) = \frac{d}{n - d}, \tag{7.21}$$
where $z! := \Gamma(z + 1)$. Moreover, as for the asymptotics of the ACF, we have $\lim_n \rho_n / n^{-1+2d} = (-d)!/(d - 1)!$ if $-d + 1 \notin \mathbb{N}$. We also have $\sum_{k=-\infty}^{\infty} \rho_k = \infty$ for $d \in (0, 1/2)$ and $\sum_{k=-\infty}^{\infty} \rho_k = 0$ for $d < 0$.
7. Let $X_i = Y \cos(i\lambda) + Z \sin(i\lambda)$, where Y and Z are independent Gaussian variables with zero means and unit variances. Show that process $(X_i)_{i=-\infty}^{\infty}$ is stationary. What is its spectral measure?
8. Show that there exists a nonstationary Gaussian process $(X_i)_{i=0}^{\infty}$ such that $\mathbf{E}\, X_i = 0$ and $\mathbf{E}\, [X_i X_j] = \min(i, j)$.

8

Sufficient statistics

Families of probability distributions. Sufficient statistic. Minimal sufficient statistic. Exponential families. Basu theorem.

In this chapter we start the discussion of mathematical statistics and its links with information theory. Mathematical statistics is focused on problems inverse to probability theory. The distinction is as follows. We toss a coin n times. A typical problem in probability theory is to compute the probability of tossing $n/2$ heads if the probability of the head in a single toss is known to be $1/2$. On the other hand, a typical problem of statistics is to estimate the probability of the head in a single toss if we obtained $n/2$ heads in the sample of n tosses.

Two fundamental notions of statistics are a parametric family of distributions and a statistic, called also an estimator. A parametric family of distributions specifies a statistical problem, whereas estimators offer its possible solutions.

Definition 8.1 (parametric family). *A* parametric family of distributions *is a family of probability distributions indexed by parameter $\theta \in \Theta$, which specify probabilities of a stochastic process $(X_i)_{i=-\infty}^{\infty}$. For discrete variables we write these distributions as*

$$P\big(X_1^n = x_1^n|\theta\big).$$

For real variables, we assume that there exists a probability density function $\rho(x_1^n|\theta)$ which satisfies

$$P\big(X_1^n \in A|\theta\big) = \int_A \rho(x_1^n|\theta)\,\mathrm{d}x_1^n,$$

where $\int \mathrm{d}x_1^n$ is the integral with respect to the n-dimensional Lebesgue measure.

Usually, parameter θ is a single real number or a vector of real numbers. It is also usually assumed that variables X_i are probabilistically independent (given the parameter θ). In that case, we call X_1^n a *random sample* of length n drawn from distribution $P(X_i = x_i|\theta)$ or $\rho(x_i|\theta)$, respectively. The first case will be called a *discrete random sample*, whereas the second will be called a *real random sample.*

Example 8.1. A random sample of length n drawn from Bernoulli distributions with success probability θ has probability distribution

$$P\big(X_1^n = x_1^n|\theta\big) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i},$$

where $x_i \in \{0,1\}$ and $\theta \in (0,1)$.

Example 8.2. A random sample of length n drawn from normal (or Gauss) distributions with expectation μ and variance σ^2 has probability density

$$\rho(x_1^n|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right],$$

where $x_i \in (-\infty,\infty)$, $\mu \in (-\infty,\infty)$, and $\sigma \in (0,\infty)$.

The theoretical problem of mathematical statistics is to point out efficient ways of estimating parameter θ given the random sample X_1^n. Any such method is a function of X_1^n. Thus we arrive at the concept of a statistic.

Definition 8.2 (statistic). *Any function $T(X_1^n)$ of the random sample X_1^n is called a* statistic*.*

A statistic the aim of which is to approximate the unknown parameter θ is called an *estimator*. The distinction between a statistic and an estimator is largely informal. We will grasp the difference on a few examples that will be presented later.

Before we study particular estimators it is advised to devote some consideration to statistics in general. The fundamental problem is whether a statistic entails all information about the parameter contained in the random sample and whether there is a statistic that does it in the most efficient way.

Definition 8.3 (sufficient statistic). *We say that statistic $T(X_1^n)$ is* sufficient *if the conditional distribution of sample X_1^n given $T(X_1^n)$ does not depend on θ. That is, for discrete X_1^n we require*

$$P\big(X_1^n = x_1^n|T(X_1^n)=t,\theta\big) = P\big(X_1^n = x_1^n|T(X_1^n)=t\big) \tag{8.1}$$

for $P(T(X_1^n)=t|\theta) \neq 0$. An analogous statement is required for real variables X_1^n using properly defined conditional densities.

Informally speaking, statistics is sufficient if it contains all information about the parameter carried by the sample. The information-theoretic sense of condition (8.1) becomes clear if we adopt the view of Bayesian statistics, which assumes that the parameter is a random variable Θ with a certain distribution $P(\Theta=\theta)$, called the prior. (In non-Bayesian statistics there is no prior distribution on the values of θ.)

Theorem 8.1. *Assume that X_1^n and Θ are discrete random variables. Condition*

$$P\big(X_1^n = x_1^n|T(X_1^n)=t,\Theta=\theta\big) = P\big(X_1^n = x_1^n|T(X_1^n)=t\big) \tag{8.2}$$

for $P(T(X_1^n)=t,\Theta=\theta) \neq 0$ is equivalent to

$$I\big(T(X_1^n);\Theta\big) = I\big(X_1^n;\Theta\big).$$

Remark: For any statistic, $T(X_1^n)$ and Θ are conditionally independent given the sample X_1^n. Hence $I(T(X_1^n);\Theta) \le I(X_1^n;\Theta)$ holds by the data-processing inequality. For a sufficient statistic, Θ and X_1^n are also conditionally independent given $T(X_1^n)$.

Proof. Conditional independence (8.2) is equivalent to $I(X_1^n;\Theta|T(X_1^n)) = 0$. Because $H(T(X_1^n)|X_1^n) = 0$ then, by the Venn diagram for three variables on page 24, $I(X_1^n;\Theta|T(X_1^n)) = 0$ is equivalent to $I(T(X_1^n);\Theta) = I(X_1^n;\Theta)$.

A convenient characterization of the sufficient statistic is given by the following proposition. This proposition can be used to effectively check whether a given statistic is sufficient both in the discrete and real case.

Theorem 8.2 (Fisher factorization theorem). *Statistic T is sufficient if and only if there exist functions g and h such that for discrete X_1^n we have for all x_1^n*

$$P\big(X_1^n = x_1^n|\theta\big) = h(x_1^n)g(\theta, T(x_1^n)), \tag{8.3}$$

whereas for real X_1^n we have for almost all x_1^n

$$\rho(x_1^n|\theta) = h(x_1^n)g(\theta, T(x_1^n)).$$

Proof. We give only the proof for the discrete case. The proof for the real case can be found in Keener (2010, Section 6.4). Let $t = T(x_1^n)$. Assume (8.3). Then

$$\begin{aligned} &P\big(X_1^n = x_1^n|T(X_1^n) = t, \theta\big) \\ &\qquad = \frac{P(X_1^n = x_1^n|\theta)}{\sum_{y_1^n:T(y_1^n)=t} P(X_1^n = y_1^n|\theta)} = \frac{h(x_1^n)}{\sum_{y_1^n:T(y_1^n)=t} h(y_1^n)} \end{aligned}$$

does not depend on θ. On the other hand, if T is sufficient, we may put $h(x_1^n) = P(X_1^n = x_1^n|T(X_1^n) = t)$ and $g(\theta, t) = P(T(X_1^n) = t|\theta)$.

Now we can exhibit sufficient statistics for the pair of parametric families introduced previously.

Example 8.3. Consider a random sample of length n drawn from Bernoulli distribution. We have

$$P\big(X_1^n = x_1^n|\theta\big) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \tag{8.4}$$

so $\sum_{i=1}^n x_i$ is a sufficient statistic.

Example 8.4. Consider a random sample of length n drawn from normal distribution. We have

$$\begin{aligned} \rho(x_1^n|\mu,\sigma) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{\mu\sum_{i=1}^n x_i}{\sigma^2} - \frac{\mu^2 n}{2\sigma^2}\right], \end{aligned}$$

so the pair $\big(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\big)$ is a sufficient statistic.

We can ask how much can we compress the information about the parameter. The following concept sets the limit of this compression.

Definition 8.4 (minimal sufficient statistics). *A sufficient statistic $T(X_1^n)$ is called* minimal *if for any other sufficient statistic $T'(X_1^n)$ there exists a function f such that $T(X_1^n) = f(T'(X_1^n))$.*

Like in the case of sufficiency, there is a theorem which allows to verify minimal sufficiency effectively.

Theorem 8.3. *T is the minimal sufficient statistic, for discrete X_1^n, if for each x_1^n and y_1^n,*

$$\frac{P(X_1^n = x_1^n|\theta)}{P(X_1^n = y_1^n|\theta)} \text{ does not depend on } \theta \iff T\big(x_1^n\big) = T\big(y_1^n\big),$$

whereas, for real X_1^n, if for almost all x_1^n and y_1^n,

$$\frac{\rho(x_1^n|\theta)}{\rho(y_1^n|\theta)} \text{ does not depend on } \theta \iff T\big(x_1^n\big) = T\big(y_1^n\big),$$

Proof. We give only the proof for the discrete case. The proof for the real case is analogous. First, we show that T is a sufficient statistic. For any value t we fix a sample $y_1^n(t)$ which achieves $T(y_1^n(t)) = t$. For an arbitrary x_1^n, let $T(x_1^n) = t$. Then we have

$$P\big(X_1^n = x_1^n|\theta\big) = \frac{P(X_1^n = x_1^n|\theta)}{P(X_1^n = y_1^n(t)|\theta)} P\big(X_1^n = y_1^n(t)|\theta\big) = h(x_1^n)g(\theta, T(x_1^n)).$$

Next, we will show that T is a function of any other statistic T'. Let x_1^n and y_1^n satisfy $T'(x_1^n) = T'(y_1^n)$. Since

$$\frac{P(X_1^n = x_1^n|\theta)}{P(X_1^n = y_1^n|\theta)} = \frac{h'(x_1^n)g'(\theta, T'(x_1^n))}{h'(y_1^n)g'(\theta, T'(y_1^n))} = \frac{h'(x_1^n)}{h'(y_1^n)}$$

does not depend on θ, we have $T(x_1^n) = T(y_1^n)$. Hence T is a function of T'.

Thus we may show that the previously exhibited sufficient statistics are minimal.

Example 8.5. Consider a random sample of length n from Bernoulli distribution. We have

$$\frac{P(X_1^n = x_1^n|\theta)}{P(X_1^n = y_1^n|\theta)} = \theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i} (1-\theta)^{\sum_{i=1}^n y_i - \sum_{i=1}^n x_i},$$

so $\sum_{i=1}^n x_i$ is a minimal sufficient statistic.

Example 8.6. Consider a random sample of length n drawn from normal distribution. We have

$$\frac{\rho(x_1^n|\mu,\sigma)}{\rho(y_1^n|\mu,\sigma)} = \exp\left[-\frac{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2}{2\sigma^2} + \frac{\mu\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)}{\sigma^2}\right],$$

so the pair $\left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)$ is a minimal sufficient statistic.

For a given parametric family, there may exist more than one minimal sufficient statistic. For example, for the sample drawn from the normal distribution not only $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is a minimal sufficient statistic but also $(\bar{X}_n, S_n^2)$, where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. As we will see in the next chapter, pair $(\bar{X}_n, S_n^2)$ is the maximum likelihood estimator of (μ, σ^2).

An important class of parametric families that admit particularly simple sufficient statistics are exponential families. They include Bernoulli and normal distributions in particular. As we will show later, exponential families also provide a link between statistics and information theory because they maximize entropy given certain linear constraints. The formal definition is as follows. (Symbol ln denotes the natural logarithm.)

Definition 8.5 (exponential family). *In the discrete case, let function $p : \mathbb{X} \to (0,\infty)$ satisfy $\sum_{x\in\mathbb{X}} p(x) < \infty$. Having functions $T_l : \mathbb{X} \to \mathbb{R}$, $l = 1, 2, ..., s$, we denote the canonical sum*

$$Z(\theta) = \sum_{x\in\mathbb{X}} p(x) \exp\left(\sum_{l=1}^{s} \theta_l T_l(x)\right) \tag{8.5}$$

and define s-parameter exponential family

$$P\big(X_1^n = x_1^n|\theta\big) = \prod_{i=1}^{n} p(x_i) \exp\left(\sum_{l=1}^{s} \theta_l T_l(x_i) - \ln Z(\theta)\right) \tag{8.6}$$

for $\theta = (\theta_1, \theta_2, ..., \theta_s) \in \Theta := \{\omega \in \mathbb{R}^s : Z(\omega) < \infty\}$. In the real case, let function $p : \mathbb{R} \to (0,\infty)$ satisfy $\int p(x)\,\mathrm{d}x < \infty$. Having functions $T_l : \mathbb{R} \to \mathbb{R}$, $l = 1, 2, ..., s$, we denote the canonical sum

$$Z(\theta) = \int p(x) \exp\left(\sum_{l=1}^{s} \theta_l T_l(x)\right) \mathrm{d}x$$

and define s-parameter exponential family

$$\rho(x_1^n|\theta) = \prod_{i=1}^{n} p(x_i) \exp\left(\sum_{l=1}^{s} \theta_l T_l(x_i) - \ln Z(\theta)\right)$$

for $\theta = (\theta_1, \theta_2, ..., \theta_s) \in \Theta := \{\theta' \in \mathbb{R}^s : Z(\theta') < \infty\}$. The s-parameter exponential family is called of full rank *if the interior of Θ is not empty and T_l do not satisfy a linear constraint of the form $\sum_{l=1}^{s} a_l T_l(x_i) = c$ for a constant c.*

Example 8.7. Bernoulli distributions form an exponential family because

$$\begin{aligned} P(X_1^n = x_1^n|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \prod_{i=1}^{n} \exp\left(x_i \ln\frac{\theta}{1-\theta} + \ln(1-\theta)\right) \\ &= \prod_{i=1}^{n} \exp\big(\eta x_i - \ln Z(\eta)\big), \end{aligned}$$

where $\eta = \ln \dfrac{\theta}{1-\theta}$ and $Z(\eta) = 1 - \theta$. Function $\eta = \eta(\theta)$ is called the logit function.

Example 8.8. Normal distributions form an exponential family because

$$\begin{aligned}\rho(x_1^n|\mu,\sigma) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x_i^2}{2\sigma^2} + \frac{\mu x_i}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] \\ &= \prod_{i=1}^n \exp\left(\alpha x_i^2 + \beta x_i - \ln Z(\alpha,\beta)\right),\end{aligned}$$

where $\alpha = -\frac{1}{2\sigma^2}$, $\beta = \frac{\mu}{\sigma^2}$, and $Z(\alpha,\beta) = \sigma\sqrt{2\pi}\exp\left[\frac{\mu^2}{2\sigma^2}\right]$.

Now we can exhibit the minimal sufficient statistic for an exponential family.

Theorem 8.4. *Consider the full rank exponential family as defined in Definition 8.5. The minimal sufficient statistic is*

$$T(x_1^n) = \left(\sum_{i=1}^n T_1(x_i), \sum_{i=1}^n T_2(x_i), ..., \sum_{i=1}^n T_s(x_i)\right). \tag{8.7}$$

Proof. We give only the proof for the discrete case since the proof for the real case is analogous. We have

$$\frac{P(X_1^n = x_1^n|\theta)}{P(X_1^n = y_1^n|\theta)} = \frac{\prod_{i=1}^n p(x_i)}{\prod_{i=1}^n p(y_i)} \exp\left(\sum_{l=1}^s \theta_l \Big(\sum_{i=1}^n T_l(x_i) - \sum_{i=1}^n T_l(y_i)\Big)\right).$$

We observe that this expression does not depend on θ_l if and only if $T(x_1^n) = T(y_1^n)$. Hence $T(x_1^n)$ is the minimal sufficient statistic.

The last result we will present in this chapter is the Basu theorem, a handy proposition for proving independence of two statistics. Let us denote the expectation

$$\mathbf{E}_\theta Y := \int Y \,\mathrm{d}P(\cdot|\theta),$$

which equals $\mathbf{E}_\theta Y = \sum_y yP(Y = y|\theta)$ in the discrete case. The conditional expectation of Y given X will be denoted $\mathbf{E}_\theta[Y|X]$. For the conditional expectation we have the smoothing identity $\mathbf{E}_\theta[\mathbf{E}_\theta[Y|X]] = \mathbf{E}_\theta Y$, see (1.3).

Definition 8.6 (complete statistic). *A statistic $T(X_1^n)$ is called* complete *if for any function s the following implication holds*

$$\boldsymbol{E}_\theta s\big(T(X_1^n)\big) = 0 \textit{ for all } \theta \implies P\big(s(T(X_1^n)) = 0|\theta\big) = 1 \textit{ for all } \theta.$$

Theorem 8.5. *If a statistic $T(X_1^n)$ is complete and sufficient then it is minimal sufficient.*

Proof. Let $\tilde{T}(X_1^n)$ be a minimal sufficient statistic. Then $\tilde{T}(X_1^n) = f(T(X_1^n))$. Define $g(\tilde{T}(X_1^n)) = \mathbf{E}_\theta\big[T(X_1^n)|\tilde{T}(X_1^n)\big]$. Function $g(\tilde{T}(X_1^n))$ is independent of θ since $\tilde{T}(X_1^n)$ is sufficient. Taking the expectation of $g(\tilde{T}(X_1^n))$, we obtain $\mathbf{E}_\theta g(\tilde{T}(X_1^n)) = \mathbf{E}_\theta T(X_1^n)$ so $\mathbf{E}_\theta\big[T(X_1^n) - g(\tilde{T}(X_1^n))\big] = 0$. But $T(X_1^n) - g(\tilde{T}(X_1^n))$ is a function of $T(X_1^n)$, so by completeness of $T(X_1^n)$ we obtain $T(X_1^n) = g(\tilde{T}(X_1^n))$ with $P(\cdot|\theta)$-probability 1. Hence there is a one-to-one correspondence between $T(X_1^n)$ and $\tilde{T}(X_1^n)$ and $T(X_1^n)$ must be also minimal sufficient.

Theorem 8.6. *Consider the full rank exponential family as defined in Definition 8.5. Statistic $T(x_1^n)$ given by (8.7) is complete. (See Barndorff-Nielsen, 1978, Lemma 8.2.)*

Example 8.9. For the sample drawn from the normal distribution with an unknown mean μ and a fixed variance σ, statistic $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is complete and sufficient.

Definition 8.7 (ancillary statistic). *A statistic $T(X_1^n)$ is called* ancillary *if its distribution does not depend on θ.*

Example 8.10. For the sample drawn from the normal distribution with an unknown mean μ and a fixed variance σ, statistic $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is ancillary.

Theorem 8.7 (Basu theorem). *If a statistic $T(X_1^n)$ is complete and sufficient whereas a statistic $S(X_1^n)$ is ancillary then $T(X_1^n)$ and $S(X_1^n)$ are independent.*

Proof. For simplicity we present the proof only for discrete $T = T(X_1^n)$ and $S = S(X_1^n)$. We may write

$$P\big(S = s|\theta\big) = \sum_t P\big(S = s|T = t, \theta\big)P\big(T = t|\theta\big)$$

Probability $P(S = s|\theta) = P(S = s)$ does not depend on θ by ancillarity, whereas $P(S = s|T = t, \theta) = P(S = s|T = t)$ does not depend on θ by sufficiency. We may write thus

$$\sum_t \big[P\big(S = s|T = t\big) - P\big(S = s\big)\big]P\big(T = t|\theta\big) = 0.$$

Quantity $[P(S = s|T = t) - P(S = s)]$ is a random variable which is a function of t and not of θ. Hence by completeness we obtain that $P(S = s|T = t) = P(S = s)$ for all t such that $P(T = t|\theta) > 0$. That is the requested claim.

Example 8.11. For the sample drawn from the normal distribution with an unknown mean μ and a fixed variance σ, statistics $\bar{X}_n$ and S_n^2 are independent.

Exercises

1. Consider a random sample X_1^n drawn from the uniform distribution
$$\rho(x_i|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x_i < \theta, \\ 0, & \text{else.} \end{cases}$$
Find a one-dimensional sufficient statistic.
2. Show that Poisson distributions
$$P(X_1^n = x_1^n|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \tag{8.8}$$
where $x_i \in \mathbb{N} \cup \{0\}$ and $\lambda > 0$, form an exponential family.
3. Show that geometric distributions
$$P\big(X_1^n = x_1^n|p\big) = \prod_{i=1}^{n} (1-p)^{x_i} p, \tag{8.9}$$
where $x_i \in \mathbb{N} \cup \{0\}$ and $0 < p \le 1$, constitute an exponential family.
4. Show that negative binomial distributions
$$P\big(X_1^n = x_1^n|r, p\big) = \prod_{i=1}^{n} \binom{k+r-1}{k} (1-p)^r p^{x_i} \tag{8.10}$$
where $x_i \in \mathbb{N} \cup \{0\}$, $r > 0$, and $0 < p < 1$, form an exponential family.
5. Show that gamma distributions
$$\rho(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{8.11}$$
where $x \ge 0$ and $\alpha, \beta > 0$, constitute an exponential family.
6. Show that beta distributions
$$\rho(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1},$$
where $x \in (0,1)$ and $\alpha, \beta > 0$, form an exponential family.
7. Suppose X_1^n are a sample from a beta distribution. Find the minimal sufficient statistic if $\alpha = 2\beta$.
8. Show that Pareto distributions
$$\rho(x|\alpha) = \begin{cases} \dfrac{\alpha}{x^{\alpha+1}}, & x \ge 1, \\ 0, & x < 1, \end{cases} \tag{8.12}$$
where $\alpha > 0$, form an exponential family.

9. Show that inverse Gauss distributions

$$\rho(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right],$$

where $x > 0$ and $\mu, \lambda > 0$, constitute an exponential family.

10. Let X_1^n be a random sample drawn from distribution

$$\rho(x_i|\theta) = \begin{cases} \frac{2x_i}{\theta^2}, & 0 \le x_i < \theta, \\ 0, & \text{else.} \end{cases}$$

   (a) Find a one-dimensional sufficient statistic T.
   (b) Determine the density of T.
   (c) Show that T is complete.

11. Consider a sample drawn from the normal distribution with an unknown mean μ and a fixed variance σ. Show that statistic $S_n^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is ancillary.

12. Let X_1^n be a sample drawn from exponential distribution

$$\rho(x_i|\beta) = \beta e^{-\beta x_i},$$

where $x_i \ge 0$ and $\beta > 0$.
   (a) Find the density of $Y_i = \beta X_i$.
   (b) Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Show that $\bar{X}_n$ and $\sum_{i=1}^n X_i^2 / \bar{X}_n^2$ are independent.

13. Consider $\mathbf{E}\, X^k = \sum_x x^k P(X = x)$, the k-th moment of variable X. Show that $\mathbf{E}\, X \in \left[-\sqrt{\mathbf{E}\, X^2}, \sqrt{\mathbf{E}\, X^2}\right]$.
Hint: Use Jensen inequality.

9

Estimators

Maximum likelihood estimator. Consistent estimators. Fisher information. Cramér-Rao inequality.

In the previous chapter, an arbitrary function of the data was called a statistic, whereas a statistic that is designed to approximate the unknown parameter was called an estimator. An important example of an estimator is the maximum likelihood estimator.

Definition 9.1 (maximum likelihood estimator). *Let* $\operatorname{argmax}_{\theta\in\Theta} f(\theta)$ *be the argument* $\theta \in \Theta$ *for which the function* f *attains the maximal value. The* maximum likelihood estimator (MLE) *is the statistic defined for discrete random sample* x_1^n *as*

$$\theta_{\mathrm{ML}}(x_1^n) := \operatorname{argmax}_{\theta\in\Theta} P\big(X_1^n = x_1^n|\theta\big), \tag{9.1}$$

whereas for real random sample x_1^n *it is defined as*

$$\theta_{\mathrm{ML}}(x_1^n) := \operatorname{argmax}_{\theta\in\Theta} \rho(x_1^n|\theta).$$

The name MLE is motivated by that probability $P(X_1^n = x_1^n|\theta)$ or density $\rho(x_1^n|\theta)$ as a function of parameter θ are called the likelihood function.

In some simple cases we can evaluate the maximum likelihood estimator analytically.

Example 9.1. Consider a sample of length n drawn from Bernoulli distribution. We have (8.4). Thus

$$0 = \left.\frac{\partial \ln P(X_1^n = x_1^n|\theta)}{\partial\theta}\right|_{\theta=\theta_{\mathrm{ML}}} = \frac{\sum_{i=1}^n x_i}{\theta_{\mathrm{ML}}} - \frac{n - \sum_{i=1}^n x_i}{1-\theta_{\mathrm{ML}}}.$$

Hence

$$\theta_{\mathrm{ML}}(x_1^n) = \frac{1}{n}\sum_{i=1}^n x_i.$$

Example 9.2. Consider a sample of length n drawn from normal distribution. We have

$$\ln \rho(x_1^n|\mu,\sigma) = -\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\ln(2\pi\sigma^2).$$

Thus

$$\begin{aligned} 0 &= \left.\frac{\partial \ln \rho(x_1^n|\mu,\sigma)}{\partial \mu}\right|_{\mu=\mu_{\mathrm{ML}},\sigma=\sigma_{\mathrm{ML}}} = \sum_{i=1}^n \frac{2(x_i-\mu_{\mathrm{ML}})}{2\sigma_{\mathrm{ML}}^2}, \\ 0 &= \left.\frac{\partial \ln \rho(x_1^n|\mu,\sigma)}{\partial \sigma}\right|_{\mu=\mu_{\mathrm{ML}},\sigma=\sigma_{\mathrm{ML}}} = \sum_{i=1}^n \frac{(x_i-\mu_{\mathrm{ML}})^2}{\sigma_{\mathrm{ML}}^3} - \frac{n}{\sigma_{\mathrm{ML}}}. \end{aligned}$$

Hence

$$\begin{aligned} \mu_{\mathrm{ML}}(x_1^n) &= \frac{1}{n}\sum_{i=1}^n x_i, \\ \sigma_{\mathrm{ML}}^2(x_1^n) &= \frac{1}{n}\sum_{i=1}^n (x_i - \mu_{\mathrm{ML}}(x_1^n))^2. \end{aligned} \tag{9.2}$$

Now we will discuss a few ideas how to understand the quality of an estimator. It seems plausible to accept that the deviation of a good estimator from the true parameter should converge to zero for the sample size tending to infinity. Thus consistency is a typically required condition.

Definition 9.2 (consistent estimator). *Estimator $T(X_1^n)$ is called* consistent in probability *if $T(X_1^n)$ converges to θ in probability, i.e., for each $\epsilon > 0$ we have*

$$\lim_{n\to\infty} P\big(|T(X_1^n) - \theta| > \epsilon|\theta\big) = 0.$$

Applying Kullback-Leibler divergence we can prove that the maximum likelihood estimator is indeed consistent. First we will exhibit the proof for a discrete parameter, which is simpler. We denote the Kullback-Leibler divergence as

$$D(\theta||\omega) = \sum_{x\in\mathbb{X}} P\big(X_i = x|\theta\big) \ln \frac{P(X_i = x|\theta)}{P(X_i = x|\omega)}.$$

In the formula above we use the natural logarithm rather than the binary one since it appears more convenient in statistical applications.

Theorem 9.1. *Let X_1^n be a random sample drawn from distribution $P(X_i = x_i|\theta)$ or $\rho(x_i|\theta)$, where θ takes values in a finite set Θ. If $\theta \neq \theta'$ implies $P(X_i|\theta) \neq P(X_i|\theta')$ or $\rho(\cdot|\theta) \neq \rho(\cdot|\theta')$ then the maximum likelihood estimator is consistent.*

Proof. We give only the proof for discrete X_1^n since the proof for real X_1^n is analogous. Let $\Theta = \{\theta, \theta_1, ..., \theta_m\}$. Assume that X_1^n is a random sample drawn from distribution $P(X_i|\theta)$. By the strong law of large numbers (1.4), we have with probability 1 that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \ln \frac{P(X_i|\theta)}{P(X_i|\theta_k)} = D(\theta||\theta_k) > 0$$

since $P(X_i|\theta) \neq P(X_i|\theta_k)$. Denoting $L_n(\theta) = P(X_1^n|\theta)$, we obtain

$$P\big(L_n(\theta) \le L_n(\theta_k)|\theta\big) = P\left(\frac{1}{n}\sum_{i=1}^{n} \ln \frac{P(X_i|\theta)}{P(X_i|\theta_k)} \le 0 \,\middle|\, \theta\right) \xrightarrow{n\to\infty} 0.$$

Using this result we derive

$$\begin{aligned} P\big(\theta_{\mathrm{ML}}(X_1^n) \neq \theta|\theta\big) &\le P\big(L_n(\theta) \le L_n(\theta_k) \text{ for some } k = 1, ..., m|\theta\big) \\ &\le \sum_{k=1}^{m} P\big(L_n(\theta) \le L_n(\theta_k)|\theta\big) \xrightarrow{n\to\infty} 0. \end{aligned}$$

Hence the maximum likelihood estimator is consistent.

To discuss consistency of the maximum likelihood estimator for a real parameter, we need to introduce a generalization of the Kullback-Leibler divergence in which the second argument is a set of parameters. For simplicity, let us restrict ourselves to discrete random variables and put

$$D(\theta||S) = \sum_{x\in\mathbb{X}} P(X_i = x|\theta) \inf_{\omega\in S} \ln \frac{P(X_i = x|\theta)}{P(X_i = x|\omega)}.$$

The idea of the theorem on consistency for a real parameter rests on the fact that $D(\theta||S)$ is usually positive if S is sufficiently small.

Theorem 9.2 (Wald theorem). *Let X_1^n be a random sample drawn from distribution $P(X_i = x_i|\theta)$, where θ takes values in the set of real numbers $\Theta = \mathbb{R}$. Suppose that for any θ the following conditions hold:*

1. *For every $\omega \neq \theta$, there is a neighborhood S_ω of ω such that $D(\theta||S_\omega) > 0$.*
2. *For some constant $a > 0$, $D(\theta||\{\omega : |\omega - \theta| > a\}) > 0$.*

Then the maximum likelihood estimator is consistent.

Remark: A similar result may be formulated for a real random sample. Moreover conditions 1–2 may be reformulated so that checking them is easier for particular parametric families. This, however, involves some additional technicalities which we want to avoid for brevity of exposition.

Proof. Assume that X_1^n is a random sample drawn from distribution $P(X_i|\theta)$. By the strong law of large numbers (1.4), we have with probability 1 that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \inf_{\omega\in S} \ln \frac{P(X_i|\theta)}{P(X_i|\omega)} = D(\theta||S).$$

Denoting $L_n(\theta) = P(X_1^n|\theta)$, we obtain for $D(\theta||S) > 0$ that

$$\begin{aligned} &P\left(L_n(\theta) \le \sup_{\omega\in S} L_n(\omega) \,\middle|\, \theta\right) = P\left(\inf_{\omega\in S} \frac{1}{n}\sum_{i=1}^{n} \ln \frac{P(X_i|\theta)}{P(X_i|\omega)} \le 0 \,\middle|\, \theta\right) \\ &\le P\left(\frac{1}{n}\sum_{i=1}^{n} \inf_{\omega\in S} \ln \frac{P(X_i|\theta)}{P(X_i|\omega)} \le 0 \,\middle|\, \theta\right) \xrightarrow{n\to\infty} 0. \end{aligned}$$

To use the above result, let $\epsilon > 0$ and define set

$$A = \{\omega : \epsilon \le |\omega - \theta| \le a\}.$$

Set A is compact (for it is closed and bounded) and the family of sets $\{S_\omega : \omega \in A\}$ forms an open cover of A. From the definition of compactness, there must exist a finite subfamily $\{S_1, ..., S_m\}$ that covers A, i.e., $A \subset \bigcup_{k=1}^m S_k$. Let us define $S_0 = \{\omega : |\omega - \theta| > a\}$ so

$$\{\omega : |\omega - \theta| \ge \epsilon\} \subset \bigcup_{k=0}^{m} S_k.$$

In the following we will use the fact that $D(\theta||S_k) > 0$.

Define event

$$F_n = \Big(L_n(\theta) \le \sup_{|\omega-\theta|\ge\epsilon} L_n(\omega) \Big).$$

Consistency of the maximum likelihood estimator will be established by showing that $P(F_n|\theta) \to 0$ as $n \to \infty$ because on the complement of F_n we have $|\theta_{\mathrm{ML}}(X_1^n) - \theta| < \epsilon$. Indeed we obtain

$$\begin{aligned} P\big(F_n|\theta\big) &\le P\Big(L_n(\theta) \le \sup_{\omega \in \bigcup_{k=0}^m S_k} L_n(\omega) \Big| \theta \Big) \\ &= P\Big(L_n(\theta) \le \sup_{\omega\in S_k} L_n(\omega) \text{ for some } k = 0, ..., m \Big| \theta \Big) \\ &\le \sum_{k=0}^{m} P\Big(L_n(\theta) \le \sup_{\omega\in S_k} L_n(\omega) \Big| \theta \Big) \xrightarrow{n\to\infty} 0, \end{aligned}$$

which completes the proof.

Whereas consistency is a desired condition, the behavior of an estimator for small samples is also important. We may rank estimators according to their bias and mean square error.

Definition 9.3 (bias). *The* bias *of an estimator $T(X_1^n)$ is defined as the expectation $\boldsymbol{E}_\theta T(X_1^n) - \theta$. The estimator is called* unbiased *if $\boldsymbol{E}_\theta T(X_1^n) = \theta$.*

Definition 9.4 (mean square error). *The* mean square error *of an estimator $T(X_1^n)$ is defined as $\boldsymbol{E}_\theta \left[T(X_1^n) - \theta\right]^2$. We say that estimator $T(X_1^n)$* dominates *estimator $T'(X_1^n)$ if*

$$\boldsymbol{E}_\theta \left[T(X_1^n) - \theta\right]^2 \le \boldsymbol{E}_\theta \left[T'(X_1^n) - \theta\right]^2.$$

The mean square error for an unbiased estimator equals its variance. A few interesting estimators such as (9.2) are, however, biased. Unbiasedness can be considered a too strong condition. For instance, to guarantee consistency, it suffices that the mean square error tends to zero.

Theorem 9.3. *Suppose that* $\lim_{n\to\infty} \boldsymbol{E}_\theta \left[T(X_1^n) - \theta\right]^2 = 0$*. Then the estimator* $T(X_1^n)$ *is consistent in probability.*

Proof. The claim follows by Markov inequality (Theorem 1.5)

$$P\big(|T(X_1^n) - \theta| > \epsilon|\theta\big) \le \frac{\mathbf{E}_\theta \left[T(X_1^n) - \theta\right]^2}{\epsilon^2}.$$

Example 9.3. Consider a sample of length n drawn from Bernoulli distribution $P(X_1^n = x_1^n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$. Both X_1 and $\bar{X}_n = \theta_{\mathrm{ML}}(X_1^n) = n^{-1}\sum_{i=1}^n X_i$ are unbiased estimators of θ. The variance of X_i equals $\theta(1-\theta)$, whereas the variance of $\bar{X}_n$ is $\theta(1-\theta)/n$. Hence $\bar{X}_n$ is consistent whereas it can be checked that X_i is not.

The considerations above raise the question what is the minimal variance of an estimator. The question is answered by the Cramér-Rao theorem. Subsequently, we will show that the maximum likelihood estimator achieves the Cramér-Rao bound asymptotically. In both theorems there appears a quantity called Fisher information.

Definition 9.5 (expected Fisher information). *For a discrete random sample, the* expected Fisher information *is defined as*

$$J_n(\theta) := \boldsymbol{E}_\theta \left[\frac{\partial}{\partial\theta} \ln P(X_1^n|\theta)\right]^2,$$

whereas, for a real random sample, we put

$$J_n(\theta) := \boldsymbol{E}_\theta \left[\frac{\partial}{\partial\theta} \ln \rho(X_1^n|\theta)\right]^2.$$

If the parameter $\theta = (\theta_1, \theta_2, ..., \theta_s)$ *is a vector then, for a discrete random sample, the expected Fisher information is defined as matrix*

$$(J_n(\theta))_{ij} := \boldsymbol{E}_\theta \left[\frac{\partial}{\partial\theta_i} \ln P(X_1^n|\theta) \cdot \frac{\partial}{\partial\theta_j} \ln P(X_1^n|\theta)\right]. \tag{9.3}$$

while, for a real random sample, we put

$$(J_n(\theta))_{ij} := \boldsymbol{E}_\theta \left[\frac{\partial}{\partial\theta_i} \ln \rho(X_1^n|\theta) \cdot \frac{\partial}{\partial\theta_j} \ln \rho(X_1^n|\theta)\right].$$

Example 9.4. Consider a sample of length n drawn from Bernoulli distribution. We have

$$P\big(X_1^n = x_1^n|\theta\big) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Hence

$$J_1(\theta) = \theta \left[\frac{\partial}{\partial\theta} \ln\theta\right]^2 + (1-\theta)\left[\frac{\partial}{\partial\theta} \ln(1-\theta)\right]^2 = \frac{1}{\theta(1-\theta)}. \tag{9.4}$$

Fisher information has a few properties that are worth mentioning in the beginning. First, there is an alternative formula for this quantity.

Theorem 9.4. *For a discrete random sample, we have*

$$\begin{aligned} \boldsymbol{E}_\theta \left[\frac{\partial}{\partial \theta} \ln P(X_1^n|\theta) \right] &= 0, \\ \boldsymbol{E}_\theta \left[-\frac{\partial^2}{\partial \theta^2} \ln P(X_1^n|\theta) \right] &= J_n(\theta). \end{aligned}$$

Remark: Random variable $V = (\partial/\partial\theta) \ln P(X_1^n|\theta)$ is called *score.*

Proof. Let us write $L(x_1^n|\theta) = P(X_1^n = x_1^n|\theta)$ as in the previous proof. First we obtain

$$\sum_{x_1^n} L(x_1^n|\theta) \frac{\partial}{\partial \theta} \ln L(x_1^n|\theta) = \sum_{x_1^n} \frac{\partial}{\partial \theta} L(x_1^n|\theta) = \frac{\partial}{\partial \theta} \sum_{x_1^n} L(x_1^n|\theta) = \frac{\partial}{\partial \theta} 1 = 0.$$

Hence the first claim follows. Second, we have

$$\begin{aligned} &\sum_{x_1^n} L(x_1^n|\theta) \left[\frac{\partial}{\partial \theta} \ln L(x_1^n|\theta) \right]^2 + \sum_{x_1^n} L(x_1^n|\theta) \frac{\partial^2}{\partial \theta^2} \ln L(x_1^n|\theta) \\ &= \sum_{x_1^n} L(x_1^n|\theta) \left[\left[\frac{1}{L(x_1^n|\theta)} \frac{\partial L(x_1^n|\theta)}{\partial \theta} \right]^2 + \frac{\partial}{\partial \theta} \frac{1}{L(x_1^n|\theta)} \frac{\partial L(x_1^n|\theta)}{\partial \theta} \right] \\ &= \sum_{x_1^n} L(x_1^n|\theta) \frac{1}{L(x_1^n|\theta)} \frac{\partial^2 L(x_1^n|\theta)}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2} \sum_{x_1^n} L(x_1^n|\theta) = 0. \end{aligned}$$

Thus we have established the second claim.

Using the previous proposition we can show that Fisher information is a linear function of the sample length.

Theorem 9.5. *For a discrete random sample, we have*

$$J_n(\theta) = n J_1(\theta).$$

Proof. Observe

$$J_n(\theta) = \mathbf{E}_\theta \left[-\frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \ln P\big(X_i|\theta\big) \right] = \sum_{i=1}^n \mathbf{E}_\theta \left[-\frac{\partial^2}{\partial \theta^2} \ln P\big(X_i|\theta\big) \right] = n J_1(\theta).$$

Moreover, we can show that Fisher information is the second derivative of Kullback-Leibler divergence with respect to the parameter.

Theorem 9.6. *For a discrete random sample, we have*

$$\left. \frac{\partial^2}{\partial \omega^2} D(\theta||\omega) \right|_{\omega=\theta} = J_1(\theta).$$

Proof. Let us write $L(x|\theta) = P(X_i = x|\theta)$. Observe that

$$\frac{\partial^2}{\partial\omega^2} D(\theta||\omega) = \frac{\partial^2}{\partial\omega^2} \sum_{x\in\mathbb{X}} L(x|\theta) \ln \frac{L(x|\theta)}{L(x|\omega)} = -\sum_{x\in\mathbb{X}} L(x|\theta) \frac{\partial^2}{\partial\omega^2} \ln L(x|\omega),$$

which equals $J_1(\theta)$ for $\omega = \theta$.

Subsequently, we demonstrate the Cramér-Rao bound.

Theorem 9.7 (Cramér-Rao theorem). *The mean square error of an estimator is bounded by the inverse of Fisher information multiplied by the squared derivative of the estimator's expectation,*

$$\boldsymbol{E}_\theta \big[T(X_1^n) - b(\theta)\big]^2 \geq \frac{[b'(\theta)]^2}{J_n(\theta)},$$

where $b(\theta) = \boldsymbol{E}_\theta T(X_1^n)$. In particular, for an unbiased estimator we have $b'(\theta) = 1$.

Proof. Let us write the estimator $T = T(X_1^n)$ and score $V = (\partial/\partial\theta) \ln P(X_1^n|\theta)$. By Schwarz inequality we have

$$\big(\mathbf{E}_\theta \big[(V - \mathbf{E}_\theta V)(T - \mathbf{E}_\theta T)\big]\big)^2 \leq \mathbf{E}_\theta \big(V - \mathbf{E}_\theta V\big)^2 \mathbf{E}_\theta \big(T - \mathbf{E}_\theta T\big)^2.$$

Notice that $\mathbf{E}_\theta V = 0$ as derived in Theorem 9.4. Thus

$$\big(\mathbf{E}_\theta \big[VT\big]\big)^2 \leq J_n(\theta) \mathbf{E}_\theta \big(T - b(\theta)\big)^2.$$

Some further algebra yields

$$\begin{aligned}
\mathbf{E}_\theta \big[VT\big] &= \sum_{x_1^n} P\big(X_1^n = x_1^n|\theta\big) \frac{\partial}{\partial\theta} \ln P\big(X_1^n = x_1^n|\theta\big) T\big(X_1^n\big) \\
&= \sum_{x_1^n} \frac{\partial}{\partial\theta} P\big(X_1^n = x_1^n|\theta\big) T\big(X_1^n\big) \\
&= \frac{\partial}{\partial\theta} \sum_{x_1^n} P\big(X_1^n = x_1^n|\theta\big) T\big(X_1^n\big) = b'(\theta).
\end{aligned}$$

Hence the claim follows.

We say is that an unbiased estimator $T(X_1^n)$ is *efficient* when it satisfies the Cramér-Rao bound with equality, i.e., if

$$\mathbf{E}_\theta \big[T(X_1^n) - \theta\big]^2 = \frac{1}{J_n(\theta)}.$$

In the following we show a simple example of an efficient estimator.

Example 9.5. Let X_1^n be a random sample from the normal distribution with expectation μ and variance σ^2. For parameter μ we obtain

$$\begin{aligned} J_n(\mu) &= n\int\left[\frac{\partial}{\partial\mu}\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\right]^2\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\mathrm{d}x \\ &= n\int\frac{(x-\mu)^2}{\sigma^4}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]=\frac{n}{\sigma^2}. \end{aligned} \tag{9.5}$$

On the other hand estimator $\bar{X}_n=\mu_{\mathrm{ML}}(X_1^n)=n^{-1}\sum_{i=1}^n X_i$ satisfies $\mathbf{E}_\theta\bar{X}_n=\mu$ and

$$\begin{aligned} \mathbf{E}_\theta\left[\bar{X}_n-\mu\right]^2 &= \mathbf{E}_\theta\bar{X}_n^2-2\mu\mathbf{E}_\theta\bar{X}_n+\mu^2 \\ &= \frac{1}{n}(\sigma^2+\mu^2)+\frac{n-1}{n}\mu^2-2\mu^2+\mu^2=\frac{\sigma^2}{n}. \end{aligned}$$

Hence estimator $\bar{X}_n$ is efficient.

An important property of the maximum likelihood estimator is that it is asymptotically unbiased and efficient. To state the respective theorem we need the concept of convergence in distribution.

Definition 9.6 (convergence in distribution). *We say that a series of random real variables $(Y_i)_{i=1}^\infty$ converges to the distribution of a random variable Y, when $\lim_{n\to\infty}P(Y_n\le r)=P(Y\le r)$ for every real number r at which the distribution function $P(Y\le r)$ is continuous.*

Theorem 9.8. *Let X_1^n be a random sample drawn from distribution $P(X_i=x_i|\theta)$, where θ takes values in a subset of real numbers $\Theta\subset\mathbb{R}$. Suppose that:*

1. *Set $A=\{x\in\mathbb{X}:P(X_i=x|\theta)>0\}$ is independent of θ.*
2. *For every $x\in A$, $(\partial^2/\partial\theta^2)P(X_i=x|\theta)$ exists and is continuous in θ.*
3. *Fisher information $J_1(\theta)$ exists and is finite.*
4. *For every θ in the interior of Θ there exists $\epsilon>0$ such that*

$$\boldsymbol{E}_\theta\left[\sup_{\omega\in[\theta-\epsilon,\theta+\epsilon]}\left|\frac{\partial^2}{\partial\omega^2}\ln P(X_i=x|\omega)\right|\right]<\infty.$$

5. *The maximum likelihood estimator $\theta_{\mathrm{ML}}(X_1^n)$ is consistent.*

Then for any θ in the interior of Θ statistic $\sqrt{n}(\theta_{\mathrm{ML}}(X_1^n)-\theta)$ converges under $P(\cdot|\theta)$ to the normal distribution with expectation 0 and variance $1/J_1(\theta)$.

Proof. The exact proof can be found in Keener (2010, Section 9.3). Here we give only a sketch. Let us denote $l_n(\theta)=\ln P(X_1^n|\theta)$ and $\theta_n=\theta_{\mathrm{ML}}(X_1^n)$. The Taylor expansion of $l_n'(\theta)$ around the maximum likelihood estimator yields

$$0=l_n'(\theta_n)=l_n'(\theta)+l_n''(\theta)(\theta_n-\theta)+\frac{1}{2}l_n'''(\theta_n^*)(\theta_n-\theta)^2,$$

where θ_n^* is an intermediate value between θ_n and θ. Solving this equation, we obtain

$$\sqrt{n}\big(\theta_n - \theta\big) = \frac{\frac{1}{\sqrt{n}} l_n'(\theta)}{-\frac{1}{n} l_n''(\theta) - \frac{1}{2n} l_n'''(\theta_n^*)(\theta_n - \theta)}.$$

Considering the numerator

$$\frac{1}{\sqrt{n}} l_n'(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln P\big(X_i|\theta\big), \tag{9.6}$$

we observe that its summands are independent identically distributed variables with expectation 0 and variance $J_1(\theta)$. So, by the central limit theorem, expression (9.6) converges in distribution to a random variable Z which has the normal distribution with expectation 0 and variance $J_1(\theta)$. On the other hand, the left part of the denominator

$$-\frac{1}{n} l_n''(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln P\big(X_i|\theta\big)$$

is an average of independent identically distributed variables converging in probability to the common expectation, which is $J_1(\theta)$. Using consistency of θ_n and other assumptions of the theorem, the right part of the denominator

$$-\frac{1}{2n} l_n'''(\theta_n^*)(\theta_n - \theta)$$

can be shown negligible. Hence $\sqrt{n}(\theta_n - \theta)$ converges in distribution to the random variable $Z/J_1(\theta)$, which has the normal distribution with expectation 0 and variance $1/J_1(\theta)$.

Exercises

1. Consider an exponential family (8.5)–(8.6). What is the maximum likelihood estimator?
2. Find the maximum likelihood estimator for the sample drawn from Poisson distribution (8.8).
3. Show that the mean square error decomposes into the sum of variance and the square of the bias,

$$\mathbf{E}_\theta\big[T(X_1^n) - \theta\big]^2 = \mathbf{E}_\theta\big[T(X_1^n) - \mathbf{E}_\theta T(X_1^n)\big]^2 + \big[\mathbf{E}_\theta T(X_1^n) - \theta\big]^2.$$

4. Let X_1^n be a random sample from the normal distribution with expectation μ and variance σ^2. Define $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Show that

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} \big(X_i - \bar{X}_n\big)^2$$

is a biased estimator of σ^2, whereas

$$\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

is an unbiased estimator.
Remark: Although estimator S_2 is biased, it has a smaller mean square error than $\bar{S}^2$. The estimator which has the smallest mean square error is

$$\hat{S}_n^2 = \frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

5. For a discrete random sample show that

$$\left.\frac{\partial^2}{\partial\omega^2}D(\omega||\theta)\right|_{\omega=\theta} = J_1(\theta).$$

Remark: In Theorem 9.6 we have considered the second derivative of $D(\theta||\omega)$.

6. For a discrete random sample and a multidimensional parameter, show that

$$\left(J_n(\theta)\right)_{ij} = \mathbf{E}_\theta\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln P\left(X_1^n|\theta\right)\right].$$

7. Consider an exponential family (8.5)–(8.6). What is the Fisher information?
8. Consider the exponential family as in the previous task. We have learned that function $\psi(\theta) = \ln Z(\theta)$ is convex. Let us introduce function $\mu_i(\theta) = \mathbf{E}_\theta T_i(X_1) = \partial\psi(\theta)/\partial\theta_i$. This function is an injection and we may define its inverse $\theta(\mu(\theta)) = \theta$. Consider the Legendre transform $\phi(\mu)$, defined as

$$\phi(\mu) = \left[\sum_{l=1}^{M}\theta_l(\mu)\mu_l - \psi(\theta(\mu))\right].$$

Show that $\theta_i(\mu) = \partial\phi(\mu)/\partial\mu_i$ and

$$P\left(X_1^n = x_1^n|\theta\right) = \prod_{i=1}^{n} p(x_i)\exp\left(-d_\phi\left(T(x_i),\mu(\theta)\right) + \phi\left(T(x_i)\right)\right),$$

where d_ϕ is the Bregman divergence for function ϕ.
Hence we see that there is a one-to-one correspondence between Bregman divergences and exponential families. The parametrization of the exponential family in terms of θ is called the canonical parametrization. The parametrization in terms of $\mu = \mu(\theta)$ is called the mean value parametrization.

9. What exponential family corresponds to the Bregman divergence equal squared Euclidean distance? What exponential family corresponds to the Bregman divergence equal Kullback-Leibler divergence?
10. Consider the mean value parametrization for an exponential family. Show that:
$$(J_1(\mu))_{ij} = \frac{\partial^2 \phi(\mu)}{\partial \mu_i \partial \mu_j}$$
and $J_1(\mu) = [J_1(\theta(\mu))]^{-1}$.
11. What is the Fisher information, the expectation $\mathbf{E}_{\theta} X_1$ and variance $\operatorname{Var} X_1$ for a sample drawn from Poisson distribution (8.8)?

10

Bayesian inference

Bayes theorem. Posterior distribution. Jeffreys prior. Improper priors. Laplace integral. Conjugate priors.

So far we have considered the setting of non-Bayesian statistics, where the parameter is a fixed unknown value. In contrast, in Bayesian statistics, it is assumed that the parameter is a random variable with a certain distribution, called the *prior distribution*. The random parameter will be denoted as Θ and the prior distribution as $\Pi(B) = P(\Theta \in B) = \int_B \pi(\theta)\,\mathrm{d}\theta$, where $\pi(\theta)$ is the prior density. When we have the prior distribution, the inference about the parameter is done using the Bayes theorem. Putting $P(X_1^n \in B|\Theta = \theta) = P(X_1^n \in B|\theta)$, we obtain the *posterior distribution*

$$P\big(\Theta \in B|X_1^n \in A\big) = \int_B \pi\big(\theta|X_1^n \in A\big)\,\mathrm{d}\theta,$$

where the posterior density $\pi(\theta|X_1^n \in A)$ is given by

$$\pi\big(\theta|X_1^n \in A\big) = \frac{P(X_1^n \in A|\theta)\pi(\theta)}{\int P(X_1^n \in A|\theta')\pi(\theta')\,\mathrm{d}\theta'}. \tag{10.1}$$

The posterior density is usually more concentrated around the correct value of the parameter than the prior density. This intuition can be substantiated by the following theorem.

Theorem 10.1 (Doob consistency theorem). *Let X_1^n be a random sample drawn from distribution $P(X_i|\theta)$, where $P(X_i|\theta) \neq P(X_i|\theta')$ for $\theta \neq \theta'$. Then for any prior distribution $\Pi = P(\Theta \in \cdot)$ and for every θ belonging to a set G such that $\Pi(G) = 1$, the sequence of posterior distributions $P(\Theta \in \cdot|X_1^n)$ (being a sequence of random variables) converges under $P(\cdot|\theta)$ in distribution to δ_θ, which is a measure concentrated on θ (i.e., $\delta_\theta(\{\theta\}) = 1$).*

The proof of this theorem can be found in van der Vaart (1998, Theorem 10.10). We omit it since it makes a heavy use of measure theory.

Knowing that the posterior is consistent, it is sensible to consider an estimator of the parameter given by maximizing the posterior density.

Definition 10.1 (maximum posterior estimator). *For a discrete random sample, the* maximum posterior estimator *(MAP) of the parameter is defined as the value for which the posterior density is the largest, i.e.,*

$$\begin{aligned}\theta_{\mathrm{MAP}}(x_1^n) &= \mathrm{argmax}_\theta\, \pi\big(\theta|X_1^n = x_1^n\big)\\ &= \mathrm{argmax}_\theta\, P\big(X_1^n = x_1^n|\theta\big)\pi(\theta).\end{aligned}$$

Observe that if $\pi(\theta)$ does not depend on θ then $\theta_{\mathrm{MAP}}(x_1^n) = \theta_{\mathrm{ML}}(x_1^n)$ where the maximum likelihood estimator $\theta_{\mathrm{ML}}(x_1^n)$ is defined in (9.1). We have shown in Chapter 9 that the maximum likelihood estimator is consistent under mild conditions. Hence we may argue that the uniform distribution $\pi(\theta) \propto 1$ is a reasonable choice of the prior. (We write $a \propto b$ if a is proportional to b.)

Differently than in the non-Bayesian setting, in Bayesian statistics we do not need to estimate the parameter to predict the next observation. First, in many cases we can compute the marginal probability

$$P\big(X_1^n \in A\big) = \int P\big(X_1^n \in A|\theta\big)\pi(\theta)\,\mathrm{d}\theta$$

and hence we may compute

$$P\big(X_{n+1} \in C|X_1^n \in A\big) = \frac{P(X_1^{n+1} \in A \times C)}{P(X_1^n \in A)}.$$

In particular, for X_i conditionally independent given Θ, we obtain

$$\begin{aligned} P\big(X_{n+1} \in C|X_1^n \in A\big) &= \frac{\int P(X_1^{n+1} \in A \times C|\theta)\pi(\theta)\,\mathrm{d}\theta}{\int P(X_1^n \in A|\theta)\pi(\theta)\,\mathrm{d}\theta} \\ &= \int P\big(X_{n+1} \in C|\theta\big)\pi\big(\theta|X_1^n \in A\big)\,\mathrm{d}\theta. \end{aligned}$$

This principle may be illustrated on the following example.

Example 10.1. Consider a random sample of length n drawn from Bernoulli distribution with the prior $\pi(\theta) = 1$. We have

$$\begin{aligned} &P\big(X_1^n = x_1^n\big) = \int P\big(X_1^n = x_1^n|\theta\big)\,\mathrm{d}\theta \\ &= \int \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\,\mathrm{d}\theta \\ &= \frac{\Gamma\big(\sum_{i=1}^n x_i + 1\big)\Gamma\big(n - \sum_{i=1}^n x_i + 1\big)}{\Gamma(n+2)} = \frac{1}{n+1}\left[\binom{n}{\sum_{i=1}^n x_i}\right]^{-1}. \end{aligned}$$

Hence

$$\begin{aligned} P\big(X_{n+1} = 1|X_1^n = x_1^n\big) &= \frac{(\sum_{i=1}^n x_i + 1)!\,(n - \sum_{i=1}^n x_i)!(n+1)!}{(n+2)!\,(\sum_{i=1}^n x_i)!\,(n - \sum_{i=1}^n x_i)!} \\ &= \frac{\sum_{i=1}^n x_i + 1}{n+2}, \end{aligned}$$

which is called the Laplace rule.

We can see that Bayesian inference is simple if we have a prior and the posterior can be computed efficiently. Now a few words should be devoted to the appropriate choice of the prior. We have seen in Theorem 10.1 that almost any choice of prior leads asymptotically to the same, correct inference. The inference can be, however, very different for finite samples. Taking the uniform prior is not the only reasonable choice. In general, the prior distribution should resume all information about the parameter that we have. If we do not have such information, the prior distribution should be chosen taking three criteria into account:

1. symmetry,
2. invariance with respect to reparametrization,
3. computational simplicity.

Each of these criteria leads to a different kind of inference.

Some interesting option is given by the Jeffreys prior. As we will see, the inference using this prior is invariant with respect to reparametrization.

Definition 10.2 (Jeffreys prior). *The* Jeffreys prior *is defined as*

$$\pi_{\text{Jeffreys}}(\theta) = \frac{\sqrt{\det J_n(\theta)}}{\int \sqrt{\det J_n(\theta)}\,\mathrm{d}\theta}, \tag{10.2}$$

where $J_n(\theta)$ is the expected Fisher information, given in formula (9.3).

If variables $X_1, X_2, X_3, \ldots$ are independent given Θ and have identical distribution then $J_n(\theta) = nJ_1(\theta)$ and the Jeffreys prior does not depend on n.

Example 10.2. Consider a random sample of length n drawn from Bernoulli distribution. From (9.4) we obtain

$$\pi_{\text{Jeffreys}}(\theta) = \frac{\theta^{-1/2}(1-\theta)^{-1/2}}{\pi}.$$

Hence

$$P\big(X_1^n = x_1^n\big) = \int P\big(X_1^n = x_1^n|\theta\big)\pi_{\text{Jeffreys}}(\theta)\,\mathrm{d}\theta = \frac{1}{n\pi}\left[\binom{n-1}{\sum_{i=1}^n x_i - 1/2}\right]^{-1}$$

and

$$P\big(X_{n+1} = 1|X_1^n = x_1^n\big) = \frac{\sum_{i=1}^n x_i + 1/2}{n+1},$$

which is called the Krichevski-Trofimov estimator.

The appeal of Jeffreys prior comes from the following fact.

Theorem 10.2. *The inference using Jeffreys prior does not depend on parametrization, i.e., the marginal distribution*

$$P\big(X_1^n \in A\big) = \int P\big(X_1^n \in A|\theta\big)\pi_{\text{Jeffreys}}(\theta)\,\mathrm{d}\theta$$

is invariant with respect to a one-to-one differentiable reparametrization.

Proof. Let us introduce parameter $\phi = \phi(\theta)$, which is a one-to-one differentiable function of θ. We have

$$\int P\big(X_1^n \in A|\theta\big)\pi_{\text{Jeffreys}}(\theta)\,\mathrm{d}\theta = \int P\big(X_1^n \in A|\phi\big)\pi(\phi)\,\mathrm{d}\phi,$$

where

$$\begin{aligned}\pi(\phi) &= \pi_{\text{Jeffreys}}(\theta)\left|\det\frac{\partial\theta_k}{\partial\phi_i}\right| \\ &\propto \sqrt{\det\frac{\partial\theta_k}{\partial\phi_i}\det\mathbf{E}_{\theta}\left[\frac{\partial\ln P(X_1^n|\theta)}{\partial\theta_k}\frac{\partial\ln P(X_1^n|\theta)}{\partial\theta_l}\right]\det\frac{\partial\theta_l}{\partial\phi_j}} \\ &= \sqrt{\det\mathbf{E}_{\phi}\left[\frac{\partial\ln P(X_1^n|\phi)}{\partial\phi_i}\frac{\partial\ln P(X_1^n|\phi)}{\partial\phi_j}\right]} \propto \pi_{\text{Jeffreys}}(\phi).\end{aligned}$$

Hence

$$\int P\big(X_1^n \in A|\theta\big)\pi_{\text{Jeffreys}}(\theta)\,\mathrm{d}\theta = \int P\big(X_1^n \in A|\phi\big)\pi_{\text{Jeffreys}}(\phi)\,\mathrm{d}\phi.$$

Sometimes in Bayesian reasoning there appear *improper priors*, which are prior densities such that $\int\pi(\theta)\,\mathrm{d}\theta = \infty$. These priors may be used for inference if $\int P(X_1^n \in A|\theta)\pi(\theta)\,\mathrm{d}\theta < \infty$. Then the posterior distribution $\pi(\theta|X_1^n \in A)$ and the conditional probability $P(X_{n+1} \in C|X_1^n \in A)$ are properly defined. We illustrate this phenomenon on an example of a Jeffreys prior.

Example 10.3. Consider a random sample drawn from the normal distribution

$$\rho(x_1^n|\mu,\sigma) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]. \tag{10.3}$$

If μ is the only unknown parameter then by (9.5) we have

$$\pi_{\text{Jeffreys}}(\mu) \propto \sqrt{J_1(\mu)} = \sqrt{\frac{1}{\sigma^2}} \propto 1.$$

Hence

$$\pi_{\text{Jeffreys}}(\mu|x_1^n) = \frac{\rho(x_1^n|\mu,\sigma)}{\rho(x_1^n|\sigma)}$$

and

$$\rho(x_{n+1}|x_1^n,\sigma) = \frac{\rho(x_1^{n+1}|\sigma)}{\rho(x_1^n|\sigma)},$$

where

$$\begin{aligned}
\rho(x_1^n|\sigma) &= \int_{-\infty}^{\infty} \rho(x_1^n|\mu,\sigma)\,\mathrm{d}\mu \\
&= \int \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{\mu\sum_{i=1}^n x_i}{\sigma^2} - \frac{\mu^2 n}{2\sigma^2}\right] \mathrm{d}\mu \\
&= \int \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{n}{2\sigma^2}\left(\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2\right)\right] \\
&\quad \times \exp\left[-\frac{\left(\sum_{i=1}^n x_i/\sqrt{n} - \mu\sqrt{n}\right)^2}{2\sigma^2}\right] \mathrm{d}\mu \\
&= \frac{1}{(\sigma\sqrt{2\pi})^{n-1}\sqrt{n}} \exp\left[-\frac{n}{2\sigma^2}\left(\frac{\sum_{i=1}^n x_i^2}{n} - \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2\right)\right]
\end{aligned}$$

Practical applications of Bayesian inference are limited by the difficulty of computing the posterior for complicated statistical models. The difficulty lies in computing the denominator in (10.1), which is an integral. In many cases the only possibility is to estimate this integral consists in using Monte Carlo methods or using the Laplace approximation. The latter approach rests on the following theorem.

Theorem 10.3. *Let X_1^n be a random sample drawn from an s-parameter exponential family (8.5)–(8.6). Let Θ_0 be a subset of $\Theta = \{\omega \in \mathbb{R}^s : Z(\omega) < \infty\}$ such that*

1. *Θ_0 is a compact subset of the interior of Θ;*
2. *The interior of Θ_0 in nonempty.*

Moreover, let π be a prior density that is continuous on Θ and strictly positive on Θ_o, i.e., $\inf_{\theta\in\Theta_0}\pi(\theta) > 0$. Finally, let a sequence $(x_i)_{i=1}^\infty$ be such that $\theta_{\mathrm{ML}}(x_1^n) \in \Theta_o$ for sufficiently large n. Then

$$\begin{aligned}
\ln P\big(X_1^n = x_1^n|\theta_{\mathrm{ML}}(x_1^n)\big) &- \ln\int P\big(X_1^n = x_1^n|\theta\big)\pi(\theta)\,\mathrm{d}\theta \\
&= \frac{s}{2}\ln\frac{n}{2\pi} + \ln\frac{\sqrt{\det J_1(\theta_{\mathrm{ML}}(x_1^n))}}{\pi(\theta_{\mathrm{ML}}(x_1^n))} + o(1), \qquad (10.4)
\end{aligned}$$

where the convergence is uniform in Θ_0.

Remark: Formula (10.4) becomes particularly simple for the Jeffreys prior (10.2). Namely, we obtain

$$\begin{aligned}
\ln P\big(X_1^n = x_1^n|\theta_{\mathrm{ML}}(X_1^n)\big) &- \ln\int P\big(X_1^n = x_1^n|\theta\big)\pi_{\mathrm{Jeffreys}}(\theta)\,\mathrm{d}\theta \\
&= \frac{s}{2}\ln\frac{n}{2\pi} + \ln\int\sqrt{\det J_1(\theta)}\,\mathrm{d}\theta + o(1).
\end{aligned}$$

Proof. We give only a sketch of the proof. The full proof can be found in Grünwald (2007, Appendix 8.A). Write $L(x_1^n|\theta) = P(X_1^n = x_1^n|\theta)$ and $\theta_n = \theta_{\mathrm{ML}}(x_1^n)$. First, we observe that

$$-\frac{1}{n}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln L(x_1^n|\theta) = \frac{\partial^2 \ln Z(\theta)}{\partial\theta_i\partial\theta_j} = (J_1(\theta))_{ij}$$

equals the expected Fisher information matrix. Because $(\partial/\partial\theta_i)\ln L(x_1^n|\theta)$ vanishes for $\theta = \theta_{\mathrm{ML}}(x_1^n)$, we may approximate

$$\ln L(x_1^n|\theta) \approx \ln L(x_1^n|\theta_n) - \frac{n}{2}(\theta-\theta_n)^T J_1(\theta_n)(\theta-\theta_n)$$

using the Taylor expansion. In the following, it can be shown that this approximation is so good that we obtain

$$\begin{aligned}\int L(x_1^n|\theta)\pi(\theta)\,\mathrm{d}\theta &\approx L(x_1^n|\theta_n)\pi(\theta_n)\\ &\qquad\times\int \exp\left[-\frac{n}{2}(\theta-\theta_n)^T J_1(\theta_n)(\theta-\theta_n)\right]\mathrm{d}\theta\\ &= L(x_1^n|\theta_n)\pi(\theta_n)\left[\frac{2\pi}{n}\right]^{s/2}\left[\det J_1(\theta_n)\right]^{-1/2}.\end{aligned}$$

Hence the claim follows.

Using the Laplace approximation is not the only trick used to make the Bayesian inference feasible. In certain special cases, it is possible to find an analytic formula for the posterior. Such a case arises for conjugate priors, which are specially designed to make the computation simple.

Definition 10.3 (conjugate prior). *A* conjugate prior *for the parametric family $P(X_1^n|\theta)$ is defined as a family of prior distributions $\pi(\theta) = \pi(\theta|\alpha)$ such that $\pi(\theta|X_1^n = x_1^n) = \pi(\theta|X_1^n = x_1^n, \alpha) = \pi(\theta|\alpha')$ for some $\alpha' = \alpha'(\alpha, x_1^n)$. Whereas θ is called a parameter, α is called a* hyperparameter*.*

In our first example, the Jeffreys prior is a special case of a conjugate prior.

Example 10.4. Consider a random sample drawn from Bernoulli distribution. Let the prior

$$\pi(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{10.5}$$

where $\alpha, \beta > 0$, be the beta distribution. Then

$$\begin{aligned}\pi\big(\theta|X_1^n = x_1^n,\alpha,\beta\big) &= \frac{\theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}}{\int \theta'^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta')^{n-\sum_{i=1}^n x_i+\beta-1}\,\mathrm{d}\theta'}\\ &= \pi\left(\theta\,\middle|\,\sum_{i=1}^n x_i+\alpha, n-\sum_{i=1}^n x_i+\beta\right)\end{aligned}$$

and

$$P\big(X_1^n = x_1^n|\alpha,\beta\big) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}\,\mathrm{d}\theta$$
$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum_{i=1}^n x_i+\alpha)\Gamma(n-\sum_{i=1}^n x_i+\beta)}{\Gamma(n+\alpha+\beta)}.$$

Hence

$$P\big(X_{n+1} = 1|X_1^n = x_1^n,\alpha,\beta\big) = \frac{\sum_{i=1}^n x_i+\alpha}{n+\alpha+\beta}.$$

In the second example, we will see that conjugate priors can be different to Jeffreys priors.

Example 10.5. Consider a random sample drawn from the normal distribution (10.3). For the unknown parameter μ we choose the conjugate prior as the normal distribution

$$\pi(\mu|\mu_0,\sigma_0) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right],$$

where $\mu_0 \in (-\infty,\infty)$ and $\sigma_0 \in (0,\infty)$. Then we obtain

$$\begin{aligned}
\pi(\mu|x_1^n,\alpha,\beta) &= \frac{\rho(x_1^n|\mu,\sigma)\pi(\mu|\mu_0,\sigma_0)}{\int \rho(x_1^n|\nu,\sigma)\pi(\nu|\mu_0,\sigma_0)\,\mathrm{d}\nu}\\
&\propto \left(\prod_{i=1}^n \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right)\exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]\\
&\propto \exp\left[\mu\left[\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right]-\frac{\mu^2}{2}\left[\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right]\right]\\
&\propto \exp\left[-\frac{1}{2}\left[\frac{\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\sqrt{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}}-\mu\sqrt{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right]^2\right]\\
&\propto \pi\left(\mu\,\middle|\,\frac{\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}},\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\right).
\end{aligned}$$

Exercises

1. *(Jeffreys prior)* Consider an exponential family (8.5)–(8.6). Find the Jeffreys prior.
2. *(Conjugate priors)* Consider a random sample drawn from Poisson distribution (8.8). Find the conjugate prior, the posterior and the distribution of X_1^n.
   Hint: The conjugate prior is gamma distribution (8.11).

3. Consider a random sample drawn from geometric distribution (8.9). Find the conjugate prior, the posterior and the distribution of X_1^n.
   Hint: The conjugate prior is beta distribution (10.5).
4. Consider a random sample drawn from negative binomial distribution (8.10). Find the conjugate prior, the posterior and the distribution of X_1^n.
   Hint: The conjugate prior is beta distribution (10.5).
5. Consider a random sample drawn from gamma distribution (8.11). Find the conjugate prior, the posterior and the distribution of X_1^n.
   Hint: The conjugate prior is gamma distribution (8.11).
6. Consider a random sample drawn from Pareto distribution (8.12). Find the conjugate prior, the posterior and the distribution of X_1^n.
   Hint: The conjugate prior is gamma distribution (8.11).
7. *(Hypothesis testing)* To compare the likelihood of two discrete hypotheses, Bayesians use the posterior ratio

   $$\frac{P(\mathcal{H}_1|D)}{P(\mathcal{H}_0|D)} = \frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_0)} \frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)},$$

   where $P(\mathcal{H}_i)$ are the prior probabilities of the hypotheses, and $P(D|\mathcal{H}_i)$ are the probabilities of obtaining data D given each hypothesis. If the ratio above is greater than 1 then hypothesis $\mathcal{H}_1$ is more likely, otherwise we prefer hypothesis $\mathcal{H}_0$. If there is no reason to prefer any hypothesis a priori, we accept $P(\mathcal{H}_1) = P(\mathcal{H}_0)$.
   Task: A certain disease attacks 0.01% population. A certain test for detecting this disease commits 0.01% errors both for healthy and ill persons. What is the probability of being ill if one receives a positive outcome of the test?

11

EM algorithm and K-means

EM algorithm. K-means algorithm. Baum-Welch algorithm.

Consider a likelihood function $\rho(y|\theta)$, where y is an observed value and θ is an unknown parameter. As we know from Chapter 9, the maximum likelihood estimator of θ given y is

$$\theta_{\mathrm{ML}}(y) = \operatorname{argmax}_\theta \rho(y|\theta). \tag{11.1}$$

Sometimes the direct maximization in (11.1) is analytically intractable. In certain of these cases we may consider an approximate procedure introduced by Dempster et al. (1977), called the expectation-maximization (EM) algorithm.

The EM algorithm may be used when there exists a latent discrete variable Z such that distributions $\rho(y|Z=z,\theta)$ and $P(Z=z|\theta)$ are particularly simple. For instance, $\rho(y|Z=z,\theta)$ may be a single-peaked distribution and $P(Z=z|\theta)$ may be the probability of that peak. We have then a mixture model

$$\rho(y|\theta) = \sum_z P\big(Z=z|\theta\big)\rho(y|Z=z,\theta),$$

which is a mixture of several peaks.

Example 11.1 (K Gaussian peaks). Let $y = y_1^N$, $Z = Z_1^N$, $\rho(y|Z,\theta) = \prod_i \rho(y_i|Z_i,\theta)$, and $P(Z|\theta) = \prod_i P(Z_i|\theta)$, where $\theta = (\tau_1, \mu_1, \sigma_1, ..., \tau_K, \mu_K, \sigma_K)$. The conditional likelihood is given as the normal distributions

$$\rho(y_i|Z_i=k,\theta) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left[-\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right]$$

and the mixture coefficients are $P(Z_i=k|\theta) = \tau_k$ where $\tau_k \geq 0$ and $\sum_{k=1}^K \tau_k = 1$. For a given sample we want to find the location and variance of the K peaks and the peak to which a given individual observation belongs. Formally, we seek for a local maximum likelihood estimate of $(\tau_1, \mu_1, \sigma_1, ..., \tau_K, \mu_K, \sigma_K)$ given sample y_1^n and the conditionally most likely value of Z_i for each y_i.

It is important to observe that we are not looking for the global maximum of the likelihood function. The global maximum is quite uninteresting. Namely, we obtain $\rho(y|\theta) = \infty$ if $\mu_1 = y_i$ and $\sigma_1 = 0$ for some i. Thus we are rather interested in finding a finite local maximum.

Now let us describe the EM algorithm. For a given mixture model, we introduce the conditional probability of the latent variable

$$P\big(Z=z|y,\theta\big) = \frac{P(Z=z|\theta)\rho(y|Z=z,\theta)}{\rho(y|\theta)} \tag{11.2}$$

and we consider function

$$\begin{aligned}Q(\theta||\theta') &= \ln\rho(y|\theta') + \sum_z P\big(Z=z|y,\theta\big)\ln P\big(Z=z|y,\theta'\big)\\ &= \sum_z P\big(Z=z|y,\theta\big)\ln\big[P\big(Z=z|\theta'\big)\rho(y|Z=z,\theta')\big],\end{aligned}$$

which is a difference of the likelihood and a cross entropy function. The EM algorithm consists in setting an initial parameter value θ_1 and iterating

$$\theta_{n+1} = \text{argmax}_{\theta'}\, Q\left(\theta_n||\theta'\right) \tag{11.3}$$

until a sufficient convergence of θ_n is achieved. The EM algorithm is worth considering only if maximization (11.3) is easy—for instance, if it can be done analytically. Once we have a sufficiently good approximation of the maximum likelihood estimator then using formula (11.2) we may also compute the conditionally most likely values of the latent variable, i.e., the mixture component which is most likely for a given observation.

The soundness of the EM algorithm follows from the fact that the likelihood grows as a function of the iteration.

Theorem 11.1. *We have*

$$\rho(y|\theta_{n+1}) \ge \rho(y|\theta_n). \tag{11.4}$$

Proof. By (11.3) and nonnegativity of Kullback-Leibler divergence, we obtain

$$\begin{aligned}0 &\le Q\left(\theta_n||\theta_{n+1}\right) - Q\left(\theta_n||\theta_n\right)\\ &= \ln\rho(y|\theta_{n+1}) + \sum_z P\big(Z=z|y,\theta_n\big)\ln P\big(Z=z|y,\theta_{n+1}\big)\\ &\quad - \ln\rho(y|\theta_n) - \sum_z P\big(Z=z|y,\theta_n\big)\ln P\big(Z=z|y,\theta_n\big)\\ &= \ln\rho(y|\theta_{n+1}) - \ln\rho(y|\theta_n) - \sum_z P\big(Z=z|y,\theta_n\big)\ln\frac{P(Z=z|y,\theta_n)}{P(Z=z|y,\theta_{n+1})}\\ &\le \ln\rho(y|\theta_{n+1}) - \ln\rho(y|\theta_n).\end{aligned}$$

It is worth noting that the EM algorithm can be generalized also to the case when the latent variable Z is continuous (the probabilities should be replaced for densities and the sums for integrals). Then inequality (11.4) stems from nonnegativity of Kullback-Leibler divergence for densities, which was discussed in Chapter 7.

It happens that the mixture of K Gaussians may be solved using the EM algorithm.

Example 11.2 (K Gaussian peaks continued). Consider the model from Example 11.1. The cross entropy takes form

$$Q(\theta||\theta') = \sum_{i=1}^N\sum_{j=1}^K \frac{\xi_{ji}}{\sum_{k=1}^K \xi_{ki}}\ln\frac{\xi'_{ji}}{\sqrt{2\pi}},$$

where we denote

$$\xi_{ki} = \frac{\tau_k}{\sigma_k} \exp\left[-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\right].$$

Now we maximize the cross entropy to obtain explicit formulae for the EM iteration. Using Lagrange multipliers, we first obtain

$$0 = \frac{\partial}{\partial \tau_j}\left[Q(\theta_n||\theta) - \lambda\Big(\sum_{k=1}^{K}\tau_k - 1\Big)\right]\Bigg|_{\theta=\theta_{n+1}} = \sum_{i=1}^{N}\frac{\xi_{1i}^{(n)}}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}\frac{1}{\tau_j^{(n+1)}} - \lambda.$$

Hence

$$\tau_j^{(n+1)} = \sum_{i=1}^{N}\frac{\xi_{ji}^{(n)}}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}.$$

Second, we have

$$0 = \frac{\partial}{\partial \mu_j}Q(\theta_n||\theta)\Bigg|_{\theta=\theta_{n+1}} = \sum_{i=1}^{N}\frac{\xi_{ji}^{(n)}}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}\frac{2(y_i - \mu_j^{(n+1)})}{2(\sigma_j^{(n+1)})^2}.$$

Thus

$$\mu_j^{(n+1)} = \sum_{i=1}^{N}\frac{\xi_{ji}^{(n)}y_i}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}\Bigg/\sum_{i=1}^{N}\frac{\xi_{ji}^{(n)}}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}. \tag{11.5}$$

Third, we obtain

$$0 = \frac{\partial}{\partial \sigma_j}Q(\theta_n||\theta)\Bigg|_{\theta=\theta_{n+1}} = \sum_{i=1}^{N}\frac{\xi_{ji}^{(n)}}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}\left(-\frac{1}{\sigma_j^{(n+1)}} + 2\frac{(y_i - \mu_j^{(n+1)})^2}{2(\sigma_j^{(n+1)})^3}\right).$$

Therefore

$$\sigma_j^{(n+1)} = \sqrt{\sum_{i=1}^{N}\frac{\xi_{ji}^{(n)}(y_i - \mu_j^{(n+1)})^2}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}\Bigg/\sum_{i=1}^{N}\frac{\xi_{ji}^{(n)}}{\sum_{k=1}^{K}\xi_{ki}^{(n)}}}.$$

To guarantee that the EM algorithm converges to a local maximum $\rho(y|\theta) < \infty$, it is necessary to set $\mu_j^{(1)} \neq y_i$ and $\sigma_j^{(1)} \neq 0$ for all i.

The EM algorithm for K Gaussian peaks is a soft assignment version of a very popular algorithm for clustering, called the K-means algorithm. The setting of the K-means algorithm is as follows. We are given some data points $y = y_1^N$ and we are interested in partitioning them into K clusters. By r_{ji} we will denote whether y_i is in the j-th cluster ($r_{ji} = 1$ if it is, $r_{ji} = 0$ if it is not). By μ_j we will denote the center of the j-th cluster. Both quantities will be computed in iteration as $r_{ji}^{(n)}$ and $\mu_j^{(n)}$ respectively. The K-means algorithm consists in the iteration

$$r_{ji}^{(n+1)} = \begin{cases} 1 & \text{if } j = \operatorname{argmin}_k \left(y_i - \mu_k^{(n)}\right)^2, \\ 0 & \text{else,} \end{cases} \tag{11.6}$$

$$\mu_j^{(n+1)} = \frac{\sum_{i=1}^n r_{ji}^{(n+1)} y_i}{\sum_{i=1}^n r_{ji}^{(n+1)}}. \tag{11.7}$$

We can see that we obtain (11.7) from (11.5) in the limit of $\sigma_k^{(n)} \to 0$.

Convergence of the K-means algorithm is guaranteed by this proposition.

Theorem 11.2. *Let*

$$J_n = \sum_{i=1}^N \sum_{j=1}^K r_{ji}^{(n)} \left(y_i - \mu_j^{(n)}\right)^2.$$

We have $J_{n+1} \le J_n$.

Proof. By (11.6) we have

$$J_n \ge \sum_{i=1}^N \sum_{j=1}^K r_{ji}^{(n+1)} \left(y_i - \mu_j^{(n)}\right)^2.$$

Subsequently, we find the minimum of the latter function with respect to $\mu_j^{(n)}$,

$$0 = \frac{\partial}{\partial \mu_j} \sum_{i=1}^N \sum_{j=1}^K r_{ji}^{(n+1)} \left(y_i - \mu_j\right)^2 = -2 \sum_{i=1}^N r_{ji}^{(n+1)} \left(y_i - \mu_j\right),$$

which implies that $\mu_j = \mu_j^{(n+1)}$. Hence we have

$$\sum_{i=1}^N \sum_{j=1}^K r_{ji}^{(n+1)} \left(y_i - \mu_j^{(n)}\right)^2 \ge J_{n+1}.$$

Another important application of the EM algorithm concerns computing parameters of a hidden Markov chain given the observed states. The respective instance of the EM algorithm is called the Baum-Welch algorithm.

Example 11.3 (Baum-Welch algorithm). Consider a hidden Markov chain with known observed states and unknown hidden states and parameters. Namely, we put $P(Z, Y|\theta) = \prod_{i=1}^N P(Z_i, Y_i|Z_{i-1}, \theta)$, where $P(Z_i = k, Y_i = l|Z_{i-1} = j, \theta) =$

θ_{jkl}, $\theta_{jkl} \geq 0$, and $\sum_{kl} \theta_{jkl} = 1$. The cross entropy takes form

$$\begin{aligned} Q(\theta\|\theta') &= \sum_z P\big(Z = z|Y = y, \theta\big) \ln P\big(Z = z, Y = y|\theta'\big) \\ &= \sum_z \frac{P(Z = z, Y = y|\theta)}{\sum_{z'} P(Z = z, Y = y|\theta)} \ln P(Z = z', Y = y|\theta') \\ &= \sum_{m=1}^N \frac{\sum_z \prod_{i=1}^N \theta_{z_{i-1},z_i,y_i} \ln \theta'_{z_{m-1},z_m,y_m}}{\sum_{z'} \prod_{i=1}^N \theta_{z'_{i-1},z'_i,y_i}}. \end{aligned}$$

To obtain explicit formulae for the EM iteration, we maximize the cross entropy. We compute

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_{jkl}} \left[Q(\theta_n\|\theta) - \lambda\Big(\sum_{kl} \theta_{jkl} - 1\Big)\right]\Bigg|_{\theta=\theta_{n+1}} \\ &= \sum_{m=1}^N \frac{\sum_z \prod_{i=1}^N \theta^{(n)}_{z_{i-1},z_i,y_i} \mathbf{1}\{(z_{m-1}, z_m, y_m) = (j,k,l)\}}{\sum_{z'} \prod_{i=1}^N \theta^{(n)}_{z'_{i-1},z'_i,y_i}} \frac{1}{\theta^{(n+1)}_{jkl}} - \lambda. \end{aligned}$$

Hence

$$\theta^{(n+1)}_{jkl} = \frac{\sum_{m=1}^N \sum_z \prod_{i=1}^N \theta^{(n)}_{z_{i-1},z_i,y_i} \mathbf{1}\{(z_{m-1}, z_m, y_m) = (j,k,l)\}}{\sum_{m=1}^N \sum_z \prod_{i=1}^N \theta^{(n)}_{z_{i-1},z_i,y_i} \mathbf{1}\{z_{m-1} = j\}}.$$

The above expression can be computed efficiently. Let us introduce forward and backward probabilities

$$\begin{aligned} \alpha_m(k) &= \sum_{z_1 \ldots z_m} \prod_{i=1}^m \theta^{(n)}_{z_{i-1},z_i,y_i} \mathbf{1}\{z_m = k\}, \\ \beta_m(j) &= \sum_{z_m \ldots z_N} \prod_{i=m}^{N-1} \theta^{(n)}_{z_i,z_{i+1},y_{i+1}} \mathbf{1}\{z_m = j\}. \end{aligned}$$

Then we have

$$\begin{aligned} \alpha_0(k) &= 1, & \alpha_{m+1}(k) &= \sum_j \theta^{(n)}_{j,k,y_{m+1}} \alpha_m(j), \\ \beta_N(k) &= 1, & \beta_{m-1}(j) &= \sum_k \theta^{(n)}_{j,k,y_m} \beta_m(k), \end{aligned}$$

and

$$\theta^{(n+1)}_{jkl} = \frac{\sum_{m=1}^N \alpha_{m-1}(j) \theta^{(n)}_{jkl} \mathbf{1}\{y_m = l\} \beta_m(k)}{\sum_{m=1}^N \alpha_{m-1}(j) \beta_{m-1}(j)}.$$

Exercises

1. Assuming $P(Z, Y|\theta) = \prod_i P(Z_i, Y_i|\theta)$, find the EM iteration for the mixture of Bernoulli distributions $P(Y_i = y|Z_i = 1, \theta) = \theta_1^y (1-\theta_1)^{1-y}$ and $P(Y_i = y|Z_i = 2, \theta) = \theta_2^y (1-\theta_2)^{1-y}$, where $P(Z_i = 1|\theta) = \tau_1$ and $P(Z_i = 2|\theta) = \tau_2 = 1 - \tau_1$.
2. We are given a sample of m male twin pairs, f female twin pairs, and o opposite sex twin pairs. Estimate the probability p that a twin pair is identical and the probability q that a child is male.
Hint: The observed data are $Y = (m, f, o)$ and $\theta = (p, q)$ is the parameter. If we knew which pairs of same-sex were identical then it would be easy to estimate p and q. Thus we may postulate the complete data $(Y, Z) = (m_1, m_2, f_1, f_2, o)$ where m_1 (f_1) is the number of male (female) identical twins and m_2 (f_2) is the number of male (female) non-identical twins. The complete likelihood function is
$$P\big(Y, Z|\theta\big) = \binom{m+f+o}{m_1, m_2, f_1, f_2, o} (pq)^{m_1} [(1-p)q^2]^{m_2} [p(1-q)]^{f_1} \times \\ \times [(1-p)(1-q)^2]^{f_2} [(1-p)2q(1-q)]^o$$
since identical twins involve one choice of sex and nonidentical twins two choices of sex. The conditional expectations of m_i and f_i given Y are
$$\mathbf{E}_\theta\big[m_1|Y\big] = m\frac{pq}{pq + (1-p)q^2},$$
$$\mathbf{E}_\theta\big[m_2|Y\big] = m\frac{(1-p)q^2}{pq + (1-p)q^2},$$
$$\mathbf{E}_\theta\big[f_1|Y\big] = f\frac{p(1-q)}{p(1-q) + (1-p)(1-q)^2},$$
$$\mathbf{E}_\theta\big[f_2|Y\big] = f\frac{(1-p)(1-q)^2}{p(1-q) + (1-p)(1-q)^2}.$$
3. Suppose that the lifetimes Y_i of lightbulbs follow exponential distribution $\rho(Y_i|\theta) = \theta^{-1}\exp(-Y_i/\theta)$. We conduct two experiments: In the first experiment, with N bulbs, the exact lifetimes $Y_1, ..., Y_N$ are recorded. In the second experiment, with M bulbs, the experimenter enters the laboratory at some time $t > 0$ and all that he registers is that Z bulbs are burning while $M - Z$ have expired. Having this data, what is the maximum likelihood estimator of θ given by the EM algorithm?
Hint: Let $E_i = 1$ when the i-th bulb from the second experiment is burning and $E_i = 0$ when the light is out. The observed data is $(Y_1, ..., Y_N, E_1, ..., E_M)$, whereas the latent variables are $(X_1, ..., X_M)$ with X_i being the unknown lifetime of the lightbulb in the second experiment. The conditional expectation of X_i given E_i is
$$\mathbf{E}_\theta\big[X_i|E_I\big] = \begin{cases} t + \theta & \text{if } E_i = 1, \\ \theta - \frac{te^{-t/\theta}}{1-e^{-t/\theta}} & \text{if } E_i = 0. \end{cases}$$

4. Let $Z_1, Z_2, ..., Z_M$, with $Z_i : \Omega \to J$, be a sequence of discrete random variables and let $Y_1, Y_2, ..., Y_M$ be a random sample of sets, where each set $Y_i : \Omega \to 2^J \setminus \emptyset$ contains the actual value of Z_i, i.e., $Z_i \in Y_i$. The objective is to guess the conditional distribution of Z_i given an event $(Y_i = A_i)_{i=1}^M$, $A_i \subset J$. In particular, we would like to know the conditionally most likely values of Z_i.

   For this task we adopt the following probability model. First, we assume that the likelihood factorizes into $P(Z, Y|\theta) = \prod_i P(Z_i, Y_i|\theta)$. Second, we assume that

$$P\big(Y_i = A|Z_i = j, \theta\big) = \begin{cases} g(A), & j \in A, \\ 0, & \text{else}, \end{cases} \tag{11.8}$$
$$P\big(Z_i = j|\theta\big) = p_j$$

   for parameter $\theta = (p_j)_{j \in J}$ and a parameter-free function $g(\cdot)$ satisfying

$$\sum_{A \in 2^J} \mathbf{1}\{j \in A\} g(A) = 1, \quad \forall j \in J. \tag{11.9}$$

   For example, let $g(A) = q^{\operatorname{card} A - 1}(1-q)^{\operatorname{card} J - \operatorname{card} A}$, where $\operatorname{card} A$ stands for the cardinality of set A and $0 \leq q \leq 1$ is a fixed number not incorporated into θ. Then the cardinalities of sets Y_i are binomially distributed, i.e., $P(\operatorname{card} Y_i - 1|\theta) \sim B(\operatorname{card} J - 1, q)$. This particular form of $g(A)$, however, is not necessary to satisfy (11.9). In fact, assumption (11.8) leads to an EM algorithm which does not depend on the specific choice of function $g(\cdot)$. Find this EM algorithm.

12

Maximum entropy

Boltzmann H-theorem. Maximum entropy modeling.

In Chapter 8 we introduced exponential families of distributions and subsequently we showed that they have many nice properties. Now we will show that exponential families have another interesting property, namely, they maximize entropy given linear constraints on the sufficient statistic. Before we discuss the maximum entropy distributions, we want, however, to make a digression and explain where the principle of maximum entropy comes from. In fact, this principle originated in physics. We will exhibit Boltzmann's H-theorem, which states that entropy of the ideal gas is a nondecreasing function of time.

Ludwig Boltzmann (1844–1906) considered time evolution of the ideal gas and in 1872 he derived an equation which implies that the entropy of the gas density does not decrease in time. This effect holds under a certain apparently intuitive assumption about the number of colliding gas particles. Here we will study a simplified version of Boltzmann's equation for which the H-theorem holds.

We will assume that the gas fills some volume and its space density is constant. In contrast, the velocity distribution in a space element will be inhomogeneous and evolving. The velocity distribution evolves because of collisions of gas particles of different velocities. To simplify the description of collision effects, we will assume that particles are ideal balls. Yet we need to assume something about the probability of collisions. It is intuitive to assume that gas particles remain in a state of a molecular chaos. Thus, following Boltzmann, we will make an important assumption that the number of collisions in a time unit is proportional to the product of densities of the colliding particles. This assumption about the number of collisions is called Stosszahlansatz.

Let us introduce the necessary notation. Symbol $\rho_t(\boldsymbol{v})$ will denote the density of particles with velocity $\boldsymbol{v}$ at while t. Moreover, $\boldsymbol{v}$ and $\boldsymbol{v}'$ will denote the velocities of particles before the collision whereas $\boldsymbol{v}''$ and $\boldsymbol{v}'''$ are the velocities of the particles after the collision. For $\boldsymbol{u}$ being the vector that joins the centers of the particles at the while of collision, $\boldsymbol{v}''$ and $\boldsymbol{v}'''$ are the functions of $\boldsymbol{v}$, $\boldsymbol{v}'$ and $\boldsymbol{u}$. In the time evolution of $\rho_t(\boldsymbol{v})$ there are two effects. First, $\rho_t(\boldsymbol{v})$ increases when two particles of velocities $\boldsymbol{v}''$ and $\boldsymbol{v}'''$ and vector $\boldsymbol{u}$ collide. Second, $\rho_t(\boldsymbol{v})$ decreases when a particle of velocity $\boldsymbol{v}$ collides with another particle. The aforementioned assumption about the number of collision, Stosszahlansatz, implies that the number of the first type collisions is proportional to $\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''')\tau(\boldsymbol{v},\boldsymbol{v}',\boldsymbol{u})$ and the number of the second type collisions is proportional to $\rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})\tau(\boldsymbol{v},\boldsymbol{v}',\boldsymbol{u})$, where $\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})$ is a certain function called cross-section. This yields equation

$$\frac{\partial \rho_t(\boldsymbol{v})}{\partial t} = \int \big[\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''') - \rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})\big]\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u}.$$

This equation is called the Boltzmann equation.

By the principles of classical mechanics, the collisions are invertible. That is, for a fixed $\boldsymbol{u}$, if $-\boldsymbol{v}''$ and $-\boldsymbol{v}'''$ are the velocities *before* the collision then $-\boldsymbol{v}$ and $-\boldsymbol{v}'$ are the velocities *after* the collision. Further analysis of the collision mechanics leads to these symmetry conditions for the cross-section $\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})$:

$$\begin{aligned}
&\int f(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{v}'', \boldsymbol{v}''')\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u} \\
&= \int f(\boldsymbol{v}', \boldsymbol{v}, \boldsymbol{v}''', \boldsymbol{v}'')\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u} \\
&= \int f(\boldsymbol{v}'', \boldsymbol{v}''', \boldsymbol{v}, \boldsymbol{v}')\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u} \\
&= \int f(\boldsymbol{v}''', \boldsymbol{v}'', \boldsymbol{v}', \boldsymbol{v})\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u}, \qquad (12.1)
\end{aligned}$$

where $f(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{v}'', \boldsymbol{v}''')$ is an arbitrary function of velocities. These equalities will be used further.

For example, we note first that the total number of particles remains constant. Denote the number of particles at while t as

$$N(t) = \int \rho_t(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}.$$

In the following we shall assume that the density $\rho_t(\boldsymbol{v})$ is sufficiently regular (namely, its derivative is continuous) so that the order of integration and differentiation can be switched.

Theorem 12.1. *We have*

$$\frac{dN(t)}{dt} = 0.$$

Proof. Observe that

$$\begin{aligned}
\frac{dN(t)}{dt} &= \int \frac{\partial \rho_t(\boldsymbol{v})}{\partial t}\,\mathrm{d}\boldsymbol{v} \\
&= \int \big[\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''') - \rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})\big]\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u}.
\end{aligned}$$

Using equality (12.1), we may symmetrize this expression as

$$\begin{aligned}
\frac{dN(t)}{dt} &= \frac{1}{2}\int \big[\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''') - \rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})\big]\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u} \\
&\quad + \frac{1}{2}\int \big[\rho_t(\boldsymbol{v})\rho_t(\boldsymbol{v}') - \rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''')\big]\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{v}'\,\mathrm{d}\boldsymbol{u} = 0.
\end{aligned}$$

Now, let us introduce the entropy of the velocity distribution:

$$H(t) = -\int \rho_t(\boldsymbol{v}) \ln \rho_t(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v}.$$

We can show that the entropy grows.

Theorem 12.2 (H-theorem). *We have*

$$\frac{dH(t)}{dt} \geq 0.$$

Proof. Notice that

$$\begin{aligned}\frac{dH(t)}{dt} &= -\int [\ln \rho_t(\boldsymbol{v}) + 1] \frac{\partial \rho_t(\boldsymbol{v})}{\partial t}\, \mathrm{d}\boldsymbol{v} \\ &= -\int [\ln \rho_t(\boldsymbol{v}) + 1] \left[\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''') - \rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})\right] \tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{v}'\, \mathrm{d}\boldsymbol{u}.\end{aligned} \tag{12.2}$$

Now using equality (12.1), we may symmetrize the expression on the right hand side of (12.2). Namely, we obtain

$$\begin{aligned}\frac{dH(t)}{dt} &= -\frac{1}{4}\int [\ln \rho_t(\boldsymbol{v}) + 1] \left[\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''') - \rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})\right] \tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{v}'\, \mathrm{d}\boldsymbol{u} \\ &\quad -\frac{1}{4}\int [\ln \rho_t(\boldsymbol{v}') + 1] \left[\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''') - \rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})\right] \tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{v}'\, \mathrm{d}\boldsymbol{u} \\ &\quad -\frac{1}{4}\int [\ln \rho_t(\boldsymbol{v}'') + 1] \left[\rho_t(\boldsymbol{v})\rho_t(\boldsymbol{v}') - \rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''')\right] \tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{v}'\, \mathrm{d}\boldsymbol{u} \\ &\quad -\frac{1}{4}\int [\ln \rho_t(\boldsymbol{v}''') + 1] \left[\rho_t(\boldsymbol{v})\rho_t(\boldsymbol{v}') - \rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''')\right] \tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{v}'\, \mathrm{d}\boldsymbol{u} \\ &= \frac{1}{4}\int \rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''')(\ln x)(x-1)\tau(\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{u})\, \mathrm{d}\boldsymbol{v}\, \mathrm{d}\boldsymbol{v}'\, \mathrm{d}\boldsymbol{u} \geq 0,\end{aligned}$$

where $x = \rho_t(\boldsymbol{v}')\rho_t(\boldsymbol{v})/\rho_t(\boldsymbol{v}'')\rho_t(\boldsymbol{v}''')$ and $(\ln x)(x-1) \geq 0$.

The fact that entropy grows is paradoxical in view of the assumptions that particle collisions are invertible. The reason for the entropy growth lies, however, in the adopted assumption about the number of collisions. The Stosszahlansatz holds true only for a particular initial state of the gas. This state is likely enough and the Boltzmann equation is a very good approximation of the time evolution of gases that we observe in nature.

The time evolution of the ideal gas density can be linked with the next problem, which is the maximum entropy modeling. Namely, we know that the entropy of the ideal gas grows but a few other statistics remain constant. We may thus suppose that the system tends to an equilibrium state $\rho(\boldsymbol{v}) = \lim_{t\to\infty} \rho_t(\boldsymbol{v})$ where the entropy is maximal given other constraints. Formally, we have

$$-\int \rho(\boldsymbol{v}) \ln \rho(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v} = \max \tag{12.3}$$

given some additional conditions. By the principles of classical mechanics, there are three constraints for the ideal gas. First, the total number of particles is constant. Thus let the density be normalized as

$$\int \rho(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} = 1. \tag{12.4}$$

Second, the total momentum (i.e., the sum of velocities) is constant. Assuming that the gas does not move globally, we obtain

$$\int \rho(\boldsymbol{v})\boldsymbol{v}\,\mathrm{d}\boldsymbol{v} = 0. \tag{12.5}$$

Third, the total kinetic energy is constant. Assuming that the average kinetic energy per particle is σ^2, we have

$$\int \rho(\boldsymbol{v})\boldsymbol{v}^2\,\mathrm{d}\boldsymbol{v} = \sigma^2. \tag{12.6}$$

We assume that there are no other constraints. As we will prove, the solution of equation (12.3) given conditions (12.4)–(12.6) is unique and is given by the three-dimensional Gauss distribution

$$\rho(\boldsymbol{v}) = \left[\frac{1}{2\pi\sigma^2}\right]^{3/2} \exp\left[-\frac{\boldsymbol{v}^2}{2\sigma^2}\right].$$

Let us show where this equation comes from. The general problem of maximum entropy modeling is as follows:

Problem 12.1 (maximum entropy). Find the probability density ρ that maximizes entropy

$$H(\rho) = -\int \rho(x)\ln\rho(x)\,\mathrm{d}x \tag{12.7}$$

given constraints:

$$\int \rho(x)\,\mathrm{d}x = 1, \tag{12.8}$$

$$\int \rho(x)T_i(x)\,\mathrm{d}x = \alpha_i, \quad 1 \le 1 \le m. \tag{12.9}$$

Similar problems of maximizing entropy given some constraints appear in many applications, in machine learning in particular. The solution is this.

Theorem 12.3. *If there exists density*

$$\rho^*(x) = \exp\left[\lambda_0^* + \sum_{i=1}^{m} \lambda_i^* T_i(x)\right],$$

where λ_i^* are chosen so that ρ^* satisfies conditions (12.8)–(12.9), then ρ^* maximizes entropy (12.7) on the space of probability densities that satisfy (12.8)–(12.9).

Remark 1: We can see that the solution of the maximum entropy problem is an exponential family with T_i being minimal sufficient statistics.
Remark 2: In problems of machine learning, we encounter discrete rather than continuous distributions. The solution of the maximum entropy problem in that case is analogous, with probabilities replacing probability densities.
Remark 3: In certain maximization problems, there exists no ρ^* that satisfies (12.8)–(12.9). In such cases there is no distribution having the maximal entropy. This happens for example for constraints $\int x^k \rho(x)\,\mathrm{d}x = \alpha_k$, where $k = 0, 1, 2, 3$. In that case we would obtain

$$\rho(x) = \exp\left[\lambda_0 + \lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3\right],$$

which cannot be normalized for any $\lambda_3 \neq 0$ because $\rho(x)$ tends to infinity for either for $x \to \infty$ or $x \to -\infty$.

Proof. Let ρ satisfy constraints (12.8)–(12.9). We observe that Kullback-Leibler divergence $D(\rho||\rho^*)$ is nonnegative and equals 0 if and only if densities ρ and ρ^* are equal. Hence we obtain

$$\begin{aligned} H(\rho) &= -\int \rho(x) \ln \rho(x)\,\mathrm{d}x \\ &= -D(\rho||\rho^*) - \int \rho(x) \ln \rho^*(x)\,\mathrm{d}x \\ &\leq -\int \rho(x) \ln \rho^*(x)\,\mathrm{d}x \\ &= -\int \rho(x) \left[\lambda_0^* + \sum_{i=1}^m \lambda_i^* T_i(x)\right] \mathrm{d}x \\ &= -\int \rho^*(x) \left[\lambda_0^* + \sum_{i=1}^m \lambda_i^* T_i(x)\right] \mathrm{d}x \\ &= -\int \rho^*(x) \ln \rho^*(x)\,\mathrm{d}x = H(\rho^*), \end{aligned}$$

with the equality if and only if ρ and ρ^* are equal. This proves the claim.

The remaining problem is to find the suitable λ_i^*. The computation involves a few steps. The first step is to consider a Lagrangian function.

Theorem 12.4. *Consider the density ρ_λ and the Lagrangian function $L(\lambda)$ defined as*

$$\rho_\lambda(x) = \exp\left[\sum_{i=1}^m \lambda_i T_i(x) - \ln Z(\lambda)\right], \tag{12.10}$$

$$L(\lambda) = \ln Z(\lambda) - \sum_{i=1}^m \lambda_i \alpha_i, \tag{12.11}$$

where the canonical sum $Z(\lambda)$ is

$$Z(\lambda) = \int \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] \mathrm{d}x.$$

Function $L(\lambda)$ has a single minimum and the vector $\lambda^* = (\lambda_1^*, ..., \lambda_k^*)$ for which density (12.10) satisfies conditions (12.9) is the solution of the equation

$$\lambda^* = \arg\min_{\lambda} L(\lambda). \tag{12.12}$$

Proof. We have

$$\begin{aligned}\frac{\partial L(\lambda)}{\partial \lambda_j} &= \frac{1}{Z(\lambda)} \int T_j(x) \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] \mathrm{d}x - \alpha_j \\ &= \int \rho_\lambda(x) T_j(x)\, \mathrm{d}x - \alpha_j.\end{aligned}$$

Hence the Lagrangian has an extremum if and only if ρ_λ satisfies conditions (12.9). Further analysis shows that there is only one extremum and it is a minimum because the Lagrangian is convex. Indeed we obtain

$$\begin{aligned}\frac{\partial^2 L(\lambda)}{\partial \lambda_j \partial \lambda_k} &= \frac{\partial}{\partial \lambda_k}\left[\frac{1}{Z(\lambda)} \int T_j(x) \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] \mathrm{d}x\right] \\ &= -\frac{1}{[Z(\lambda)]^2}\left[\int T_k(x) \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] \mathrm{d}x\right] \\ &\qquad \times \left[\int T_j(x) \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] \mathrm{d}x\right] \\ &\quad + \frac{1}{Z(\lambda)} \int T_k(x) T_j(x) \exp\left[\sum_{i=1}^{m} \lambda_i T_i(x)\right] \mathrm{d}x.\end{aligned}$$

Writing $\mathbf{E}_\lambda T = \int \rho_\lambda(x) T(x)\, \mathrm{d}x$, we have

$$\frac{\partial^2 L(\lambda)}{\partial \lambda_j \partial \lambda_k} = \mathbf{E}_\lambda (T_j T_k) - \mathbf{E}_\lambda T_j \mathbf{E}_\lambda T_k = \mathbf{E}_\lambda \left[T_j - \mathbf{E}_\lambda T_j\right]\left[T_k - \mathbf{E}_\lambda T_k\right].$$

We observe that the second derivative of the Lagrangian is a covariance matrix, which is nonnegative definite, i.e.,

$$\sum_{j,k=1}^{m} a_j \frac{\partial^2 L(\lambda)}{\partial \lambda_j \partial \lambda_k} a_k = \mathbf{E}_\lambda \left[\sum_{j=1}^{m} a_j \left[T_j - \mathbf{E}_\lambda T_j\right]\right]^2 \geq 0.$$

Hence the Lagrangian is convex and the there is only one extremum.

The final step to find λ_i^* is to minimize the Lagrangian $L(\lambda)$. In many problems of machine learning this can be only done numerically. The suitable minimization can be performed using generic minimization algorithms, e.g. minimization by conjugate gradients, or algorithms dedicated for function (12.11), e.g. the iterative scaling (Berger et al., 1996).

Exercises

1. *(Entropy growth)* Show that $H(X_n|X_1) \geq H(X_{n-1}|X_1)$ for a stationary Markov chain.
2. Let the value of X be an ordered n-tuple and let T be a random permutation of an n-tuple, probabilistically independent from X. Show that $H(TX) \geq H(X)$.
3. *(Differential entropy)* Let $\rho(x)$ be a probability density limited to a certain volume V, i.e, $\rho(x) = 0$ for $x \notin V$. Show that

$$-\int_V \rho(x) \ln \rho(x) \leq \ln V,$$

   where the equality holds if and only if $\rho(x) = 1/V$ for almost all $x \in V$.
4. *(Maximum entropy)* Show that every probability density is a maximum entropy density under a certain constraint.
   Hint: Show that ρ^* is the maximum entropy density under constraint $\int \rho(x) \ln \rho^*(x)\,\mathrm{d}x = \alpha$. Find the appropriate α.
5. Let $\rho(x)$ be a probability density of a nonnegative random variable with mean μ. What is the maximum value of entropy $H(\rho)$ in that case?
6. Prove Lemma 6.3.
7. Consider *cross entropy*

$$H(\rho||\sigma) := -\int \rho(x) \ln \sigma(x)\,\mathrm{d}x = H(\rho) + D(\rho||\sigma).$$

   Let ρ satisfy conditions (12.8)–(12.9) and let ρ_λ be of form (12.10). Show that cross entropy $H(\rho||\rho_\lambda)$ achieves minimum for $\rho_\lambda = \rho^*$ where ρ^* is the maximum entropy distribution for problem (12.8)–(12.9).

13

Kolmogorov complexity

Kolmogorov complexity. The information-theoretic Gödel theorem. Incompressibility method. Oscillations of Kolmogorov complexity.

In Shannon's information theory, the amount of information carried by a random variable depends on the ascribed probability distribution. Andrey Kolmogorov (1903–1987), the founding father of modern probability theory, remarked that it is extremely hard to imagine a reasonable probability distribution for utterances in natural language but we can still estimate their information content (Kolmogorov, 1965). For that reason Kolmogorov proposed to define the information content of a particular string in a purely algorithmic way. A similar approach has been proposed a year earlier by Ray Solomonoff (1926–2009), who sought for optimal inductive inference (Solomonoff, 1964). The newly emerged research paradigm has been called algorithmic information theory. In the following chapters we will be occupied with rudiments of that theory.

The fundamental object of algorithmic information theory is Kolmogorov complexity of a string. The Kolmogorov complexity constitutes an algorithmic analogue of Shannon entropy. It is defined as the length of the shortest program for a Turing machine such that the machine prints out the string and halts. The Turing machine itself is defined as a deterministic finite state automaton which moves along one or more infinite tapes filled with symbols from a fixed finite alphabet and which may read and write individual symbols. The automaton has a distinct start state, from which the computation begins, and a distinct halt state, at which the computation ends. The concrete value of Kolmogorov complexity depends on the used Turing machine but many properties of Kolmogorov complexity are universal.

Formally, a Turing machine is defined as a 6-tuple $T = (Q, s, h, \Gamma, B, \delta)$, where

1. Q is a finite, nonempty set of states,
2. $s \in Q$ is the start state,
3. $h \in Q$ is the halt state,
4. Γ is a finite, nonempty set of symbols,
5. $B \in \Gamma$ is the blank symbol,
6. $\delta : Q \setminus \{h\} \times \Gamma \to Q \times \Gamma \times \{L, R\}$ is a function called a transition function, where L is the left shift and R is the right shift.

This formal description translates to machine operation in the following way. The machine is given an infinite tape $(X_i)_{i \in \mathbb{Z}}$ filled with symbols from Γ. In this chapter we will assume that $\Gamma = \{0, 1, B\}$. The initial state of the tape is

$X_1^{|p|+1} = pB$, where $p \in \{0,1\}^*$, and $X_i = 0$ for other X_i. The initial machine state is s and the machine reads symbol X_1. Subsequently, the machine shifts in discrete steps along the tape in the way prescribed by the transition function. Namely, if the machine is in state a and reads symbol b then, for $\delta(a,b) = (a', b', M)$, the machine writes symbol b' on the tape, moves along the tape one symbol to the left or to the right if $M = L$ or $M = R$, respectively, and assumes state a' in the next step. This procedure takes place until the machine reaches the halt state h. Then the computation stops.

The set of such Turing machines will be denoted as $\mathcal{T}$. In the following we will say that machine $T \in \mathcal{T}$ halts on input (program) $p \in \{0,1\}^*$ and returns string $w \in \{0,1\}^*$ if:

(A) The head in the start state reads symbol X_1, and the initial state of the tape is $X_1^{|p|+1} = pB$. We put $X_i = 0$ for $i < 1$ and $i > |p| + 1$.
(B) The head in the halt state reads symbol X_j, and the final state of the tape is $X_j^{|w|+j} = wB$.

Assuming that condition (A) is satisfied, we write

$$T(p) = \begin{cases} w, & \text{if condition (B) is satisfied,} \\ \infty, & \text{if condition (B) is not satisfied for any } w \in \{0,1\}^* \end{cases}$$

The initial and the final state of machine T is depicted in Figure 7.

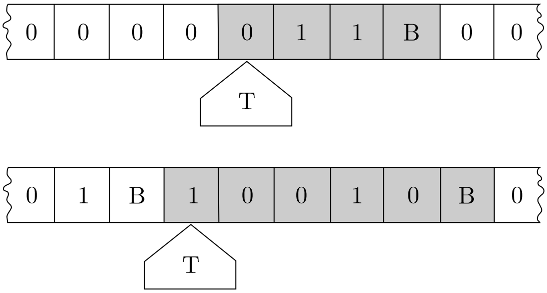

Fig. 7. The initial and the final state of machine T for $T(011) = 10010$.

Definition 13.1 (Kolmogorov complexity). Kolmogorov complexity $C_T(w)$ *of a string* $w \in \{0,1\}^*$ *with respect to machine* $T \in \mathcal{T}$ *is defined as*

$$C_T(w) := \min_{p \in \{0,1\}^*} \{|p| : T(p) = w\}.$$

In the following, we will discuss Kolmogorov complexity with respect to a universal machine.

Definition 13.2 (universal Turing machine). *Machine $U \in \mathcal{T}$ is called* universal *if for each machine $T \in \mathcal{T}$ there exists a string $u \in \{0,1\}^*$ such that*

$$U(up) = T(p)$$

for all $p \in \{0,1\}^*$.

Universal machines exist. The proof is not complicated but tedious. For a sketch of the proof we refer to Li and Vitányi (2008, Example 1.7.4). Briefly speaking, machine U operates in this way: Given u it first enumerates descriptions of machines in set $\mathcal{T}$ until it finds a description of machine T identified by index u. Then, having the description of machine T, machine U simulates machine T on input p.

The reason for considering universal machines is that Kolmogorov complexity for a universal machine is almost invariant up to an additive constant.

Theorem 13.1 (invariance theorem). *For any two universal machines U and U' there exists a constant c such that*

$$|C_U(w) - C_{U'}(w)| \le c$$

for any string $w \in \{0,1\}^*$.

Proof. We have $U(up) = U'(p)$ and $U'(u'p) = U(p)$ for certain strings u and u'. Hence $C_U(w) \le C_{U'}(w) + |u|$ and $C_{U'}(w) \le C_U(w) + |u'|$.

Thus for further considerations we will choose a certain universal machine U as a reference machine to determine Kolmogorov complexity.

Definition 13.3 (Kolmogorov complexity II). *Let $U \in \mathcal{T}$ be a selected universal machine. We will put*

$$C(w) := C_U(w).$$

This quantity will be called (plain) Kolmogorov complexity.

Besides complexity of strings, we will also discuss complexity of arbitrary discrete objects such as rational numbers or automata. Those objects are formally defined as finitely nested n-tuples of binary strings and natural numbers. Thus, let $\mathbb{A}$ be the set of finitely nested n-tuples of binary strings and natural numbers. Moreover let $E(n)$ be the Elias omega code of natural number n. We will write $\bar{w} := E(|w|)w$ for strings $w \in \{0,1\}^*$. Notation $\langle a_1, a_2, .., a_{n-1}, a_n\rangle$ denotes an n-tuple, where $a_i \in \mathbb{A}$. By recursion, we define the coding function $\phi : \mathbb{A} \to \{0,1\}^*$ as

$$\begin{cases} \phi\left(\langle a_1, a_2, .., a_{n-1}, a_n\rangle\right) := 101\phi(a_1)0\phi(a_2)0...0\phi(a_{n-1})0\phi(a_n)100, \\ \phi(w) := 110\bar{w}, \text{ if } w \in \{0,1\}^*, \\ \phi(n) := 111\bar{w}, \text{ if } n \in \mathbb{N}, \text{ and } w \text{ is the binary expansion of number } n. \end{cases}$$

Moreover, we put $\phi(\infty) := \infty$, where ∞ denotes the indefinite value. Kolmogorov complexity of an object $a \in \mathbb{A} \setminus \{0,1\}^*$ will be defined as $C_T(a) := C_T(\phi(a))$ and $C(a) := C(\phi(a))$, respectively.

It is also important to introduce the concept of a computable function.

Definition 13.4 (computable function). *We say that machine T or program p* computes *partial function $f : \mathbb{A} \to \mathbb{A} \cup \{\infty\}$ if $T(\phi(a)) = \phi(f(a))$ or $U(p\phi(a)) = \phi(f(a))$, respectively, for all $a \in \mathbb{A}$. A function that can be computed by a program is called* recursive *or* computable*.*

Having all these definitions, we may prove a few interesting theorems. First, we will give a simple bound for Kolmogorov complexity of a string.

Theorem 13.2. *There exists a constant c such that $C(w) \leq |w| + c$ for any string w.*

Proof. A certain program that generates string w has form "print w".

The second result states that Kolmogorov complexity cannot be computed in general. Some people consider this property a drawback. We would like however to put forward an alternative interpretation, namely, that incomputability of Kolmogorov complexity is just yet another interesting property.

Theorem 13.3. *Kolmogorov complexity $C(\cdot)$ is not a computable function.*

Proof. Assume that there exists a program q which computes $C(w)$ for any w. Then there exists a program p which uses q as a subroutine to print out the shortest string w such that $C(w) > |p|$. Namely, such a program p inspects strings w sorted according to their length, computes $C(w)$ using subroutine q and checks whether $C(w) > |p|$. It is obvious that this inequality will hold for a certain w because $C(w)$ is unbounded. But by the definition of Kolmogorov complexity, we have $C(w) \leq |p|$ for the same string. Hence our assumption about the existence of program q was false.

A similar search for the shortest element appears in the proof of the information-theoretic Gödel theorem, another fundamental result in the algorithmic information theory. A formal inference system is a finite collection of axioms and inference rules. The system is called *consistent* if it is not possible to prove both a statement and its negation, whereas the system is called *sound* if only true propositions can be proved. (Thus a sound system is consistent.) According to the information-theoretic Gödel theorem, in any sound formal inference system it is not possible to prove incompressibility of any string which is substantially longer than the definition of this formal system. A binary string is called *incompressible* if $C(x_1^n) \geq n$. The following theorem is due to Chaitin (1975b).

Theorem 13.4 (information-theoretic Gödel theorem). *For any sound formal inference system, there exists a constant K such that propositions "$C(w) > K$" are unprovable in that system.*

Proof. Let us assume that for any number K there exists a proof of proposition "$C(w) > K$". Then we may construct a program of length of L which searches all proofs of the formal system to find the first proof that a certain string w has the complexity greater than K and then prints out that string. Then we have $C(w) \leq L$. Since $L < c + \log K$, we obtain a contradiction for sufficiently large K's.

Although incompressibility is unprovable, there are infinitely many incompressible strings because there are not so many short enough programs. A string x_1^n will be called *c-incompressible* if $C(x_1^n) \geq n - c$. In particular, a 0-incompressible string will be called incompressible.

Theorem 13.5. *There exist at least $2^n - 2^{n-c} + 1$ distinct c-incompressible strings of length n.*

Proof. There exist at most $2^{n-c} - 1$ distinct programs of length strictly smaller than $n - c$ and there exists 2^n distinct strings of length n. Subtracting the latter from the former, we obtain the desired bound.

In particular, there is at least one incompressible string of length n and at least a half of strings of length n is 1-incompressible.

The existence of incompressible strings may be used for nonconstructive proofs of a few asymptotic bounds. This technique is called *incompressibility method*. The first example will concern the density of prime numbers.

Theorem 13.6. *Let $\pi(n)$ be the number of primes that do not exceed n. For infinitely many n we have $\pi(n) \geq \log n/(\log \log n + c)$.*

Proof. Let $p_1, p_2, ..., p_{\pi(n)}$ be the primes that are smaller than or equal n. We may represent $n = p_1^{e_1} p_2^{e_2} ... p_{\pi(n)}^{e_{\pi(n)}}$, where $e_i \leq \log n$ because $p_i \geq 2$. Thus the length of the shortest description of n is upper bounded by $\pi(n) \log \log n + O(1)$. If n has an incompressible binary representation then $\log n \leq \pi(n) \log \log n + O(1)$. Hence we obtain the desired bound.

The given bound is not very good because we have $\lim_{n\to\infty} \frac{\pi(n)}{n/\ln n} = 1$.

The second example of incompressibility method concerns language and automata theory. A *formal language* L is a subset of the set of strings Γ^*, i.e., $L \subset \Gamma^*$. We say that a formal language L is *regular* if it is recognized by a certain deterministic finite state automaton, i.e., when $w \in L$ holds if and only if the automaton accepts w. The formal definition of the latter concept is as follows. A *deterministic finite state automaton* is a 5-tuple $M = (Q, s, h, \Gamma, \delta)$, where

1. Q is a finite, nonempty set of states,
2. s is the start state,
3. h is the halt state,
4. Γ is a finite, nonempty set of symbols,
5. $\delta : Q \setminus \{h\} \times \Gamma \to Q$ is a function called a transition function.

We say that the automaton M accepts a word $x_1^n \in \Gamma^*$ if there exists a sequence of states $r_0^n \in Q^*$ such that $r_0 = s$, $r_{i+1} = \delta(r_i, x_{i+1})$ for $i \in \{0, 1, ..., n-1\}$, and $r_n = h$.

Theorem 13.7. *Language $\left\{0^k 1^k : k \in \mathbb{N}\right\}$ is not regular.*

Proof. Assume that language $\{0^k1^k : k \in \mathbb{N}\}$ is regular. Then the language is recognized by a certain deterministic finite state automaton. The number k or the string 1^k can be encoded by specifying this automaton and the state of the automaton after reading the string 0^k. We obtain a contradiction if $C(k)$ is greater than the double Kolmogorov complexity of the automaton plus a constant.

The technique used in the above proposition may be generalized to a lemma that characterizes regular languages.

Lemma 13.1. *Let language $L \subset \{0,1\}^*$ be regular and $L_x = \{y : xy \in L\}$. Then there exists a constant c such that for each x we have $C(y) \leq C(n) + c$ if y is the n-th element of L_x. (We assume that the elements of L_x are sorted according to length if they are of different length and lexicographically if they have the same length.)*

Proof. Let y be the n-th element of L_x. String y can be described by (i) specifying the deterministic finite state automaton that recognizes L, (ii) giving the state of the automaton after reading x, and (iii) giving the number n. The first two parts require c bits, whereas the third part requires $C(n)$ bits. Hence we obtain $C(y) \leq C(n) + c$.

Using this lemma we can obtain another known result.

Theorem 13.8. *Language $\{1^p : p \text{ is prime}\}$ is not regular.*

Proof. Assume that $L = \{1^p : p \text{ is prime}\}$ is regular. Let us consider $xy = 1^p$ where p is the $(k+1)$-st prime. Put $x = 1^{p'}$ where p is the k-th prime. Then $y = 1^{p-p'}$ is the first element of L_x. Hence by Lemma 13.1 we have $C(p - p') \leq C(y) + O(1) \leq O(1)$. But differences $p - p'$ are unbounded hence $C(p - p')$ is also unbounded. We obtained a contradiction so our assumption was false.

Finally, let us consider properties of incompressible strings again. Although there are infinitely many incompressible strings, there is no such infinite sequence $(x_i)_{i=1}^{\infty}$ that $C(x_1^n) = n - c$ for each n. This phenomenon is called *oscillations of Kolmogorov complexity.*

Theorem 13.9. *Let $(x_i)_{i=1}^{\infty}$ be an infinite binary sequence. For infinitely many M we have $C(x_1^M) \leq M - \log M + O(1)$.*

Proof. Consider an arbitrary m. Let x_1^m be the binary expansion of a number n stripped of the initial digit 1. Then we have $C(x_1^{m+n}) \leq O(1) + C(x_{m+1}^{m+n}) \leq O(1) + n$ because we may compute x_1^m given (the length of) x_{m+1}^{m+n}. Put $M = m + n$. We have $m \geq \log n + O(1) = \log(M - m) + O(1) \geq \log M + O(1)$. Hence $C(x_1^M) \leq n + O(1) \leq M - \log M + O(1)$.

Exercises

1. *(Continuity of Kolmogorov complexity)* Show that for natural numbers x, y we have $|C(x+y) - C(x)| \leq 2 \log y + c$.
2. Let string x satisfy $C(x) \geq n - c$, where $n = |x|$. Show that $C(y), C(z) \geq n/2 - c$ for $x = yz$ and $|y| = |z|$.
3. Assume that the elements of $\{1, .., n\}$ are uniformly distributed with probability $1/n$. Compute the expected value of $C(x)$, $1 \leq x \leq n$.
4. Prove that language $L = \{xx : x \in \mathbb{X}^*\}$ is not regular.
5. Let $x^R = x_n x_{n-1} ... x_1$ be the reflection of string $x = x_1 x_2 ... x_n$. Prove that language $L = \left\{xx^R : x \in \mathbb{X}^*\right\}$ is not regular.
   Hint: Consider $w = (01)^n$.
6. Let # denote a symbol out of the alphabet which is used to write down strings x, y, and z. Prove that language $L = \{x\#y\#z : xy = z\}$ is not regular.
   Hint: Consider $w = 0^n\#\#$.
7. Prove that language $L = \{x\#y : x \text{ appears in a discontinuous way in } y\}$ is not regular.
   Hint: Consider $w = 0^n\#$.
8. Prove that language $L = \{0^n 1^m : m > n\}$ is not regular.
   Hint: Consider $w = 0^n$.
9. Prove that language $L = \{x\#y : \text{at least half of } x \text{ appears in } y\}$ is not regular.
   Hint: Consider $w = 0^{2n}\#$.
10. Let $\text{GCD}(i, j)$ be the greatest common divisor of i and j. Prove that language $L = \left\{0^i 1^j : \text{GCD}(i, j) = 1\right\}$ is not regular.
    Hint: Consider $w = 0^{(p-1)!}11$, where p is the n-th prime number.

14

Prefix-free complexity

Prefix-free Kolmogorov complexity. Links between Kolmogorov complexity and entropy. Algorithmic probability. Symmetry of algorithmic information.

In Chapter 13, we have discussed the so called plain Kolmogorov complexity. A bit different definition of Kolmogorov complexity is convenient to discuss links between complexity and entropy. This complexity is called the prefix-free Kolmogorov complexity. The difference to the plain complexity lies in using a different Turing machine. The fundamental idea is to force that the accepted programs form a prefix-free set. This can be done in a few ways. Here we will use the construction by Gregory Chaitin (1947–), described in his seminal paper Chaitin (1975a).

Thus we will consider a machine that has two tapes: a bidirectional tape $(X_i)_{i\in\mathbb{Z}}$ filled with symbols 0, 1, and B (blank symbol) and a unidirectional tape $(Y_k)_{k\in\mathbb{N}}$ filled with symbols 0 and 1. The head of the machine may move in both directions along tape $(X_i)_{i\in\mathbb{Z}}$ and only in the direction of growing k along tape $(Y_k)_{k\in\mathbb{N}}$. Tape $(X_i)_{i\in\mathbb{Z}}$ can be both read and written, tape $(Y_k)_{k\in\mathbb{N}}$ is read only. The set of such machines is denoted $\mathcal{S}$. We say that machine $S \in \mathcal{S}$ halts on input $(p, q) \in \{0,1\}^* \times (\{0,1\}^* \cup \{0,1\}^{\mathbb{N}})$ and returns string $w \in \{0,1\}^*$ if:

(A) The head in the start state matches symbols Y_1 and X_1, and the initial state of the tapes is $Y_1^{|p|} = p$ and $X_1^{|q|+1} = qB$ or $X_1^{\infty} = q$ if q is an infinite sequence. Besides we put $X_i = 0$ for $i < 1$ and $i > |q| + 1$ if q is finite.

(B) The head in the halt state matches symbols $Y_{|p|}$ and X_j, and the final state of the tape $(X_i)_{i\in\mathbb{Z}}$ is $X_j^{|w|+j} = wB$.

Assuming that condition (A) is satisfied, we write

$$S(p|q) = \begin{cases} w, & \text{if condition (B) is satisfied,} \\ \infty, & \text{if condition (B) is not satisfied for any } w \in \{0,1\}^*. \end{cases}$$

In other words, $S(p|q) = \infty$, if the machine does not reach the halt state or reaches the halt state with the head matching symbol Y_k, where $k \neq |p|$. The initial and the final state of machine S is depicted in Figure 8. It can be easily seen that for a given q the set of strings p such that machine S halts on input (p, q) is prefix-free. Such strings are called *self-delimiting programs*.

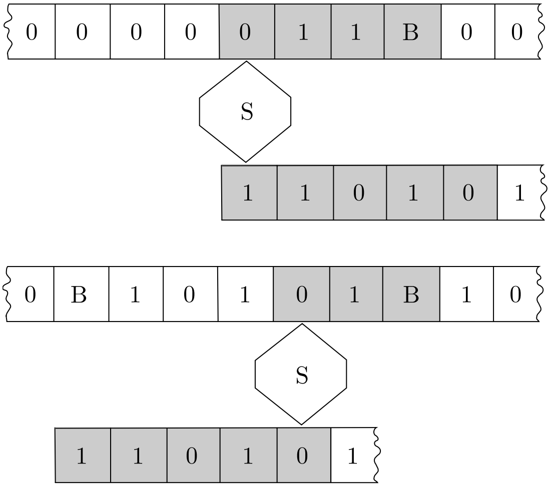

Fig. 8. The initial and the final state of machine S for $S(11010|011) = 01$.

Definition 14.1 (prefix-free complexity). Prefix-free conditional Kolmogorov complexity $K_S(w|q)$ *of a string* $w \in \{0,1\}^*$ *given string* $q \in \{0,1\}^*$ *and machine* $S \in \mathcal{S}$ *is defined as*

$$K_S(w|q) := \min_{p \in \{0,1\}^*} \big\{ |p| : S(p|q) = w \big\}.$$

Analogously as in Chapter 13, we introduce universal machines.

Definition 14.2 (universal machine). *Machine* $V \in \mathcal{S}$ *is called* universal *if for each machine* $S \in \mathcal{S}$ *there exists a string* $u \in \{0,1\}^*$ *such that*

$$V(up|q) = S(p|q)$$

for all $p \in \{0,1\}^*$ *and* $q \in \{0,1\}^* \cup \{0,1\}^{\mathbb{N}}$.

Such universal machines exist. Again, we will assume this fact without proof.

We also have the invariance theorem as for the plain complexity.

Theorem 14.1 (invariance theorem). *For any two universal machines* V *and* V' *there exists a constant* c *such that*

$$|K_V(w|q) - K_{V'}(w|q)| \le c$$

for any string $w \in \{0,1\}^*$ *and* $q \in \{0,1\}^* \cup \{0,1\}^{\mathbb{N}}$.

Proof. We have $V(up|q) = V'(p|q)$ and $V'(u'p|q) = V(p)$ for certain strings u and u'. Hence $K_V(w|q) \le K_{V'}(w|q) + |u|$ and $K_{V'}(w|q) \le K_V(w|q) + |u'|$.

Thus for further consideration we will choose a certain universal machine as a reference to determine prefix-free complexity.

Definition 14.3 (prefix-free complexity II). *Let $V \in \mathcal{S}$ be a certain universal machine. We put*

$$K(w|q) := K_V(w|q).$$

This quantity is called prefix-free conditional Kolmogorov complexity*.*

Moreover, unconditional complexities are defined as

$$K_S(w) := K_S(w|\lambda),$$
$$K(w) := K(w|\lambda),$$

where λ is the empty string. Similarly, we will use shorthand $V(p) := V(p|\lambda)$ for any other function V.

Now let us discuss complexities of objects different to strings. Let $\mathbb{A}$ be the set of discrete objects and $\phi : \mathbb{A} \to \{0,1\}^*$ be the coding function defined in Chapter 13. In this chapter we allow Turing machines that also process infinite binary sequences. Such sequences can be identified with tuples of both discrete objects and real numbers. Let $\mathbb{B}$ be the set of such objects and let $\psi : \mathbb{B} \to \{0,1\}^* \cup \{0,1\}^{\mathbb{N}}$ be the appropriate coding function, respectively. The complexity of objects $a \in \mathbb{A} \setminus \{0,1\}^*$ and $b \in \mathbb{B} \setminus (\{0,1\}^* \cup \{0,1\}^{\mathbb{N}})$ will be defined as

$$K_S(a|b) := K_S(\phi(a)|\psi(b)),$$
$$K(a|b) := K(\phi(a)|\psi(b)).$$

Analogous convention is used if one of the objects is a string from $\{0,1\}^*$ or an infinite sequence from $\{0,1\}^{\mathbb{N}}$. To increase readability, the notation of functions ϕ and ψ in any similar contexts will be also suppressed.

We will also define complexity of functions. First, the Kolmogorov complexity of a discrete partial function $f : \{0,1\}^* \cup \{0,1\}^{\mathbb{N}} \to \{0,1\}^* \cup \{\infty\}$ is defined as

$$K(f) := \min_p \left\{ |p| : \forall_{u \in \{0,1\}^* \cup \{0,1\}^{\mathbb{N}}} V(p|u) = f(u) \right\}.$$

In the above definition, we assume that the minimum of the empty set is infinity. In a similar way, we also define

$$K(f) := K(\phi \circ f \circ \psi^{-1})$$

for other discrete-valued partial functions $f : \mathbb{B} \to \mathbb{A} \cup \{\infty\}$. Hence we may extend the definition of a computable function from Chapter 13 as follows.

Definition 14.4 (computable discrete function). *A discrete-valued function $f : \mathbb{B} \to \mathbb{A} \cup \{\infty\}$ is called* computable *if $K(f) < \infty$.*

The more interesting case is that of real-valued functions. Let $\mathbb{Q} \subset \mathbb{A}$ be the set of rational numbers. We have three important cases:

Definition 14.5 (lower semicomputable real function). *A real-valued function $f : \mathbb{B} \to \mathbb{R}$ is called* lower semicomputable *if there is a computable function $A : \mathbb{B} \times \mathbb{N} \to \mathbb{Q}$ which satisfies $A(x, k+1) \geq A(x, k)$ for any $k \geq 1$ and*

$$\forall_{x \in \mathbb{B}} \lim_{k \to \infty} A(x, k) = f(x).$$

Definition 14.6 (upper semicomputable real function). *A real-valued function $f : \mathbb{B} \to \mathbb{R}$ is called* upper semicomputable *if there is a computable function $A : \mathbb{B} \times \mathbb{N} \to \mathbb{Q}$ which satisfies $A(x, k+1) \leq A(x, k)$ for any $k \geq 1$ and*

$$\forall_{x \in \mathbb{B}} \lim_{k \to \infty} A(x, k) = f(x).$$

Definition 14.7 (computable real function). *A real-valued function $f : \mathbb{B} \to \mathbb{R}$ is called* computable *if there is a computable function $A : \mathbb{B} \times \mathbb{N} \to \mathbb{Q}$ which satisfies*

$$\forall_{k \in \mathbb{N}} \forall_{x \in \mathbb{B}} \left| f(x) - A(x, k) \right| < 1/k.$$

In all three cases we put

$$K(f) := \min_{A} K(A).$$

An important example of an upper semicomputable function is Kolmogorov complexity. A function which is both upper and lower semicomputable can be shown computable. It can be shown that prefix-free Kolmogorov complexity is not computable, as in the plain case.

In further considerations, machines from set $\mathcal{T}$ will be called *plain* and machines from set $\mathcal{S}$ will be called *prefix-free.* We will also write $p \overset{+}{<} q$ and $p \overset{+}{>} q$ if there exists a constant c such that $p \leq q + c$ and $p \geq q + c$ holds respectively for all p and q. We will write $p \overset{\pm}{=} q$ when we have both $p \overset{+}{<} q$ and $p \overset{+}{>} q$. Thus the claim of Theorem 14.1 may be restated as

$$K_V(w|q) \overset{\pm}{=} K_{V'}(w|q).$$

Now let us discuss a few theorems that concern prefix-free complexity. The first is a bound for the prefix-free complexity which is analogous to Theorem 13.2. Let us remind that $E(n)$ is the Elias omega code for number n and $\bar{w} = E(|w|)w$.

Theorem 14.2. *We have*

$$K(w|q) \overset{+}{<} |\bar{w}| \, .$$

Proof. There is a prefix-free machine S which satisfies $S(\bar{w}|q) = w$. It generates w as follows. First it reads the Elias omega code for number $n = |w|$ from the unidirectional tape. Then it copies string w to the bidirectional tape. After copying the last symbol of w the machine halts because it knows its length from reading code $E(n)$. Hence we have the desired bound.

The second bound links prefix-free complexity and plain complexity.

Theorem 14.3. *Let $C(\cdot)$ be the plain Kolmogorov complexity and let $K(\cdot)$ be the prefix-free Kolmogorov complexity. We have*

$$C(w) \le K(w) \overset{+}{<} C(w) + K(C(w)).$$

Proof. The left bound follows because if $S(p) = w$ for a certain prefix-free machine S then $T(p) = w$ for a certain plain machine T. The proof of the right bound is as follows. Let p be the shortest program that satisfies $U(p) = w$ for the plain universal machine U and let p' be the shortest program that satisfies $V(p') = \phi(|p|)$ for the prefix-free universal machine V. Then $S(p'p) = w$ for a certain prefix-free machine S which does not depend on w. Hence $K(w) \overset{+}{<} |p'p|$.

In contrast to the plain Kolmogorov complexity, the prefix-free complexity is close to entropy. The following theorem is the first step to see it.

Theorem 14.4. *We have inequalities:*

1. $K(w|u) \overset{+}{<} K(w) \overset{+}{<} K(\langle u, w\rangle)$,
2. $K(uw) \overset{+}{<} K(\langle u, w\rangle) \overset{+}{<} K(u) + K(w|u) \overset{+}{<} K(u) + K(w)$,
3. $K(f(w)) \overset{+}{<} K(w) + K(f)$,
4. $K(w) \overset{+}{<} -\log p(w) + K(p)$ *for a distribution* $\sum_w p(w) \le 1$.

Proof.

1. $K(w|u) \overset{+}{<} K(w)$ because a certain program that computes w given u has form "ignore u and execute the shortest program that computes w".

   $K(w) \overset{+}{<} K(\langle u, w\rangle)$ because a certain program that computes w has form "execute the shortest program that computes $\langle u, w\rangle$ and compute w from $\langle u, w\rangle$".
2. $K(uw) \overset{+}{<} K(\langle u, w\rangle)$ because a certain program that computes uw has form "execute the shortest program that computes $\langle u, w\rangle$ and compute uw from $\langle u, w\rangle$".

   Similarly, $K(\langle u, w\rangle) \overset{+}{<} K(u) + K(w|u)$ because a certain program that computes $\langle u, w\rangle$ has form "execute the shortest program that computes u and the shortest program that computes w given u and from that compute $\langle u, w\rangle$".

   The last inequality follows from $K(w|u) \overset{+}{<} K(w)$.
3. The inequality follows from the fact that a certain program that computes $f(w)$ has form "execute the shortest program that computes w and to the result apply the program that computes $f(w)$ given w".
4. The inequality follows from the fact that a certain program that computes w has form "having a program that computes $w \mapsto p(w)$, take the Shannon-Fano code word for w with respect to p and compute w from it."

Now we can demonstrate the exact link between prefix-free Kolmogorov complexity and entropy. Namely, the expectation of prefix-free complexity is close to entropy for computable probability distributions.

Theorem 14.5. *We have*

$$H(p) \le \sum_w p(w)K(w) \overset{+}{<} H(p) + K(p),$$

where $H(p) = -\sum_w p(w) \log p(w)$ *is the entropy of a distribution* p.

Proof. The left inequality follows because $K(w)$ is the length of a prefix-free code. The right inequality follows by inequality $K(w) \overset{+}{<} -\log p(w) + K(p)$.

The parallels between prefix-free complexity and entropy can be drawn further. The following theorem is an analogue of the chain rule $H(X, Y) = H(X) + H(Y|X)$.

Theorem 14.6. *We have*

$$K(\langle u, w\rangle) \overset{\pm}{=} K(u) + K(w| \langle u, K(u)\rangle).$$

In the proposition above, it is easy to show that the left hand side is smaller than the right hand side. The proof of the converse inequality is harder but it rests on an interesting use of Kraft inequality and algorithmic probability. Thus we will demonstrate the entire reasoning.

Definition 14.8 (recursively enumerable function). *For a subset* $\mathbb{S} \subset \mathbb{A} \times \mathbb{A}$ *let us denote projections* $\mathbb{S}_1^b := \{a : (a, b) \in \mathbb{S}\}$, $\mathbb{S}_1 := \{a : \exists_b (a, b) \in \mathbb{S}\}$, *and* $\mathbb{S}_2 := \{b : \exists_a (a, b) \in \mathbb{S}\}$. *A function* $W(\cdot|\cdot) : \mathbb{S} \to \mathbb{N}$ *will be called* conditionally recursively enumerable, *if there exists a computable function* $f(\cdot|\cdot) : \mathbb{N} \times \mathbb{S}_2 \to \mathbb{S}_1 \cup \{\xi\}$, *where* $\xi \in \mathbb{A} \setminus \mathbb{S}_1$ *and*

$$\operatorname{card}\big\{m \in \mathbb{N} : f(m|b) = a\big\} = W(a|b)$$

for each $b \in \mathbb{S}_2$ *and* $a \in \mathbb{S}_1^b$, *where* $\operatorname{card} A$ *is the cardinality of set* A. *We say respectively that* $f(\cdot|\cdot)$ enumerates $W(\cdot|\cdot)$.

Function $f(\cdot|\cdot)$ may return a value ξ out of set $\mathbb{S}_1$, but because there is only one such value, we may effectively identify that it does not belong to $\mathbb{S}_1$.

Now let us introduce a generalization of Theorem 3.6 for codes that can be effectively *decoded*, but not necessarily effectively encoded.

Theorem 14.7 (effective Kraft inequality). *If a function* $W : \mathbb{S} \to \mathbb{N}$, *where* $\mathbb{S} \subset (\{0,1\}^* \times \mathbb{N}) \times \{0,1\}^*$, *is conditionally recursively enumerable and satisfies*

$$\sum_{(w,n)\in \mathbb{S}_1^q} 2^{-n} W(w, n|q) \le 1 \tag{14.1}$$

then there exists a prefix-free machine $S \in \mathcal{S}$ *such that for all* $q \in \mathbb{S}_2$ *we have*

$$\operatorname{card}\big\{p \in \{0,1\}^* : S(p|q) = w, |p| = n\big\} = W(w, n|q). \tag{14.2}$$

Proof. The proof is analogous to the proof of Theorem 3.6. The subtle difference is that the ordering of code values is now chosen as the order of pairs (w, n) produced by the computable function that enumerates function $W(\cdot|\cdot)$. In this way, we obtain a computable function that enumerates triples (p, w, q) where p are prefix-free code words. Subsequently, the function is used for the construction of machine S. Namely, machine S on input (p', q') enumerates triples (p, w, q) until it finds triple (p', w', q') and then it outputs w'.

The second ingredient for the proof of Theorem 14.6 are algorithmic probabilities.

Definition 14.9. *We define* algorithmic probabilities *as*

$$\Pi_S(w|q) := \sum_{p \in \{0,1\}^*} 2^{-|p|} \mathbf{1}\{S(p|q) = w\}.$$

By Kraft inequality (Theorem 3.3), we have

$$\sum_{w \in \{0,1\}^*} \Pi_S(w|q) = \Omega_S(q) \le 1,$$

where

$$\Omega_S(q) := \sum_{p \in \{0,1\}^*} 2^{-|p|} \mathbf{1}\{S(p|q) \neq \infty\}$$

is called halting probability*.*

By definition, we have $\Pi_S(w|q) \ge 2^{-K_S(w|q)}$, so

$$-\log \Pi_S(w|q) \le K_S(w|q).$$

In contrast, Theorem 14.7 implies the converse inequality for a universal machine.

Theorem 14.8 (coding theorem). *For any prefix-free machine $S' \in \mathcal{S}$,*

$$K(w|q) \overset{+}{<} -\log \Pi_{S'}(w|q).$$

Proof. Define function $W(w, n|q) = \mathbf{1}\big\{\Pi_{S'}(w|q) > 2^{-n+1}\big\}$. Function $W(w, n|q)$ is conditionally recursively enumerable because, by simulating all programs in parallel, we can determine that $\Pi_{S'}(w|q) > 2^{-n+1}$ for any n for which it is true. Moreover, we have

$$\sum_{n \in \mathbb{N}} 2^{-n} W(w, n|q) = 2^{-\lceil -\log \Pi_{S'}(w|q) \rceil} \le \Pi_{S'}(w|q).$$

Thus function $W(w, n|q)$ satisfies (14.1). In consequence, there exists a machine $S \in \mathcal{S}$ such that (14.2) is satisfied. The smallest n for which $\Pi_{S'}(w|q) > 2^{-n+1}$ holds is $\lceil -\log \Pi_{S'}(w|q) \rceil + 1$. Hence we obtain $K_S(w|q) = \lceil -\log \Pi_{S'}(w|q) \rceil + 1$. Therefore the claim follows by inequality $K_S(w|q) \overset{+}{>} K(w|q)$.

Denoting $\Pi(w|q) := \Pi_V(w|q)$ for a universal machine V, we have

$$K(w|q) \overset{+}{=} -\log \Pi(w|q). \tag{14.3}$$

Since we know that Kolmogorov complexity is close to algorithmic probability, let us inspect properties of the latter.

In the following, we will write $p \overset{*}{<} q$ and $p \overset{*}{>} q$ if there is a constant $c > 0$ such that $p \le cq$ and $p \ge cq$ holds respectively for all p and q. We will write $p \overset{*}{=} q$ when we have both $p \overset{*}{<} q$ and $p \overset{*}{>} q$. By Theorem 14.8 we have

$$\Pi(w|q) \overset{*}{>} \Pi_S(w|q),$$

an important fact concerning algorithmic probability. Another important fact is that algorithmic probability behaves almost like a usual probability distribution.

Theorem 14.9. *We have*

$$\Pi(u|q) \overset{*}{=} \sum_{w \in \{0,1\}^*} \Pi(\langle u, w\rangle |q). \tag{14.4}$$

Proof. On the one hand, there exists a machine S such that $S(p|q) = u$ if $V(p|q) = \langle u, w\rangle$. Hence $\Pi(u|q) \overset{*}{>} \Pi_S(u|q) \ge \sum_{w\in\{0,1\}^*} \Pi(\langle u, w\rangle |q)$. On the other hand there exists a machine S such that $S(p|q) = \langle u, u\rangle$ if $V(p|q) = u$. Hence $\sum_{w\in\{0,1\}^*} \Pi(\langle u, w\rangle |q) \ge \Pi(\langle u, u\rangle |q) \overset{*}{>} \Pi_S(\langle u, u\rangle |q) \ge \Pi(u|q)$.

Now we may prove the chain rule for Kolmogorov complexity.

Proof of Theorem 14.6: First we will prove the easier inequality

$$K(\langle u, w\rangle) \overset{+}{<} K(u) + K(w| \langle u, K(u)\rangle).$$

Let p be the shortest program that satisfies $V(p) = u$ and let p' be the shortest program that satisfies $V(p'| \langle u, K(u)\rangle) = w$. Then there exists a prefix-free machine S that satisfies $S(pp') = \langle u, w\rangle$. Hence we obtain the claim.

Next, we will show the harder inequality

$$K(\langle u, w\rangle) - K(u) \overset{+}{>} K(w| \langle u, K(u)\rangle).$$

By (14.3) and (14.4) we know, that there exists a constant c such that for all u and w we have

$$2^{K(u)-c} \sum_{w\in\{0,1\}^*} \Pi(\langle u, w\rangle) \le 1.$$

Define function $W'(p) = \mathbf{1}\{V(p) \neq \infty\}$. It is conditionally recursively enumerable. Let it be enumerated by function $f'(\cdot)$. Subsequently, let us put $q = \langle u, K(u)\rangle$ and define function

$$f(n|q) = \begin{cases} (w, |p| - K(u) + c), & \text{if } f'(n) = p \text{ and } V(p) = \langle u, w\rangle, \\ \xi, & \text{otherwise.} \end{cases}$$

This function is computable and enumerates a certain function $W(\cdot|\cdot)$ which satisfies (14.1). Thus there exists a prefix-free Turing machine S which satisfies (14.2). Hence we obtain $K(w|q) \overset{+}{<} K_S(w|q) \leq K(\langle u, w\rangle) - K(u) + c$. □

The parallel between the chain rule $H(X,Y) = H(X) + H(Y|X)$ and Theorem 14.6 remains incomplete because in the algorithmic version there appears term $K(w|\langle u, K(u)\rangle)$ rather than $K(w|u)$. Although $K(w|\langle u, K(u)\rangle)$ differs from $K(w|u)$, in the next theorem we can see that $K(\langle u, K(u)\rangle)$ and $K(u)$ are approximately equal.

Theorem 14.10. *We have*

$$K(\langle w, K(w)\rangle) \overset{\pm}{=} K(w).$$

Proof. From the shortest program that computes w, we may reconstruct both w and $K(w)$. Hence $K(\langle w, K(w)\rangle) \overset{+}{<} K(w)$. On the hand, we have $K(\langle w, K(w)\rangle) \overset{+}{>} K(w)$ from Theorem 14.4.1.

We may also define an algorithmic analogue of mutual information.

Definition 14.10. *We define* algorithmic information *between strings u and w as*

$$I(u; w) = K(w) - K(w|\langle u, K(u)\rangle).$$

By Theorem 14.6 algorithmic information is symmetric.

Theorem 14.11. *We have*

$$I(u; w) \overset{\pm}{=} I(w; u).$$

Proof. Observe that

$$\begin{aligned} I(u; w) &= K(w) - K(w|\langle u, K(u)\rangle) \\ &\overset{\pm}{=} K(w) + K(u) - K(\langle u, w\rangle) \\ &\overset{\pm}{=} K(u) - K(u|\langle w, K(w)\rangle) = I(w; u). \end{aligned}$$

Exercises

1. Let $n = |w|$. Show that:
   (a) $K(w) \overset{+}{<} n + 2\log n$;
   (b) $K(w|n) \overset{+}{<} n$.
2. Define $K^+(x) = \max\{K(y) : y \leq x\}$. Show that $K^+(x) \overset{\pm}{=} \log x + K(\lceil \log x \rceil)$.
3. Let $n = |w|$ and let w and n be incompressible. Show that $K(w, n) \overset{\pm}{=} K(w) + K(n|w) \overset{\pm}{=} K(n) + K(w|n)$.

4. Having $\langle u, K(u)\rangle$ we may effectively enumerate all programs of the shortest length that compute u. Let u^* be the first program in the enumeration. Prove that $K(w|\langle u, K(u)\rangle) \overset{\pm}{=} K(w|u^*)$ and $K(w^*) \overset{\pm}{=} |w^*| = K(w)$.
5. Show that $K(w|p) \overset{+}{<} -\log p(w)$ and $H(p) \overset{\pm}{=} \sum_w p(w)K(w|p)$ for any probability distribution p.
6. Information distance is defined as

$$\mathrm{ID}(u, w) = K(\langle u, w\rangle) - \min\big\{K(u), K(w)\big\}.$$

Show that this quantity satisfies the axioms of a distance up to a constant.

7. Show that $K(K(K(u))|u, K(u)) \overset{\pm}{=} 0$.
8. Normalized information distance is defined as

$$\mathrm{NID}(u, w) = \frac{K(\langle u, w\rangle) - \min\{K(u), K(w)\}}{\max\{K(u), K(w)\}}.$$

Show that this quantity satisfies the axioms of a normalized distance up to a constant. A distance d is called normalized if $0 \le d(u, w) \le 1$.
Hint: To prove the triangle inequality, consider three cases: (a) $K(z) \le \max\{K(u), K(w)\}$, (b) $K(z) \ge \max\{K(u), K(w)\}$, $K(u|z^*)+K(z|w^*) \le K(u) \ge K(w)$, (c) $K(z) \ge \max\{K(u), K(w)\}$, $K(u|z^*) + K(z|w^*) \ge K(u) \ge K(w)$.

9. We have $\sum_{u\in\{0,1\}^*} 2^{-K(u|w)} \le 1$ by the Kraft inequality. Show that $\sum_{w\in\{0,1\}^*} 2^{-K(u|w)} = \infty$.

15

Random sequences

Barron theorem. Various characterizations of Martin-Löf random sequences. Halting probability. Optimality of Bayesian inference for random parameters.

In this chapter we will consider infinite sequences that are typical outcomes of a probability measure in an intuitive sense. We will show that the set of such sequences, called (Martin-Löf) random sequences, has probability one, whereas random sequences for a uniform measure are incompressible. Subsequently, we will exhibit a few other characterizations of random sequences. Moreover, we will show that Bayesian inference is optimal if and only if the parameter is random with respect to the prior. The last result proves that the concept of Kolmogorov complexity is highly relevant for statistics.

To begin, the sequences will be written down in boldface as $\boldsymbol{x} = (x_1, x_2, x_3, ...)$, where $x_i \in \{0,1\}$. The boldface symbol $\boldsymbol{P}$ will denote a probability measure on infinite sequences. We will write $\boldsymbol{P}(w) = \boldsymbol{P}\big(\{\boldsymbol{x} : x_1^{|w|} = w\}\big)$ and λ will denote the empty string as previously. Thus $\boldsymbol{P}$ satisfies

$$\begin{aligned}\boldsymbol{P}(\lambda) &= 1,\\ \boldsymbol{P}(w) &\geq 0,\\ \boldsymbol{P}(w) &= \sum_{a\in\{0,1\}} \boldsymbol{P}(wa)\end{aligned}$$

Moreover, by a computable measure we will understand a measure such that function $\boldsymbol{P} : \{0,1\}^* \ni w \mapsto \boldsymbol{P}(w) \in \mathbb{R}$ is computable.

Now we want to discuss typical outcomes of a computable probability measure. Using the Borel-Cantelli lemma, we can show that the length of a prefix-free code asymptotically always exceeds pointwise entropy. This can be interpreted as a strengthening of the source coding inequality—Theorem 3.4.

Theorem 15.1 (Barron theorem). *Let $B : \{0,1\}^* \to \{0,1\}^*$ be a prefix-free code. Then for any probability measure $\boldsymbol{P}$ we have*

$$\boldsymbol{P}\Big(\Big\{\boldsymbol{x} : \lim_{m\to\infty} [|B(x_1^m)| + \log \boldsymbol{P}(x_1^m)] = \infty\Big\}\Big) = 1. \tag{15.1}$$

Proof. Let us write

$$W(x_1^m) = \frac{2^{-|B(x_1^m)|}}{\boldsymbol{P}(x_1^m)2^{-n}}.$$

The Markov inequality (Theorem 1.5) asserts that

$$\boldsymbol{P}\big(\{\boldsymbol{x} : W(x_1^m) \geq 1\}\big) \leq \sum_{x_1^m} \boldsymbol{P}(x_1^m)W(x_1^m).$$

Thus by Kraft inequality (Theorem 3.3) we obtain

$$\begin{aligned}
&\sum_{m=1}^{\infty} \boldsymbol{P}\big(\{\boldsymbol{x} : |B(x_1^m)| + \log \boldsymbol{P}(x_1^m) \leq n\}\big) \\
&\qquad = \sum_{m=1}^{\infty} \boldsymbol{P}\big(\{\boldsymbol{x} : W(x_1^m) \geq 1\}\big) \\
&\qquad \leq \sum_{m=1}^{\infty} \sum_{x_1^m} \boldsymbol{P}(x_1^m)W(x_1^m) \\
&\qquad \leq \sum_{m=1}^{\infty} \sum_{x_1^m} 2^{-|B(x_1^m)|+n} \leq 2^n < \infty.
\end{aligned}$$

Hence from the Borel-Cantelli lemma (Theorem 1.4) we obtain

$$\boldsymbol{P}\big(\{\boldsymbol{x} : |B(x_1^m)| + \log \boldsymbol{P}(x_1^m) \leq n \text{ for infinitely many } m\}\big) = 0.$$

The n in this statement is arbitrary so we get

$$\boldsymbol{P}\big(\{\boldsymbol{x} : \liminf_{m\to\infty} [|B(x_1^m)| + \log \boldsymbol{P}(x_1^m)] < \infty\}\big) = 0.$$

In consequence, the claim (15.1) follows.

Considering the concept of a random sequence, it is intuitive to require that the set of random sequences has probability one. Guided by Barron's theorem, let us adopt the following definition, in which $K(w)$ is the prefix-free Kolmogorov complexity of string w.

Definition 15.1 (random sequence). *We say that a sequence $\boldsymbol{x}$ is* (Martin-Löf) random *for a computable probability measure $\boldsymbol{P}$ when*

$$\lim_{m\to\infty} [K(x_1^m) + \log \boldsymbol{P}(x_1^m)] = \infty. \tag{15.2}$$

The set of Martin-Löf random sequences is denoted as

$$\mathcal{R}_{\boldsymbol{P}} := \big\{\boldsymbol{x} : \lim_{m\to\infty} [K(x_1^m) + \log \boldsymbol{P}(x_1^m)] = \infty\big\}.$$

Since the prefix-free Kolmogorov complexity is a length of a prefix-free code, the set of Martin-Löf random sequences has measure one, i.e., $\boldsymbol{P}(\mathcal{R}_{\boldsymbol{P}}) = 1$, by Theorem 15.1. In the definition of random sequences we have restricted to computable measures because expression $K(x_1^m) + \log \boldsymbol{P}(x_1^m)$ grows too fast for noncomputable measures and a different definition of a random sequence is more appropriate then.

Example 15.1 (uniform measure). Let $\boldsymbol{P}(x_1^m) = 2^{-m}$ be the uniform measure. Then the respective Martin-Löf random sequences satisfy

$$\lim_{m\to\infty} \left[K(x_1^m) - m\right] = \infty.$$

This means that random sequences for a uniform measure are exactly those that are incompressible.

Example 15.2 (point measure). Let probability measure $\boldsymbol{P}$ be concentrated on a sequence $\boldsymbol{y}$, i.e., $\boldsymbol{P}(y_1^m) = 1$. Then the set of Martin-Löf random sequences is the singleton $\{\boldsymbol{y}\}$.

Historically, the concept of a random sequence was developed by Per Martin-Löf (1942–) by means of algorithmic tests. It was proved later by Claus-Peter Schnorr that this approach is equivalent to the definition motivated by Barron's theorem. Let us inspect this piece of theory.

Definition 15.2 (recursively enumerable set). *Set $S \subset \{0,1\}^*$ is called* recursively enumerable *if there exists a computable function $f : \mathbb{N} \to \{0,1\}^*$ such that $f(\mathbb{N}) = S$.*

Definition 15.3 (Martin-Löf test). *Let $U \subset \mathbb{N} \times \{0,1\}^*$ and $U_n := \{w : (n,w) \in U\}$. Set U is called a* Martin-Löf test *for a measure $\boldsymbol{P}$ if*

1. *U is recursively enumerable,*
2. *$\boldsymbol{P}(\tilde{U}_n) \le 2^{-n}$, where*

$$\tilde{U}_n := \left\{\boldsymbol{x} : w \text{ is a prefix of } \boldsymbol{x} \text{ and } w \in U_n\right\}.$$

In the informal motivation, sets $\tilde{U}_n$ are sets of sequences that approximate non-random sequences. As we will see, a sequence is random if it is not contained in the intersection of these sets.

We will also need another concept, Solovay's tests, which are a slight relaxation of Martin-Löf tests.

Definition 15.4 (Solovay test). *Let $V \subset \mathbb{N}\times\{0,1\}^*$ and $V_n := \{w : (n,w) \in V\}$. Set V is called a* Solovay test *for a measure $\boldsymbol{P}$ if*

1. *V is recursively enumerable,*
2. *$\sum_{n=1}^{\infty} \boldsymbol{P}(\tilde{V}_n) < \infty$, where*

$$\tilde{V}_n := \left\{\boldsymbol{x} : w \text{ is a prefix of } \boldsymbol{x} \text{ and } w \in V_n\right\}.$$

As the next theorem states, a sequence is random if it does not pass any Martin-Löf or Solovay test. The concept of passing a test is made precise in the proposition.

Theorem 15.2 (Schnorr theorem). *If measure $\boldsymbol{P}$ is computable then the following conditions are equivalent:*

1. *Sequence* $\boldsymbol{x}$ *is Martin-Löf random.*
2. *For any Solovay test* V*,* $\boldsymbol{x}$ *is contained in finitely many* $\tilde{V}_n$*.*
3. *For any Martin-Löf test* U*, we have* $\boldsymbol{x} \notin \bigcap_{n=1}^{\infty} \tilde{U}_n$*.*
4. *Sequence* $\boldsymbol{x}$ *is* weakly Martin-Löf random, *i.e., it satisfies*

$$\inf_{m\in\mathbb{N}} \left[K(x_1^m) + \log \boldsymbol{P}(x_1^m)\right] > -\infty. \tag{15.3}$$

Remark: Equivalence of (15.2) and (15.3) reveals a gap in the "randomness deficiency". Namely, expression $K(x_1^m) + \log \boldsymbol{P}(x_1^m)$ either tends to infinity or goes arbitrarily negative.

Proof. We will demonstrate that 1. $\implies$ 4. $\implies$ 3. $\implies$ 2. $\implies$ 1.

1. $\implies$ 4.: Obviously, $K(x_1^m) + \log \boldsymbol{P}(x_1^m)$ is bounded below if it tends to infinity.

4. $\implies$ 3.: Suppose that $\boldsymbol{x} \in \bigcap_{n=1}^{\infty} \tilde{U}_n$ for a certain Martin-Löf test U. We will show that $\boldsymbol{x}$ is not weakly Martin-Löf random. Without loss of generality, assume that sets U_n are prefix-free. Then we have

$$\begin{aligned}\sum_{n=2}^{\infty}\sum_{w\in U_{n^2}} 2^{n+\lfloor \log \boldsymbol{P}(w)\rfloor} &\le \sum_{n=2}^{\infty}\sum_{w\in U_{n^2}} 2^{n+\log \boldsymbol{P}(w)} \\ &= \sum_{n=2}^{\infty} 2^n \boldsymbol{P}(\tilde{U}_{n^2}) \le \sum_{n=2}^{\infty} 2^{n-n^2} \le 1.\end{aligned}$$

Hence in view of Theorem 14.7 there exists a prefix-free Turing machine S such that

$$\begin{aligned}&\operatorname{card}\left\{p \in \{0,1\}^* : S(p) = w, |p| = m\right\} \\ &\qquad = \operatorname{card}\left\{n \ge 2 : m = -\lfloor \log \boldsymbol{P}(w)\rfloor - n, w \in U_{n^2}\right\}.\end{aligned}$$

(The function on the right hand side is recursively enumerable.) In consequence,

$$K_S(w) \le -\lfloor \log \boldsymbol{P}(w)\rfloor - n$$

for any $w \in U_{n^2}$ and $n \ge 2$. Since $\boldsymbol{x} \in \bigcap_{n=1}^{\infty} \tilde{U}_n$, there are infinitely many prefixes $x_1^{m(n)}$ such that $x_1^{m(n)} \in U_{n^2}$. This yields

$$K\big(x_1^{m(n)}\big) + \log \boldsymbol{P}\big(x_1^{m(n)}\big) \overset{+}{<} K\big(x_1^{m(n)}\big) + \left\lfloor \log \boldsymbol{P}\big(x_1^{m(n)}\big)\right\rfloor < -n$$

for an arbitrary n. Thus $\boldsymbol{x}$ is not weakly Martin-Löf random.

3. $\implies$ 2.: Suppose that $\boldsymbol{x}$ is contained in infinitely many $\tilde{V}_n$ for a certain Solovay test V. We will show that $\boldsymbol{x} \in \bigcap_{n=1}^{\infty} \tilde{U}_n$ for a certain Martin-Löf test U. For $\sum_{n=1}^{\infty} \boldsymbol{P}(\tilde{V}_n) \le C$, the test U is constructed as

$$\tilde{U}_n = \left\{\boldsymbol{y} : \boldsymbol{y} \text{ is at least in } 2^n C \text{ of } \tilde{V}_i\right\}.$$

Then $\boldsymbol{x}$ belongs to all $\tilde{U}_n$ and $\boldsymbol{P}(\tilde{U}_n) \le 2^{-n}$ so U is a Martin-Löf test.

2. $\implies$ 1.: Suppose that $\boldsymbol{x}$ is not Martin-Löf random. We will show that $\boldsymbol{x}$ belongs to infinitely many $\tilde{V}_m$ for a certain Solovay test V. We know that $K(x_1^m) + \log \boldsymbol{P}(x_1^m) \le n$ for a certain n and infinitely many m. Thus it belongs to infinitely many $\tilde{V}_m$ if we set

$$V_m = \left\{ w \in \{0,1\}^m : K(w) + \log \boldsymbol{P}(w) \le n \right\}.$$

For this choice of V_m, set V is recursively enumerable. It suffices to prove the second condition for the Solovay test. Denote

$$W(w) = \frac{2^{-K(w)}}{\boldsymbol{P}(w) 2^{-n}}.$$

Then by the Markov and Kraft inequality, we obtain

$$\begin{aligned} \sum_{m=1}^{\infty} \boldsymbol{P}(\tilde{V}_m) &= \sum_{m=1}^{\infty} \boldsymbol{P}\left(\left\{ \boldsymbol{y} : W(y_1^m) \ge 1 \right\}\right) \\ &\le \sum_{m=1}^{\infty} \sum_{y_1^m} \boldsymbol{P}(y_1^m) W(y_1^m) \\ &\le \sum_{m=1}^{\infty} \sum_{y_1^m} 2^{-K(y_1^m)+n} \le 2^n < \infty, \end{aligned}$$

as required.

Having Schnorr's theorem, we can show that the binary expansion of the halting probability is Martin-Löf random with respect to the uniform measure. First, let us formulate a useful lemma.

Lemma 15.1. *Consider halting probability*

$$\Omega = \sum_{p \colon V(p) \ne \infty} 2^{-|p|}.$$

Let Ω_1^n be the first n digits of Ω and let p be a string of a length smaller than n. Given Ω_1^n we may decide whether machine V stops on input p.

Proof. We have $0.\Omega_1^n \le \Omega < 0.\Omega_1^n + 2^{-n}$. Let us simulate the computation of machine V on all inputs shorter than n. Namely, in the i-th step we execute the j-th step of computations for all k-th inputs which satisfy $j + k = i$. In the beginning of the simulations, we set the approximation of Ω as $\Omega' := 0$. When V halts for a certain input p, we improve the approximation by setting $\Omega' := \Omega' + 2^{-|p|}$. At a certain instant, Ω' becomes equal or greater than $0.\Omega_1^n$. Then it becomes clear that machine V will not halt on any other input shorter than n and we may decide on the halting problem.

In view of Lemma 15.1, the number Ω encodes the solution of the halting problem in the most dense way. Now we will show that this implies that the binary expansion of Ω is random.

Theorem 15.3. *We have* $K(\Omega_1^n) \overset{+}{>} n$.

Proof. By Lemma 15.1, we can enumerate, given Ω_1^n, all programs shorter than n for which machine V halts. For any w which is not computed by these programs we have $K(w) > n$. We can construct a computable function ϕ which computes one of these w given Ω_1^n. Hence $K(\phi(\Omega_1^n)) \geq n$, which implies $K(\Omega_1^n) \geq n - c$ for a certain c.

Another important way of defining Martin-Löf random sequences leads through so called impossibility levels. This characterization is useful in stating some further properties of random sequences.

Definition 15.5 (impossibility level). *We define the impossibility level of a sequence* $\boldsymbol{x}$ *with respect to a computable measure* $\boldsymbol{P}$ *as*

$$\mathcal{I}(\boldsymbol{x}; \boldsymbol{P}) := \sup_{n \in \mathbb{N}} \frac{2^{-K(x^n)}}{\boldsymbol{P}(x^n)}.$$

Theorem 15.4. *We have* $\mathcal{I}(\boldsymbol{x}; \boldsymbol{P}) < \infty$ *if and only if* $\boldsymbol{x}$ *is Martin-Löf random.*

Proof. The claim follows by Schnorr's theorem (Theorem 15.2).

As we will see, the impossibility level is an example of a unit integrable function.

Definition 15.6 (unit integrable function). *We say that function* $f : \{0,1\}^{\mathbb{N}} \to \mathbb{R}$ *is a unit integrable function for a computable measure* $\boldsymbol{P}$ *if* f *is lower semicomputable, nonnegative, and*

$$\int f \, \mathrm{d}\boldsymbol{P} \leq 1.$$

Theorem 15.5. *Impossibility level* $\mathcal{I}(\boldsymbol{x}; \boldsymbol{P})$ *is a unit integrable function.*

Proof. It is easy to see that $\mathcal{I}(\boldsymbol{x}; \boldsymbol{P})$ is nonnegative and lower semicomputable. As for the third property, we obtain

$$\begin{aligned}
\int \mathcal{I}(\boldsymbol{x}; \boldsymbol{P}) \, \mathrm{d}\boldsymbol{P}(\boldsymbol{x}) &= \int \left[\sup_{n \in \mathbb{N}} \frac{2^{-K(x^n)}}{\boldsymbol{P}(x^n)} \right] \mathrm{d}\boldsymbol{P}(\boldsymbol{x}) \\
&\leq \int \left[\sum_{n=1}^{\infty} \frac{2^{-K(x^n)}}{\boldsymbol{P}(x^n)} \right] \mathrm{d}\boldsymbol{P}(\boldsymbol{x}) \\
&= \sum_{n=1}^{\infty} \frac{2^{-K(x^n)}}{\boldsymbol{P}(x^n)} \boldsymbol{P}(x^n) \\
&= \sum_{n=1}^{\infty} 2^{-K(x^n)} \leq 1.
\end{aligned}$$

Moreover, the impossibility level dominates all other unit integrable functions.

Theorem 15.6. *We have $\mathcal{I}(\boldsymbol{x};\boldsymbol{P}) \overset{*}{>} f(\boldsymbol{x})$ for any unit integrable function f.*

The proof can be found in Li and Vitányi (2008, Theorem 4.5.5).

In view of Theorem 15.6 we may define Martin-Löf sequences as such that $f(\boldsymbol{x}) < \infty$ for any unit integrable function f. This characterization plays an important role in the derivation of the following remarkable property. Namely, Martin-Löf sequences are exactly those sequences for which measure $\boldsymbol{P}$ is the optimal compressor in the pool of all semicomputable semimeasures.

Definition 15.7 (semimeasure). *A* semimeasure $\boldsymbol{M}$ *is a function that satisfies*

$$\begin{aligned} \boldsymbol{M}(\lambda) &\leq 1, \\ \boldsymbol{M}(w) &\geq 0, \\ \boldsymbol{M}(w) &\geq \sum_{a\in\{0,1\}} \boldsymbol{M}(wa). \end{aligned}$$

We have two important theorems.

Theorem 15.7. *There exists a lower semicomputable semimeasure $\boldsymbol{M}$, such that $\boldsymbol{U}(x_1^n) \overset{*}{<} \boldsymbol{M}(x_1^n)$ for any lower semicomputable semimeasure $\boldsymbol{U}$.*

The proof can be found in Li and Vitányi (2008, Theorem 4.5.1)

Theorem 15.8. *Let $\boldsymbol{M}$ be the lower semicomputable semimeasure as in the previous theorem. For an $\epsilon > 0$ and a computable measure $\boldsymbol{P}$, we have*

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{P}) \overset{*}{<} \liminf_{n\to\infty} \frac{\boldsymbol{M}(x_1^n)}{\boldsymbol{P}(x_1^n)} \overset{*}{<} \sup_{n\in\mathbb{N}} \frac{\boldsymbol{M}(x_1^n)}{\boldsymbol{P}(x_1^n)} \overset{*}{<} \left[\mathcal{I}(\boldsymbol{x};\boldsymbol{P})\right]^{1+\epsilon}.$$

The proof can be found in Vovk and V'yugin (1994, Theorem 1 and Lemma 3).

By Theorem 15.8, we have

$$\sup_{n\in\mathbb{N}} \frac{\boldsymbol{M}(x_1^n)}{\boldsymbol{P}(x_1^n)} < \infty$$

if and only if the sequence $\boldsymbol{x}$ is random with respect to the measure $\boldsymbol{P}$. Since $\boldsymbol{U}(x^n) \overset{*}{<} \boldsymbol{M}(x^n)$ for any semicomputable semimeasure $\boldsymbol{U}$, this result states that Martin-Löf random sequences are exactly those sequences that are optimally compressed by a computable measure $\boldsymbol{P}$ in the pool of all semicomputable semimeasures (up to a multiplicative constant).

Finally, we will show how the concept of Martin-Löf randomness sheds light on the power and perils of effective Bayesian inference. Suppressing a few technically difficult proofs to references, let us report the general ideas since they are highly interesting. The framework is as follows. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, ...)$, where $\theta_i \in \{0, 1\}$, denote the parameter, which is equivalently interpreted as a binary expansion of a real number, also denoted as $\boldsymbol{\theta}$. The notion of computable probability measures can be generalized to families of probability measures via probability kernels.

Definition 15.8 (probability kernel). *Function* $\boldsymbol{P}(\cdot|\cdot) : \{0,1\}^* \times \{0,1\}^{\mathbb{N}} \to \mathbb{R}$ *is called a* probability kernel *if* $\boldsymbol{P}(\cdot|\boldsymbol{\theta})$ *is a probability measure for each* $\boldsymbol{\theta}$*.*

A probability kernel $\boldsymbol{P}$ will be called computable if function $\boldsymbol{P}(\cdot|\cdot) : \{0,1\}^* \times \{0,1\}^{\mathbb{N}} \ni (w,\boldsymbol{\theta}) \mapsto \boldsymbol{P}(w|\boldsymbol{\theta}) \in \mathbb{R}$ is computable. Let us observe that, even for a computable kernel, the conditional measure $\boldsymbol{P}(\cdot|\boldsymbol{\theta})$ need not be computable if we fix a particular value of the parameter $\boldsymbol{\theta}$. Typically this measure is not computable if the parameter is Martin-Löf random. Hence the optimal computable compressor of data that are typical of a conditional measure $\boldsymbol{P}(\cdot|\boldsymbol{\theta})$ is different to this conditional measure.

To find the optimal compressor of data typical of a noncomputable measure, let us introduce the setting of Bayesian inference. Let $\boldsymbol{Q}$ be a computable measure called the prior and let $\boldsymbol{P}$ be a computable probability kernel. Measure

$$\boldsymbol{Y}(w) = \int \boldsymbol{P}(w|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{Q}(\boldsymbol{\theta}) \tag{15.4}$$

is computable and will be called the Bayesian measure. Subsequently, we will show that, whenever the parameter can be effectively estimated, Bayesian measure $\boldsymbol{Y}$ is the optimal computable compressor of data typical of conditional measure $\boldsymbol{P}(\cdot|\boldsymbol{\theta})$ if and only if the parameter $\boldsymbol{\theta}$ is typical of the prior $\boldsymbol{Q}$.

First, let us recall that the set of Martin-Löf random sequences $\mathcal{R}_{\boldsymbol{Y}}$ is the maximal set of sequences that are optimally compressed by measure $\boldsymbol{Y}$ in the pool of all semicomputable semimeasures. The following theorem states that for almost all parameters $\boldsymbol{\theta}$, typical outcomes of conditional measures $\boldsymbol{P}(\cdot|\boldsymbol{\theta})$ are also optimally compressed by measure $\boldsymbol{Y}$.

Theorem 15.9. *For (15.4), we have*

$$\boldsymbol{Q}\big(\{\boldsymbol{\theta} : \boldsymbol{P}(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}) = 1\}\big) = 1.$$

Proof. Let $\mathcal{G}_n = \{\boldsymbol{\theta} : \boldsymbol{P}(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}) \ge 1 - 1/n\}$. We have

$$\begin{aligned} 1 = \boldsymbol{Y}\big(\mathcal{R}_{\boldsymbol{Y}}\big) &= \int_{\mathcal{G}_n} \boldsymbol{P}\big(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}\big)\, \mathrm{d}\boldsymbol{Q}(\boldsymbol{\theta}) + \int_{\{0,1\}^{\mathbb{N}} \setminus \mathcal{G}_n} \boldsymbol{P}\big(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}\big)\, \mathrm{d}\boldsymbol{Q}(\boldsymbol{\theta}) \\ &\le \boldsymbol{Q}\big(\mathcal{G}_n\big) + \boldsymbol{Q}\big(\{0,1\}^{\mathbb{N}} \setminus \mathcal{G}_n\big)(1 - 1/n) \\ &= 1 - n^{-1}\boldsymbol{Q}\big(\{0,1\}^{\mathbb{N}} \setminus \mathcal{G}_n\big). \end{aligned}$$

Hence $\boldsymbol{Q}(\{0,1\}^{\mathbb{N}} \setminus \mathcal{G}_n) \le 0$ so $\boldsymbol{Q}(\mathcal{G}_n) = 1$. Denote

$$\mathcal{G} = \big\{\boldsymbol{\theta} : \boldsymbol{P}(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}) = 1\big\} = \bigcap_{n \in \mathbb{N}} \mathcal{G}_n.$$

By continuity, we obtain $\boldsymbol{Q}(\mathcal{G}) = \inf_{n \in \mathbb{N}} \boldsymbol{Q}(\mathcal{G}_n) = 1$.

In the following, we want to show that for certain probability kernels Theorem 15.9 can be strengthened as

$$\boldsymbol{P}(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta} \in \mathcal{R}_{\boldsymbol{Q}}, \\ 0 & \text{if } \boldsymbol{\theta} \not\in \mathcal{R}_{\boldsymbol{Q}}. \end{cases} \tag{15.5}$$

Let us observe that (15.5) need not hold for an arbitrary probability kernel. For instance if $\boldsymbol{P}(\cdot|\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, we obtain $\boldsymbol{P}(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}) = 1$ for all $\boldsymbol{\theta}$ regardless of the prior. We will show that a sufficient condition to obtain (15.5) is that the probability kernel admits an effective estimator.

Definition 15.9 (effectively strictly consistent estimator). *An estimator $T(x_1^n)$ is called* effectively strictly consistent *if there exists a computable function $N(\epsilon, \delta)$ such that for each $\boldsymbol{\theta}$ and for all ϵ and δ we have*

$$\boldsymbol{P}\big(\big\{\boldsymbol{x} : \sup_{n \geq N(\epsilon,\delta)} |T(x_1^n) - \boldsymbol{\theta}| > \epsilon\big\} \,\big|\, \boldsymbol{\theta}\big) \leq \delta.$$

Now we have to introduce Martin-Löf random sequences with respect to a computable probability kernel. In the respective definition, $K(w|\boldsymbol{\theta})$ stands for the prefix-free Kolmogorov complexity of string w given the infinite sequence $\boldsymbol{\theta}$.

Definition 15.10 (conditionally random sequence). *We say that a sequence $\boldsymbol{x}$ is* conditionally (Martin-Löf) random *for a computable probability kernel $\boldsymbol{P}$ with a parameter $\boldsymbol{\theta}$ when*

$$\lim_{m \to \infty} \big[K(x_1^m|\boldsymbol{\theta}) + \log \boldsymbol{P}(x_1^m|\boldsymbol{\theta})\big] = \infty.$$

The set of Martin-Löf random sequences for a given parameter $\boldsymbol{\theta}$ is denoted as

$$\mathcal{R}_{\boldsymbol{P}|\boldsymbol{\theta}} := \big\{\boldsymbol{x} : \lim_{m \to \infty} \big[K(x_1^m|\boldsymbol{\theta}) + \log \boldsymbol{P}(x_1^m|\boldsymbol{\theta})\big] = \infty\big\}.$$

Since the prefix-free Kolmogorov complexity is the length of a prefix-free code, the respective sets of conditionally Martin-Löf random sequences have measure one, i.e., $\boldsymbol{P}(\mathcal{R}_{\boldsymbol{P}|\boldsymbol{\theta}}|\boldsymbol{\theta}) = 1$, by Theorem 15.1. Moreover, analogues of Theorems 15.2, 15.6, and 15.8 can be also established for conditionally random sequences.

Some useful property of a conditionally random sequence is that an effectively consistent estimator converges to the right value of the parameter.

Theorem 15.10. *If a computable probability kernel $\boldsymbol{P}$ admits an effectively strictly consistent estimator $T(x_1^n)$ then*

$$\lim_{n \to \infty} T(x_1^n) = \boldsymbol{\theta}$$

for each $\boldsymbol{x} \in \mathcal{R}_{\boldsymbol{P}|\boldsymbol{\theta}}$.

The proof can be found in V'yugin (2007, Proposition 1).

Another fact that we shall use is a decomposition of the set of sequences which are random for the Bayesian measure. This set decomposes into sets of sequences which are random for the conditional measures. It is remarkable that the decomposition ranges only over parameters which are random with respect to the prior.

Theorem 15.11. *For a computable probability kernel $\boldsymbol{P}$ and a computable prior $\boldsymbol{Q}$, define the Bayesian measure (15.4). We have*

$$\mathcal{R}_{\boldsymbol{Y}} = \bigcup_{\boldsymbol{\theta}\in\mathcal{R}_{\boldsymbol{Q}}} \mathcal{R}_{\boldsymbol{P}|\boldsymbol{\theta}}. \tag{15.6}$$

Proof. Define the conditional impossibility level

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{P}|\boldsymbol{\theta}) := \sup_{n\in\mathbb{N}} \frac{2^{-K(x^n|\boldsymbol{\theta})}}{\boldsymbol{P}(x^n|\boldsymbol{\theta})}.$$

By the analogue of Theorem 15.2 for conditionally random sequences, impossibility level $\mathcal{I}(\boldsymbol{x};\boldsymbol{P}|\boldsymbol{\theta})$ is finite if and only if sequence $\boldsymbol{x}$ is conditionally random for parameter $\boldsymbol{\theta}$. This should be combined with the following fact. By Vovk and V'yugin (1993, Corollary 4), impossibility levels satisfy an analogue of the chain rule for Kolmogorov complexity (Theorem 14.6). Namely, for an $\epsilon > 0$, we have

$$\inf_{\boldsymbol{\theta}} \left[\mathcal{I}\big(\boldsymbol{x};\boldsymbol{P}|\boldsymbol{\theta}\big)\,\mathcal{I}(\boldsymbol{\theta};\boldsymbol{Q})\right] \overset{*}{<} \mathcal{I}\big(\boldsymbol{x};\boldsymbol{Y}\big) \overset{*}{<} \inf_{\boldsymbol{\theta}} \left[\mathcal{I}\big(\boldsymbol{x};\boldsymbol{P}|\boldsymbol{\theta}\big)\,\left[\mathcal{I}(\boldsymbol{\theta};\boldsymbol{Q})\right]^{1+\epsilon}\right]. \tag{15.7}$$

Hence $\mathcal{I}(\boldsymbol{x};\boldsymbol{Y})$ is finite if and only if $\mathcal{I}(\boldsymbol{x};\boldsymbol{P}|\boldsymbol{\theta})$ and $\mathcal{I}(\boldsymbol{\theta};\boldsymbol{Q})$ are finite. This implies the claim.

Remark: Vovk and V'yugin proved (15.7) using the analogue of Theorem 15.6 for conditionally unit integrable functions. Independently, decomposition (15.6) has also been proved by Takahashi (2008, Theorem 4.2 and 5.3) using a different method.

Now we may state the optimality result.

Theorem 15.12. *For a computable probability kernel $\boldsymbol{P}$ and a computable prior $\boldsymbol{Q}$, define the Bayesian measure (15.4). If the probability kernel admits an effectively strictly consistent estimator, then we have dichotomy*

$$\boldsymbol{P}(\mathcal{R}_{\boldsymbol{Y}}|\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \boldsymbol{\theta}\in\mathcal{R}_{\boldsymbol{Q}}, \\ 0 & \text{if } \boldsymbol{\theta}\not\in\mathcal{R}_{\boldsymbol{Q}}. \end{cases} \tag{15.8}$$

Proof. In view of Theorem 15.10, sets $\mathcal{R}_{\boldsymbol{P}|\boldsymbol{\theta}}$ are disjoint. We also have $\boldsymbol{P}(\mathcal{R}_{\boldsymbol{P}|\boldsymbol{\theta}}|\boldsymbol{\theta}) = 1$ for each $\boldsymbol{\theta}$. Hence (15.8) follows from (15.6).

Thus we have demonstrated that whenever the parameter can be effectively estimated then the Bayesian measure gives the optimal compression of data that are random with respect to a conditional measure if and only if the parameter is random with respect to the prior. This statement is useful when the conditional measure is not computable for a fixed parameter. Moreover, once we know where Bayesian compression fails, we should systematically adjust the prior to our hypotheses about the algorithmic complexity of a parameter in an application.

Exercises

1. Let sequence $\boldsymbol{x} = x_1x_2x_3...$ be Martin-Löf random with respect to the uniform distribution $\boldsymbol{P}(x_1^n) = 2^{-n}$. Show that for any n, sequences $\boldsymbol{y}$, where $y_i = x_{i+n}$ for $i \geq 1$, and $\boldsymbol{z}$, where $z_{i+n} = x_i$ for $i \geq 1$, are also random.
2. A sequence $\boldsymbol{x}$ is called computable if there exists a computable function ϕ such that $\phi(n) = x_n$ for each n. Show that for a computable sequence $x^{\mathbb{N}}$ we have $K(x_1^n) \leq K(n) + c$.
3. Using Kraft inequality, show that $K(n) \geq \log n + \log\log n$ for infinitely many n.
4. Show that a real function f is computable if it is both lower and upper semicomputable.

Solutions of selected exercises

Chapter 1

1. The probability that the car remains behind a randomly selected door out of the set of three doors equals 1/3. In particular, we can assume that the initially chosen door, say door A, was selected at random. Thus the probability that the car is behind the door A equals 1/3. Now let us compute the probability p that the car remains behind the other door that was not opened by Monty Hall, say door B. Door B is not selected at random so we cannot assume that $p = 1/3$. In fact, the car is either behind the door A or door B so $1/3 + p = 1$. Hence $p = 2/3$. Thus, in the second phase of the quiz, it is advisable to choose door B rather than door A.

7. Let $A^c = \Omega \setminus A$ denote the complement of set A. We have

$$P\big(\limsup_{n\to\infty} A_n\big) = \lim_{n\to\infty} P\Big(\bigcup_{k=n}^{\infty} A_k\Big) = 1 - \lim_{n\to\infty} P\Big(\bigcap_{k=n}^{\infty} A_k^c\Big).$$

In the following, we obtain

$$\begin{aligned} P\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) &= \prod_{k=n}^{\infty} P(A_k^c) = \prod_{k=n}^{\infty} \big(1 - P(A_k)\big) \\ &\le \prod_{k=n}^{\infty} \exp\left(-P(A_k)\right) = \exp\Big(-\sum_{k=n}^{\infty} P(A_k)\Big) = 0. \end{aligned}$$

9. By Markov inequality, we have

$$P\bigg(\bigg|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\bigg| > \epsilon\bigg) \le \frac{\mathbf{E}\,\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2}{\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2},$$

which tends to 0 as $n \to \infty$.

10. Let $\epsilon > 0$. By Markov inequality, we obtain

$$P\bigg(\bigg|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\bigg| > \epsilon\bigg) \le \frac{\mathbf{E}\,\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^4}{\epsilon^4} = \frac{n\mu_4}{n^4\epsilon^4} + \binom{4}{2}\frac{n(n-1)\sigma^4}{n^4\epsilon^4}.$$

Hence

$$\sum_{n=1}^{\infty} P\bigg(\bigg|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\bigg| > \epsilon\bigg) < \infty.$$

Thus Borel-Cantelli lemma yields

$$P\bigg(\limsup_{n\to\infty}\bigg|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\bigg| > \epsilon\bigg) = 0.$$

In consequence, (1.4) follows since ϵ was chosen arbitrarily.

Chapter 2

8. The mutual information between X and (Y, Z) can be written in two ways

$$\begin{aligned} I\big(X;Y,Z\big) &= I\big(X;Z\big) + I\big(X;Y|Z\big) \\ &= I\big(X;Y\big) + I\big(X;Z|Y\big). \end{aligned}$$

Because X and Z are conditionally independent given Y, we have $I(X;Z|Y) = 0$. On the other hand, $I(X;Y|Z) \geq 0$. Hence $I(X;Y) \geq I(X;Z)$.

16. Let $p(i) = P(X = i)$. Because $-\log p$ is decreasing, whereas $-p\log p$ is increasing for $p \leq 1/4$, we have

$$\begin{aligned} H\big(X\big) &= -\sum_{i:p(i)>2^{-i}} p(i)\log p(i) - \sum_{i:p(i)\leq 2^{-i}} p(i)\log p(i) \\ &\leq \sum_i ip(i) + \sum_i i2^{-i} < \infty. \end{aligned}$$

Chapter 3

2. For a finite $\mathbb{X}$, let l be the maximal length of $B(x)$. If the Kraft inequality is strict then some string of length l does not contain any prefix in set $\{B(x) : x \in \mathbb{X}\}$. Hence we may enlarge $\{B(x) : x \in \mathbb{X}\}$ with that string. For an infinite code the situation is different. Consider for instance set $W = \left\{www : w \in \{0,1\}^+\right\}$ and let us delete from W all strings whose proper prefixes belong to W. The so obtained set is maximal prefix-free and its Kraft sum is less than $\sum_{w\in W} 2^{-|w|} = \sum_{i=1} 2^i 2^{-3i} = \sum_{i=1} 4^{-i} = 1/3$. Hence the set is not complete.

9. The decoding algorithm is as follows:
   (a) Begin with $N = 1$.
   (b) Read the next block of digits, which can be 0, or 1 followed by N digits. If the read block is 0 then return N, being the encoded number, and stop. If the read block is 1 followed by N digits then let N be the value of the block interpreted as the binary expansion.
   (c) Apply the previous step for the next block.

Chapter 4

2. We have

$$\begin{aligned} f(n) &= f(0) + \sum_{k=1}^{n} \left(\varDelta f(0) + \sum_{j=1}^{k} \varDelta^2 f(j) \right) \\ &= n\varDelta f(0) + \sum_{k=1}^{n}\sum_{j=1}^{k} \varDelta^2 f(j) = n\varDelta f(0) + \sum_{k=1}^{n} (n-k+1)\varDelta^2 f(k). \end{aligned}$$

Then

$$\begin{aligned} & f(n) + f(m) - f(n+m) \\ &= \sum_{k=1}^{\infty} \big[(n-k+1)\mathbf{1}\{k \le n\} + (m-k+1)\mathbf{1}\{k \le m\} \\ &\qquad - (n+m-k+1)\mathbf{1}\{k \le n+m\}\big] \varDelta^2 f(k) \ge 0. \end{aligned}$$

3. The term $n^{-1}a_n$ is the arithmetic mean of the first n increments. Averages of a decreasing sequence decrease to the same limit, so (a) implies (b) and the equality of the respective limits. If a_n were not increasing, there would be a negative increment and all following increments would be negative, leading eventually to negative values of a_n (assumed to be nonnegative). Further (b) implies $(m+n)^{-1}a_{m+n} \le \min\{m^{-1}a_m, n^{-1}a_n\}$, hence

$$\frac{a_{m+n}}{m+n} \le \frac{m}{m+n}\frac{a_m}{m} + \frac{n}{m+n}\frac{a_n}{n} = \frac{a_m + a_n}{m+n},$$

which yields (c).
Next, assume (c). If $m = kn + r$ with $0 \le r < n$ then

$$\frac{a_m}{m} \le \frac{ka_n}{m} + \frac{a_r}{m} \le \frac{a_n}{n} + \frac{na_1}{m}.$$

Hence for every n, $\limsup_{m\to\infty} m^{-1}a_m \le n^{-1}a_n$. This implies (d).

4. We have

$$\begin{aligned} & f(p_1 n + p_2 m) - p_1 f(n) - p_2 f(m) \\ &= \sum_{k=1}^{\infty} \big[(p_1 n + p_2 m - k + 1)\mathbf{1}\{k \le p_1 n + p_2 m\} \\ &\qquad - p_1(n-k+1)\mathbf{1}\{k \le n\} - p_2(m-k+1)\mathbf{1}\{k \le m\}\big] \varDelta^2 f(k) \ge 0. \end{aligned}$$

Chapter 5

2. We have

$$\begin{aligned} \mathbf{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_i - \mu\right]^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(|i-j|) = \frac{\sigma(0)}{n} + \frac{2}{n^2}\sum_{1\le k\le l\le n}\sigma(l-k) \\ &= \frac{\sigma(0)}{n} + \sum_{k=1}^{n}\frac{k\sigma(n-k-1)}{n^2} = \int_0^1 \frac{\lceil xn\rceil}{n}\sigma(n - \lceil xn\rceil - 1)\,\mathrm{d}x. \end{aligned}$$

We have $\frac{\lceil xn\rceil}{n}\sigma(n-\lceil xn\rceil-1) \le \sigma(0)$ for $x \in (0,1)$, where $\sigma(0)$ is integrable on the section $(0,1)$. If $\sigma(n)$ tends to 0 for $n \to \infty$, then by the dominated convergence theorem (Theorem 1.8),

$$\begin{aligned} & \lim_{n\to\infty}\int_0^1 \frac{\lceil xn\rceil}{n}\sigma\big(n - \lceil xn\rceil - 1\big)\,\mathrm{d}x \\ &= \int_0^1 \lim_{n\to\infty}\left[\frac{\lceil xn\rceil}{n}\sigma\big(n - \lceil xn\rceil - 1\big)\right]\mathrm{d}x = 0. \end{aligned}$$

3. Let $A_1, A_2, A_3, ..., A_m$ be the maximal closed sets of states. Process $(X_i)_{i=-\infty}^{\infty}$ would be ergodic if $X_1 \in A_l$ for a fixed l. Hence we obtain

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k &= \sum_{l=1}^{m} \mathbf{1}\{X_1 \in A_l\} \int X_1 \, \mathrm{d}P\big(\cdot|X_1 \in A_l\big) \\ &= \sum_{l=1}^{m} \mathbf{1}\{X_1 \in A_l\} \sum_{x} xP\big(X_1 = x|X_1 \in A_l\big).\end{aligned}$$

4. The increase of our capital is $W_n/W_0 = \prod_{i=1}^{n} b_{K_i} x_{K_i}$, which may be written down as

$$\log W_n/W_0 = \sum_{i=1}^{n} \log b_{K_i} + \sum_{i=1}^{n} \log x_{K_i}.$$

Applying the ergodic theorem, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log W_n/W_0 = \sum_{k=1}^{q} p_k \log b_k + \sum_{i=1}^{q} p_k \log x_k.$$

The only possibility of maximizing the capital increase lies in minimizing *cross entropy*

$$H(p||b) := -\sum_{k=1}^{q} p_k \log b_k = H(p) + D(p||b) \geq 0.$$

Entropy $H(p)$ is out of our control. In contrast, Kullback-Leibler divergence $D(p||b)$ is minimized for $b_k = p_k$. Thus we see that $b_k = p_k$ is the optimal strategy and it does not depend on bookmakers' stakes x_k. The return of our capital does depend on x_k, however. In the limit, we get $2^{-n[H(p)+H(p||x)]}$ dollars in n races from each invested dollar.

Chapter 7

1. By stationarity,

$$P_{\{2,...,n\}} X_1 = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{j+1}, \qquad P_{\{2,...,n\}} X_{n+1} = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{n+1-j}.$$

The best linear predictor $P_{\{1,...,n\}} X_{n+1}$ may be decomposed as the sum of the best linear predictor $P_{\{2,...,n\}} X_{n+1}$ and a term proportional to innovation $X_1 - P_{\{2,...,n\}} X_1$. In view of this we have

$$P_{\{1,...,n\}} X_{n+1} = P_{\{2,...,n\}} X_{n+1} + a(X_1 - P_{\{2,...,n\}} X_1),$$

where

$$a = \frac{\operatorname{Cov}(X_{n+1}, X_1 - P_{\{2,\dots,n\}}X_1)}{\big||X_1 - P_{\{2,\dots,n\}}X_1\big||^2}$$
$$= \frac{\operatorname{Cov}(X_{n+1} - P_{\{2,\dots,n\}}X_1, X_1 - P_{\{2,\dots,n\}}X_1)}{\big||X_1 - P_{\{2,\dots,n\}}X_1\big||^2} = \alpha(n).$$

Hence

$$\alpha(n) = \frac{\rho(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\rho(n-j)}{v_{n-1}},$$

which is (7.8). Besides,

$$P_{\{1,\dots,n\}}X_{n+1} = \alpha(n)X_1 + \sum_{j=1}^{n-1}(\phi_{n-1,j} - \alpha(n)\phi_{n-1,n-j})X_{n+1-j},$$

so we obtain (7.9)–(7.10). It remains to derive (7.11). Indeed we have

$$v_n = \frac{\big||X_{n+1} - P_{\{1,\dots,n\}}X_{n+1}\big||^2}{||X_{n+1}||^2}$$
$$= \frac{\big||X_{n+1} - P_{\{2,\dots,n\}}X_{n+1} - \alpha(n)(X_1 - P_{\{2,\dots,n\}}X_1)\big||^2}{||X_{n+1}||^2}$$
$$= v_{n-1} - 2\alpha(n)^2 v_{n-1} + \alpha(n)^2 v_{n-1} = \left[1 - \alpha(n)^2\right] v_{n-1}.$$

6. Observe that $(\alpha(n))_{n\in\mathbb{N}}$ defined in (7.21) satisfies $|\alpha(n)| \le 1$ for $n \ge 1$ and $d \in (-\infty, 1/2)$ and therefore is a PACF of a Gaussian process. PACF determines the coefficients ϕ_{nk} and $\rho(k)$ through iterations (7.9) and (7.18) uniquely given the initial conditions $\phi_{n0} = -1$, $\phi_{n,n+1} = 0$, and $\rho(0) = 1$. These initial conditions are clearly satisfied. Hence to demonstrate that (7.20) and (7.19) are the appropriate coefficients pertaining to the process, it suffices to check that (7.9) is satisfied for $n \ge 1$ and $0 \le j \le n$ given (7.20)–(7.21) and that (7.18) is satisfied for $n \ge 1$ given (7.19)–(7.20). Indeed, for (7.20) and (7.21) we obtain (7.9):

$$\phi_{n-1,k} - \frac{d}{n-d}\,\phi_{n-1,n-k}$$
$$= -\left[\binom{n}{k}\frac{n-k}{n} - \frac{d}{n-d}\binom{n}{k}\frac{k}{n}\right]\frac{(k-d-1)!(n-1-d-k)!}{(-d-1)!(n-1-d)!}$$
$$= -\binom{n}{k}\frac{n-k-d}{n-d}\cdot\frac{(k-d-1)!(n-1-d-k)!}{(-d-1)!(n-1-d)!}$$
$$= -\binom{n}{k}\frac{(k-d-1)!(n-d-k)!}{(-d-1)!(n-d)!} = \phi_{nk}.$$

On the other hand, for (7.19) and (7.20) we obtain (7.18):

$$\begin{aligned}\sum_{k=0}^{n} \phi_{nk}\rho(n-k) &= -\sum_{k=0}^{n} \binom{n}{k} \frac{\prod_{i=1}^{k}(i-d-1)\cdot\prod_{i=1}^{n-k}(i+d-1)}{\prod_{i=1}^{n}(i-d)} \\ &= -\left[\binom{n-d}{n}\right]^{-1} \sum_{k=0}^{n} \binom{k-d-1}{k}\binom{n-k+d-1}{n-k} \\ &= -\left[\binom{n-d}{n}\right]^{-1} \sum_{k=0}^{n} \binom{d}{k}\binom{-d}{n-k}(-1)^n \\ &= -\left[\binom{n-d}{n}\right]^{-1} \binom{d-d}{n}(-1)^n = 0,\end{aligned}$$

where we used the upper negation formula $\binom{r}{k} = (-1)^k\binom{k-r-1}{k}$ and the Cauchy formula $\sum_{k=0}^{n}\binom{r}{k}\binom{s}{n-k} = \binom{r+s}{n}$. Both formulae hold for all $r, s \in \mathbb{R}$ (Graham et al., 1994, Chapter 5, Table 202).

In the sequel, let us establish the asymptotics for the autocorrelations. For $-d+1 \not\in \mathbb{N}$ we have $\rho_n = (-d)!(n+d-1)!/(n-d)!(d-1)!$. By the Stirling approximation $\lim_{|z|\to\infty} \Gamma(z)[e^{-z}z^{z-1/2}\sqrt{2\pi}]^{-1} = 1$ we have $\lim_n n^x\Gamma(n)/\Gamma(n+x) = 1$ so $\lim_n \rho_n/n^{-1+2d} = (-d)!/(d-1)!$.

Finally, we inspect the sum of autocorrelations. Notice that $\sum_{k=-n}^{n}\rho_k = \prod_{i=1}^{n}(i+d)/(i-d)$ follows by induction on n. Hence $\sum_{k=-\infty}^{\infty}\rho_k = \infty$ for $d > 0$ and $\sum_{k=-\infty}^{\infty}\rho_k = 0$ for $d < 0$.

Chapter 9

1. We have

$$0 = \left.\frac{d\ln P(X_1^n = x_1^n|\theta)}{d\theta_l}\right|_{\theta=\theta_{\mathrm{ML}}(x_1^n)} = \left(\sum_{i=1}^{n} T_l(x_i) - n\left.\frac{d\ln Z(\theta)}{d\theta_l}\right|_{\theta=\theta_{\mathrm{ML}}(x_1^n)}\right).$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n} T_l(x_i) = \left.\frac{d\ln Z(\theta)}{d\theta_l}\right|_{\theta=\theta_{\mathrm{ML}}(x_1^n)} = \mathbf{E}_{\theta_{\mathrm{ML}}(x_1^n)} T_l(X_1),$$

which gives the maximum likelihood estimator in an implicit form.

4. We have

$$\begin{aligned} \mathbf{E}_\theta X_i &= \mu, & \mathbf{E}_\theta X_i^2 &= \sigma^2 + \mu^2, \\ \mathbf{E}_\theta \bar{X}_n &= \mu, & \mathbf{E}_\theta \bar{X}_n^2 &= \frac{1}{n}(\sigma^2+\mu^2) + \frac{n-1}{n}\mu^2. \end{aligned}$$

Hence

$$\mathbf{E}_\theta S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_\theta\big(X_i - \bar{X}_n\big)^2 = \mathbf{E}_\theta X_i^2 - 2\mathbf{E}_\theta\big(X_i\bar{X}_n\big) + \mathbf{E}_\theta\bar{X}_n^2$$

$$= \sigma^2 + \mu^2 - \frac{2}{n}(\sigma^2 + \mu^2) - \frac{2(n-1)}{n}\mu^2 + \frac{1}{n}(\sigma^2 + \mu^2) + \frac{n-1}{n}\mu^2$$
$$= \frac{n-1}{n}\sigma^2.$$

5. Let us write $L(x|\theta) = P(X_i = x|\theta)$. Observe that

$$\begin{aligned}
\frac{\partial}{\partial\omega} D(\omega||\theta) &= \frac{\partial}{\partial\omega} \sum_{x\in\mathbb{X}} L(x|\omega) \ln \frac{L(x|\omega)}{L(x|\theta)} \\
&= \sum_{x\in\mathbb{X}} \frac{\partial L(x|\omega)}{\partial\omega} \ln \frac{L(x|\omega)}{L(x|\theta)} + \sum_{x\in\mathbb{X}} \frac{\partial L(x|\omega)}{\partial\omega} \\
&= \sum_{x\in\mathbb{X}} \frac{\partial L(x|\omega)}{\partial\omega} \ln \frac{L(x|\omega)}{L(x|\theta)} + \frac{\partial}{\partial\omega} \sum_{x\in\mathbb{X}} L(x|\omega) \\
&= \sum_{x\in\mathbb{X}} \frac{\partial L(x|\omega)}{\partial\omega} \ln \frac{L(x|\omega)}{L(x|\theta)} + \frac{\partial}{\partial\omega} 1 \\
&= \sum_{x\in\mathbb{X}} \frac{\partial L(x|\omega)}{\partial\omega} \ln \frac{L(x|\omega)}{L(x|\theta)}.
\end{aligned}$$

Differentiating again, we obtain

$$\begin{aligned}
\frac{\partial^2}{\partial\omega^2} D(\omega||\theta) &= \frac{\partial}{\partial\omega} \sum_{x\in\mathbb{X}} \frac{\partial L(x|\omega)}{\partial\omega} \ln \frac{L(x|\omega)}{L(x|\theta)} \\
&= \sum_{x\in\mathbb{X}} \frac{\partial^2 L(x|\omega)}{\partial\omega^2} \ln \frac{L(x|\omega)}{L(x|\theta)} + \sum_{x\in\mathbb{X}} \frac{1}{L(x|\omega)} \left[\frac{\partial L(x|\omega)}{\partial\omega}\right]^2.
\end{aligned}$$

In the formula above, the first term vanishes for $\omega = \theta$, whereas the second one equals $J_1(\theta)$.

7. We have

$$(J_1(\theta))_{ij} = \frac{\partial^2 \ln Z(\theta)}{\partial\theta_i \partial\theta_j} = \mathbf{E}_\theta \big[T_i(X_1) - \mathbf{E}_\theta T_i(X_1)\big]\big[T_j(X_1) - \mathbf{E}_\theta T_j(X_1)\big],$$

where

$$\mathbf{E}_\theta T_i(X_1) = \frac{\partial \ln Z(\theta)}{\partial\theta_i}.$$

Hence $\ln Z(\theta)$ is a convex function.

Chapter 10

1. We have

$$\pi_{\text{Jeffreys}}(\theta) \propto \sqrt{J_1(\theta)} = \sqrt{\det \frac{\partial^2 \ln Z(\theta)}{\partial\theta_i \partial\theta_j}}.$$

2. Let

$$\pi(\lambda|\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

be the gamma distribution. We have

$$\begin{aligned}\pi(\lambda|X_1^n=x_1^n,\alpha,\beta)&=\frac{\lambda^{\sum_{i=1}^n x_i}e^{-n\lambda}\lambda^{\alpha-1}e^{-\beta\lambda}}{\int\lambda'^{\sum_{i=1}^n x_i}e^{-n\lambda'}\lambda'^{\alpha-1}e^{-\beta\lambda'}\,\mathrm{d}\lambda'}\\&=\pi\left(\lambda\,\middle|\,\sum_{i=1}^n x_i+\alpha,n+\beta\right)\end{aligned}$$

and

$$\begin{aligned}P\big(X_1^n=x_1^n|\alpha,\beta\big)&=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\int\frac{\lambda^{\sum_{i=1}^n x_i}e^{-n\lambda}}{\prod_{i=1}^n x_i!}\lambda^{\alpha-1}e^{-\beta\lambda}\,\mathrm{d}\lambda\\&=\frac{1}{\prod_{i=1}^n x_i!}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\frac{\Gamma\left(\sum_{i=1}^n x_i+\alpha\right)}{(n+\beta)^{\sum_{i=1}^n x_i+\alpha}}.\end{aligned}$$

7. We have

$$\begin{aligned}\frac{P(\text{ill}|\text{positive})}{P(\text{healthy}|\text{positive})}&=\frac{P(\text{positive}|\text{ill})P(\text{ill})}{P(\text{positive}|\text{healthy})P(\text{healthy})}\\&=\frac{0.9999\cdot 0.0001}{0.0001\cdot 0.9999}=1.\end{aligned}$$

Hence $P(\text{ill}|\text{positive})=1/2$.

Chapter 11

4. Let us compute the EM iteration. Cross entropy $Q(\theta',\theta)$ takes form

$$Q(\theta',\theta)=\sum_i\sum_j P\big(Z_i=j|Y_i=A_i,\theta'\big)\ln P\big(Z_i=j,Y_i=A_i|\theta\big).$$

Moreover, (11.8) assures that $P(Y_i=A_i|\theta)=g(A_i)P(Z_i\in A_i|\theta)$ and

$$P\big(Z_i=j|Y_i=A_i,\theta\big)=P\big(Z_i=j|Z_i\in A_i,\theta\big).$$

Let us write $p_{ji}^{(n)}=P(Z_i=j|Z_i\in A_i,\theta_n)$. Recalling constrained minimization using Langrange multipliers, we have that iteration (11.3) is equivalent to

$$0=\frac{\partial}{\partial p_j}\left[Q(\theta_n,\theta)-\lambda\bigg(\sum_{j'\in J}p_{j'}-1\bigg)\right]\Bigg|_{\theta=\theta_{n+1}}=\frac{\sum_{i=1}^M p_{ji}^{(n)}}{p_j^{(n+1)}}-\lambda.$$

If the Lagrange multiplier λ is assigned the value that satisfies constraint $\sum_{j \in J} p_{j'} = 1$, we obtain iteration

$$p_{ji}^{(n)} = \begin{cases} p_j^{(n)} / \sum_{j' \in A_i} p_{j'}^{(n)}, & j \in A_i, \\ 0, & \text{else}, \end{cases} \tag{15.9}$$

$$p_j^{(n+1)} = \frac{1}{M} \sum_{i=1}^{M} p_{ji}^{(n)}. \tag{15.10}$$

As an initial value we may take $p_j^{(1)} = [\operatorname{card} J]^{-1}$.
Iteration (15.9)–(15.10) maximizes locally the log-likelihood

$$L(\theta) := \ln P((Y_i = A_i)_{i=1}^{M} | \theta) = \ln \left[\prod_{i=1}^{M} \frac{P(Z_i \in A_i | \theta)}{g(A_i)} \right], \tag{15.11}$$

or simply $L^{(n+1)} \geq L^{(n)}$ for

$$L^{(n)} := L(\theta_n) + \sum_{i=1}^{M} \ln g(A_i) = \sum_{i=1}^{M} \ln \left[\sum_{j \in A_i} p_j^{(n)} \right], \quad n \geq 2.$$

Moreover, there is no need to care for the initialization of iteration (15.9)–(15.10) since the local maxima of function (15.11) form a convex set $\mathcal{M}$, i.e., $\theta, \theta' \in \mathcal{M} \implies q\theta + (1-q)\theta' \in \mathcal{M}$ for $0 \leq q \leq 1$. Hence that function is, of course, constant on $\mathcal{M}$. To show this, observe that the domain of log-likelihood (15.11) is a convex compact set $\mathcal{P} = \left\{ \theta : \sum_j p_j = 1, \ p_j \geq 0 \right\}$. The second derivative of L reads

$$L_{jj'}(\theta) := \frac{\partial^2 L(\theta)}{\partial p_j \partial p_{j'}} = - \sum_{i=1}^{M} \frac{\mathbf{1}\{j \in A_i\} \mathbf{1}\{j' \in A_i\}}{\left(\sum_{j'' \in A_i} p_{j''} \right)^2}.$$

Since matrix $\{L_{jj'}\}$ is negative definite, i.e., $\sum_{jj'} a_j L_{jj'}(\theta) a_{j'} \leq 0$, function L is concave. As a general fact, a continuous function L achieves its supremum on a compact set $\mathcal{P}$ (Rudin, 1974, Theorem 2.10). If additionally L is concave and its domain $\mathcal{P}$ is convex then the local maxima of L form a convex set $\mathcal{M} \subset \mathcal{P}$, where L is constant and achieves its supremum (Boyd and Vandenberghe, 2004, Section 4.2.2).

Chapter 12

1. $H(X_n | X_1) \geq H(X_n | X_1, X_2) = H(X_n | X_2) = H(X_{n-1} | X_1)$.
2. $H(TX) \geq H(TX | T) = H(T^{-1}TX | T) = H(X | T) = H(X)$.
6. Entropy of a variable U is maximized for the fixed expectation when the variable has geometric distribution $P(U = k) = (1 - p)^k p$. The expectation for that distribution is $\mu = (1 - p)/p$ and the entropy equals $H(U) = -\log p - \frac{1-p}{p} \log(1 - p)$. Substituting $p = 1/(\mu + 1)$, we obtain $H(U) = (\mu + 1) \log(\mu + 1) - \mu \log \mu$.

7. By formulae (12.10)–(12.11) we have

$$H(\rho||\rho_\lambda) = -\int \rho(x)\left[\sum_{i=1}^{m} \lambda_i T_i(x) - \ln Z(\lambda)\right] \mathrm{d}x = L(\lambda)$$

if ρ satisfies (12.8)–(12.9). Hence from (12.12), $H(\rho||\rho_\lambda)$ achieves minimum for $\rho_\lambda = \rho^*$.

Chapter 13

4. Consider language $L_w = \{u : wu \in L\}$. For $w = 0^n1$ the first element of L_w is $u = 0^n1$. If L were regular, we would have $C(u) = O(1)$ by Lemma 13.1. In this way we obtain a contradiction for an incompressible n.

Chapter 14

6. We have

$$\begin{aligned} K(u|w^*) &\overset{+}{<} K(\langle u, z\rangle\,|w^*) \overset{+}{=} K(\langle u, z, w\rangle) - K(w) \\ &\overset{+}{<} K(z) + K(u|z^*) + K(w|z^*) - K(w) \\ &\overset{+}{=} K(\langle w, z\rangle) - K(w) + K(u|z^*) \\ &\overset{+}{=} K(u|z^*) + K(z|w^*). \end{aligned}$$

Hence we obtain the triangle inequality

$$\begin{aligned} \mathrm{ID}(u, w) &= \max\big\{K(u|w^*), K(w|u^*)\big\} \\ &\overset{+}{<} \max\big\{K(u|z^*) + K(z|w^*), K(w|z^*) + K(z|u^*)\big\} \\ &\le \max\big\{K(u|z^*), K(z|u^*)\big\} + \max\big\{K(z|w^*), K(w|z^*)\big\} \\ &= \mathrm{ID}(u, z) + \mathrm{ID}(z, w). \end{aligned}$$

7. Let $k = K(u)$. We have

$$K\big(K(k)|u^*\big) \overset{+}{<} K\big(K(k)|k^*\big) + K(k|u^*) \overset{+}{=} 0.$$

Chapter 15

3. Let $f(n) = \log n + \log\log n$. If we had $K(n) < f(n)$ for all $n \ge N$ then

$$\sum_{n\ge N} 2^{-f(n)} \le \sum_{n\ge N} 2^{-K(n)} \le 1$$

But $\sum_n 2^{-f(n)} = \sum_n (n \log n)^{-1} = \infty$ so $\sum_{n\ge N} 2^{-f(n)} = \infty$. Hence $K(n) < f(n)$ for infinitely many n.

Bibliography

O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* New York: John Wiley, 1978.

A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71, 1996.

P. Billingsley. *Probability and Measure.* New York: John Wiley, 1979.

C. M. Bishop. *Pattern Recognition and Machine Learning.* New York: Springer, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge: Cambridge University Press, 2004.

L. Breiman. *Probability.* Philadephia: SIAM, 1992.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* New York: Springer, 1987.

G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22:329–340, 1975a.

G. J. Chaitin. Randomness and mathematical proof. *Scientific American*, 232 (5):47–52, 1975b.

G. J. Chaitin. *Algorithmic Information Theory.* Cambridge: Cambridge University Press, 1987.

T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd ed.* New York: John Wiley, 2006.

S. Dégerine and S. Lambert-Lacroix. Characterization of the partial autocorrelation function of a nonstationary time series. *Journal of Multivariate Analysis*, 87:46–59, 2003.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39:185–197, 1977.

J. Durbin. The fitting of time series models. *Review of the International Statistical Institute*, 28:233–244, 1960.

D. Gillman and R. L. Rivest. Complete variable-length "fix-free" codes. *Designs, Codes and Cryptography*, 5:109–114, 1995.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics. A Foundation for Computer Science.* Reading: Addison-Wesley, 1994.

U. Grenander and G. Szegő. *Toeplitz Forms and Their Applications.* Berkeley: University of California Press, 1958.

P. D. Grünwald. *The Minimum Description Length Principle.* Cambridge, MA: The MIT Press, 2007.

O. Kallenberg. *Foundations of Modern Probability.* New York: Springer, 1997.

R. W. Keener. *Theoretical Statistics. Topics for a Core Course.* New York: Springer, 2010.

J. C. Kieffer and E. Yang. Grammar-based codes: A new class of universal lossless source codes. *IEEE Transactions on Information Theory*, 46:737–754, 2000.

D. Knuth. The complexity of songs. *Communications of the ACM*, 27:345–348, 1984.

A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications, 3rd ed.* New York: Springer, 2008.

A. I. McLeod. Hyperbolic decay time series. *The Journal of Time Series Analysis*, 19:473–483, 1998.

F. L. Ramsey. Characterization of the partial autocorrelation function. *The Annals of Statistics*, 2:1296–1301, 1974.

W. Rudin. *Real and complex analysis.* New York: McGraw-Hill, 1974.

I. Schur. Über Potenzreihen, die im Innern des Einheitskreises beschrankt sind. *Journal für die Reine und Angewandte Mathematik*, 147:205–232, 1917.

C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 30:379–423,623–656, 1948.

C. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64, 1951.

R. J. Solomonoff. A formal theory of inductive inference, part 1 and part 2. *Information and Control*, 7:1–22, 224–254, 1964.

H. Takahashi. On a definition of random sequences with respect to conditional probability. *Information and Computation*, 206:1375–1382, 2008.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge: Cambridge University Press, 1998.

V. G. Vovk and V. V. V'yugin. On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society, series B*, 55:253–266, 1993.

V. G. Vovk and V. V. V'yugin. Prequential level of impossibility with some applications. *Journal of the Royal Statistical Society, series B*, 56:115–123, 1994.

V. V. V'yugin. On empirical meaning of randomness with respect to a real parameter. In *Computer Science — Theory and Applications*, pages 387–396. New York: Springer, 2007.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.

Index

K-means algorithm, 104
π-λ theorem, 39
π-system, 39
σ-field, 7

absolutely continuous measure, 9
algorithmic information, 133
algorithmic probability, 131
ancillary statistic, 77
autocorrelation function, 65

Barron theorem, 135
Basu theorem, 77
Baum-Welch algorithm, 104
Bernoulli distribution, 71
best linear predictor, 65
beta distribution, 78
bias, 84
binary tree, 33
block, 42
block entropy, 42
Boltzmann equation, 110
Borel-Cantelli lemma, 11
Bregman divergence, 26, 90

chain rule, 25
Chebyshev inequality, 14
code, 27
code tree, 33
coding theorem, 131
comma-separated code, 28
complete code, 32
complete statistic, 76
computable function, 120, 127, 128
concave function, 18
conditional entropy, 19
conditional expectation, 10
conditional independence, 22
conditional mutual information, 21
conditional probability, 10
conditionally Martin-Löf random sequence, 143
conjugate prior, 98
consistent estimator, 82
convergence in distribution, 88
convex function, 18
Cramér-Rao theorem, 87
cross entropy, 115

data-processing inequality, 25
deterministic finite state automaton, 121
differential entropy, 63
discrete random sample, 71
dominated convergence theorem, 12
Doob consistency theorem, 93
Durbin-Levinson algorithm, 66
dynamical system, 39

effective Kraft inequality, 130
effectively strictly consistent estimator, 143
Elias omega code, 38, 119, 128
entropy, 16, 17, 63
entropy rate, 43
ergodic dynamical system, 48
ergodic stationary process, 48
ergodic theorem, 48
excess entropy, 44
exchangeable process, 70
expectation, 8
expectation-maximization algorithm, 102
expected Fisher information, 85
exponential distribution, 79
exponential family, 75

Fano inequality, 25
Fatou lemma, 12
filtration, 13
Fisher factorization theorem, 73
fixed-length code, 28
formal language, 121

gamma distribution, 78
Gauss distribution, 72
Gaussian process, 65
generalized Pythagoras theorem, 26
generalized variance, 68
generated σ-field, 39

geometric distribution, 78

H-theorem, 111
Hahn-Jordan decomposition, 10
halting probability, 131, 139
Herglotz theorem, 68
Huffman code, 34

impossibility level, 140
improper prior, 96
incompressibility method, 121
independent random variables, 15
information distance, 134
information-theoretic Gödel theorem, 120
injection, 27
innovation, 65
innovation variance, 68
invariance theorem, 119, 126
invariant algebra, 47
invariant measure, 39
inverse Gauss distribution, 79

Jeffreys prior, 95
Jensen inequality, 18

Kolmogorov complexity, 118, 119
Kolmogorov formula, 69
Kraft inequality, 29
Krichevski-Trofimov estimator, 95
Kullback-Leibler divergence, 18, 64

Lagrange multipliers, 103
Lagrangian function, 113
Laplace rule, 94
Lebesgue-Radon-Nikodym theorem, 9
Lempel-Ziv code, 58
Levy law, 13
lower semicomputable function, 128

Markov chain, 22
Markov inequality, 12
Martin-Löf random sequence, 136
Martin-Löf test, 137
martingale, 13
maximal ergodic theorem, 49
maximum entropy modeling, 112
maximum likelihood estimator, 81
maximum posterior estimator, 93
mean square error, 84
measurable function, 9
measure, 9
minimal sufficient statistics, 74
monotone convergence theorem, 12
Monty Hall paradox, 13
moving average representation, 70
mutual information, 20
mutually singular measure, 9

negative binomial distribution, 78
normal distribution, 72
normalized information distance, 134

oscillations of Kolmogorov complexity, 122

parametric family of distributions, 71
Pareto distribution, 78
partial autocorrelation function, 65
path, 33
Poisson distribution, 78
posterior distribution, 93
prefix-free code, 30
prefix-free Kolmogorov complexity, 126, 127
prior distribution, 93
probability kernel, 142
probability measure, 7, 135
probability space, 7
process theorem, 40

random sample, 71
random variable, 8
real random sample, 71
recursive function, 120
recursively enumerable function, 130
recursively enumerable set, 137
regular language, 121
relative entropy, 18

Schnorr theorem, 137
self-delimiting program, 125
semimeasure, 141
Shannon-Fano code, 32
Shannon-McMillan-Breiman theorem, 52
Solovay test, 137
source coding inequality, 30
stationary process, 40
statistic, 72
Stirling approximation, 25
stochastic process, 8

Stosszahlansatz, 109
strong law of large numbers, 14, 47
subadditive function, 46
sufficient statistic, 72
suffix-free code, 30

triple information, 24
Turing machine, 117, 125

unbiased estimator, 84
uniquely decodable code, 27
unit integrable function, 140
universal code, 57
universal Turing machine, 119, 126
upper semicomputable function, 128

Venn diagram, 24

Wald theorem, 83
weak law of large numbers, 14
weakly ergodic stationary process, 54
weakly stationary process, 54
weighted code tree, 34

Yule-Walker equations, 69

Ziv inequality, 59