

Lectures on Entropy. Part I.

VOJKAN JAKŠIĆ

Department of Mathematics and Statistics
McGill University
805 Sherbrooke Street West
Montreal, QC, H3A 2K6, Canada

© 2018 Vojkan Jakšić
All Rights Reserved

June 20, 2018

Contents

1	Introduction	5
1.1	Notes and references.	6
2	Elements of probability	7
2.1	Prologue: integration on finite sets	7
2.2	Probability on finite sets	9
2.3	Law of large numbers	10
2.4	Cumulant generating function	11
2.5	Rate function	12
2.6	Cramér’s theorem	14
2.7	Notes and references	17
3	Boltzmann–Gibbs–Shannon entropy	19
3.1	Preliminaries	19
3.2	Definition and basic properties	20
3.3	Covering exponents and source coding	22
3.4	Why is the entropy natural?	24
3.4.1	Split additivity characterization	24
3.4.2	Sub-additivity characterization	27
3.5	Rényi entropy	34
3.6	Why is the Rényi entropy natural?	35
3.7	Notes and references	36
4	Relative entropy	39
4.1	Definition and basic properties	39
4.2	Variational principles	44
4.3	Stein’s Lemma	46
4.4	Fluctuation relation	48
4.5	Jensen-Shannon entropy and metric	49
4.6	Rényi’s relative entropy	53
4.7	Hypothesis testing	56
4.8	Asymmetric hypothesis testing	59

4.9	Notes and references	64
5	Why is the relative entropy natural?	67
5.1	Introduction	67
5.2	Proof of Theorem 5.1	69
5.3	Proof of Theorem 5.2	71
5.4	Sanov's theorem	74
5.5	Notes and references	79
6	Fisher entropy	81
6.1	Definition and basic properties	81
6.2	Entropic geometry	82
6.3	Chentsov's theorem	85
6.4	Notes and references	88
7	Parameter estimation	89
7.1	Introduction	89
7.2	Basic facts	89
7.3	Two remarks	92
7.4	The maximum likelihood estimator	94
7.5	Notes and references	100
	Bibliography	103

Chapter 1

Introduction

These lecture notes concern information-theoretic notions of entropy. They are intended for, and have been successfully taught to, undergraduate students interested in research careers. Besides basic notions of analysis related to convergence that are typically taught in the first or second year of undergraduate studies, no other background is needed to read the notes. The notes might be also of interest to any mathematically inclined reader who wishes to learn basic facts about notions of entropy in an elementary setting.

As the title indicates, this is the first in a planned series of four lecture notes. The Part II concerns notions of entropy in study of statistical mechanics, and III/IV are the quantum information theory/quantum statistical mechanics counterparts of I/II. All four parts target similar audience and are on a similar technical level. Eventually, Parts I-IV together are intended to be an introductory chapter to a comprehensive volume dealing with the topic of entropy from a certain point of view on which I will elaborate below.

The research program that leads to these lecture notes concerns the elusive notion of entropy in non-equilibrium statistical mechanics. It is for this pursuit that the notes are preparing a research-oriented reader, and it is the pursuit to which the later more advanced topics hope to contribute. Thus, it is important to emphasize that the choice of topics and their presentation have a specific motivation which may not be obvious until at least the Part II of the lecture notes is completed. Needless to say, the lecture notes can be read independently of its motivation, as they provide a concise, elementary, and mathematically rigorous introduction to the topics they cover.

The theme of this Part I is the Boltzmann–Gibbs–Shannon (BGS) entropy of a finite probability distribution (p_1, \dots, p_n) , and its various deformations such as the Rényi entropy, the relative entropy, and the relative Rényi entropy. The BGS entropy and the relative entropy have intuitive and beautiful axiomatic characterizations discussed in Section 3.4 and Chapter 5. The Rényi entropies also have axiomatic characterizations, but those are perhaps less natural, and we shall not discuss them in detail. Instead, we shall motivate the Rényi entropies by the so-called Large Deviation Principle (LDP) in probability theory. The link between the LDP and notions of entropy runs deep and will play a central role in this lecture notes. For this reason Cramér’s theorem is proven right away in the introductory Chapter 2 (the more involved proof of Sanov’s theorem is given in Section 5.4). It is precisely this emphasis on the LDP that makes this lecture notes somewhat unusual in comparison with other introductory presentations of the information-theoretic entropy.

The Fisher entropy and a related topic of parameter estimation are also an important part of this lecture notes. The historical background and most of applications of these topics are in the field of statistics. There is a hope that they may play an important role in study of entropy in non-equilibrium statistical mechanics, and that is the reason for including them in the lecture notes. Again, Chapters 6 and 7 can be read independently of this motivation by anyone interested in an elementary introduction to the Fisher entropy and parameter estimation.

These notes are work in progress, and additional topics may be added in the future.

The notes benefited from the comments of numerous McGill undergraduate students who attended the sem-

inars and courses in which I have taught the presented material. I am grateful for their help and for their enthusiasm which to a large extent motivated my decision to prepare the notes for publication. In particular, I am grateful to Sherry Chu, Wissam Ghantous, and Jane Panangaden whose McGill's undergraduate summer research projects were linked to the topics of the lecture notes and whose research reports helped me in writing parts of the notes. I am also grateful to Laurent Bruneau, Noé Cuneo, Tomas Langsetmo, Renaud Raquépas and Armen Shirikyan for comments and suggestions. I wish to thank Jacques Hurtubise and David Stephens who, as the chairmans of the McGill Department of Mathematics and Statistics, enabled me to teach the material of the notes in a course format. Finally, I am grateful to Marisa Rossi for her exceptional hospitality and support during the period when Chapter 7 was written.

This research that has led to this lecture notes was partly funded by NSERC, *Agence Nationale de la Recherche* through the grant NONSTOPS (ANR-17-CE40-0006-01, ANR-17-CE40-0006-02, ANR-17-CE40-0006-03), the CNRS collaboration grant *Fluctuation theorems in stochastic systems*, and the *Initiative d'excellence Paris-Seine*.

1.1 Notes and references.

Shannon's seminal 1948 paper [Sha], reprinted in [ShaWe], remains a must-read for anyone interested in notions of entropy. Khintchine's reworking of the mathematical foundations of Shannon's theory in early 1950's, summarized in the monograph [Khi], provides a perspective on the early mathematically rigorous developments of the subject. For further historical perspective we refer the reader to [Ver] and the detailed list of references provided there. There are many books dealing with entropy and information theory. The textbook [CovTh] is an excellent introduction to the subject, [Bill, Gra, Shi] are recommended to mathematically more advanced reader. Another instructive reference is [CsiKö], where a substantial part of the material covered in this lecture notes is left as an exercise for the reader!

Discussions of a link between information and statistical mechanics preceded Shannon's work. Although Weaver's remark¹ on page 3 of [ShaWe] appears to be historically inaccurate, the discussions of the role of information in foundations of statistical mechanics goes back at least to the work of L. Szillard [Szi] in 1929, see also <https://plato.stanford.edu/entries/information-entropy/>, and remains to this day a hotly disputed subject; see [GHLS] for a recent discussion. An early discussion can be found in [Jay1, Jay2]. The textbook [Mer] gives an additional perspective on this topic.

In contrast to equilibrium statistical mechanics whose mathematically rigorous foundations, based on the 19th century works of Boltzmann and Gibbs, were laid in 1960's and 70's, the physical and mathematical theory of non-equilibrium statistical mechanics remains in its infancy. The introduction of non-equilibrium steady states and the discovery of the fluctuation relations in context of chaotic dynamical systems in early 1990's (see [JPR] for references) revolutionized our understanding of some important corners of the field, and have generated an enormous amount of theoretical, experimental, and numerical works with applications extending to chemistry and biology. The research program of Claude-Alain Pillet and myself mentioned in the introduction is rooted in these developments.² In this program, the search for a notion of entropy for systems out of equilibrium plays a central role. The planned four parts lecture notes are meant as an introduction to this search, with this Part I focusing on the information-theoretic notions of entropy.

¹" Dr. Shannon's work roots back, as von Neumann has pointed out, to Boltzmann's observation, in some of his work on statistical physics (1894), that entropy is related to "missing information," inasmuch as it is related to the number of alternatives which remain possible to a physical system after all the macroscopically observable information concerning it has been recorded."

²The references to results of this program are not relevant for this Part I of the lectures and they will be listed in the latter installements.

Chapter 2

Elements of probability

2.1 Prologue: integration on finite sets

Let Ω be a finite set. Generic element of Ω is denoted by ω . When needed, we will enumerate elements of Ω as $\Omega = \{\omega_1, \dots, \omega_L\}$, where $|\Omega| = L$.

A measure on Ω is a map

$$\mu : \Omega \rightarrow \mathbb{R}_+ = [0, \infty[.$$

The pair (Ω, μ) is called measurable space. The measure of $S \subset \Omega$ is

$$\mu(S) = \sum_{\omega \in S} \mu(\omega).$$

By definition, $\mu(\emptyset) = 0$.

Let $f : \Omega \rightarrow \mathbb{C}$ be a function. The integral of f over $S \subset \Omega$ is defined by

$$\int_S f d\mu = \sum_{\omega \in S} f(\omega) \mu(\omega).$$

Let Ω and \mathcal{E} be two finite sets and $T : \Omega \rightarrow \mathcal{E}$ a map. Let μ be a measure on Ω . For $\zeta \in \mathcal{E}$ set

$$\mu_T(\zeta) = \mu(T^{-1}(\zeta)) = \sum_{\omega: T(\omega)=\zeta} \mu(\omega).$$

μ_T is a measure on \mathcal{E} induced by (μ, T) . If $f : \mathcal{E} \rightarrow \mathbb{C}$, then

$$\int_{\mathcal{E}} f d\mu_T = \int_{\Omega} f \circ T d\mu.$$

If $f : \Omega \rightarrow \mathbb{C}$, we denote by μ_f the measure on the set of values $\mathcal{E} = \{f(\omega) \mid \omega \in \Omega\}$ induced by (Ω, f) . μ_f is called the distribution measure of the function f .

We denote by

$$\begin{aligned} \Omega^N &= \{\omega = (\omega_1, \dots, \omega_N) \mid \omega_k \in \Omega\}, \\ \mu_N(\omega) &= (\omega_1, \dots, \omega_N) = \mu(\omega_1) \cdots \mu(\omega_N), \end{aligned}$$

the N -fold product set and measure of the pair (Ω, μ) .

Let Ω_l, Ω_r be two finite sets and μ a measure on $\Omega_l \times \Omega_r$. The marginals of μ are measures $\mu_{l/r}$ on Ω_l, Ω_r defined by

$$\mu_l(\omega) = \sum_{\omega' \in \Omega_r} \mu(\omega, \omega'), \quad \omega \in \Omega_l,$$

$$\mu_r(\omega) = \sum_{\omega' \in \Omega_l} \mu(\omega', \omega), \quad \omega \in \Omega_r.$$

If μ_l/r are measures on Ω_l/r . we denote by $\mu_l \otimes \mu_r$ the product measure defined by

$$\mu_l \otimes \mu_r(\omega, \omega') = \mu_l(\omega)\mu_r(\omega').$$

The support of the measure μ is the set

$$\text{supp } \mu = \{\omega \mid \mu(\omega) \neq 0\}.$$

Two measures μ_1 and μ_2 are mutually singular, denoted $\mu_1 \perp \mu_2$, iff $\text{supp } \mu_1 \cap \text{supp } \mu_2 = \emptyset$. A measure μ_1 is absolutely continuous w.r.t. another measure μ_2 , denoted $\mu_1 \ll \mu_2$, iff $\text{supp } \mu_1 \subset \text{supp } \mu_2$, that is, iff $\mu_2(\omega) = 0 \Rightarrow \mu_1(\omega) = 0$. If $\mu_1 \ll \mu_2$, the Radon-Nikodym derivative of μ_1 w.r.t. μ_2 is defined by

$$\Delta_{\mu_1|\mu_2}(\omega) = \begin{cases} \frac{\mu_1(\omega)}{\mu_2(\omega)} & \text{if } \omega \in \text{supp } \mu_1 \\ 0 & \text{if } \omega \notin \text{supp } \mu_1. \end{cases}$$

Note that

$$\int_{\Omega} f \Delta_{\mu_1|\mu_2} d\mu_2 = \int_{\Omega} f d\mu_1.$$

Two measures μ_1 and μ_2 are called equivalent iff $\text{supp } \mu_1 = \text{supp } \mu_2$.

Let μ, ρ be two measures on Ω . Then there exists a unique decomposition (called the Lebesgue decomposition) $\mu = \mu_1 + \mu_2$, where $\mu_1 \ll \rho$ and $\mu_2 \perp \rho$. Obviously,

$$\mu(\omega) = \begin{cases} \mu(\omega) & \text{if } \omega \in \text{supp } \rho \\ 0 & \text{if } \omega \notin \text{supp } \rho, \end{cases} \quad \mu_2(\omega) = \begin{cases} 0 & \text{if } \omega \in \text{supp } \rho \\ \mu(\omega) & \text{if } \omega \notin \text{supp } \rho. \end{cases}$$

A measure μ is called faithful if $\mu(\omega) > 0$ for all $\omega \in \Omega$.

Proposition 2.1 Let $f : \Omega \rightarrow \mathbb{R}_+$, $a > 0$, and $S_a = \{\omega \mid f(\omega) \geq a\}$. Then

$$\mu(S_a) \leq \frac{1}{a} \int_{\Omega} f d\mu.$$

Proof. The statement is obvious is $S_a = \emptyset$. If S_a is non-empty,

$$\mu(S_a) = \sum_{\omega \in S_a} \mu(\omega) \leq \frac{1}{a} \sum_{\omega \in S_a} f(\omega)\mu(\omega) \leq \frac{1}{a} \int_{\Omega} f d\mu.$$

□

We recall the Minkowski inequality

$$\left(\int_{\Omega} |f + g|^p d\mu \right)^{1/p} \leq \left(\int_{\Omega} |f|^p d\mu \right)^{1/p} + \left(\int_{\Omega} |g|^p d\mu \right)^{1/p},$$

where $p \geq 1$, and the Hölder inequality

$$\int_{\Omega} f g d\mu \leq \left(\int_{\Omega} |f|^p d\mu \right)^{1/p} \left(\int_{\Omega} |g|^q d\mu \right)^{1/q},$$

where $p, q \geq 1$, $p^{-1} + q^{-1} = 1$. For $p = q = 2$ the Hölder inequality reduces to the Cauchy-Schwarz inequality.

If $f : \Omega \rightarrow]-\infty, \infty]$ or $[-\infty, \infty[$, we again set $\int_{\Omega} f d\mu = \sum_{\omega} f(\omega)\mu(\omega)$ with the convention that $0 \cdot (\pm\infty) = 0$.

2.2 Probability on finite sets

We start with a change of vocabulary adapted to the probabilistic interpretation of measure theory.

A measure P on a finite set Ω is called a probability measure if $P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1$. The pair (Ω, P) is called probability space. A set $S \subset \Omega$ is called an event and $P(S)$ is the probability of the event S . Points $\omega \in \Omega$ are sometimes called elementary events.

A perhaps most basic example of a probabilistic setting is a fair coin experiment, where a coin is tossed N times and the outcomes are recorded as Head = 1 and Tail = -1. The set of outcomes is

$$\Omega = \{\omega = (\omega_1, \dots, \omega_N) \mid \omega_k = \pm 1\},$$

and

$$P(\omega = (\omega_1, \dots, \omega_N)) = \frac{1}{2^N}.$$

Let S be the event that k Heads and $N - k$ Tails are observed. The binomial formula gives

$$P(S) = \binom{N}{k} \frac{1}{2^N}.$$

As another example, let

$$S_j = \left\{ \omega = (\omega_1, \dots, \omega_N) \mid \sum_k \omega_k = j \right\},$$

where $-N \leq j \leq N$. $P(S_j) = 0$ if $N + j$ is odd. If $N + j$ is even, then

$$P(S_j) = \binom{N}{\frac{N+j}{2}} \frac{1}{2^N}.$$

A function $X : \Omega \rightarrow \mathbb{R}$ is called random variable.

The measure P_X induced by (P, X) is called the probability distribution of X . The expectation of X is

$$E(X) = \int_{\Omega} X dP.$$

The moments of X are

$$M_k = E(X^k), \quad k = 1, 2, \dots,$$

and the moment generating function is

$$M(\alpha) = E(e^{\alpha X}) = \sum_{\omega \in \Omega} e^{\alpha X(\omega)} P(\omega),$$

where $\alpha \in \mathbb{R}$. Obviously,

$$M_k = \frac{d^k}{d\alpha^k} M(\alpha) \Big|_{\alpha=0}.$$

The cumulant generating function of X is

$$C(\alpha) = \log E(e^{\alpha X}) = \log \left(\sum_{\omega \in \Omega} e^{\alpha X(\omega)} P(\omega) \right).$$

The cumulants of X are

$$C_k = \frac{d^k}{d\alpha^k} C(\alpha) \Big|_{\alpha=0}, \quad k = 1, 2, \dots$$

$C_1 = M_1 = E(X)$ and

$$C_2 = E(X^2) - E(X)^2 = E((X - E(X))^2).$$

C_2 is called the variance of X and is denoted by $\text{Var}(X)$. Note that $\text{Var}(X) = 0$ iff X is constant on $\text{supp } P$. When we wish to indicate the dependence of the expectation and variance on the underlying measure P , we shall write $E_P(X)$, $\text{Var}_P(X)$, etc.

Exercise 2.1. The sequences $\{M_k\}$ and $\{C_k\}$ determine each other, i.e., there are functions F_k and G_k such that

$$C_k = F_k(M_1, \dots, M_k), \quad M_k = G_k(C_1, \dots, C_k).$$

Describe recursive relations that determine F_k and G_k .

In probabilistic setup Proposition 2.1 takes the form

$$P(\{\omega \in \Omega \mid |X(\omega)| \geq a\}) \leq \frac{1}{a} E(|X|), \quad (2.1)$$

and is often called Markov or Chebyshev inequality. We shall often use a shorthand and abbreviate the l.h.s in (2.1) as $P\{|X(\omega)| \geq a\}$, etc.

2.3 Law of large numbers

Let (Ω, P) be a probability space and $X : \Omega \rightarrow \mathbb{R}$ a random variable. On the product probability space (Ω^N, P_N) we define

$$\mathcal{S}_N(\omega = (\omega_1, \dots, \omega_N)) = \sum_{k=1}^N X(\omega_k).$$

We shall refer to the following results as the *Law of large numbers (LLN)*.

Proposition 2.2 For any $\epsilon > 0$,

$$\lim_{N \rightarrow \infty} P_N \left\{ \left| \frac{\mathcal{S}_N(\omega)}{N} - E(X) \right| \geq \epsilon \right\} = 0.$$

Remark 2.1 An equivalent formulation of the LLN is that for any $\epsilon > 0$,

$$\lim_{N \rightarrow \infty} P_N \left\{ \left| \frac{\mathcal{S}_N(\omega)}{N} - E(X) \right| \leq \epsilon \right\} = 1.$$

Proof. Denote by E_N the expectation w.r.t. P_N . Define $X_k(\omega) = X(\omega_k)$ and note that $E_N(X_k) = E(X)$, $E_N(X_k^2) = E(X^2)$, $E_N(X_k X_j) = E(X)^2$ for $k \neq j$. Then

$$\begin{aligned} P_N \left\{ \left| \frac{\mathcal{S}_N(\omega)}{N} - E(X) \right| \geq \epsilon \right\} &= P_N \left\{ \left(\frac{\mathcal{S}_N(\omega)}{N} - E(X) \right)^2 \geq \epsilon^2 \right\} \\ &\leq \frac{1}{\epsilon^2} E_N \left(\left(\frac{\mathcal{S}_N(\omega)}{N} - E(X) \right)^2 \right) \\ &= \frac{1}{N^2 \epsilon^2} E_N \left(\sum_{k,j} (X_k - E(X_k))(X_j - E(X_j)) \right) \\ &= \frac{1}{N \epsilon^2} \text{Var}(X), \end{aligned}$$

and the statement follows. \square

2.4 Cumulant generating function

Let (Ω, P) be a probability space and $X : \Omega \rightarrow \mathbb{R}$ a random variable. In this section we shall study in some detail the properties of the cumulant generating function

$$C(\alpha) = \log E(e^{\alpha X}).$$

To avoid discussion of trivialities, until the end of this chapter we shall assume that X is not constant on $\text{supp } P$, i.e. that X assumes at least two distinct values on $\text{supp } P$. Obviously, the function $C(\alpha)$ is infinitely differentiable and

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} C'(\alpha) &= \max_{\omega} X(\omega), \\ \lim_{\alpha \rightarrow -\infty} C'(\alpha) &= \min_{\omega} X(\omega). \end{aligned} \tag{2.2}$$

Proposition 2.3 $C''(\alpha) > 0$ for all α . In particular, the function C is strictly convex.

Remark 2.2 By strictly convex we mean that C' is strictly increasing, i.e., that the graph of C does not have a flat piece.

Proof. Set

$$Q_{\alpha}(\omega) = \frac{e^{\alpha X(\omega)} P(\omega)}{\sum_{\omega} e^{\alpha X(\omega)} P(\omega)}, \tag{2.3}$$

and note that Q_{α} is a probability measure on Ω equivalent to P .

One easily verifies that

$$C'(\alpha) = E_{Q_{\alpha}}(X), \quad C''(\alpha) = \text{Var}_{Q_{\alpha}}(X).$$

The second identity yields the statement. □

Proposition 2.4 C extends to an analytic function in the strip

$$|\text{Im } \alpha| < \frac{\pi}{2 \max_{\omega} |X(\omega)|}. \tag{2.4}$$

Proof. Obviously, the function $\alpha \mapsto E(e^{\alpha X})$ is entire analytic. If $\alpha = a + ib$, then

$$E(e^{\alpha X}) = \sum_{\omega \in \Omega} e^{aX(\omega)} \cos(bX(\omega)) P(\omega) + i \sum_{\omega \in \Omega} e^{aX(\omega)} \sin(bX(\omega)) P(\omega).$$

If $|bX(\omega)| < \pi/2$ for all ω , then the real part of $E(e^{\alpha X})$ is strictly positive. It follows that the function

$$\text{Log } E(e^{\alpha X}),$$

where Log is the principal branch of complex logarithm, is analytic in the strip (2.4) and the statement follows. □

Remark 2.3 Let $\Omega = \{-1, 1\}$, $P(-1) = P(1) = 1/2$, $X(1) = 1$, $X(-1) = -1$. Then

$$C(\alpha) = \log \cosh \alpha.$$

Since $\cosh(\pi i/2) = 0$, we see that Proposition 2.4 is an optimal result.

2.5 Rate function

We continue with the framework of the previous section. The *rate function* of the random variable X is defined by

$$I(\theta) = \sup_{\alpha \in \mathbb{R}} (\alpha\theta - C(\alpha)), \quad \theta \in \mathbb{R}.$$

In the language of convex analysis, I is the Fenchel-Legendre transform of the cumulant generating function C . Obviously, $I(\theta) \geq 0$ for all θ . Set

$$m = \min_{\omega} X(\omega), \quad M = \max_{\omega} X(\omega),$$

and recall the relations (2.2). By the intermediate value theorem, for any θ in $]m, M[$ there exists unique $\alpha(\theta) \in \mathbb{R}$ such that

$$\theta = C'(\alpha(\theta)).$$

The function

$$\alpha(\theta) = (C')^{-1}(\theta)$$

is infinitely differentiable on $]m, M[$, strictly increasing on $]m, M[$, $\alpha(\theta) \downarrow -\infty$ iff $\theta \downarrow m$, and $\alpha(\theta) \uparrow \infty$ iff $\theta \uparrow M$.

Exercise 2.2. Prove that the function $]m, M[\ni \theta \mapsto \alpha(\theta)$ is real-analytic.
Hint: Apply the analytic implicit function theorem.

Proposition 2.5 (1) For $\theta \in]m, M[$,

$$I(\theta) = \alpha(\theta)\theta - C(\alpha(\theta)).$$

(2) The function I is infinitely differentiable on $]m, M[$.

(3) $I'(\theta) = \alpha(\theta)$. In particular, I' is strictly increasing on $]m, M[$ and

$$\lim_{\theta \downarrow m} I'(\theta) = -\infty, \quad \lim_{\theta \uparrow M} I'(\theta) = \infty.$$

(4) $I''(\theta) = 1/C''(\alpha(\theta))$.

(5) $I(\theta) = 0$ iff $\theta = E(X)$.

Proof. To prove (1), note that for $\theta \in]m, M[$ the function

$$\frac{d}{d\alpha}(\alpha\theta - C(\alpha)) = \theta - C'(\alpha)$$

vanishes at $\alpha(\theta)$, is positive for $\alpha < \alpha(\theta)$, and is negative for $\alpha > \alpha(\theta)$. Hence, the function $\alpha \mapsto \alpha\theta - C(\alpha)$ has the global maximum at $\alpha = \alpha(\theta)$ and Part (1) follows. Parts (2), (3) and (4) are obvious. To prove (5), note that if $I(\theta) = 0$ for some $\theta \in]m, M[$, then, since I is non-negative, we also have $0 = I'(\theta) = \alpha(\theta)$, and the relation $\theta = C'(\alpha(\theta)) = C'(0) = E(X)$ follows. On the other hand, if $\theta = E(X) = C'(0)$, then $\alpha(\theta) = 0$, and $I(\theta) = -C(0) = 0$. \square

Exercise 2.3. Prove that the function I is real-analytic in $]m, M[$.

Let

$$S_m = \{\omega \in \Omega \mid X(\omega) = m\}, \quad S_M = \{\omega \in \Omega \mid X(\omega) = M\}.$$

Proposition 2.6 (1) $I(\theta) = \infty$ for $\theta \notin [m, M]$.

(2)

$$I(m) = \lim_{\theta \downarrow m} I(\theta) = -\log P(S_m),$$

$$I(M) = \lim_{\theta \uparrow M} I(\theta) = -\log P(S_M).$$

Proof. (1) Suppose that $\theta > M$. Then

$$\frac{d}{d\alpha}(\alpha\theta - C(\alpha)) = \theta - C'(\alpha) > \theta - M.$$

Integrating this inequality over $[0, \alpha]$ we derive

$$\alpha\theta - C(\alpha) > (\theta - M)\alpha,$$

and so

$$I(\theta) = \sup_{\alpha \in \mathbb{R}} (\alpha\theta - C(\alpha)) = \infty.$$

The case $\theta < m$ is similar.

(2) We shall prove only the second formula, the proof of the first is similar. Since the function $\alpha M - C(\alpha)$ is increasing,

$$I(M) = \lim_{\alpha \rightarrow \infty} (\alpha M - C(\alpha)).$$

Since

$$C(\alpha) = \alpha M + \log P(S_M) + \log(1 + A(\alpha)), \quad (2.5)$$

where

$$A(\alpha) = \frac{1}{P(S_M)} \sum_{\omega \notin S_M} e^{\alpha(X(\omega) - M)} P(\omega),$$

we derive that $I(M) = -\log P(S_M)$.

Since $C'(\alpha(\theta)) = \theta$, Part (1) of Proposition 2.5 gives that

$$\lim_{\theta \uparrow M} I(\theta) = \lim_{\alpha \rightarrow \infty} (\alpha C'(\alpha) - C(\alpha)).$$

Write

$$C'(\alpha) = M \frac{1 + B(\alpha)}{1 + A(\alpha)}, \quad (2.6)$$

where

$$B(\alpha) = \frac{1}{MP(S_M)} \sum_{\omega \notin S_M} X(\omega) e^{\alpha(X(\omega) - M)} P(\omega).$$

The formulas (2.5) and (2.6) yield

$$\alpha C'(\alpha) - C(\alpha) = \alpha M \frac{B(\alpha) - A(\alpha)}{1 + A(\alpha)} - \log P(S_M) - \log(1 + A(\alpha)).$$

Since $A(\alpha)$ and $B(\alpha)$ converge to 0 as $\alpha \rightarrow \infty$,

$$\lim_{\theta \uparrow M} I(\theta) = \lim_{\alpha \rightarrow \infty} (\alpha C'(\alpha) - C(\alpha)) = -\log P(S_M).$$

□

Proposition 2.7

$$C(\alpha) = \sup_{\theta \in \mathbb{R}} (\theta\alpha - I(\theta)). \quad (2.7)$$

Proof. To avoid confusion, fix $\alpha = \alpha_0$. Below, $\alpha(\theta) = (C')^{-1}(\theta)$ is as in Proposition 2.5.

The supremum in (2.7) is achieved at θ_0 satisfying

$$\alpha_0 = I'(\theta_0).$$

Since $I'(\theta_0) = \alpha(\theta_0)$, we have $\alpha_0 = \alpha(\theta_0)$, and

$$I(\theta_0) = \theta_0 \alpha(\theta_0) - C(\alpha(\theta_0)) = \theta_0 \alpha_0 - C(\alpha_0).$$

Hence

$$\sup_{\theta \in \mathbb{R}} (\theta \alpha_0 - I(\theta)) = \alpha_0 \theta_0 - I(\theta_0) = C(\alpha_0).$$

□

Returning to the example of Remark 2.3, $m = -1$, $M = 1$, $C(\alpha) = \log \cosh \alpha$, and $C'(\alpha) = \tanh \alpha$. Hence, for $\theta \in]-1, 1[$,

$$\alpha(\theta) = \tanh^{-1}(\theta) = \frac{1}{2} \log \frac{1+\theta}{1-\theta}.$$

It follows that

$$I(\theta) = \theta \alpha(\theta) - C(\alpha(\theta)) = \frac{1}{2}(1+\theta) \log(1+\theta) + \frac{1}{2}(1-\theta) \log(1-\theta).$$

2.6 Cramér's theorem

This section is devoted to the proof of Cramér's theorem:

Theorem 2.8 For any interval $[a, b]$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \in [a, b] \right\} = - \inf_{\theta \in [a, b]} I(\theta).$$

Remark 2.4 To prove this result without loss of generality we may assume that $[a, b] \subset [m, M]$.

Remark 2.5 Note that

$$\inf_{\theta \in [a, b]} I(\theta) = \begin{cases} 0 & \text{if } \mathbb{E}(X) \in [a, b] \\ I(a) & \text{if } a > \mathbb{E}(X) \\ I(b) & \text{if } b < \mathbb{E}(X), \end{cases}$$

and that

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} = M \right\} &= \log P(S_M) = -I(M), \\ \lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} = m \right\} &= \log P(S_m) = -I(m). \end{aligned}$$

We start the proof with

Proposition 2.9 (1) For $\theta \geq \mathbb{E}(X)$,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq \theta \right\} \leq -I(\theta).$$

(2) For $\theta \leq \mathbb{E}(X)$,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \leq \theta \right\} \leq -I(\theta).$$

Remark 2.6 Note that if $\theta < \mathbb{E}(X)$, then by the LLN

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq \theta \right\} = 0.$$

Similarly, if $\theta > \mathbb{E}(X)$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \leq \theta \right\} = 0.$$

Proof. For $\alpha > 0$,

$$\begin{aligned} P_N \{ \mathcal{S}_N(\omega) \geq N\theta \} &= P_N \left\{ e^{\alpha \mathcal{S}_N(\omega)} \geq e^{\alpha N\theta} \right\} \\ &\leq e^{-\alpha N\theta} \mathbb{E}_N \left(e^{\alpha \mathcal{S}_N(\omega)} \right) \\ &= e^{-\alpha N\theta} \mathbb{E} \left(e^{\alpha X} \right)^N \\ &= e^{N(C(\alpha) - \alpha\theta)}. \end{aligned}$$

It follows that

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq \theta \right\} \leq \inf_{\alpha > 0} (C(\alpha) - \alpha\theta) = -\sup_{\alpha > 0} (\alpha\theta - C(\alpha)).$$

If $\theta \geq \mathbb{E}(X)$, then $\alpha\theta - C(\alpha) \leq 0$ for $\alpha \leq 0$ and

$$\sup_{\alpha > 0} (\alpha\theta - C(\alpha)) = \sup_{\alpha \in \mathbb{R}} (\alpha\theta - C(\alpha)) = I(\theta).$$

This yields Part (1). Part (2) follows by applying Part (1) to the random variable $-X$. \square

Exercise 2.4. Using Proposition 2.9 prove that for any $\epsilon > 0$ there exist $\gamma_\epsilon > 0$ and N_ϵ such that for $N \geq N_\epsilon$,

$$P_N \left\{ \left| \frac{\mathcal{S}_N(\omega)}{N} - E(X) \right| \geq \epsilon \right\} \leq e^{-\gamma_\epsilon N}.$$

Proposition 2.10 (1) For $\theta \geq \mathbb{E}(X)$,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq \theta \right\} \geq -I(\theta).$$

(2) For $\theta \leq \mathbb{E}(X)$,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \leq \theta \right\} \geq -I(\theta).$$

Remark 2.7 Note that Part (1) trivially holds if $\theta < \mathbb{E}(X)$. Similarly, Part (2) trivially holds if $\theta > \mathbb{E}(X)$.

Proof. We again need to prove only Part (1) (Part (2) follows by applying Part (1) to the random variable $-X$). If $\theta \geq M$, the statement is obvious and so without loss of generality we may assume that $\theta \in [\mathbb{E}(X), M[$. Fix such θ and choose s and $\epsilon > 0$ such that $\theta < s - \epsilon < s + \epsilon < M$.

Let Q_α be the probability measure introduced in the proof of Proposition 2.3, and let $Q_{\alpha, N}$ be the induced product probability measure on Ω^N . The measures P_N and $Q_{\alpha, N}$ are equivalent, and for $\omega \in \text{supp } P_N$

$$\Delta_{P_N | Q_{\alpha, N}}(\omega) = e^{-\alpha \mathcal{S}_N(\omega) + NC(\alpha)}.$$

We now consider the measure $Q_{\alpha,N}$ for $\alpha = \alpha(s)$. Recall that

$$C'(\alpha(s)) = s = \mathbb{E}_{Q_{\alpha(s)}}(X).$$

Set

$$T_N = \left\{ \omega \in \Omega^N \mid \frac{\mathcal{S}_N(\omega)}{N} \in [s - \epsilon, s + \epsilon] \right\},$$

and note that the LLN implies

$$\lim_{N \rightarrow \infty} Q_{\alpha(s),N}(T_N) = 1. \quad (2.8)$$

The estimates

$$\begin{aligned} P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq \theta \right\} &\geq P_N(T_N) = \int_{T_N} \Delta_{P_N|Q_{\alpha(s),N}} dQ_{\alpha(s),N} \\ &= \int_{T_N} e^{-\alpha(s)\mathcal{S}_N + NC(\alpha(s))} dQ_{\alpha(s),N} \\ &\geq e^{N(C(\alpha(s)) - s\alpha(s) - \epsilon|\alpha(s)|)} Q_{\alpha(s),N}(T_N) \end{aligned}$$

and (2.8) give

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq \theta \right\} \geq C(\alpha(s)) - s\alpha(s) - \epsilon|\alpha(s)| = -I(s) - \epsilon|\alpha(s)|.$$

The statement now follows by taking first $\epsilon \downarrow 0$ and then $s \downarrow \theta$. \square

Combining Propositions 2.9 and 2.10 we derive

Corollary 2.11 For $\theta \geq \mathbb{E}(X)$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq \theta \right\} = -I(\theta).$$

For $\theta \leq \mathbb{E}(X)$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \leq \theta \right\} = -I(\theta).$$

We are now ready to complete

Proof of Theorem 2.8. If $\mathbb{E}(X) \in]a, b[$ the result follows from the LLN. Suppose that $M > a \geq \mathbb{E}(X)$. Then

$$P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \in [a, b] \right\} = P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq a \right\} - P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} > b \right\}.$$

It follows from Corollary 2.11 that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \left[1 - \frac{P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} > b \right\}}{P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq a \right\}} \right] = 0, \quad (2.9)$$

and so

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \in [a, b] \right\} = \lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \frac{\mathcal{S}_N(\omega)}{N} \geq a \right\} = -I(a).$$

The case $m < b \leq \mathbb{E}(X)$ is similar. \square

Exercise 2.5. Write down the proof of (2.9) and of the case $m < b \leq \mathbb{E}(X)$.

Exercise 2.6. Consider the example introduced in Remark 2.3 and prove Cramér's theorem in this special case by using Stirling's formula and a direct combinatorial argument.
Hint: See Theorem 1.3.1 in [EII].

2.7 Notes and references

Although it is assumed that the student reader had no previous exposure to probability theory, a reading of additional material could be helpful at this point. Recommended textbooks are [Chu, RohSa, Ross].

For additional information and original references regarding Cramer's theorem we refer the reader to Chapter 2 of [DeZe]. Reader interested to learn more about theory of large deviations may consult classical references [dHoll, DeZe, EII], and the lecture notes of S.R.S. Varadhan <https://math.nyu.edu/~varadhan/LDP.html>.

It is possible to give a combinatorial proof of Theorem 2.8, as indicated in the Exercise 2.6. The advantage of the argument presented in this chapter is that it naturally extends to a proof of much more general results (such as the Gärtner-Ellis theorem) which will be discussed in the Part II of the lecture notes.

Chapter 3

Boltzmann–Gibbs–Shannon entropy

3.1 Preliminaries

Let Ω be a finite set, $|\Omega| = L$, and let $\mathcal{P}(\Omega)$ be the collection of all probability measures on Ω . $\mathcal{P}(\Omega)$ is naturally identified with the set

$$\mathcal{P}_L = \left\{ (p_1, \dots, p_L) \mid p_k \geq 0, \sum_{k=1}^L p_k = 1 \right\} \quad (3.1)$$

(the identification map is $P \mapsto (P(\omega_1), \dots, P(\omega_L))$). We shall often use this identification without further notice. A convenient metric on $\mathcal{P}(\Omega)$ is the variational distance

$$d_V(P, Q) = \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|. \quad (3.2)$$

We denote by $\mathcal{P}_f(\Omega)$ the set of all faithful probability measures on $\mathcal{P}(\Omega)$ (recall that $P \in \mathcal{P}_f(\Omega)$ iff $P(\omega) > 0$ for all $\omega \in \Omega$). $\mathcal{P}_f(\Omega)$ coincides with the interior of $\mathcal{P}(\Omega)$ and is identified with

$$\mathcal{P}_{L,f} = \left\{ (p_1, \dots, p_L) \mid p_k > 0, \sum_{k=1}^L p_k = 1 \right\}.$$

Note that $\mathcal{P}(\Omega)$ and $\mathcal{P}_f(\Omega)$ are convex sets.

The probability measure P is called *pure* if $P(\omega) = 1$ for some $\omega \in \Omega$. The *chaotic* probability measure is $P_{\text{ch}}(\omega) = 1/L, \omega \in \Omega$.

We shall often make use of Jensen's inequality. This inequality states that if $f : [a, b] \rightarrow \mathbb{R}$ is concave, then for $x_k \in [a, b], k = 1, \dots, n$, and $(p_1, \dots, p_n) \in \mathcal{P}_{n,f}$ we have

$$\sum_{k=1}^n p_k f(x_k) \leq f\left(\sum_{k=1}^n p_k x_k\right). \quad (3.3)$$

Moreover, if f is strictly concave the inequality is strict unless $x_1 = \dots = x_n$. A similar statement holds for convex functions.

Exercise 3.1. Prove Jensen's inequality.

3.2 Definition and basic properties

The *entropy function* (sometimes called the *information function*) of $P \in \mathcal{P}(\Omega)$ is¹

$$S_P(\omega) = -c \log P(\omega), \quad (3.4)$$

where $c > 0$ is a constant that does not depend on P or Ω , and $-\log 0 = \infty$. The function S_P takes values in $[0, \infty]$. The *Boltzmann–Gibbs–Shannon entropy* (in the sequel we will often call it just *entropy*) of P is

$$S(P) = \int_{\Omega} S_P dP = -c \sum_{\omega \in \Omega} P(\omega) \log P(\omega). \quad (3.5)$$

The value of the constant c is linked to the choice of units (or equivalently, the base of logarithm). The natural choice in the information theory is $c = 1/\log 2$ (that is, the logarithm is taken in the base 2). The value of c plays no role in these lecture notes, and from now on we set $c = 1$ and call

$$S(P) = - \sum_{\omega \in \Omega} P(\omega) \log P(\omega)$$

the Boltzmann–Gibbs–Shannon entropy of P . We note, however, that the constant c will reappear in the axiomatic characterizations of entropy given in Theorems 3.4 and 3.5.

The basic properties of entropy are:

Proposition 3.1 (1) $S(P) \geq 0$ and $S(P) = 0$ iff P is pure.

(2) $S(P) \leq \log L$ and $S(P) = \log L$ iff $P = P_{\text{ch}}$.

(3) The map $\mathcal{P}(\Omega) \ni P \mapsto S(P)$ is continuous and concave, that is, if p_k 's are as in (3.3) and $P_k \in \mathcal{P}(\Omega)$, then

$$p_1 S(P_1) + \cdots + p_n S(P_n) \leq S(p_1 P_1 + \cdots + p_n P_n), \quad (3.6)$$

with equality iff $P_1 = \cdots = P_n$.

(4) The concavity inequality (3.6) has the following "almost convexity" counterpart:

$$S(p_1 P_1 + \cdots + p_n P_n) \leq p_1 S(P_1) + \cdots + p_n S(P_n) + S(p_1, \dots, p_n),$$

with equality iff $\text{supp } P_k \cap \text{supp } P_j = \emptyset$ for $k \neq j$.

Proof. Parts (1) and (3) follow from the obvious fact that the function $[0, 1] \ni x \mapsto -x \log x$ is continuous, strictly concave, non-negative, and vanishing iff $x = 0$ or $x = 1$. Part (2) follows from Jensen's inequality.

¹ Regarding the choice of logarithm, in the introduction of [Sha] Shannon comments: "(1) It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc. (2) It is nearer to our intuitive feeling as to the proper measure. This is closely related to (1) since we intuitively measure entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information. (3) It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities."

Part (4) follows from the monotonicity of $\log x$:

$$\begin{aligned}
S(p_1 P_1 + \cdots + p_n P_n) &= \sum_{\omega \in \Omega} \sum_{k=1}^n -p_k P_k(\omega) \log \left(\sum_{j=1}^n p_j P_j(\omega) \right) \\
&\leq \sum_{\omega \in \Omega} \sum_{k=1}^n -p_k P_k(\omega) \log (p_k P_k(\omega)) \\
&= \sum_{k=1}^n p_k \left(\sum_{\omega \in \Omega} -P_k(\omega) \log P_k(\omega) \right) - \sum_{k=1}^n \left(\sum_{\omega \in \Omega} P_k(\omega) \right) p_k \log p_k \\
&= \sum_{k=1}^n p_k S(P_k) + S(p_1, \dots, p_n).
\end{aligned}$$

The equality holds if for all ω and $k \neq j$, $p_k P_k(\omega) > 0 \Rightarrow p_j P_j(\omega) = 0$, which is equivalent to $\text{supp } P_k \cap \text{supp } P_j = \emptyset$ for all $k \neq j$. \square

Suppose that $\Omega = \Omega_l \times \Omega_r$ and let $P_{l/r}$ be the marginals of $P \in \mathcal{P}(\Omega)$. For a given $\omega \in \text{supp } P_l$ the conditional probability measure $P_{r|l}^\omega$ on Ω_r is defined by

$$P_{r|l}^\omega(\omega') = \frac{P(\omega, \omega')}{P_l(\omega)}.$$

Note that

$$\sum_{\omega \in \text{supp } P_l} P_l(\omega) P_{r|l}^\omega = P_r.$$

Proposition 3.2 (1)

$$S(P) = S(P_l) + \sum_{\omega \in \Omega_l} P_l(\omega) S(P_{r|l}^\omega).$$

(2) *The entropy is strictly sub-additive:*

$$S(P) \leq S(P_l) + S(P_r),$$

with the equality iff $P = P_l \otimes P_r$.

Proof. Part (1) and the identity $S(P_l \otimes P_r) = S(P_l) + S(P_r)$ follow by direct computation. To prove (2), note that Part (3) of Proposition 3.1 gives

$$\sum_{\omega \in \text{supp } P_l} P_l(\omega) S(P_{r|l}^\omega) \leq S \left(\sum_{\omega \in \text{supp } P_l} P_l(\omega) P_{r|l}^\omega \right) = S(P_r),$$

and so it follows from Part (1) that $S(P) \leq S(P_l) + S(P_r)$ with the equality iff all the probability measures $P_{r|l}^\omega$, $\omega \in \text{supp } P_l$, are equal. Thus, if the equality holds, then for all $(\omega, \omega') \in \Omega_l \times \Omega_r$, $P(\omega, \omega') = C(\omega') P_l(\omega)$. Summing over ω 's gives that $P = P_l \otimes P_r$. \square

Exercise 3.2. The Hartley entropy of $P \in \mathcal{P}(\Omega)$ is defined by

$$S_H(P) = \log |\{\omega \mid P(\omega) > 0\}|.$$

1. Prove that the Hartley entropy is also strictly sub-additive: $S_H(P) \leq S_H(P_l) + S_H(P_r)$, with the equality iff $P = P_l \otimes P_r$.
2. Show that the map $P \mapsto S_H(P)$ is not continuous if $L \geq 2$.

3.3 Covering exponents and source coding

To gain further insight into the concept of entropy, assume that P is faithful and consider the product probability space (Ω^N, P_N) . For given $\epsilon > 0$ let

$$\begin{aligned} T_{N,\epsilon} &= \left\{ \omega = (\omega_1, \dots, \omega_N) \in \Omega^N \mid \left| \frac{S_P(\omega_1) + \dots + S_P(\omega_N)}{N} - S(P) \right| < \epsilon \right\} \\ &= \left\{ \omega \in \Omega^N \mid \left| -\frac{\log P_N(\omega)}{N} - S(P) \right| < \epsilon \right\} \\ &= \left\{ \omega \in \Omega^N \mid e^{-N(S(P)+\epsilon)} < P_N(\omega) < e^{-N(S(P)-\epsilon)} \right\}. \end{aligned}$$

The LLN gives

$$\lim_{N \rightarrow \infty} P_N(T_{N,\epsilon}) = 1.$$

We also have the following obvious bounds on the cardinality of $T_{N,\epsilon}$:

$$P_N(T_{N,\epsilon})e^{N(S(P)-\epsilon)} < |T_{N,\epsilon}| < e^{N(S(P)+\epsilon)}.$$

It follows that

$$S(P) - S(P_{\text{ch}}) - \epsilon \leq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \frac{|T_{N,\epsilon}|}{|\Omega|^N} \leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log \frac{|T_{N,\epsilon}|}{|\Omega|^N} \leq S(P) - S(P_{\text{ch}}) + \epsilon.$$

This estimate implies that if $P \neq P_{\text{ch}}$, then, as $N \rightarrow \infty$, the measure P_N is "concentrated" and "equipartitioned" on the set $T_{N,\epsilon}$ whose size is "exponentially small" with respect to the size of Ω^N .

We continue with the analysis of the above concepts. Let $\gamma \in]0, 1[$ be fixed. The (N, γ) covering exponent is defined by

$$c_N(\gamma) = \min \{ |A| \mid A \subset \Omega^N, P_N(A) \geq \gamma \}. \quad (3.7)$$

One can find $c_N(\gamma)$ according to the following algorithm:

- (a) List the events $\omega = (\omega_1, \dots, \omega_N)$ in order of decreasing probabilities.
- (b) Count the events until the first time the total probability is $\geq \gamma$.

Proposition 3.3 For all $\gamma \in]0, 1[$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log c_N(\gamma) = S(P).$$

Proof. Fix $\epsilon > 0$ and recall the definition of $T_{N,\epsilon}$. For N large enough, $P_N(T_{N,\epsilon}) \geq \gamma$, and so for such N 's,

$$c_N(\gamma) \leq |T_{N,\epsilon}| \leq e^{N(S(P)+\epsilon)}.$$

It follows that

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log c_N(\gamma) \leq S(P).$$

To prove the lower bound, let $A_{N,\gamma}$ be a set for which the minimum in (3.7) is achieved. Let $\epsilon > 0$. Note that

$$\liminf_{N \rightarrow \infty} P_N(T_{N,\epsilon} \cap A_{N,\gamma}) \geq \gamma. \quad (3.8)$$

Since for $P_N(\omega) \leq e^{-N(S(P)-\epsilon)}$ for $\omega \in T_{N,\epsilon}$,

$$P_N(T_{N,\epsilon} \cap A_{N,\gamma}) = \sum_{\omega \in T_{N,\epsilon} \cap A_{N,\gamma}} P_N(\omega) \leq e^{-N(S(P)-\epsilon)} |T_{N,\epsilon} \cap A_{N,\gamma}|.$$

Hence,

$$|A_{N,\gamma}| \geq e^{N(S(P)-\epsilon)} P_N(T_{N,\epsilon} \cap A_{N,\gamma}),$$

and it follows from (3.8) that

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log c_N(\gamma) \geq S(P) - \epsilon.$$

Since $\epsilon > 0$ is arbitrary,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log c_N(\gamma) \geq S(P),$$

and the proposition is proven. \square

We finish this section with a discussion of Shannon's source coding theorem. Given a pair of positive integers N, M , the *encoder* is a map

$$F_N : \Omega^N \rightarrow \{0, 1\}^M.$$

The *decoder* is a map

$$G_N : \{0, 1\}^M \rightarrow \Omega^N.$$

The error probability of the coding pair (F_N, G_N) is

$$P_N \{G_N \circ F_N(\omega) \neq \omega\}.$$

If this probability is less than some prescribed $1 > \epsilon > 0$, we shall say that the coding pair is ϵ -good. Note that to any ϵ -good coding pair one can associate the set

$$A = \{\omega \mid G_N \circ F_N(\omega) = \omega\}$$

which satisfies

$$P_N(A) \geq 1 - \epsilon, \quad |A| \leq 2^M. \quad (3.9)$$

On the other hand, if $A \subset \Omega^N$ satisfies (3.9), we can associate to it an ϵ -good pair (F_N, G_N) by setting F_N to be one-one on A (and arbitrary otherwise), and $G_N = F_N^{-1}$ on $F_N(A)$ (and arbitrary otherwise).

In the source coding we wish to find M that minimizes the compression coefficients M/N subject to an allowed ϵ -error probability. Clearly, the optimal M is

$$M_N = \lceil \log_2 \min \{|A| \mid A \subset \Omega^N, P_N(A) \geq 1 - \epsilon\} \rceil,$$

where $\lceil \cdot \rceil$ denotes the greatest integer part. Shannon's source coding theorem now follows from Proposition 3.3: the limiting optimal compression coefficient is

$$\lim_{N \rightarrow \infty} \frac{M_N}{N} = \frac{1}{\log 2} S(P).$$

3.4 Why is the entropy natural?

Set $\mathcal{P} = \cup_{\Omega} \mathcal{P}(\Omega)$. In this section we shall consider functions $\mathfrak{S} : \mathcal{P} \rightarrow \mathbb{R}$ that satisfy properties that correspond intuitively to those of *entropy* as a measure of *randomness* of probability measures. The goal is to show that those intuitive natural demands uniquely specify \mathfrak{S} up to a choice of units, that is, that for some $c > 0$ and all $P \in \mathcal{P}$, $\mathfrak{S}(P) = cS(P)$.

We describe first three basic properties that any candidate for \mathfrak{S} should satisfy. The first is the positivity and non-triviality requirement: $\mathfrak{S}(P) \geq 0$ and this inequality is strict for at least one $P \in \mathcal{P}$. The second is that if $|\Omega_1| = |\Omega_2|$ and $\theta : \Omega_1 \rightarrow \Omega_2$ is a bijection, then for any $P \in \mathcal{P}(\Omega_1)$, $\mathfrak{S}(P) = \mathfrak{S}(P \circ \theta)$. In other words, the entropy of P should not depend on the labeling of the elementary events. This second requirement gives that \mathfrak{S} is completely specified by its restriction $\mathfrak{S} : \cup_{L \geq 1} \mathcal{P}_L \rightarrow [0, \infty[$ which satisfies

$$\mathfrak{S}(p_1, \dots, p_L) = \mathfrak{S}(p_{\pi(1)}, \dots, p_{\pi(L)}) \quad (3.10)$$

for any $L \geq 1$ and any permutation π of $\{1, \dots, L\}$. In the proof of Theorem 3.5 we shall also assume that

$$\mathfrak{S}(p_1, \dots, p_L, 0) = \mathfrak{S}(p_1, \dots, p_L) \quad (3.11)$$

for all $L \geq 1$ and $(p_1, \dots, p_L) \in \mathcal{P}_L$. In the literature, the common sense assumption (3.11) is sometimes called *expansibility*.

Throughout this section we shall assume that the above three properties hold. We remark that the assumptions of Theorem 3.5 actually imply the positivity and non-triviality requirement.

3.4.1 Split additivity characterization

If Ω_1, Ω_2 are two disjoint sets, we denote by $\Omega_1 \oplus \Omega_2$ their union (the symbol \oplus is used to emphasize the fact that the sets are disjoint). If μ_1 is a measure on Ω_1 and μ_2 is a measure on Ω_2 , then $\mu = \mu_1 \oplus \mu_2$ is a measure on $\Omega_1 \oplus \Omega_2$ defined by $\mu(\omega) = \mu_1(\omega)$ if $\omega \in \Omega_1$ and $\mu(\omega) = \mu_2(\omega)$ if $\omega \in \Omega_2$. Two measurable spaces $(\Omega_1, \mu_1), (\Omega_2, \mu_2)$ are called disjoint if the sets Ω_1, Ω_2 , are disjoint.

The split additivity characterization has its roots in the identity

$$S(p_1 P_1 + \dots + p_n P_n) = p_1 S(P_1) + \dots + p_n S(P_n) + S(p_1, \dots, p_n)$$

which holds if $\text{supp} P_k \cap \text{supp} P_j = \emptyset$ for $k \neq j$.

Theorem 3.4 *Let $\mathfrak{S} : \mathcal{P} \rightarrow [0, \infty[$ be a function such that:*

- (a) \mathfrak{S} is continuous on \mathcal{P}_2 .
- (b) For any finite collection of disjoint probability spaces (Ω_j, P_j) , $j = 1, \dots, n$, and any $(p_1, \dots, p_n) \in \mathcal{P}_n$,

$$\mathfrak{S} \left(\bigoplus_{k=1}^n p_k P_k \right) = \sum_{k=1}^n p_k \mathfrak{S}(P_k) + \mathfrak{S}(p_1, \dots, p_n). \quad (3.12)$$

Then there exists $c > 0$ such that for all $P \in \mathcal{P}$,

$$\mathfrak{S}(P) = cS(P). \quad (3.13)$$

Remark 3.1 If the positivity and non-triviality assumptions are dropped, then the proof gives that (3.13) holds for some $c \in \mathbb{R}$.

Remark 3.2 The split-additivity property (3.12) is sometimes called the chain rule for entropy. It can be verbalized as follows: if the initial choices $(1, \dots, n)$, realized with probabilities (p_1, \dots, p_n) , are split into sub-choices described by probability spaces (Ω_k, P_k) , $k = 1, \dots, n$, then the new entropy is the sum of the initial entropy and the entropies of sub-choices weighted by their probabilities.

Proof. In what follows, $\bar{P}_n \in \mathcal{P}_n$ denotes the chaotic probability measure

$$\bar{P}_n = \left(\frac{1}{n}, \dots, \frac{1}{n} \right),$$

and

$$f(n) = \mathfrak{S}(\bar{P}_n) = \mathfrak{S} \left(\frac{1}{n}, \dots, \frac{1}{n} \right).$$

We split the argument into six steps.

Step 1. $\mathfrak{S}(1) = \mathfrak{S}(0, 1) = 0$.

Suppose that $|\Omega| = 2$ and let $P = (q_1, q_2) \in \mathcal{P}_2$. Writing $\Omega = \Omega_1 \oplus \Omega_2$ where $|\Omega_1| = |\Omega_2| = 1$ and taking $P_1 = (1)$, $P_2 = (1)$, $p_1 = q_1$, $p_2 = q_2$, we get $\mathfrak{S}(q_1, q_2) = \mathfrak{S}(1) + \mathfrak{S}(q_1, q_2)$, and so $\mathfrak{S}(1) = 0$. Similarly, the relations

$$\begin{aligned} \mathfrak{S}(0, q_1, q_2) &= q_1 \mathfrak{S}(0, 1) + q_2 \mathfrak{S}(1) + \mathfrak{S}(q_1, q_2), \\ \mathfrak{S}(0, q_1, q_2) &= 0 \cdot \mathfrak{S}(1) + 1 \cdot \mathfrak{S}(q_1, q_2) + \mathfrak{S}(0, 1), \end{aligned}$$

yield that $\mathfrak{S}(0, 1) = q_1 \mathfrak{S}(0, 1)$ for all q_1 , and so $\mathfrak{S}(0, 1) = 0$.

Step 2. $f(nm) = f(n) + f(m)$.

Take $\Omega = \Omega_1 \oplus \dots \oplus \Omega_m$ with $|\Omega_k| = n$ for all $1 \leq k \leq m$, and set $P_k = \bar{P}_n$, $p_k = 1/m$. It then follows from (3.12) that $f(nm) = m \cdot \frac{1}{m} f(n) + f(m) = f(n) + f(m)$.

Step 3. $\lim_{n \rightarrow \infty} (f(n) - f(n-1)) = 0$.

In the proof of this step we shall make use of the following elementary result regarding convergence of the Cesàro means: if $(a_n)_{n \geq 1}$ is a converging sequence of real numbers and $\lim_{n \rightarrow \infty} a_n = a$, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n a_k = a.$$

As an exercise, prove this result.

Set $d_n = f(n) - f(n-1)$, $\delta_n = \mathfrak{S}(\frac{1}{n}, 1 - \frac{1}{n})$. Since $f(1) = \mathfrak{S}(1) = 0$,

$$f(n) = d_n + \dots + d_2.$$

The relation (3.12) gives

$$f(n) = \left(1 - \frac{1}{n}\right) f(n-1) + \delta_n,$$

and so

$$n\delta_n = nd_n + f(n-1).$$

It follows that

$$\sum_{k=2}^n k\delta_k = nf(n) = n(d_n + f(n-1)) = n(n\delta_n - (n-1)d_n),$$

which yields

$$d_n = \delta_n - \frac{1}{n(n-1)} \sum_{k=2}^{n-1} k\delta_k.$$

By Step 1, $\lim_{n \rightarrow \infty} \delta_n = 0$. Obviously,

$$0 \leq \frac{1}{n(n-1)} \sum_{k=2}^{n-1} k\delta_k \leq \frac{1}{n} \sum_{k=2}^{n-1} \delta_k,$$

and we derive

$$\lim_{n \rightarrow \infty} \frac{1}{n(n-1)} \sum_{k=2}^{n-1} k \delta_k = 0.$$

It follows that $\lim_{n \rightarrow \infty} d_n = 0$.

Step 4. There is a constant c such that $f(n) = c \log n$ for all n .

By Step 2, for any $k \geq 1$,

$$\frac{f(n^k)}{\log n^k} = \frac{n}{\log n}.$$

Hence, to prove the statement it suffices to show that the limit

$$c = \lim_{n \rightarrow \infty} \frac{f(n)}{\log n}$$

exists. To prove that, we will show that $g(n)$ defined by

$$g(n) = f(n) - \frac{f(2)}{\log 2} \log n \quad (3.14)$$

satisfies

$$\lim_{n \rightarrow \infty} \frac{g(n)}{\log n} = 0.$$

The choice of integer 2 in (3.14) is irrelevant, and the argument works with 2 replaced by any integer $m \geq 2$.

Obviously, $g(nm) = g(n) + g(m)$ and $g(1) = g(2) = 0$. Set $\xi_m = g(m) - g(m-1)$ if n is odd, $\xi_m = 0$ if m is even. By Step 3, $\lim_{m \rightarrow \infty} \xi_m = 0$. Let $n > 1$ be given. Write $n = 2n_1 + r_1$, where $r_1 = 0$ or $r_1 = 1$. Then

$$g(n) = \zeta_n + g(2n_1) = \zeta_n + g(n_1),$$

where we used that $g(2) = 0$. If $n_1 > 1$, write again $n_1 = 2n_2 + r_2$, where $r_2 = 0$ or $r_2 = 1$, so that

$$g(n_1) = \zeta_{n_1} + g(n_2).$$

This procedure terminates after k_0 steps, that is, when we reach $n_{k_0} = 1$. Obviously,

$$k_0 \leq \frac{\log n}{\log 2}, \quad g(n) = \sum_{k=0}^{k_0-1} \zeta_{n_k},$$

where we set $n_0 = n$. Let $\epsilon > 0$ and m_ϵ be such that for $m \geq m_\epsilon$ we have $|\xi_m| < \epsilon / \log 2$. Then

$$\frac{|g(n)|}{\log n} \leq \frac{1}{\log n} \left(\sum_{m \leq m_\epsilon} |\xi_m| \right) + \epsilon \frac{k_0 \log 2}{\log n} \leq \frac{1}{\log n} \left(\sum_{m \leq m_\epsilon} |\xi_m| \right) + \epsilon.$$

It follows that

$$\limsup_{n \rightarrow \infty} \frac{|g(n)|}{\log n} \leq \epsilon.$$

Since $\epsilon > 0$ is arbitrary, the proof is complete.

Step 5. If c is as in Step 4, then

$$\mathfrak{S}(q_1, q_2) = cS(q_1, q_2).$$

Let $\Omega = \Omega_1 \oplus \Omega_2$ with $|\Omega_1| = m$, $|\Omega_2| = m - n$. Applying (3.12) to $P_1 = \bar{P}_n$, $P_2 = \bar{P}_{n-m}$, $p_1 = \frac{n}{m}$, $p_2 = \frac{m-n}{m}$, we derive

$$f(m) = \frac{n}{m} f(n) + \frac{m-n}{m} f(m-n) + \mathfrak{S} \left(\frac{n}{m}, \frac{m-n}{m} \right).$$

Step 4 gives that

$$\mathfrak{S}\left(\frac{n}{m}, \frac{m-n}{m}\right) = cS\left(\frac{n}{m}, \frac{m-n}{m}\right).$$

Since this relation holds for any $m < n$, the continuity of \mathfrak{S} and S on \mathcal{P}_2 yields the statement.

Step 6. We now complete the proof by induction on $|\Omega|$. Suppose that $\mathfrak{S}(P) = cS(P)$ holds for all $P \in \mathcal{P}(\Omega)$ with $|\Omega| = n - 1$, where c is as in Step 4. Let $P = (p_1, \dots, p_n)$ be a probability measure on $\Omega = \Omega_{n-1} \oplus \Omega_1$, where $|\Omega_{n-1}| = n - 1$, $|\Omega_1| = 1$. Without loss of generality we may assume that $q_n < 1$. Applying (3.12) with

$$P_1 = \left(\frac{q_1}{1-q_n}, \dots, \frac{q_{n-1}}{1-q_n}\right),$$

$P_2 = (1), p_1 = 1 - q_n, p_2 = q_n$, we derive

$$\mathfrak{S}(P) = cS(P_1) + cS(p_1, p_2) = cS(P).$$

This completes the proof. The non-triviality assumption yields that $c > 0$. \square

3.4.2 Sub-additivity characterization

The sub-additivity of entropy described in Proposition 3.2 is certainly a very intuitive property. If the entropy quantifies randomness of a probability measure P , or equivalently, the amount of information gained by an outcome of a probabilistic experiment described by P , than the product of marginals $P_l \otimes P_r$ is certainly more random than $P \in \mathcal{P}(\Omega_l \times \Omega_r)$. The Boltzmann–Gibbs–Shannon entropy S and the Hartley entropy S_H introduced in Exercise 3.2 are strictly sub-additive, and so is any linear combination

$$\mathfrak{S} = cS + CS_H, \quad (3.15)$$

where $c \geq 0$, $C \geq 0$, and at least one of these constants is strictly positive. It is a remarkable fact that the strict sub-additivity requirement together with the obvious assumption (3.11) selects (3.15) as the only possible choices for entropy. We also note the strict sub-additivity assumption selects the sign of the constants in (3.15), and that here we can omit the assumption (a) of Theorem 3.4.

Theorem 3.5 *Let $\mathfrak{S} : \mathcal{P} \rightarrow [0, \infty[$ be a strictly sub-additive map, namely if $\Omega = \Omega_l \times \Omega_r$ and $P \in \mathcal{P}(\Omega)$, then*

$$\mathfrak{S}(P) \leq \mathfrak{S}(P_l) + \mathfrak{S}(P_r)$$

with equality iff $P = P_l \otimes P_r$. Then there are constants $c \geq 0, C \geq 0, c + C > 0$, such that for all $P \in \mathcal{P}$,

$$\mathfrak{S}(P) = cS(P) + CS_H(P). \quad (3.16)$$

If in addition \mathfrak{S} is continuous on \mathcal{P}_2 , then $C = 0$ and $\mathfrak{S} = cS$ for some $c > 0$.

Proof. We denote by \mathfrak{S}_n the restriction of \mathfrak{S} to \mathcal{P}_n . Note that the sub-additivity implies that

$$\begin{aligned} \mathfrak{S}_{2n}(p_{11}, p_{12}, \dots, p_{n1}, p_{n2}) &\leq \mathfrak{S}_2(p_{11} + \dots + p_{n1}, p_{12} + \dots + p_{n2}) \\ &\quad + \mathfrak{S}_n(p_{11} + p_{12}, \dots, p_{n1} + p_{n2}). \end{aligned} \quad (3.17)$$

For $x \in [0, 1]$ we set $\bar{x} = 1 - x$. The function

$$F(x) = \mathfrak{S}_2(\bar{x}, x) \quad (3.18)$$

will play an important role in the proof. It follows from (3.10) that $F(x) = F(\bar{x})$. By taking $P_l = P_r = (1, 0)$, we see that

$$2F(0) = \mathfrak{S}(P_l) + \mathfrak{S}(P_r) = \mathfrak{S}(P_l \otimes P_r) = \mathfrak{S}(1, 0, 0, 0) = \mathfrak{S}(1, 0) = F(0),$$

and so $F(0) = 0$.

We split the proof into eight steps.

Step 1. For all $q, r \in [0, 1]$ and $(p, p_3, \dots, p_n) \in \mathcal{P}_{n-1}$, $n \geq 3$, one has

$$\begin{aligned} \mathfrak{S}_2(\bar{q}, q) - \mathfrak{S}_2(\bar{p}\bar{q} + p\bar{r}, \bar{p}q + pr) &\leq \mathfrak{S}_n(p\bar{q}, pq, p_3, \dots, p_n) - \mathfrak{S}_n(p\bar{r}, pr, p_3, \dots, p_n) \\ &\leq \mathfrak{S}_2(\bar{p}\bar{r} + p\bar{q}, \bar{p}r + rq) - \mathfrak{S}_2(\bar{r}, r). \end{aligned} \quad (3.19)$$

By interchanging q and r , it suffices to prove the first inequality in (3.19). We have

$$\begin{aligned} \mathfrak{S}_2(\bar{q}, q) + \mathfrak{S}_n(p\bar{r}, pr, p_3, \dots, p_n) &= \mathfrak{S}_{2n}(\bar{q}p\bar{r}, qp\bar{r}, \bar{q}pr, qpr, \bar{q}p_3, qp_3, \dots, \bar{q}p_n, qp_n) \\ &= \mathfrak{S}_{2n}(\bar{q}p\bar{r}, \bar{q}pr, qp\bar{r}, qpr, \bar{q}p_3, qp_3, \dots, \bar{q}p_n, qp_n) \\ &\leq \mathfrak{S}_2(\bar{q}p\bar{r} + qp\bar{r} + \bar{q}(p_3 + \dots + p_n), \bar{q}pr + qpr + q(p_3 + \dots + p_n)) \\ &\quad + \mathfrak{S}_n(\bar{q}p\bar{r} + \bar{q}pr, qp\bar{r} + qpr, \bar{q}p_3 + qp_3, \dots, \bar{q}p_n + qp_n) \\ &= \mathfrak{S}_2(\bar{p}\bar{q} + p\bar{r}, \bar{p}q + pr) + S_n(p\bar{r}, pr, p_3, \dots, p_n). \end{aligned}$$

The first equality follows from (3.10) and the first inequality from (3.17). The final equality is elementary (we used that $p + p_3 + \dots + p_n = 1$).

Step 2. The function F , defined by (3.18), is increasing on $[0, 1/2]$, decreasing on $[1/2, 1]$, and is continuous and concave on $]0, 1[$. Moreover, for $q \in]0, 1[$ the left and right derivatives

$$D^+F(q) = \lim_{h \downarrow 0} \frac{F(q+h) - F(q)}{h}, \quad D^-F(q) = \lim_{h \uparrow 0} \frac{F(q+h) - F(q)}{h}$$

exist, are finite, and $D^+F(q) \geq D^-F(q)$.

We first establish the monotonicity statement. Note that the inequality of Step 1

$$\mathfrak{S}_2(\bar{q}, q) - \mathfrak{S}_2(\bar{p}\bar{q} + p\bar{r}, \bar{p}q + pr) \leq \mathfrak{S}_2(\bar{p}\bar{r} + p\bar{q}, \bar{p}r + rq) - \mathfrak{S}_2(\bar{r}, r) \quad (3.20)$$

with $r = \bar{q}$ gives

$$\begin{aligned} 2\mathfrak{S}_2(\bar{q}, q) &\leq \mathfrak{S}_2((1-p)(1-q) + pq, (1-p)q + p(1-q)) \\ &\quad + \mathfrak{S}_2((1-p)q + p(1-q), (1-p)(1-q) + pq), \end{aligned}$$

or equivalently, that

$$F(q) \leq F((1-p)q + p(1-q)). \quad (3.21)$$

Fix $q \in [0, 1/2]$ and note that $[0, 1] \ni p \mapsto (1-p)q + p(1-q)$ is the parametrization of the interval $[q, 1-q]$. Since $F(q) = F(1-q)$, we derive that $F(q) \leq F(x)$ for $x \in [q, 1/2]$, and that $F(x) \geq F(1-q)$ for $x \in [1/2, q]$. Thus, F is increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$. In particular, for all $x \in [0, 1]$,

$$F(1/2) \geq F(x) \geq 0, \quad (3.22)$$

where we used that $F(0) = F(1) = 0$.

We now turn to the continuity and concavity, starting with continuity first. The inequality (3.20) with $p = 1/2$ gives that for any $q, r \in [0, 1]$,

$$\frac{1}{2}F(q) + \frac{1}{2}F(r) \leq F\left(\frac{1}{2}q + \frac{1}{2}r\right). \quad (3.23)$$

Fix now $q \in]0, 1[$, set $\lambda_n = 2^{-n}$ and, starting with large enough n so that $q \pm \lambda_n \in [0, 1]$, define

$$\Delta_n^+(q) = \frac{F(q + \lambda_n) - F(q)}{\lambda_n}, \quad \Delta_n^-(q) = \frac{F(q - \lambda_n) - F(q)}{-\lambda_n}.$$

It follows from (3.23) that the sequence $\Delta_n^+(q)$ is increasing, that the sequence $\Delta_n^-(q)$ is decreasing, and that $\Delta_n^+(q) \leq \Delta_n^-(q)$ (write down the details!). Hence, the limits

$$\lim_{n \rightarrow \infty} \Delta_n^+(q), \quad \lim_{n \rightarrow \infty} \Delta_n^-(q)$$

exists, are finite, and

$$\lim_{n \rightarrow \infty} F(q \pm \lambda_n) = F(q). \quad (3.24)$$

The established monotonicity properties of F yield that the limits $\lim_{h \downarrow 0} F(q+h)$ and $\lim_{h \uparrow 0} F(q+h)$ exist. Combining this observation with (3.24), we derive that

$$\lim_{h \rightarrow 0} F(q+h) = F(q),$$

and so F is continuous on $]0, 1[$. We now prove the concavity. Replacing r with $(q+r)/2$ in (3.23), we get that

$$\lambda F(q) + (1-\lambda)F(r) \leq F(\lambda q + (1-\lambda)r) \quad (3.25)$$

holds for $\lambda = 3/4$, while replacing q with $(q+r)/2$ shows that (3.25) holds for $\lambda = 1/4$. Continuing in this way shows that (3.25) holds for all dyadic fractions $\lambda = k/2^n$, $1 \leq k \leq 2^n$, $n = 1, 2, \dots$. Since dyadic fractions are dense in $[0, 1]$, the continuity of F yields that (3.25) holds for $\lambda \in [0, 1]$ and $q, r \in]0, 1[$. Finally, to prove the statement about the derivatives, fix $q \in]0, 1[$ and for $h > 0$ small enough consider the functions

$$\Delta^+(h) = \frac{F(q+h) - F(q)}{h}, \quad \Delta^-(h) = \frac{F(q-h) - F(q)}{-h}.$$

The concavity of F gives that the function $h \mapsto \Delta^+(h)$ is increasing, that $h \mapsto \Delta^-(h)$ is increasing, and that $\Delta^+(h) \leq \Delta^-(h)$. This establishes the last claim of the Step 2 concerning left and right derivatives of F on $]0, 1[$.

Step 3. There exist functions $\mathcal{R}_n : \mathcal{P}_n \rightarrow \mathbb{R}$, $n \geq 2$, such that

$$\mathfrak{S}_n(p\bar{q}, pq, p_3, \dots, p_n) = pF(q) + \mathcal{R}_{n-1}(p, p_3, \dots, p_n) \quad (3.26)$$

for all $q \in]0, 1[$, $(p, p_3, \dots, p_n) \in \mathcal{P}_{n-1}$ and $n \geq 2$.

To prove this, note that the Step 1 and the relation $F(x) = F(\bar{x})$ give

$$\begin{aligned} \frac{F(\bar{p}q + pq) - F(\bar{p}q + pr)}{q-r} &\leq \frac{S_n(p\bar{q}, pq, p_3, \dots, p_n) - S_n(p\bar{r}, pr, p_3, \dots, p_n)}{q-r} \\ &\leq \frac{F(pq + \bar{p}r) - F(pr + \bar{p}r)}{q-r} \end{aligned} \quad (3.27)$$

for $0 < r < q < 1$ and $(p, p_3, \dots, p_n) \in \mathcal{P}_n$. Fix $(p, p_3, \dots, p_n) \in \mathcal{P}_n$ and set

$$L(q) = \mathfrak{S}_n(p\bar{q}, pq, p_3, \dots, p_n).$$

Taking $q \downarrow r$ in (3.27) we get

$$pD^-F(r) = D^-L(r),$$

while taking $r \uparrow q$ gives

$$pD^+F(q) = D^+L(q).$$

Since $D^\pm F(q)$ is finite by Step 2, we derive that the function $L(q) - pF(q)$ is differentiable on $]0, 1[$ with vanishing derivative. Hence, for $q \in]0, 1[$,

$$L(q) = pF(q) + \mathcal{R}_{n-1}(p, p_3, \dots, p_n),$$

where the constant \mathcal{R}_{n-1} depends on the values (p, p_3, \dots, p_n) we have fixed in the above argument.

Step 4. There exist constants $c \geq 0$ and C such that for all $q \in]0, 1[$,

$$F(q) = cS(1 - q, q) + C. \quad (3.28)$$

We start the proof by taking $(p_1, p_2, p_3) \in \mathcal{P}_{3,f}$. Setting

$$p = p_1 + p_2, \quad q = \frac{p_2}{p_1 + p_2},$$

we write

$$\mathfrak{S}_3(p_1, p_2, p_3) = \mathfrak{S}_3(p\bar{q}, pq, p_3).$$

It then follows from Step 3 that

$$\mathfrak{S}_3(p_1, p_2, p_3) = (p_1 + p_2)\mathfrak{S}_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) + \mathcal{R}_2(p_1 + p_2, p_3). \quad (3.29)$$

By (3.10) we also have

$$\mathfrak{S}_3(p_1, p_2, p_3) = \mathfrak{S}_3(p_1, p_3, p_2) = (p_1 + p_3)\mathfrak{S}_2\left(\frac{p_1}{p_1 + p_3}, \frac{p_3}{p_1 + p_3}\right) + \mathcal{R}_2(p_1 + p_3, p_3). \quad (3.30)$$

Setting $G(x) = \mathcal{R}_2(\bar{x}, x)$, $x = p_3$, $y = p_2$, we rewrite (3.29)=(3.30) as

$$(1 - x)F\left(\frac{y}{1 - x}\right) + G(x) = (1 - y)F\left(\frac{x}{1 - y}\right) + G(y), \quad (3.31)$$

where $x, y \in]0, 1[$ and $x + y < 1$. The rest of the proof concerns analysis of the functional equation (3.31).

Since F is continuous on $]0, 1[$, fixing one variable one easily deduces from (3.31) that G is also continuous on $]0, 1[$. Let $0 < a < b < 1$ and fix $y \in]0, 1 - b[$. It follows that (verify this!)

$$\frac{x}{1 - y} \in \left]a, \frac{b}{1 - y}\right] \subset]0, 1[, \quad \frac{y}{1 - x} \in \left]y, \frac{y}{1 - b}\right] \subset]0, 1[.$$

Integrating (3.31) with respect to x over $[a, b]$ we derive

$$\begin{aligned} (b - a)G(y) &= \int_a^b G(y)dx \\ &= \int_a^b G(x)dx + \int_a^b (1 - x)F\left(\frac{y}{1 - x}\right)dx - (1 - y) \int_a^b F\left(\frac{x}{1 - y}\right)dx \\ &= \int_a^b G(x)dx + y^2 \int_{y/(1-a)}^{y/(1-b)} s^{-3}F(s)ds - (1 - y)^2 \int_{a/(1-y)}^{b/(1-y)} F(t)dt, \end{aligned} \quad (3.32)$$

where we have used the change of variable

$$s = \frac{y}{1 - x}, \quad t = \frac{x}{1 - y}. \quad (3.33)$$

It follows that G is differentiable on $]0, b[$. Since $0 < b < 1$ is arbitrary, G is differentiable on $]0, 1[$.

The change of variable (3.33) maps bijectively $\{(x, y) \mid x, y > 0\}$ to $\{(s, t) \mid s, t \in]0, 1[\}$ (verify this!), and in this new variables the functional equation (3.31) reads

$$F(t) = \frac{1 - t}{1 - s}F(s) + \frac{1 - st}{1 - s} \left[G\left(\frac{t - st}{1 - st}\right) - G\left(\frac{s - st}{1 - st}\right) \right]. \quad (3.34)$$

Fixing s , we see that the differentiability of G implies the differentiability of F on $]0, 1[$. Returning to (3.32), we get that G is twice differentiable on $]0, 1[$, and then (3.34) gives that F is also twice differentiable on $]0, 1[$. Continuing in this way we derive that both F and G are infinitely differentiable on $]0, 1[$. Differentiating (3.31) first with respect to x and then with respect to y gives

$$\frac{y}{(1-x)^2} F''\left(\frac{y}{1-x}\right) = \frac{x}{(1-y)^2} F''\left(\frac{x}{1-y}\right). \quad (3.35)$$

The substitution (3.33) gives that for $s, t \in]0, 1[$,

$$s(1-s)F''(s) = t(1-t)F''(t).$$

It follows that for some $c \in \mathbb{R}$,

$$t(1-t)F''(t) = -c.$$

Integration gives

$$F(t) = cS(1-t, t) + Bt + C.$$

Since $F(t) = F(\bar{t})$, we have $B = 0$, and since F is increasing on $[0, 1/2]$, we have $c \geq 0$. This completes the proof of the Step 4. Note that as a by-product of the proof we have derived that for some constant D ,

$$G(x) = F(x) + D, \quad x \in]0, 1[. \quad (3.36)$$

To prove (3.36), note that (3.28) gives that F satisfies the functional equation

$$(1-x)F\left(\frac{y}{1-x}\right) + F(x) = (1-y)F\left(\frac{x}{1-y}\right) + F(y).$$

Combining this equation with (3.31) we derive that for $x, y > 0, 0 < x + y < 1$,

$$G(x) - F(x) = G(y) - F(y).$$

Hence, $G(x) - F(x) = D_y$ for $x \in]0, 1 - y[$. If $y_1 < y_2$, we must have $D_{y_1} = D_{y_2}$, and so $D = D_y$ does not depend on y , which gives (3.36).

Step 5. For any $n \geq 2$ there exists constant $C(n)$ such that for $(p_1, \dots, p_n) \in \mathcal{P}_{n,f}$,

$$\mathfrak{S}_n(p_1, \dots, p_n) = cS(p_1, \dots, p_n) + C(n), \quad (3.37)$$

where $c \geq 0$ is the constant from the Step 4.

In the Step 4 we established (3.37) for $n = 2$ (we set $C(2) = C$), and so we assume that $n \geq 3$. Set $p = p_1 + p_2, q = p_2/(p_1 + p_2)$. It then follows from Steps 3 and 4 that

$$\begin{aligned} \mathfrak{S}_n(p_1, \dots, p_n) &= (p_1 + p_2)\mathfrak{S}_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) + \mathcal{R}_{n-1}(p_1 + p_2, p_3, \dots, p_n) \\ &= (p_1 + p_2)cS\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) + \widehat{\mathcal{R}}_{n-1}(p_1 + p_2, p_3, \dots, p_n), \end{aligned} \quad (3.38)$$

where $\widehat{\mathcal{R}}_{n-1}(p, p_3, \dots, p_n) = pC_2 + \mathcal{R}_{n-1}(p, p_3, \dots, p_n)$. Note that since \mathcal{R}_{n-1} is invariant under the permutations of the variables (p_3, \dots, p_n) (recall (3.38)), so is $\widehat{\mathcal{R}}_{n-1}$. The invariance of \mathfrak{S}_n under the permutation of the variables gives

$$\mathfrak{S}_n(p_1, \dots, p_n) = (p_1 + p_3)cS\left(\frac{p_1}{p_1 + p_3}, \frac{p_3}{p_1 + p_3}\right) + \widehat{\mathcal{R}}_{n-1}(p_1 + p_3, p_2, p_4, \dots, p_n),$$

and so

$$\begin{aligned} (p_1 + p_2)cS\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) - (p_1 + p_3)cS\left(\frac{p_1}{p_1 + p_3}, \frac{p_3}{p_1 + p_3}\right) \\ = \widehat{\mathcal{R}}_{n-1}(p_1 + p_2, p_3, \dots, p_n) - \widehat{\mathcal{R}}_{n-1}(p_1 + p_3, p_2, p_4, \dots, p_n). \end{aligned} \quad (3.39)$$

Until the the end of the proof when we wish to indicate the number of variables in the Boltzmann–Gibbs–Shannon entropy we will write $S_n(p_1, \dots, p_n)$. One easily verifies that

$$\begin{aligned} S_n(p_1, \dots, p_n) &= (p_1 + p_2)S_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) + S_{n-1}(p_1 + p_2, p_3, \dots, p_n) \\ &= (p_1 + p_3)S_2\left(\frac{p_1}{p_1 + p_3}, \frac{p_3}{p_1 + p_3}\right) + S_{n-1}(p_1 + p_3, p_2, p_4, \dots, p_n), \end{aligned}$$

and so

$$\begin{aligned} (p_1 + p_2)S_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) - (p_1 + p_3)S_2\left(\frac{p_1}{p_1 + p_3}, \frac{p_3}{p_1 + p_3}\right) \\ = S_{n-1}(p_1 + p_2, p_3, \dots, p_n) - S_{n-1}(p_1 + p_3, p_2, p_4, \dots, p_n). \end{aligned} \quad (3.40)$$

Since in the formulas (3.39) and (3.40) $S = S_2$, we derive that the function

$$T_{n-1}(p, q, p_4, \dots, p_n) = \widehat{\mathcal{R}}_{n-1}(p, q, p_4, \dots, p_n) - cS_{n-1}(p, q, p_4, \dots, p_n)$$

satisfies

$$T_{n-1}(p_1 + p_2, p_3, p_4, \dots, p_n) = T_{n-1}(p_1 + p_3, p_2, p_4, \dots, p_n) \quad (3.41)$$

for all $(p_1, \dots, p_n) \in \mathcal{P}_{n,f}$. Moreover, by construction, $T_{n-1}(p, q, p_4, \dots, p_n)$ is invariant under the permutation of the variables (q, p_4, \dots, p_n) . Set $s = p_1 + p_2 + p_3$. Then (3.41) reads as

$$T_{n-1}(s - p_3, p_3, p_4, \dots, p_n) = T_{n-1}(s - p_2, p_2, p - p_4, \dots, p_n).$$

Hence, the map

$$]0, s[\ni p \mapsto T_{n-1}(s - p, p, p_4, \dots, p_n)$$

is constant. By the permutation invariance, the maps

$$]0, s[\ni p \mapsto T_{n-1}(s - p, p_3, \dots, p_{m-1}, p, p_{m+1}, \dots)$$

are also constant. Setting $s = p_1 + p_2 + p_3 + p_4$, we deduce that the map

$$(p_3, p_4) \mapsto T_{n-1}(s - p_3 - p_4, p_3, p_4, \dots, p_n)$$

with domain $p_3 > 0, p_4 > 0, p_3 + p_4 < s$, is constant. Continuing inductively, we conclude that the map

$$(p_3, \dots, p_n) \mapsto T_{n-1}(1 - (p_3 + \dots + p_n), p_3, p_4, \dots, p_n)$$

with domain $p_k > 0, \sum_{k=3}^n p_k < 1$ is constant. Hence, the map

$$\mathcal{P}_{n,f} \ni (p_1, \dots, p_n) \mapsto T_{n-1}(p_1 + p_2, p_3, \dots, p_n)$$

is constant, and we denote the value it assumes by $C(n)$. Returning now to (3.38), we conclude the proof of (3.37):

$$\begin{aligned} \mathfrak{S}_n(p_1, \dots, p_n) &= (p_1 + p_2)cS_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) + \widehat{\mathcal{R}}_{n-1}(p_1 + p_2, p_3, \dots, p_n) \\ &= (p_1 + p_2)cS_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) + cS_{n-1}(p_1 + p_2, p_3, \dots, p_n) + C(n) \\ &= cS_n(p_1, \dots, p_n) + C(n). \end{aligned} \quad (3.42)$$

Step 6. $C(n + m) = C(n)C(m)$ for $n, m \geq 2$, and

$$\liminf_{n \rightarrow \infty} (C(n + 1) - C(n)) = 0. \quad (3.43)$$

If $P_l \in \mathcal{P}_n$ and $P_r \in \mathcal{P}_m$, then the identity $\mathfrak{S}_{nm}(P_l \times P_r) = \mathfrak{S}_n(P_l) + \mathfrak{S}_m(P_r)$ and (3.37) give that $C(n+m) = C(n) + C(m)$. To prove (3.43), suppose that $n \geq 3$ and take in (3.19) $q = 1/2$, $r = 0$, $p = p_3 = \dots = p_n = 1/(n-1)$. Then, combining (3.19) with Step 5, we derive

$$\begin{aligned} & F\left(\frac{1}{2}\right) - F\left(\frac{n-2}{2(n-1)}\right) \\ & \leq \mathfrak{S}_n\left(\frac{1}{2(n-1)}, \frac{1}{2(n-1)}, \frac{1}{n-1}, \dots, \frac{1}{n-1}\right) - \mathfrak{S}_n\left(\frac{1}{n-1}, 0, \frac{1}{n-1}, \dots, \frac{1}{n-1}\right) \\ & = cS_n\left(\frac{1}{2(n-1)}, \frac{1}{2(n-1)}, \frac{1}{n-1}, \dots, \frac{1}{n-1}\right) - cS_{n-1}\left(\frac{1}{n-1}, \frac{1}{n-1}, \dots, \frac{1}{n-1}\right) \\ & \quad + C(n) - C(n-1) \\ & = \frac{\log 2}{n-1} + C(n) - C(n-1). \end{aligned}$$

The first inequality in (3.22) gives

$$0 \leq \frac{\log 2}{n-1} + C(n) - C(n-1),$$

and the statement follows.

Step 7. There is a constant $C \geq 0$ such that for all $n \geq 2$, $C(n) = C \log n$.

Fix $\epsilon > 0$ and $n > 1$. Let $k \in \mathbb{N}$ be such that for all integers $p \geq n^k$, $C(p+1) - C(p) \geq -\epsilon$. It follows that for $p \geq p^k$ and $j \in \mathbb{N}$,

$$C(p+j) - C(p) = \sum_{i=1}^j (C(p+i) - C(p+i-1)) \geq -j\epsilon.$$

Fix now $p \geq n^k$ and let $m \in \mathbb{N}$ be such that $n^m \leq p < n^{m+1}$. Obviously, $m \geq k$. Write

$$p = a_m n^m + a_{m-1} n^{m-1} + \dots + a_1 p + a_0,$$

where a_k 's are integers such that $1 \leq a_m < n$ and $0 \leq a_k < n$ for $k < m$. It follows that

$$C(p) > C(a_m n^m + \dots + a_1 n) - n\epsilon = C(n) + C(a_m n^{m-1} + \dots + a_2 n + a_1) - n\epsilon.$$

Continuing inductively, we derive that

$$C(p) > (m-k+1)C(n) + C(a_m n^{k-1} + a_{m-1} n^{k-2} + \dots + a_{m-k+1}) - (m-k+1)\epsilon.$$

If $M = \max_{2 \leq j \leq n^{k+1}} |C(j)|$, then the last inequality gives

$$C(p) > (m-k+1)C(n) - M - (m-k+1)\epsilon.$$

By the choice of m , $\log p \leq (m+1) \log n$, and so

$$\liminf_{p \rightarrow \infty} \frac{C(p)}{\log p} \geq \frac{C(n)}{\log n}.$$

Since

$$\liminf_{n \rightarrow \infty} \frac{C(p)}{\log p} \leq \liminf_{j \rightarrow \infty} \frac{C(n^j)}{\log n^j} = \frac{C(n)}{n},$$

we derive that for all $n \geq 2$,

$$C(n) = C \log n,$$

where

$$C = \liminf_{p \rightarrow \infty} \frac{C(p)}{p}.$$

It remains to show that $C \geq 0$. Since

$$F(x) = cS_2(1-x, x) + C \log 2,$$

we have $\lim_{x \downarrow 0} F(x) = C \log 2$, and (3.22) yields that $C \geq 0$.

Step 8. We now conclude the proof. Let $P = (p_1, \dots, p_n) \in \mathcal{P}_n$. Write

$$P = (p_{j_1}, \dots, p_{j_k}, 0, \dots, 0),$$

where $p_{j_m} > 0$ for $m = 1, \dots, k$. Then

$$\mathfrak{S}_n(P) = \mathfrak{S}_k(p_{j_1}, \dots, p_{j_k}) = cS_k(p_{j_1}, \dots, p_{j_k}) + C \log k = cS_n(P) + CS_H(P).$$

Since \mathfrak{S}_n is strictly sub-additive, we must have $c + C > 0$. The final statement is a consequence of the fact that S_H is not continuous on \mathcal{P}_n for $n \geq 2$. \square

3.5 Rényi entropy

Let Ω be a finite set and $P \in \mathcal{P}(\Omega)$. For $\alpha \in]0, 1[$ we set

$$S_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{\omega \in \Omega} P(\omega)^\alpha \right).$$

$S_\alpha(P)$ is called the Rényi entropy of P .

Proposition 3.6 (1) $\lim_{\alpha \uparrow 1} S_\alpha(P) = S(P)$.

(2) $\lim_{\alpha \downarrow 0} S_\alpha(P) = S_H(P)$.

(3) $S_\alpha(P) \geq 0$ and $S_\alpha(P) = 0$ iff P is pure.

(4) $S_\alpha(P) \leq \log |\Omega|$ with equality iff $P = P_{\text{ch}}$.

(5) The map $]0, 1[\ni \alpha \mapsto S_\alpha(P)$ is decreasing and is strictly decreasing unless $P = P_{\text{ch}}$.

(6) The map $\mathcal{P}(\Omega) \ni P \mapsto S_\alpha(P)$ is continuous and concave.

(7) If $P = P_l \otimes P_r$ is a product measure on $\Omega = \Omega_l \times \Omega_r$, then $S_\alpha(P) = S_\alpha(P_l) + S_\alpha(P_r)$.

(8) The map $\alpha \mapsto S_\alpha(P)$ extends to a real analytic function on \mathbb{R} by the formulas $S_1(P) = S(P)$ and

$$S_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{\omega \in \text{supp} P} P(\omega)^\alpha \right), \quad \alpha \neq 1.$$

Exercise 3.3. Prove Proposition 3.6.

Exercise 3.4. Describe properties of $S_\alpha(P)$ for $\alpha \notin]0, 1[$.

Exercise 3.5. Let $\Omega = \{-1, 1\} \times \{-1, 1\}$, $0 < p, q < 1$, $p + q = 1$, $p \neq q$, and

$$\begin{aligned} P_\epsilon(-1, -1) &= pq + \epsilon, & P_\epsilon(-1, 1) &= p(1 - q) - \epsilon, \\ P_\epsilon(1, -1) &= (1 - p)q - \epsilon, & P_\epsilon(1, 1) &= (1 - p)(1 - q) + \epsilon. \end{aligned}$$

Show that for $\alpha \neq 1$ and small non-zero ϵ ,

$$S_\alpha(P_\epsilon) > S_\alpha(P_{\epsilon,l}) + S_\alpha(P_{\epsilon,r}).$$

Hence, Rényi entropy is not sub-additive (compare with Theorem 3.5).

3.6 Why is the Rényi entropy natural?

In introducing $S_\alpha(P)$ Rényi was motivated by a concept of generalized means. Let $w_k > 0$, $\sum_{k=1}^n w_k = 1$ be weights and $G :]0, \infty[\rightarrow]0, \infty[$ a continuous strictly increasing function. We shall call such G a *mean function*. The G -mean of strictly positive real numbers x_1, \dots, x_n is

$$S_G(x_1, \dots, x_n) = G^{-1} \left(\sum_{k=1}^n w_k G(x_k) \right).$$

Set $\mathcal{P}_f = \cup_{n \geq 1} \mathcal{P}_{n,f}$.

One then has:

Theorem 3.7 Let $\mathfrak{S} : \mathcal{P}_f \rightarrow [0, \infty[$ be a function with the following properties.

(a) If $P = P_l \otimes P_r$, then $\mathfrak{S}(P) = \mathfrak{S}(P_l) + \mathfrak{S}(P_r)$.

(b) There exists a mean function G such that for all $n \geq 1$ and $P = (p_1, \dots, p_n) \in \mathcal{P}_{n,f}$,

$$\mathfrak{S}(p_1, \dots, p_n) = G^{-1} (\mathbb{E}_P(G(S_P))) = G^{-1} \left(\sum_{k=1}^n p_k G(-\log p_k) \right).$$

(c) $\mathfrak{S}(p, 1 - p) \rightarrow 0$ as $p \rightarrow 0$.

Then there exists $\alpha > 0$ and a constant $c \geq 0$ such that for all $P \in \mathcal{P}_f$,

$$\mathfrak{S}(P) = cS_\alpha(P).$$

Remark 3.3 The assumption (c) excludes the possibility $\alpha \leq 0$.

Remark 3.4 If in addition one requires that the map $\mathcal{P}_{n,f} \ni P \rightarrow \mathfrak{S}(P)$ is concave for all $n \geq 1$, then $\mathfrak{S}(P) = cS_\alpha(P)$ for some $\alpha \in]0, 1]$.

Although historically important, we find that Theorem 3.7 (and any other axiomatic characterization of the Rényi entropy) is less satisfactory than the powerful characterizations of the Boltzmann–Gibbs–Shannon entropy given in Section 3.4. Taking Boltzmann–Gibbs–Shannon entropy for granted, an alternative understanding of the Rényi entropy arises through Cramér’s theorem for the entropy function S_P . For the purpose of this interpretation, without loss of generality we may assume that $P \in \mathcal{P}(\Omega)$ is faithful. Set

$$\hat{S}_\alpha(P) = \log \left(\sum_{\omega \in \Omega} [P(\omega)]^{1-\alpha} \right), \quad \alpha \in \mathbb{R}. \quad (3.44)$$

Obviously, for $\alpha \in \mathbb{R}$,

$$\widehat{S}_\alpha(P) = \alpha S_{1-\alpha}(P). \quad (3.45)$$

The naturalness of the choice (3.44) stems from the fact that the function $\alpha \mapsto \widehat{S}_\alpha(P)$ is the cumulant generating function of $S_P(\omega) = -\log P(\omega)$ with respect to P ,

$$\widehat{S}_\alpha(P) = \log \mathbb{E}_P(e^{\alpha S_P}). \quad (3.46)$$

Passing to the products (Ω^N, P_N) , the LLN gives that for any $\epsilon > 0$,

$$\lim_{N \rightarrow \infty} P_N \left\{ \omega = (\omega_1, \dots, \omega_N) \in \Omega^N \mid \left| \frac{S_P(\omega_1) + \dots + S_P(\omega_N)}{N} - S(P) \right| \geq \epsilon \right\} = 0. \quad (3.47)$$

It follows from Cramér's theorem that the rate function

$$I(\theta) = \sup_{\alpha \in \mathbb{R}} (\alpha\theta - \widehat{S}_\alpha(P)), \quad \theta \in \mathbb{R}, \quad (3.48)$$

controls the fluctuations that accompany the limit (3.47):

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \omega = (\omega_1, \dots, \omega_N) \in \Omega^N \mid \frac{S_P(\omega_1) + \dots + S_P(\omega_N)}{N} \in [a, b] \right\} = - \inf_{\theta \in [a, b]} I(\theta). \quad (3.49)$$

We shall adopt a point of view that the relations (3.45), (3.48), and (3.49) constitute the foundational basis for introduction of the Rényi entropy. In accordance with this interpretation, the traditional definition of the Rényi entropy is somewhat redundant, and one may as well work with $\widehat{S}_\alpha(P)$ from the beginning and call it the *Rényi entropy* of P (or α -entropy of P when there is a danger of confusion).

The basic properties of the map $\alpha \mapsto \widehat{S}_\alpha(P)$ follow from (3.46) and results described in Section 2.4. Note that $S_0(P) = 0$ and $S_1(P) = \log |\Omega|$. The map $\mathcal{P}_f(\Omega) \ni P \mapsto \widehat{S}_\alpha(P)$ is convex for $\alpha \notin [0, 1]$ and concave for $\alpha \in]0, 1[$.

3.7 Notes and references

The celebrated expression (3.5) for entropy of a probability measure goes back to 1870's and works of Boltzmann and Gibbs on the foundations of statistical mechanics. This will be discussed in more detail in Part II of the lecture notes. Shannon has rediscovered this expression in his work on foundations of mathematical information theory [Sha]. The results of Section 3.2 and 3.3 go back to this seminal work. Regarding Exercise 3.2, Hartley entropy was introduced in [Har]. Hartley's work has partly motivated Shannon's [Sha].

Shannon was also first to give an axiomatization of entropy. The axioms in [Sha] are the continuity of \mathfrak{S} on \mathcal{P}_n for all n , the split-additivity (3.12), and the monotonicity $\mathfrak{S}(\overline{P}_{n+1}) < \mathfrak{S}(\overline{P}_n)$, where $\overline{P}_k \in \mathcal{P}_k$ is the chaotic probability measures. Shannon then proved that the only functions \mathfrak{S} satisfying these properties are cS , $c > 0$. Theorem 3.4 is in spirit of Shannon's axiomatization, with the monotonicity axiom $\mathfrak{S}(\overline{P}_{n+1}) < \mathfrak{S}(\overline{P}_n)$ dropped and the continuity requirement relaxed; see Chapter 2 in [AczDa] for additional information and Theorem 2.2.3 in [Thi] whose proof we roughly followed. We leave it as an exercise for the reader to simplify the proof of Theorem 3.4 under additional Shannon's axioms.

Shannon comments in [Sha] on the importance of his axiomatization as

This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications.

The others beg to differ on its importance, and axiomatizations of entropies became an independent research direction, starting with early works of Khintchine [Khi] and Faddeev [Fadd]. Much of these efforts are summarized in the monograph [AczDa], see also [Csi].

The magnificent Theorem 3.5 is due to Aczél, Forte, and Ng [AcFoNg]. I was not able to simplify their arguments and the proof of Theorem 3.5 follows closely the original paper. The Step 7 is due to [Kát]. The proof of Theorem 3.5 can be also found in [AczDa], Section 4.4. An interesting exercise that may elucidate a line of thought that has led to the proof of Theorem 3.5 is to simplify various steps of the the proof by making additional regularity assumptions.

Rényi entropy has been introduced in [Rén]. Theorem 3.7 was proven in [Dar]; see Chapter 5 in [AczDa] for additional information.

Chapter 4

Relative entropy

4.1 Definition and basic properties

Let Ω be a finite set and $P, Q \in \mathcal{P}(\Omega)$. If $P \ll Q$, the relative entropy function of the pair (P, Q) is defined for $\omega \in \text{supp}P$ by

$$cS_{P|Q}(\omega) = cS_Q(\omega) - cS_P(\omega) = c \log P(\omega) - c \log Q(\omega) = c \log \Delta_{P|Q}(\omega),$$

where $c > 0$ is a constant that does not depend on Ω, P, Q . The relative entropy of P with respect to Q is

$$S(P|Q) = c \int_{\text{supp}P} S_{P|Q} dP = c \sum_{\omega \in \text{supp}P} P(\omega) \log \frac{P(\omega)}{Q(\omega)}. \quad (4.1)$$

If P is not absolutely continuous with respect to Q (i.e., $Q(\omega) = 0$ and $P(\omega) > 0$ for some ω), we set

$$S(P|Q) = \infty.$$

The value of the constant c will play no role in the sequel, and we set $c = 1$. As in the case of entropy, the constant c will reappear in the axiomatic characterizations of relative entropy (see Theorems 5.1 and 5.2).

Note that

$$S(P|P_{\text{ch}}) = -S(P) + \log |\Omega|.$$

Proposition 4.1 $S(P|Q) \geq 0$ and $S(P|Q) = 0$ iff $P = Q$.

Proof. We need to consider only the case $P \ll Q$. By Jensen's inequality,

$$\sum_{\omega \in \text{supp}P} P(\omega) \log \frac{Q(\omega)}{P(\omega)} \leq \log \left(\sum_{\omega \in \text{supp}P} Q(\omega) \right),$$

and so

$$\sum_{\omega \in \text{supp}P} P(\omega) \log \frac{Q(\omega)}{P(\omega)} \leq 0$$

with equality iff $P = Q$. □

The next result refines the previous proposition. Recall that the variational distance $d_V(P, Q)$ is defined by (3.2).

Theorem 4.2

$$S(P|Q) \geq \frac{1}{2} d_V(P, Q)^2. \quad (4.2)$$

The equality holds iff $P = Q$.

Proof. We start with the elementary inequality

$$(1+x)\log(1+x) - x \geq \frac{1}{2} \frac{x^2}{1+\frac{x}{3}}, \quad x \geq -1. \quad (4.3)$$

This inequality obviously holds for $x = -1$, so we may assume that $x > -1$. Denote the l.h.s by $F(x)$ and the r.h.s. by $G(x)$. One verifies that $F(0) = F'(0) = G(0) = G'(0) = 0$, and that

$$F''(x) = \frac{1}{1+x}, \quad G''(x) = \left(1 + \frac{x}{3}\right)^{-3}.$$

Obviously, $F''(x) > G''(x)$ for $x > -1, x \neq 0$. Integrating this inequality we derive that $F'(x) > G'(x)$ for $x > 0$ and $F'(x) < G'(x)$ for $x \in]-1, 0[$. Integrating these inequalities we get $F(x) \geq G(x)$ and that equality holds iff $x = 0$.

We now turn to the proof of the theorem. We need only to consider the case $P \ll Q$. Set

$$X(\omega) = \frac{P(\omega)}{Q(\omega)} - 1,$$

with the convention that $0/0 = 0$. Note that $\int_{\Omega} X dQ = 0$ and that

$$S(P|Q) = \int_{\Omega} ((X+1)\log(X+1) - X) dQ.$$

The inequality (4.3) implies

$$S(P|Q) \geq \frac{1}{2} \int_{\Omega} \frac{X^2}{1+\frac{X}{3}} dQ, \quad (4.4)$$

with the equality iff $P = Q$. Note that

$$\int_{\Omega} \left(1 + \frac{X}{3}\right) dQ = 1,$$

and that Cauchy-Schwarz inequality gives

$$\int_{\Omega} \frac{X^2}{1+\frac{X}{3}} dQ = \left(\int_{\Omega} \left(1 + \frac{X}{3}\right) dQ\right) \left(\int_{\Omega} \frac{X^2}{1+\frac{X}{3}} dQ\right) \geq \left(\int_{\Omega} |X| dQ\right)^2 = d_V(P, Q)^2. \quad (4.5)$$

Combining (4.4) and (4.5) we derive the statement. \square

Exercise 4.1. Prove that the estimate (4.2) is the best possible in the sense that

$$\inf_{P \neq Q} \frac{S(P|Q)}{d_V(P, Q)^2} = \frac{1}{2}.$$

Set

$$\mathcal{A}(\Omega) = \{(P, Q) \mid P, Q \in \mathcal{P}(\Omega), P \ll Q\}. \quad (4.6)$$

One easily verifies that $\mathcal{A}(\Omega)$ is a convex subset of $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$. Obviously,

$$\mathcal{A}(\Omega) = \{(P, Q) \mid S(P|Q) < \infty\}.$$

Note also that $\mathcal{P}(\Omega) \times \mathcal{P}_f(\Omega)$ is a dense subset of $\mathcal{A}(\Omega)$.

Proposition 4.3 *The map*

$$\mathcal{A}(\Omega) \ni (P, Q) \mapsto S(P|Q)$$

is continuous, and the map

$$\mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \ni (P, Q) \mapsto S(P|Q) \quad (4.7)$$

is lower semicontinuous.

Exercise 4.2. Prove the above proposition. Show that if $|\Omega| > 1$ and Q is a boundary point of $\mathcal{P}(\Omega)$, then there is a sequence $P_n \rightarrow Q$ such that $\lim_{n \rightarrow \infty} S(P_n|Q) = \infty$. Hence, the map (4.7) is not continuous except in the trivial case $|\Omega| = 1$.

Proposition 4.4 *The relative entropy is jointly convex: for $\lambda \in]0, 1[$ and $P_1, P_2, Q_1, Q_2 \in \mathcal{P}(\Omega)$,*

$$S(\lambda P_1 + (1 - \lambda)P_2 | \lambda Q_1 + (1 - \lambda)Q_2) \leq \lambda S(P_1|Q_1) + (1 - \lambda)S(P_2|Q_2). \quad (4.8)$$

Moreover, if the r.h.s. in (4.8) is finite, the equality holds iff for $\omega \in \text{supp } Q_1 \cap \text{supp } Q_2$ we have $P_1(\omega)/Q_1(\omega) = P_2(\omega)/Q_2(\omega)$.

Remark 4.1 In particular, if $Q_1 \perp Q_2$ and the r.h.s. in (4.8) is finite, then $P_1 \perp P_2$ and the equality holds in (4.8). On the other hand, if $Q_1 = Q_2 = Q$ and Q is faithful,

$$S(\lambda P_1 + (1 - \lambda)P_2 | Q) \leq \lambda S(P_1|Q) + (1 - \lambda)S(P_2|Q).$$

with the equality iff $P_1 = P_2$. An analogous statement holds if $P_1 = P_2 = P$ and P is faithful.

Proof. We recall the following basic fact: if $g :]0, \infty[\rightarrow \mathbb{R}$ is concave, then the function

$$G(x, y) = xg\left(\frac{y}{x}\right) \quad (4.9)$$

is jointly concave on $]0, \infty[\times]0, \infty[$. Indeed, for $\lambda \in]0, 1[$,

$$\begin{aligned} & G(\lambda x_1 + (1 - \lambda)x_2, \lambda y_1 + (1 - \lambda)y_2) \\ &= (\lambda x_1 + (1 - \lambda)x_2)g\left(\frac{\lambda x_1 y_1 + (1 - \lambda)x_2 y_2}{\lambda x_1 + (1 - \lambda)x_2}\right) \\ &\geq \lambda G(x_1, y_1) + (1 - \lambda)G(x_2, y_2), \end{aligned} \quad (4.10)$$

and if g is strictly concave, the inequality is strict unless $\frac{y_1}{x_1} = \frac{y_2}{x_2}$.

We now turn to the proof. Without loss of generality we may assume that $P_1 \ll Q_1$ and $P_2 \ll Q_2$. One easily shows that then also $\lambda P_1 + (1 - \lambda)P_2 \ll \lambda Q_1 + (1 - \lambda)Q_2$. For any $\omega \in \Omega$ we have that

$$\begin{aligned} & (\lambda P_1(\omega) + (1 - \lambda)P_2(\omega)) \log \frac{\lambda P_1(\omega) + (1 - \lambda)P_2(\omega)}{\lambda Q_1(\omega) + (1 - \lambda)Q_2(\omega)} \\ &\leq \lambda P_1(\omega) \log \frac{P_1(\omega)}{Q_1(\omega)} + (1 - \lambda)P_2(\omega) \log \frac{P_2(\omega)}{Q_2(\omega)}. \end{aligned} \quad (4.11)$$

To establish this relation, note that if $P_1(\omega) = P_2(\omega) = 0$, then (4.11) holds with the equality. If $P_1(\omega) = 0$ and $P_2(\omega) > 0$, the inequality (4.11) is strict unless $Q_1(\omega) = 0$, and similarly in the case $P_1(\omega) > 0$, $P_2(\omega) = 0$. If $P_1(\omega) > 0$ and $P_2(\omega) > 0$, then taking $g(t) = \log t$ in (4.9) and using the joint concavity of G gives that (4.11) holds and that the inequality is strict unless $P_1(\omega)/Q_1(\omega) = P_2(\omega)/Q_2(\omega)$. Summing (4.11) over ω we derive the statement. The discussion of the cases where the equality holds in (4.8) is simple and is left to the reader. \square

The relative entropy is super-additive in the following sense:

Proposition 4.5 For any P and $Q = Q_l \otimes Q_r$ in $\mathcal{P}(\Omega_l \times \Omega_r)$,

$$S(P_l|Q_l) + S(P_r|Q_r) \leq S(P|Q). \quad (4.12)$$

Moreover, if the r.h.s. in (4.12) is finite, the equality holds iff $P = P_l \otimes P_r$.

Proof. We may assume that $P \ll Q$, in which case one easily verifies that $P_l \ll Q_l$ and $P_r \ll Q_r$. One computes

$$S(P|Q) - S(P_l|Q_l) - S(P_r|Q_r) = S(P_l) + S(P_r) - S(P),$$

and the result follows from Proposition 3.2. \square

In general, for $P, Q \in \mathcal{P}(\Omega_l \times \Omega_r)$ it is *not* true that $S(P|Q) \geq S(P_l|Q_l) + S(P_r|Q_r)$ even if $P = P_l \otimes P_r$.

Exercise 4.3. Find an example of faithful $P = P_l \otimes P_r, Q \in \mathcal{P}(\Omega_l \times \Omega_r)$ where $|\Omega_l| = |\Omega_r| = 2$ such that

$$S(P|Q) < S(P_l|Q_l) + S(P_r|Q_r).$$

Let $\Omega = (\omega_1, \dots, \omega_L), \widehat{\Omega} = \{\hat{\omega}_1, \dots, \hat{\omega}_{\widehat{L}}\}$ be two finite sets. A matrix of real numbers $[\Phi(\omega, \hat{\omega})]_{(\omega, \hat{\omega}) \in \Omega \times \widehat{\Omega}}$ is called *stochastic* if $\Phi(\omega, \hat{\omega}) \geq 0$ for all pairs $(\omega, \hat{\omega})$ and

$$\sum_{\hat{\omega} \in \widehat{\Omega}} \Phi(\omega, \hat{\omega}) = 1$$

for all $\omega \in \Omega$. A stochastic matrix induces a map $\Phi : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\widehat{\Omega})$ by

$$\Phi(P)(\hat{\omega}) = \sum_{\omega \in \Omega} P(\omega) \Phi(\omega, \hat{\omega}).$$

We shall refer to Φ as the *stochastic map* induced by the stochastic matrix $[\Phi(\omega, \hat{\omega})]$. One can interpret the elements of Ω and $\widehat{\Omega}$ as states of two stochastic systems and $P(\omega)$ as probability that the state ω is realized. $\Phi(\omega, \hat{\omega})$ is interpreted as the *transition probability*, i.e. the probability that in a unit of time the system will make a transition from the state ω to the state $\hat{\omega}$. With this interpretation, the probability that the state $\hat{\omega}$ is realized after the transition has taken place is $\Phi(P)(\hat{\omega})$.

Note that if $[\Phi(\omega, \hat{\omega})]_{(\omega, \hat{\omega}) \in \Omega \times \widehat{\Omega}}$ and $[\widehat{\Phi}(\hat{\omega}, \hat{\hat{\omega}})]_{(\hat{\omega}, \hat{\hat{\omega}}) \in \widehat{\Omega} \times \widehat{\widehat{\Omega}}}$ are stochastic matrices, then their product is also stochastic matrix and that the induced stochastic map is $\widehat{\Phi} \circ \Phi$. Another elementary property of stochastic maps is:

Proposition 4.6 $d_V(\Phi(P), \Phi(Q)) \leq d_V(P, Q)$.

Exercise 4.4. Prove Proposition 4.6. When the equality holds?

The following result is deeper.

Proposition 4.7

$$S(\Phi(P)|\Phi(Q)) \leq S(P|Q). \quad (4.13)$$

Remark 4.2 In information theory, the inequality (4.13) is sometimes called the *data processing inequality*. We shall refer to it as the *stochastic monotonicity*. If the relative entropy is interpreted as a measure of *distinguishability* of two probability measures, then the inequality asserts that probability measures are less distinguishable after an application of a stochastic map.

Proof. We start with the so called *log-sum inequality*: If $a_j, b_j, j = 1, \dots, M$, are non-negative numbers, then

$$\sum_{j=1}^M a_j \log \frac{a_j}{b_j} \geq \sum_{j=1}^M a_j \log \frac{\sum_{k=1}^M a_k}{\sum_{k=1}^M b_k}, \quad (4.14)$$

with the usual convention that $0 \log 0/x = 0$. If $b_j = 0$ and $a_j > 0$ for some j , then l.h.s is ∞ and there is nothing to prove. If $a_j = 0$ for all j again there is nothing to prove. Hence, without loss of generality we may assume that $\sum_j a_j > 0, \sum b_j > 0$, and $b_j = 0 \Rightarrow a_j = 0$. Set $p = (p_1, \dots, p_M), p_k = a_k / \sum_j a_j, q = (q_1, \dots, q_M), q_k = b_k / \sum_j b_j$. Then the inequality (4.14) is equivalent to

$$S(p|q) \geq 0.$$

This observation and Proposition 4.1 prove (4.14).

We now turn to the proof. Clearly, we need only to consider the case $P \ll Q$. Then

$$\begin{aligned} S(\Phi(P)|\Phi(Q)) &= \sum_{\hat{\omega} \in \hat{\Omega}} \Phi(P)(\hat{\omega}) \log \frac{\Phi(P)(\hat{\omega})}{\Phi(Q)(\hat{\omega})} \\ &= \sum_{\hat{\omega} \in \hat{\Omega}} \sum_{\omega \in \Omega} P(\omega) \Phi(\omega, \hat{\omega}) \log \frac{\sum_{\omega' \in \Omega} P(\omega') \Phi(\omega', \hat{\omega})}{\sum_{\omega' \in \Omega} Q(\omega') \Phi(\omega', \hat{\omega})} \\ &\leq \sum_{\hat{\omega} \in \hat{\Omega}} \sum_{\omega \in \Omega} P(\omega) \Phi(\omega, \hat{\omega}) \log \frac{P(\omega)}{Q(\omega)} \\ &= S(P|Q), \end{aligned}$$

where the third step follows from the log-sum inequality. \square \square

Exercise 4.5. A stochastic matrix $[\Phi(\omega, \hat{\omega})]$ is called doubly stochastic if

$$\sum_{\omega \in \Omega} \Phi(\omega, \hat{\omega}) = \frac{|\Omega|}{|\hat{\Omega}|}$$

for all $\hat{\omega} \in \hat{\Omega}$. Prove that $S(P) \leq S(\Phi(P))$ for all $P \in \mathcal{P}(\Omega)$ iff $[\Phi(\omega, \hat{\omega})]$ is doubly stochastic. Hint: Use that $\Phi(P_{\text{ch}}) = \hat{P}_{\text{ch}}$ iff $[\Phi(\omega, \hat{\omega})]$ is doubly stochastic.

Exercise 4.6. Suppose that $\Omega = \hat{\Omega}$. Let $\gamma = \min_{(\omega_1, \omega_2)} \Phi(\omega_1, \omega_2)$ and suppose that $\gamma > 0$.

1. Show that $S(\Phi(P)|\Phi(Q)) = S(P|Q)$ iff $P = Q$.

2. Show that

$$d_V(\Phi(P), \Phi(Q)) \leq (1 - \gamma) d_V(P, Q).$$

3. Using Part 2 show that there exists unique probability measure \bar{Q} such that $\Phi(\bar{Q}) = \bar{Q}$. Show that \bar{Q} is faithful and that for any $P \in \mathcal{P}(\Omega)$,

$$d_V(\Phi^n(P), \bar{Q}) \leq (1 - \gamma)^n d_V(P, \bar{Q}),$$

where $\Phi^2 = \Phi \circ \Phi$, etc.

Hint: Follow the proof of the Banach fixed point theorem.

Exercise 4.7. The stochastic monotonicity yields the following elegant proof of Theorem 4.2.

1. Let $P, Q \in \mathcal{P}(\Omega)$ be given, where $|\Omega| \geq 2$. Let $T = \{\omega : P(\omega) \geq Q(\omega)\}$ and

$$p = (p_1, p_2) = (P(T), P(T^c)), \quad q = (q_1, q_2) = (Q(T), Q(T^c)),$$

be probability measures on $\widehat{\Omega} = \{1, 2\}$. Find a stochastic map $\Phi : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\widehat{\Omega})$ such that $\Phi(P) = p$, $\Phi(Q) = q$.

2. Since $S(P|Q) \geq S(p|q)$ and $d_V(P, Q) = d_V(p, q)$, observe that to prove Theorem 4.2 it suffices to show that for all $p, q \in \mathcal{P}(\widehat{\Omega})$,

$$S(p|q) \geq \frac{1}{2} d_V(p, q)^2. \quad (4.15)$$

3. Show that (4.15) is equivalent to the inequality

$$x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} \geq 2(x-y)^2, \quad (4.16)$$

where $0 \leq y \leq x \leq 1$. Complete the proof by establishing (4.16).

Hint: Fix $x > 0$ and consider the function

$$F(y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} - 2(x-y)^2$$

on $]0, x[$. Since $F(x) = 0$, it suffices to show that $F'(y) \leq 0$ for $y \in]0, x[$. Direct computation gives $F'(y) \leq 0 \Leftrightarrow y(1-y) \leq \frac{1}{4}$ and the statement follows.

The log-sum inequality used in the proof Proposition 4.7 leads to the following refinement of Proposition 4.4.

Proposition 4.8 Let $P_1, \dots, P_n, Q_1, \dots, Q_n \in \mathcal{P}(\Omega)$ and $p = (p_1, \dots, p_n), q = (q_1, \dots, q_n) \in \mathcal{P}_n$. Then

$$S(p_1 P_1 + \dots + p_n P_n | q_1 Q_1 + \dots + q_n Q_n) \leq p_1 S(P_1 | Q_1) + \dots + p_n S(P_n | Q_n) + S(p|q). \quad (4.17)$$

If the r.h.s. in (4.17) is finite, then the equality holds iff for all j, k such that $q_j > 0, q_k > 0$,

$$\frac{p_j P_j(\omega)}{q_j Q_j(\omega)} = \frac{p_k P_k(\omega)}{q_k Q_k(\omega)}$$

holds for all $\omega \in \text{supp } Q_k \cap \text{supp } Q_j$.

Exercise 4.8. Deduce Proposition 4.8 from the log-sum inequality.

4.2 Variational principles

The relative entropy is characterized by the following variational principle.

Proposition 4.9

$$S(P|Q) = \sup_{X: \Omega \rightarrow \mathbb{R}} \left(\int_{\Omega} X dP - \log \int_{\text{supp } P} e^X dQ \right). \quad (4.18)$$

If $S(P|Q) < \infty$, then the supremum is achieved, and each maximizer is equal to $S_{P|Q} + \text{const}$ on $\text{supp}P$ and is arbitrary otherwise.

Proof. Suppose that $Q(\omega_0) = 0$ and $P(\omega_0) > 0$ for some $\omega_0 \in \Omega$. Set $X_n(\omega) = n$ if $\omega = \omega_0$ and zero otherwise. Then

$$\int_{\Omega} X_n dP = nP(\omega_0), \quad \int_{\text{supp}P} e^{X_n} dQ = Q(\text{supp}P).$$

Hence, if P is not absolutely continuous w.r.t. Q the relation (4.18) holds since both sides are equal to ∞ .

Suppose now that $P \ll Q$. For given $X : \Omega \rightarrow \mathbb{R}$ set

$$Q_X(\omega) = \frac{e^{X(\omega)} Q(\omega)}{\sum_{\omega' \in \text{supp}P} e^{X(\omega')} Q(\omega')}$$

if $\omega \in \text{supp}P$ and zero otherwise. $Q_X \in \mathcal{P}(\Omega)$ and

$$S(P|Q_X) = S(P|Q) - \left(\int_{\Omega} X dP - \log \int_{\text{supp}P} e^X dQ \right).$$

Hence,

$$S(P|Q) \geq \int_{\Omega} X dP - \log \int_{\text{supp}P} e^X dQ$$

with equality iff $P = Q_X$. Obviously, $P = Q_X$ iff $X = S_{P|Q} + \text{const}$ on $\text{supp}P$ and is arbitrary otherwise. \square

Exercise 4.9. Show that

$$S(P|Q) = \sup_{X: \Omega \rightarrow \mathbb{R}} \left(\int_{\Omega} X dP - \log \int_{\Omega} e^X dQ \right). \quad (4.19)$$

When is the supremum achieved? Use (4.19) to prove that the map $(P, Q) \mapsto S(P|Q)$ is jointly convex.

Proposition 4.10 *The following dual variational principle holds: for $X : \Omega \rightarrow \mathbb{R}$ and $Q \in \mathcal{P}(\Omega)$,*

$$\log \int_{\Omega} e^X dQ = \max_{P \in \mathcal{P}(\Omega)} \left(\int_{\Omega} X dP - S(P|Q) \right).$$

The maximizer is unique and is given by

$$P_{X,Q}(\omega) = \frac{e^{X(\omega)} Q(\omega)}{\sum_{\omega' \in \Omega} e^{X(\omega')} Q(\omega')}.$$

Proof. For any $P \ll Q$,

$$\log \int_{\Omega} e^X dQ - \int_{\Omega} X dP + S(P|Q) = S(P|P_{X,Q}),$$

and the result follows from Proposition 4.1. \square

Setting $Q = P_{\text{ch}}$ in Propositions 4.9 and 4.10, we derive the variational principle for entropy and the respective dual variational principle.

Proposition 4.11 (1)

$$S(P) = \inf_{X: \Omega \rightarrow \mathbb{R}} \left(\log \left(\sum_{\omega \in \Omega} e^{X(\omega)} \right) - \int_{\Omega} X dP \right).$$

The infimum is achieved if P is faithful and $X = -S_P + \text{const}$.

(2) For any $X : \Omega \rightarrow \mathbb{R}$,

$$\log \left(\sum_{\omega \in \Omega} e^{X(\omega)} \right) = \max_{P \in \mathcal{P}(\Omega)} \left(\int_{\Omega} X dP + S(P) \right).$$

The maximizer is unique and is given by

$$P(\omega) = \frac{e^{X(\omega)}}{\sum_{\omega' \in \Omega} e^{X(\omega')}}.$$

4.3 Stein's Lemma

Let $P, Q \in \mathcal{P}(\Omega)$ and let P_N, Q_N be the induced product probability measures on Ω^N . For $\gamma \in]0, 1[$ the Stein exponents are defined by

$$s_N(\gamma) = \min \{ Q_N(T) \mid T \subset \Omega^N, P_N(T) \geq \gamma \}. \quad (4.20)$$

The following result is often called *Stein's Lemma*.

Theorem 4.12

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log s_N(\gamma) = -S(P|Q).$$

Remark 4.3 If $Q = P_{\text{ch}}$, then Stein's Lemma reduces to Proposition 3.3. In fact, the proofs of the two results are very similar.

Proof. We deal first with the case $S(P|Q) < \infty$. Set $S_{P|Q}(\omega) = 0$ for $\omega \notin \text{supp} P$ and

$$S_N(\omega = (\omega_1, \dots, \omega_N)) = \sum_{j=1}^N S_{P|Q}(\omega_j).$$

For given $\epsilon > 0$ let

$$R_{N,\epsilon} = \left\{ \omega \in \Omega^N \mid \frac{S_N(\omega)}{N} \geq S(P|Q) - \epsilon \right\}.$$

By the LLN,

$$\lim_{N \rightarrow \infty} P_N(R_{N,\epsilon}) = 1,$$

and so for N large enough, $P_N(R_{N,\epsilon}) \geq \gamma$. We also have

$$Q_N(R_{N,\epsilon}) = Q_N \left\{ e^{S_N(\omega)} \geq e^{NS(P|Q) - N\epsilon} \right\} \leq e^{N\epsilon - NS(P|Q)} \mathbb{E}_{Q_N}(e^{S_N}).$$

Since

$$\mathbb{E}_{Q_N}(e^{S_N}) = \left(\int_{\Omega} \Delta_{P|Q} dQ \right)^N = 1,$$

we derive

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log s_N(\gamma) \leq -S(P|Q) + \epsilon.$$

Since $\epsilon > 0$ is arbitrary,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log s_N(\gamma) \leq -S(P|Q).$$

To prove the lower bound, let $U_{N,\gamma}$ be the set for which the minimum in (4.20) is achieved. Let $\epsilon > 0$ be given and let

$$D_{N,\epsilon} = \left\{ \omega \in \Omega^N \mid \frac{S_N(\omega)}{N} \leq S(P|Q) + \epsilon \right\}.$$

Again, by the LLN,

$$\lim_{N \rightarrow \infty} P_N(D_{N,\epsilon}) = 1,$$

and so for N large enough, $P_N(D_{N,\epsilon}) \geq \gamma$. We then have

$$\begin{aligned} P_N(U_{N,\gamma} \cap D_{N,\epsilon}) &= \int_{U_{N,\gamma} \cap D_{N,\epsilon}} \Delta_{P_N|Q_N} dQ_N = \int_{U_{N,\gamma} \cap D_{N,\epsilon}} e^{S_N} dQ_N \\ &\leq e^{NS(P|Q)+N\epsilon} Q_N(U_{N,\gamma} \cap D_{N,\epsilon}) \\ &\leq e^{NS(P|Q)+N\epsilon} Q_N(U_{N,\gamma}). \end{aligned}$$

Since

$$\liminf_{N \rightarrow \infty} P_N(U_{N,\gamma} \cap D_{N,\epsilon}) \geq \gamma,$$

we have

$$\liminf_{N \rightarrow \infty} \frac{1}{N} s_N(\gamma) \geq -S(P|Q) - \epsilon.$$

Since $\epsilon > 0$ is arbitrary,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} s_N(\gamma) \geq -S(P|Q).$$

This proves Stein's Lemma in the case $S(P|Q) < \infty$.

We now deal with the case $S(P|Q) = \infty$. For $0 < \delta < 1$ set $Q_\delta = (1 - \delta)Q + \delta P$. Obviously, $S(P|Q_\delta) < \infty$. Let $s_{N,\delta}(\gamma)$ be the Stein exponent of the pair (P, Q_δ) . Then

$$s_{N,\delta}(\gamma) \geq (1 - \delta)^N s_N(\gamma),$$

and

$$-S(P|Q_\delta) = \lim_{N \rightarrow \infty} \frac{1}{N} \log s_{N,\delta}(\gamma) \geq \log(1 - \delta) + \liminf_{N \rightarrow \infty} \frac{1}{N} \log s_N(\gamma).$$

The lower semicontinuity of relative entropy gives $\lim_{\delta \rightarrow 0} S(P|Q_\delta) = \infty$, and so

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log s_N(\gamma) = \infty = -S(P|Q).$$

□

Exercise 4.10. Prove the following variant of Stein's Lemma. Let

$$\begin{aligned} \underline{s} &= \inf_{(T_N)} \left\{ \liminf_{N \rightarrow \infty} \frac{1}{N} Q_N(T_N) \mid \lim_{N \rightarrow \infty} P_N(T_N^c) = 0 \right\}, \\ \bar{s} &= \inf_{(T_N)} \left\{ \limsup_{N \rightarrow \infty} \frac{1}{N} Q_N(T_N) \mid \lim_{N \rightarrow \infty} P_N(T_N^c) = 0 \right\}, \end{aligned}$$

where the infimum is taken over all sequences $(T_N)_{N \geq 1}$ of sets such that $T_N \subset \Omega^N$ for all $N \geq 1$. Then

$$\underline{s} = \bar{s} = -S(P|Q).$$

4.4 Fluctuation relation

Let Ω be a finite set and $P \in \mathcal{P}_f(\Omega)$. Let $\Theta : \Omega \rightarrow \Omega$ be a bijection such that

$$\Theta^2(\omega) = \Theta \circ \Theta(\omega) = \omega \quad (4.21)$$

for all ω . We set $P_\Theta(\omega) = P(\Theta(\omega))$. Obviously, $P_\Theta \in \mathcal{P}_f(\Omega)$. The relative entropy function

$$S_{P|P_\Theta}(\omega) = \log \frac{P(\omega)}{P_\Theta(\omega)}$$

satisfies

$$S_{P|P_\Theta}(\Theta(\omega)) = -S_{P|P_\Theta}(\omega), \quad (4.22)$$

and so the set of values of $S_{P|P_\Theta}$ is symmetric with respect to the origin. On the other hand,

$$S(P|P_\Theta) = \mathbb{E}_P(S_{P|P_\Theta}) \geq 0$$

with equality iff $P = P_\Theta$. Thus, the probability measure P "favours" positive values of $S_{P|P_\Theta}$. Proposition 4.13 below is a refinement of this observation.

Let Q be the probability distribution of the random variable $S_{P|P_\Theta}$ w.r.t. P . We recall that Q is defined by

$$Q(s) = P\{\omega \mid S_{P|P_\Theta}(\omega) = s\}.$$

Obviously, $Q(s) \neq 0$ iff $Q(-s) \neq 0$.

The following result is known as the *fluctuation relation*.

Proposition 4.13 For all s ,

$$Q(-s) = e^{-s}Q(s).$$

Proof. For any α ,

$$\begin{aligned} \mathbb{E}_P(e^{-\alpha S_{P|P_\Theta}}) &= \sum_{\omega \in \Omega} [P_\Theta(\omega)]^\alpha [P(\omega)]^{1-\alpha} \\ &= \sum_{\omega \in \Omega} [P_\Theta(\Theta(\omega))]^\alpha [P(\Theta(\omega))]^{1-\alpha} \\ &= \sum_{\omega \in \Omega} [P(\omega)]^\alpha [P_\Theta(\omega)]^{1-\alpha} \\ &= \mathbb{E}_P(e^{-(1-\alpha)S_{P|P_\Theta}}). \end{aligned}$$

Hence, if $\mathcal{S} = \{s \mid Q(s) \neq 0\}$,

$$\sum_{s \in \mathcal{S}} e^{-\alpha s} Q(s) = \sum_{s \in \mathcal{S}} e^{-(1-\alpha)s} Q(s) = \sum_{s \in \mathcal{S}} e^{(1-\alpha)s} Q(-s),$$

and so

$$\sum_{s \in \mathcal{S}} e^{-\alpha s} (Q(s) - e^s Q(-s)) = 0. \quad (4.23)$$

Since (4.23) holds for all real α , we must have that $Q(s) - e^s Q(-s) = 0$ for all $s \in \mathcal{S}$, and the statement follows. \square

Remark 4.4 The assumption that P is faithful can be omitted if one assumes in addition that Θ preserves $\text{supp}P$. If this is the case, one can replace Ω with $\text{supp}P$, and the above proof applies.

Exercise 4.11. Prove that the fluctuation relation implies (4.22).

Exercise 4.12. This exercise is devoted to a generalization of the fluctuation relation which has also found fundamental application in physics. Consider a family $\{P_X\}_{X \in \mathbb{R}^n}$ of probability measures on Ω indexed by vectors $X = (X_1, \dots, X_n) \in \mathbb{R}^n$. Set

$$\mathcal{E}_X(\omega) = \log \frac{P_X(\omega)}{P_X(\Theta_X(\omega))},$$

where Θ_X satisfies (4.21). Suppose that $\mathcal{E}_0 = 0$ and consider a decomposition

$$\mathcal{E}_X = \sum_{k=1}^n X_k \mathfrak{F}_{X,k}, \quad (4.24)$$

where the random variables $\mathfrak{F}_{X,k}$ satisfy

$$\mathfrak{F}_{X,k} \circ \Theta_X = -\mathfrak{F}_{X,k}. \quad (4.25)$$

We denote by Q_X the probability distribution of the vector random variable $(\mathfrak{F}_{X,1}, \dots, \mathfrak{F}_{X,n})$ with respect to P_X : for $s = (s_1, \dots, s_n) \in \mathbb{R}^n$,

$$Q_X(s) = P_X \{ \omega \in \Omega \mid \mathcal{F}_{X,1} = s_1, \dots, \mathcal{F}_{X,n} = s_n \}.$$

We also denote $\mathcal{S} = \{s \in \mathbb{R}^n \mid Q_X(s) \neq 0\}$ and, for $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, set

$$G(X, Y) = \sum_{s \in \mathcal{S}} e^{-\sum_k s_k Y_k} Q_X(s).$$

1. Prove that a decomposition (4.24) satisfying (4.25) always exists and that, except in trivial cases, is never unique.
2. Prove that $Q_X(s) \neq 0$ iff $Q_X(-s) \neq 0$.
3. Prove that

$$G(X, Y) = G(X, X - Y).$$

4. Prove that

$$Q_X(-s) = e^{-\sum_k s_k X_k} Q_X(s).$$

4.5 Jensen-Shannon entropy and metric

The Jensen-Shannon entropy of two probability measures $P, Q \in \mathcal{P}(\Omega)$ is

$$\begin{aligned} S_{JS}(P|Q) &= S(M(P, Q)) - \frac{1}{2}S(P) - \frac{1}{2}S(Q) \\ &= \frac{1}{2}(S(P|M(P, Q)) + S(Q|M(P, Q))), \end{aligned}$$

where

$$M(P, Q) = \frac{P + Q}{2}.$$

The Jensen-Shannon entropy can be viewed as a measure of concavity of the entropy. Obviously, $S_{JS}(P|Q) \geq 0$ with equality iff $P = Q$. In addition:

Proposition 4.14 (1)

$$S_{JS}(P|Q) \leq \log 2,$$

with equality iff $P \perp Q$.

(2)

$$\frac{1}{8}d_V(P, Q)^2 \leq S_{JS}(P|Q) \leq d_V(P, Q) \log \sqrt{2}.$$

The first inequality is saturated iff $P = Q$ and the second iff $P = Q$ or $P \perp Q$.

Proof. Part (1) follows from

$$\begin{aligned} S_{JS}(P|Q) &= \frac{1}{2} \sum_{\omega \in \Omega} \left(P(\omega) \log \left(\frac{2P(\omega)}{P(\omega) + Q(\omega)} \right) + Q(\omega) \log \left(\frac{2Q(\omega)}{P(\omega) + Q(\omega)} \right) \right) \\ &\leq \frac{1}{2} \sum_{\omega \in \Omega} (P(\omega) + Q(\omega)) \log 2 \\ &= \log 2. \end{aligned}$$

To prove (2), we start with the lower bound:

$$\begin{aligned} S_{JS}(P|Q) &= \frac{1}{2}S(P|M(P, Q)) + \frac{1}{2}S(Q|M(P, Q)) \\ &\geq \frac{1}{4}d_V(P, M(P, Q))^2 + \frac{1}{4}d_V(Q, M(P, Q))^2 \\ &= \frac{1}{8} \left(\sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| \right)^2 = \frac{1}{8}d_V(P|Q)^2, \end{aligned}$$

where the inequality follows from Theorem 4.2.

To prove the upper bound, set $S_+ = \{\omega \mid P(\omega) \geq Q(\omega)\}$, $S_- = \{\omega \mid P(\omega) < Q(\omega)\}$. Then

$$\begin{aligned} S_{JS}(P|Q) &= \frac{1}{2} \sum_{\omega \in S_+} \left(P(\omega) \log \left(\frac{2P(\omega)}{P(\omega) + Q(\omega)} \right) - Q(\omega) \log \left(\frac{P(\omega) + Q(\omega)}{2Q(\omega)} \right) \right) \\ &\quad + \frac{1}{2} \sum_{\omega \in S_-} \left(Q(\omega) \log \left(\frac{2Q(\omega)}{P(\omega) + Q(\omega)} \right) - P(\omega) \log \left(\frac{P(\omega) + Q(\omega)}{2P(\omega)} \right) \right) \\ &\leq \frac{1}{2} \sum_{\omega \in S_+} (P(\omega) - Q(\omega)) \log \left(\frac{2P(\omega)}{P(\omega) + Q(\omega)} \right) \\ &\quad + \frac{1}{2} \sum_{\omega \in S_-} (Q(\omega) - P(\omega)) \log \left(\frac{2Q(\omega)}{P(\omega) + Q(\omega)} \right) \\ &\leq \frac{1}{2} \sum_{\omega \in S_+} (P(\omega) - Q(\omega)) \log 2 + \frac{1}{2} \sum_{\omega \in S_-} (Q(\omega) - P(\omega)) \log 2 \\ &= d_V(P, Q) \log \sqrt{2}. \end{aligned}$$

In the first inequality we have used that for $P(\omega) \neq 0$ and $Q(\omega) \neq 0$,

$$\frac{P(\omega) + Q(\omega)}{2P(\omega)} \geq \frac{2Q(\omega)}{P(\omega) + Q(\omega)},$$

and the same inequality with P and Q interchanged.

The cases where equality holds in Parts (1) and (2) are easily identified from the above argument and we leave the formal proof as an exercise for the reader. \square

Set

$$d_{JS}(P, Q) = \sqrt{S_{JS}(P, Q)}.$$

Theorem 4.15 d_{JS} is a metric on $\mathcal{P}(\Omega)$.

Remark 4.5 If $|\Omega| \geq 2$, then S_{JS} is not a metric on $\mathcal{P}(\Omega)$. To see that, pick $\omega_1, \omega_2 \in \Omega$ and define $P, Q, R \in \mathcal{P}(\Omega)$ by $P(\omega_1) = 1, Q(\omega_2) = 1, R(\omega_1) = R(\omega_2) = \frac{1}{2}$. Then

$$S_{JS}(P|Q) = \log 2 > \frac{3}{2} \log \frac{4}{3} = S_{JS}(P|R) + S_{JS}(R|Q).$$

Remark 4.6 In the sequel we shall refer to d_{SJ} as the *Jensen-Shannon* metric.

Proof. Note that only the triangle inequality needs to be proved. Set $\mathbb{R}_+ =]0, \infty[$.

For $p, q \in \mathbb{R}_+$ let

$$L(p, q) = p \log \left(\frac{2p}{p+q} \right) + q \log \left(\frac{2q}{p+q} \right).$$

Since the function $F(x) = x \log x$ is strictly convex, writing

$$L(p, q) = (p+q) \left[\frac{1}{2} F \left(\frac{2p}{p+q} \right) + \frac{1}{2} F \left(\frac{2q}{p+q} \right) \right]$$

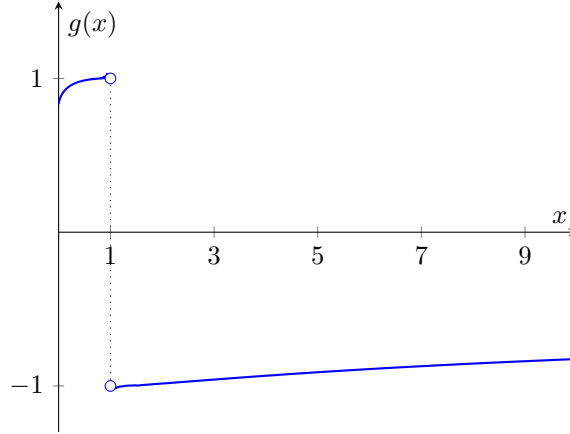
and applying the Jensen inequality to the expression in the brackets, we derive that $L(p, q) \geq 0$ with equality iff $p = q$. Our goal is to prove that for all $p, q, r \in \mathbb{R}_+$,

$$L(p, q) \leq \sqrt{L(p, r)} + \sqrt{L(r, q)}. \quad (4.26)$$

This yields the triangle inequality for d_{JS} as follows. If $P, Q, R \in \mathcal{P}_f(\Omega)$, (4.26) and Minkowski's inequality give

$$\begin{aligned} d_{JS}(P, Q) &= \left(\sum_{\omega \in \Omega} \sqrt{L(P(\omega), Q(\omega))^2} \right)^{\frac{1}{2}} \\ &\leq \left(\sum_{\omega \in \Omega} \left(\sqrt{L(P(\omega), R(\omega))} + \sqrt{L(R(\omega), Q(\omega))} \right)^2 \right)^{\frac{1}{2}} \\ &\leq \left(\sum_{\omega \in \Omega} \sqrt{L(P(\omega), R(\omega))^2} \right)^{\frac{1}{2}} + \left(\sum_{\omega \in \Omega} \sqrt{L(R(\omega), Q(\omega))^2} \right)^{\frac{1}{2}} \\ &= d_{JS}(P, R) + d_{JS}(R, Q). \end{aligned}$$

This yields the triangle inequality on $\mathcal{P}_f(\Omega)$. Since the map $(P, Q) \mapsto d_{JS}(P, Q)$ is continuous, the triangle inequality extends to $\mathcal{P}(\Omega)$.



The proof of (4.26) is an elaborate calculus exercise. The relation is obvious if $p = q$. Since $L(p, q) = L(q, p)$, it suffices to consider the case $p < q$. We fix such p and q and set

$$f(r) = \sqrt{L(p, r)} + \sqrt{L(r, q)}.$$

Then

$$f'(r) = \frac{1}{2\sqrt{L(p, r)}} \log\left(\frac{2r}{p+r}\right) + \frac{1}{2\sqrt{L(r, q)}} \log\left(\frac{2r}{r+q}\right).$$

Define $g : \mathbb{R}_+ \setminus \{1\} \rightarrow \mathbb{R}$ by

$$g(x) = \frac{1}{\sqrt{L(x, 1)}} \log\left(\frac{2}{x+1}\right),$$

One easily verifies that

$$f'(r) = \frac{1}{2\sqrt{r}} \left(g\left(\frac{p}{r}\right) + g\left(\frac{q}{r}\right) \right). \quad (4.27)$$

We shall need the following basic properties of g , clearly displayed in the above graph:

- (a) $g > 0$ on $]0, 1[$, $g < 0$ on $]1, \infty[$.
- (b) $\lim_{x \uparrow 1} g(x) = 1$, $\lim_{x \downarrow 1} g(x) = -1$. This follows from $\lim_{x \rightarrow 1} [g(x)]^2 = 1$, which can be established by applying l'Hopital's rule twice.
- (c) $g'(x) > 0$ for $x \in \mathbb{R}_+ \setminus \{1\}$. To prove this one computes

$$g'(x) = -\frac{h(x)}{(x+1)L(x, 1)^{3/2}},$$

where

$$h(x) = 2x \log\left(\frac{2x}{x+1}\right) + 2 \log\left(\frac{2}{x+1}\right) + (x+1) \log\left(\frac{2x}{x+1}\right) \log\left(\frac{2}{x+1}\right).$$

One further computes

$$h'(x) = \log\left(\frac{2x}{x+1}\right) \log\left(\frac{2}{x+1}\right) + \log\left(\frac{2x}{x+1}\right) + \frac{1}{x} \log\left(\frac{2}{x+1}\right),$$

$$h''(x) = -\frac{1}{x+1} \log\left(\frac{2x}{x+1}\right) - \frac{1}{x^2(x+1)} \log\left(\frac{2}{x+1}\right).$$

Note that $h(1) = h'(1) = h''(1) = 0$. The inequality $\log t \geq (t-1)/t$, which holds for all $t > 0$, gives

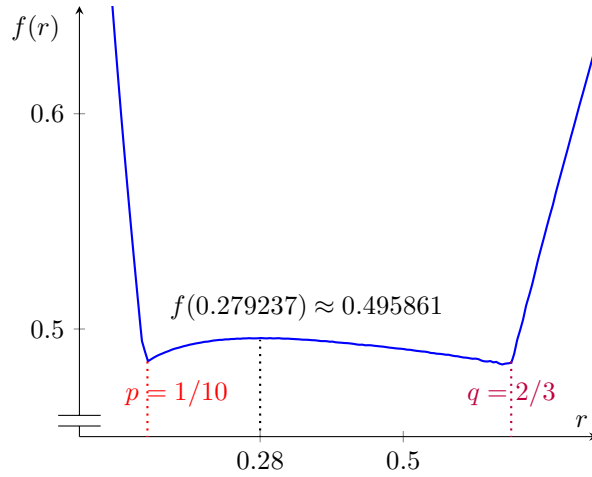
$$h''(x) \leq -\frac{1}{x+1} \left(1 - \frac{x+1}{2x}\right) - \frac{1}{x^2(x+1)} \left(1 - \frac{x+1}{2}\right) = -\frac{(x-1)^2}{2x^2(x+1)}.$$

Hence $h''(x) < 0$ for $x \in \mathbb{R}_+ \setminus \{1\}$, and the statement follows.

(d) Note that (a), (b) and (c) give that $0 < g(x) < 1$ on $]0, 1[$ and $-1 < g(x) < 0$ on $]1, \infty[$.

It follows from (a) that $f'(r) < 0$ for $r \in]0, p[$, $f'(r) > 0$ for $r > q$, and so $f(r)$ is decreasing on $]0, p[$ and increasing on $]q, \infty[$. Hence, for $r < p$ and $r > q$, $f(r) > f(p)$, which gives (4.26) for those r 's. To deal with the case $p < r < q$, set $m(r) = g(p/r) + g(q/r)$. It follows from (b) that $m'(r) < 0$ for $p < r < q$, while (b) and (d) give $m(p+) = 1 + g(q/p) > 0$, $m(q-) = -1 + g(p/q) < 0$. Hence $f'(r)$ has precisely one zero r_m in the interval $]p, q[$. Since $f'(p+) > 0$, $f'(q-) > 0$, $f(r)$ is increasing in $]p, r_m[$ and decreasing on $]r_m, q[$. On the first interval, $f(r) \geq f(p)$, and on the second interval $f(r) \geq f(q)$, which gives that (4.26) also holds for $p < r < q$. \square

The graph of $r \mapsto f(r)$ is plotted below for $p = \frac{1}{10}$ and $q = \frac{2}{3}$. In this case $r_m \approx 0.28$.



4.6 Rényi's relative entropy

Let Ω be a finite set and $P, Q \in \mathcal{P}(\Omega)$. For $\alpha \in]0, 1[$ we set

$$S_\alpha(P|Q) = \frac{1}{\alpha-1} \log \left(\sum_{\omega \in \Omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} \right).$$

$S_\alpha(P|Q)$ is called Rényi's relative entropy of P with respect to Q . Note that

$$S_\alpha(P|P_{\text{ch}}) = S_\alpha(P) + \log |\Omega|.$$

Proposition 4.16 (1) $S_\alpha(P|Q) \geq 0$.

(2) $S_\alpha(P|Q) = \infty$ iff $P \perp Q$ and $S_\alpha(P|Q) = 0$ iff $P = Q$.

(3)

$$S_\alpha(P|Q) = \frac{\alpha}{1-\alpha} S_{1-\alpha}(Q|P).$$

(4)

$$\lim_{\alpha \uparrow 1} S_\alpha(P|Q) = S(P|Q).$$

(5) Suppose that $P \not\ll Q$. Then the function $]0, 1[\ni \alpha \mapsto S_\alpha(P|Q)$ is strictly increasing(6) The map $(P, Q) \mapsto S_\alpha(P|Q) \in [0, \infty]$ is continuous and jointly convex.(7) Let $\Phi : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\hat{\Omega})$ be a stochastic map. Then for all $P, Q \in \mathcal{P}(\Omega)$,

$$S_\alpha(\Phi(P)|\Phi(Q)) \leq S_\alpha(P|Q).$$

(8) If $S(P|Q) < \infty$, then $\alpha \mapsto S_\alpha(P|Q)$ extends to a real-analytic function on \mathbb{R} .**Proof.** Obviously, $S_\alpha(P|Q) = \infty$ iff $P \perp Q$. In what follows, if $P \not\ll Q$, we set

$$T = \text{supp } P \cap \text{supp } Q.$$

An application of Jensen's inequality gives

$$\begin{aligned} \sum_{\omega \in \Omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} &= Q(T) \sum_{\omega \in T} \left(\frac{P(\omega)}{Q(\omega)} \right)^\alpha \frac{Q(\omega)}{Q(T)} \\ &\leq Q(T) \left(\sum_{\omega \in T} \frac{P(\omega)}{Q(\omega)} \frac{Q(\omega)}{Q(T)} \right)^\alpha \\ &= Q(T)^{1-\alpha} P(T)^\alpha. \end{aligned}$$

Hence, $\sum_{\omega \in \Omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} \leq 1$ with the equality iff $P = Q$, and Parts (1), (2) follow.

Part (3) is obvious. To prove (4), note that

$$\lim_{\alpha \uparrow 1} \sum_{\omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} = P(T),$$

and that $P(T) = 1$ iff $P \ll Q$. Hence, if P is not absolutely continuous with respect to Q , then $\lim_{\alpha \uparrow 1} S_\alpha(P|Q) = \infty = S(P|Q)$. If $P \ll Q$, an application of L'Hopital rule gives $\lim_{\alpha \uparrow 1} S_\alpha(P|Q) = S(P|Q)$.

To prove (5), set

$$F(\alpha) = \log \left(\sum_{\omega \in \Omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} \right),$$

and note that $\mathbb{R} \ni \alpha \mapsto F(\alpha)$ is a real-analytic strictly convex function satisfying $F(0) \leq 0$, $F(1) \leq 0$. We have

$$\frac{dS_\alpha(P|Q)}{d\alpha} = \frac{F'(\alpha)(\alpha - 1) - (F(\alpha) - F(1))}{(\alpha - 1)^2} - \frac{F(1)}{(\alpha - 1)^2}.$$

By the mean-value theorem, $F(\alpha) - F(1) = (\alpha - 1)F'(\zeta_\alpha)$ for some $\zeta_\alpha \in]\alpha, 1[$. Since F' is strictly increasing, $F'(\alpha) < F'(\zeta_\alpha)$ and

$$\frac{dS_\alpha(P|Q)}{d\alpha} > 0$$

for $\alpha \in]0, 1[$.The continuity part of (6) are obvious. The proof of the joint convexity is the same as the proof of Proposition 4.4 (one now takes $g(t) = t^\alpha$) and is left as an exercise for the reader.

We now turn to Part (7). First, we have

$$[\Phi(P)(\hat{\omega})]^\alpha [\Phi(Q)(\hat{\omega})]^{1-\alpha} \geq \sum_{\omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} \Phi(\omega, \hat{\omega}).$$

This inequality is obvious if the r.h.s. is equal to zero. Otherwise, let

$$R = \{\omega \mid P(\omega)Q(\omega)\Phi(\omega, \hat{\omega}) > 0\}.$$

Then

$$\begin{aligned} [\Phi(P)(\hat{\omega})]^\alpha [\Phi(Q)(\hat{\omega})]^{1-\alpha} &\geq \left(\sum_{\omega \in R} P(\omega)\Phi(\omega, \hat{\omega}) \right)^\alpha \left(\sum_{\omega \in R} Q(\omega)\Phi(\omega, \hat{\omega}) \right)^{1-\alpha} \\ &= \left(\frac{\sum_{\omega \in R} P(\omega)\Phi(\omega, \hat{\omega})}{\sum_{\omega \in R} Q(\omega)\Phi(\omega, \hat{\omega})} \right)^\alpha \sum_{\omega \in R} Q(\omega)\Phi(\omega, \hat{\omega}) \\ &\geq \sum_{\omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} \Phi(\omega, \hat{\omega}), \end{aligned}$$

where in the last step we have used the joint concavity of the function $(x, y) \mapsto x(y/x)^\alpha$ (recall proof of Proposition 4.4). Hence,

$$\sum_{\hat{\omega}} [\Phi(P)(\hat{\omega})]^\alpha [\Phi(Q)(\hat{\omega})]^{1-\alpha} \geq \sum_{\hat{\omega}} \sum_{\omega} P(\omega)^\alpha Q(\omega)^{1-\alpha} \Phi(\omega, \hat{\omega}) = \sum_{\omega} P(\omega)^\alpha Q(\omega)^{1-\alpha},$$

and Part (7) follows.

It remains to prove Part (8). For $\alpha \in \mathbb{R} \setminus \{1\}$ set

$$\mathfrak{S}_\alpha(P|Q) = \frac{1}{\alpha - 1} \log \left(\sum_{\omega \in T} P(\omega)^\alpha Q(\omega)^{1-\alpha} \right).$$

Obviously, $\alpha \mapsto \mathfrak{S}_\alpha(P|Q)$ is real-analytic on $\mathbb{R} \setminus \{1\}$. Since

$$\lim_{\alpha \uparrow 1} \mathfrak{S}_\alpha(P|Q) = \lim_{\alpha \downarrow 1} \mathfrak{S}_\alpha(P|Q) = S(P|Q),$$

$\alpha \mapsto \mathfrak{S}_\alpha(P|Q)$ extends to a real-analytic function on \mathbb{R} with $\mathfrak{S}_1(P|Q) = S(P|Q)$. Finally, Part (8) follows from the observation that $S_\alpha(P|Q) = \mathfrak{S}_\alpha(P|Q)$ for $\alpha \in]0, 1[$. □

Following on the discussion at the end of Section 3.6, we set

$$\widehat{S}_\alpha(P|Q) = \log \left(\sum_{\omega \in T} P(\omega)^\alpha Q(\omega)^{1-\alpha} \right), \quad \alpha \in \mathbb{R}.$$

If $P \ll Q$, then

$$\widehat{S}_\alpha(P|Q) = \log \mathbb{E}_Q(e^{\alpha S_{P|Q}}), \quad (4.28)$$

and so $\widehat{S}_\alpha(P|Q)$ is the cumulant generating function for the relative entropy function $S_{P|Q}$ defined on the probability space (T, P) . The discussion at the end of 3.6 can be now repeated verbatim (we will return to this point in Section 5.1). Whenever there is no danger of the confusion, we shall also call $\widehat{S}_\alpha(P|Q)$ Rényi's relative entropy of the pair (P, Q) . Note that

$$\widehat{S}_\alpha(P_{\text{ch}}|P) = \widehat{S}_\alpha(P) - \alpha \log |\Omega|. \quad (4.29)$$

Some care is needed in transposing the properties listed in Proposition 4.16 to $\widehat{S}_\alpha(P|Q)$. This point is discussed in the Exercise 4.14.

Exercise 4.13.

1. Describe the subset of $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$ on which the function $(P, Q) \mapsto S_\alpha(P|Q)$ is strictly convex.
2. Describe the subset of $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$ on which $S_\alpha(\Phi(P)|\Phi(Q)) < S_\alpha(P|Q)$.
3. Redo the Exercise 4.2 in Section 4.1 and reprove Proposition 4.7 following the proofs of Parts (7) and (8) of Proposition 4.16. Describe the subset of $\mathcal{P}(\Omega)$ on which

$$S(\Phi(P)|\Phi(Q)) < S(P|Q).$$

Exercise 4.14. Prove the following properties of $\widehat{S}_\alpha(P|Q)$.

1. $\widehat{S}_\alpha(P|Q) = -\infty$ iff $P \perp Q$.

In the remaining statements we shall suppose that $P \not\perp Q$.

2. The function $\mathbb{R} \ni \alpha \mapsto \widehat{S}_\alpha(P|Q)$ is real-analytic and convex. This function is trivial (i.e., identically equal to zero) iff $P = Q$. If P/Q not constant on $T = \text{supp}P \cap \text{supp}Q$, then the function $\alpha \mapsto \widehat{S}_\alpha(P|Q)$ is strictly convex.

3. If $Q \ll P$, then

$$\left. \frac{d\widehat{S}_\alpha(P|Q)}{d\alpha} \right|_{\alpha=0} = -S(Q|P).$$

If $P \ll Q$, then

$$\left. \frac{d\widehat{S}_\alpha(P|Q)}{d\alpha} \right|_{\alpha=1} = S(P|Q).$$

4. If P and Q are mutually absolutely continuous, then $\widehat{S}_0(P|Q) = \widehat{S}_1(P|Q) = 0$, $\widehat{S}_\alpha(P|Q) \leq 0$ for $\alpha \in [0, 1]$, and $\widehat{S}_\alpha(P|Q) \geq 0$ for $\alpha \notin [0, 1]$. Moreover,

$$\widehat{S}_\alpha(P|Q) \geq \max\{-\alpha S(Q|P), (\alpha - 1)S(P|Q)\}.$$

5. For $\alpha \in]0, 1[$ the function $(P, Q) \mapsto \widehat{S}_\alpha(P|Q)$ is continuous and jointly concave. Moreover, for any stochastic matrix Φ ,

$$\widehat{S}_\alpha(\Phi(P)|\Phi(Q)) \geq \widehat{S}_\alpha(P|Q).$$

Exercise 4.15. Prove that the fluctuation relation of Section 4.4 is equivalent to the following statement: for all $\alpha \in \mathbb{R}$,

$$\widehat{S}_\alpha(P|P_\Theta) = \widehat{S}_{1-\alpha}(P|P_\Theta).$$

4.7 Hypothesis testing

Let Ω be a finite set and P, Q two distinct probability measures on Ω . We shall assume that P and Q are faithful.

Suppose that we know a priori that a probabilistic experiment is with probability p described by P and with probability $1-p$ by Q . By performing an experiment we wish to decide with minimal error probability what is the correct probability measure. For example, suppose that we are given two coins, one fair ($P(\text{Head}) = P(\text{Tail}) = 1/2$) and one unfair ($Q(\text{Head}) = s, Q(\text{Tail}) = 1-s, s > 1/2$). We pick coin randomly (hence $p = 1/2$). The experiment is a coin toss. After tossing a coin we wish to decide with minimal error probability whether we picked the fair or the unfair coin. The correct choice of obvious: if the outcome is Head, pick Q , if the outcome is Tail, pick P .

The following procedure is known as *hypothesis testing*. A *test* T is a subset of Ω . On the basis of the outcome of the experiment with respect to T one chooses between P or Q . More precisely, if the outcome of the experiment is in T , one chooses Q (Hypothesis I: Q is correct) and if the outcome is not in T , one chooses P (Hypothesis II: P is correct). $P(T)$ is the conditional error probability of accepting I if II is true and $Q(T^c)$ is the conditional error probability of accepting II if I is true. The average error probability is

$$D_p(P, Q, T) = pP(T) + (1-p)Q(T^c),$$

and we are interested in minimizing $D_p(P, Q, T)$ w.r.t. T . Let

$$D_p(P, Q) = \inf_T D_p(P, Q, T).$$

The Bayesian distinguishability problem is to identify tests T such that $D_p(P, Q, T) = D_p(P, Q)$. Let

$$T_{\text{opt}} = \{\omega \mid pP(\omega) \leq (1-p)Q(\omega)\}.$$

Proposition 4.17 (1) T_{opt} is a minimizer of the function $T \mapsto D_p(P, Q, T)$. If T is another minimizer, then $T \subset T_{\text{opt}}$ and $pP(\omega) = (1-p)Q(\omega)$ for $\omega \in T_{\text{opt}} \setminus T$.

(2)

$$D_p(P, Q) = \int_{\Omega} \min\{1-p, p\Delta_{P|Q}(\omega)\}dQ.$$

(3) For $\alpha \in]0, 1[$,

$$D_p(P, Q) \leq p^\alpha(1-p)^{1-\alpha}e^{\widehat{S}_\alpha(P|Q)}.$$

(4)

$$D_p(P, Q) \geq \int_{\Omega} \frac{p\Delta_{P|Q}}{1 + \frac{p}{1-p}\Delta_{P|Q}}dQ.$$

Remark 4.7 Part (1) of this proposition is called Neyman-Pearson lemma. Part (3) is called Chernoff bound.

Proof.

$$D_p(P, Q, T) = 1-p - \sum_{\omega \in T} ((1-p)Q(\omega) - pP(\omega)) \geq 1-p - \sum_{\omega \in T_{\text{opt}}} ((1-p)Q(\omega) - pP(\omega)),$$

and Part (1) follows. Part (2) is a straightforward computation. Part (3) follows from (2) and the bound $\min\{x, y\} \leq x^\alpha y^{1-\alpha}$ that holds for $x, y \geq 0$ and $\alpha \in]0, 1[$. Part (4) follows from (2) and the obvious estimate

$$\min\{1-p, p\Delta_{P|Q}(\omega)\} \geq \frac{p\Delta_{P|Q}}{1 + \frac{p}{1-p}\Delta_{P|Q}}.$$

□

Obviously, the errors are smaller if the hypothesis testing is based on repeated experiments. Let P_N and Q_N be the respective product probability measures on Ω^N .

Theorem 4.18

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log D_p(P_N, Q_N) = \min_{\alpha \in [0,1]} \widehat{S}_\alpha(P|Q).$$

Proof. By Part (2) of the last proposition, for any $\alpha \in]0, 1[$,

$$D_p(P_N, Q_N) \leq p^\alpha (1-p)^{1-\alpha} e^{\widehat{S}_\alpha(P_N|Q_N)} = p^\alpha (1-p)^{1-\alpha} e^{N \widehat{S}_\alpha(P|Q)},$$

and so

$$\frac{1}{N} \log D_p(P_N, Q_N) \leq \min_{\alpha \in [0,1]} \widehat{S}_\alpha(P|Q).$$

This yields the upper bound:

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log D_p(P_N|Q_N) \leq \min_{\alpha \in [0,1]} \widehat{S}_\alpha(P|Q).$$

To prove the lower bound we shall make use of the lower bound in Cramér's theorem (Corollary 2.11). Note first that the function

$$x \mapsto \frac{px}{1 + \frac{p}{1-p}x}$$

is increasing on \mathbb{R}_+ . Let $\theta > 0$ be given. By Part (4) of the last proposition,

$$D_p(P_N, Q_N) \geq \frac{pe^{N\theta}}{1 + \frac{p}{1-p}e^{N\theta}} Q_N \{ \omega \in \Omega^N \mid \Delta_{P_N|Q_N}(\omega) \geq e^{N\theta} \}.$$

Hence,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log D_p(P_N|Q_N) \geq \liminf_{N \rightarrow \infty} \frac{1}{N} \log Q_N \{ \omega \in \Omega^N \mid \log \Delta_{P_N|Q_N}(\omega) \geq N\theta \}. \quad (4.30)$$

Let $X = \log \Delta_{P|Q}$ and $\mathcal{S}_N(\omega) = \sum_{k=1}^N X(\omega_k)$. Note that $\mathcal{S}_N = \log \Delta_{P_N|Q_N}$. The cumulant generating function of X w.r.t. Q is

$$\log \mathbb{E}_Q(e^{\alpha X}) = \widehat{S}_\alpha(P|Q).$$

Since $\mathbb{E}_Q(X) = -S(Q|P) < 0$ and $\theta > 0$, it follows from Corollary 2.11 that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log Q_N \{ \omega \in \Omega^N \mid \log \Delta_{P_N|Q_N}(\omega) \geq N\theta \} \geq -I(\theta) \quad (4.31)$$

Since

$$\frac{d\widehat{S}_\alpha}{d\alpha} \Big|_{\alpha=0} = -S(Q|P) < 0, \quad \frac{d\widehat{S}_\alpha}{d\alpha} \Big|_{\alpha=1} = S(P|Q) > 0,$$

the rate function $I(\theta)$ is continuous around zero, and it follows from (4.30) and (4.31) that

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log D_p(P_N|Q_N) \geq -I(0) = -\sup_{\alpha \in \mathbb{R}} (-\widehat{S}_\alpha(P|Q)).$$

Since $\widehat{S}_\alpha(P|Q) \leq 0$ for $\alpha \in [0, 1]$ and $\widehat{S}_\alpha(P|Q) \geq 0$ for $\alpha \notin [0, 1]$,

$$-\sup_{\alpha \in \mathbb{R}} (-\widehat{S}_\alpha(P|Q)) = \min_{\alpha \in [0,1]} \widehat{S}_\alpha(P|Q),$$

and the lower bound follows:

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log D_p(P_N|Q_N) \geq \min_{\alpha \in [0,1]} \widehat{S}_\alpha(P|Q).$$

□

4.8 Asymmetric hypothesis testing

We continue with the framework and notation of the previous section. The asymmetric hypothesis testing concerns individual error probabilities $P_N(T_N)$ (type I-error) and $Q_N(T_N^c)$ (type II-error). For $\gamma \in]0, 1[$ the Stein error exponents are defined by

$$s_N(\gamma) = \min \{ P(T_N) \mid T_N \subset \Omega^N, Q(T_N^c) \leq \gamma \}.$$

Theorem 4.12 gives

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log s_N(\gamma) = -S(Q|P).$$

The Hoeffding error exponents are similar to Stein's exponents, but with a tighter constraint on the family $(T_N)_{N \geq 1}$ of tests which are required to ensure exponential decay of type-II errors with a minimal rate $s > 0$. They are defined as

$$\begin{aligned} \bar{h}(s) &= \inf_{(T_N)} \left\{ \limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \mid \limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) \leq -s \right\}, \\ \underline{h}(s) &= \inf_{(T_N)} \left\{ \liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \mid \limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) \leq -s \right\}, \\ h(s) &= \inf_{(T_N)} \left\{ \lim_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \mid \limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) \leq -s \right\}, \end{aligned}$$

where in the last case the infimum is taken over all sequences of tests $(T_N)_{N \geq 1}$ for which the limit

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N)$$

exists. The analysis of these exponents is centred around the function

$$\psi(s) = \inf_{\alpha \in [0, 1[} \frac{s\alpha + \widehat{S}_\alpha(Q|P)}{1 - \alpha}, \quad s \geq 0.$$

We first describe some basic properties of ψ .

Proposition 4.19 (1) ψ is continuous on $[0, \infty[$, $\psi(0) = -S(Q|P)$ and $\psi(s) = 0$ for $s \geq S(P|Q)$.

(2) ψ is strictly increasing and strictly concave on $[0, S(P|Q)]$, and real analytic on $]0, S(P|Q)[$.

(3)

$$\lim_{s \downarrow 0} \psi'(s) = \infty, \quad \lim_{s \uparrow S(P|Q)} \psi'(s) = \left[\widehat{S}''_\alpha(Q|P) \Big|_{\alpha=0} \right]^{-1}.$$

(4) For $\theta \in \mathbb{R}$ set

$$\varphi(\theta) = \sup_{\alpha \in [0, 1]} \left(\theta\alpha - \widehat{S}_\alpha(Q|P) \right), \quad \hat{\varphi}(\theta) = \varphi(\theta) - \theta.$$

Then for all $s \geq 0$,

$$\psi(s) = -\varphi(\hat{\varphi}^{-1}(s)). \tag{4.32}$$

Proof. Throughout the proof we shall often use Part 3 of the Exercise 4.14.

We shall prove Parts (1)-(3) simultaneously. Set

$$F(\alpha) = \frac{s\alpha + \widehat{S}_\alpha(Q|P)}{1 - \alpha}.$$

Then

$$F'(\alpha) = \frac{G(\alpha)}{(1-\alpha)^2},$$

where $G(\alpha) = s + \widehat{S}_\alpha(Q|P) + (1-\alpha)\widehat{S}'_\alpha(Q|P)$. Furthermore, $G'(\alpha) = (1-\alpha)\widehat{S}''_\alpha(Q|P)$ and so $G'(\alpha) > 0$ for $\alpha \in]0, 1[$. Note that $G(0) = s - S(P|Q)$ and $G(1) = s$. It follows that if $s = 0$, then $G(\alpha) < 0$ for $\alpha \in]0, 1[$ and $F(\alpha)$ is decreasing on $]0, 1[$. Hence,

$$\psi(0) = \lim_{\alpha \rightarrow 1} \frac{\widehat{S}_\alpha(Q|P)}{1-\alpha} = -S(Q|P).$$

On the other hand, if $0 < s < S(P|Q)$, then $G(0) < 0$, $G(1) > 0$, and so there exists unique $\alpha_*(s) \in]0, 1[$ such that

$$G(\alpha_*(s)) = 0. \quad (4.33)$$

In this case,

$$\psi(s) = \frac{s\alpha_*(s) + \widehat{S}_{\alpha_*(s)}(Q|P)}{1-\alpha_*(s)} = -s - \widehat{S}'_{\alpha_*(s)}(Q|P). \quad (4.34)$$

If $s \geq S(P|Q)$, then $G(\alpha) \geq 0$ for $\alpha \in]0, 1[$, and $\psi(s) = F(0) = 0$. The analytic implicit function theorem yields that $s \mapsto \alpha_*(s)$ is analytic on $]0, S(P|Q)[$, and so ψ is real-analytic on $]0, S(P|Q)[$. The identity

$$0 = G(\alpha_*(s)) = s + \widehat{S}_{\alpha_*(s)}(Q|P) + (1-\alpha_*(s))\widehat{S}'_{\alpha_*(s)}(Q|P), \quad (4.35)$$

which holds for $s \in]0, S(P|Q)[$, gives that

$$\alpha'_*(s) = -\frac{1}{(1-\alpha_*(s))G'(\alpha_*(s))}, \quad (4.36)$$

and so $\alpha'_*(s) < 0$ for $s \in]0, S(P|Q)[$. One computes

$$\psi'(s) = \frac{\alpha_*(s) - s\alpha'_*(s)}{(1-\alpha_*(s))^2}, \quad (4.37)$$

and so ψ is strictly increasing on $]0, S(P|Q)[$ and hence on $[0, S(P|Q)]$. Since $\alpha_*(s)$ is strictly decreasing on $]0, S(P|Q)[$, the limits

$$\lim_{s \downarrow 0} \alpha_*(s) = x, \quad \lim_{s \uparrow S(P|Q)} \alpha_*(s) = y,$$

exist. Obviously, $x, y \in [0, 1]$, $x > y$, and the definition of G and α_* give that

$$\widehat{S}_x(Q|P) + (1-x)\widehat{S}'_x(Q|P) = 0, \quad S(P|Q) + \widehat{S}_y(Q|P) + (1-y)\widehat{S}'_y(Q|P) = 0. \quad (4.38)$$

We proceed to show that $x = 1$ and $y = 0$. Suppose that $x < 1$. The mean value theorem gives that for some $z \in]x, 1[$

$$-\widehat{S}_x(Q|P) = \widehat{S}_1(Q|P) - \widehat{S}_x(Q|P) = (1-x)\widehat{S}'_z(Q|P) > (1-x)\widehat{S}'_z(Q|P), \quad (4.39)$$

where we used that $\alpha \mapsto \widehat{S}'_\alpha(Q|P)$ is strictly increasing. Obviously, (4.39) contradicts the first equality in (4.38), and so $x = 1$. Similarly, if $y > 0$,

$$\begin{aligned} S(P|Q) + \widehat{S}_y(Q|P) + (1-y)\widehat{S}'_y(Q|P) &> S(P|Q) + \widehat{S}_y(Q|P) + (1-y)\widehat{S}'_0(Q|P) \\ &= \widehat{S}_y(Q|P) - y\widehat{S}'_0(Q|P) > 0, \end{aligned}$$

contradicting the second equality in (4.38). Since $x = 1$ and $y = 0$, (4.36) and (4.37) yield Part (3). Finally, to prove that ψ is strictly concave on $[0, S(P|Q)]$ (in view of real analyticity of ψ on $]0, S(P|Q)[$), it suffices to show that ψ' is not constant on $]0, S(P|Q)[$. That follows from Part (3), and the proofs of Parts (1)-(3) are complete.

We now turn to Part (4). The following basic properties of the "restricted Legendre transform" φ are easily proven following the arguments in Section 2.5 and we leave the details as an exercise for the reader: φ is continuous, non-negative and convex on \mathbb{R} , $\varphi(\theta) = 0$ for $\theta \leq -S(P|Q)$, φ is real analytic, strictly increasing and strictly convex on $] -S(P|Q), S(Q|P)[$, and $\varphi(\theta) = \theta$ for $\theta \geq S(Q|P)$. The properties of $\hat{\varphi}$ are now deduced from those of φ and we mention the following: $\hat{\varphi}$ is convex, continuous and decreasing, $\hat{\varphi}(\theta) = \theta$ for $\theta \leq -S(P|Q)$, and $\hat{\varphi}(\theta) = 0$ for $\theta \geq S(Q|P)$. Moreover, the map $\hat{\varphi} :] -\infty, S(Q|P)] \rightarrow [0, \infty[$ is a bijection, and we denote by $\hat{\varphi}^{-1}$ its inverse. For $s \geq S(P|Q)$, $\hat{\varphi}^{-1}(s) = -s$ and $\varphi(-s) = 0$, and so (4.32) holds for $s \geq S(P|Q)$. Since $\hat{\varphi}^{-1}(0) = S(Q|P)$ and $\varphi(S(Q|P)) = S(Q|P)$, (4.32) also holds for $s = 0$.

It remains to consider the case $s \in]0, S(P|Q)[$. The map $\hat{\varphi} :] -S(P|Q), S(Q|P)[\rightarrow]0, S(P|Q)[$ is a strictly decreasing bijection. Since

$$-\varphi(\hat{\varphi}^{-1}(s)) = -s - \hat{\varphi}^{-1}(s),$$

it follows from (4.34) that it suffices to show that

$$\hat{\varphi}^{-1}(s) = \hat{S}'_{\alpha_*(s)}(Q|P),$$

or equivalently, that

$$\varphi(\hat{S}'_{\alpha_*(s)}(Q|P)) = -s - \hat{S}'_{\alpha_*(s)}(Q|P). \quad (4.40)$$

Since on $] -S(P|Q), S(Q|P)[$ the function φ coincides with the Legendre transform of $\hat{S}_\alpha(P|Q)$, it follows from Part (1) of Proposition 2.5 that

$$\varphi(\hat{S}'_{\alpha_*(s)}(Q|P)) = \alpha_*(s)\hat{S}'_{\alpha_*(s)}(Q|P) - \hat{S}_{\alpha_*(s)}(Q|P),$$

and (4.40) follows from (4.35). \square

Exercise 4.16. Prove the properties of φ and $\hat{\varphi}$ that were stated and used in the proof of Part (4) of Proposition 4.19.

The next result sheds additional light on the function ψ . For $\alpha \in [0, 1]$ we define $R_\alpha \in \mathcal{P}(\Omega)$ by

$$R_\alpha(\omega) = \frac{Q(\omega)^\alpha P(\omega)^{1-\alpha}}{\sum_{\omega'} Q(\omega')^\alpha P(\omega')^{1-\alpha}}.$$

Proposition 4.20 (1) For all $s \geq 0$,

$$\psi(s) = -\inf \{S(R|P) \mid R \in \mathcal{P}(\Omega), S(R|Q) \leq s\}. \quad (4.41)$$

(2) For any $s \in]0, S(P|Q)[$,

$$S(R_{\alpha_*(s)}|Q) = s, \quad S(R_{\alpha_*(s)}|P) = -\psi(s),$$

where $\alpha_*(s)$ is given by (4.33).

Proof. Denote by $\phi(s)$ the r.h.s. in (4.41). Obviously, $\phi(0) = -S(Q|P)$ and $\phi(s) = 0$ for $s \geq S(P|Q)$. So we need to prove that $\psi(s) = \phi(s)$ for $s \in]0, S(P|Q)[$.

For any $R \in \mathcal{P}(\Omega)$ and $\alpha \in [0, 1]$,

$$S(R|R_\alpha) = \alpha S(R|Q) + (1 - \alpha)S(R|P) + \hat{S}_\alpha(Q|P).$$

If R is such that $S(R|Q) \leq s$ and $\alpha \in]0, 1[$, then

$$\frac{S(R|R_\alpha)}{1-\alpha} \leq \frac{\alpha s + \widehat{S}_\alpha(Q|P)}{1-\alpha} + S(R|P).$$

Since $S(R|R_\alpha) \geq 0$,

$$\inf_{\alpha \in]0, 1[} \frac{\alpha s + \widehat{S}_\alpha(Q|P)}{1-\alpha} + S(R|P) \geq 0.$$

This gives that $\phi(s) \leq \psi(s)$. If Part (2) holds, then also $\phi(s) \geq \psi(s)$ for all $s \in]0, S(P|Q)[$, and we have the equality $\phi = \psi$. To prove Part (2), a simple computation gives

$$S(R_\alpha|Q) = -(1-\alpha)\widehat{S}'_\alpha(Q|P) - \widehat{S}_\alpha(Q|P), \quad S(R_\alpha|Q) = S(R_\alpha|P) + \widehat{S}'_\alpha(Q|P).$$

After setting $\alpha = \alpha_*(s)$ in these equalities, Part (2) follows from (4.35) and (4.34). \square

The main result of this section is

Theorem 4.21 For all $s > 0$,

$$\overline{h}(s) = \underline{h}(s) = h(s) = \psi(s). \quad (4.42)$$

Proof. Note that the functions \overline{h} , \underline{h} , h are non-negative and increasing on $]0, \infty[$ and that

$$\underline{h}(s) \leq \overline{h}(s) \leq h(s) \quad (4.43)$$

for all $s > 0$.

We shall prove that for all $s \in]0, S(P|Q)[$,

$$h(s) \leq \psi(s), \quad \underline{h}(s) \geq \psi(s). \quad (4.44)$$

In view of (4.43), that proves (4.42) for $s \in]0, S(P|Q)[$. Assuming that (4.44) holds, the relations $h(s) \leq h(S(P|Q)) \leq 0$ for $s \in]0, S(P|Q)[$ and

$$\lim_{s \uparrow S(P|Q)} h(s) = \lim_{s \uparrow S(P|Q)} \psi(s) = 0$$

give that $h(S(P|Q)) = 0$. Since h is increasing, $h(s) = 0$ for $s \geq S(P|Q)$ and so $h(s) = \psi(s)$ for $s \geq S(P|Q)$. In the same way one shows that $\overline{h}(s) = \underline{h}(s) = \psi(s)$ for $s \geq S(P|Q)$.

We now prove the first inequality in (4.44). Recall that the map $\hat{\varphi} :]-S(P|Q), S(Q|P)[\rightarrow]0, S(P|Q)[$ is a bijection. Fix $s \in]0, S(P|Q)[$ and let $\theta \in]-S(P|Q), S(Q|P)[$ be such that $\hat{\varphi}(\theta) = s$. Let

$$T_N(\theta) = \{\omega \in \Omega^N \mid Q_N(\omega) \geq e^{N\theta} P_N(\omega)\}. \quad (4.45)$$

Then

$$P_N(T_N(\theta)) = P_N \left\{ \omega = (\omega_1, \dots, \omega_N) \in \Omega^N \mid \frac{1}{N} \sum_{j=1}^N S_{Q|P}(\omega_j) \geq \theta \right\}.$$

Since the cumulant generating function for $S_{Q|P}$ with respect to P is $\widehat{S}_\alpha(Q|P)$, and the rate function I for $S_{Q|P}$ with respect to P coincides with φ on $]S(P|Q), S(Q|P)[$, it follows from Part (1) of Corollary 2.11 that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N(\theta)) = -\varphi(\theta). \quad (4.46)$$

Similarly,

$$Q_N([T_N(\theta)]^c) = Q_N \left\{ \omega = (\omega_1, \dots, \omega_N) \in \Omega^N \mid \frac{1}{N} \sum_{j=1}^N S_{Q|P}(\omega_j) < \theta \right\}.$$

The cumulant generating function for $S_{Q|P}$ with respect to Q is $\widehat{S}_{\alpha+1}(Q|P)$, and the rate function for $S_{Q|P}$ with respect to Q on $]S(P|Q), S(Q|P)[$ is $\widehat{\varphi}$. Part (2) of Corollary 2.11 yields

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log Q_N([T_N(\theta)]^c) = -\widehat{\varphi}(\theta). \quad (4.47)$$

The relations (4.46) and (4.47) yield that $h(\widehat{\varphi}(\theta)) \leq -\varphi(-\theta)$. Since $\widehat{\varphi}(\theta) = s$, the first inequality (4.44) follows from Part (4) of Proposition 4.19.

We now turn to the second inequality in (4.44). For $\theta \in]-S(P|Q), S(Q|P)[$ and $T_N \subset \Omega^N$ we set

$$D_N(T_N, \theta) = Q_N([T_N]^c) + e^{\theta N} P_N(T_N).$$

Arguing in the same way as in the proof of Parts (1)-(3) of Proposition 4.17, one shows that for any T_N ,

$$D_N(T_N, \theta) \geq D_N(T_N(\theta), \theta).$$

The relations (4.46) and (4.47) yield

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log D_N(T_N(\theta), \theta) = -\widehat{\varphi}(\theta).$$

Fix now $s \in]0, S(P|Q)[$ and let $\theta \in]-S(P|Q), S(Q|P)[$ be such that $\widehat{\varphi}(\theta) = s$. Let $(T_N)_{N \geq 1}$ be a sequence of tests such that

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) \leq -s.$$

Then, for any θ' satisfying $\theta < \theta' < S(Q|P)$ we have

$$\begin{aligned} -\widehat{\varphi}(\theta') &= \lim_{N \rightarrow \infty} \frac{1}{N} \log \left(Q_N([T_N(\theta')]^c) + e^{\theta' N} P_N(T_N(\theta')) \right) \\ &\leq \liminf_{N \rightarrow \infty} \frac{1}{N} \log \left(Q_N(T_N^c) + e^{\theta' N} P_N(T_N) \right) \\ &\leq \max \left(\liminf_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c), \theta' + \liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \right) \\ &\leq \max \left(-\widehat{\varphi}(\theta), \theta' + \liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \right). \end{aligned} \quad (4.48)$$

Since $\widehat{\varphi}$ is strictly decreasing on $] -S(P|Q), S(Q|P)[$ we have that $-\widehat{\varphi}(\theta') > -\varphi(\theta)$, and (4.48) gives

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \geq -\theta' - \widehat{\varphi}(\theta') = -\varphi(\theta').$$

Taking $\theta' \downarrow \theta$, we derive

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \geq -\varphi(\theta) = -\varphi(\widehat{\varphi}^{-1}(s)) = \psi(s),$$

and so $\underline{h}(s) \geq \psi(s)$. \square

Remark 4.8 Theorem 4.21 and its proof give the following. For any sequence of tests $(T_N)_{N \geq 1}$ such that

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) \leq -s \quad (4.49)$$

one has

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \geq \psi(s).$$

On the other hand, if $s \in]0, S(P|Q)[$, $\widehat{\varphi}(\theta) = s$, and $T_N(\theta)$ is defined by (4.45), then

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N([T_N(\theta)]^c) = -s \quad \text{and} \quad \lim_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N(\theta)) = \psi(s).$$

Exercise 4.17. Set

$$\bar{h}(0) = \inf_{(T_N)} \left\{ \limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \mid \limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) < 0 \right\},$$

$$\underline{h}(0) = \inf_{(T_N)} \left\{ \liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \mid \limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) < 0 \right\},$$

$$h(0) = \inf_{(T_N)} \left\{ \lim_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N) \mid \limsup_{N \rightarrow \infty} \frac{1}{N} \log Q_N(T_N^c) < 0 \right\},$$

where in the last case the infimum is taken over all sequences of tests $(T_N)_{N \geq 1}$ for which the limit

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N(T_N)$$

exists. Prove that

$$\bar{h}(0) = \underline{h}(0) = h(0) = -S(Q|P).$$

Compare with Exercise 4.10.

4.9 Notes and references

The relative entropy $S(P|P_{\text{ch}})$ already appeared in Shannon's work [Sha]. The definition (4.1) is commonly attributed to Kullback and Leibler [KullLe], and the relative entropy is sometimes called the Kullback-Leibler divergence. From a historical perspective, it is interesting to note that the symmetrized relative entropy $S(P|Q) + S(Q|P)$ was introduced by Jeffreys in [Jeff] (see Equation (1)) in 1946.

The basic properties of the relative entropy described in Section 4.1 are so well-known that it is difficult to trace the original sources. The statement of Proposition 4.1 is sometimes called Gibbs's inequality and sometimes Shannon's inequality. For the references regarding Theorem 4.2 and Exercise 4.7 see Exercise 17 in Chapter 3 of [CsiKö] (note the typo regarding the value of the constant c).

The variational principles discussed in Section 4.2 are of fundamental importance in statistical mechanics and we postpone their discussion to Part II of the lecture notes.

The attribution of Theorem 4.12 to statistician Charles Stein appears to be historically inaccurate; for a hilarious account of the events that has led to this see the footnote on the page 85 of [John]. Theorem 4.12 was proven by Hermann Chernoff in [Che]. To avoid further confusion, we have used the usual terminology. To the best of my knowledge, the Large Deviations arguments behind the proof of Stein's Lemma, which were implicit in the original work [Che], were brought to the surface for the first time in [Ana, Sow], allowing for a substantial generalization of the original results.¹ Our proof follows [Sow].

The Fluctuation Relation described in Section 4.4 is behind the spectacular developments in non-equilibrium statistical mechanics mentioned in the Introduction. We will return to this topic in Part II of the lecture notes.

The choice of the name for Jensen-Shannon entropy (or divergence) and metric is unclear; see [Lin]. To the best of my knowledge, Theorem 4.15 was first proven in [EndSc, ÖstVa]. Our proof follows closely [EndSc]. For additional information see [FugTo].

The definition of the Rényi relative entropy is usually attributed to [Rén], although the "un-normalized" $\hat{S}_\alpha(P|Q)$ already appeared in the work of Chernoff [Che] in 1952.

The hypothesis testing is an essential procedure in statistics. Its relevance to modern developments in

¹By this I mean that essentially the same argument yields the proof of Stein's Lemma in a very general probabilistic setting.

non-equilibrium statistical mechanics will be discussed in Part II of the lecture notes. Theorem 4.18 is due to Chernoff [Che]. As in the case of Stein's Lemma, the LDP based proof allows to considerably generalize the original result. The Hoeffding error exponents were first introduced and studied in [Hoe] and the previous remarks regarding the proof applies to them as well. For additional information about hypothesis testing see [LeRo].

Chapter 5

Why is the relative entropy natural?

5.1 Introduction

This chapter is a continuation of Section 3.4 and concerns naturalness of the relative entropy.

1. Operational interpretation. Following on Shannon's quote in Section 3.7, Stein's Lemma gives an operational interpretation of the relative entropy $S(P|Q)$. Chernoff and Hoeffding error exponents, Theorems 4.18 and 4.21, give an operational interpretation of Rényi's relative entropy $\widehat{S}_\alpha(P|Q)$ and, via formula (4.29), of Rényi's entropy $\widehat{S}_\alpha(P)$ as well. Note that this operational interpretation of Rényi's entropies is rooted in the LDP's for respective entropy functions which are behind the proofs of Theorems 4.18 and 4.21.

2. Axiomatic characterizations. Recall that $\mathcal{A}(\Omega) = \{(P, Q) \in \mathcal{P}(\Omega) \mid P \ll Q\}$. Set $\mathcal{A} = \cup_{\Omega} \mathcal{A}(\Omega)$. The axiomatic characterizations of relative entropy concern choice of a function $\mathfrak{S} : \mathcal{A} \rightarrow \mathbb{R}$ that should qualify as a measure of *entropic distinguishability* of a pair $(P, Q) \in \mathcal{A}$. The goal is to show that intuitive natural demands uniquely specify \mathfrak{S} up to a choice of units, namely that for some $c > 0$ and all $(P, Q) \in \mathcal{A}$, $\mathfrak{S}(P, Q) = cS(P|Q)$.

We list basic properties that any candidate \mathfrak{S} for relative entropy should satisfy. The obvious ones are

$$\mathfrak{S}(P, P) = 0, \quad \mathfrak{S}(P, Q) \geq 0, \quad \exists (P, Q) \text{ such that } \mathfrak{S}(P, Q) > 0. \quad (5.1)$$

Another obvious requirement is that if $|\Omega_1| = |\Omega_2|$ and $\theta : \Omega_1 \rightarrow \Omega_2$ is a bijection, then for any $(P, Q) \in \mathcal{A}$,

$$\mathfrak{S}(P, Q) = \mathfrak{S}(P \circ \theta, Q \circ \theta).$$

In other words, the distinguishability of a pair (P, Q) should not depend on the labeling of the elementary events. This requirement gives that \mathfrak{S} is completely specified by its restriction $\mathfrak{S} : \cup_{L \geq 1} \mathcal{A}_L \rightarrow [0, \infty[$, where

$$\mathcal{A}_L = \{(p_1, \dots, p_L), (q_1, \dots, q_L)\} \in \mathcal{P}_L \times \mathcal{P}_L \mid q_k = 0 \Rightarrow p_k = 0\},$$

and that this restriction satisfies

$$\mathfrak{S}((p_1, \dots, p_L), (q_1, \dots, q_L)) = \mathfrak{S}((p_{\pi(1)}, \dots, p_{\pi(L)}), (q_{\pi(1)}, \dots, q_{\pi(L)})) \quad (5.2)$$

for any $L \geq 1$ and any permutation π of $\{1, \dots, L\}$. In the proofs of Theorems 5.1 and 5.2 we shall assume that (5.1) and (5.2) are satisfied.

Split additivity characterization. This axiomatic characterization is the relative entropy analog of Theorem 3.4, and has its roots in the identity (recall Proposition 4.8)

$$\begin{aligned} S(p_1 P_1 + \dots + p_n P_n \mid q_1 Q_1 + \dots + q_n Q_n) \\ = p_1 S(P_1 \mid Q_1) + \dots + p_n S(P_n \mid Q_n) + S((p_1, \dots, p_n) \mid (q_1, \dots, q_n)) \end{aligned}$$

which holds if $(\text{supp}P_j \cup \text{supp}Q_j) \cap (\text{supp}P_k \cup \text{supp}Q_k) = \emptyset$ for all $j \neq k$.

Theorem 5.1 Let $\mathfrak{S} : \mathcal{A} \rightarrow [0, \infty[$ be a function such that:

(a) \mathfrak{S} is continuous on \mathcal{A}_2 .

(b) For any finite collection of disjoint sets Ω_j , $j = 1, \dots, n$, any $(P_j, Q_j) \in \mathcal{A}(\Omega_j)$, and any $p = (p_1, \dots, p_n)$, $q = (q_1, \dots, q_n) \in \mathcal{P}_n$,

$$\mathfrak{S}\left(\bigoplus_{k=1}^n p_k P_k, \bigoplus_{k=1}^n q_k Q_k\right) = \sum_{k=1}^n p_k \mathfrak{S}(P_k, Q_k) + \mathfrak{S}(p|q). \quad (5.3)$$

Then there exists $c > 0$ such that for all $(P, Q) \in \mathcal{A}$,

$$\mathfrak{S}(P, Q) = cS(P|Q). \quad (5.4)$$

Remark 5.1 If the positivity and non-triviality assumptions are dropped, then the proof gives that (5.4) holds for some $c \in \mathbb{R}$.

Exercise 5.1. Following on Remark 3.2, can you verbalize the split-additivity property (5.3)?

We shall prove Theorem 5.1 in Section 5.2. The vanishing assumption $\mathfrak{S}(P, P) = 0$ for all P plays a very important role in the argument. Note that

$$\mathfrak{S}(P, Q) = - \sum_{\omega} P(\omega) \log Q(\omega)$$

satisfies (a) and (b) of Theorem 5.1 and assumptions (5.1) apart from $\mathfrak{S}(P, P) = 0$.

Stochastic monotonicity + super additivity characterization. This characterization is related to Theorem 3.5, although its proof is both conceptually different and technically simpler. The characterization asserts that two intuitive requirements, the stochastic monotonicity (Proposition 4.7) and super-additivity (Proposition 4.12) uniquely specify relative entropy.

Theorem 5.2 Let $\mathfrak{S} : \mathcal{A} \rightarrow [0, \infty[$ be a function such that:

(a) \mathfrak{S} is continuous on \mathcal{A}_L for all $L \geq 1$.

(b) For any $P, Q \in \mathcal{A}(\Omega)$ and any stochastic map $\Phi : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\hat{\Omega})$ (note that $(\Phi(P), \Phi(Q)) \in \mathcal{A}(\hat{\Omega})$),

$$\mathfrak{S}(\Phi(P), \Phi(Q)) \leq \mathfrak{S}(P, Q). \quad (5.5)$$

(c) For any P and $Q = Q_l \otimes Q_r$ in $\mathcal{A}(\Omega_l \times \Omega_r)$,

$$\mathfrak{S}(P_l, Q_l) + \mathfrak{S}(P_r, Q_r) \leq \mathfrak{S}(P, Q), \quad (5.6)$$

with the equality iff $P = P_l \otimes P_r$.

Then there exists $c > 0$ such that for all $(P, Q) \in \mathcal{A}$,

$$\mathfrak{S}(P, Q) = cS(P|Q). \quad (5.7)$$

We shall prove Theorem 5.2 in Section 5.3. Note that neither assumptions (a) \wedge (b) nor (a) \wedge (c) are sufficient to deduce (5.7): (a) and (b) hold for the Rényi relative entropy $(P, Q) \mapsto S_\alpha(P, Q)$ if $\alpha \in]0, 1[$ ((c) fails here), while (a) and (c) hold for the entropy $(P, Q) \mapsto S(P)$ ((b) fails here, recall Exercise 4.5).

4. Sanov's theorem. This result is a deep refinement of Crámer's theorem and the basic indicator of the central role the relative entropy plays in the theory of Large Deviations. We continue with our framework: Ω is a finite set and P a given probability measure on Ω . We shall assume that P is faithful.

To avoid confusion, we shall occasionally denote the generic element of Ω with a letter a (and list the elements of Ω as $\Omega = \{a_1, \dots, a_L\}$). For $\omega \in \Omega$ we denote by $\delta_\omega \in \mathcal{P}(\Omega)$ the pure probability measure concentrated at ω : $\delta_\omega(a) = 1$ if $a = \omega$ and zero otherwise. For $\omega = (\omega_1, \dots, \omega_N)$ we set

$$\delta_\omega = \frac{1}{N} \sum_{k=1}^N \delta_{\omega_k}.$$

Obviously, $\delta_\omega \in \mathcal{P}(\Omega)$ and

$$\delta_\omega(a) = \frac{\text{the number of times } a \text{ appears in the sequence } \omega = (\omega_1, \dots, \omega_N)}{N}.$$

Sanov's theorem concerns the statistics of the map $\Omega^N \ni \omega \mapsto \delta_\omega \in \mathcal{P}(\Omega)$ w.r.t. the product probability measure P_N . The starting point is the corresponding law of large numbers.

Proposition 5.3 For any $\epsilon > 0$,

$$\lim_{N \rightarrow \infty} P_N \{ \omega \in \Omega^N \mid d_V(\delta_\omega, P) \geq \epsilon \} = 0.$$

Sanov's theorem concerns fluctuations in the above LLN, or more precisely, for a given $\Gamma \subset \mathcal{P}(\Omega)$, it estimates the probabilities

$$P_N \{ \omega \in \Omega^N \mid \delta_\omega \in \Gamma \}$$

in the limit of large N .

Theorem 5.4 For any closed set $\Gamma \subset \mathcal{P}(\Omega)$,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \{ \omega \in \Omega^N \mid \delta_\omega \in \Gamma \} \leq - \inf_{Q \in \Gamma} S(Q|P),$$

and for any open set $\Gamma \subset \mathcal{P}(\Omega)$,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \{ \omega \in \Omega^N \mid \delta_\omega \in \Gamma \} \geq - \inf_{Q \in \Gamma} S(Q|P).$$

We shall prove Proposition 5.3 and Theorem 5.4 in Section 5.4 where the reader can also find additional information about Sanov's theorem.

5.2 Proof of Theorem 5.1

The function

$$F(t) = \mathfrak{G}((1, 0), (t, 1 - t)), \quad t \in]0, 1],$$

will play an important role in the proof. Obviously, F is continuous on $]0, 1]$ and $F(1) = 0$.

We split the proof into five steps.

Step 1. Let $(P, Q) \in \mathcal{A}(\Omega)$, where $\Omega = \{\omega_1, \dots, \omega_n\}$, and suppose that $P(\omega_j) = 0$ for $j > k$. Set $\Omega_1 = \{\omega_1, \dots, \omega_k\}$, $P_1(\omega_j) = P(\omega_j)$, and

$$Q_1(\omega_j) = \frac{Q(\omega_j)}{Q(\omega_1) + \dots + Q(\omega_k)}.$$

It is obvious that $(P_1, Q_1) \in \mathcal{A}(\Omega_1)$. We then have

$$\mathfrak{S}(P, Q) = F(q_1 + \dots + q_k) + \mathfrak{S}(P_1, Q_1). \quad (5.8)$$

Note that if $k = n$, then (5.8) follows from $F(1) = 1$. Otherwise, write $\Omega = \Omega_1 \oplus \Omega_2$, with $\Omega_2 = \{\omega_{k+1}, \dots, \omega_n\}$. Take any $P_2 \in \mathcal{P}(\Omega_2)$, write

$$(P, Q) = (1 \cdot P_1 \oplus 0 \cdot P_2, tQ_1 \oplus (1-t)Q_2),$$

where $t = q_1 + \dots + q_k$, Q_2 is arbitrary if $t = 1$, and $Q_2(\omega_j) = Q(\omega_j)/(1-t)$ if $t < 1$, and observe that the statement follows from (5.5).

Step 2. $F(ts) = F(t) + F(s)$ for all $s, t \in]0, 1]$.

Consider $\mathfrak{S}((1, 0, 0), (ts, t(1-s), 1-t))$. Applying Step 1 with $k = 1$ we get

$$\mathfrak{S}((1, 0, 0), (ts, t(1-s), 1-t)) = F(ts) + \mathfrak{S}((1), (1)) = F(ts).$$

Applying Step 1 with $k = 2$ gives

$$\mathfrak{S}((1, 0, 0), (ts, t(1-s), 1-t)) = F(t) + \mathfrak{S}((1, 0), (s, 1-s)) = F(t) + F(s),$$

and the statement follows.

Step 3. For some $c \in \mathbb{R}$, $F(t) = -c \log t$ for all $t \in]0, 1]$.

Set $H(s) = F(e^{-s})$. Then H is continuous on $[0, \infty[$ and satisfies $H(s_1 + s_2) = H(s_1) + H(s_2)$. It is now a standard exercise to show that $H(s) = cs$ where $c = H(1)$. Setting $t = e^{-s}$ gives $F(t) = -c \log t$.

This is the only point where the regularity assumption (a) has been used (implying the continuity of F), and so obviously (a) can be relaxed.¹ Note that (5.1) implies $c \geq 0$.

Step 4. We now prove that for any $n \geq 2$ and any pair $(p, q) \in \mathcal{A}_n$ of faithful probability measures,

$$\mathfrak{S}(p, q) = cS(p|q), \quad (5.9)$$

where c is the constant from Step 3.

Let $p = (p_1, \dots, p_n)$, $q = (q_1, \dots, q_n)$, and choose $t \in]0, 1]$ such that $q_k - tp_k \geq 0$ for all k . Set

$$K = \mathfrak{S}((p_1, \dots, p_n, 0, \dots, 0), (tp_1, \dots, tp_n, q_1 - tp_1, \dots, q_n - tp_n)).$$

It follows from Steps 1 and 3 that

$$K = F(t) + \mathfrak{S}(p, p) = -c \log t. \quad (5.10)$$

On the other hand, (5.2) and (5.3) yield

$$\begin{aligned} K &= \mathfrak{S}((p_1, 0, \dots, p_n, 0), (tp_1, q_1 - tp_1, \dots, tp_n, q_n - tp_n)) \\ &= \mathfrak{S}\left((p_1(1, 0), \dots, p_n(1, 0)), \left(q_1 \left(\frac{tp_1}{q_1}, 1 - \frac{tp_1}{q_1}\right), \dots, q_n \left(\frac{tp_n}{q_n}, 1 - \frac{tp_n}{q_n}\right)\right)\right) \\ &= \sum_{k=1}^n p_k F\left(\frac{tp_k}{q_k}\right) + \mathfrak{S}(p, p), \end{aligned}$$

¹It suffices that F is Borel measurable.

and it follows from Step 3 that

$$K = -c \log t - cS(p|q) + \mathfrak{S}(p, q). \quad (5.11)$$

Comparing (5.10) and (5.11) we derive (5.9).

Step 5. We now show that (5.9) also holds for non-faithful p 's and complete the proof of Theorem 5.1. By (5.2) we may assume that $p_j > 0$ for $j \leq k$ and $p_j = 0$ for $j > k$, where $k < n$. Then, setting $s = q_1 + \cdots + q_k$, Steps 1 and 3 yield

$$\mathfrak{S}(p, q) = -c \log s + \mathfrak{S}((p_1, \dots, p_k), (q_1/s, \dots, q_k/s)),$$

and it follows from Step 4 that

$$\mathfrak{S}(p, q) = -c \log s + cS((p_1, \dots, p_k)|(q_1/s, \dots, q_k/s)).$$

On the other hand, a direct computation gives

$$S(p|q) = -\log s + S((p_1, \dots, p_k)|(q_1/s, \dots, q_k/s)),$$

and so $\mathfrak{S}(p, q) = cS(p|q)$.

The non-triviality assumption that \mathfrak{S} is not vanishing on \mathcal{A} gives that $c > 0$.

5.3 Proof of Theorem 5.2

We shall need the following preliminary result which is of independent interest and which we will prove at the end of this section. Recall that if P is a probability measure on Ω , then $P_N = P \otimes \cdots \otimes P$ is the product probability measure on $\Omega^N = \Omega \times \cdots \times \Omega$.

Proposition 5.5 *Suppose that $(P, Q) \in \mathcal{A}(\Omega)$ and $(\widehat{P}, \widehat{Q}) \in \mathcal{A}(\widehat{\Omega})$ are such that $S(P|Q) > S(\widehat{P}|\widehat{Q})$. Then there exists a sequence of stochastic maps $(\Phi_N)_{N \geq 1}$, $\Phi_N : \mathcal{P}(\Omega^N) \rightarrow \mathcal{P}(\widehat{\Omega}^N)$ such that $\Phi_N(Q_N) = \widehat{Q}_N$ for all $N \geq 1$ and*

$$\lim_{N \rightarrow \infty} d_V(\Phi_N(P_N), \widehat{P}_N) = 0.$$

We now turn to the proof of Theorem 5.2. Recall our standing assumptions (5.1). Let $(P^{(0)}, Q^{(0)}) \in \mathcal{A}$ be such that $\mathfrak{S}(P^{(0)}, Q^{(0)}) > 0$, and let $c > 0$ be such that

$$\mathfrak{S}(P^{(0)}, Q^{(0)}) = cS(P^{(0)}|Q^{(0)}).$$

Let $(P, Q) \in \mathcal{A}$, $P \neq Q$, be given and let L, M, L', M' be positive integers such that

$$\frac{L'}{M'} S(P^{(0)}|Q^{(0)}) < S(P|Q) < \frac{L}{M} S(P^{(0)}|Q^{(0)}). \quad (5.12)$$

We work first with the r.h.s. of this inequality which can be rewritten as

$$S(P_M|Q_M) < S(P_L^{(0)}|Q_L^{(0)}).$$

It follows from Proposition 5.5 that there exists a sequence of stochastic maps $(\Phi_N)_{N \geq 1}$ such that $\Phi_N(Q_{LN}^{(0)}) = Q_{MN}$ and

$$\lim_{N \rightarrow \infty} d_V(\Phi_N(P_L^{(0)}), P_{MN}) = 0. \quad (5.13)$$

We now turn to $\mathfrak{S}(P, Q)$ and note that

$$\begin{aligned}
M\mathfrak{S}(P, Q) &= \mathfrak{S}(P_M, Q_M) = \frac{1}{N}\mathfrak{S}(P_{MN}, Q_{MN}) \\
&= \frac{1}{N} \left[\mathfrak{S}(P_{MN}, Q_{MN}) - \mathfrak{S}(\Phi_N(P_L^{(0)}), Q_{MN}) \right] + \frac{1}{N}\mathfrak{S}(\Phi_N(P_L^{(0)}), \Phi_N(Q_{LN}^{(0)})) \\
&\leq \frac{1}{N} \left[\mathfrak{S}(P_{MN}, Q_{MN}) - \mathfrak{S}(\Phi_N(P_L^{(0)}), Q_{MN}) \right] + \frac{1}{N}\mathfrak{S}(P_{LN}^{(0)}, Q_{LN}^{(0)}) \\
&= \frac{1}{N} \left[\mathfrak{S}(P_{MN}, Q_{MN}) - \mathfrak{S}(\Phi_N(P_L^{(0)}), Q_{MN}) \right] + L\mathfrak{S}(P_L^{(0)}, Q^{(0)}).
\end{aligned} \tag{5.14}$$

Write $Q_{MN} = Q_M \otimes \cdots \otimes Q_M$ and denote by $R_{k,N}$ the marginal of $\Phi_N(P_L^{(0)})$ with the respect to the k -th component of this decomposition. Assumption (c) gives

$$\frac{1}{N} \left[\mathfrak{S}(P_{MN}, Q_{MN}) - \mathfrak{S}(\Phi_N(P_L^{(0)}), Q_{MN}) \right] \leq \frac{1}{N} \sum_{k=1}^N [\mathfrak{S}(P_M, Q_M) - \mathfrak{S}(R_{k,N}, Q_M)]. \tag{5.15}$$

One easily shows that (5.13) implies that for any k ,

$$\lim_{N \rightarrow \infty} d_V(R_{k,N}, P_M) = 0. \tag{5.16}$$

It then follows from (5.15) that

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \left[\mathfrak{S}(P_{MN}, Q_{MN}) - \mathfrak{S}(\Phi_N(P_L^{(0)}), Q_{MN}) \right] \leq 0. \tag{5.17}$$

Returning to (5.14), (5.17) yields

$$\mathfrak{S}(P, Q) \leq \frac{L}{M}\mathfrak{S}(P^{(0)}, Q^{(0)}) = \frac{L}{M}cS(P^{(0)}|Q^{(0)}). \tag{5.18}$$

Since the only constraint regarding the choice of L and M is that (5.12) holds, we derive from (5.18) that

$$\mathfrak{S}(P, Q) \leq cS(P|Q).$$

Starting with the l.h.s. of the inequality (5.12) and repeating the above argument one derives that $\mathfrak{S}(P, Q) \geq cS(P|Q)$. Hence, $\mathfrak{S}(P, Q) = cS(P|Q)$ for all $(P, Q) \in \mathcal{A}$ with $P \neq Q$. Since this relation holds trivially for $P = Q$, the proof is complete. \square

Exercise 5.2. Prove that (5.13) implies (5.16).

Proof of Proposition 5.5. The statement is trivial if $\hat{P} = \hat{Q}$, so we assume that $\hat{P} \neq \hat{Q}$ (hence $S(\hat{P}|\hat{Q}) > 0$). Let t, \hat{t} be such that

$$S(\hat{P}|\hat{Q}) < \hat{t} < t < S(P|Q).$$

It follows from Stein's Lemma that one can find a sequence of sets $(T_N)_{N \geq 1}$, $T_N \subset \Omega_N$, such that

$$\lim_{N \rightarrow \infty} P_N(T_N) = 1, \quad Q_N(T_N) \leq C_1 e^{-Nt},$$

for some constant $C_1 > 0$. Let $\Psi_N : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\{0, 1\})$ be a stochastic map induced by the matrix

$$\Psi_N(\omega, 0) = \chi_{T_N}(\omega), \quad \Psi_N(\omega, 1) = \chi_{T_N^c}(\omega),$$

where χ_{T_N} and $\chi_{T_N^c}$ are the characteristic functions of T_N and its complement T_N^c . It follows that

$$\Psi_N(P_N) = (p_N, \bar{p}_N), \quad \Psi(Q_N) = (q_N, \bar{q}_N),$$

where

$$p_N = P_N(T_N), \quad q_N = Q(T_N).$$

Obviously $\bar{p}_N = 1 - p_N$, $\bar{q}_N = 1 - q_N$.

It follows again from Stein's Lemma that one can find a sequence of sets $(\hat{T}_N)_{N \geq 1}$, $\hat{T}_N \subset \hat{\Omega}_N$, such that

$$\lim_{N \rightarrow \infty} \hat{P}_N(\hat{T}_N) = 1, \quad Q_N(\hat{T}_N^c) > C_2 e^{-N\hat{t}},$$

for some constant $C_2 > 0$. We now construct a stochastic map $\hat{\Psi}_N : \mathcal{P}(\{0, 1\}) \rightarrow \mathcal{P}(\hat{\Omega})$ as follows. Let $\delta_0 = (1, 0)$, $\delta_1 = (0, 1)$. We set first

$$\hat{\Psi}_N(\delta_0)(\omega) = \frac{\hat{P}_N(\omega)}{\sum_{\omega' \in \hat{T}_N} \hat{P}_N(\omega')} \quad \text{if } \omega \in \hat{T}_N,$$

$\hat{\Psi}_N(\delta_0)(\omega) = 0$ otherwise, and observe that

$$d_V(\hat{\Psi}_N(\delta_0), \hat{P}_N) \leq \hat{P}_N(\hat{T}_N^c) + \frac{1 - \hat{P}_N(\hat{T}_N)}{\hat{P}_N(\hat{T}_N)}.$$

Hence,

$$\lim_{N \rightarrow \infty} d_V(\hat{\Psi}_N(\delta_0), \hat{P}_N) = 0.$$

Let

$$D_N(\omega) = \hat{Q}_N(\omega) - q_N \Phi_N(\delta_0)(\omega).$$

If $\omega \notin \hat{T}_N$, then obviously $D_N(\omega) = \hat{Q}_N(\omega) \geq 0$, and if $\omega \in \hat{T}_N$,

$$D_N(\omega) \geq C_2^{-\hat{t}N} - c_1 e^{-\hat{t}N}.$$

Since $0 < \hat{t} < t$, there is N_0 such that for $N \geq N_0$ and all $\omega \in \hat{\Omega}$, $D_N(\omega) \geq 0$. From now on we assume that $N \geq N_0$, set

$$\hat{\Psi}_N(\delta_1) = \frac{1}{\bar{q}_N} (Q_N - q_N \Phi_N(\delta_0)),$$

and define $\hat{\Psi}_N : \mathcal{P}(\{0, 1\}) \rightarrow \mathcal{P}(\hat{\Omega})$ by

$$\hat{\Psi}_N(p, q) = p \Psi(\delta_0) + q \Psi(\delta_1).$$

The map $\hat{\Psi}_N$ is obviously stochastic and

$$\hat{\Psi}_N(q_N, \bar{q}_N) = \hat{Q}_N.$$

Moreover,

$$\begin{aligned} d_V(\hat{\Psi}_N(p_N, \bar{p}_N), \hat{P}_N) &\leq d_V(\hat{\Psi}_N(p_N, \bar{p}_N), \hat{\Psi}_N(\delta_0)) + d_V(\hat{\Psi}_N(\delta_0), \hat{P}_N) \\ &\leq 2(1 - p_N) + d_V(\hat{\Psi}_N(\delta_0), \hat{P}_N), \end{aligned}$$

and so

$$\lim_{N \rightarrow \infty} d_V(\hat{\Psi}_N(p_N, \bar{p}_N), \hat{P}_N) = 0.$$

For $N < N_0$ we take for Φ_N an arbitrary stochastic map satisfying $\Phi_N(Q_N) = \hat{Q}_N$ and for $N \geq N_0$ we set $\Phi_N = \hat{\Psi}_N \circ \Psi_N$. Then $\Phi_N(Q_N) = \hat{Q}_N$ for all $N \geq 1$ and

$$\lim_{N \rightarrow \infty} d_V(\Phi_N(P_N), \hat{P}_N) = 0,$$

proving the proposition. \square

Exercise 5.3. Write down the stochastic matrix that induces $\widehat{\Psi}_N$.

5.4 Sanov's theorem

We start with

Proof of Proposition 5.3. Recall that $L = |\Omega|$. We have

$$d_V(\delta_\omega, P) = \sum_{a \in \Omega} \left| \frac{\sum_{k=1}^N \delta_{\omega_k}(a)}{N} - P(a) \right|,$$

and

$$\{\omega \in \Omega^N \mid d_V(\delta_\omega, P) \geq \epsilon\} \subset \bigcup_{a \in \Omega} \left\{ \omega \in \Omega^N \mid \left| \frac{\sum_{k=1}^N \delta_{\omega_k}(a)}{N} - P(a) \right| \geq \frac{\epsilon}{L} \right\}.$$

Hence,

$$P_N \{\omega \in \Omega^N \mid d_V(\delta_\omega, P) \geq \epsilon\} \leq \sum_{a \in \Omega} P_N \left\{ \omega \in \Omega^N \mid \left| \frac{\sum_{k=1}^N \delta_{\omega_k}(a)}{N} - P(a) \right| \geq \frac{\epsilon}{L} \right\}. \quad (5.19)$$

For given $a \in \Omega$, consider a random variable $X : \Omega \rightarrow \mathbb{R}$ defined by $X(\omega) = \delta_\omega(a)$. Obviously, $\mathbb{E}(X) = P(a)$ and the LLN yields that

$$\lim_{N \rightarrow \infty} P_N \left\{ \omega \in \Omega^N \mid \left| \frac{\sum_{k=1}^N \delta_{\omega_k}(a)}{N} - P(a) \right| \geq \frac{\epsilon}{L} \right\} = 0.$$

The proposition follows by combining this observation with inequality (5.19). \square

We now turn to the proof of Sanov's theorem. Recall the assumption that P is faithful. We start with the upper bound.

Proposition 5.6 Suppose that $\Gamma \subset \mathcal{P}(\Omega)$ is a closed set. Then

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \{\omega \in \Omega^N \mid \delta_\omega \in \Gamma\} \leq - \inf_{Q \in \Gamma} S(Q|P).$$

Remark 5.2 Recall that the map $\mathcal{P}(\Omega) \ni Q \mapsto S(Q|P) \in [0, \infty[$ is continuous (P is faithful). Since Γ is compact, there exists $Q_m \in \mathcal{P}(\Omega)$ such that

$$\inf_{Q \in \Gamma} S(Q|P) = S(Q_m|P).$$

Proof. Let $\epsilon > 0$ be given. Let $Q \in \Gamma$. By Exercise 4.9,

$$S(Q|P) = \sup_{X: \Omega \rightarrow \mathbb{R}} \left(\int_{\Omega} X dQ - \log \int_{\Omega} e^X dP \right).$$

Hence, we can find X such that

$$S(Q|P) - \epsilon < \int_{\Omega} X dQ - \log \int_{\Omega} e^X dP.$$

Let

$$U_\epsilon(Q) = \left\{ Q' \in \mathcal{P}(\Omega) \mid \left| \int_{\Omega} X dQ - \int_{\Omega} X dQ' \right| < \epsilon \right\}.$$

Since the map $\mathcal{P}(\Omega) \ni Q' \mapsto \int_{\Omega} X dQ'$ is continuous, $U_{\epsilon}(Q)$ is an open subset of $\mathcal{P}(\Omega)$. We now estimate

$$\begin{aligned}
P_N \{ \delta_{\omega} \in U_{\epsilon}(Q) \} &= P_N \left\{ \left| \int_{\Omega} X dQ - \int_{\Omega} X d\delta_{\omega} \right| < \epsilon \right\} \\
&\leq P_N \left\{ \int_{\Omega} X d\delta_{\omega} > \int_{\Omega} X dQ - \epsilon \right\} \\
&= P_N \left\{ \sum_{k=1}^N X(\omega_k) > N \int_{\Omega} X dQ - N\epsilon \right\} \\
&= P_N \left\{ e^{\sum_{k=1}^N X(\omega_k)} > e^{N \int_{\Omega} X dQ - N\epsilon} \right\} \\
&\leq e^{-N \int_{\Omega} X dQ + N\epsilon} \mathbb{E}(e^X)^N \\
&= e^{-N \int_{\Omega} X dQ + N \log \int_{\Omega} e^X dP + N\epsilon} \\
&\leq e^{-NS(Q|P) + 2N\epsilon}
\end{aligned}$$

Since Γ is compact, we can find $Q_1, \dots, Q_M \in \Gamma$ such that

$$\Gamma \subset \bigcup_{j=1}^M U_{\epsilon}(Q_j).$$

Then

$$\begin{aligned}
P_N \{ \delta_{\omega} \in \Gamma \} &\leq \sum_{j=1}^M P_N \{ \delta_{\omega} \in U_{\epsilon}(Q_j) \} \\
&\leq e^{2N\epsilon} \sum_{j=1}^M e^{-NS(Q_j|P)} \\
&\leq e^{2N\epsilon} M e^{-N \inf_{Q \in \Gamma} S(Q|P)}.
\end{aligned}$$

Hence

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \{ \omega \in \Omega^N \mid \delta_{\omega} \in \Gamma \} \leq - \inf_{Q \in \Gamma} S(Q|P) + 2\epsilon.$$

Since $\epsilon > 0$ is arbitrary, the statement follows. \square

We now turn to the lower bound.

Proposition 5.7 For any open set $\Gamma \subset \mathcal{P}(\Omega)$,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \{ \omega \in \Omega^N \mid \delta_{\omega} \in \Gamma \} \geq - \inf_{Q \in \Gamma} S(Q|P).$$

Proof. Let $Q \in \Gamma$ be faithful. Recall that $S_{Q|P} = \log \Delta_{Q|P}$ and

$$\int_{\Omega} S_{P|Q} d\delta_{\omega} = \frac{S_{Q|P}(\omega_1) + \dots + S_{Q|P}(\omega_N)}{N}.$$

Let $\epsilon > 0$ and

$$R_{N,\epsilon} = \left\{ \delta_{\omega} \in \Gamma \mid \left| \int_{\Omega} S_{Q|P} d\delta_{\omega} - S(Q|P) \right| < \epsilon \right\}.$$

Then

$$\begin{aligned} P_N \{\delta_\omega \in \Gamma\} &\geq P_N(R_{N,\epsilon}) = \int_{R_{N,\epsilon}} \Delta_{P_N|Q_N} dQ_N = \int_{R_{N,\epsilon}} \Delta_{Q_N|P_N}^{-1} dQ_N \\ &= \int_{R_{N,\epsilon}} e^{-\sum_{k=1}^N S_{Q|P}(\omega_k)} dQ_N \\ &\geq e^{-NS(Q|P) - N\epsilon} Q_N(R_{N,\epsilon}). \end{aligned}$$

Note that for ϵ small enough (Γ is open!)

$$R_{N,\epsilon} \supset \left\{ \omega \in \Omega^N \mid d_V(Q, \delta_\omega) < \epsilon \right\} \cap \left\{ \omega \in \Omega^N \mid \left| \frac{S_{Q|P}(\omega_1) + \cdots + S_{Q|P}(\omega_N)}{N} - S(Q|P) \right| < \epsilon \right\}.$$

By the LLN,

$$\lim_{N \rightarrow \infty} Q_N(R_{N,\epsilon}) = 1.$$

Hence, for any faithful $Q \in \Gamma$,

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \{ \omega \in \Omega^N \mid \delta_\omega \in \Gamma \} \geq -S(Q|P). \quad (5.20)$$

Since Γ is open and the map $\mathcal{P}(\Omega) \ni Q \rightarrow S(Q|P)$ is continuous,

$$\inf_{Q \in \Gamma \cap \mathcal{P}_f(\Omega)} S(Q|P) = \inf_{Q \in \Gamma} S(Q|P). \quad (5.21)$$

The relations (5.20) and (5.21) imply

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \{ \omega \in \Omega^N \mid \delta_\omega \in \Gamma \} \geq - \inf_{Q \in \Gamma} S(Q|P).$$

□

Exercise 5.4. Prove the identity (5.21).

A set $\Gamma \in \mathcal{P}(\Omega)$ is called *Sanov-nice* if

$$\inf_{Q \in \text{int } \Gamma} S(Q|P) = \inf_{Q \in \text{cl } \Gamma} S(Q|P),$$

where int/cl stand for the interior/closure. If Γ is Sanov-nice, then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \{ \omega \in \Omega^N \mid \delta_\omega \in \Gamma \} = - \inf_{Q \in \Gamma} S(Q|P).$$

Exercise 5.5.

1. Prove that any open set $\Gamma \subset \mathcal{P}(\Omega)$ is Sanov-nice.
2. Suppose that $\Gamma \subset \mathcal{P}(\Omega)$ is convex and has non-empty interior. Prove that Γ is Sanov-nice.

We now show that Sanov's theorem implies Cramér's theorem. The argument we shall use is an example of the powerful *contraction principle* in theory of Large Deviations.

Suppose that in addition to Ω and P we are given a random variable $X : \Omega \rightarrow \mathbb{R}$. C and I denote the cumulant generating function and the rate function of X . Note that

$$\frac{\mathcal{S}_N(\omega)}{N} = \frac{X(\omega_1) + \cdots + X(\omega_N)}{N} = \int_{\Omega} X d\delta_{\omega}.$$

Hence, for any $S \subset \mathbb{R}$,

$$\frac{\mathcal{S}_N(\omega)}{N} \in S \Leftrightarrow \delta_{\omega} \in \Gamma_S,$$

where

$$\Gamma_S = \left\{ Q \in \mathcal{P}(\Omega) \mid \int_{\Omega} X dQ \in S \right\}.$$

Exercise 5.6. Prove that

$$\text{int } \Gamma_S = \Gamma_{\text{int}S}, \quad \text{cl } \Gamma_S = \Gamma_{\text{cl}S}.$$

Sanov's theorem and the last exercise yield

Proposition 5.8 For any $S \subset \mathbb{R}$,

$$\begin{aligned} - \inf_{Q \in \Gamma_{\text{int}S}} S(Q|P) &\leq \liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \omega \in \Omega^N \mid \frac{\mathcal{S}_N(\omega)}{N} \in S \right\} \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \omega \in \Omega^N \mid \frac{\mathcal{S}_N(\omega)}{N} \in S \right\} \leq - \inf_{Q \in \Gamma_{\text{cl}S}} S(Q|P), \end{aligned}$$

To relate this result to Cramér's theorem we need:

Proposition 5.9 For any $S \subset \mathbb{R}$,

$$\inf_{\theta \in S} I(\theta) = \inf_{Q \in \Gamma_S} S(Q|P). \quad (5.22)$$

Proof. Let $Q \in \mathcal{P}(\Omega)$. An application of Jensen's inequality gives that for all $\alpha \in \mathbb{R}$,

$$\begin{aligned} C(\alpha) &= \log \left(\sum_{\omega \in \Omega} e^{\alpha X(\omega)} P(\omega) \right) \\ &\geq \log \left(\sum_{\omega \in \text{supp}Q} e^{\alpha X(\omega)} \frac{P(\omega)}{Q(\omega)} Q(\omega) \right) \\ &\geq \sum_{\omega \in \text{supp}Q} Q(\omega) \log \left[e^{\alpha X(\omega)} \frac{P(\omega)}{Q(\omega)} \right]. \end{aligned}$$

Hence,

$$C(\alpha) \geq \alpha \int_{\Omega} X dQ - S(Q|P). \quad (5.23)$$

If Q is such that $\theta_0 = \int_{\Omega} X dQ \in S$, then (5.23) gives

$$S(Q|P) \geq \sup_{\alpha \in \mathbb{R}} (\alpha \theta_0 - C(\alpha)) = I(\theta_0) \geq \inf_{\theta \in S} I(\theta),$$

and so

$$\inf_{Q \in \Gamma_S} S(Q|P) \geq \inf_{\theta \in S} I(\theta). \quad (5.24)$$

On the other hand, if $\theta \in]m, M[$, where $m = \min_{\omega \in \Omega} X(\omega)$ and $M = \max_{\omega \in \Omega} X(\omega)$, and $\alpha = \alpha(\theta)$ is such that $C'(\alpha(\theta)) = \theta$, then, with Q_α defined by (2.3) (recall also the proof of Cramer's theorem), $\theta = \int_{\Omega} X dQ_\alpha$ and $S(Q_\alpha|P) = \alpha\theta - C(\alpha) = I(\theta)$. Hence, if $S \subset]m, M[$, then for any $\theta_0 \in S$, $\inf_{Q \in \Gamma_S} S(Q|P) \leq I(\theta_0)$, and so

$$\inf_{Q \in \Gamma_S} S(Q|P) \leq \inf_{\theta \in S} I(\theta). \quad (5.25)$$

It follows from (5.24) and (5.25) that (5.22) holds for $S \subset]m, M[$. One checks directly that

$$I(m) = \inf_{Q: \int_{\Omega} X dQ = m} S(Q|P), \quad I(M) = \inf_{Q: \int_{\Omega} X dQ = M} S(Q|P). \quad (5.26)$$

If $S \cap [m, M] = \emptyset$, then both sides in (5.22) are ∞ (by definition, $\inf \emptyset = \infty$). Hence,

$$\inf_{\theta \in S} I(\theta) = \inf_{\theta \in S \cap [m, M]} I(\theta) = \inf_{Q \in \Gamma_{S \cap [m, M]}} S(Q|P) = \inf_{Q \in \Gamma_S} S(Q|P),$$

and the statement follows. \square

Exercise 5.7. Prove the identities (5.26).

Propositions 5.8 and 5.9 yield the following generalization of Cramér's theorem:

Theorem 5.10 For any $S \subset \mathbb{R}$,

$$\begin{aligned} - \inf_{\theta \in \text{int} S} I(\theta) &\leq \liminf_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \omega \in \Omega^N \mid \frac{\mathcal{S}_N(\omega)}{N} \in S \right\} \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \omega \in \Omega^N \mid \frac{\mathcal{S}_N(\omega)}{N} \in S \right\} \leq - \inf_{\theta \in \text{cl} S} I(\theta). \end{aligned}$$

A set S is called *Cramer-nice* if

$$\inf_{\theta \in \text{int} S} I(\theta) = \inf_{\theta \in \text{cl} S} I(\theta).$$

Obviously, if S is Cramer-nice, then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log P_N \left\{ \omega \in \Omega^N \mid \frac{\mathcal{S}_N(\omega)}{N} \in S \right\} = - \inf_{\theta \in S} I(\theta).$$

Exercise 5.8.

1. Is it true that any open/closed interval is Cramér-nice?
2. Prove that any open set $S \subset]m, M[$ is Cramér-nice.
3. Describe all open sets that are Cramér-nice.

5.5 Notes and references

Theorem 5.1 goes back to the work of Hobson [Hob] in 1969. Following in Shannon's step, Hobson has proved Theorem 5.1 under the additional assumptions that \mathfrak{S} is continuous on \mathcal{A}_L for all $L \geq 1$, and that the function

$$(n, n_0) \mapsto \mathfrak{S} \left(\left(\frac{1}{n}, \dots, \frac{1}{n}, 0, \dots, 0 \right), \left(\frac{1}{n_0}, \dots, \frac{1}{n_0} \right) \right),$$

defined for $n \leq n_0$, is an increasing function of n_0 and a decreasing function of n . Our proof of Theorem 5.1 follows closely [Lei] where the reader can find additional information about the history of this result.

The formulation and the proof of Theorem 5.2 are based on the recent works [Mat, WiGaEi].

For additional information about axiomatizations of relative entropy we refer the reader to Section 7.2 in [AczDa].

Regarding Sanov's theorem, for the original references and additional information we refer the reader to [DeZe, CovTh]. In these monographs one can also find a purely combinatorial proof of Sanov's theorem and we urge the reader to study this alternative proof. As in the case of Cramér's theorem, the proof presented here has the advantage that it extends to a much more general setting that will be discussed in the Part II of the lecture notes.

Chapter 6

Fisher entropy

6.1 Definition and basic properties

Let Ω be a finite set and $[a, b]$ a bounded closed interval in \mathbb{R} . To avoid trivialities, we shall always assume that $|\Omega| = L > 1$. Let $\{P_\theta\}_{\theta \in [a, b]}$, $P_\theta \in \mathcal{P}_f(\Omega)$, be a family of faithful probability measures on Ω indexed by points $\theta \in [a, b]$. We shall assume that the functions $[a, b] \ni \theta \mapsto P_\theta(\omega)$ are C^2 (twice continuously differentiable) for all $\omega \in \Omega$. The expectation and variance with respect to P_θ are denoted by \mathbb{E}_θ and Var_θ . The entropy function is denoted by $S_\theta = -\log P_\theta$. The derivatives w.r.t. θ are denoted as $\dot{f}(\theta) = \partial_\theta f(\theta)$, $\ddot{f}(\theta) = \partial_\theta^2 f(\theta)$, etc. Note that

$$\dot{S}_\theta = -\frac{\dot{P}_\theta}{P_\theta}, \quad \ddot{S}_\theta = -\frac{\ddot{P}_\theta}{P_\theta} + \frac{\dot{P}_\theta^2}{P_\theta^2}, \quad \mathbb{E}_\theta(\dot{S}_\theta) = 0.$$

The Fisher entropy of P_θ is defined by

$$\mathcal{I}(\theta) = \mathbb{E}_\theta([\dot{S}_\theta]^2) = \sum_{\omega \in \Omega} \frac{[\dot{P}_\theta(\omega)]^2}{P_\theta(\omega)}.$$

Obviously,

$$\mathcal{I}(\theta) = \text{Var}_\theta(\dot{S}_\theta) = \mathbb{E}_\theta(\ddot{S}_\theta).$$

Example 6.1 Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable and

$$P_\theta(\omega) = \frac{e^{\theta X(\omega)}}{\sum_{\omega'} e^{\theta X(\omega')}}.$$

Then

$$\mathcal{I}(\theta) = \text{Var}_\theta(X).$$

The Fisher entropy arises by considering local relative entropy distortion of P_θ . Fix $\theta \in I$ and set

$$L(\epsilon) = S(P_{\theta+\epsilon}|P_\theta), \quad R(\epsilon) = S(P_\theta|P_{\theta+\epsilon}).$$

The functions $\epsilon \mapsto L(\epsilon)$ and $\epsilon \mapsto R(\epsilon)$ are well-defined in a neighbourhood of θ (relative to the interval $[a, b]$). An elementary computation yields:

Proposition 6.1

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon^2} L(\epsilon) = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon^2} R(\epsilon) = \frac{1}{2} \mathcal{I}(\theta).$$

In terms of the Jensen-Shannon entropy and metric we have

Proposition 6.2

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon^2} S_{\text{JS}}(P_{\theta+\epsilon}, P_\theta) = \frac{1}{4} \mathcal{I}(\theta),$$

$$\lim_{\epsilon \rightarrow 0} \frac{1}{|\epsilon|} d_{\text{JS}}(P_{\theta+\epsilon}, P_\theta) = \frac{1}{2} \sqrt{\mathcal{I}(\theta)}.$$

Exercise 6.1. Prove Propositions 6.1 and 6.2.

Since the relative entropy is stochastically monotone, Proposition 6.1 implies that the Fisher entropy is also stochastically monotone. More precisely, let $[\Phi(\omega, \hat{\omega})]_{(\omega, \hat{\omega}) \in \Omega \times \hat{\Omega}}$ be a stochastic matrix and $\Phi : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\hat{\Omega})$ the induced stochastic map. Set

$$\hat{P}_\theta = \Phi(P_\theta),$$

and note that \hat{P}_θ is faithful. Let $\hat{\mathcal{I}}(\theta)$ be the Fisher entropy of \hat{P}_θ . Then

$$\hat{\mathcal{I}}(\theta) = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon^2} S(\hat{P}_{\theta+\epsilon} | \hat{P}_\theta) \leq \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon^2} S(P_{\theta+\epsilon} | P_\theta) = \mathcal{I}(\theta).$$

The inequality $\hat{\mathcal{I}}(\theta) \leq \mathcal{I}(\theta)$ can be directly proven as follows. Since the function $x \mapsto x^2$ is convex, the Jensen inequality yields

$$\begin{aligned} \left(\sum_{\omega} \Phi(\omega, \hat{\omega}) \dot{P}_\theta(\omega) \right)^2 &= \left(\sum_{\omega} \Phi(\omega, \hat{\omega}) P_\theta(\omega) \frac{\dot{P}_\theta(\omega)}{P_\theta(\omega)} \right)^2 \\ &\leq \left(\sum_{\omega} \Phi(\omega, \hat{\omega}) \frac{[\dot{P}_\theta(\omega)]^2}{P_\theta(\omega)} \right) \left(\sum_{\omega} \Phi(\omega, \hat{\omega}) P_\theta(\omega) \right). \end{aligned}$$

Hence,

$$\begin{aligned} \hat{\mathcal{I}}(\theta) &= \sum_{\hat{\omega}} \left(\sum_{\omega} \Phi(\omega, \hat{\omega}) P_\theta(\omega) \right)^{-1} \left(\sum_{\omega} \Phi(\omega, \hat{\omega}) \dot{P}_\theta(\omega) \right)^2 \\ &\leq \sum_{\hat{\omega}} \sum_{\omega} \Phi(\omega, \hat{\omega}) P_\theta(\omega) \frac{[\dot{P}_\theta(\omega)]^2}{P_\theta(\omega)} \\ &= \mathcal{I}(\theta). \end{aligned}$$

6.2 Entropic geometry

We continue with the framework of the previous section. In this section we again identify $\mathcal{P}_f(\Omega)$ with

$$\mathcal{P}_{L,f} = \left\{ (p_1, \dots, p_L) \in \mathbb{R}^L \mid p_k > 0, \sum_k p_k = 1 \right\}.$$

We view $\mathcal{P}_{L,f}$ as a surface in \mathbb{R}^L and write $p = (p_1, \dots, p_L)$. The family $\{P_\theta\}_{\theta \in [a,b]}$ is viewed as a map (we will also call it a path)

$$[a, b] \ni \theta \mapsto p_\theta = (p_{\theta 1}, \dots, p_{\theta L}) \in \mathcal{P}_{L,f},$$

where $p_{\theta k} = P_{\theta}(\omega_k)$. For the purpose of this section it suffices to assume that all such path are C^1 (that is, continuously differentiable). The tangent vector $\dot{p}_{\theta} = (\dot{p}_{\theta 1}, \dots, \dot{p}_{\theta L})$ satisfies $\sum_k \dot{p}_{\theta k} = 0$ and hence belongs to the hyperplane

$$\mathcal{T}_L = \left\{ \zeta = (\zeta_1, \dots, \zeta_L) \mid \sum_k \zeta_k = 0 \right\}.$$

The tangent space of the surface $\mathcal{P}_{L,f}$ is $T_L = \mathcal{P}_{L,f} \times \mathcal{T}_L$.

A Riemannian structure (abbreviated RS) on $\mathcal{P}_{L,f}$ is a family $g_L = \{g_{L,p}(\cdot, \cdot)\}_{p \in \mathcal{P}_L}$ of real inner products on \mathcal{T}_L such that for all $\zeta, \eta \in \mathcal{T}_L$ the map

$$\mathcal{P}_L \ni p \mapsto g_{L,p}(\zeta, \eta) \quad (6.1)$$

is continuous. The geometric notions (angles, length of curves, curvature...) on \mathcal{P}_L are defined with respect to the RS (to define some of them one needs additional regularity of the maps (6.1)). For example, the energy of the path $\theta \mapsto p_{\theta}$ is

$$\mathcal{E}([p_{\theta}]) = \int_a^b g_{L,p_{\theta}}(\dot{p}_{\theta}, \dot{p}_{\theta}) d\theta,$$

and its length is

$$\mathcal{L}([p_{\theta}]) = \int_a^b \sqrt{g_{L,p_{\theta}}(\dot{p}_{\theta}, \dot{p}_{\theta})} d\theta.$$

Jensen's inequality for integrals (which is proven by applying Jensen's inequality to Riemann sums) gives that

$$\mathcal{L}([p_{\theta}]) \geq [(b-a)\mathcal{E}([p_{\theta}])]^{1/2}. \quad (6.2)$$

The Fisher Riemannian structure (abbreviated FRS) is defined by

$$g_p^F(\zeta, \eta) = \sum_k \frac{1}{p_k} \zeta_k \eta_k.$$

In this case,

$$g_{p(\theta)}^F(\dot{p}_{\theta}, \dot{p}_{\theta}) = \mathcal{I}(\theta),$$

where $\mathcal{I}(\theta)$ is the Fisher entropy of P_{θ} . Hence,

$$\mathcal{E}([p_{\theta}]) = \int_a^b \mathcal{I}(\theta) d\theta, \quad \mathcal{L}([p_{\theta}]) = \int_a^b \sqrt{\mathcal{I}(\theta)} d\theta.$$

We have the following general bounds:

Proposition 6.3

$$\int_a^b \mathcal{I}(\theta) d\theta \geq \frac{1}{b-a} d_V(p_a, p_b)^2, \quad \int_a^b \sqrt{\mathcal{I}(\theta)} d\theta \geq d_V(p_a, p_b), \quad (6.3)$$

where d_V is the variational distance defined by (3.2).

Remark 6.1 The first inequality in (6.3) yields the "symetrized" version of Theorem 4.2. Let $p, q \in \mathcal{P}_{L,f}$ and consider the path $p_{\theta} = \theta p + (1-\theta)q$, $\theta \in [0, 1]$. Then

$$\int_0^1 \mathcal{I}(\theta) d\theta = S(p|q) + S(q|p),$$

and the first inequality in (6.3) gives

$$S(p|q) + S(q|p) \geq d_V(p, q)^2.$$

Proof. To prove the first inequality, note that Jensen's inequality gives

$$\mathcal{I}(\theta) = \sum_{k=1}^L \frac{\dot{p}_{\theta k}^2}{p_{\theta k}} = \sum_{k=1}^L \left[\frac{\dot{p}_{\theta k}}{p_{\theta k}} \right]^2 p_{\theta k} \geq \left(\sum_{k=1}^L |\dot{p}_{\theta k}| \right)^2. \quad (6.4)$$

Hence,

$$\int_a^b \mathcal{I}(\theta) d\theta \geq \int_a^b \left(\sum_{k=1}^L |\dot{p}_{\theta k}| \right)^2 d\theta \geq \frac{1}{b-a} \left(\sum_{k=1}^L \int_a^b |\dot{p}_{\theta k}| d\theta \right)^2,$$

where the second inequality follows from Jensen's integral inequality. The last inequality and

$$\int_a^b |\dot{p}_{\theta k}| d\theta \geq \left| \int_a^b \dot{p}_{\theta k} d\theta \right| = |p_{bk} - p_{ak}| \quad (6.5)$$

yield the statement.

Note that the first inequality in (6.3) and (6.2) imply the second. Alternatively, the second inequality follows immediately from (6.4) and (6.5). \square

The geometry induced by the FRS can be easily understood in terms of the surface

$$\mathfrak{S}_L = \{s = (s_1, \dots, s_L) \in \mathbb{R}^L \mid s_k > 0, \sum_k s_k^2 = 1\}.$$

The respective tangent space is $\mathfrak{S}_L \times \mathbb{R}^{L-1}$ which we equip with the Euclidian RS

$$e_s(\zeta, \eta) = \sum_k \zeta_k \eta_k.$$

Note that $e_s(\zeta, \eta)$ does not depend on $s \in \mathfrak{S}_L$ and we will drop the subscript s . Let now $\theta \mapsto p_\theta = (p_{\theta 1}, \dots, p_{\theta L})$ be a path connecting $p = (p_1, \dots, p_L)$ and $q = (q_1, \dots, q_L)$ in $\mathcal{P}_{L,f}$. Then,

$$\theta \mapsto s_\theta = (\sqrt{p_{\theta 1}}, \dots, \sqrt{p_{\theta L}})$$

is a path in \mathfrak{S}_L connecting $s = (\sqrt{p_1}, \dots, \sqrt{p_L})$ and $u = (\sqrt{q_1}, \dots, \sqrt{q_L})$. The map $[p_\theta] \mapsto [s_\theta]$ is a bijective correspondences between all C^1 -paths in $\mathcal{P}_{L,f}$ connecting p and q and all C^1 -paths in \mathfrak{S}_L connecting s and u . Since

$$e(\dot{s}_\theta, \dot{s}_\theta) = \frac{1}{4} g_{p(\theta)}^F(\dot{p}_\theta, \dot{p}_\theta) = \frac{1}{4} \mathcal{I}(\theta),$$

the geometry on $\mathcal{P}_{L,f}$ induced by the FRS is identified with the Euclidian geometry of \mathfrak{S}_L via the map $[p_\theta] \mapsto [s_\theta]$.

Exercise 6.2. The geodesic distance between $p, q \in \mathcal{P}_{L,f}$ w.r.t. the FRS is defined by

$$\gamma(p, q) = \inf \int_a^b \sqrt{g_{p(\theta)}^F(\dot{p}_\theta, \dot{p}_\theta)} d\theta, \quad (6.6)$$

where inf is taken over all C^1 -paths $[a, b] \ni \theta \mapsto p_\theta \in \mathcal{P}_{L,f}$ such that $p_a = p$ and $p_b = q$. Prove that

$$\gamma(p, q) = \arccos \left(\sum_{k=1}^L \sqrt{p_k q_k} \right).$$

Show that the r.h.s. in (6.6) has a unique minimizer and identify this minimizer.

The obvious hint for a solution of this exercise is to use the correspondence between the Euclidian geometry of the sphere and the FRS geometry of $\mathcal{P}_{L,f}$. We leave it to the interested reader familiar with basic notions of differential geometry to explore this connection further. For example, can you compute the sectional curvature of $\mathcal{P}_{L,f}$ w.r.t. the FRS?

6.3 Chentsov's theorem

Let $(g_L)_{L \geq 2}$ be a sequence of RS, where g_L is a RS on $\mathcal{P}_{L,f}$. The sequence $(g_L)_{L \geq 2}$ is called stochastically monotone if for any $L, \widehat{L} \geq 2$ and any stochastic map $\Phi : \mathcal{P}_{L,f} \rightarrow \mathcal{P}_{\widehat{L},f}$,

$$g_{\widehat{L},\Phi(p)}(\Phi(\zeta), \Phi(\zeta)) \leq g_{L,p}(\zeta, \zeta)$$

for all $p \in \mathcal{P}_{L,f}$ and $\zeta \in \mathcal{T}_L$. Here we used that, in the obvious way, Φ defines a linear map $\Phi : \mathbb{R}^L \mapsto \mathbb{R}^{\widehat{L}}$ which maps \mathcal{T}_L to $\mathcal{T}_{\widehat{L}}$.

Proposition 6.4 *The sequence $(g_L^F)_{L \geq 1}$ of the FRS is stochastically monotone.*

Proof. The argument is a repetition of the direct proof of the inequality $\mathcal{I}(\theta) \leq \widehat{\mathcal{I}}(\theta)$ given in Section 6.1. The details are as follows.

Let $[\Phi(i, j)]_{1 \leq i \leq L, 1 \leq j \leq \widehat{L}}$ be a stochastic matrix defining $\Phi : \mathcal{P}_{L,f} \rightarrow \mathcal{P}_{\widehat{L},f}$, i.e., for any $v = (v_1, \dots, v_L) \in \mathbb{R}^L$, $\Phi(v) \in \mathbb{R}^{\widehat{L}}$ is given by

$$(\Phi(v))_j = \sum_{i=1}^L \Phi(i, j) v_i.$$

For $p \in \mathcal{P}_L$ and $\zeta \in \mathcal{T}_L$ the convexity gives

$$\begin{aligned} \left(\sum_i \Phi(i, j) \zeta_i \right)^2 &= \left(\sum_i \Phi(i, j) p_i \frac{\zeta_i}{p_i} \right)^2 \leq \left(\sum_i \Phi(i, j) \frac{\zeta_i^2}{p_i} \right) \left(\sum_i \Phi(i, j) p_i \right) \\ &= \left(\sum_i \Phi(i, j) \frac{\zeta_i^2}{p_i} \right) (\Phi(p))_j. \end{aligned}$$

Hence,

$$\begin{aligned} g_{\widehat{L}}^F(\Phi(\zeta), \Phi(\zeta)) &= \sum_j \frac{1}{(\Phi(p))_j} \left(\sum_i \Phi(i, j) \zeta_i \right)^2 \\ &\leq \sum_j \sum_i \Phi(i, j) \frac{\zeta_i^2}{p_i} = \sum_i \frac{\zeta_i^2}{p_i} = g_{L,p}^F(\zeta, \zeta). \end{aligned}$$

□

The main result of this section is:

Theorem 6.5 *Suppose that a sequence $(g_L)_{L \geq 2}$ is stochastically monotone. Then there exists a constant $c > 0$ such that $g_L = c g_L^F$ for all $L \geq 2$.*

Proof. We start the proof by extending each $g_{L,p}$ to a bilinear map $G_{L,p}$ on $\mathbb{R}^L \times \mathbb{R}^L$ as follows. Set $\nu_L = (1, \dots, 1) \in \mathbb{R}^L$ and note that any $v \in \mathbb{R}^L$ can be uniquely written as $v = a \nu_L + \zeta$, where $a \in \mathbb{R}$ and $\zeta \in \mathcal{T}_L$. If $v = a \nu_L + \zeta$ and $w = a' \nu_L + \zeta'$, we set

$$G_{L,p}(v, w) = g_{L,p}(\zeta, \zeta').$$

The map $G_{L,p}$ is obviously bilinear, symmetric ($G_{L,p}(v, w) = G_{L,p}(w, v)$), and non-negative ($G_{L,p}(v, v) \geq 0$). In particular, the polarization identity holds:

$$G_{L,p}(v, w) = \frac{1}{4} (G_{L,p}(v+w, v+w) - G_{L,p}(v-w, v-w)). \quad (6.7)$$

Note however that $G_{L,p}$ is not an inner product since $G_{L,p}(\nu_L, \nu_L) = 0$.

In what follows $p_{L,\text{ch}}$ denotes the chaotic probability distribution in \mathcal{P}_L , i.e., $p_{L,\text{ch}} = (1/L, \dots, 1/L)$. A basic observation is that if the stochastic map $\Phi : \mathcal{P}_{L,f} \rightarrow \mathcal{P}_{\widehat{L},f}$ is stochastically invertible (that is, there exists a stochastic map $\Psi : \mathcal{P}_{\widehat{L},f} \rightarrow \mathcal{P}_{L,f}$ such that $\Phi \circ \Psi(p) = p$ for all $p \in \mathcal{P}_{L,f}$) and $\Phi(p_{L,\text{ch}}) = p_{\widehat{L},\text{ch}}$, then for all $v, w \in \mathbb{R}^L$,

$$G_{\widehat{L},p_{\widehat{L},\text{ch}}}(\Phi(v), \Phi(w)) = G_{L,p_{L,\text{ch}}}(v, w). \quad (6.8)$$

To prove this, note that since Φ preserves the chaotic probability distribution, we have that $\Phi(\nu_L) = L\widehat{L}^{-1}\nu_{\widehat{L}}$. Then, writing $v = a\nu_L + \zeta$, we have

$$\begin{aligned} G_{L,p_{L,\text{ch}}}(v, v) &= g_{L,p_{L,\text{ch}}}(\zeta, \zeta) \geq g_{\widehat{L},p_{\widehat{L},\text{ch}}}(\Phi(\zeta), \Phi(\zeta)) \\ &= G_{\widehat{L},p_{\widehat{L},\text{ch}}}\left(aL\widehat{L}^{-1}\nu_{\widehat{L}} + \Phi(\zeta), aL\widehat{L}^{-1}\nu_{\widehat{L}} + \Phi(\zeta)\right) \\ &= G_{\widehat{L},p_{\widehat{L},\text{ch}}}(a\Phi(\nu_L) + \Phi(\zeta), a\Phi(\nu_L) + \Phi(\zeta)) \\ &= G_{\widehat{L},p_{\widehat{L},\text{ch}}}(\Phi(v), \Phi(v)). \end{aligned} \quad (6.9)$$

If $\Psi : \mathcal{P}_{\widehat{L},f} \rightarrow \mathcal{P}_{L,f}$ is the stochastic inverse of Φ , then $\Psi(p_{\widehat{L},\text{ch}}) = p_{L,\text{ch}}$ and so by repeating the above argument we get

$$G_{\widehat{L},p_{\widehat{L},\text{ch}}}(\Phi(v), \Phi(v)) \geq G_{L,p_{L,\text{ch}}}(\Psi(\Phi(v)), \Psi(\Phi(v))) = G_{L,p_{L,\text{ch}}}(v, v). \quad (6.10)$$

The inequalities (6.9) and (6.10) yield (6.8) in the case $v = w$. The polarization identity (6.7) then yields the statement for all vectors v and w .

We proceed to identify $G_{\widehat{L},p_{\widehat{L},\text{ch}}}$ and $g_{\widehat{L},p_{\widehat{L},\text{ch}}}$. The identity (6.8) will play a central role in this part of the argument. Let $e_{L,k}$, $k = 1, \dots, L$, be the standard basis of \mathbb{R}^L . Let π be a permutation of $\{1, \dots, L\}$. Then for all $1 \leq j, k \leq L$,

$$G_{p_{L,\text{ch}}}(e_{L,j}, e_{L,k}) = G_{p_{L,\text{ch}}}(e_{L,\pi(j)}, e_{L,\pi(k)}). \quad (6.11)$$

To establish (6.11), we use (6.8) with $\Phi : \mathcal{P}_{L,f} \rightarrow \mathcal{P}_{L,f}$ defined by

$$\Phi((p_1, \dots, p_L)) = (p_{\pi(1)}, \dots, p_{\pi(L)}).$$

Note that Φ is stochastically invertible with the inverse

$$\Psi((p_1, \dots, p_L)) = (p_{\pi^{-1}(1)}, \dots, p_{\pi^{-1}(L)}),$$

and that $\Phi(p_{L,\text{ch}}) = p_{L,\text{ch}}$. An immediate consequence of the (6.11) is that for all k, j ,

$$G_{p_{L,\text{ch}}}(e_{L,j}, e_{L,j}) = G_{p_{L,\text{ch}}}(e_{L,k}, e_{L,k}), \quad (6.12)$$

and that for all pairs $(j, k), (j', k')$ with $j \neq j'$ and $k \neq k'$,

$$G_{p_{L,\text{ch}}}(e_{L,j}, e_{L,k}) = G_{p_{L,\text{ch}}}(e_{L,j'}, e_{L,k'}). \quad (6.13)$$

We introduce the constants

$$c_L = G_{p_{L,\text{ch}}}(e_{L,j}, e_{L,j}), \quad b_L = G_{p_{L,\text{ch}}}(e_{L,j}, e_{L,k}),$$

where $j \neq k$. By (6.12) and (6.13), these constants do not depend on the choice of j, k . We now show that there exist constants $c, b \in \mathbb{R}$ such that for all $L \geq 2$, $c_L = cL + b$ and $b_L = b$. To prove this, let $L, L' \geq 2$ and consider the stochastic map $\Phi : \mathcal{P}_{L,f} \rightarrow \mathcal{P}_{L'L',f}$ defined by

$$\Phi((p_1, \dots, p_L)) = \left(\frac{p_1}{L'}, \dots, \frac{p_1}{L'}, \dots, \frac{p_L}{L'}, \dots, \frac{p_L}{L'}\right),$$

where each term p_k/L' is repeated L' times. This map is stochastically invertible with the inverse

$$\Psi\left((p_1^{(1)}, \dots, p_{L'}^{(1)}, \dots, p_1^{(L)}, \dots, p_{L'}^{(L)})\right) = \left(\sum_{k=1}^{L'} p_k^{(1)}, \dots, \sum_{k=1}^{L'} p_k^{(L)}\right).$$

Since $\Phi(p_{L,\text{ch}}) = p_{LL',\text{ch}}$, (6.8) holds. Combining (6.8) with the definition b_L , we derive that

$$b_L = b_{LL'} = b_{L'}.$$

Set $b = b_L$. Then, for $L, L' \geq 2$, (6.8) and the definition of c_L give

$$c_L = \frac{1}{L'} c_{LL'} + \frac{L'(L'-1)}{(L')^2} b_{LL'} = \frac{1}{L'} c_{LL'} + \frac{L'(L'-1)}{(L')^2} b,$$

and so

$$c_L - b = \frac{1}{L'} (c_{LL'} - b).$$

Hence,

$$\frac{1}{L} (c_L - b) = \frac{1}{LL'} (c_{LL'} - b) = \frac{1}{L'} (c_{L'} - b),$$

and we conclude that

$$c_L = cL + b$$

for some $c \in \mathbb{R}$. It follows that for $v, w \in \mathbb{R}^L$,

$$G_{p_{L,\text{ch}}}(v, w) = cL \sum_{k=1}^L v_k w_k + b \left(\sum_{k=1}^L v_k \right) \left(\sum_{k=1}^L w_k \right).$$

and that for $\zeta, \eta \in \mathcal{T}_L$,

$$g_{p_{L,\text{ch}}}(\zeta, \eta) = cL \sum_{k=1}^L \zeta_k \eta_k. \quad (6.14)$$

The last relation implies in particular that $c > 0$. Note that (6.14) can be written as $g_{L,p_{L,\text{ch}}} = c g_{L,p_{\text{ch}}}^F$, proving the statement of the theorem for the special values $p = p_{L,\text{ch}}$.

The rest of the argument is based on the relation (6.14). By essentially repeating the proof of the identity (6.8) one easily shows that if $\Phi : \mathcal{P}_{L,\text{f}} \rightarrow \mathcal{P}_{\widehat{L},\text{f}}$ is stochastically invertible, then for all $p \in \mathcal{P}_{L,\text{f}}$ and $\zeta, \eta \in \mathcal{T}_L$,

$$g_{L,\Phi(p)}(\Phi(\zeta), \Phi(\eta)) = g_{L,p}(\zeta, \eta). \quad (6.15)$$

Let now $\bar{p} = (\bar{p}_1, \dots, \bar{p}_L) \in \mathcal{P}_{L,\text{f}}$ be such that all \bar{p}_k 's are rational numbers. We can write

$$\bar{p} = \left(\frac{\ell_1}{L'}, \dots, \frac{\ell_L}{L'} \right).$$

where all ℓ_k 's are integers ≥ 1 and $\sum_k \ell_k = L'$. Let $\Phi : \mathcal{P}_{L,\text{f}} \rightarrow \mathcal{P}_{L',\text{f}}$ be a stochastic map defined by

$$\Phi((p_1, \dots, p_L)) = \left(\frac{p_1}{\ell_1}, \dots, \frac{p_1}{\ell_1}, \dots, \frac{p_L}{\ell_L}, \dots, \frac{p_L}{\ell_L} \right),$$

where each term p_k/ℓ_k is repeated ℓ_k times. The map Φ is stochastically invertible and its inverse is

$$\Psi((p_1^{(1)}, \dots, p_{\ell_1}^{(1)}, \dots, p_1^{(\ell_L)}, \dots, p_{\ell_L}^{(\ell_L)})) = \left(\sum_{k=1}^{\ell_1} p_k^{\ell_1}, \dots, \sum_{k=1}^{\ell_L} p_k^{\ell_L} \right).$$

Note that $\Phi(\bar{p}) = p_{L',\text{ch}}$, and so

$$g_{L,\bar{p}}(\zeta, \eta) = g_{L',p_{L',\text{ch}}}(\Phi(\zeta), \Phi(\eta)) = c \sum_{k=1}^L \frac{L'}{\ell_k} \zeta_k \eta_k = c g_{L,\bar{p}}^F(\zeta, \eta). \quad (6.16)$$

Since the set of all \bar{p} 's in $\mathcal{P}_{L,\text{f}}$ whose all components are rational is dense in $\mathcal{P}_{L,\text{f}}$ and since the map $p \mapsto g_{L,p}(\zeta, \eta)$ is continuous, it follows from (6.16) that for all $L \geq 2$ and all $p \in \mathcal{P}_{L,\text{f}}$,

$$g_{L,p} = c g_{L,p}^F.$$

This completes the proof of Chentsov's theorem. □

6.4 Notes and references

The Fisher entropy (also often called Fisher information) was introduced by Fisher in [Fis1] and plays a fundamental role in statistics (this is the topic of the next chapter). Although Fisher's work precedes Shannon's by twenty three years, it apparently played no role in the genesis of the information theory. The first mentioning of the Fisher entropy in context of information theory goes back to [KullLe] where Proposition 6.1 was stated.

The geometric interpretation of the Fisher entropy is basically built in its definition. We shall return to this point in the Part II of the lecture notes where the reader can find references to the vast literature on this topic.

Chentsov's theorem goes back to [Cen]. Our proof is based on the elegant arguments of Campbel [Cam].

Chapter 7

Parameter estimation

7.1 Introduction

Let \mathcal{A} be a set and $\{P_\theta\}_{\theta \in \mathcal{A}}$ a family of probability measures on a finite set Ω . We shall refer to the elements of \mathcal{A} as *parameters*. Suppose that a probabilistic experiment is described by one unknown member of this family. By performing a trial we wish to choose the unknown parameter θ such that P_θ is the most likely description of the experiment. To predict θ one chooses a function $\hat{\theta} : \Omega \rightarrow \mathcal{A}$ which, in the present context, is called an *estimator*. If the outcome of a trial is $\omega \in \Omega$, then the value $\theta = \hat{\theta}(\omega)$ is the prediction of the unknown parameter and the probability. Obviously, a reasonable estimator should satisfy a reasonable requirements, and we will return to this point shortly.

The hypothesis testing, described in Section 4.7, is the simplest non-trivial example of the above setting with $\mathcal{A} = \{0, 1\}$, $P_0 = P$ and $P_1 = Q$ (we also assume that the priors are $p = q = 1/2$.) The estimators are identified with characteristic functions $\hat{\theta} = \chi_T$, $T \subset \Omega$. With an obvious change of vocabulary, the mathematical theory described in Section 4.7 can be viewed as a theory of parameter estimation in the case where \mathcal{A} has two elements.

Here we shall assume that \mathcal{A} is a bounded closed interval $[a, b]$ and we shall explore the conceptual and mathematical aspects the continuous set of parameters brings to the problem of estimation. The Fisher entropy will play an important role in this development. We continue with the notation and assumptions introduced in the beginning of Section 6.1, and start with some preliminaries.

A *loss function* is a map $L : \mathbb{R} \times [a, b] \rightarrow \mathbb{R}_+$ such that $L(x, \theta) \geq 0$ and $L(x, \theta) = 0$ iff $x = \theta$. To a given loss function and the estimator $\hat{\theta}$, one associates the *risk function* by

$$R(\hat{\theta}, \theta) = E_\theta(L(\hat{\theta}, \theta)) = \sum_{\omega \in \Omega} L(\hat{\theta}(\omega), \theta) P_\theta(\omega).$$

Once a choice of the loss function is made, the goal is to find an estimator that will minimize the risk function subject to appropriate consistency requirements.

We shall work only with the quadratic loss function $L(x, \theta) = (x - \theta)^2$. In this case, the risk function is

$$E_\theta((\hat{\theta} - \theta)^2) = \text{Var}_\theta(\hat{\theta}).$$

7.2 Basic facts

The following general estimate is known as the Cramér-Rao bound.

Proposition 7.1 For any estimator $\hat{\theta}$ and all $\theta \in [a, b]$,

$$\frac{[\dot{E}_{\theta}(\hat{\theta})]^2}{\mathcal{I}(\theta)} \leq E_{\theta}((\hat{\theta} - \theta)^2).$$

Proof.

$$\dot{E}_{\theta}(\hat{\theta}) = \sum_{\omega \in \Omega} \hat{\theta}(\omega) \dot{P}_{\theta}(\omega) = \sum_{\omega \in \Omega} (\hat{\theta}(\omega) - \theta) \dot{P}_{\theta}(\omega).$$

Writing $\dot{P}_{\theta}(\omega) = \dot{P}_{\theta}(\omega) \sqrt{P_{\theta}(\omega)} / \sqrt{P_{\theta}(\omega)}$ and applying the Cauchy-Schwartz inequality one gets

$$\begin{aligned} |\dot{E}_{\theta}(\hat{\theta})| &\leq \left(\sum_{\omega \in \Omega} (\hat{\theta}(\omega) - \theta)^2 P_{\theta}(\omega) \right)^{1/2} \left(\sum_{\omega \in \Omega} \frac{[\dot{P}_{\theta}(\omega)]^2}{P_{\theta}(\omega)} \right)^{1/2} \\ &= \left(E_{\theta}((\hat{\theta} - \theta)^2) \right)^{1/2} \sqrt{\mathcal{I}(\theta)}. \end{aligned}$$

□

As in the case of hypothesis testing, multiple trials improve the errors in the parameter estimation. Passing to the product space Ω^N and the product probability measure $P_{\theta N}$, and denoting by $E_{\theta N}$ the expectation w.r.t. $P_{\theta N}$, the Cramér-Rao bound takes the following form.

Proposition 7.2 For any estimator $\hat{\theta}_N : \Omega^N \rightarrow [a, b]$ and all $\theta \in [a, b]$,

$$\frac{1}{N} \frac{[\dot{E}_{\theta N}(\hat{\theta}_N)]^2}{\mathcal{I}(\theta)} \leq E_{\theta N}((\hat{\theta}_N - \theta)^2).$$

Proof.

$$\begin{aligned} \dot{E}_{\theta N}(\hat{\theta}_N) &= \sum_{\omega=(\omega_1, \dots, \omega_N) \in \Omega^N} \sum_{k=1}^N (\hat{\theta}_N(\omega) - \theta) P_{\theta}(\omega_1) \cdots \dot{P}_{\theta}(\omega_k) \cdots P_{\theta}(\omega_N) \\ &= \sum_{\omega=(\omega_1, \dots, \omega_N) \in \Omega^N} \left(\sum_{k=1}^N \frac{\dot{P}_{\theta}(\omega_k)}{P_{\theta}(\omega_k)} \right) (\hat{\theta}_N(\omega) - \theta) P_{\theta N}(\omega). \end{aligned}$$

Applying the Cauchy-Schwarz inequality

$$\int_{\Omega^N} fg dP_{\theta N} \leq \left(\int_{\Omega^N} f^2 dP_{\theta N} \right)^{1/2} \left(\int_{\Omega^N} g^2 dP_{\theta N} \right)^{1/2}$$

with

$$f(\omega) = \sum_{k=1}^N \frac{\dot{P}_{\theta}(\omega_k)}{P_{\theta}(\omega_k)}, \quad g(\omega) = \hat{\theta}_N(\omega) - \theta,$$

one gets

$$\begin{aligned} |\dot{E}_{\theta N}(\hat{\theta})| &\leq \left(\sum_{\omega \in \Omega^N} (\hat{\theta}_N(\omega) - \theta)^2 P_{\theta N}(\omega) \right)^{1/2} \left(\sum_{\omega=(\omega_1, \dots, \omega_N)} \sum_{k=1}^N \frac{[\dot{P}_{\theta}(\omega_k)]^2}{[P_{\theta}(\omega_k)]^2} P_{\theta N}(\omega) \right)^{1/2} \\ &= \left(E_{\theta N}((\hat{\theta}_N - \theta)^2) \right)^{1/2} \sqrt{N\mathcal{I}(\theta)}. \end{aligned}$$

□

We now describe the *consistency* requirement. In a nutshell, the consistency states that if the experiment is described by P_θ , then the estimator should statistically return the value θ . An ideal consistency would be $E_{\theta N}(\hat{\theta}_N) = \theta$ for all $\theta \in [a, b]$. However, it is clear that in our setting such estimator cannot exist. Indeed, using that $\hat{\theta}$ takes values in $[a, b]$, the relations $E_{a N}(\hat{\theta}_N) = a$ and $E_{b N}(\hat{\theta}_N) = b$ give that $\hat{\theta}_N(\omega) = a$ and $\hat{\theta}_N(\omega) = b$ for all $\omega \in \Omega^N$. Requiring $E_{\theta N}(\hat{\theta}_N) = \theta$ only for $\theta \in]a, b[$ does not help, and the remaining possibility is to formulate the consistency in an asymptotic setting.

Definition 7.3 A sequence of estimators $\hat{\theta}_N : \Omega^N \rightarrow [a, b]$, $N = 1, 2, \dots$, is called *consistent* if

$$\lim_{N \rightarrow \infty} E_{\theta N}(\hat{\theta}_N) = \theta$$

for all $\theta \in [a, b]$, and *uniformly consistent* if

$$\lim_{N \rightarrow \infty} \sup_{\theta \in [a, b]} E_{\theta N}(|\hat{\theta} - \theta|) = 0.$$

Finally, we introduce the notion of *efficiency*.

Definition 7.4 Let $\hat{\theta}_N : \Omega^N \rightarrow [a, b]$, $N = 1, 2, \dots$ be a sequence of estimators. A continuous function $\mathcal{E} :]a, b[\rightarrow \mathbb{R}_+$ is called the *efficiency* of $(\hat{\theta}_N)_{N \geq 1}$ if

$$\lim_{N \rightarrow \infty} N E_{\theta N} \left((\hat{\theta} - \theta)^2 \right) = \mathcal{E}(\theta) \quad (7.1)$$

for all $\theta \in]a, b[$. The sequence $(\hat{\theta}_N)_{N \geq 1}$ is called *uniformly efficient* if in addition for any $[a', b'] \subset]a, b[$,

$$\lim_{N \rightarrow \infty} \sup_{\theta \in [a', b']} \left| N E_{\theta N} \left((\hat{\theta} - \theta)^2 \right) - \mathcal{E}(\theta) \right| = 0. \quad (7.2)$$

To remain on a technically elementary level, we will work only with uniformly efficient estimators. The reason for staying away from the boundary points a and b in the definition of efficiency is somewhat subtle and we will elucidate it in Remark 7.2.

Proposition 7.5 Let $(\hat{\theta}_N)_{N \geq 1}$ be a uniformly efficient consistent sequence of estimators. Then its efficiency \mathcal{E} satisfies

$$\mathcal{E}(\theta) \geq \frac{1}{\mathcal{I}(\theta)}$$

for all $\theta \in]a, b[$.

Proof. Fix $\theta_1, \theta_2 \in]a, b[$, $\theta_1 < \theta_2$. The consistency gives

$$\theta_2 - \theta_1 = \lim_{N \rightarrow \infty} \left[E_{\theta_2 N}(\hat{\theta}_N) - E_{\theta_1 N}(\hat{\theta}_N) \right]. \quad (7.3)$$

The Cramér-Rao bound yields the estimate

$$\begin{aligned} E_{\theta_2 N}(\hat{\theta}_N) - E_{\theta_1 N}(\hat{\theta}_N) &= \int_{\theta_1}^{\theta_2} \dot{E}_{\theta N}(\hat{\theta}_N) d\theta \leq \int_{\theta_1}^{\theta_2} |\dot{E}_{\theta N}(\hat{\theta}_N)| d\theta \\ &\leq \int_{\theta_1}^{\theta_2} \left[N \mathcal{I}(\theta) E_{\theta N} \left((\hat{\theta}_N - \theta)^2 \right) \right]^{1/2} d\theta. \end{aligned} \quad (7.4)$$

Finally, the uniform efficiency gives

$$\begin{aligned} \lim_{N \rightarrow \infty} \int_{\theta_1}^{\theta_2} \left[N\mathcal{I}(\theta)E_{\theta N} \left((\hat{\theta}_N - \theta)^2 \right) \right]^{1/2} d\theta &= \int_{\theta_1}^{\theta_2} \lim_{N \rightarrow \infty} \left[N\mathcal{I}(\theta)E_{\theta N} \left((\hat{\theta}_N - \theta)^2 \right) \right]^{1/2} d\theta \\ &= \int_{\theta_1}^{\theta_2} \sqrt{\mathcal{I}(\theta)\mathcal{E}(\theta)} d\theta. \end{aligned} \quad (7.5)$$

Combining (7.3), (7.4), and (7.5), we derive that

$$\theta_2 - \theta_1 \leq \int_{\theta_1}^{\theta_2} \sqrt{\mathcal{I}(\theta)\mathcal{E}(\theta)} d\theta$$

for all $a \leq \theta_1 < \theta_2 \leq b$. Hence, $\sqrt{\mathcal{I}(\theta)\mathcal{E}(\theta)} \geq 1$ for all $\theta \in]a, b[$, and the statement follows. \square

In Section 7.4 we shall construct a uniformly consistent and uniformly efficient sequence of estimators whose efficiency is equal to $1/\mathcal{I}(\theta)$ for all $\theta \in]a, b[$. This sequence of estimators saturates the bound of Proposition 7.5 and in that sense is the best possible one. In Remark 7.2 we shall also exhibit a concrete example of such estimator sequence for which the limit (7.1) also exists for $\theta = a$ and satisfies $\mathcal{E}(a) < 1/\mathcal{I}(a)$. This shows that Proposition 7.5 is an optimal result.

7.3 Two remarks

The first remark is that the existence of a consistent estimator sequence obviously implies that

$$\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}. \quad (7.6)$$

In Section 7.4 we shall assume that (7.6) holds and refer to it as the *identifiability* property of our starting family of probability measures $\{P_\theta\}_{\theta \in [a, b]}$.

The second remark concerns the LLN adapted to the parameter setting, which will play a central role in the proofs of the next section. This variant of the LLN is of independent interest, and for this reason we state it and prove it separately.

Proposition 7.6 *Let $X_\theta : \Omega \rightarrow \mathbb{R}$, $\theta \in [a, b]$, be random variables such that the map $[a, b] \ni \theta \mapsto X_\theta(\omega)$ is continuous for all $\omega \in \Omega$. Set*

$$\mathcal{S}_{\theta N}(\omega = (\omega_1, \dots, \omega_N)) = \sum_{k=1}^N X_\theta(\omega_k).$$

Then for any $\epsilon > 0$,

$$\lim_{N \rightarrow \infty} \sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_{\theta'}(X_{\theta'}) \right| \geq \epsilon \right\} = 0. \quad (7.7)$$

Moreover, (7.7) can be refined as follows. For any $\epsilon > 0$ there are constants $C_\epsilon > 0$ and $\gamma_\epsilon > 0$ such that for all $N \geq 1$,

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_{\theta'}(X_{\theta'}) \right| \geq \epsilon \right\} \leq C_\epsilon e^{-\gamma_\epsilon N}. \quad (7.8)$$

Remark 7.1 The point of this result is uniformity in θ and θ' . Note that

$$\lim_{N \rightarrow \infty} P_{\theta N} \left\{ \omega \in \Omega^N \mid \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_{\theta'}(X_{\theta'}) \right| \geq \epsilon \right\} = 0$$

is the statement of the LLN, while

$$P_{\theta N} \left\{ \omega \in \Omega^N \mid \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_{\theta}(X_{\theta'}) \right| \geq \epsilon \right\} \leq C_{\epsilon} e^{-\gamma_{\epsilon} N},$$

with C_{ϵ} and γ_{ϵ} depending on θ, θ' , is the statement of the strong LLN formulated in Exercise 2.4.

Proof. By uniform continuity, there exists $\delta > 0$ such that for all $u, v \in [a, b]$ satisfying $|u - v| < \delta$ one has

$$\sup_{u' \in [a, b]} |E_{u'}(X_u) - E_{u'}(X_v)| < \frac{\epsilon}{4} \quad \text{and} \quad \sup_{\omega \in \Omega} |X_u(\omega) - X_v(\omega)| < \frac{\epsilon}{4}.$$

Let $a = \theta'_0 < \theta'_1 < \dots < \theta'_n = b$ be such that $\theta'_k - \theta'_{k-1} < \delta$. Then, for all $\theta \in [a, b]$,

$$\left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_{\theta}(X_{\theta'}) \right| \geq \epsilon \right\} \subset \bigcup_{k=1}^n \left\{ \omega \in \Omega^N \mid \left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - E_{\theta}(X_{\theta'_k}) \right| \geq \frac{\epsilon}{2} \right\}. \quad (7.9)$$

It follows that (recall the proof of the LLN, Proposition 2.2)

$$\begin{aligned} P_{\theta N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_{\theta}(X_{\theta'}) \right| \geq \epsilon \right\} &\leq \sum_{k=1}^n P_{\theta N} \left\{ \omega \in \Omega^N \mid \left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - E_{\theta}(X_{\theta'_k}) \right| \geq \frac{\epsilon}{2} \right\} \\ &\leq \frac{4}{\epsilon^2} \sum_{k=1}^n E_{\theta N} \left(\left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - E_{\theta}(X_{\theta'_k}) \right|^2 \right) \\ &\leq \frac{4}{\epsilon^2} \frac{1}{N} \sum_{k=1}^n E_{\theta} \left(|X_{\theta'_k} - E_{\theta}(X_{\theta'_k})|^2 \right). \end{aligned} \quad (7.10)$$

Setting

$$C = \max_{1 \leq k \leq n} \max_{\theta, \theta' \in [a, b]} E_{\theta} \left(|X_{\theta'} - E_{\theta}(X_{\theta'})|^2 \right),$$

we derive that

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_{\theta}(X_{\theta'}) \right| \geq \epsilon \right\} \leq \frac{4}{\epsilon^2} \frac{Cn}{N},$$

and (7.7) follows.

The proof of (7.8) also starts with (7.9) and follows the argument of Proposition 2.9 (recall the Exercise 2.4). The details are as follows. Let $\alpha > 0$. Then for any θ and k ,

$$\begin{aligned} P_{\theta N} \left\{ \omega \in \Omega^N \mid \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - E_{\theta}(X_{\theta'_k}) \geq \frac{\epsilon}{2} \right\} &= P_{\theta N} \left\{ \omega \in \Omega^N \mid \mathcal{S}_{\theta'_k N}(\omega) \geq N \frac{\epsilon}{2} + N E_{\theta}(X_{\theta'_k}) \right\} \\ &= P_{\theta N} \left\{ \omega \in \Omega^N \mid e^{\alpha \mathcal{S}_{\theta'_k N}(\omega)} \geq e^{\alpha N \epsilon / 2} e^{\alpha N E_{\theta}(X_{\theta'_k})} \right\} \\ &\leq e^{-\alpha N \epsilon / 2} e^{-\alpha N E_{\theta}(X_{\theta'_k})} E_{\theta N} \left(e^{\alpha \mathcal{S}_{\theta'_k N}(\omega)} \right) \\ &\leq e^{-\alpha N \epsilon / 2} e^{-\alpha N E_{\theta}(X_{\theta'_k})} e^{N C_{\theta}^{(k)}(\alpha)}, \end{aligned} \quad (7.11)$$

where

$$C_{\theta}^{(k)}(\alpha) = \log E_{\theta} \left(e^{\alpha X_{\theta'_k}} \right).$$

We write

$$C_\theta^{(k)}(\alpha) - \alpha E_\theta(X_{\theta'_k}) = \int_0^\alpha \left[\left(C_\theta^{(k)} \right)'(u) - E_\theta(X_{\theta'_k}) \right] du,$$

and estimate

$$|C_\theta^{(k)}(\alpha) - \alpha E_\theta(X_{\theta'_k})| \leq \alpha \sup_{u \in [0, \alpha]} \left| \left(C_\theta^{(k)} \right)'(u) - E_\theta(X_{\theta'_k}) \right|.$$

Since $\left(C_\theta^{(k)} \right)'(0) = E_\theta(X_{\theta'_k})$, the uniform continuity gives

$$\lim_{\alpha \rightarrow 0} \sup_{\theta \in [a, b]} \sup_{u \in [0, \alpha]} \left| \left(C_\theta^{(k)} \right)'(u) - E_\theta(X_{\theta'_k}) \right| = 0.$$

It follows that there exists $\alpha_\epsilon^+ > 0$ such that for all $k = 1, \dots, n$,

$$\sup_{\theta \in [a, b]} \left| C_\theta^{(k)}(\alpha_\epsilon^+) - \alpha_\epsilon^+ E_\theta(X_{\theta'_k}) \right| \leq \frac{\epsilon}{4},$$

and (7.11) gives that for all k ,

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \frac{S_{\theta'_k N}(\omega)}{N} - E_\theta(X_{\theta'_k}) \geq \frac{\epsilon}{2} \right\} \leq e^{-\alpha_\epsilon^+ N \epsilon / 4}.$$

Going back to first inequality in (7.10), we conclude that

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left(\frac{S_{\theta' N}(\omega)}{N} - E_\theta(X_{\theta'}) \right) \geq \epsilon \right\} \leq n e^{-\alpha_\epsilon^+ N \epsilon / 4}. \quad (7.12)$$

By repeating the above argument (or by simply applying the final estimate (7.12) to the random variables $-X_\theta$), one derives

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \inf_{\theta' \in [a, b]} \left(\frac{S_{\theta' N}(\omega)}{N} - E_\theta(X_{\theta'}) \right) \leq -\epsilon \right\} \leq n e^{-\alpha_\epsilon^- N \epsilon / 4} \quad (7.13)$$

for a suitable $\alpha_\epsilon^- > 0$. Finally, since

$$\begin{aligned} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{S_{\theta' N}(\omega)}{N} - E_\theta(X_{\theta'}) \right| \geq \epsilon \right\} &\subset \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left(\frac{S_{\theta' N}(\omega)}{N} - E_\theta(X_{\theta'}) \right) \geq \epsilon \right\} \\ &\cup \left\{ \omega \in \Omega^N \mid \inf_{\theta' \in [a, b]} \left(\frac{S_{\theta' N}(\omega)}{N} - E_\theta(X_{\theta'}) \right) \leq -\epsilon \right\}, \end{aligned}$$

(7.8) follows from (7.12) and (7.13). \square

Exercise 7.1. Prove the relation (7.9).

7.4 The maximum likelihood estimator

For each N and $\omega = (\omega_1, \dots, \omega_N) \in \Omega^N$, consider the function

$$[a, b] \ni \theta \mapsto P_{\theta N}(\omega_1, \dots, \omega_N) \in]0, 1[. \quad (7.14)$$

By continuity, this function achieves its global maximum on the interval $[a, b]$. We denote by $\hat{\theta}_{ML, N}(\omega)$ a point where this maximum is achieved (in the case where there are several such points, we select one

arbitrarily but always choosing $\hat{\theta}_{ML,N}(\omega) \in]a, b[$ whenever such possibility exists). This defines a random variable

$$\hat{\theta}_{ML,N} : \Omega^N \rightarrow [a, b]$$

that is called the *maximum likelihood estimator* (abbreviated MLE) of order N . We shall also refer to the sequence $(\hat{\theta}_{ML,N})_{N \geq 1}$ as the MLE.

Note that maximizing (7.14) is equivalent to minimizing the entropy function

$$[a, b] \ni \theta \mapsto S_{\theta N}(\omega) = \sum_{k=1}^N -\log P_{\theta}(\omega_k).$$

Much of our analysis of the MLE will make use of this elementary observation and will be centred around the entropy function $S_{\theta N}$. We set

$$S(\theta, \theta') = E_{\theta}(S_{\theta'}) = - \sum_{\omega \in \Omega} P_{\theta}(\omega) \log P_{\theta'}(\omega).$$

Obviously, $S(\theta, \theta) = S(P_{\theta})$ and

$$S(\theta, \theta') - S(\theta, \theta) = S(P_{\theta} | P_{\theta'}). \quad (7.15)$$

The last relation and the identifiability (7.6), which we assume throughout, give that

$$S(\theta, \theta') > S(\theta, \theta) \quad \text{for} \quad \theta \neq \theta'. \quad (7.16)$$

Applying Proposition 7.6 to $X_{\theta} = -\log P_{\theta}$, we derive

Proposition 7.7 *For any $\epsilon > 0$,*

$$\lim_{N \rightarrow \infty} \sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{S_{\theta' N}(\omega)}{N} - S(\theta, \theta') \right| \geq \epsilon \right\} = 0.$$

Moreover, for any $\epsilon > 0$ there is $C_{\epsilon} > 0$ and $\gamma_{\epsilon} > 0$ such that for all $N \geq 1$,

$$\sup_{\theta \in [a, b]} P_{\theta' N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{S_{\theta' N}(\omega)}{N} - S(\theta, \theta') \right| \geq \epsilon \right\} \leq C_{\epsilon} e^{-\gamma_{\epsilon} N}.$$

The first result of this section is:

Theorem 7.8 *For any $\epsilon > 0$,*

$$\lim_{N \rightarrow \infty} \sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid |\hat{\theta}_{ML,N}(\omega) - \theta| \geq \epsilon \right\} = 0.$$

Moreover, for any $\epsilon > 0$ there exists $C_{\epsilon} > 0$ and $\gamma_{\epsilon} > 0$ such that for all $N \geq 1$,

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid |\hat{\theta}_{ML,N}(\omega) - \theta| \geq \epsilon \right\} \leq C_{\epsilon} e^{-\gamma_{\epsilon} N}.$$

Proof. Let

$$I_{\epsilon} = \{(u, v) \in [a, b] \times [a, b] \mid |u - v| \geq \epsilon\}.$$

It follows from (7.16) and continuity that

$$\delta = \sup_{(u, v) \in I_{\epsilon}} [S(u, v) - S(u, u)] > 0. \quad (7.17)$$

Fix $\theta \in [a, b]$ and set $I_\epsilon(\theta) = \{\theta' \in [a, b] \mid |\theta - \theta'| \geq \epsilon\}$. Let

$$A = \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in I_\epsilon(\theta)} \left| \frac{S_{\theta'N}(\omega)}{N} - S(\theta, \theta') \right| < \frac{\delta}{2} \right\},$$

$$B = \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b] \setminus I_\epsilon(\theta)} \left| \frac{S_{\theta'N}(\omega)}{N} - S(\theta, \theta') \right| < \frac{\delta}{2} \right\}.$$

For $\omega \in A$ and $\theta' \in I_\epsilon(\theta)$,

$$\frac{S_{\theta'N}(\omega)}{N} < S(\theta, \theta') + \frac{\delta}{2} \leq S(\theta, \theta) - \frac{\delta}{2}. \quad (7.18)$$

On the other hand, for $\omega \in B$ and $\theta \in [a, b] \setminus I_\epsilon(\theta)$,

$$\frac{S_{\theta'N}(\omega)}{N} > S(\theta, \theta') - \frac{\delta}{2} \geq S(\theta, \theta) - \frac{\delta}{2}. \quad (7.19)$$

Since $\hat{\theta}_{ML,N}(\omega)$ minimizes the map $[a, b] \ni \theta' \mapsto S_{\theta'N}(\omega)$,

$$\omega \in A \cap B \quad \Rightarrow \quad |\hat{\theta}_{ML,N}(\omega) - \theta| < \epsilon.$$

It follows that

$$\left\{ \omega \in \Omega^N \mid |\hat{\theta}_{ML,N}(\omega) - \theta| \geq \epsilon \right\} \subset A^c \cup B^c = \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{S_{\theta'N}(\omega)}{N} - S(\theta, \theta') \right| \geq \frac{\delta}{2} \right\},$$

and so

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid |\hat{\theta}_{ML,N}(\omega) - \theta| \geq \epsilon \right\} \leq \sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \sup_{\theta' \in [a, b]} \left| \frac{S_{\theta'N}(\omega)}{N} - S(\theta, \theta') \right| \geq \frac{\delta}{2} \right\}.$$

Since δ depends only on the choice of ϵ (recall (7.17)), the last inequality and Proposition 7.7 yield the statement. \square

Theorem 7.8 gives that the MLE is consistent in a very strong sense, and in particular that is uniformly consistent.

Corollary 7.9

$$\lim_{N \rightarrow \infty} \sup_{\theta \in [a, b]} E_{\theta N}(|\hat{\theta}_{ML,N} - \theta|) = 0.$$

Proof. Let $\epsilon > 0$. Then

$$\begin{aligned} E_{\theta N}(|\hat{\theta}_{ML,N} - \theta|) &= \int_{\Omega^N} |\hat{\theta}_{ML,N} - \theta| dP_{\theta N} \\ &= \int_{|\hat{\theta}_{ML,N} - \theta| < \epsilon} |\hat{\theta}_{ML,N} - \theta| dP_{\theta N} + \int_{|\hat{\theta}_{ML,N} - \theta| \geq \epsilon} |\hat{\theta}_{ML,N} - \theta| dP_{\theta N} \\ &\leq \epsilon + (b - a) P_{\theta N} \left\{ \omega \in \Omega^N \mid |\hat{\theta}_{ML,N}(\omega) - \theta| \geq \epsilon \right\}. \end{aligned}$$

Hence,

$$\sup_{\theta \in [a, b]} E_{\theta N}(|\hat{\theta}_{ML,N} - \theta|) \leq \epsilon + (b - a) \sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid |\hat{\theta}_{ML,N}(\omega) - \theta| \geq \epsilon \right\},$$

and the result follows from Proposition 7.8. \square

We note that so far all results of this section hold under the sole assumptions that the maps $[a, b] \ni \theta \mapsto P_\theta(\omega)$ are continuous for all $\omega \in \Omega$ and that the identifiability condition (7.6) is satisfied.

We now turn to study of the efficiency of the MLN and prove the second main result of this section. We strengthen our standing assumptions and assume that the maps $[a, b] \ni \theta \mapsto P_\theta(\omega)$ are C^3 for all $\omega \in \Omega$.

Theorem 7.10 *Suppose that $[a', b'] \subset]a, b[$. Then*

$$\lim_{N \rightarrow \infty} \sup_{\theta \in [a', b']} \left| NE_{\theta N}(|\hat{\theta}_{ML,N} - \theta|^2) - \frac{1}{\mathcal{I}(\theta)} \right| = 0.$$

Proof. Recall that

$$[a, b] \ni \theta \mapsto S_{\theta N}(\omega = (\omega_1, \dots, \omega_N)) = - \sum_{k=1}^N \log P_\theta(\omega_k)$$

achieves its minimum at $\hat{\theta}_{ML,N}(\omega)$ and that $\hat{\theta}_{ML,N}(\omega) \in]a, b[$ unless a strict minimum is achieved at either a or b . Let

$$B_N(a) = \left\{ \omega \in \Omega^N \mid \hat{\theta}_{ML,N}(\omega) = a \right\}, \quad B_N(b) = \left\{ \omega \in \Omega^N \mid \hat{\theta}_{ML,N}(\omega) = b \right\},$$

and

$$\zeta = \min \left(\inf_{\theta \in [a', b']} S(P_\theta | P_a), \inf_{\theta \in [a', b']} S(P_\theta | P_b) \right).$$

Since the maps $\theta \mapsto S(P_\theta | P_a)$, $\theta \mapsto S(P_\theta | P_b)$ are continuous, the identifiability (7.6) yields that $\zeta > 0$. Then, for $\theta \in [a', b']$,

$$\begin{aligned} P_{\theta N}(B_N(a)) &\leq P_{\theta N} \left\{ \omega \in \Omega^N \mid \frac{1}{N} \sum_{k=1}^N \log \frac{P_\theta(\omega_k)}{P_a(\omega_k)} < 0 \right\} \\ &\leq P_{\theta N} \left\{ \omega \in \Omega^N \mid \frac{1}{N} \sum_{k=1}^N \log \frac{P_\theta(\omega_k)}{P_a(\omega_k)} - S(P_\theta | P_a) \leq -\zeta \right\}, \end{aligned}$$

and similarly,

$$P_{\theta N}(B_N(b)) \leq P_{\theta N} \left\{ \omega \in \Omega^N \mid \frac{1}{N} \sum_{k=1}^N \log \frac{P_\theta(\omega_k)}{P_b(\omega_k)} - S(P_\theta | P_b) \leq -\zeta \right\}.$$

Proposition 7.7 now yields that for some constants $K_\zeta > 0$ and $k_\zeta > 0$,

$$\sup_{\theta \in [a', b']} P_{\theta N}(B_N(a) \cup B_N(b)) \leq K_\zeta e^{-k_\zeta N}$$

for all $N \geq 1$. A simple but important observation is that if $\omega \notin B_N(a) \cup B_N(b)$, then $\hat{\theta}_{ML,N}(\omega) \in]a, b[$ and so

$$\dot{S}_{\hat{\theta}_{ML,N}(\omega)N}(\omega) = 0. \quad (7.20)$$

The Taylor expansion gives that for any $\omega \in \Omega^N$ and $\theta \in [a, b]$ there is $\theta'(\omega)$ between $\hat{\theta}_{ML,N}(\omega)$ and θ such that

$$\dot{S}_{\hat{\theta}_{ML,N}(\omega)N}(\omega) - \dot{S}_{\theta N}(\omega) = (\hat{\theta}_{ML,N}(\omega) - \theta) \left[\ddot{S}_{\theta N} + \frac{1}{2} (\hat{\theta}_{ML,N}(\omega) - \theta) \ddot{S}_{\theta'(\omega)N} \right]. \quad (7.21)$$

Write

$$E_{\theta N} \left(\left(\dot{S}_{\hat{\theta}_{ML,N}(\omega)N}(\omega) - \dot{S}_{\theta N}(\omega) \right)^2 \right) = L_N(\theta) + E_{\theta N} \left(\left[\dot{S}_{\theta N} \right]^2 \right),$$

where

$$L_N(\theta) = E_{\theta N} \left(\left[\dot{S}_{\hat{\theta}_{ML,N}(\omega)N} \right]^2 \right) + 2E_{\theta N} \left(\dot{S}_{\hat{\theta}_{ML,N}(\omega)N} \dot{S}_{\theta N} \right). \quad (7.22)$$

It follows from (7.20) that in (7.22) $E_{\theta N}$ reduces to integration over $B_N(a) \cup B_N(b)$, and we arrive at the estimate

$$\sup_{\theta \in [a', b']} |L_N(\theta)| \leq KN^2 \sup_{\theta \in [a', b']} P_{\theta N}(B_N(a) \cup B_N(b)) \leq KN^2 K_{\zeta} e^{-k_{\zeta} N} \quad (7.23)$$

for some uniform constant $K > 0$, where by uniform we mean that K does not depend on N . It is easy to see that one can take

$$K = 3 \sup_{\theta \in [a, b], \omega \in \Omega} \left(\frac{\dot{P}_{\theta}(\omega)}{P_{\theta}(\omega)} \right)^2.$$

In Exercise 7.2 the reader is asked to estimate other uniform constant that will appear in the proof.

Squaring both sides in (7.21), taking the expectation, and dividing both sides with N^2 , we derive the identity

$$\frac{1}{N^2} L_N(\theta) + \frac{1}{N^2} E_{\theta N} \left(\left[\dot{S}_{\theta N} \right]^2 \right) = E_{\theta N} \left((\hat{\theta}_{ML,N} - \theta)^2 \left[\frac{\ddot{S}_{\theta N}}{N} + \frac{1}{2N} (\hat{\theta}_{ML,N} - \theta) \ddot{S}_{\theta' N} \right]^2 \right). \quad (7.24)$$

An easy computation gives

$$\frac{1}{N^2} E_{\theta N} \left(\left[\dot{S}_{\theta N} \right]^2 \right) = \frac{1}{N} \mathcal{I}(\theta).$$

Regarding the right hand side in (7.24), we write it as

$$E_{\theta N} \left((\hat{\theta}_{ML,N} - \theta)^2 \left[\frac{\ddot{S}_{\theta N}}{N} \right]^2 \right) + R_N(\theta),$$

where the remainder $R_N(\theta)$ can be estimated as

$$|R_N(\theta)| \leq C_1 E_{\theta N} \left(|\hat{\theta}_{ML,N} - \theta|^3 \right) \quad (7.25)$$

for some uniform constant $C_1 > 0$.

With these simplifications, an algebraic manipulation of the identity (7.24) gives

$$NE_{\theta N} \left((\hat{\theta}_{ML,N} - \theta)^2 \right) - \frac{1}{\mathcal{I}(\theta)} = -D_N(\theta) - \frac{NR_N(\theta)}{\mathcal{I}(\theta)^2} + \frac{1}{N} \frac{L_N(\theta)}{\mathcal{I}(\theta)^2}, \quad (7.26)$$

where

$$D_N(\theta) = NE_{\theta N} \left((\hat{\theta}_{ML} - \theta)^2 \left(\left[\frac{\ddot{S}_{\theta N}}{N} \right]^2 \frac{1}{\mathcal{I}(\theta)^2} - 1 \right) \right). \quad (7.27)$$

Writing

$$\left[\frac{\ddot{S}_{\theta N}}{N} \right]^2 \frac{1}{\mathcal{I}(\theta)^2} - 1 = \frac{1}{\mathcal{I}(\theta)^2} \left(\frac{\ddot{S}_{\theta N}}{N} + \mathcal{I}(\theta) \right) \left(\frac{\ddot{S}_{\theta N}}{N} - \mathcal{I}(\theta) \right)$$

and using that $\mathcal{I}(\theta)$ is continuous and strictly positive on $[a, b]$, we derive the estimate

$$|D_N(\theta)| \leq C_2 NE_{\theta N} \left((\hat{\theta}_{ML} - \theta)^2 \left| \frac{\ddot{S}_{\theta N}}{N} - \mathcal{I}(\theta) \right| \right) \quad (7.28)$$

for some uniform constant $C_2 > 0$.

Fix $\epsilon > 0$, and choose $C_\epsilon > 0$ and $\gamma_\epsilon > 0$ such that

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid |\hat{\theta}_{ML, N}(\omega) - \theta| \geq \epsilon \right\} \leq C_\epsilon e^{-\gamma_\epsilon N}, \quad (7.29)$$

$$\sup_{\theta \in [a, b]} P_{\theta N} \left\{ \omega \in \Omega^N \mid \left| \frac{\ddot{S}_{\theta N}}{N} - \mathcal{I}(\theta) \right| \geq \epsilon \right\} \leq C_\epsilon e^{-\gamma_\epsilon N}. \quad (7.30)$$

Here, (7.29) follows from Theorem 7.8, while (7.30) follows from Proposition 7.7 applied to $X_\theta = -\frac{d^2}{d\theta^2} \log P_\theta$ (recall that $E_\theta(X_\theta) = \mathcal{I}(\theta)$).

Let $\delta = \inf_{u \in [a, b]} \mathcal{I}(u)$. Then, for all $\theta \in [a, b]$,

$$\begin{aligned} \frac{N|\mathcal{R}_N(\theta)|}{\mathcal{I}(\theta)^2} &\leq \frac{C_1 N}{\delta^2} \int_{\Omega^N} |\hat{\theta}_{ML, N} - \theta|^3 dP_{\theta N} \\ &= \frac{C_1 N}{\delta^2} \int_{|\hat{\theta}_{ML, N} - \theta| < \epsilon} |\hat{\theta}_{ML, N} - \theta|^3 dP_{\theta N} + \frac{C_1 N}{\delta^2} \int_{|\hat{\theta}_{ML, N} - \theta| \geq \epsilon} |\hat{\theta}_{ML, N} - \theta|^3 dP_{\theta N} \\ &\leq \epsilon \frac{C_1}{\delta^2} N E_{\theta N} \left((\hat{\theta}_{ML, N} - \theta)^2 \right) + \frac{C_1 (b-a)^3 N}{\delta^2} C_\epsilon e^{-\gamma_\epsilon N}. \end{aligned} \quad (7.31)$$

Similarly, splitting the integral (that is, $E_{\theta N}$) on the r.h.s. of (7.28) into the sum of integrals over the sets

$$\left| \frac{\ddot{S}_{\theta N}}{N} - \mathcal{I}(\theta) \right| < \epsilon, \quad \left| \frac{\ddot{S}_{\theta N}}{N} - \mathcal{I}(\theta) \right| \geq \epsilon,$$

we derive that for all $\theta \in [a, b]$,

$$|D_N(\theta)| \leq \epsilon C_2 N E_{\theta N} \left((\hat{\theta}_{ML, N} - \theta)^2 \right) + C_2 C_3 N C_\epsilon e^{-\gamma_\epsilon N}, \quad (7.32)$$

where $C_3 > 0$ is a uniform constant. Returning to (7.26) and taking $\epsilon = \epsilon_0$ such that

$$\epsilon_0 \frac{C_1}{\delta^2} < \frac{1}{4}, \quad \epsilon_0 C_2 < \frac{1}{4},$$

the estimates (7.23), (7.31), and (7.32) give that for all $\theta \in [a', b']$,

$$N E_{\theta N} \left((\hat{\theta}_{ML, N} - \theta)^2 \right) \leq \frac{2}{\mathcal{I}(\theta)} + C'_{\epsilon_0} N e^{-\gamma_{\epsilon_0} N} + \frac{2K}{\delta^2} K_\zeta e^{-k_\zeta N},$$

where $C'_{\epsilon_0} > 0$ is a uniform constant (that of course depends on ϵ_0). It follows that

$$C' = \sup_{N \geq 1} \sup_{\theta \in [a', b']} N E_{\theta N} \left((\hat{\theta}_{ML, N} - \theta)^2 \right) < \infty. \quad (7.33)$$

Returning to (7.31), (7.32), we then have that for any $\epsilon > 0$,

$$\sup_{\theta \in [a', b']} \frac{N|\mathcal{R}_N(\theta)|}{\mathcal{I}(\theta)^2} \leq \epsilon \frac{C_1}{\delta^2} C' + \frac{C_1 (b-a)^3 N}{\delta^2} C_\epsilon e^{-\gamma_\epsilon N}, \quad (7.34)$$

$$\sup_{\theta \in [a', b']} |D_N(\theta)| \leq \epsilon C_2 C' + C_2 C_3 N C_\epsilon e^{-\gamma_\epsilon N}. \quad (7.35)$$

Finally, returning once again to (7.26), we derive that for any $\epsilon > 0$,

$$\begin{aligned} \sup_{\theta \in [a', b']} \left| N E_{\theta N} \left((\hat{\theta}_{ML, N} - \theta)^2 \right) - \frac{1}{\mathcal{I}(\theta)} \right| &\leq \sup_{\theta \in [a', b']} |D_N(\theta)| + \sup_{\theta \in [a', b']} \frac{N|\mathcal{R}_N(\theta)|}{\mathcal{I}(\theta)^2} + \sup_{\theta \in [a', b']} \frac{|L_N(\theta)|}{N\mathcal{I}(\theta)^2} \\ &\leq \epsilon C'' + C''_\epsilon N e^{-\gamma_\epsilon N} + K K_\zeta e^{-k_\zeta N}, \end{aligned}$$

where $C'' > 0$ is a uniform constant and $C_\epsilon'' > 0$ depends only on ϵ . Hence,

$$\limsup_{N \rightarrow \infty} \sup_{\theta \in [a', b']} \left| NE_{\theta N} \left((\hat{\theta}_{ML,N} - \theta)^2 \right) - \frac{1}{\mathcal{I}(\theta)} \right| \leq \epsilon C''.$$

Since $\epsilon > 0$ is arbitrary, the result follows. \square

Exercise 7.2. Write an explicit estimate for all uniform constants that have appeared in the above proof.

Remark 7.2 The proof of Theorem 7.10 hints at the special role the boundary points a and b of the chosen parameter interval may play in study of the efficiency. The MLE is selected with respect to the $[a, b]$ and $\hat{\theta}_{ML,N}(\omega)$ may take value a or b without the derivative $\dot{S}_{\hat{\theta}_{ML,N}(\omega)N}(\omega)$ vanishing. That forces the estimation of the probability of the set $B_N(a) \cup B_N(b)$ and the argument requires that θ stays away from the boundary points. If the parameter interval is replaced by a circle, there would be no boundary points and the above proof then gives that the uniform efficiency of the MLE holds with respect to the entire parameter set. One may wonder whether a different type of argument may yield the same result in the case of $[a, b]$. The following example shows that this is not the case.

Let $\Omega = \{0, 1\}$ and let $P_\theta(0) = 1 - \theta$, $P_\theta(1) = \theta$, where $\theta \in]0, 1[$. One computes $\mathcal{I}(\theta) = (\theta - \theta^2)^{-1}$. If $[a, b] \subset]0, 1[$ is selected as the estimation interval, the MLE $\hat{\theta}_{ML,N}$ takes the following form:

$$\begin{aligned} \hat{\theta}_{ML,N}(\omega_1, \dots, \omega_N) &= \frac{\omega_1 + \dots + \omega_N}{N} & \text{if } \frac{\omega_1 + \dots + \omega_N}{N} \in [a, b], \\ \hat{\theta}_{ML,N}(\omega_1, \dots, \omega_N) &= a & \text{if } \frac{\omega_1 + \dots + \omega_N}{N} < a, \\ \hat{\theta}_{ML,N}(\omega_1, \dots, \omega_N) &= b & \text{if } \frac{\omega_1 + \dots + \omega_N}{N} > b. \end{aligned}$$

We shall indicate the dependence of $\hat{\theta}_{ML,N}$ on $[a, b]$ by $\hat{\theta}_{ML,N}^{[a,b]}$. It follows from Theorem 7.10 that

$$\lim_{N \rightarrow \infty} NE_{(\theta=1/2)N} \left(\left(\hat{\theta}_{ML,N}^{[\frac{1}{3}, \frac{2}{3}]} - \frac{1}{2} \right)^2 \right) = \left[\mathcal{I} \left(\frac{1}{2} \right) \right]^{-1} = \frac{1}{4}.$$

On the other hand, a moment's reflection shows that

$$\frac{1}{2} E_{(\theta=1/2)N} \left(\left(\hat{\theta}_{ML,N}^{[\frac{1}{3}, \frac{2}{3}]} - \frac{1}{2} \right)^2 \right) = E_{(\theta=1/2)N} \left(\left(\hat{\theta}_{ML,N}^{[\frac{1}{2}, \frac{2}{3}]} - \frac{1}{2} \right)^2 \right),$$

and so

$$\lim_{N \rightarrow \infty} NE_{(\theta=1/2)N} \left(\left(\hat{\theta}_{ML,N}^{[\frac{1}{2}, \frac{2}{3}]} - \frac{1}{2} \right)^2 \right) = \frac{1}{8}.$$

Thus, in this case even the bound of Proposition 7.5 fails at the boundary point $1/2$ at which the MLE becomes "superefficient". In general, such artificial boundary effects are difficult to quantify and we feel it is best that they are excluded from the theory. These observations hopefully elucidate our definition of efficiency which excludes the boundary points of the interval of parameters.

7.5 Notes and references

For additional information and references about parameter estimation the reader may consult [LeCa, Vaa]. For additional information about the Cramér-Rao bound and its history we refer the reader to the respective Wikipedia and Scholarpedia articles.

The modern theory of the MLE started with the seminal work of Fisher [Fis1]; for the fascinating history of the subject see [Sti]. Our analysis of the MLE follows the standard route, but I have followed no particular reference. In particular, I am not aware whether Theorem 7.10 as formulated have appeared previously in the literature.

Bibliography

- [AczDa] J. Aczél and Z. Daróczy.: *On Measures of Information and Their Characterizations*. Academic Press, 1975.
- [Ana] V. Anantharam.: A large deviations approach to error exponents in source coding and hypothesis testing. *IEEE Trans. Inf. Theory* **36**, 938-943 (1990).
- [AcFoNg] J. Aczél, B. Forte, and C. T. Ng.: Why the Shannon and Hartley entropies are 'natural'. *Adv. Appl. Prob.* **6**, 131-146 (1974)
- [Bill] P. Billingsley.: *Ergodic Theory and Information* John Wiley & Sons, 1965.
- [Cam] L.L. Campbel.: An extended Cencov characterization of the information metric. *Proc. AMS* **98**, 135-141 (1996).
- [Cen] N.N. Cencov.: *Statistical decision rules and optimal inference*. Trans. Math. Monographs **53**, AMS, 1981
- [Che] H. Chernoff.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **23**, 493 (1952)
- [Chu] K.L. Chung.: *A Course in Probability Theory*. Academic Press, 2001.
- [Csi] I. Csiszár.: Axiomatic characterizations of information measures. *Entropy* **10**, 261-273 (2008).
- [CsiLo] I. Csiszár and G. Longo.: On the error exponent for source coding and for testing simple statistical hypotheses. *Studia Sci. Math. Hungarica* **6**, 181 (1971).
- [CsiKö] I. Csiszár and J. Körner.: *Information Theory*. Academic Press, 1981.
- [CovTh] T.A. Cover and J.A. Thomas.: *Elements of Information Theory*. John Wiley & Sons, 1991.
- [Dar] Z. Daróczy.: Über mittelwerte und entropien vollständiger wahrscheinlichkeitsverteilungen. *Acta Math. Acad. Sci. Hungar.* **15**, 203-210 (1964).
- [dHoll] F. den Hollander.: *Large Deviations*. AMS, 2000.
- [DeZe] A. Dembo and O. Zeitouni.: *Large deviations techniques and applications*. Springer (1998).
- [EndSc] D.M. Endres and J. E. Schindelin.: A new metric for probability distributions. *IEEE Trans. Inf. Theory* **49**, 1858-1860 (2003).
- [Ell] R.S. Ellis.: *Entropy, Large Deviations, and Statistical Mechanics*. Springer, 1985. Reprinted in the series Classics of Mathematics, 2006.
- [Fadd] D. K. Faddeev.: On the concept of entropy of a finite probabilistic scheme. *Uspehi Mat. Nauk* **1**, 227-231 (1956).
- [Fis1] R. A. Fisher.: On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222**, 309-368 (1921).

- [Fis2] R. A. Fisher.: Theory of statistical estimation. Proc. Cambridge Philos. Soc. **22**, 700-725 (1925).
- [FugTo] B. Fugledge and F. Topsoe.: Jensen-Shannon divergence and Hilbert space embedding. : Proceedings of International Symposium on Information Theory, ISIT 2004.
- [GHLS] S. Goldstein, D.A. Huse, J.L. Lebowitz, P. Sartori.: Statistical mechanics and thermodynamics of large and small systems. Preprint, <https://arxiv.org/pdf/1712.08961.pdf>.
- [Gra] R. M. Gray.: *Entropy and Information Theory*. Springer, 2011.
- [Har] R.V.L. Hartley.: Transmission of information. Bell System Technical Journal **7**, 535-563 (1928).
- [Hob] A. Hobson.: A new theorem of information theory. J. Stat. Phys. **1**, 383-391 (1969).
- [Hoe] H. Hoeffding.: Asymptotically optimal tests for multinomial distributions. Ann. Math. Statist. **36**, 369 (1965).
- [JPR] V. Jakšić, C-A. Pillet and L. Rey-Bellet.: Entropic fluctuations in statistical mechanics: I. Classical dynamical systems. Nonlinearity **24**, 699 (2011).
- [Jay1] E.T. Jaynes.: Information theory and statistical mechanics. Phys. Rev. **106**, 620-630 (1957).
- [Jay2] E.T. Jaynes.: Information theory and statistical mechanics II. Phys. Rev. **108**, 171 (1957).
- [John] D. Johnson.: *Statistical Signal Processing*. <https://cpb-us-e1.wpmucdn.com/blogs.rice.edu/dist>
- [Jeff] H. Jeffreys.: An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. A **186**, 453-461 (1946).
- [Kát] I. Kátai.: A remark on additive arithmetical functions. Ann. Univ. Sci. Budapest, Etdvds Sect. Math. **12**, 81-83 (1967).
- [Khi] A. Ya. Khinchin.: *Mathematical Foundations of Information Theory*. Dover Publications, 1957.
- [KullLe] S. Kullback and R.A. Leibler.: On information and sufficiency. Ann. Math. Statist. **22**, 79-86 (1951).
- [LeCa] E.L. Lehmann and G. Cassela.: *Theory of Point Estimation*. Springer, 1998.
- [LeRo] E.L. Lehmann and J.P. Romano: *Testing Statistical Hypotheses*. Springer, 2005.
- [Lei] T. Leinster.: A short characterization of relative entropy. Preprint, <https://arxiv.org/pdf/1712.04903.pdf>.
- [Mer] N. Merhav.: *Statistical Physics and Information Theory*. Foundations and Trends in Communications and Information Theory **6**, (2009). Now Publishers, Hanover, MA.
- [Lin] J. Lin.: Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory **27**, 145-151 (1991).
- [ÖstVa] F. Österreicher and I. Vajda.: A new class of metric divergences on probability spaces and its statistical applications. Ann. Inst. Statist. Math. **55**, 639-653 (2003).
- [Sha] C.E. Shannon.: A mathematical theory of communication. The Bell System Technical Journal **27**, 379-423, 623-656 (1948).
- [ShaWe] C.E. Shannon and W. Weaver.: *The Mathematical Theory of Communication*. The University of Illinois Press, 1964.
- [Shi] P.C. Shields.: *The Ergodic Theory of Discrete Sample Paths*. AMS (1991).
- [Sti] S.M. Stiegler.: The epic story of maximum likelihood. Statistical Science **22**, 598-620 (2007).

- [Sow] R. Sowers.: Stein's lemma—a large deviation approach. Naval research laboratory report 9185 (1989).
- [Mat] K. Matsumoto.: Reverse test and characterization of quantum relative entropy. Preprint, <https://arxiv.org/pdf/1010.1030.pdf>.
- [Rén] A. Rényi.: On measures of information and entropy. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, Univ. California Press, Berkeley (1961).
- [RohSa] V.K. Rohtagi and A.K. Md. E. Saleh.: *An Introduction to Probability and Statistics*. John Wiley & Sons, 2015.
- [Ross] S. Ross.: *First Course in Probability*. Pearson, 2014.
- [Szi] L. Szilard.: On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Zeitschrift fur Physik* **53**, 840-856 (1929). English translation in *The Collected Works of Leo Szilard: Scientific Papers*, B.T. Feld and G. Weiss Szilard (eds.), Cambridge, Massachusetts: MIT Press, 1972, pp. 103-129.
- [Thi] W. Thirring.: *Quantum Mathematical Physics. Atoms, Molecules, and Large Systems*. Springer, 2002.
- [Vaa] A.W. van der Vaart.: *Asymptotic Statistics*. Cambridge University Press, 1998.
- [Ver] S. Verdú.: Fifty years of Shannon theory. *IEEE Trans. Inf. Theory* **44**, 2057-2078 (1998).
- [WiGaEi] H. Wilming, R. Gallego, J. Eisert.: Axiomatic characterization of the quantum relative entropy and free energy. *Entropy* **19**, 241-244 (2017).