Foundations and Trends™ in Communications and Information Theory

Volume 1 Issue 4, 2004

Editorial Scope

Foundations and Trends™ in Communications and Information Theory will publish survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and image compression
- Wireless communications

Information for Librarians

Foundations and Trends™ in Communications and Information Theory, 2004, Volume 1, 4 issues. ISSN paper version 1567-2190 (USD 200 N. America; EUR 200 Outside N. America). ISSN online version 1567-2328 (USD 250 N. America; EUR 250 Outside N. America). Also available as a combined paper and online subscription (USD 300 N. America; EUR 300 Outside N. America).

Information Theory and Statistics: A Tutorial

Imre Csiszár

Rényi Institute of Mathematics,
Hungarian Academy of Sciences
POB 127, H-1364 Budapest,
Hungary

Paul C. Shields

Professor Emeritus of Mathematics,
University of Toledo,
Ohio,
USA

now

the essence of knowledge

**Foundations and Trends™ in
Communications and Information Theory**

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

Printed on acid-free paper

ISSNs: Paper version 1567-2190; Electronic version 1567-2328

Foundations and Trends™ in
Communications and Information Theory
Vol 1, No 4 (2004) 417-528

now
the essence of knowledge

Information Theory and Statistics: A Tutorial

I. Csiszár[1] and P.C. Shields[2]

[1] *Rényi Institute of Mathematics, Hungarian Academy of Sciences, POB 127, H-1364 Budapest, Hungary, csiszar@renyi.hu*
[2] *University of Toledo, Ohio, USA, paul.shields@utoledo.edu*

Abstract

This tutorial is concerned with applications of information theory concepts in statistics, in the finite alphabet setting. The information measure known as information divergence or Kullback-Leibler distance or relative entropy plays a key role, often with a geometric flavor as an analogue of squared Euclidean distance, as in the concepts of I-projection, I-radius and I-centroid. The topics covered include large deviations, hypothesis testing, maximum likelihood estimation in exponential families, analysis of contingency tables, and iterative algorithms with an "information geometry" background. Also, an introduction is provided to the theory of universal coding, and to statistical inference via the minimum description length principle motivated by that theory.

Table of Contents

Preface

This tutorial is concerned with applications of information theory concepts in statistics. It originated as lectures given by Imre Csiszár at the University of Maryland in 1991 with later additions and corrections by Csiszár and Paul Shields.

Attention is restricted to finite alphabet models. This excludes some celebrated applications such as the information theoretic proof of the dichotomy theorem for Gaussian measures, or of Sanov's theorem in a general setting, but considerably simplifies the mathematics and admits combinatorial techniques. Even within the finite alphabet setting, no efforts were made at completeness. Rather, some typical topics were selected, according to the authors' research interests. In all of them, the information measure known as information divergence (I-divergence) or Kullback–Leibler distance or relative entropy plays a basic role. Several of these topics involve "information geometry", that is, results of a geometric flavor with I-divergence in the role of squared Euclidean distance.

In Section 2, a combinatorial technique of major importance in information theory is applied to large deviation and hypothesis testing problems. The concept of I-projections is addressed in Sections 3 and

4, with applications to maximum likelihood estimation in exponential families and, in particular, to the analysis of contingency tables. Iterative algorithms based on information geometry, to compute I-projections and maximum likelihood estimates, are analyzed in Section 5. The statistical principle of minimum description length (MDL) is motivated by ideas in the theory of universal coding, the theoretical background for efficient data compression. Sections 6 and 7 are devoted to the latter. Here, again, a major role is played by concepts with a geometric flavor that we call I-radius and I-centroid. Finally, the MDL principle is addressed in Section 8, based on the universal coding results.

Reading this tutorial requires no prerequisites beyond basic probability theory. Measure theory is needed only in the last three Sections, dealing with processes. Even there, no deeper tools than the martingale convergence theorem are used. To keep this tutorial self-contained, the information theoretic prerequisites are summarized in Section 1, and the statistical concepts are explained where they are first used. Still, while prior exposure to information theory and/or statistics is not indispensable, it is certainly useful. Very little suffices, however, say Chapters 2 and 5 of the Cover and Thomas book [8] or Sections 1.1, 1.3, 1.4 of the Csiszár-Körner book [15], for information theory, and Chapters 1–4 and Sections 9.1–9.3 of the book by Cox and Hinckley [9], for statistical theory.

1

Preliminaries

The symbol $A = \{a_1, a_2, \dots, a_{|A|}\}$ denotes a finite set of cardinality $|A|$; x_m^n denotes the sequence $x_m, x_{m+1}, \dots, x_n$, where each $x_i \in A$; A^n denotes the set of all x_1^n; A^∞ denotes the set of all infinite sequences $x = x_1^\infty$, with $x_i \in A, i \geq 1$; and A^* denotes the set of all finite sequences drawn from A. The set A^* also includes the empty string Λ. The concatenation of $u \in A^*$ and $v \in A^* \cup A^\infty$ is denoted by uv. A finite sequence u is a prefix of a finite or infinite sequence w, and we write $u \prec w$, if $w = uv$, for some v.

The *entropy* $H(P)$ of a probability distribution $P = \{P(a), a \in A\}$ is defined by the formula

$$H(P) = -\sum_{a \in A} P(a) \log P(a).$$

Here, as elsewhere in this tutorial, *base two logarithms* are used and $0 \log 0$ is defined to be 0. Random variable notation is often used in this context. For a random variable X with values in a finite set, $H(X)$ denotes the entropy of the distribution of X. If Y is another random variable, not necessarily discrete, the *conditional entropy* $H(X|Y)$ is defined as the average, with respect to the distribution of Y, of the entropy of the conditional distribution of X, given $Y = y$. The *mutual* *information* between X and Y is defined by the formula

$$I(X \wedge Y) = H(X) - H(X|Y).$$

If Y (as well as X) takes values in a finite set, the following alternative formulas are also valid.

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

For two distributions P and Q on A, *information divergence* (*I-divergence*) or *relative entropy* is defined by

$$D(P\|Q) = \sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)}.$$

A key property of I-divergence is that it is nonnegative and zero if and only if $P = Q$. This is an instance of the *log-sum inequality*, namely, that for arbitrary nonnegative numbers $p_1, \ldots, p_t$ and $q_1, \ldots, q_t$,

$$\sum_{i=1}^{t} p_i \log \frac{p_i}{q_i} \geq \left(\sum_{i=1}^{t} p_i \right) \log \frac{\sum_{i=1}^{t} p_i}{\sum_{i=1}^{t} q_i}$$

with equality if and only if $p_i = cq_i, 1 \leq i \leq t$. Here $p \log \frac{p}{q}$ is defined to be 0 if $p = 0$ and $+\infty$ if $p > q = 0$.

Convergence of probability distributions, $P_n \to P$, means pointwise convergence, that is, $P_n(a) \to P(a)$ for each $a \in A$. Topological concepts for probability distributions, continuity, open and closed sets, etc., are meant for the topology of pointwise convergence. Note that the entropy $H(P)$ is a continuous function of P, and the I-divergence $D(P\|Q)$ is a lower semi-continuous function of the pair (P, Q), continuous at each (P, Q) with strictly positive Q.

A *code* for symbols in A, with image alphabet B, is a mapping $C: A \mapsto B^*$. Its *length function* $L: A \mapsto N$ is defined by the formula

$$C(a) = b_1^{L(a)}.$$

In this tutorial, it will be assumed, unless stated explicitly otherwise, that the image alphabet is binary, $B = \{0, 1\}$, and that all codewords $C(a)$, $a \in A$, are distinct and different from the empty string Λ.

Often, attention will be restricted to codes satisfying the *prefix condition* that $C(a) \prec C(\tilde{a})$ never holds for $a \neq \tilde{a}$ in A. These codes, called *prefix codes*, have the desirable properties that each sequence in A^* can be uniquely decoded from the concatenation of the codewords of its symbols, and each symbol can be decoded "instantaneously", that is, the receiver of any sequence $w \in B^*$ of which $u = C(x_1) \dots C(x_i)$ is a prefix need not look at the part of w following u in order to identify u as the code of the sequence $x_1 \dots x_i$.

Of fundamental importance is the following fact.

Lemma 1.1. A function $L\colon A \mapsto N$ is the length function of some prefix code if and only if it satisfies the so-called *Kraft inequality*

$$\sum_{a \in A} 2^{-L(a)} \leq 1.$$

Proof. Given a prefix code $C\colon A \mapsto B^*$, associate with each $a \in A$ the number $t(a)$ whose dyadic expansion is the codeword $C(a) = b_1^{L(a)}$, that is, $t(a) = 0.b_1 \dots b_{L(a)}$. The prefix condition implies that $t(\tilde{a}) \notin [t(a), t(a) + 2^{-L(a)})$ if $\tilde{a} \neq a$, thus the intervals $[t(a), t(a) + 2^{-L(a)})$, $a \in A$, are disjoint. As the total length of disjoint subintervals of the unit interval is at most 1, it follows that $\sum 2^{-L(a)} \leq 1$.

Conversely, suppose a function $L\colon A \mapsto N$ satisfies $\sum 2^{-L(a)} \leq 1$. Label A so that $L(a_i) \leq L(a_{i+1})$, $i < |A|$. Then $t(i) = \sum_{j<i} 2^{-L(a_j)}$ can be dyadically represented as $t(i) = 0.b_1 \dots b_{L(a_i)}$, and $C(a_i) = b_1^{L(a_i)}$ defines a prefix code with length function L. □

A key consequence of the lemma is Shannon's *noiseless coding theorem.*

Theorem 1.1. Let P be a probability distribution on A. Then each prefix code has expected length

$$E(L) = \sum_{a \in A} P(a)L(a) \geq H(P).$$

Furthermore, there is a prefix code with length function $L(a) = \lceil -\log P(a) \rceil$; its expected length satisfies

$$E(L) < H(P) + 1.$$

Proof. The first assertion follows by applying the log-sum inequality to $P(a)$ and $2^{-L(a)}$ in the role of p_i and q_i and making use of $\sum P(a) = 1$ and $\sum 2^{-L(a)} \leq 1$. The second assertion follows since $L(a) = \lceil -\log P(a) \rceil$ obviously satisfies the Kraft inequality. □

By the following result, even non-prefix codes cannot "substantially" beat the entropy lower bound of Theorem 1.1. This justifies the practice of restricting theoretical considerations to prefix codes.

Theorem 1.2. The length function of a not necessarily prefix code $C\colon A \mapsto B^*$ satisfies

$$\sum_{a \in A} 2^{-L(a)} \leq \log |A|, \tag{1.1}$$

and for any probability distribution P on A, the code has expected length

$$E(L) = \sum_{a \in A} P(a)L(a) \geq H(P) - \log\log |A|.$$

Proof. It suffices to prove the first assertion, for it implies the second assertion via the log-sum inequality as in the proof of Theorem 1.1. To this end, we may assume that for each $a \in A$ and $i < L(a)$, every $u \in B^i$ is equal to $C(\tilde{a})$ for some $\tilde{a} \in A$, since otherwise $C(a)$ can be replaced by an $u \in B^i$, increasing the left side of (1.1). Thus, writing

$$|A| = \sum_{i=1}^{m} 2^i + r, \qquad m \geq 1,\ 0 \leq r < 2^{m+1},$$

it suffices to prove (1.1) when each $u \in B^i$, $1 \leq i \leq m$, is a codeword, and the remaining r codewords are of length $m+1$. In other words, we have to prove that

$$m + r2^{-(m+1)} \leq \log |A| = \log(2^{m+1} - 2 + r),$$

or

$$r2^{-(m+1)} \leq \log(2 + (r-2)2^{-m}).$$

This trivially holds if $r = 0$ or $r \geq 2$. As for the remaining case $r = 1$, the inequality

$$2^{-(m+1)} \leq \log(2 - 2^{-m})$$

is verified by a trite calculation for $m = 1$, and then it holds even more for $m > 1$. □

The above concepts and results extend to codes for n-length messages or *n-codes*, that is, to mappings $C\colon A^n \mapsto B^*$, $B = \{0, 1\}$. In particular, the length function $L\colon A^n \mapsto N$ of an n-code is defined by the formula $C(x_1^n) = b_1^{L(x_1^n)}$, $x_1^n \in A^n$, and satisfies

$$\sum_{x_1^n \in A^n} 2^{-L(x_1^n)} \leq n \log |A|;$$

and if $C\colon A^n \mapsto B^*$ is a prefix code, its length function satisfies the Kraft inequality

$$\sum_{x_1^n \in A^n} 2^{-L(x_1^n)} \leq 1\ .$$

Expected length $E(L) = \sum_{x_1^n \in A^n} P_n(x_1^n) L(x_1^n)$ for a probability distribution P_n on A^n, of a prefix n-code satisfies

$$E(L) \geq H(P_n)\ ,$$

while

$$E(L) \geq H(P_n) - \log n - \log\log |A|$$

holds for any n-code.

An important fact is that, for any probability distribution P_n on A^n, the function $L(x_1^n) = \lceil -\log P_n(x_1^n) \rceil$ satisfies the Kraft inequality. Hence there exists a prefix n-code whose length function is $L(x_1^n)$ and whose expected length satisfies $E(L) < H(P_n) + 1$. Any such code is called a *Shannon code* for P_n.

Supposing that the limit

$$\overline{H} = \lim_{n \to \infty} \frac{1}{n}\, H(P_n)$$

exists, it follows that for any n-codes $C_n\colon A^n \mapsto B^*$ with length functions $L_n\colon A^n \mapsto N$, the expected length per symbol satisfies

$$\liminf_{n\to\infty} \frac{1}{n} E(L_n) \geq \overline{H} \ ;$$

moreover, the expected length per symbol of a Shannon code for P_n converges to $\overline{H}$ as $n \to \infty$.

We close this introduction with a discussion of arithmetic codes, which are of both practical and conceptual importance. An *arithmetic code* is a sequence of n-codes, $n = 1, 2, \ldots$ defined as follows.

Let $Q_n, \ n = 1, 2, \ldots$ be probability distributions on the sets A^n satisfying the consistency conditions

$$Q_n(x_1^n) = \sum_{a\in A} Q_{n+1}(x_1^n a);$$

these are necessary and sufficient for the distributions Q_n to be the marginal distributions of a process (for process concepts, see Appendix). For each n, partition the unit interval $[0, 1)$ into subintervals $J(x_1^n) = [\ell(x_1^n), r(x_1^n))$ of length $r(x_1^n) - \ell(x_1^n) = Q_n(x_1^n)$ in a nested manner, i. e., such that $\{J(x_1^n a)\colon \ a \in A\}$ is a partitioning of $J(x_1^n)$, for each $x_1^n \in A^n$. Two kinds of arithmetic codes are defined by setting $C(x_1^n) = z_1^m$ if the endpoints of $J(x_1^n)$ have binary expansions

$$\ell(x_1^n) = .z_1 z_2 \cdots z_m 0 \cdots, \quad r(x_1^n) = .z_1 z_2 \cdots z_m 1 \cdots,$$

and $\widetilde{C}(x_1^n) = z_1^{\widetilde{m}}$ if the midpoint of $J(x_1^n)$ has binary expansion

$$\frac{1}{2}\Big(\ell(x_1^n) + r(x_1^n)\Big) = .z_1 z_2 \cdots z_{\widetilde{m}} \cdots, \ \widetilde{m} = \lceil -\log Q_n(x_1^n) \rceil + 1. \quad (1.2)$$

Since clearly $\ell(x_1^n) \leq .z_1 z_2 \cdots z_{\widetilde{m}}$ and $r(x_1^n) \geq .z_1 z_2 \cdots z_{\widetilde{m}} + 2^{-\widetilde{m}}$, we always have that $C(x_1^n)$ is a prefix of $\widetilde{C}(x_1^n)$, and the length functions satisfy $L(x_1^n) < \widetilde{L}(x_1^n) = \lceil -\log Q_n(x_1^n) \rceil + 1$. The mapping $C\colon A^n \mapsto B^*$ is one-to-one (since the intervals $J(x_1^n)$ are disjoint) but not necessarily a prefix code, while $\widetilde{C}(x_1^n)$ is a prefix code, as one can easily see.

In order to determine the codeword $C(x_1^n)$ or $\widetilde{C}(x_1^n)$, the nested partitions above need not be actually computed, it suffices to find the interval $J(x_1^n)$. This can be done in steps, the i-th step is to partition the interval $J(x_1^{i-1})$ into $|A|$ subintervals of length proportional to the conditional probabilities $Q(a|x_1^{i-1}) = Q_i(x_1^{i-1}a)/Q_{i-1}(x_1^{i-1}),\ a \in A$. Thus, providing these conditional probabilities are easy to compute, the encoding is fast (implementation issues are relevant, but not considered here). A desirable feature of the first kind of arithmetic codes is that they operate on-line, i.e., sequentially, in the sense that $C(x_1^n)$ is always a prefix of $C(x_1^{n+1})$. The conceptual significance of the second kind of codes $\widetilde{C}(x_1^n)$ is that they are practical prefix codes effectively as good as Shannon codes for the distribution Q_n, namely the difference in length is only 1 bit. Note that strict sense Shannon codes may be of prohibitive computational complexity if the message length n is large.

2

Large deviations, hypothesis testing

2.1 Large deviations via types

An important application of information theory is to the theory of large deviations. A key to this application is the theory of types. The *type* of a sequence $x_1^n \in A^n$ is just another name for its empirical distribution $\hat{P} = \hat{P}_{x_1^n}$, that is, the distribution defined by

$$\hat{P}(a) = \frac{|\{i\colon x_i = a\}|}{n}, \; a \in A.$$

A distribution P on A is called an *n-type* if it is the type of some $x_1^n \in A^n$. The set of all $x_1^n \in A^n$ of type P is called the *type class of the n-type P* and is denoted by $\mathcal{T}_P^n$.

Lemma 2.1. The number of possible n-types is $\begin{pmatrix} n + |A| - 1 \\ |A| - 1 \end{pmatrix}$.

Proof. Left to the reader. □

Lemma 2.2. For any n-type P

$$\begin{pmatrix} n + |A| - 1 \\ |A| - 1 \end{pmatrix}^{-1} 2^{nH(P)} \le |\mathcal{T}_P^n| \le 2^{nH(P)}.$$

Proof. Let $A = \{a_1, a_2, \ldots, a_t\}$, where $t = |A|$. By the definition of types we can write $P(a_i) = k_i/n, i = 1, 2, \ldots, t$, with $k_1 + k_2 + \ldots + k_t = n$, where k_i is the number of times a_i appears in x_1^n for any fixed $x_1^n \in \mathcal{T}_P^n$. Thus we have

$$|\mathcal{T}_P^n| = \frac{n!}{k_1!k_2! \cdots k_t!}.$$

Note that

$$n^n = (k_1 + \ldots + k_t)^n = \sum \frac{n!}{j_1! \cdots j_t!} k_1^{j_1} \cdots k_t^{j_t},$$

where the sum is over all t-tuples $(j_1, \ldots, j_t)$ of nonnegative integers such that $j_1 + \ldots + j_t = n$. The number of terms is $\begin{pmatrix} n + |A| - 1 \\ |A| - 1 \end{pmatrix}$, by Lemma 2.1, and the largest term is

$$\frac{n!}{k_1!k_2! \cdots k_t!} k_1^{k_1} k_2^{k_2} \cdots k_t^{k_t},$$

for if $j_r > k_r,\ j_s < k_s$ then decreasing j_r by 1 and increasing j_s by 1 multiplies the corresponding term by

$$\frac{j_r}{k_r} \frac{k_s}{1 + j_s} \geq \frac{j_r}{k_r} > 1.$$

The lemma now follows from the fact that the sum is bounded below by its largest term and above by the largest term times the number of terms, and noting that

$$\frac{n^n}{k_1^{k_1} k_2^{k_2} \cdots k_t^{k_t}} = \prod_{i=1}^{t} \left(\frac{k_i}{n}\right)^{-k_i} = \prod_{i=1}^{t} P(a_i)^{-nP(a_i)} = 2^{nH(P)}.$$

□

The next result connects the theory of types with general probability theory. For any distribution P on A, let P^n denote the distribution of n independent drawings from P, that is $P^n(x_1^n) = \prod_{i=1}^n P(x_i),\ x_1^n \in A^n$.

Lemma 2.3. For any distribution P on A and any n-type Q

$$\frac{P^n(x_1^n)}{Q^n(x_1^n)} = 2^{-nD(Q\|P)}, \text{ if } x_1^n \in \mathcal{T}_Q^n,$$

$$\begin{pmatrix} n + |A| - 1 \\ |A| - 1 \end{pmatrix}^{-1} 2^{-nD(Q\|P)} \leq P^n(\mathcal{T}_Q^n) \leq 2^{-nD(Q\|P)}.$$

Corollary 2.1. Let $\hat{P}_n$ denote the empirical distribution (type) of a random sample of size n drawn from P. Then

$$\text{Prob}(D(\hat{P}_n\|P) \geq \delta) \leq \binom{n+|A|-1}{|A|-1} 2^{-n\delta},\ \forall \delta > 0.$$

Proof. If $x_1^n \in \mathcal{T}_Q^n$ the number of times $x_i = a$ is just $nQ(a)$, so that

$$\frac{P^n(x_1^n)}{Q^n(x_1^n)} = \prod_a \left(\frac{P(a)}{Q(a)}\right)^{nQ(a)} = 2^{\left(n\sum_a Q(a)\log\frac{P(a)}{Q(a)}\right)} = 2^{-nD(Q\|P)},$$

that is,

$$P^n(\mathcal{T}_Q^n) = Q^n(\mathcal{T}_Q^n) 2^{-nD(Q\|P)}.$$

Here $Q^n(\mathcal{T}_Q^n) \geq \binom{n+|A|-1}{|A|-1}^{-1}$, by Lemma 2.2 and the fact that $Q^n(x_1^n) = 2^{-nH(Q)}$ if $x_1^n \in \mathcal{T}_Q^n$. The probability in the Corollary equals the sum of $P^n(\mathcal{T}_Q^n)$ for all n-types Q with $D(Q\|P) \geq \delta$, thus Lemmas 2.1 and 2.3 yield the claimed bound. □

The empirical distribution $\hat{P}_n$ in the Corollary converges to P with probability 1 as $n \to \infty$, by the law of large numbers, or by the very Corollary (and Borel–Cantelli). The next result, the finite alphabet special case of the celebrated Sanov theorem, is useful for estimating the (exponentially small) probability that $\hat{P}_n$ belongs to some set Π of distributions that does not contain the true distribution P.

We use the notation $D(\Pi\|P) = \inf_{Q\in\Pi} D(Q\|P)$.

Theorem 2.1 (Sanov's Theorem.). Let Π be a set of distributions on A whose closure is equal to the closure of its interior. Then for the empirical distribution of a sample from a strictly positive distribution P on A,

$$-\frac{1}{n}\log \mathit{Prob}\left(\hat{P}_n \in \Pi\right) \to D(\Pi\|P).$$

Proof. Let $\mathcal{P}_n$ be the set of possible n-types and let $\Pi_n = \Pi \cap \mathcal{P}_n$. Lemma 2.3 implies that

$$\text{Prob}\ (\hat{P}_n \in \Pi_n) = P^n\left(\cup_{Q\in\Pi_n}\mathcal{T}_Q^n\right)$$

is upper bounded by

$$\binom{n+|A|-1}{|A|-1} 2^{-nD(\Pi_n\|P)}$$

and lower bounded by

$$\binom{n+|A|-1}{|A|-1}^{-1} 2^{-nD(\Pi_n\|P)}.$$

Since $D(Q\|P)$ is continuous in Q, the hypothesis on Π implies that $D(\Pi_n\|P)$ is arbitrarily close to $D(\Pi\|P)$ if n is large. Hence the theorem follows. □

Example 2.1. Let f be a given function on A and set $\Pi = \{Q: \sum_a Q(a)f(a) > \alpha\}$ where $\alpha < \max_a f(a)$. The set Π is open and hence satisfies the hypothesis of Sanov's theorem. The empirical distribution of a random sample $X_1, ..., X_n$ belongs to Π iff $(1/n)\sum_i f(X_i) > \alpha$, since $\sum_a \hat{P}_n(a)f(a) = (1/n)\sum_i f(X_i)$. Thus we obtain the large deviations result

$$-\frac{1}{n}\log \operatorname{Prob}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) > \alpha\right) \to D(\Pi\|P).$$

In this case, $D(\Pi\|P) = D(\operatorname{cl}(\Pi)\|P) = \min D(Q\|P)$, where the minimum is over all Q for which $\sum Q(a)f(a) \geq \alpha$. In particular, for $\alpha > \sum P(a)f(a)$ we have $D(\Pi\|P) > 0$, so that, the probability that $(1/n)\sum_1^n f(X_i) > \alpha$ goes to 0 exponentially fast.

It is instructive to see how to calculate the exponent $D(\Pi\|P)$ for the preceding example. Consider the exponential family of distributions $\tilde{P}$ of the form $\tilde{P}(a) = cP(a)2^{tf(a)}$, where $c = (\sum_a P(a)2^{tf(a)})^{-1}$. Clearly $\sum_a \tilde{P}(a)f(a)$ is a continuous function of the parameter t and this function tends to $\max f(a)$ as $t \to \infty$. (Check!) As $t = 0$ gives $\tilde{P} = P$, it follows by the assumption

$$\sum_a P(a)f(a) < \alpha < \max_a f(a)$$

that there is an element of the exponential family, with $t > 0$, such that $\sum \tilde{P}(a)f(a) = \alpha$. Denote such a $\tilde{P}$ by Q^*, so that,

$$Q^*(a) = c^* P(a) 2^{t^* f(a)},\ t^* > 0,\ \sum_a Q^*(a)f(a) = \alpha.$$

We claim that

$$D(\Pi\|P) = D(Q^*\|P) = \log c^* + t^*\alpha. \tag{2.1}$$

To show that $D(\Pi\|P) = D(Q^*\|P)$ it suffices to show that $D(Q\|P) > D(Q^*\|P)$ for every $Q \in \Pi$, i. e., for every Q for which $\sum_a Q(a)f(a) > \alpha$. A direct calculation gives

$$\begin{aligned} D(Q^*\|P) &= \sum_a Q^*(a) \log \frac{Q^*(a)}{P(a)} \\ &= \sum_a Q^*(a) \left[\log c^* + t^* f(a)\right] = \log c^* + t^*\alpha \end{aligned} \tag{2.2}$$

and

$$\sum_a Q(a) \log \frac{Q^*(a)}{P(a)} = \sum_a Q(a) \left[\log c^* + t^* f(a)\right] > \log c^* + t^*\alpha.$$

Hence

$$D(Q\|P) - D(Q^*\|P) > D(Q\|P) - \sum_a Q(a) \log \frac{Q^*(a)}{P(a)} = D(Q\|Q^*) > 0.$$

This completes the proof of (2.1).

Remark 2.1. Replacing P in (2.2) by any $\tilde{P}$ of the exponential family, i. e., $\tilde{P}(a) = cP(a)2^{tf(a)}$, we get that

$$\begin{aligned} &D(Q^*\|\tilde{P}) = \\ &\qquad \log \frac{c^*}{c} + (t^* - t)\alpha = \log c^* + t^*\alpha - (\log c + t\alpha). \end{aligned}$$

Since $D(Q^*\|\tilde{P}) > 0$ for $\tilde{P} \neq Q^*$, it follows that

$$\log c + t\alpha = -\log \sum_a P(a) 2^{tf(a)} + t\alpha$$

attains its maximum at $t = t^*$. This means that the "large deviations exponent"

$$\lim_{n\to\infty} \left[-\frac{1}{n} \log \operatorname{Prob} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) > \alpha) \right) \right]$$

can be represented also as

$$\max_{t\geq 0} \left[-\log \sum_a P(a) 2^{tf(a)} + t\alpha \right].$$

This latter form is the one usually found in textbooks, with the formal difference that logarithm and exponentiation with base e rather than base 2 are used. Note that the restriction $t \geq 0$ is not needed when $\alpha > \sum_a P(a)f(a)$, because, as just seen, the unconstrained maximum is attained at $t^* > 0$. However, the restriction to $t \geq 0$ takes care also of the case when $\alpha \leq \sum_a P(a)f(a)$, when the exponent is equal to 0.

2.2 Hypothesis testing

Let us consider now the problem of hypothesis testing. Suppose the statistician, observing independent drawings from an unknown distribution P on A, wants to test the "null-hypothesis" that P belongs to a given set Π of distributions on A. A (nonrandomized) test of sample size n is determined by a set $C \subseteq A^n$, called the *critical region*; the null-hypothesis is accepted if the observed sample x_1^n does not belong to C. Usually the test is required to have type 1 error probability not exceeding some $\epsilon > 0$, that is, $P^n(C) \leq \epsilon$, for all $P \in \Pi$. Subject to this constraint, it is desirable that the type 2 error probability, that is $P(A^n - C)$, when $P \notin \Pi$, be small, either for a specified $P \notin \Pi$ ("testing against a simple alternative hypothesis") or, preferably, for all $P \notin \Pi$.

Theorem 2.2. Let P_1 and P_2 be any two distributions on A, let α be a positive number, and for each $n \geq 1$ suppose $B_n \subseteq A^n$ satisfies $P_1^n(B_n) \geq \alpha$. Then

$$\liminf_{n\to\infty} \frac{1}{n} \log P_2^n(B_n) \geq -D(P_1 \| P_2).$$

The assertion of Theorem 2.2 and the special case of Theorem 2.3, below, that there exists sets $B_n \subset A^n$ satisfying

$$P_1^n(B_n) \to 1, \quad \frac{1}{n} \log P_2^n(B_n) = -D(P_1 \| P_2),$$

are together known as *Stein's lemma.*

Remark 2.2. On account of the log-sum inequality (see Section 1), we have for any $B \subset A^n$

$$\begin{aligned} P_1^n(B) \log \frac{P_1^n(B)}{P_2^n(B)} + P_1^n(A^n - B) \log \frac{P_1^n(A^n - B)}{P_2^n(A^n - B)} &\leq D(P_1^n \| P_2^n) \\ &= n\, D(P \| Q), \end{aligned}$$

a special case of the lumping property in Lemma 4.1, Section 4. Since $t \log t + (1-t)\log(1-t) \geq -1$ for each $0 \leq t \leq 1$, it follows that

$$\log P_2^n(B) \geq -\frac{nD(P\|Q)+1}{P_1^n(B)}\ .$$

Were the hypothesis $P_1^n(B_n) \geq \alpha$ of Theorem 2.2 strengthened to $P_1^n(B_n) \to 1$, the assertion of that theorem would immediately follow from the last inequality.

Proof of Theorem 2.2. With $\delta_n = \frac{|A| \log n}{n}$, say, Corollary 2.1 gives that the empirical distribution $\hat{P}_n$ of a sample drawn from P_1 satisfies $\text{Prob}(D(\hat{P}_n \| P_1) \geq \delta_n) \to 0$. This means that the P_1^n-probability of the union of the type classes $\mathcal{T}_Q^n$ with $D(Q\|P_1) < \delta_n$ approaches 1 as $n \to \infty$. Thus the assumption $P_1^n(B_n) \geq \alpha$ implies that the intersection of B_n with the union of these type classes has P_1^n-probability at least $\alpha/2$ when n is large, and consequently there exists n-types Q_n with $D(Q_n \| P_1) < \delta_n$ such that

$$P_1^n(B_n \cap \mathcal{T}_{Q_n}^n) \geq \frac{\alpha}{2} P_1^n(\mathcal{T}_{Q_n}^n).$$

Since samples in the same type class are equiprobable under P^n for each distribution P on A, the last inequality holds for P_2 in place of P_1. Hence, using Lemma 2.3,

$$P_2^n(B_n) \geq \frac{\alpha}{2} P_2^n(\mathcal{T}_{Q_n}^n) \geq \frac{\alpha}{2} \begin{pmatrix} n + |A| - 1 \\ |A| - 1 \end{pmatrix} 2^{-nD(Q_n \| P_2)}.$$

As $D(Q_n\|P_1) < \delta_n \to 0$ implies that $D(Q_n\|P_2) \to D(P_1\|P_2)$, this completes the proof of Theorem 2.2. □

Theorem 2.3. For testing the null-hypothesis that $P \in \Pi$, where Π is a closed set of distributions on A, the tests with critical region

$$C_n = \left\{x_1^n\colon \inf_{P\in\Pi} D(\hat{P}_{x_1^n}\|P) \geq \delta_n\right\},\ \delta_n = \frac{|A|\log n}{n},$$

have type 1 error probability not exceeding ϵ_n, where $\epsilon_n \to 0$, and for each $P_2 \notin \Pi$, the type 2 error probability goes to 0 with exponential rate $D(\Pi\|P_2)$.

Proof. The assertion about type 1 error follows immediately from Corollary 2.1. To prove the remaining assertion, note that for each $P_2 \notin \Pi$, the type 2 error probability $P_2(A^n - C_n)$ equals the sum of $P_2^n(\mathcal{T}_Q^n)$ for all n-types Q such that $\inf\limits_{P\in\Pi} D(Q\|P) < \delta_n$. Denoting the minimum of $D(Q\|P_2)$ for these n-types by ξ_n, it follows by Lemmas 2.1 and 2.3, that

$$P_2^n(A^n - C_n) \leq \left(\begin{array}{c} n+|A|-1 \\ |A|-1 \end{array}\right) 2^{-n\xi_n}.$$

A simple continuity argument gives $\lim\limits_{n\to\infty} \xi_n = \inf\limits_{P\in\Pi} D(P\|P_2) = D(\Pi\|P_2)$, and hence

$$\limsup_{n\to\infty} \frac{1}{n} \log P_2^n(A^n - C_n) \leq -D(\Pi\|P_2).$$

As noted in Remark 2.3, below, the opposite inequality also holds, hence

$$\lim_{n\to\infty} \frac{1}{n} \log P_2^n(A^n - C_n) = -D(\Pi\|P_2),$$

which completes the proof of the theorem. □

Remark 2.3. On account of Theorem 2.2, for any sets $C_n \subseteq A^n$, such that $P^n(C_n) \leq \epsilon < 1$, for all $P \in \Pi, n \geq 1$, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log P_2^n(A^n - C_n) \geq -D(\Pi\|P_2), \forall P_2 \notin \Pi.$$

Hence, the tests in Theorem 2.3 are asymptotically optimal against all alternatives $P_2 \notin \Pi$. The assumption that Π is closed guarantees that $D(\Pi \| P_2) > 0$, whenever $P_2 \notin \Pi$. Dropping that assumption, the type 2 error probability still goes to 0 with exponential rate $D(\Pi \| P_2)$ for P_2 not in the closure of Π, but may not go to 0 for P_2 on the boundary of Π. Finally, it should be mentioned that the criterion $\inf\limits_{P \in \Pi} D(\hat{P}_{x_1^n} \| P) \geq \delta_n$ defining the critical region of the tests in Theorem 2.3 is equivalent, by Lemma 2.3, to

$$\frac{\sup_{P \in \Pi} P^n(x_1^n)}{Q^n(x_1^n)} \leq 2^{-n\delta_n} = n^{-|A|},\ Q = \hat{P}_{x_1^n}.$$

Here the denominator is the maximum of $P^n(x_1^n)$ for all distributions P on A, thus the asymptotically optimal tests are *likelihood ratio tests* in statistical terminology.

We conclude this subsection by briefly discussing *sequential tests*. In a sequential test, the sample size is not predetermined, rather $x_1, x_2, \ldots$ are drawn sequentially until a stopping time N that depends on the actual observations, and the null hypothesis is accepted or rejected on the basis of the sample x_1^N of random size N.

A *stopping time* N for sequentially drawn $x_1, x_2, \ldots$ is defined, for our purposes, by the condition $x_1^N \in G$, for a given set $G \subset A^*$ of finite sequences that satisfies the prefix condition: no $u \in G$ is a prefix of another $v \in G$; this ensures uniqueness in the definition of N. For existence with probability 1, when $x_1, x_2, \ldots$ are i.i.d. drawings from a distribution P on A, we assume that

$$P^\infty(\{x_1^\infty : x_1^\infty \succ u \ \text{ for some } \ u \in G\}) = 1$$

where P^∞ denotes the infinite product measure on A^∞. When several possible distributions are considered on A, as in hypothesis testing, this condition in assumed for all of them. As $P^\infty(\{x_1^\infty : x_1^\infty \succ u\}) = P^n(u)$ if $u \in A^n$, the last condition equivalently means that

$$P^N(u) = P^n(u) \quad \text{if} \quad u \in G \cap A^n$$

defines a probability distribution P^N on G.

A *sequential test* is specified by a stopping time N or a set $G \subset A^*$ as above, and by a set $C \subset G$. Sequentially drawing $x_1, x_2, \ldots$ until the stopping time N, the null hypothesis is rejected if $x_1^N \in C$, and accepted if $x_1^N \in G - C$. Thus, the set of possible samples is G and the critical region is C.

Let us restrict attention to testing a simple null hypothesis P_1 against a simple alternative P_2, where P_1 and P_2 are strictly positive distributions on A. Then, with the above notation, the type 1 and type 2 error probabilities, denoted by α and β, are given by

$$\alpha = P_1^N(C), \quad \beta = P_2^N(G - C).$$

The log-sum inequality implies the bound

$$\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta} \leq \sum_{u \in G} P_1^N(u) \log \frac{P_1^N(u)}{P_2^N(u)} = D(P_1^N \| P_2^N)$$

whose special case, for N equal to a constant n, appears in Remark 2.2. Here the right hand side is the expectation of

$$\log \frac{P_1^N(x_1^N)}{P_2^N(x_1^N)} = \sum_{i=1}^{N} \log \frac{P_1(x_i)}{P_2(x_i)}$$

under P_1^N or, equivalently, under the infinite product measure P_1^∞ on A^∞. As the terms of this sum are i.i.d. under P_1^∞, with finite expectation $D(P_1 \| P_2)$, and their (random) number N is a stopping time, Wald's identity gives

$$D(P_1^N \| P_2^N) = E_1(N) D(P_1 \| P_2)$$

whenever the expectation $E_1(N)$ of N under P_1^∞ (the average sample size when hypothesis P_1 is true) is finite. It follows as in Remark 2.2 that

$$\log \beta \geq - \frac{E_1(N)\ D(P_1 \| P_2) + 1}{1 - \alpha}\ ,$$

thus sequential tests with type 1 error probability $\alpha \to 0$ cannot have type 2 error probability exponentially smaller than $2^{-E_1(N)\ D(P_1 \| P_2)}$. In this sense, sequential tests are not superior to tests with constant sample size.

On the other hand, sequential tests can be much superior in the sense that one can have $E_1(N) = E_2(N) \to \infty$ and *both* error probabilities decreasing at the best possible exponential rates, namely with exponents $D(P_1\|P_2)$ and $D(P_2\|P_1)$. This would immediately follow if in the bound

$$\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha)\log\frac{1-\alpha}{\beta} \le D(P_1^N\|P_2^N)$$

and its counterpart obtained by reversing the roles of P_1 and P_2, the equality could be achieved for tests with $E_1(N) = E_2(N) \to \infty$.

The condition of equality in the log-sum inequality gives that both these bounds hold with the equality if and only if the probability ratio $\frac{P_1^N(u)}{P_2^N(u)}$ is constant for $u \in C$ and also for $u \in G - C$. That condition cannot be met exactly, in general, but it is possible to make $\frac{P_1^N(u)}{P_2^N(u)}$ "nearly constant" on both C and $G - C$.

Indeed, consider the *sequential probability ratio test*, with stopping time N equal to the smallest n for which

$$c_1 < \frac{P_1^n(x_1^n)}{P_2^n(x_1^n)} < c_2$$

does not hold, where $c_1 < 1 < c_2$ are given constants, and with the decision rule that P_1 or P_2 is accepted according as the second or the first inequality is violated at this stopping time. For $C \subset G \subset A^*$ implicitly defined by this description, it is obvious that

$$\frac{P_1^N(a)}{P_2^N(a)} \in (c_1 m, c_1] \text{ if } a \in C, \quad \frac{P_1^N(a)}{P_2^N(a)} \in [c_2, c_2 M) \text{ if } a \in G - C,$$

where m and M are the minimum and maximum of the ratio $\frac{P_1(a)}{P_2(a)}$, $a \in A$. The fact that $\frac{P_1^N(a)}{P_2^N(a)}$ is nearly constant in this sense both for $a \in C$ and $a \in G - C$, is sufficient to show that the mentioned two bounds "nearly" become equalities, asymptotically, if $E_1(N) = E_2(N) \to \infty$ (the latter can be achieved by suitable choice of c_1 and c_2 with $c_1 \to 0$, $c_2 \to \infty$). Then, both the type 1 and type 2 error probabilities of these sequential probability ratio tests go to 0 with the best possible exponential rates. The details are omitted.

3

I-projections

Information divergence of probability distributions can be interpreted as a (nonsymmetric) analogue of squared Euclidean distance. With this interpretation, several results in this Section are intuitive "information geometric" counterparts of standard results in Euclidean geometry, such as the inequality in Theorem 3.1 and the identity in Theorem 3.2.

The *I-projection* of a distribution Q onto a (non-empty) closed, convex set Π of distributions on A is the $P^* \in \Pi$ such that

$$D(P^*\|Q) = \min_{P\in\Pi} D(P\|Q).$$

In the sequel we suppose that $Q(a) > 0$ for all $a \in A$. The function $D(P\|Q)$ is then continuous and strictly convex in P, so that P^* exists and is unique.

The *support* of the distribution P is the set $S(P) = \{a\colon P(a) > 0\}$. Since Π is convex, among the supports of elements of Π there is one that contains all the others; this will be called the *support of* Π and denoted by $S(\Pi)$.

Theorem 3.1. $S(P^*) = S(\Pi)$, and $D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q)$ for all $P \in \Pi$.

Of course, if the asserted inequality holds for some $P^* \in \Pi$ and all $P \in \Pi$ then P^* must be the I-projection of Q onto Π.

Proof. For arbitrary $P \in \Pi$, by the convexity of Π we have $P_t = (1-t)P^* + tP \in \Pi$, for $0 \leq t \leq 1$, hence for each $t \in (0,1)$,

$$0 \leq \frac{1}{t}\left[D(P_t\|Q) - D(P^*\|Q)\right] = \frac{d}{dt}D(P_t\|Q)\,|_{t=\tilde{t}}\ ,$$

for some $\tilde{t} \in (0,t)$. But

$$\frac{d}{dt}D(P_t\|Q) = \sum_a (P(a) - P^*(a)) \log \frac{P_t(a)}{Q(a)},$$

and this converges (as $t \downarrow 0$) to $-\infty$ if $P^*(a) = 0$ for some $a \in S(P)$, and otherwise to

$$\sum_a (P(a) - P^*(a)) \log \frac{P^*(a)}{Q(a)}. \tag{3.1}$$

It follows that the first contingency is ruled out, proving that $S(P^*) \supseteq S(P)$, and also that the quantity (3.1) is nonnegative, proving the claimed inequality. □

Now we examine some situations in which the inequality of Theorem 3.1 is actually an equality. For any given functions $f_1, f_2, \ldots, f_k$ on A and numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$, the set

$$\mathcal{L} = \{P\colon \sum_a P(a)f_i(a) = \alpha_i, 1 \leq i \leq k\},$$

if non-empty, will be called a *linear family* of probability distributions. Moreover, the set $\mathcal{E}$ of all P such that

$$P(a) = cQ(a)\exp\left(\sum_1^k \theta_i f_i(a)\right), \text{ for some } \theta_1, \ldots, \theta_k,$$

will be called an *exponential family* of probability distributions; here Q is any given distribution and

$$c = c(\theta_1, \ldots, \theta_k) = \left(\sum_a Q(a)\exp\left(\sum_1^k \theta_i f_i(a)\right)\right)^{-1}.$$

We will assume that $S(Q) = A$; then $S(P) = A$ for all $P \in \mathcal{E}$. Note that $Q \in \mathcal{E}$. The family $\mathcal{E}$ depends on Q, of course, but only in a weak manner, for any element of $\mathcal{E}$ could play the role of Q. If necessary to emphasize this dependence on Q we shall write $\mathcal{E} = \mathcal{E}_Q$.

Linear families are closed sets of distributions, exponential families are not. Sometimes it is convenient to consider the closure $\mathrm{cl}(\mathcal{E})$ of an exponential family $\mathcal{E}$.

Theorem 3.2. The I-projection P^* of Q onto a linear family $\mathcal{L}$ satisfies the *Pythagorean identity*

$$D(P\|Q) = D(P\|P^*) + D(P^*\|Q),\ \forall P \in \mathcal{L}.$$

Further, if $S(\mathcal{L}) = A$ then $\mathcal{L} \cap \mathcal{E}_Q = \{P^*\}$, and, in general, $\mathcal{L} \cap \mathrm{cl}(\mathcal{E}_Q) = \{P^*\}$.

Corollary 3.1. For a linear family $\mathcal{L}$ and exponential family $\mathcal{E}$, defined by the same functions $f_1, ..., f_k$, the intersection $\mathcal{L} \cap \mathrm{cl}(\mathcal{E})$ consists of a single distribution P^*, and

$$D(P\|Q) = D(P\|P^*) + D(P^*\|Q),\ \forall P \in \mathcal{L}, Q \in \mathrm{cl}(\mathcal{E}).$$

Proof of Theorem 3.2. By the preceding theorem, $S(P^*) = S(\mathcal{L})$. Hence for every $P \in \mathcal{L}$ there is some $t < 0$ such that $P_t = (1-t)P^* + tP \in \mathcal{L}$. Therefore, we must have $(d/dt)D(P_t\|Q)|_{t=0} = 0$, that is, the quantity (3.1) in the preceding proof is equal to 0, namely,

$$\sum_a (P(a) - P^*(a)) \log \frac{P^*(a)}{Q(a)} = 0,\ \forall P \in \mathcal{L}. \tag{3.2}$$

This proves that P^* satisfies the Pythagorean identity.

By the definition of linear family, the distributions $P \in \mathcal{L}$, regarded as $|A|$-dimensional vectors, are in the orthogonal complement $\mathcal{F}^\perp$ of the subspace $\mathcal{F}$ of $R^{|A|}$, spanned by the k vectors $f_i(\cdot) - \alpha_i, 1 \le i \le k$. If $S(\mathcal{L}) = A$ then the distributions $P \in \mathcal{L}$ actually span the orthogonal complement of $\mathcal{F}$ (any subspace of $R^{|A|}$ that contains a strictly positive vector is spanned by the probability vectors in that subspace; the proof is left to the reader.) Since the identity (3.2) means that the vector

$$\log \frac{P^*(\cdot)}{Q(\cdot)} - D(P^* \| Q)$$

is orthogonal to each $P \in \mathcal{L}$, it follows that this vector belongs to $(\mathcal{F}^\perp)^\perp = \mathcal{F}$. This proves that $P^* \in \mathcal{E}$, if $S(\mathcal{L}) = A$.

Next we show that any distribution $P^* \in \mathcal{L} \cap \mathrm{cl}(\mathcal{E}_Q)$ satisfies (3.2). Since (3.2) is equivalent to the Pythagorean identity, this will show that $\mathcal{L} \cap \mathrm{cl}(\mathcal{E}_Q)$, if nonempty, consists of the single distribution equal to the I-projection of Q onto $\mathcal{L}$. Now, let $P_n \in \mathcal{E}, P_n \to P^* \in \mathcal{L}$. By the definition of $\mathcal{E}$,

$$\log \frac{P_n(a)}{Q(a)} = \log c_n + (\log e) \sum_{i=1}^{k} \theta_{i,n} f_i(a).$$

As $P \in \mathcal{L}, P^* \in \mathcal{L}$ implies $\sum P(a) f_i(a) = \sum P^*(a) f_i(a), i = 1, \ldots, k$, it follows that

$$\sum_a (P(a) - P^*(a)) \log \frac{P_n(a)}{Q(a)} = 0,\ \forall P \in \mathcal{L}.$$

Since $P_n \to P^*$, this gives (3.2).

To complete the proof of the theorem it remains to show that $\mathcal{L} \cap \mathrm{cl}(\mathcal{E})$ is always nonempty. Towards this end, let P_n^* denote the I-projection of Q onto the linear family

$$\mathcal{L}_n = \left\{ P \colon \sum_{a \in A} P(a) f_i(a) = \left(1 - \frac{1}{n}\right) \alpha_i + \frac{1}{n} \sum_{a \in A} Q(a) f_i(a), i = 1, \ldots, k \right\}.$$

Since $(1 - \frac{1}{n})P + \frac{1}{n}Q \in \mathcal{L}_n$ if $P \in \mathcal{L}$, here $S(\mathcal{L}_n) = A$ and therefore $P_n^* \in \mathcal{E}$. Thus the limit of any convergent subsequence of $\{P_n^*\}$ belongs to $\mathcal{L} \cap \mathrm{cl}(\mathcal{E})$. □

Proof of Corollary 3.1. Only the validity of the Pythagorean identity for $Q \in \mathrm{cl}(\mathcal{E})$ needs checking. Since that identity holds for $Q \in \mathcal{E}$, taking limits shows that the identity holds also for the limit of a sequence $Q_n \in \mathcal{E}$, that is, for each Q in $\mathrm{cl}(\mathcal{E})$. □

Remark 3.1. A minor modification of the proof of Theorem 3.2 shows that the I-projection P^* of Q to a linear family with $S(\mathcal{L}) = B \subset A$ is of the form

$$P^*(a) = \begin{cases} cQ(a)\exp\left(\sum_1^k \theta_i f_i(a)\right) & \text{if } a \in B \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

This and Theorem 3.2 imply that $\mathrm{cl}(\mathcal{E}_Q)$ consists of distributions of the form (3.3), with $B = S(\mathcal{L})$ for suitable choice of the constants $\alpha_1, \dots, \alpha_k$ in the definition of $\mathcal{L}$. We note without proof that also, conversely, all such distributions belong to $\mathrm{cl}(\mathcal{E}_Q)$.

Next we show that I-projections are relevant to maximum likelihood estimation in exponential families.

Given a sample $x_1^n \in A^n$ drawn from an unknown distribution supposed to belong to a feasible set Π of distributions on A, a *maximum likelihood estimate (MLE)* of the unknown distribution is a maximizer of $P^n(x_1^n)$ subject to $P \in \Pi$; if the maximum is not attained the MLE does not exist.

Lemma 3.1. An MLE is the same as a minimizer of $D(\widehat{P}\|P)$ for P in the set of feasible distributions, where $\widehat{P}$ is the empirical distribution of the sample.

Proof. Immediate from Lemma 2.3. □

In this sense, an MLE can always be regarded as a "reverse I-projection". In the case when Π is an exponential family, the MLE equals a proper I-projection, though not of $\hat{P}$ onto Π.

Theorem 3.3. Let the set of feasible distributions be the exponential family

$$\mathcal{E} = \left\{P\colon P(a) = c(\theta_1, \dots, \theta_k)Q(a)\exp(\sum_{i=1}^k \theta_i f_i(a)), (\theta_1, \dots, \theta_k) \in R^k\right\},$$

where $S(Q) = A$. Then, given a sample $x_1^n \in A^n$, the MLE is unique and equals the I-projection P^* of Q onto the linear family

$$\mathcal{L} = \{P\colon \sum_a P(a)f_i(a) = \frac{1}{n}\sum_{j=1}^n f_i(x_j),\ 1 \le i \le k\},$$

provided $S(\mathcal{L}) = A$. If $S(\mathcal{L}) \neq A$, the MLE does not exist, but P^* will be the MLE in that case if $\mathrm{cl}(\mathcal{E})$ rather than $\mathcal{E}$ is taken as the set of feasible distributions.

Proof. The definition of $\mathcal{L}$ insures that $\hat{P} \in \mathcal{L}$. Hence by Theorem 3.2 and its Corollary,

$$D(\hat{P}\|P) = D(\hat{P}\|P^*) + D(P^*\|P),\ \forall P \in \mathrm{cl}(\mathcal{E}).$$

Also by Theorem 3.2, $P^* \in \mathcal{E}$ if and only if $S(\mathcal{L}) = A$, while always $P^* \in \mathrm{cl}(\mathcal{E})$. Using this, the last divergence identity gives that the minimum of $D(\hat{P}\|P)$ subject to $P \in \mathcal{E}$ is uniquely attained for $P = P^*$, if $S(\mathcal{L}) = A$, and is not attained if $S(\mathcal{L}) \neq A$, while P^* is always the unique minimizer of $D(\hat{P}\|P)$ subject to $P \in \mathrm{cl}(\mathcal{E})$. On account of Lemma 3.1, this completes the proof of the theorem. □

We conclude this subsection with a counterpart of Theorem 3.1 for "reverse I-projections." The reader is invited to check that the theorem below is also an analogue of one in Euclidean geometry.

Let us be given a distribution P and a closed convex set Π of distributions on A such that $S(P) \subseteq S(\Pi)$. Then there exists $Q^* \in \Pi$ attaining the (finite) minimum $\min_{Q\in\Pi} D(P\|Q)$; this Q^* is unique if $S(P) = S(\Pi)$, but need not be otherwise.

Theorem 3.4. A distribution $Q^* \in \Pi$ minimizes $D(P\|Q)$ subject to $Q \in \Pi$ if and only if for all distributions P' on A and $Q' \in \Pi$,

$$D(P'\|Q') + D(P'\|P) \geq D(P'\|Q^*).$$

Proof. The "if" part is obvious (take $P' = P$.) To prove the "only if" part, $S(P') \subseteq S(Q') \cap S(P)$ may be assumed else the left hand side is infinite. We claim that

$$\sum_{a\in S(P)} P(a)\left(1 - \frac{Q'(a)}{Q^*(a)}\right) \geq 0. \tag{3.4}$$

Note that (3.4) and $S(P) \supseteq S(P')$ imply

$$\sum_{a \in S(P')} P'(a)\Big(1 - \frac{P(a)Q'(a)}{P'(a)Q^*(a)}\Big) \geq 0,$$

which, on account of $\log \frac{1}{t} \geq (1-t)\log e$, implies in turn

$$\sum_{a \in S(P')} P'(a) \log \frac{P'(a)Q^*(a)}{P(a)Q'(a)} \geq 0.$$

The latter is equivalent to the inequality in the statement of the theorem, hence it suffices to prove the claim (3.4).

Now set $Q_t = (1-t)Q^* + tQ' \in \mathcal{Q}, 0 \leq t \leq 1$. Then

$$0 \leq \frac{1}{t}\left[D(P\|Q_t) - D(P\|Q^*)\right] = \frac{d}{dt} D(P\|Q_t)\,|_{t=\tilde{t}},\ 0 < \tilde{t} \leq t.$$

With $t \to 0$ it follows that

$$\begin{aligned} 0 &\leq \lim_{\tilde{t}\to 0} \sum_{a \in S(P)} P(a) \frac{(Q^*(a) - Q'(a))\log e}{(1-\tilde{t})Q^*(a) + \tilde{t}Q'(a)} \\ &= \sum_{a \in S(P)} P(a) \frac{Q^*(a) - Q'(a)}{Q^*(a)} \log e. \end{aligned}$$

This proves the claim (3.4) and completes the proof of Theorem 3.4. □

4

f-Divergence and contingency tables

Let $f(t)$ be a convex function defined for $t > 0$, with $f(1) = 0$. The f*-divergence* of a distribution P from Q is defined by

$$D_f(P\|Q) = \sum_a Q(x) f\left(\frac{P(x)}{Q(x)}\right).$$

Here we take $0f(\frac{0}{0}) = 0,\ f(0) = \lim_{t\to 0} f(t),\ 0f(\frac{a}{0}) = \lim_{t\to 0} tf(\frac{a}{t}) = a\lim_{u\to\infty} \frac{f(u)}{u}$.

Some examples include the following.

(1) $f(t) = t\log t \ \Rightarrow D_f(P\|Q) = D(P\|Q)$.
(2) $f(t) = -\log t \Rightarrow D_f(P\|Q) = D(Q\|P)$.
(3) $f(t) = (t-1)^2$

$$\Rightarrow D_f(P\|Q) = \sum_a \frac{(P(a) - Q(a))^2}{Q(a)}.$$

(4) $f(t) = 1 - \sqrt{t}$

$$\Rightarrow D_f(P\|Q) = 1 - \sum_a \sqrt{P(a)Q(a)}.$$

(5) $f(t) = |t-1| \Rightarrow D_f(P\|Q) = |P - Q| = \sum_a |P(a) - Q(a)|$.

In addition to information divergence obtained in (1), (2), the f-divergences in (3), (4), (5) are also often used in statistics. They are called χ^2-divergence, Hellinger distance, and variational distance, respectively.

The analogue of the log-sum inequality is

$$\sum_i b_i f\left(\frac{a_i}{b_i}\right) \geq bf\left(\frac{a}{b}\right),\ a = \sum a_i,\ b = \sum b_i, \tag{4.1}$$

where if f is strictly convex at $c = a/b$, the equality holds iff $a_i = cb_i$, for all i. Using this, many of the properties of the information divergence $D(P\|Q)$ extend to general f-divergences, as shown in the next lemma. Let $\mathcal{B} = \{B_1, B_2, \ldots, B_k\}$ be a partition of A and let P be a distribution on A. The distribution defined on $\{1, 2, \ldots, k\}$ by the formula

$$P^{\mathcal{B}}(i) = \sum_{a \in B_i} P(a),$$

is called the $\mathcal{B}$-*lumping* of P.

Lemma 4.1. $D_f(P\|Q) \geq 0$ and if f is strictly convex at $t = 1$ then $D_f(P\|Q) = 0$ only when $P = Q$. Further, $D_f(P\|Q)$ is a convex function of the pair (P, Q), and the lumping property, $D_f(P\|Q) \geq D_f(P^{\mathcal{B}}\|Q^{\mathcal{B}})$ holds for any partition $\mathcal{B}$ of A.

Proof. The first assertion and the lumping property obviously follow from the analogue of the log-sum inequality, (4.1). To prove convexity, let $P = \alpha P_1 + (1-\alpha)P_2, Q = \alpha Q_1 + (1-\alpha)Q_2$. Then P and Q are lumpings of distributions $\widetilde{P}$ and $\widetilde{Q}$ defined on the set $A \times \{1, 2\}$ by $\widetilde{P}(a, 1) = \alpha P_1(a), \widetilde{P}(a, 2) = (1-\alpha)P_2(a)$, and similarly for $\widetilde{Q}$. Hence by the lumping property,

$$D_f(P\|Q) \leq D_f(\widetilde{P}\|\widetilde{Q}) = \alpha D_f(P_1\|Q_1) + (1-\alpha)D_f(P_2\|Q_2). \qquad \square$$

A basic theorem about f-divergences is the following approximation by the χ^2-divergence $\chi^2(P, Q) = \sum (P(a) - Q(a))^2/Q(a)$.

Theorem 4.1. If f is twice differentiable at $t = 1$ and $f''(1) > 0$ then for any Q with $S(Q) = A$ and P " close" to Q we have

$$D_f(P\|Q) \sim \frac{f''(1)}{2}\chi^2(P, Q).$$

Formally, $D_f(P\|Q)/\chi^2(P, Q) \to f''(1)/2$ as $P \to Q$.

Proof. Since $f(1) = 0$, Taylor's expansion gives

$$f(t) = f'(1)(t-1) + \frac{f''(1)}{2}(t-1)^2 + \epsilon(t)(t-1)^2,$$

where $\epsilon(t) \to 0$ as $t \to 1$. Hence

$$\begin{aligned} Q(a)f\left(\frac{P(a)}{Q(a)}\right) = \\ & f'(1)(P(a)-Q(a)) + \frac{f''(1)}{2}\frac{(P(a)-Q(a))^2}{Q(a)} \\ & +\epsilon\left(\frac{P(a)}{Q(a)}\right)\frac{(P(a)-Q(a))^2}{Q(a)}. \end{aligned}$$

Summing over $a \in A$ then establishes the theorem. □

Remark 4.1. The same proof works even if Q is not fixed, replacing $P \to Q$ by $P - Q \to 0$, provided that no $Q(a)$ can become arbitrarily small. However, the theorem (the "asymptotic equivalence" of f-divergences subject to the differentiability hypotheses) does not remain true if Q is not fixed and the probabilities of $Q(a)$ are not bounded away from 0.

Corollary 4.1. Let $f_0 = 1, f_1, \ldots, f_{|A|-1}$ be a basis for $R^{|A|}$ (regarded as the linear space of all real-valued functions on A), orthonormal with respect to the inner product $< g,h >_Q = \sum_a Q(a)g(a)h(a)$. Then, under the hypotheses of Theorem 4.1,

$$D_f(P\|Q) \sim \frac{f''(1)}{2}\sum_{i=1}^{|A|-1}\left(\sum_a P(a)f_i(a)\right)^2,$$

and, for the linear family

$$\mathcal{L}(\alpha) = \{P\colon \sum_a P(a)f_i(a) = \alpha_i,\ 1 \le i \le k\},$$

with $\alpha = (\alpha_1 \ldots, \alpha_k)$ approaching the zero vector,

$$\min_{P\in\mathcal{L}(\alpha)} D_f(P\|Q) \sim \frac{f''(1)}{2}\sum_{i=1}^{k}\alpha_i^2.$$

Proof. On account of Theorem 4.1, it suffices to show that

$$\chi^2(P,Q) = \sum_{i=1}^{|A|-1} \left(\sum_a P(a) f_i(a) \right)^2 \tag{4.2}$$

and, at least when $\alpha = (\alpha_1 \dots, \alpha_k)$ is sufficiently close to the zero vector,

$$\min_{P \in \mathcal{L}(\alpha)} \chi^2(P,Q) = \sum_{i=1}^{k} \alpha_i^2. \tag{4.3}$$

Now, $\chi^2(P,Q) = \sum_a Q(a) \left(\frac{P(a)}{Q(a)} - 1 \right)^2$ is the squared norm of the function g defined by $g(a) = \frac{P(a)}{Q(a)} - 1$ with respect to the given inner product, and that equals $\sum_{i=0}^{|A|-1} < g, f_i >_Q^2$. Here

$$\begin{aligned} < g, f_0 >_Q &= \sum_a (P(a) - Q(a)) = 0 \\ < g, f_i >_Q &= \sum_a (P(a) - Q(a)) f_i(a) \\ &= \sum_a P(a) f_i(a),\ 1 \le i \le |A| - 1, \end{aligned}$$

the latter since $< f_0, f_i >_Q = 0$ means that $\sum_a Q(a) f_i(a) = 0$. This proves (4.2), and (4.3) then obviously follows if some $P \in \mathcal{L}(\alpha)$ satisfies $\sum_a P(a) f_i(a) = 0,\ \ k+1 \le i \le |A| - 1$. Finally, the assumed orthonormality of $1, f_1, \dots, f_{|A|-1}$ implies that P defined by $P(a) = Q(a)(1 + \sum_{i=1}^{k} \alpha_i f_i(a))$ satisfies the last conditions, and this P is a distribution in $\mathcal{L}(\alpha)$ provided it is nonnegative, which is certainly the case if α is sufficiently close to the zero vector. □

One property distinguishing information divergence among f-divergences is transitivity of projections, as summarized in the following lemma. It can, in fact, be shown that the only f-divergence for which either of the two properties of the lemma holds is the informational divergence.

Lemma 4.2. Let P^* be the I-projection of Q onto a linear family $\mathcal{L}$. Then

(i) For any convex subfamily $\mathcal{L}' \subseteq \mathcal{L}$ the I-projections of Q and of P^* onto $\mathcal{L}'$ are the same.
(ii) For any "translate" $\mathcal{L}'$ of $\mathcal{L}$, the I-projections of Q and of P^* onto $\mathcal{L}'$ are the same, provided $S(\mathcal{L}) = A$.

$\mathcal{L}'$ is called a translate of $\mathcal{L}$ if it is defined in terms of the same functions f_i, but possibly different α_i.

Proof. By the Pythagorean identity

$$D(P\|Q) = D(P\|P^*) + D(P^*\|Q),\ P \in \mathcal{L},$$

it follows that on any subset of $\mathcal{L}$ the minimum of $D(P\|Q)$ and of $D(P\|P^*)$ are achieved by the same P. This establishes (i).

The exponential family corresponding to a translate of $\mathcal{L}$ is the same as it is for $\mathcal{L}$. Since $S(\mathcal{L}) = A$, we know by Theorem 3.2 that P^* belongs to this exponential family. But every element of the exponential family has the same I-projection onto $\mathcal{L}'$, which establishes (ii). □

In the following theorem, $\hat{P}_n$ denotes the empirical distribution of a random sample of size n from a distribution Q with $S(Q) = A$, that is, the type of the sequence $(X_1, \ldots, X_n)$ where $X_1, X_2, \ldots$ are independent random variables with distribution Q.

Theorem 4.2. Given real valued functions $f_1, \ldots, f_k, (1 \leq k < |A| - 1)$ on A such that $f_0 = 1, f_1, \ldots, f_k$ are linearly independent, let P_n^* be the I-projection of Q onto the (random) linear family

$$\mathcal{L}_n = \{P\colon \sum_a P(a) f_i(a) = \frac{1}{n} \sum_{j=1}^{n} f_i(X_j),\ 1 \leq i \leq k\}.$$

Then

$$D(\hat{P}_n\|Q) = D(\hat{P}_n\|P_n^*) + D(P_n^*\|Q),$$

each term multiplied by $\frac{2n}{\log e}$ has a χ^2 limiting distribution with $|A| - 1, |A| - 1 - k$, respectively k, degrees of freedom, and the right hand side terms are asymptotically independent.

The χ^2 distribution with k degrees of freedom is defined as the distribution of the sum of squares of k independent random variables having the standard normal distribution.

Proof of Theorem 4.2. The decomposition of $D(\hat{P}_n \| Q)$ is a special case of the Pythagorean identity, see Theorem 3.2, since clearly $\hat{P}_n \in \mathcal{L}_n$. To prove the remaining assertions, assume that $f_0 = 1, f_1, \dots, f_k$ are orthonormal for the inner product defined in Corollary 4.1. This does not restrict generality since the family $\mathcal{L}_n$ depends on $f_1, \dots, f_k$ through the linear span of $1, f_1, \dots, f_k$, only. Further, take additional functions $f_{k+1}, \dots, f_{|A|-1}$ on A to obtain a basis for $R^{|A|}$, orthonormal for the considered inner product. Then, since $\hat{P}_n \to Q$ in probability, Corollary 4.1 applied to $f(t) = t \log t$, with $f''(1) = \log e$, gives

$$\begin{aligned} D(\hat{P}_n \| Q) &\sim \frac{\log e}{2} \sum_{i=1}^{|A|-1} \left(\sum_a \hat{P}_n(a) f_i(a) \right)^2 \\ &= \frac{\log e}{2} \sum_{i=1}^{|A|-1} \left(\frac{1}{n} \sum_{j=1}^{n} f_i(X_j) \right)^2, \\ D(P_n^* \| Q) &= \min_{P \in \mathcal{L}_n} D(P \| Q) \sim \frac{\log e}{2} \sum_{i=1}^{k} \left(\frac{1}{n} \sum_{j=1}^{n} f_i(X_j) \right)^2. \end{aligned}$$

Here, asymptotic equivalence $\sim$ of random variables means that their ratio goes to 1 in probability, as $n \to \infty$.

By the assumed orthonormality of $f_0 = 1, f_1, \dots, f_{|A|-1}$, for X with distribution Q the real valued random variables $f_i(X),\ 1 \le i \le |A|-1$, have zero mean and their covariance matrix is the $(|A|-1) \times (|A|-1)$ identity matrix. It follows by the central limit theorem that the joint distribution of the random variables

$$Z_{n,i} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f_i(X_j),\ 1 \le i \le |A|-1$$

converges, as $n \to \infty$, to the joint distribution of $|A|-1$ independent random variables having the standard normal distribution.

As the asymptotic relations established above give

$$\frac{2n}{\log e} D(\hat{P}_n \| Q) \sim \sum_{i=1}^{|A|-1} Z_{n,i}^2, \quad \frac{2n}{\log e} D(P_n^* \| Q) \sim \sum_{i=1}^{k} Z_{n,i}^2,$$

and these imply by the Pythagorean identity that

$$\frac{2n}{\log e} D(\hat{P}_n \| P_n^*) \sim \sum_{i=k+1}^{|A|-1} Z_{n,i}^2,$$

all the remaining claims follow. □

Remark 4.2. $D(\hat{P}_n \| P_n^*)$ is the preferred statistic for testing the hypothesis that the sample has come from a distribution in the exponential family

$$\mathcal{E} = \left\{ P\colon\ P(a) = cQ(a) \exp\left(\sum_{i=1}^{k} \theta_i f_i(a) \right), (\theta_1, \dots, \theta_k) \in R^k \right\}.$$

Note that $D(\hat{P}_n \| P_n^*)$ equals the infimum of $D(\hat{P}_n \| P)$, subject to $P \in \mathcal{E}$, by Corollary 3.1 in Section 3, and the test rejecting the above hypothesis when $D(\hat{P}_n \| P_n^*)$ exceeds a threshold is a likelihood ratio test, see Remark 2.3 in Subsection 2.2. In this context, it is relevant that the limiting distribution of $\frac{2n}{\log e} D(\hat{P}_n \| P_n^*)$ is the same no matter which member of $\mathcal{E}$ the sample is coming from, as any $P \in \mathcal{E}$ could play the role of Q in Theorem 4.2.

Note also that Theorem 4.2 easily extends to further decompositions of $D(\hat{P}_n \| Q)$. For example, taking additional functions $f_{k+1}, \dots, f_\ell$ with $1, f_1, \dots, f_\ell$ linearly independent, let P_n^{**} be the common I-projection of Q and P_n^* to

$$\mathcal{L}_1 = \left\{ P\colon \sum_a P(a) f_i(a) = \frac{1}{n} \sum_{j=1}^{n} f_i(X_j),\ 1 \le i \le \ell \right\}.$$

Then

$$D(\hat{P}_n \| Q) = D(\hat{P}_n \| P_n^{**}) + D(P_n^{**} \| P_n^*) + D(P_n^* \| Q),$$

the right hand side terms multiplied by $\frac{2n}{\log e}$ have χ^2 limiting distributions with degrees of freedom $|A| - 1 - \ell, \ell - k, k$ respectively, and these terms are asymptotically independent.

Table 4.1 A 2-dimensional contingency table

$x(0,0)$	$x(0,1)$	$\cdots$	$x(0,r_2)$	$x(0\cdot)$
$x(1,0)$	$x(1,1)$	$\cdots$	$x(1,r_2)$	$x(1\cdot)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x(r_1,0)$	$x(r_1,1)$	$\cdots$	$x(r_1,r_2)$	$x(r_1\cdot)$
$x(\cdot 0)$	$x(\cdot 1)$	$\cdots$	$x(\cdot r_2)$	n

Now we apply some of these ideas to the analysis of contingency tables. A 2-dimensional contingency table is indicated in Table 4.1. The sample data have two features, with categories $0,\ldots,r_1$ for the first feature and $0,\ldots,r_2$ for the second feature. The cell counts

$$x(j_1,j_2),\ 0\le j_1\le r_1,\ 0\le j_2\le r_2$$

are nonnegative integers; thus in the sample there were $x(j_1,j_2)$ members that had category j_1 for the first feature and j_2 for the second. The table has two marginals with marginal counts

$$x(j_1\cdot)=\sum_{j_2=0}^{r_2}x(j_1,j_2),\ x(\cdot j_2)=\sum_{j_1=0}^{r_1}x(j_1,j_2).$$

The sum of all the counts is

$$n=\sum_{j_1}x(j_1\cdot)=\sum_{j_2}x(\cdot j_2)=\sum_{j_1}\sum_{j_2}x(j_1,j_2).$$

The term contingency table comes from this example, the cell counts being arranged in a table, with the marginal counts appearing at the margins. Other forms are also commonly used, e. g., the counts are replaced by the empirical probabilities $\hat{p}(j_1,j_2)=x(j_1,j_2)/n$, and the marginal counts are replaced by the marginal empirical probabilities $\hat{P}(j_1.)=x(j_1.)/n$ and $\hat{P}(.j_2)=x(.j_2)/n$.

In the general case the sample has d *features* of interest, with the ith feature having categories $0,1,\ldots,r_i$. The d-tuples $\omega=(j_1,\ldots,j_d)$ are called *cells*; the corresponding *cell count* $x(\omega)$ is the number of members of the sample such that, for each i, the ith feature is in the

j_ith category. The collection of possible cells will be denoted by Ω. The empirical distribution is defined by $\hat{p}(\omega) = x(\omega)/n$, where $n = \sum_\omega x(\omega)$ is the sample size. By a d-dimensional contingency table we mean either the aggregate of the cell counts $x(\omega)$, or the empirical distribution $\hat{p}$, or sometimes any distribution P on Ω (mainly when considered as a model for the "true distribution" from which the sample came.)

The *marginals* of a contingency table are obtained by restricting attention to those features i that belong to some given set $\gamma \subset \{1, 2, \dots, d\}$. Formally, for $\gamma = (i_1, \dots, i_k)$ we denote by $\omega(\gamma)$ the γ-projection of $\omega = (j_1, \dots, j_d)$, that is, $\omega(\gamma) = (j_{i_1}, j_{i_2}, \dots, j_{i_k})$. The γ-marginal of the contingency table is given by the marginal counts

$$x(\omega(\gamma)) = \sum_{\omega' : \omega'(\gamma) = \omega(\gamma)} x(\omega')$$

or the corresponding empirical distribution $\hat{p}(\omega(\gamma)) = x(\omega(\gamma))/n$. In general the γ-marginal of any distribution $\{P(\omega) : \omega \in \Omega\}$ is defined as the distribution P_γ defined by the marginal probabilities

$$P_\gamma(\omega(\gamma)) = \sum_{\omega' : \omega'(\gamma) = \omega(\gamma)} P(\omega').$$

A d-dimensional contingency table has d one-dimensional marginals, $d(d-1)/2$ two-dimensional marginals, ..., corresponding to the subsets of $\{1, \dots, d\}$ of one, two, ..., elements.

For contingency tables the most important linear families of distributions are those defined by fixing certain γ-marginals, for a family Γ of sets $\gamma \subset \{1, \dots, d\}$. Thus, denoting the fixed marginals by $\bar{P}_\gamma, \gamma \in \Gamma$, we consider

$$\mathcal{L} = \{P : P_\gamma = \bar{P}_\gamma, \gamma \in \Gamma\}.$$

The exponential family (through any given Q) that corresponds to this linear family $\mathcal{L}$ consists of all distributions that can be represented in product form as

$$P(\omega) = cQ(\omega) \prod_{\gamma \in \Gamma} a_\gamma(\omega(\gamma)). \tag{4.4}$$

In particular, if $\mathcal{L}$ is given by fixing the one-dimensional marginals (i. e., Γ consists of the one point subsets of $\{1, \dots, d\}$) then the corresponding exponential family consists of the distributions of the form

$$P(i_1, \dots, i_d) = cQ(i_1, \dots, i_d)a_1(i_1) \cdots a_d(i_d).$$

The family of all distributions of the form (4.4) is called a *log-linear family* with *interactions* $\gamma \in \Gamma$. In most applications, Q is chosen as the uniform distribution; often the name "log-linear family" is restricted to this case. Then (4.4) gives that the log of $P(\omega)$ is equal to a sum of terms, each representing an "interaction" $\gamma \in \Gamma$, for it depends on $\omega = (j_1, \dots, j_d)$ only through $\omega(\gamma) = (j_{i_1}, \dots, j_{i_k})$, where $\gamma = (i_1, \dots, i_k)$.

A log-linear family is also called a *log-linear model.* It should be noted that the representation (4.4) is not unique, because it corresponds to a representation in terms of linearly dependent functions. A common way of achieving uniqueness is to postulate $a_\gamma(\omega(\gamma)) = 1$ whenever at least one component of $\omega(\gamma)$ is equal to 0. In this manner a unique representation of the form (4.4) is obtained, provided that with every $\gamma \in \Gamma$ also the subsets of γ are in Γ. Log-linear models of this form are called *hierarchical models.*

Remark 4.3. The way we introduced log-linear models shows that restricting to the hierarchical ones is more a notational than a real restriction. Indeed, if some γ-marginal is fixed then so are the γ'-marginals for all $\gamma' \subseteq \gamma$.

In some cases of interest it is desirable to summarize the information content of a contingency table by its γ-marginals, $\gamma \in \Gamma$. In such cases it is natural to consider the linear family $\mathcal{L}$ consisting of those distributions whose γ-marginals equal those of the empirical distribution, $\hat{P}$. If a prior guess Q is available, then we accept the I-projection P^* of Q onto $\mathcal{L}$ as an estimate of the true distribution. By Theorem 3.2, this P^* equals the intersection of the log-linear family (4.4), or its closure, with the linear family $\mathcal{L}$. Also, P^* equals the maximum likelihood estimate of the true distribution if it is assumed to belong to the family (4.4).

By Theorem 2.3, an asymptotically optimal test of the null-hypothesis that the true distribution belongs to the log-linear family $\mathcal{E}$ with interactions $\gamma \in \Gamma$ consists in rejecting the null-hypothesis if

$$D(\hat{P}\|P^*) = \min_{P \in \mathcal{E}} D(\hat{P}\|P)$$

is "large". Unfortunately the numerical bounds obtained in Theorem 2.3 appear too crude for most applications, and the rejection criterion there, namely $D(\hat{P}\|P^*) \geq |\Omega|\frac{\log n}{n}$, admits false acceptance too often. A better criterion is suggested by the result in Theorem 4.2 (see also Remark 4.3) that $\frac{2n}{\log e}D(\hat{P}\|P^*)$ has χ^2 limit distribution, with specified degrees of freedom, if the null hypothesis is true. Using this theorem, the null-hypothesis is rejected if $(2n/\log e)D(\hat{P}\|P^*)$ exceeds the threshold found in the table of the χ^2 distribution for the selected level of significance. Of course, the type 1 error probability of the resulting test will be close to the desired one only when the sample size n is sufficiently large for the distribution of the test statistic to be close to its χ^2 limit. The question of how large n is needed is important but difficult, and will not be entered here.

Now we look at the problem of *outliers*. A lack of fit (i. e., $D(\hat{P}\|P^*)$ "large") may be due not to the inadequacy of the model tested, but to outliers. A cell ω_0 is considered to be an outlier in the following case: Let $\mathcal{L}$ be the linear family determined by the γ-marginals (say $\gamma \in \Gamma$) of the empirical distribution $\hat{P}$, and let $\mathcal{L}'$ be the subfamily of $\mathcal{L}$ consisting of those $P \in \mathcal{L}$ that satisfy $P(\omega_0) = \hat{P}(\omega_0)$. Let P^{**} be the I-projection of P^* onto $\mathcal{L}'$. Ideally, we should consider ω_0 as an outlier if $D(P^{**}\|P^*)$ is "large", for if $D(P^{**}\|P^*)$ is close to $D(\hat{P}\|P^*)$ then $D(\hat{P}\|P^{**})$ will be small by the Pythagorean identity. Now by the lumping property (Lemma 4.1):

$$D(P^{**}\|P^*) \geq \hat{P}(\omega_0)\log\frac{\hat{P}(\omega_0)}{P^*(\omega_0)} + \Big(1 - \hat{P}(\omega_0\Big)\log\frac{\hat{P}(\omega_0)}{P^*(\omega_0)},$$

and we declare ω_0 as an outlier if the right-hand side of this inequality is "large", that is, after scaling by $(2n/\log e)$, it exceeds the critical value of χ^2 with one degree of freedom.

If the above method produces only a few outliers, say $\omega_0, \omega_1, \ldots, \omega_\ell$, we consider the subset $\tilde{\mathcal{L}}$ of $\mathcal{L}$ consisting of those $P \in \mathcal{L}$ that satisfy $P(\omega_j) = \hat{P}(\omega_j)$ for $j = 0, \ldots, \ell$. If the I-projection of P^* onto $\tilde{\mathcal{L}}$ is already "close" to $\hat{P}$, we accept the model and attribute the original lack of fit to the outliers. Then the "outlier" cell counts $x(\omega_j)$ are deemed unreliable and they may be adjusted to $nP^*(\omega_j)$.

Similar techniques are applicable in the case when some cell counts are missing.

5

Iterative algorithms

In this Section we discuss iterative algorithms to compute I-projections and to minimize I-divergence between two convex sets of distributions, as well as to estimate a distribution from incomplete data.

5.1 Iterative scaling

The I-projection to a linear family $\mathcal{L}$ is very easy to find if $\mathcal{L}$ is determined by a partition $\mathcal{B} = (B_1, \dots, B_k)$ of A and consists of all those distributions P whose $\mathcal{B}$-lumping is a given distribution $(\alpha_1, \dots, \alpha_k)$ on $\{1, \dots, k\}$. Indeed, then $D(P\|Q) \geq D(P^{\mathcal{B}}\|Q^{\mathcal{B}}) = \sum \alpha_i \log \alpha_i / Q(B_i)$ for each $P \in \mathcal{L}$, by the lumping property (see Lemma 4.1), and here the equality holds for P^* defined by

$$P^*(a) = c_i Q(a),\ a \in B_i,\ \text{where } c_i = \frac{\alpha_i}{Q(B_i)}. \tag{5.1}$$

It follows that P^* obtained by “scaling” Q as above is the I-projection of Q to $\mathcal{L}$.

In the theory of contingency tables, see Section 4, lumpings occur most frequently as marginals. Accordingly, when $\mathcal{L}$ is defined by prescribing some γ-marginal of P, say $\mathcal{L}_\gamma = \{P\colon P_\gamma = \overline{P}_\gamma\}$, where $\gamma \subset \{1,\dots,d\}$, the I-projection P^* of Q to $\mathcal{L}_\gamma$ is obtained by scaling Q to adjust its γ-marginal: $P^*(\omega) = Q(\omega)\overline{P}_\gamma(\omega(\gamma))/Q_\gamma(\omega(\gamma))$. Suppose next that $\mathcal{L}$ can be represented as the intersection of families $\mathcal{L}_i, i = 1,\dots,m$, each of form as above. Then, on account of Theorem 5.1, below, and the previous paragraph, I-projections to $\mathcal{L}$ can be computed by iterative scaling. This applies, in particular, to I-projections to families defined by prescribed marginals, required in the analysis of contingency tables: For $\mathcal{L} = \{P_\gamma = \overline{P}_\gamma, \gamma \in \Gamma\}, \Gamma = \{\gamma_1,\dots,\gamma_m\}$, the I-projection of Q to $\mathcal{L}$ equals the limit of the sequence of distributions $P^{(n)}$ defined by iterative scaling, that is, $P^{(0)} = Q$, and $P^{(n)}(\omega) = P^{(n-1)}(\omega)\overline{P}_{\gamma_n}(\omega(\gamma_n))/P^{(n-1)}_{\gamma_n}(\omega(\gamma_n))$, where $\gamma_1,\gamma_2,\dots$ is a cyclic repetition of Γ.

Suppose $\mathcal{L}_1,\dots,\mathcal{L}_m$, are given linear families and generate a sequence of distributions P_n as follows: Set $P_0 = Q$ (any given distribution with support $S(Q) = A$), let P_1 be the I-projection of P_0 onto $\mathcal{L}_1$, P_2 the I-projection of P_1 onto $\mathcal{L}_2$, and so on, where for $n > m$ we mean by $\mathcal{L}_n$ that $\mathcal{L}_i$ for which $i \equiv n \pmod m$; i. e., $\mathcal{L}_1,\dots,\mathcal{L}_m$ is repeated cyclically.

Theorem 5.1. If $\cap_{i=1}^m \mathcal{L}_i = \mathcal{L} \neq \emptyset$ then $P_n \to P^*$, the I-projection of Q onto $\mathcal{L}$.

Proof. By the Pythagorean identity, see Theorem 3.2, we have for every $P \in \mathcal{L}$ (even for $P \in \mathcal{L}_n$) that

$$D(P\|P_{n-1}) = D(P\|P_n) + D(P_n\|P_{n-1}), n = 1,2,\dots$$

Adding these equations for $1 \le n \le N$ we get that

$$D(P\|Q) = D(P\|P_0) = D(P\|P_N) + \sum_{n=1}^{N} D(P_n\|P_{n-1}).$$

By compactness there exists a subsequence $P_{N_k} \to P'$, say, and then from the preceding inequality we get for $N_k \to \infty$ that

$$D(P\|Q) = D(P\|P') + \sum_{n=1}^{\infty} D(P_n\|P_{n-1}). \tag{5.2}$$

Since the series in (5.2) is convergent, its terms go to 0, hence also the variational distance $|P_n - P_{n-1}| = \sum_a |P_n(a) - P_{n-1}(a)|$ goes to 0 as $n \to \infty$. This implies that together with $P_{N_k} \to P'$ we also have

$$P_{N_k+1} \to P', P_{N_k+2} \to P', \dots, P_{N_k+m} \to P'.$$

Since by the periodic construction, among the m consecutive elements,

$$P_{N_k}, P_{N_k+1}, \dots, P_{N_k+m-1}$$

there is one in each $\mathcal{L}_i, i = 1, 2, \dots, m$, it follows that $P' \in \cap \mathcal{L}_i = \mathcal{L}$.

Since $P' \in \mathcal{L}$ it may be substituted for P in (5.2) to yield

$$D(P' \| Q) = \sum_{i=1}^{\infty} D(P_n \| P_{n-1}).$$

With this, in turn, (5.2) becomes

$$D(P\|Q) = D(P\|P') + D(P'\|Q),$$

which proves that P' equals the I-projection P^* of Q onto $\mathcal{L}$. Finally, as P' was the limit of an arbitrary convergent subsequence of the sequence P_n, our result means that every convergent subsequence of P_n has the same limit P^*. Using compactness again, this proves that $P_n \to P^*$ and completes the proof of the theorem. □

In the general case when $\mathcal{L} = \cap_{i=1}^m \mathcal{L}_i$ but no explicit formulas are available for I-projections to the families $\mathcal{L}_i$, Theorem 5.1 need not directly provide a practical algorithm for computing the I-projection to $\mathcal{L}$. Still, with a twist, Theorem 5.1 does lead to an iterative algorithm, known as *generalized iterative scaling* (or the SMART algorithm) to compute I-projections to general linear families and, in particular, MLE's for exponential families, see Theorem 3.3.

Generalized iterative scaling requires that the linear family

$$\mathcal{L} = \left\{ P \colon \sum_{a \in A} P(a) f(a) = \alpha_i, \quad 1 \le i \le k \right\}$$

be given in terms of functions f_i that satisfy

$$f_i(a) \ge 0, \quad \sum_{i=1}^{k} f_i(a) = 1, \quad a \in A\ ; \tag{5.3}$$

accordingly, $(\alpha_1, \dots, \alpha_k)$ has to be a probability vector. This does not restrict generality, for if $\mathcal{L}$ is initially represented in terms of any functions $\widetilde{f}_i$, these can be replaced by $f_i = C\widetilde{f}_i + D$ with suitable constants C and D to make sure that $f_i \ge 0$ and $\sum_{i=1}^k f_i(a) \le 1$; if the last inequality is strict for some $a \in A$, one can replace k by $k+1$, and introduce an additional function $f_{k+1} = 1 - \sum_{i=1}^k f_i$.

Theorem 5.2. Assuming (5.3), the nonnegative functions b_n on A defined recursively by

$$b_0(a) = Q(a), \quad b_{n+1}(a) = b_n(a) \prod_{i=1}^k \left(\frac{\alpha_i}{\beta_{n,i}}\right)^{f_i(a)}, \quad \beta_{n,i} = \sum_{a \in A} b_n(a) f_i(a)$$

converge to the I-projection P^* of Q to $\mathcal{L}$, that is, $P^*(a) = \lim_{n\to\infty} b_n(a),\ a \in A$.

Intuitively, in generalized iterative scaling the values $b_n(a)$ are updated using all constraints in each step, via multiplications by weighted geometric means of the analogues $\alpha_i/\beta_{n,i}$ of the ratios in (5.1) that have been used in standard iterative scaling, taking one constraint into account in each step. Note that the functions b_n need not be probability distributions, although their limit is.

Proof of Theorem 5.2. Consider the product alphabet $\widetilde{A} = A \times \{1, \dots, k\}$, the distribution $\widetilde{Q} = \{Q(a)f_i(a),\ (a,i) \in \widetilde{A}\}$, and the linear family $\widetilde{\mathcal{L}}$ of those distributions $\widetilde{P}$ on $\widetilde{A}$ that satisfy $\widetilde{P}(a,i) = P(a)f_i(a)$ for some $P \in \mathcal{L}$. Since for such $\widetilde{P}$ we have $D(\widetilde{P}\|\widetilde{Q}) = D(P\|Q)$, the I-projection of $\widetilde{Q}$ to $\widetilde{\mathcal{L}}$ equals $\widetilde{P}^* = \{P^*(a)f_i(a)\}$ where P^* is the I-projection of Q to $\mathcal{L}$.

Note that $\widetilde{\mathcal{L}} = \widetilde{\mathcal{L}}_1 \cap \widetilde{\mathcal{L}}_2$ where $\widetilde{\mathcal{L}}_1$ is the set of all distributions $\widetilde{P} = \{\widetilde{P}(a,i)\}$ whose marginal on $\{1, \dots, k\}$ is equal to $(\alpha_1, \dots, \alpha_k)$, and

$$\widetilde{\mathcal{L}_2} = \{\widetilde{P}\colon \widetilde{P}(a,i) = P(a)f_i(a), \quad P \text{ any distribution on } A\}.$$

It follows by Theorem 5.1 that the sequence of distributions $\widetilde{P_0}, \widetilde{P_0}', \widetilde{P_1}, \widetilde{P_1}', \dots$ on $\widetilde{A}$ defined iteratively, with $\widetilde{P_0} = \widetilde{Q}$, by

$$\widetilde{P_n'} = I\text{-projection to } \widetilde{\mathcal{L}}_1 \text{ of } \widetilde{P_n}, \qquad \widetilde{P}_{n+1} = I\text{-projection to } \widetilde{\mathcal{L}}_2 \text{ of } \widetilde{P_n}'$$

converges to $\widetilde{P^*}$. In particular, writing $\widetilde{P_n}(a,i) = P_n(a)f_i(a)$, we have $P_n \to P^*$. The theorem will be proved if we show that $P_n(a) = c_n b_n(a)$, where $c_n \to 1$ as $n \to \infty$.

Now, by the first paragraph of this subsection, $\widetilde{P_n'}$ is obtained from $\widetilde{P_n}$ by scaling, thus

$$\widetilde{P_n'}(a,i) = \frac{\alpha_i}{\gamma_{n,i}} P_n(a) f_i(a), \quad \gamma_{n,i} = \sum_{a\in A} P_n(a) f_i(a) \ .$$

To find $\widetilde{P}_{n+1}$, note that for each $\widetilde{P} = \{P(a)f_i(a)\}$ in $\widetilde{\mathcal{L}}_2$ we have, using (5.3),

$$\begin{aligned} D(\widetilde{P}\|\widetilde{P_n'}) &= \sum_{a\in A}\sum_{i=1}^{k} P(a)f_i(a)\log\left(\frac{P(a)}{P_n(a)}\Big/\frac{\alpha_i}{\gamma_{n,i}}\right) \\ &= \sum_{a\in A} P(a)\log\frac{P(a)}{P_n(a)} - \sum_{a\in A} P(a)\sum_{i=1}^{k} f_i(a)\log\frac{\alpha_i}{\gamma_{n,i}} \\ &= \sum_{a\in A} P(a)\log\frac{P(a)}{P_n(a)\prod\limits_{i=1}^{k}\left(\frac{\alpha_i}{\gamma_{n,i}}\right)^{f_i(a)}} \ . \end{aligned}$$

This implies, by the log-sum inequality, that the minimum of $D(\widetilde{P}\|\widetilde{P}_n')$ subject to $\widetilde{P} \in \widetilde{\mathcal{L}_2}$ is attained by $\widetilde{P}_{n+1} = \{P_{n+1}(a)f_i(a)\}$ with

$$P_{n+1}(a) = c_{n+1}P_n(a)\prod_{i=1}^{k}\left(\frac{\alpha_i}{\gamma_{n,i}}\right)^{f_i(a)}$$

where c_{n+1} is a normalizing constant. Comparing this with the recursion defining b_n in the statement of the theorem, it follows by induction that $P_n(a) = c_n b_n(a), n = 1, 2, \ldots$.

Finally, $c_n \to 1$ follows since the above formula for $D(\widetilde{P}\|\widetilde{P}_n')$ gives $D(\widetilde{P}_{n+1}\|\widetilde{P}_n') = \log c_{n+1}$, and $D(\widetilde{P}_{n+1}\|\widetilde{P}_n') \to 0$ as in the proof of Theorem 5.1. □

5.2 Alternating divergence minimization

In this subsection we consider a very general alternating minimization algorithm which, in particular, will find the minimum divergence between two convex sets $\mathcal{P}$ and $\mathcal{Q}$ of distributions on a finite set A.

In the general considerations below, $\mathcal{P}$ and $\mathcal{Q}$ are arbitrary sets and $D(P,Q)$ denotes an extended real-valued function on $\mathcal{P} \times \mathcal{Q}$ which satisfies the following conditions.

(a) $-\infty < D(P,Q) \leq +\infty,\ P \in \mathcal{P}, Q \in \mathcal{Q}$.
(b) $\forall P \in \mathcal{P}, \exists Q^* = Q^*(P) \in \mathcal{Q}$ such that $\min\limits_{Q \in \mathcal{Q}} D(P,Q) = D(P,Q^*)$.
(c) $\forall Q \in \mathcal{Q}, \exists P^* = P^*(Q) \in \mathcal{P}$ such that $\min\limits_{P \in \mathcal{P}} D(P,Q) = D(P^*,Q)$.

A problem of interest in many situations is to determine

$$D_{\min} \overset{\text{def}}{=} \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} D(P,Q). \tag{5.4}$$

A naive attempt to solve this problem would be to start with some $Q_0 \in \mathcal{Q}$ and recursively define

$$P_n = P^*(Q_{n-1}),\ Q_n = Q^*(P_n),\ n = 1,2,\ldots \tag{5.5}$$

hoping that $D(P_n, Q_n) \to D_{\min}$, as $n \to \infty$.

We show that, subject to some technical conditions, the naive iteration scheme (5.5) determines the infimum in (5.4). This is stated as the following theorem.

Theorem 5.3. Suppose there is a nonnegative function $\delta(P,P')$ defined on $\mathcal{P} \times \mathcal{P}$ with the following properties:

(i) "three-points property,"

$$\delta(P, P^*(Q)) + D(P^*(Q), Q) \leq D(P,Q), \quad \forall P \in \mathcal{P}, Q \in \mathcal{Q},$$

(ii) "four-points property," for $P \in \mathcal{P}$ with $\min\limits_{Q \in \mathcal{Q}} D(P\|Q) < \infty$,

$$D(P',Q') + \delta(P',P) \geq D(P', Q^*(P)), \quad \forall P' \in \mathcal{P},\ Q' \in \mathcal{Q}.$$

(iii) $\delta(P^*(Q), P_1) < \infty$ for $Q \in \mathcal{Q}$ with $\min\limits_{P \in \mathcal{P}} D(P,Q) < \infty$.

Then, if $\min\limits_{P \in \mathcal{P}} D(P, Q_0) < \infty$, the iteration (5.5) produces (P_n, Q_n) such that

$$\lim_{n \to \infty} D(P_n, Q_n) = \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} D(P,Q) = D_{\min}. \tag{5.6}$$

Under the additional hypotheses that (iv) $\mathcal{P}$ is compact, (v) $D(P, Q^*(P))$ is a lower semi-continuous function of P, and (vi) $\delta(P, P_n) \to 0$ iff $P_n \to P$, we also have $P_n \to P_\infty$, where $D(P_\infty, Q^*(P_\infty)) = D_{\min}$; moreover, $\delta(P_\infty, P_n) \downarrow 0$ and

$$D(P_{n+1}, Q_n) - D_{\min} \leq \delta(P_\infty, P_n) - \delta(P_\infty, P_{n+1}). \tag{5.7}$$

Proof. We have, by the three-points property,

$$\delta(P, P_{n+1}) + D(P_{n+1}, Q_n) \leq D(P, Q_n),$$

and, by the four-points property

$$D(P, Q_n) \leq D(P, Q) + \delta(P, P_n),$$

for all $P \in \mathcal{P}, Q \in \mathcal{Q}$. Hence

$$\delta(P, P_{n+1}) \leq D(P, Q) - D(P_{n+1}, Q_n) + \delta(P, P_n) \tag{5.8}$$

We claim that the iteration (5.5) implies the basic limit result (5.6). Indeed, since

$$D(P_1, Q_0) \geq D(P_1, Q_1) \geq D(P_2, Q_1) \geq D(P_2, Q_2) \geq \dots$$

by definition, if (5.6) were false there would exist Q and $\epsilon > 0$ such that $D(P_{n+1}, Q_n) > D(P^*(Q), Q) + \epsilon, n = 1, 2, \dots$. Then the inequality (5.8) applied with this Q and $P^*(Q)$ would give $\delta(P^*(Q), P_{n+1}) \leq \delta(P^*(Q), P_n) - \epsilon$, for $n = 1, 2, \dots$, contradicting assumption (iii) and the nonnegativity of δ.

Supposing also the assumptions (iv)-(vi), pick a convergent subsequence of $\{P_n\}$, say $P_{n_k} \to P_\infty \in \mathcal{P}$. Then by (v) and (5.6),

$$D(P_\infty, Q^*(P_\infty)) \leq \liminf_{k \to \infty} D(P_{n_k}, Q_{n_k}) = D_{\min},$$

and by the definition of $D_{\min}$, here the equality must hold. By (5.8) applied to $D(P, Q) = D(P_\infty, Q^*(P_\infty)) = D_{\min}$, it follows that

$$\delta(P_\infty, P_{n+1}) \leq D_{\min} - D(P_{n+1}, Q_n) + \delta(P_\infty, P_n),$$

proving (5.7). This last inequality also shows that $\delta(P_\infty, P_{n+1}) \leq \delta(P_\infty, P_n)$, $n = 1, 2, \dots$, and, since $\delta(P_\infty, P_{n_k}) \to 0$, by (vi), this proves that $\delta(P_\infty, P_n) \downarrow 0$. Finally, again by (vi), the latter implies that $P_n \to P_\infty$. □

Next we want to apply the theorem to the case when $\mathcal{P}$ and $\mathcal{Q}$ are convex, compact sets of measures on a finite set A (in the remainder of this subsection by a *measure* we mean a nonnegative, finite-valued measure, equivalently, a nonnegative, real-valued function on A), and $D(P,Q) = D(P\|Q) = \sum_a P(a)\log(P(a)/Q(a))$, a definition that makes sense even if the measures do not sum to 1. The existence of minimizers $Q^*(P)$ and $P^*(Q)$ of $D(P\|Q)$ with P or Q fixed is obvious.

We show now that with

$$\delta(P,P') = \delta(P\|P') \overset{\text{def}}{=} \sum_{a\in A}\left[P(a)\log\frac{P(a)}{P'(a)} - (P(a) - P'(a))\log e\right],$$

which is nonnegative term-by-term, all assumptions of Theorem 5.3 are satisfied, with the possible exception of assumption (iii) to which we will return later.

Indeed, the three-points and four-points properties have already been established in the case when the measures in question are probability distributions, see Theorems 3.1 and 3.4. The proofs of these two theorems easily extend to the present more general case.

Of assumptions (iv)–(vi), only (v) needs checking, that is, we want to show that if $P_n \to P$ then $\min_{Q\in\mathcal{Q}} D(P\|Q) \le \liminf_{n\to\infty} D(P_n\|Q_n)$, where $Q_n = Q^*(P_n)$. To verify this, choose a subsequence such that $D(P_{n_k}\|Q_{n_k}) \to \liminf_{n\to\infty} D(P_n\|Q_n)$ and Q_{n_k} converges to some $Q^* \in \mathcal{Q}$. The latter and $P_{n_k} \to P$ imply that $D(P\|Q^*) \le \lim_{k\to\infty} D(P_{n_k}\|Q_{n_k})$, and the assertion follows.

Returning to the question whether assumption (iii) of Theorem 5.3 holds in our case, note that $\delta(P^*(Q)\|P_1) = \delta(P^*(Q)\|P^*(Q_0))$ is finite if the divergence $D(P^*(Q)\|P^*(Q_0))$ is finite on account of the three-points property (i). Now, for each $Q \in \mathcal{Q}$ with $\inf_{P\in\mathcal{P}} D(P\|Q) < \infty$ whose support is contained in the support of Q_0, the inclusions $S(P^*(Q)) \subseteq S(Q) \subseteq S(Q_0)$ imply that $D(P^*(Q)\|P^*(Q_0)$ is finite. This means that assumption is always satisfied if Q_0 has maximal support, that is, $S(Q_0) = S(\mathcal{Q})$. Thus we have arrived at

Corollary 5.1. Suppose $\mathcal{P}$ and $\mathcal{Q}$ are convex compact sets of measures on a finite set A such that there exists $P \in \mathcal{P}$ with $S(P) \subseteq S(\mathcal{Q})$, and let $D(P,Q) = D(P\|Q), \delta(P,Q) = \delta(P\|Q)$. Then all assertions of Theorem 5.3 are valid, provided the iteration (5.5) starts with a $Q_0 \in \mathcal{Q}$ of maximal support.

Note that under the conditions of the corollary, there exists a unique minimizer of $D(P\|Q)$ subject to $P \in \mathcal{P}$, unless $D(P\|Q) = +\infty$ for every $P \in \mathcal{P}$. There is a unique minimizer of $D(P\|Q)$ subject to $Q \in \mathcal{Q}$ if $S(P) = S(\mathcal{Q})$, but not necessarily if $S(P)$ is a proper subset of $S(\mathcal{Q})$; in particular, the sequences P_n, Q_n defined by the iteration (5.5) need not be uniquely determined by the initial $Q_0 \in \mathcal{Q}$. Still, $D(P_n\|Q_n) \to D_{\min}$ always holds, P_n always converges to some $P_\infty \in \mathcal{P}$ with $\min_{Q\in\mathcal{Q}} D(P_\infty\|Q) = D_{\min}$, and each accumulation point of $\{Q_n\}$ attains that minimum (the latter can be shown as assumption (v) of Theorem 5.3 was verified above). If $D(P_\infty, Q)$ is minimized for a unique $Q_\infty \in \mathcal{Q}$, then $Q_n \to Q_\infty$ can also be concluded.

The following consequence of (5.7) is also worth noting, for it provides a stopping criterion for the iteration (5.5).

$$
\begin{aligned}
D(P_{n+1}\|Q_n) - D_{\min} \le \delta(P_\infty\|P_n) - \delta(P_\infty\|P_{n+1}) = \\
&= \sum_{a\in A} P_\infty(a) \log \frac{P_{n+1}(a)}{P_n(a)} + \sum_{a\in A} [P_n(a) - P_{n+1}(a)] \log e \\
&\le \left(\max_{P\in\mathcal{P}} P(A)\right) \max_{a\in A} \log \frac{P_{n+1}(a)}{P_n(a)} + [P_n(A) - P_{n+1}(A)] \log e
\end{aligned}
$$

where $P(A) \overset{\text{def}}{=} \sum_{a\in A} P(a)$; using this, the iteration can be stopped when the last bound becomes smaller than a prescribed $\epsilon > 0$. The criterion becomes particularly simple if $\mathcal{P}$ consists of probability distributions.

Corollary 5.1 can be applied, as we show below, to minimizing I-divergence when either the first or second variable is fixed and the other variable ranges over the image of a "nice" set of measures on a larger alphabet. Here "nice" sets of measures are those for which the divergence minimization is "easy."

For a mapping $T: A \mapsto B$ and measures P on A, write P^T for the image of P on B, that is, $P^T(b) = \sum_{a:Ta=b} P(a)$. For a set $\mathcal{P}$ of measures on A write $\mathcal{P}^T = \{P^T : P \in \mathcal{P}\}$.

Problem 1. Given a measure $\overline{P}$ on B and a convex set $\mathcal{Q}$ of measures on A, minimize $D(\overline{P}\|\overline{Q})$ subject to $\overline{Q} \in \mathcal{Q}^T$.

Problem 2. Given a measure $\overline{Q}$ on B and a convex set $\mathcal{P}$ of measures on A, minimize $D(\overline{P}\|\overline{Q})$ subject to $\overline{P} \in \mathcal{P}^T$.

Lemma 5.1. The minimum in Problem 1 equals $D_{\min} = \min_{P\in\mathcal{P}, Q\in\mathcal{Q}}$ for $\mathcal{P} = \{P : P^T = \overline{P}\}$ and the given $\mathcal{Q}$, and if (P^*, Q^*) attains $D_{\min}$ then Q^{*T} attains the minimum in Problem 1.

A similar result holds for Problem 2, with the roles of P and Q interchanged.

Proof. The lumping property of Lemma 4.1, which also holds for arbitrary measures, gives

$$D(P^T\|Q^T) \le D(P\|Q), \text{ with equality if } \frac{P(a)}{Q(a)} = \frac{P^T(b)}{Q^T(b)},\ b = Ta.$$

From this it follows that if $\mathcal{P} = \{P : P^T = \overline{P}\}$ for a given $\overline{P}$, then the minimum of $D(P\|Q)$ subject to $P \in \mathcal{P}$ (for Q fixed) is attained for $P^* = P^*(Q)$ with

$$P^*(a) = \frac{Q(a)}{Q^T(b)}\overline{P}(b),\ b = Ta \tag{5.9}$$

and this minimum equals $D(\overline{P}\|Q^T)$. A similar result holds also for minimizing $D(P\|Q)$ subject to $Q \in \mathcal{Q}$ (for P fixed) in the case when $\mathcal{Q} = \{Q : Q^T = \overline{Q}\}$ for a given $\overline{Q}$, in which case the minimizer $Q^* = Q^*(P)$ is given by

$$Q^*(a) = \frac{P(a)}{P^T(b)}\overline{Q}(b),\ b = Ta \tag{5.10}$$

The assertion of the lemma follows. □

Example 5.1 (Decomposition of mixtures.). Let $\overline{P}$ be a probability distribution and let $\mu_1, \ldots, \mu_k$ be arbitrary measures on a finite set B. The goal is to minimize $D(\overline{P}\|\sum_k c_i\mu_i)$ for weight vectors

$(c_1,\dots,c_k)$ with nonnegative components that sum to 1. If $\mu_1,\dots,\mu_k$ are probability distributions and $\overline{P}$ is the empirical distribution of a sample drawn from the mixture $\sum_i c_i\mu_i$ then the goal is identical to finding the MLE of the weight vector $(c_1,\dots,c_k)$.

This example fits into the framework of Problem 1, above, by setting $A = \{1,\dots,k\}\times B$, $T(i,b) = b$, $\mathcal{Q} = \{Q{:}\, Q(i,b) = c_i\mu_i(b)\}$. Thus we consider the iteration (5.5) as in Corollary 5.1, with $\mathcal{P} = \{P{:}\, \sum_i P(i,b) = \overline{P}(b), b\in B\}$ and $\mathcal{Q}$ above, assuming for nontriviality that $S(\overline{P}) \subseteq \cup_i S(\mu_i)$ (equivalent to the support condition in Corollary 5.1 in our case). As Corollary 5.1 requires starting with $Q_0\in\mathcal{Q}$ of maximal support, we assume $Q_0(i,b) = c_i^0\mu_i(b)$, $c_i^0 > 0$, $i = 1,\dots,k$. To give the iteration explicitly, note that if $Q_{n-1}(i,b) = c_i^{n-1}\mu_i(b)$ is already defined then P_n is obtained, according to (5.9), as

$$P_n(i,b) = \frac{Q_{n-1}(i,b)}{Q_{n-1}^T(b)}\overline{P}(b) = \frac{c_i^{n-1}\mu_i(b)}{\sum_j c_j^{n-1}\mu_j(b)}\overline{P}(b).$$

To find $Q_n\in\mathcal{Q}$ minimizing $D(P_n\|Q)$, put $P_n(i) = \sum_{b\in B} P_n(i,b)$ and use $Q(i,b) = c_i\mu_i(b)$ to write

$$\begin{aligned} D(P_n\|Q) &= \sum_{i=1}^k\sum_{b\in B} P_n(i,b)\log\frac{P_n(i,b)}{c_i\mu_i(b)} \\ &= \sum_{i=1}^k P_n(i)\log\frac{P_n(i)}{c_i} + \sum_{i=1}^k\sum_{b\in B} P_n(i,b)\log\frac{P_n(i,b)}{P_n(i)\mu_i(b)}. \end{aligned}$$

This is minimized for $c_i^n = P_n(i)$, hence the recursion for c_i^n will be

$$c_i^n = c_i^{n-1}\sum_{b\in B}\frac{\mu_i(b)\overline{P}(b)}{\sum_j c_j^{n-1}\mu_j(b)}.$$

Finally, we show that $(c_1^n,\dots,c_k^n)$ converges to a minimizer $(c_1^*,\dots,c_k^*)$ of $D(\overline{P}\|\sum_k c_i\mu_i)$. Indeed, P_n converges to a limit P_∞ by Corollary 5.1, hence $c_i^n = P_n(i)$ also has a limit c_i^* and $Q_n\to Q^*$ with $Q^*(i,b) = c_i^*\mu_i(b)$. By the passage following Corollary 5.1, (P_∞, Q^*) attains $D_{\min} = \min_{P\in\mathcal{P},Q\in\mathcal{Q}} D(P\|Q)$, and then, by Lemma 5.1, $Q^{*T} = \sum_i c_i^*\mu_i$ attains $\min_{\overline{Q}\in\mathcal{Q}^T} D(\overline{P}\|\overline{Q}) = D_{\min}$.

Remark 5.1. A problem covered by Example 5.1 is that of finding weights $c_i > 0$ of sum 1 that maximize the expectation of $\log\sum_i c_iX_i$, where $X_1, \dots, X_k$ are given nonnegative random variables defined on a finite probability space $(B, \overline{P})$. Indeed, then

$$E(\log \sum_i c_i X_i) = -D(\overline{P} \| \sum_i c_i \mu_i),$$

for $\mu_i(b) = \overline{P}(b) X_i(b)$. In this case, the above iteration takes the form

$$c_i^n = c_i^{n-1} E(\frac{X_i}{\sum_j c_j^n X_j}),$$

which is known as Cover's portfolio optimization algorithm. We note without proof that the algorithm works also for nondiscrete $X_1, \dots, X_k$.

Remark 5.2. The counterpart of the problem of Example 5.1, namely, the minimization of $D(\sum_k c_i \mu_i \| \overline{Q})$ can be solved similarly. Then the iteration of Corollary 5.1 has to be applied to the set $\mathcal{P}$ consisting of the measures of the form $P(i, b) = c_i \mu_i(b)$ and to $\mathcal{Q} = \{Q \colon \sum_i Q(i, b) = \overline{Q}(b), b \in B\}$. Actually, the resulting iteration is the same as that in the proof of Theorem 5.2 (assuming the μ_i and $\overline{Q}$ are probability distributions), with notational difference that the present $i, b, c_i, \mu_i(b), \overline{Q}(b), P_n \in \mathcal{P}, Q_n \in \mathcal{Q}$ correspond to $a, i, P(a), f_i(a), \alpha_i, \widetilde{P}_n \in \widetilde{\mathcal{L}}_2, \widetilde{P}'_n \in \widetilde{\mathcal{L}}_1$ there. To see this, note that while the even steps of the two iterations are conceptually different divergence minimizations (with respect to the second, respectively, first variable, over the set denoted by $\mathcal{Q}$ or $\widetilde{\mathcal{L}}_1$), in fact both minimizations require the same scaling, see (5.9), (5.10).

This observation gives additional insight into generalized iterative scaling, discussed in the previous subsection. Note that Theorem 5.2 involves the assumption $\mathcal{L} \neq \emptyset$ (as linear families have been defined to be non-empty, see Section 3), and that assumption is obviously necessary. Still, the sequence $\{P_n\}$ in the proof of Theorem 5.2 is well defined also if $\mathcal{L} = \emptyset$, when $\widetilde{\mathcal{L}}_1$ and $\widetilde{\mathcal{L}}_2$ in that proof are disjoint. Now, the above observation and Corollary 5.1 imply that P_n converges to a limit P^* also in that case, moreover, $P = P^*$ minimizes the I-divergence from $(\alpha_1, \dots, \alpha_k)$ of distributions $(\gamma_1, \dots, \gamma_k)$ such that $\gamma_i = \sum_a P(a) f_i(a), 1 \leq i \leq k$, for some probability distribution P on A.

5.3 The EM algorithm

The expectation–maximization or EM algorithm is an iterative method frequently used in statistics to estimate a distribution supposed to belong to a given set $\mathcal{Q}$ of distributions on a set A when, instead of a "full" sample $x_1^n = x_1 \dots x_n \in A^n$ from the unknown distribution, only an "incomplete" sample $y_1^n = y_1 \dots y_n \in B^n$ is observable. Here $y_i = Tx_i$ for a known mapping $T : A \mapsto B$. As elsewhere in this tutorial, we restrict attention to the case of finite A, B.

The EM algorithm produces a sequence of distributions $Q_k \in \mathcal{Q}$, regarded as consecutively improved estimates of the unknown distribution, iterating the following steps E and M, starting with an arbitrary $Q_0 \in \mathcal{Q}$.

Step E: Calculate the conditional expectation $P_k = E_{Q_{k-1}}(\hat{P}_n | y_1^n)$ of the empirical distribution $\hat{P}_n$ of the unobservable full sample, conditioned on the observed incomplete sample, pretending the true distribution equals the previous estimate Q_{k-1}.

Step M: Calculate the MLE of the distribution the full sample x_1^n is coming from, pretending that the empirical distribution of x_1^n equals P_k calculated in Step E. Set Q_k equal to this MLE.

Here, motivated by Lemma 3.1, by "MLE pretending the empirical distribution equals P_k" we mean the minimizer of $D(P_k||Q)$ subject to $Q \in \mathcal{Q}$, even if P_k is not a possible empirical distribution (implicitly assuming that a minimizer exists; if there are several, any one of them may be taken). For practicality, we assume that step M is easy to perform; as shown below, step E is always easy.

The EM algorithm is, in effect, an alternating divergence minimization, see the previous subsection. To verify this, it suffices to show that P_k in Step E minimizes the divergence $D(P||Q_{k-1})$ subject to $P \in \mathcal{P}$, for $\mathcal{P} = \{P : P^T = \hat{P}_n^T\}$, where P^T denotes the image of P under the mapping $T: A \mapsto B$. Actually, we claim that for any distribution Q on A, the conditional expectation $P = E_Q(\hat{P}_n | y_1^n)$ attains the minimum of $D(P||Q)$ subject to $P^T = \hat{P}_n^T$.

Now, writing $\delta(x, a) = 1$ if $x = a$ and 0 otherwise, we have for any

$a \in A$

$$\begin{aligned} E_Q(\hat{P}_n(a)|y_1^n) &= E_Q\left(\frac{1}{n}\sum_{i=1}^{n}\delta(x_i,a)\Big|y_1^n\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E_Q(\delta(x_i,a)|y_i) = \frac{|\{i : y_i = Ta\}|}{n}\frac{Q(a)}{Q^T(Ta)}. \end{aligned}$$

Indeed, under the condition that $y_i = Tx_i$ is given, the conditional expectation of $\delta(x_i, a)$, that is, the conditional probability of $x_i = a$, equals $Q(a)/Q^T(Ta)$ if $y_i = Ta$, and zero otherwise. As the empirical distribution of y_1^n is equal to $\hat{P}_n^T$, this means that $P = E_Q(\hat{P}_n|y_1^n)$ is given by

$$P(a) = \hat{P}_n^T(Ta)\frac{Q(a)}{Q^T(Ta)}.$$

Since $D(P||Q) \geq D(P^T||Q^T)$ for any P (by the lumping property, see Lemma 4.1), and for P given above here clearly the equality holds, it follows that $P = E_Q(\hat{P}_n|y_1^n)$ minimizes $D(P||Q)$ subject to $P^T = \hat{P}_n^T$, as claimed.

An immediate consequence is that

$$D(P_1||Q_0) \geq D(P_1||Q_1) \geq D(P_2||Q_1) \geq D(P_2||Q_2) \geq \ldots$$

In particular, as $D(P_k||Q_{k-1}) = D(\hat{P}_n^T||Q_{k-1}^T)$, the sequence $D(\hat{P}_n^T||Q_k^T)$ is always non-increasing and hence converges to a limit as $k \to \infty$.

In the ideal case, this limit equals

$$\min_{Q\in\mathcal{Q}} D(\hat{P}_n^T||Q^T) = \min_{P\in\mathcal{P},Q\in\mathcal{Q}} D(P||Q) = D_{\min}$$

where $\mathcal{P} = \{P : P^T = \hat{P}_n^T\}$. In this ideal case, supposing some $\bar{Q}$ in $\mathcal{Q}^T = \{Q^T : Q \in \mathcal{Q}\}$ is the unique minimizer of $D(\hat{P}_n^T||Q')$ subject to $Q' \in cl(\mathcal{Q}^T)$, it also holds that $Q_k^T \to \bar{Q}$. Indeed, for any convergent subsequence of Q_k^T, its limit $Q' \in cl(\mathcal{Q}^T)$ satisfies

$$D(\hat{P}_n^T||Q') = \lim_{k\to\infty} D(\hat{P}_n^T||Q_k^T) = D_{\min},$$

hence $Q' = \bar{Q}$ by the uniqueness assumption. Note that $\bar{Q}$ is the MLE of the distribution governing the elements $y_i = Tx_i$ of the incomplete sample y_1^n, by Lemma 3.1.

If the set $\mathcal{Q}$ of feasible distributions is convex and compact, the above ideal situation always obtains if the EM algorithm is started with a $Q_0 \in \mathcal{Q}$ of maximal support, by Corollary 5.1 in the previous subsection. Then by the last paragraph, supposing the minimizer of $D(\hat{P}_n^T||Q')$ subject to $Q' \in \mathcal{Q}^T$ is unique (for which $S(\hat{P}_n^T) = S(\mathcal{Q}^T)$ is a sufficient condition), the EM algorithm always provides a sequence of distributions Q_k whose T-images approach that minimizer, that is, the MLE of the distribution underlying the incomplete sample. This implies, in turn, that the distributions Q_k themselves converge to a limit, because the P_k's obtained in the steps E as

$$P_k(a) = \hat{P}_n^T(Ta)\frac{Q_{k-1}(a)}{Q_{k-1}^T(Ta)}$$

do converge to a limit, by Corollary 5.1.

An example of the EM algorithm with convex, compact $\mathcal{Q}$ is the decomposition of mixtures in Example 5.1. It should be noted, however, that in most situations where the EM algorithm is used in statistics, the set $\mathcal{Q}$ of feasible distributions is not convex. Then Corollary 5.1 does not apply, and the ideal case $D(\hat{P}_n^T||Q_k^T) \to D_{\min}$ need not obtain; indeed, the iteration often gets stuck at a local optimum. A practical way to overcome this problem is to run the algorithm with several different choices of the initial Q_0.

6

Universal coding

A Shannon code for a distribution P_n on A^n has the length function $\lceil -\log P_n(x_1^n) \rceil$ and produces expected length within 1 bit of the entropy lower bound $H(P_n)$; it therefore provides an almost optimal method for coding if it is known that the data x_1^n is governed by P_n. In practice, however, the distribution governing the data is usually not known, though it may be reasonable to assume that the data are coming from an unknown member of a known class $\mathcal{P}$ of processes, such as the i.i.d. or Markov or stationary processes. Then it is desirable to use "universal" codes that perform well no matter which member of $\mathcal{P}$ is the true process. In this Section, we introduce criteria of "good performance" of codes relative to a process. We also describe universal codes for the classes of i.i.d. and Markov processes, and for some others, which are almost optimal in a strong sense and, in addition, are easy to implement.

By a *process* with alphabet A we mean a Borel probability measure P on A^∞, that is, a probability measure on the σ-algebra generated by the cylinder sets

$$[a_1^n] = \{x_1^\infty : x_1^n = a_1^n\}, \qquad a_1^n \in A^n,\ n = 1, 2, \dots\ ;$$

see the Appendix for a summary of process concepts. The marginal distribution P_n on A^n of a process P is defined by

$$P_n(a_1^n) = P([a_1^n]), \qquad a_1^n \in A^n;$$

we also write briefly $P(a_1^n)$ for $P_n(a_1^n)$.

Simple examples are the i.i.d. processes with

$$P(a_1^n) = \prod_{t=1}^{n} P(a_t), \qquad a_1^n \in A^n,$$

and the Markov chains with

$$P(a_1^n) = P_1(a_1) \prod_{t=2}^{n} P(a_t|a_{t-1}), \qquad a_1^n \in A^n,$$

where $P_1 = \{P_1(a) \colon a \in A\}$ is an *initial distribution*, and $\{P(a|\tilde{a}), a \in A, \tilde{a} \in A\}$ is a *transition probability matrix*, that is, $P(\cdot|\tilde{a})$ is a probability distribution on A for each $\tilde{a} \in A$. *Stationary processes* are those that satisfy

$$P(\{x_1^\infty \colon x_{i+1}^{i+n} = a_1^n\}) = P([a_1^n]), \text{ for each } i, n, \text{ and } a_1^n \in A^n.$$

6.1 Redundancy

The *ideal codelength* of a message $x_1^n \in A^n$ coming from a process P is defined as $-\log P(x_1^n)$. For an arbitrary n-code $C_n \colon A^n \mapsto B^*$, $B = \{0, 1\}$, the difference of its length function from the "ideal" will be called the *redundancy function* $R = R_{P,C_n}$:

$$R(x_1^n) = L(x_1^n) + \log P(x_1^n).$$

The value $R(x_1^n)$ for a particular x_1^n is also called the *pointwise redundancy.*

One justification of this definition is that a Shannon code for P_n, with length function equal to the rounding of the "ideal" to the next integer, attains the least possible expected length of a prefix code $C_n \colon A^n \mapsto B^*$, up to 1 bit (and the least possible expected length of any n-code up to $\log n$ plus a constant), see Section 1. Note that while the expected redundancy

$$E_P(R) = E_P(L) - H(P_n)$$

is non-negative for each prefix code $C_n\colon A^n \mapsto B^*$, the redundancy function takes also negative values, in general. The next theorem shows, however, that pointwise redundancy can never be "substantially" negative for large n, with P-probability 1. This provides additional justification of the definition above.

In the sequel, the term *code* will either mean an n-code $C_n\colon A^n \mapsto B^*$, or a sequence $\{C_n\colon n = 1, 2, \ldots\}$ of n-codes. The context will make clear which possibility is being used. A code $\{C_n\colon n = 1, 2, \ldots\}$ is said to be a prefix code if each C_n is one, and strongly prefix if $C_m(y_1^m) \prec C_n(x_1^n)$ can hold only when $y_1^m \prec x_1^n$.

Theorem 6.1. Given an arbitrary process P and code $\{C_n\colon n = 1, 2, \ldots\}$ (not necessarily prefix),

$$R(x_1^n) \geq -c_n \quad \text{eventually almost surely,}$$

for any sequence of numbers $\{c_n\}$ with $\sum n2^{-c_n} < +\infty$, e.g., for $c_n = 3\log n$. Moreover, if the code is strongly prefix, or its length function satisfies $L(x_1^n) \geq -\log Q(x_1^n)$ for some process Q, then

$$E_P(\inf_n R(x_1^n)) > -\infty.$$

Proof. Let

$$A_n(c) = \{x_1^n\colon R(x_1^n) < -c\} = \{x_1^n\colon 2^{L(x_1^n)}P(x_1^n) < 2^{-c}\}.$$

Then

$$P_n(A_n(c)) = \sum_{x_1^n \in A^n(c)} P(x_1^n) < 2^{-c} \sum_{x_1^n \in A_n(c)} 2^{-L(x_1^n)} \leq 2^{-c} \log |A^n|$$

where, in the last step, we used Theorem 1.2. Hence

$$\sum_{n=1}^{\infty} P_n(A_n(c_n)) \leq \log |A| \cdot \sum_{n=1}^{\infty} n\ 2^{-c_n},$$

and the first assertion follows by the Borel–Cantelli principle.

The second assertion will be established if we show that for codes with either of the stated properties

$$P(\{x_1^\infty\colon \inf{}_n\ R(x_1^n) < -c\}) < 2^{-c}, \qquad \forall\ c > 0\ ,$$

or in other words,

$$\sum_{n=1}^{\infty} P_n(B_n(c)) < 2^{-c}$$

where

$$B_n(c) = \{x_1^n : R(x_1^n) < -c,\ R(x_1^k) \geq -c,\ k < n\}.$$

As in the proof of the first assertion,

$$P_n(B_n(c)) < 2^{-c} \sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)},$$

hence it suffices to show that

$$\sum_{n=1}^{\infty} \sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)} \leq 1.$$

If $\{C_n : n = 1, 2, \ldots\}$ is a strongly prefix code, the mapping $C : (\cup_{n=1}^{\infty} B_n(c)) \mapsto B^*$ defined by $C(x_1^n) = C_n(x_1^n)$, $x_1^n \in B_n(c)$, satisfies the prefix condition, and the claim holds by the Kraft inequality. If $L(x_1^n) \geq -\log Q(x_1^n)$, we have

$$\sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)} \leq \sum_{x_1^n \in B_n(c)} Q(x_1^n) = Q_n(B_n(c)),$$

and the desired inequality follows since

$$\sum_{n=1}^{\infty} Q_n(B_n(c)) = Q(\{x_1^{\infty} : \inf{}_n\ R(x_1^n) < -c\}) \leq 1.$$

□

In the literature, different concepts of universality, of a code $\{C_n : n = 1, 2, \ldots\}$ for a given class $\mathcal{P}$ of processes, have been used. A weak concept requires the convergence to 0 of the expected redundancy per symbol,

$$\frac{1}{n} E_P(R_{P,C_n}) \to 0, \text{ for each } P \in \mathcal{P}; \tag{6.1}$$

stronger concepts require uniform convergence to 0, for $P \in \mathcal{P}$, of either $(1/n)E_P(R_{P,C_n})$ or of $(1/n) \max_{x_1^n \in A^n} R_{P,C_n}(x_1^n)$.

In the context of "strong" universality, natural figures of merit of a code $C_n\colon A^n \mapsto B^*$ (for a given class of processes) are the worst case expected redundancy

$$\overline{R}_{C_n} = \sup_{P\in\mathcal{P}} E_P(R_{P,C_n})$$

and the worst case maximum redundancy

$$R^*_{C_n} = \sup_{P\in\mathcal{P}} \max_{x_1^n\in A^n} R_{P,C_n}(x_1^n).$$

Example 6.1. For the class of i.i.d. processes, natural universal codes are obtained by first encoding the type of x_1^n, and then identifying x_1^n within its type class via enumeration. Formally, for x_1^n of type Q, let $C_n(x_1^n) = C(Q)C_Q(x_1^n)$, where $C\colon\mathcal{P}_n \mapsto B^*$ is a prefix code for n-types ($\mathcal{P}_n$ denotes the set of all n-types), and for each $Q \in \mathcal{P}_n$, $C_Q\colon T^n_Q \mapsto B^*$ is a code of fixed length $\lceil\log|\mathcal{T}^n_Q|\rceil$. This code is an example of what are called *two-stage codes*. The redundancy function $R_{P,C_n} = L(Q) + \lceil\log|\mathcal{T}^n_Q|\rceil + \log P(x_1^n)$ of the code C_n equals $L(Q) + \log P^n(\mathcal{T}^n_Q)$, up to 1 bit, where $L(Q)$ denotes the length function of the type code $C\colon\mathcal{P}_n \mapsto B^*$. Since $P^n(\mathcal{T}^n_Q)$ is maximized for $P = Q$, it follows that for x_1^n in $\mathcal{T}_Q$, the maximum pointwise redundancy of the code C_n equals $L(Q) + \log Q^n(\mathcal{T}^n_Q)$, up to 1 bit.

Consider first the case when the type code has fixed length $L(Q) = \lceil\log|\mathcal{P}_n|\rceil$. This is asymptotically equal to $(|A| - 1)\log n$ as $n \to \infty$, by Lemma 2.1 and Stirling's formula. For types Q of sequences x_1^n in which each $a \in A$ occurs a fraction of time bounded away from 0, one can see via Stirling's formula that $\log Q^n(\mathcal{T}^n_Q)$ is asymptotically $-((|A| - 1)/2)\log n$. Hence for such sequences, the maximum redundancy is asymptotically $((|A| - 1)/2)\log n$. On the other hand, the maximum for x_1^n of $L(Q) + \log Q^n(\mathcal{T}^n_Q)$ is attained when x_1^n consists of identical symbols, when $Q^n(\mathcal{T}^n_Q) = 1$; this shows that $R^*_{C_n}$ is asymptotically $(|A| - 1)\log n$ in this case.

Consider next the case when $C\colon\mathcal{P}_n \mapsto B^*$ is a prefix code of length function $L(Q) = \lceil\log(c_n/Q^n(\mathcal{T}^n_Q))\rceil$ with

$c_n = \sum_{Q\in\mathcal{P}_n} Q^n(\mathcal{T}^n_Q)$; this is a bona-fide length function, satisfying the Kraft inequality. In this case $L(Q) + \log Q^n(\mathcal{T}^n_Q)$ differs from

$\log c_n$ by less than 1, for each Q in $\mathcal{P}_n$, and we obtain that $R^*_{C_n}$ equals $\log c_n$ up to 2 bits. We shall see that this is essentially best possible (Theorem 6.2), and in the present case $R^*_{C_n} = ((|A|-1)/2)\log n + O(1)$ (Theorem 6.3).

Note that to any prefix code $C_n\colon A^n \mapsto B^*$, with length function $L(x_1^n)$, there is a probability distribution Q_n or A^n such that $L(x_1^n) \geq -\log Q_n(x_1^n)$ (one can take $Q_n(x_1^n) = c2^{-L(x_1^n)}$, with $c \geq 1$, using the Kraft inequality). Conversely, to any distribution Q_n on A^n there exists a prefix code with length function $L(x_1^n) < -\log Q_n(x_1^n)+1$ (a Shannon code for Q_n). It follows that for any class $\mathcal{P}$ of processes with alphabet A, the least possible value of $\overline{R}_{C_n}$ or $R^*_{C_n}$ for prefix codes $C_n\colon A^n \mapsto B^*$ "almost" equals

$$\overline{R}_n = \min_{Q_n} \sup_{P\in\mathcal{P}} \sum_{x_1^n\in A^n} P(x_1^n)\log\frac{P(x_1^n)}{Q_n(x_1^n)} \tag{6.2}$$

or

$$R^*_n = \min_{Q_n} \sup_{P\in\mathcal{P}} \max_{x_1^n\in A^n} \log\frac{P(x_1^n)}{Q_n(x_1^n)}. \tag{6.3}$$

More exactly, each prefix code $C_n : A^n \mapsto B^*$ has worst case expected and maximal redundancy not smaller than $\overline{R}_n$ and R^*_n, respectively, and a Shannon code for a Q_n attaining the minimum in (6.2) or (6.3) achieves this lower bound up to 1 bit. In particular, for a class $\mathcal{P}$ of processes, there exist "strongly universal codes" with expected or maximum redundancy per symbol converging to 0 uniformly for $P \in \mathcal{P}$, if and only if $\overline{R}_n = o(n)$ or $R^*_n = o(n)$, respectively.

Our next theorem identifies the minimizer in (6.3) and the value R^*_n. The related problem for $\overline{R}_n$ will be treated in Section 7.

We use the following notation. Given a class $\mathcal{P}$ of processes with alphabet A, we write

$$P_{\mathrm{ML}}(x_1^n) \overset{\mathrm{def}}{=} \sup_{P\in\mathcal{P}} P(x_1^n), \qquad x_1^n \in A^n,$$

where the subscript on P_{ML} emphasizes its interpretation as the maximizer of $P(x_1^n)$ subject to $P \in \mathcal{P}$ (if it exists), that is, as the maximum likelihood estimate of the process $P \in \mathcal{P}$ that generates x_1^n. The normalized form

$$NML_n(a_1^n) \stackrel{\text{def}}{=} P_{\text{ML}}(a_1^n) / \sum_{x_1^n \in A^n} P_{\text{ML}}(x_1^n), \qquad a_1^n \in A^n,$$

is called the *the normalized maximum likelihood distribution.*

Theorem 6.2. For any class $\mathcal{P}$ of processes with finite alphabet A, the minimum in (6.3) is attained for $Q_n = NML_n$, the normalized maximum likelihood distribution, and

$$R_n^* = \log \sum_{x_1^n \in A^n} P_{\text{ML}}(x_1^n).$$

Proof. For arbitrary Q_n,

$$\sup_{P \in \mathcal{P}} \max_{x_1^n \in A^n} \log \frac{P(x_1^n)}{Q_n(x_1^n)} = \log \max_{x_1^n \in A^n} \frac{P_{\text{ML}}(x_1^n)}{Q_n(x_1^n)}.$$

Here

$$\max_{x_1^n \in A^n} \frac{P_{\text{ML}}(x_1^n)}{Q_n(x_1^n)} \geq \sum_{x_1^n \in A^n} Q_n(x_1^n) \frac{P_{\text{ML}}(x_1^n)}{Q_n(x_1^n)} = \sum_{x_1^n \in A^n} P_{\text{ML}}(x_1^n),$$

with equality if $Q_n = NML_n$. □

6.2 Universal codes for certain classes of processes

While Shannon codes for the distributions NML_n, $n = 1, 2, \ldots$ are optimal for the class $\mathcal{P}$ within 1 bit, with respect to the maximum redundancy criterion, by Theorem 6.2, they are typically not practical from the implementation point of view. We will show that for some simple but important classes $\mathcal{P}$ there exist easily implementable arithmetic codes $\{C_n : n = 1, 2, \ldots\}$ which are nearly optimal, in the sense that

$$R_{C_n}^* \leq R_n^* + \text{ constant}, \qquad n = 1, 2, \ldots \tag{6.4}$$

Recall that an arithmetic code (of the second kind, see equation (1.2) in Section 1) determined by the marginals $Q_n, n = 1, 2, \ldots$ of any process

Q, is a prefix code $\{C_n : n = 1, 2, \ldots\}$ with length function $L(x_1^n) = \lceil -\log Q(x_1^n) \rceil + 1$. Note that the obvious idea to take a process Q with marginals $Q_n = NML_n$ does not work, since such a process typically does not exist (that is, the distributions NML_n, $n = 1, 2, \ldots$, do not meet the consistency criteria for a process).

Below we describe suitable "coding processes" Q, and for the corresponding arithmetic codes we prove upper bounds to $R^*_{C_n}$. For the class of i.i.d processes, we also determine R^*_n, up to a constant, and establish the bound (6.4) for our code. For other classes, the proof of the claimed near optimality will be completed in the next subsection, where we also prove near optimality in the expected redundancy sense.

In the rest of this subsection, we assume with no loss of generality that $A = \{1, \ldots, k\}$.

Consider first the case when $\mathcal{P}$ is the class of i.i.d. processes with alphabet A. Let the "coding process" be the process Q whose marginal distributions $Q_n = \{Q(x_1^n) : x_1^n \in A^n\}$ are given by

$$Q(x_1^n) = \prod_{t=1}^{n} \frac{n(x_t|x_1^{t-1}) + \frac{1}{2}}{t - 1 + \frac{k}{2}},$$

where $n(i|x_1^{t-1})$ denotes the number of occurrences of the symbol i in the "past" x_1^{t-1}. Equivalently,

$$Q(x_1^n) = \frac{\prod_{i=1}^{k} [(n_i - \frac{1}{2})(n_i - \frac{3}{2}) \cdots \frac{1}{2}]}{(n - 1 + \frac{k}{2})(n - 2 + \frac{k}{2}) \cdots \frac{k}{2}}, \tag{6.5}$$

where $n_i = n(i|x_1^n)$, and $(n_i - \frac{1}{2})(n_i - \frac{3}{2}) \ldots \frac{1}{2} = 1$, by definition, if $n_i = 0$.

Note that the conditional probabilities needed for arithmetic coding are given by the simple formula

$$Q(i|x_1^{t-1}) = \frac{n(i|x_1^{t-1}) + \frac{1}{2}}{t - 1 + \frac{k}{2}}.$$

Intuitively, this $Q(i|x_1^{t-1})$ is an estimate of the probability of i from the "past" x_1^{t-1}, under the assumption that the data come from an unknown $P \in \mathcal{P}$. The unbiased estimate $n(i|x_1^{t-1})/(t-1)$ would be inappropriate here, since an admissible coding process requires $Q(i|x_1^{t-1}) > 0$ for each possible x_1^{t-1} and i.

Remark 6.1. It is not intuitively obvious at this point in our discussion why exactly 1/2 is the "right" bias term that admits the strong redundancy bound below. Later, in Section 7, we establish a deep connection between minimax expected redundancy $\bar{R}_n$ and mixture distributions with respect to priors (which is closely connected to Bayesian ideas); the 1/2 then arises from using a specific prior. For now we note only that replacing 1/2 by 1 in formula (6.5) leads to the coding distribution $Q(x_1^n) = \frac{\prod_1^n n_i!}{(n-1+k)(n-2+k)\cdots k}$ which equals $\frac{1}{|P_n|\cdot|\mathcal{T}_Q|}$, if $x_1^n \in \mathcal{T}_Q$, see Lemma 2.1. In this case, the length function is the same (up to 2 bits) as the first, suboptimal, version of the two-stage code in Example 6.1.

We claim that the arithmetic code determined by the process Q satisfies

$$R^*_{C_n} \leq \frac{k-1}{2} \log n \ + \ \text{constant}. \tag{6.6}$$

Since the length function is $L(x_1^n) = \lceil \log Q(x_1^n) \rceil + 1$, our claim will be established if we prove

Theorem 6.3. For Q determined by (6.5), and any i.i.d. process P with alphabet $A = \{1, \ldots, k\}$,

$$\frac{P(x_1^n)}{Q(x_1^n)} \leq K_0 \ n^{\frac{k-1}{2}}, \qquad \forall \ x_1^n \in A^n,$$

where K_0 is a constant depending on the alphabet size k only.

Proof. We begin by noting that given $x_1^n \in A^n$, the i.i.d. process with largest $P(x_1^n)$ is that whose one-dimensional distribution equals the empirical distribution $(n_1/n, \ldots, n_k/n)$ of x_1^n, see Lemma 2.3, and hence

$$P(x_1^n) \leq P_{\mathrm{ML}}(x_1^n) = \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i}.$$

In a moment we will use a combinatorial argument to establish the bound

$$\prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n} \ \leq \ \frac{\prod_{i=1}^{k} [(n_i - \frac{1}{2})(n_i - \frac{3}{2}) \cdots \frac{1}{2}]}{(n - \frac{1}{2})(n - \frac{3}{2}) \cdots \frac{1}{2}}. \tag{6.7}$$

This is enough to yield the desired result, for if it is true, then we can use $P(x_1^n)/Q(x_1^n) \leq P_{\mathrm{ML}}(x_1^n)/Q(x_1^n)$ and the Q-formula, (6.5), to obtain

$$\frac{P(x_1^n)}{Q(x_1^n)} \leq \prod_{j=1}^{n} \frac{n + \frac{k}{2} - j}{n + \frac{1}{2} - j}, \qquad \forall\ x_1^n \in A^n. \tag{6.8}$$

If the alphabet size k is odd, the product here simplifies, and is obviously of order $n^{\frac{k-1}{2}}$. If k is even, using

$$\begin{aligned} (n - \tfrac{1}{2})(n - \tfrac{3}{2}) \cdots \tfrac{1}{2} &= \frac{(2n-1)(2n-3)\cdots 1}{2^n} \\ &= \frac{(2n)!}{2^{2n} n!} = \frac{2n(2n-1)\cdots(n+1)}{2^{2n}}, \end{aligned} \tag{6.9}$$

the product in (6.8) can be rewritten as

$$\frac{(n + \frac{k}{2} - 1)!/(\frac{k}{2} - 1)!}{(2n)!/2^{2n} n!},$$

and Stirling's formula gives that this is of order $n^{\frac{k-1}{2}}$. Hence, it indeed suffices to prove (6.7).

To prove (6.7), we first use (6.9) to rewrite it as

$$\prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^n \leq \frac{\prod_{i=1}^{k}[2n_i(2n_i - 1)\cdots(n_i + 1)]}{2n(2n+1)\cdots(n+1)}, \tag{6.10}$$

which we wish to establish for k-tuples of non-negative integers n_i with sum n. This will be done if we show that it is possible to assign to each $\ell = 1, \ldots, n$ in a one-to-one manner, a pair $(i, j),\ 1 \leq i \leq k,\ 1 \leq j \leq n$, such that

$$\frac{n_i}{n} \leq \frac{n_i + j}{n + \ell}. \tag{6.11}$$

Now, for any given ℓ and i, (6.11) holds iff $j \geq n_i \ell / n$. Hence the number of those $1 \leq j \leq n_i$ that satisfy (6.11) is greater than $n_i - n_i \ell / n$, and the total number of pairs $(i, j),\ 1 \leq i \leq k,\ 1 \leq j \leq n$, satisfying (6.11) is greater than

$$\sum_{i=1}^{k} \left(n_i - \frac{n_i}{n}\ell\right) = n - \ell.$$

It follows that if we assign to $\ell = n$ any (i, j) satisfying (6.11) (i. e., i may be chosen arbitrarily and $j = n_i$), then recursively assign to each

$\ell = n-1, n-2$, etc., a pair (i,j) satisfying (6.11) that was not assigned previously, we never get stuck; at each step there will be at least one "free" pair (i,j) (because the total number of pairs (i,j) satisfying (6.11) is greater than $n-\ell$, the number of pairs already assigned.) This completes the proof of the theorem. □

Remark 6.2. The above proof has been preferred for it gives a sharp bound, namely, in equation (6.7) the equality holds if x_1^n consists of identical symbols, and this bound could be established by a purely combinatorial argument. An alternative proof via Stirling's formula, however, yields both upper and lower bounds. Using equation (6.9), the numerator in equation (6.5) can be written as

$$\prod_{i:n_i\neq 0} \frac{(2n_i)!}{2^{2n_i} n!},$$

which, by Stirling's formula, is bounded both above and below by constant times $e^{-n}\prod_{i:n_i\neq 0} n_i^{n_i}$. The denominator in equation (6.5) can also be expressed by factorials (trivially if k is even, and via equation (6.9) if k is odd), and Stirling's formula shows that it is bounded both above and below by a constant times $e^{-n}n^{n+\frac{k-1}{2}}$. This admits the conclusion that $P_{\mathrm{ML}}(x_1^n)/Q(x_1^n)$ is bounded both above and below by a constant times $n^{\frac{k-1}{2}}$, implying

Theorem 6.4. For the class of i.i.d processes,

$$R_n^* = \log \sum_{x_1^n \in A^n} P_{\mathrm{ML}}(x_1^n) = \frac{k-1}{2}\log n + O(1).$$

Consequently, our code satisfying equation (6.6) is nearly optimal in the sense of equation (6.4).

Next, let $\mathcal{P}$ be the class of Markov chains with alphabet $A = \{1, \ldots, k\}$. We claim that for this class, the arithmetic code determined by the process Q below satisfies

$$R_{C_n}^* \ \leq \ \frac{k(k-1)}{2}\log n \ + \ \text{constant}. \tag{6.12}$$

Let the "coding process" be that Q whose marginal distributions $Q_n = \{Q(x_1^n)\colon x_1^n \in A^n\}$ are given by

$$Q(x_1^n) = \frac{1}{k}\prod_{i=1}^{k} \frac{\prod_{j=1}^{k}[(n_{ij}-1/2)(n_{ij}-3/2)\dots 1/2]}{(n_i-1+k/2)(n_i-2+k/2)\dots k/2}; \tag{6.13}$$

here n_{ij} is the number of times the pair i, j occurs in adjacent places in x_1^n, and $n_i = \sum_j n_{ij}$. Note that n_i is now the number of occurrences of i in the block x_1^{n-1} (rather than in x_1^n as before). The conditional Q-probabilities needed for arithmetic coding are given by

$$Q(j|x_1^{t-1}) = \frac{n_{t-1}(i,j)+\frac{1}{2}}{n_{t-1}(i)+\frac{k}{2}}, \text{ if } x_{t-1}=i,$$

where $n_{t-1}(i,j)$ and $n_{t-1}(i)$ have similar meaning as n_{ij} and n_i above, with x_1^{t-1} in the role of x_1^n.

Similarly to the i.i.d. case, to show that the arithmetic code determined by Q above satisfies (6.12) for the class of Markov chains, it suffices to prove

Theorem 6.5. For Q determined by (6.13) and any Markov chain with alphabet $A = \{1.\dots,k\}$,

$$\frac{P(x_1^n)}{Q(x_1^n)} \le K_1\, n^{\frac{k(k-1)}{2}}, \qquad \forall\, x_1^n \in A^n,$$

where K_1 is a constant depending on k only.

Proof. For any Markov chain, the probability of $x_1^n \in A^n$ is of form

$$P(x_1^n) = P_1(x_1)\prod_{t=2}^{n} P(x_t|x_{t-1}) = P_1(x_1)\prod_{i=1}^{k}\prod_{j=1}^{k} P(j|i)^{n_{ij}}.$$

This and (6.13) imply that

$$\frac{P(x_1^n)}{Q(x_1^n)} \le k\prod_{i=1}^{k}\left[\prod_{j=1}^{k} P(j|i)^{n_{ij}} \Big/ \frac{\prod_{j=1}^{k}[(n_{ij}-1/2)(n_{ij}-3/2)\dots 1/2]}{(n_i-1+k/2)(n_i-2+k/2)\dots k/2}\right].$$

Here, when $n_i \neq 0$, the square bracket is the same as the ratio in Theorem 6.3 for a sequence $x_1^{n_i} \in A^{n_i}$ with empirical distribution

$(n_{i1}/n_i, \ldots, n_{ik}/n_i)$, and an i.i.d. process with one-dimensional distribution $P(\cdot|i)$. Hence, it follows from Theorem 6.3 that

$$\frac{P(x_1^n)}{Q(x_1^n)} \le k \prod_{n_i \neq 0} \left[K_0 \; n_i^{\frac{k-1}{2}} \right] \le (k \; K_0^k) n^{\frac{k(k-1)}{2}}.$$

□

Consider next the class of Markov chains of order at most m, namely of those processes P (with alphabet $A = \{1, \ldots, k\}$) for which the probabilities $P(x_1^n)$, $x_1^n \in A^n$, $n \ge m$ can be represented as

$$P(x_1^n) = P_m(x_1^m) \prod_{t=m+1}^{n} P(x_t | x_{t-m}^{t-1}),$$

where $P(\cdot|a_1^m)$ is a probability distribution for each $a_1^n \in A^n$. The Markov chains considered before correspond to $m = 1$. To the analogy of that case we now define a "coding process" Q whose marginal distributions Q_n, $n \ge m$, are given by

$$Q(x_1^n) = \frac{1}{k^m} \prod_{a_1^m \in A^m} \frac{\prod_{j=1}^{k} [(n_{a_1^m j} - 1/2)(n_{a_1^m j} - 3/2) \ldots 1/2]}{(n_{a_1^m} - 1 + k/2)(n_{a_1^m} - 2 + k/2) \ldots k/2}, \tag{6.14}$$

where $n_{a_1^m j}$ denotes the number of times the block $a_1^m j$ occurs in x_1^n, and $n_{a_1^m} = \sum_j n_{a_1^m j}$ is the number of times the block a_1^m occurs in x_1^{n-1}.

The same argument as in the proof of Theorem 6.5 gives that for Q determined by (6.14), and any Markov chain of order m,

$$\frac{P(x_1^n)}{Q(x_1^n)} \le K_m \; n^{\frac{k^m(k-1)}{2}}, \qquad K_m = k^m \; K_0^{k^m}. \tag{6.15}$$

It follows that the arithmetic code determined by Q in (6.14) is a universal code for the class of Markov chains of order m, satisfying

$$R_{C_n}^* \le \frac{k^m(k-1)}{2} \log n \; + \; \text{constant}. \tag{6.16}$$

Note that the conditional Q-probabilities needed for arithmetic coding are now given by

$$Q(j|x_1^{t-1}) = \frac{n_{t-1}(a_1^m, j) + \frac{1}{2}}{n_{t-1}(a_1^m) + \frac{k}{2}}, \quad \text{if} \quad x_{t-m}^{t-1} = a_1^m,$$

where $n_{t-1}(a_1^m, j)$ and $n_{t-1}(a_1^m)$ are defined similarly to $n_{a_1^m j}$ and $n_{a_1^m}$, with x_1^{t-1} in the role of x_1^n.

A subclass of the Markov chains of order m, often used in statistical modeling, is specified by the assumption that the transition probabilities $P(j|a_1^m)$ depend on a_1^m through a "context function" $f(a_1^m)$ that has less than k^m possible values, say $1, \ldots, s$. For $m < t \leq n$, the t'th symbol in a sequence $x_1^n \in A^n$ is said to occur in context ℓ if $f(x_{t-m}^{t-1}) = \ell$. A suitable coding process for this class, determined by the context function f, is defined by

$$Q(x_1^m) = \frac{1}{k^m} \prod_{\ell=1}^{s} \frac{\prod_{j=1}^{k} [(n_{\ell,j} - 1/2)(n_{\ell,j} - 3/2) \ldots 1/2]}{(n_\ell - 1 + k/2)(n_\ell - 2 + k/2) \ldots k/2},$$

where $n_{\ell,j}$ denotes the number of times j occurs in context ℓ in the sequence x_1^n, and $n_\ell = \sum\limits_{j=1}^{k} n_{\ell,j}$. The arithmetic code determined by this process Q satisfies, for the present class,

$$R^*_{C_n} \leq \frac{s(k-1)}{2} \log n \ + \ \text{constant}, \tag{6.17}$$

by the same argument as above. The conditional Q-probabilities needed for arithmetic coding are now given by

$$Q(j|x_1^{t-1}) = \frac{n_{t-1}(\ell, j) + \frac{1}{2}}{n_{t-1}(\ell) + \frac{k}{2}}, \qquad \text{if} \qquad f(x_{t-m}^{t-1}) = \ell.$$

Finally, let $\mathcal{P}$ be the class of all stationary processes with alphabet $A = \{1, \ldots, k\}$. This is a "large" class that does not admit strong sense universal codes, that is, the convergence in (6.1) cannot be uniform for any code, see Example 8.3 in Section 8. We are going to show, however, that the previous universal codes designed for Markov chains of order m perform "reasonably well" also for the class $\mathcal{P}$ of stationary processes, and can be used to obtain universal codes for $\mathcal{P}$ in the weak sense of (6.1).

To this end, we denote by $Q^{(m)}$ the coding process defined by (6.14) tailored to the class of Markov chains of order m (in particular, $Q^{(0)}$ is the process defined by (6.5)), and by $\{C_n^m : n = 1, 2, \ldots\}$ the arithmetic code determined by the process $Q^{(m)}$.

Theorem 6.6. Let $P \in \mathcal{P}$ have entropy rate $\overline{H} = \lim_{n\to\infty} H_m$ where, with $\{X_n\}$ denoting a representation of P, $H_m = H(X_{m+1}|X_1,\ldots,X_m)$. Then

$$\frac{1}{n}E_P(R_{P,C_n^m}) \leq H_m - \overline{H} + \frac{k^m(k-1)}{2}\,\frac{\log n}{n} + \frac{c_m}{n},$$

where c_m is a constant depending only on m and the alphabet size k, with $c_m = O(k^m)$ as $m \to \infty$.

Corollary 6.1. For any sequence of integers $m_n \to \infty$ with $m_n \leq \alpha \log n$, $\alpha < 1/\log k$, the prefix code $\{C_n^{m_n}: n = 1,2,\ldots\}$ satisfies (6.1). Moreover, the arithmetic code determined by the mixture process

$$Q = \sum_{m=0}^{\infty} \alpha_m Q^{(m)} \quad (\text{with} \quad \alpha_m > 0, \sum \alpha_m = 1)$$

also satisfies (6.1).

Proof. Given a stationary process P, let $P^{(m)}$ denote its m'th Markov approximation, that is, the stationary Markov chain of order m with

$$P^{(m)}(x_1^n) = P(x_1^m)\prod_{t=m+1}^{n} P(x_t|x_{t-m}^{t-1}), \qquad x_1^n \in A^n,$$

where

$$P(x_t|x_{t-m}^{t-1}) = \text{Prob}\{X_t = x_t|X_{t-m}^{t-1} = x_{t-m}^{t-1}\}.$$

The bound (6.15) applied to $P^{(m)}$ in the role of P gives

$$\begin{aligned}\log\frac{P(x_1^n)}{Q^{(m)}(x_1^n)} &= \log\frac{P(x_1^n)}{P^{(m)}(x_1^n)} + \log\frac{P^{(m)}(x_1^n)}{Q^{(m)}(x_1^n)}\\ &\leq \log\frac{P(x_1^n)}{P^{(m)}(x_1^n)} + \frac{k^m(k-1)}{2}\log n + \log K_m,\end{aligned}$$

where $\log K_m = O(k^m)$.

Note that the expectation under P of $\log P(x_1^n)$ equals $-H(P_n)$, and that of $\log P(x_t|x_{t-m}^{t-1})$ equals $-H_m$. Hence for the code C_n^m with length function $L(x_1^n) = \lceil -\log Q^{(m)}(x_1^n)\rceil + 1$, the last bound gives that

$$\begin{aligned}E_P(R_{P,C_n}) &< E_P\left(\log\frac{P(x_1^n)}{Q^{(m)}(x_1^n)}\right) + 2\\ &\leq -H(P_n) + H(P_m) + (n-m)H_m + \tfrac{k^m(k-1)}{2}\log n + \log K_m + 2.\end{aligned}$$

Since

$$H(P_n) - H(P_m) = H(X_1^n) - H(X_1^m) = \sum_{i=m}^{n-1} H(X_{i+1}|X_1^i) \geq (n-m)\overline{H},$$

the assertion of the theorem follows. □

The corollary is immediate, noting for the second assertion that $Q(x_1^n) \geq \alpha_m Q^{(m)}(x_1^m)$ implies

$$\log \frac{P(x_1^n)}{Q(x_1^n)} \leq \log \frac{P(x_1^n)}{Q^{(m)}(x_1^n)} - \log \alpha_m.$$

Remark 6.3. The last inequality implies that for Markov chains of any order m, the arithmetic code determined by $Q = \sum_{m=0}^{\infty} \alpha_m Q^{(m)}$ performs effectively as well as that determined by $Q^{(m)}$, the coding process tailored to the class of Markov chains of order m: the increase in pointwise redundancy is bounded by a constant (depending on m). Of course, the situation is similar for other finite or countable mixtures of coding processes. For example, taking a mixture of coding processes tailored to subclasses of the Markov chains of order m corresponding to different context functions, the arithmetic code determined by this mixture will satisfy the bound (6.17) whenever the true process belongs to one of the subclasses with s possible values of context function. Such codes are sometimes called *twice universal.* Their practicality depends on how easily the conditional probabilities of the mixture process, needed for arithmetic coding, can be calculated. This issue is not entered here, but we note that for the case just mentioned (with a natural restriction on the admissible context functions) the required conditional probabilities can be calculated via a remarkably simple "context weighting algorithm".

7

Redundancy bounds

In this Section, we address code performance for a class of processes with respect to the expected redundancy criterion. We also show that the universal codes constructed for certain classes in the previous Section are optimal within a constant, both for the maximum and expected redundancy criteria.

As noted in the previous Section, the least possible worst case expected redundancy $\overline{R}_{C_n}$, attainable for a given class $\mathcal{P}$ of processes by prefix codes $C_n\colon A^n \mapsto B^*$, exceeds by less than 1 bit the value

$$\overline{R}_n = \min_{Q_n} \sup_{P \in \mathcal{P}} D(P_n \| Q_n), \tag{7.1}$$

see (6.2). Moreover, a distribution Q_n^* attaining this minimum is effectively an optimal coding distribution for n-length messages tailored to the class $\mathcal{P}$, in the sense that a Shannon code for Q_n^* attains the least possible worst case expected redundancy within 1 bit.

Next we discuss a remarkable relationship of the expression (7.1) to the seemingly unrelated concepts of mutual information and channel capacity. As process concepts play no role in this discussion, we shall simply consider some set Π of probability distributions on A, and its I-divergence radius, defined as the minimum for Q of $\sup_{P \in \Pi} D(P\|Q)$.

Later the results will be applied to A^n and the set of marginal distributions on A^n of the processes $P \in \mathcal{P}$, in the role of A and Π.

7.1 I-radius and channel capacity

The I-*radius* of a set Π of distributions on A is the minimum, for distributions Q on A, of $\sup_{P \in \Pi} D(P \| Q)$. If the minimum is attained by a unique $Q = Q^*$ (as we shall show, this is always the case), the minimizer Q^* is called the I-*centroid* of the set Π.

In the following lemma and theorems, we consider "parametric" sets of probability distributions $\Pi = \{P_\theta, \theta \in \Theta\}$, where Θ is a Borel subset of R^k, for some $k \geq 1$, and $P_\theta(a)$ is a measurable function of θ for each $a \in A$.

In information theory parlance, $\{P_\theta, \theta \in \Theta\}$ defines a *channel* with input alphabet Θ and output alphabet A: when an input $\theta \in \Theta$ is selected, the output is governed by the distribution $P_\theta = \{P_\theta(a), a \in A\}$. If the input is selected at random according a probability measure ν on Θ, the information that the output provides for the input is measured by the mutual information

$$I(\nu) = H(Q_\nu) - \int H(P_\theta)\nu(d\theta),$$

where $Q_\nu = \{Q_\nu(a)\colon\ a \in A\}$ is the "output distribution" on A corresponding to the "input distribution" ν, that is,

$$Q_\nu(a) = \int P_\theta(a)\nu(d\theta),\ a \in A.$$

The supremum of the mutual information $I(\nu)$ for all probability measures ν on Θ is the *channel capacity*. A measure ν_0 is a *capacity achieving distribution* if $I(\nu_0) = \sup_\nu I(\nu)$.

Lemma 7.1. For arbitrary distributions Q on A and ν on Θ,

$$\int D(P_\theta \| Q)\nu(d\theta) = I(\nu) + D(Q_\nu \| Q).$$

Proof. Both sides equal $+\infty$ if $S(Q)$, the support of Q, does not contain $S(P_\theta)$ for ν-almost all $\theta \in \Theta$. If it does we can write

$$\begin{aligned}\int D(P_\theta \| Q)\nu(d\theta) &= \int \Big(\sum_{a\in A} P_\theta(a) \log \frac{P_\theta(a)}{Q(a)} \Big) \nu(d\theta) \\ &= \int \Big(\sum_{a\in A} P_\theta(a) \log \frac{P_\theta(a)}{Q_\nu(a)} \Big) \nu(d\theta) + \int \Big(\sum_{a\in A} P_\theta(a) \log \frac{Q_\nu(a)}{Q(a)} \Big) \nu(d\theta).\end{aligned}$$

Using the definition of Q_ν, the first term of this sum is equal to $I(\nu)$, and the second term to $D(Q_\nu \| Q)$. □

Theorem 7.1. For arbitrary distributions Q on A and ν on Θ,

$$\sup_{\theta\in\Theta} D(P_\theta \| Q) \geq I(\nu),$$

with equality if and only if ν is a capacity achieving distribution and $Q = Q_\nu$.

Proof. The inequality follows immediately from Lemma 7.1, as does the necessity of the stated condition of equality. To prove sufficiency, suppose on the contrary that there is a capacity achieving distribution ν_0 such that $D(P_{\theta_0} \| Q_{\nu_0}) > I(\nu_0)$, for some $\theta_0 \in \Theta$.

Denote by ν_1 the point mass at θ_0 and set $\nu_t = (1-t)\nu_0 + t\nu_1$, $0 < t < 1$. Then by the definition of $I(\nu)$,

$$I(\nu_t) = H(Q_{\nu_t}) - (1-t) \int H(P_\theta)\nu_0(d\theta) - tH(P_{\theta_0}),$$

so that,

$$\frac{d}{dt} I(\nu_t) = \frac{d}{dt} H(Q_{\nu_t}) + \int H(P_\theta)\nu_0(d\theta) - H(P_{\theta_0}).$$

Since $Q_{\nu_t} = (1-t)Q_{\nu_0} + tP_{\theta_0}$, simple calculus gives that

$$\frac{d}{dt} H(Q_{\nu_t}) = \sum_a (Q_{\nu_0}(a) - P_{\theta_0}(a)) \log Q_{\nu_t}(a).$$

It follows that

$$\lim_{t\downarrow 0} \frac{d}{dt} I(\nu_t) = -I(\nu_0) + D(P_{\theta_0} \| Q_{\nu_0}) > 0,$$

contradicting the assumption that ν_0 is capacity achieving (which implies that $I(\nu_t) \leq I(\nu_0)$). The proof of the theorem is complete. $\square$

Note that any set Π of distributions on A which is a Borel subset of $R^{|A|}$ (with the natural identification of distributions with points in $R^{|A|}$) has a natural parametric representation, with $\Theta = \Pi$ and $\theta \mapsto P_\theta$ the identity mapping. This motivates consideration, for probability measures μ on Π, of the mutual information

$$I(\mu) = H(Q_\mu) - \int_\Pi H(P)\mu(dP), \quad Q_\mu = \int_\Pi P\mu(dP). \tag{7.2}$$

Lemma 7.2. For any closed set Π of distributions on A, there exists a probability measure μ_0 concentrated on a finite subset of Π of size $m \leq |A|$ that maximizes $I(\mu)$. If a parametric set of distributions $\{P_\theta, \theta \in \Theta\}$ is closed, there exists a capacity achieving distribution ν_0 concentrated on a finite subset of Θ of size $m \leq |A|$.

Proof. If Π is a closed (hence compact) subset of $R^{|A|}$, the set of all probability measures on Π is compact in the usual topology of weak convergence, where $\mu_n \to \mu$ means that $\int\limits_\Pi \Phi(P)\mu_n(dP) \to \int\limits_\Pi \Phi(P)\mu(dP)$ for every continuous function Φ on Π. Since $I(\mu)$ is continuous in that topology, its maximum is attained.

Theorem 7.1 applied with the natural parametrization of Π gives that if μ^* maximizes $I(\mu)$ then $Q^* = Q_{\mu^*}$ satisfies $D(P\|Q^*) \leq I(\mu^*)$ for each $\theta \in \Theta$. Since $I(\mu^*) = \int\limits_\Pi D(P\|Q^*)\mu^*(dP)$, by Lemma 4.2, it follows that $D(P\|Q^*) = I(\mu^*)$ for μ^*-almost all $P \in \Pi$, thus $Q^* = \int\limits_\Pi P\mu^*(dP)$ belongs to the convex hull of the set of those $P \in \Pi$ that satisfy $D(P\|Q) = I(\mu^*)$. Since the probability distributions on A belong to an $(|A| - 1)$-dimensional affine subspace of $R^{|A|}$, this implies by Caratheodory's theorem that Q^* is a convex combination of $m \leq |A|$ member of the above set, that is, $Q^* = \sum\limits_{i=1}^{m} \alpha_i P_i$ where the distributions $P_i \in \Pi$ satisfy $D(P_i\|Q^*) = I(\mu^*)$, $i = 1, \ldots, m$. Then the probability measure μ_0 concentrated on $\{P_1, \ldots, P_m\}$ that assigns weight α_i to P_i, satisfies $I(\mu_0) = I(\mu^*)$, completing the proof of the first assertion.

The second assertion follows by applying the one just proved to $\Pi = \{P_\theta, \theta \in \Theta\}$, because any probability measure ν on Θ and its image μ on Π under the mapping $\theta \mapsto P_\theta$ satisfy $I(\nu) = I(\mu)$, and any measure concentrated on a finite subset $\{P_{\theta_1}, \dots, P_{\theta_m}\}$ of Π is the image of one concentrated on $\{\theta_1, \dots, \theta_m\} \subseteq \Theta$. □

Corollary 7.1. Any set Π of probability distributions on A has an I-centroid, that is, a unique Q^* attains the minimum of $\sup\limits_{P \in \Pi} D(P\|Q)$.

Proof: For Π closed, the existence of I-centroid follows from the fact that the maximum of $I(\mu)$ is attained, by Theorem 7.1 applied with the natural parametrization of Π. For arbitrary Π, it suffices to note that the I-centroid of the closure of Π is also the I-centroid on Π, since $\sup_{P \in \Pi} D(P\|Q) = \sup_{P \in c\ell(\Pi)} D(P\|Q)$, for any Q. □

Theorem 7.2. For any parametric set of distributions $\{P_\theta, \theta \in \Theta\}$, the I-radius equals the channel capacity $\sup I(\nu)$, and Q_{ν_n} converges to the I-centroid Q^* whenever $I(\nu_n) \to \sup I(\nu)$.

Proof: Let Π denote the closure of $\{P_\theta, \theta \in \Theta\}$. Then both sets have the same I-radius, whose equality to $\sup I(\nu)$ follows from Theorem 7.1 and Lemma 7.2 if we show that to any probability measure μ_0 concentrated on a finite subset $\{P_1, \dots, P_m\}$ of Π, there exist probability measures ν_n on Θ with $I(\nu_n) \to I(\mu_0)$.

Such ν_n's can be obtained as follows. Take sequences of distributions in $\{P_\theta, \theta \in \Theta\}$ that converge to the P_i's, say $P_{\theta_{i,n}} \to P_i$, $i = 1, \dots, m$. Let ν_n be the measure concentrated on $\{\theta_{1,n}, \dots, \theta_{m,n}\}$, giving the same weight to $\theta_{i,n}$ that μ_0 gives to P_i. □

Finally, we establish a lower bound to channel capacity, more exactly, to the mutual information $I(\nu)$ for a particular choice of ν, that will be our key tool to bounding worst case expected redundancy from below. Given a parametric set $\{P_\theta, \theta \in \Theta\}$ of distributions on A, a mapping $\hat{\theta}\colon A \mapsto \Theta$ is regarded as a good estimator of the parameter θ

if the mean square error

$$E_\theta\|\theta - \hat{\theta}\|^2 = \sum_{x\in A} P_\theta(x)\|\theta - \hat{\theta}(x)\|^2$$

is small for each $\theta \in \Theta$. We show that if a good estimator exists, the channel capacity cannot be too small.

Theorem 7.3. If the parameter set $\Theta \subseteq R^k$ has Lebesgue measure $0 < \lambda(\Theta) < \infty$, and an estimator $\hat{\theta}\colon A \mapsto \Theta$ exists with

$$E_\theta\|\theta - \hat{\theta}\|^2 \le \varepsilon \;\text{ for each }\; \theta \in \Theta \;,$$

then for ν equal to the uniform distribution on Θ,

$$I(\nu) \ge \frac{k}{2}\log\frac{k}{2\pi e\varepsilon} + \log\lambda(\Theta).$$

To prove this theorem, we need some standard facts from information theory, stated in the next two lemmas. The *differential entropy* of a random variable X with values in R^k that has a density $f(x)$, is defined as

$$H(X) = -\int f(x)\log f(x)dx;$$

thus H denotes entropy as before in the discrete case, and differential entropy in the continuous case. The conditional differential entropy of X given a random variable Y with values in a finite set A (more general cases will not be needed below), is defined similarly as

$$H(X|Y) = \sum_{a\in A} P(a)\Big[-\int f(x|a)\log f(x|a)dx\Big],$$

where $P(a)$ is the probability of $Y = a$, and $f(x|a)$ is the conditional density of X on the condition $Y = a$.

Lemma 7.3. For X and Y as above, $I(X \wedge Y) = H(X) - H(X|Y)$. Moreover, if Z is a function of Y then $H(X|Y) \le H(X|Z) \le H(X)$.

Proof. By the definition of mutual information of random variables, one with values in a finite set and the other arbitrary, see Section 1,

$$I(X \wedge Y) = H(Y) - H(Y|X) = H(P) - \int H(P(\cdot|x))f(x)dx,$$

where $P(\cdot|x)$ denotes the conditional distribution of Y on the condition $X = x$. Substituting the formula for the latter, $P(a|x) = P(a)f(x|a)/f(x)$, into the above equation, the claimed identity

$$I(X \wedge Y) = -\int f(x)\log f(x)dx + \sum_{a\in A} P(a)\int f(x|a)\log f(x|a)dx$$

follows by simple algebra.

Next, if Z is a function of Y, for each possible value c of Z let $A(c)$ denote the set of possible values of Y when $Z = c$. Then the conditional density of X on the condition $Z = c$ is given by

$$g(x|c) = \frac{\sum_{a\in A(c)} P(a)f(x|a)}{\sum_{a\in A(c)} P(a)},$$

and Jensen's inequality for the concave function $-t\log t$ yields that

$$\sum_{a\in A(c)} P(a)(-f(x|a)\log f(x|a)) \le (\sum_{a\in A(c)} P(a))(-g(x|c)\log g(x|c)).$$

Hence, by integrating and summing for all possible c, the claim $H(X|Y) \le H(X|Z)$ follows. Finally, $H(X|Z) \le H(X)$ follows similarly. □

Lemma 7.4. A k-dimensional random variable $V = (V_1, \ldots, V_k)$ with $E\|V\|^2 \le k\sigma^2$ has maximum differential entropy if $V_1, \ldots, V_k$ are independent and have Gaussian distribution with mean 0 and variance σ^2, and this maximum entropy is $\frac{k}{2}\log(2\pi e\sigma^2)$.

Proof. The integral analogue of the log-sum inequality is

$$\int a(x)\log\frac{a(x)}{b(x)}\,dx \ge a\log\frac{a}{b},\ a = \int a(x)dx, b = \int b(x)dx,$$

valid for any non-negative integrable functions on R^k. Letting $a(x)$ be any k-dimensional density for which $E\|V\|^2 \le k\sigma^2$, and $b(x)$ be the

Gaussian density $\prod_i (2\pi\sigma^2)^{-1/2} e^{(-x_i^2/2\sigma^2)}$, this inequality gives

$$\int a(x)\log a(x)dx - \int a(x)\Big[(k/2)\log(2\pi\sigma^2) + \sum(x_i^2/2\sigma^2)\log e\Big]dx \geq 0.$$

Here $\int a(x)(\sum x_i^2)dx \leq k\sigma^2$ by assumption, hence the assertion

$$-\int a(x)\log a(x)dx \leq (k/2)\log(2\pi e\sigma^2)$$

follows, with equality if $a(x) = b(x)$. □

Proof of Theorem 7.3. Let X be a random variable uniformly distributed on Θ, and let Y be the channel output corresponding to input X, that is, a random variable with values in A whose conditional distribution on the condition $X = \theta$ equals P_θ. Further, let $Z = \hat{\theta}(Y)$. Then, using Lemma 7.3,

$$\begin{aligned} I(\nu) = I(X \wedge Y) &= H(X) - H(X|Y) \\ &\geq H(X) - H(X|Z) = H(X) - H(X - Z|Z) \\ &\geq H(X) - H(X - Z). \end{aligned} \tag{7.3}$$

The hypothesis on the estimator $\hat{\theta}$ implies that

$$E\|X - Z\|^2 = E(E\|X - Z\|^2|X) = \int E_\theta\|\theta - \hat{\theta}\|^2\nu(d\theta) \leq \varepsilon.$$

Hence, by Lemma 7.4 applied with $\sigma^2 = \varepsilon/k$,

$$H(X - Z) \leq \frac{k}{2}\log\frac{2\pi e\varepsilon}{k}\ .$$

On account of the inequality (7.3), where $H(X) = \log\lambda(\Theta)$, this completes the proof of the theorem. □

7.2 Optimality results

Returning to the problem of least possible worst case expected redundancy, it follows from Corollary 7.1 that for any class $\mathcal{P}$ of processes with alphabet A, there exists, for each n, a unique Q_n^* attaining the minimum in (7.1). As discussed before, this I-centroid of the set $\{P_n : P \in \mathcal{P}\}$ of the marginals on A^n of the processes in $\mathcal{P}$ is effectively an optimal coding distribution for n-length messages, tailored to the class $\mathcal{P}$. When $\mathcal{P}$ is a *parametric class of processes*, that is, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where Θ is a Borel subset of R^k, for some $k \geq 1$, and $P_\theta(a_1^n)$ is a measurable function of θ for each n and $a_1^n \in A^n$, Theorem 7.1 identifies the I-centroid Q_n^* as

$$Q_n^*(x_1^n) = \int P_{\theta,n}(x_1^n)\nu_n(d\theta), \qquad x_1^n \in A^n$$

where ν_n is a capacity achieving distribution for the channel determined by $\{P_{\theta,n}, \theta \in \Theta\}$ provided that a capacity achieving distribution exists; a sufficient condition for the latter is the closedness of the set $\{P_{\theta,n}, \theta \in \Theta\}$ of the marginal distributions on A^n, see Lemma 7.2.

Typically, ν_n does depend on n, and no process exists of which Q_n^* would be the marginal on A^n for $n = 1, 2, \ldots$ (a similar inconvenience occurred also in the context of Theorem 6.2). Still, for important process classes $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, there exists a probability measure ν on Θ not depending on n, such that the marginals $Q_n = \{Q(x_1^n), x_1^n \in A^n\}$ of the "mixture process" $Q = \int P_\theta \nu(d\theta)$ given by

$$Q(x_1^n) = \int P_{\theta,n}(x_1^n)\nu(d\theta), \qquad x_1^n \in A^n,\ n = 1, 2, \ldots \tag{7.4}$$

attain the minimum of $\sup_{\theta\in\Theta} D(P_{\theta,n}\|Q_n)$ within a constant. Then Q is a "nearly optimal coding process": the arithmetic code determined by Q attains the least possible worst case expected redundancy for $\mathcal{P}$, within a constant. Typical examples are the coding processes tailored to the classes of i.i.d. and Markov processes, treated in the previous subsection. We now show that these are mixture processes as in (7.4). Their "near optimality" will be proved later on.

First, let $\mathcal{P}$ be the class of i.i.d. processes with alphabet $A = \{1, \ldots, k\}$, parametrized by $\Theta = \{(p_1, \ldots, p_{k-1}) : p_i \geq 0,\ \sum_{i=1}^{k-1} p_i \leq 1\}$, with

$$P_\theta(x_1^n) = \prod_{i=1}^{k} p_i^{n_i}, \qquad n_i = |\{1 \leq t \leq n : x_t = i\}|;$$

here, for $\theta = (p_1, \ldots, p_{k-1})$, $p_k = 1 - (p_1 + \ldots + p_{k-1})$.

Let ν be the Dirichlet distribution on Θ with parameters $\alpha_i > 0$,

$i = 1, \ldots, k$, whose density function, with the notation above, is

$$f_{\alpha_1,\ldots,\alpha_k}(\theta) \stackrel{\text{def}}{=} \frac{\Gamma(\sum\limits_{i=1}^{k} \alpha_i)}{\prod\limits_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1},$$

where $\Gamma(s) = \int\limits_0^\infty x^{s-1} e^{-x} dx$. Then (7.4) gives

$$\begin{aligned} Q(x_1^n) &= \int\limits_\Theta P_\theta(x_1^n) f_{\alpha_1,\ldots,\alpha_k}(\theta) d\theta = \frac{\Gamma(\sum\limits_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int\limits_\Theta \prod_{i=1}^{k} p_i^{n_i + \alpha_i - 1} d\theta \\ &= \frac{\Gamma(\sum\limits_{i=1}^{k} \alpha_i)}{\prod\limits_{i=1}^{k} \Gamma(\alpha_i)} \cdot \frac{\prod\limits_{i=1}^{k} \Gamma(n_i + \alpha_i)}{\Gamma(\sum\limits_{i=1}^{k} (n_i + \alpha_i))} \cdot \int_\Theta f_{n_1+\alpha_1,\ldots,n_k+\alpha_k}(\theta)\, d\theta \\ &= \frac{\prod\limits_{i=1}^{k} [(n_i + \alpha_i - 1)(n_i + \alpha_i - 2) \ldots \alpha_i]}{(n + \sum\limits_{i=1}^{k} \alpha_i - 1)(n + \sum\limits_{i=1}^{k} \alpha_i - 2) \ldots (\sum\limits_{i=1}^{k} \alpha_i)}, \end{aligned}$$

where the last equality follows since the integral of a Dirichlet density is 1, and the Γ-function satisfies the functional equation $\Gamma(s+1) = s\Gamma(s)$. In particular, if $\alpha_1 = \ldots = \alpha_k = \frac{1}{2}$, the mixture process $Q = \int P_\theta \nu(d\theta)$ is exactly the coding process tailored to the i.i.d. class $\mathcal{P}$, see (6.5).

Next, let $\mathcal{P}$ the class of Markov chains with alphabet $A = \{1, \ldots, k\}$, with initial distribution equal to the uniform distribution on A, say, parametrized by $\Theta = \{(p_{ij})_{1\le i\le k, 1\le j\le k-1} \colon p_{ij} \ge 0, \sum\limits_{j=1}^{k-1} p_{ij} \le 1\}$:

$$P_\theta(x_1^n) = \frac{1}{|A|} \prod_{i=1}^{k} \prod_{j=1}^{k} p_{ij}^{n_{ij}}, \qquad n_{ij} = |\{1 \le t \le n-1 \colon x_t = i,\ x_{t+1} = j\}|,$$

where, for $\theta = (p_{ij})$, $p_{ik} = 1 - (p_{i1} + \ldots + p_{ik-1})$. Let ν be the Cartesian product of k Dirichlet $(\frac{1}{2}, \ldots, \frac{1}{2})$ distributions, that is, a distribution on Θ under which the rows of the matrix (p_{ij}) are independent and

Dirichlet $(\frac{1}{2},\ldots,\frac{1}{2})$ distributed. The previous result implies that the corresponding mixture process $Q = \int P_\theta \nu(d\theta)$ equals the coding process tailored to the Markov class $\mathcal{P}$, see (6.13).

Similarly, the coding process tailored to the class of m'th order Markov chains, see (6.14), or to its subclass determined by a context function, can also be represented as $Q = \int P_\theta \nu(d\theta)$, with ν equal to a Cartesian product of Dirichlet $(\frac{1}{2},\ldots,\frac{1}{2})$ distributions.

To prove "near optimality" of any code, a lower bound to $\overline{R}_n$ in equation (7.1) is required. Such bound can be obtained applying Theorems 7.1 and 7.3, with A^n in the role of A.

Theorem 7.4. Let $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ be a parametric class of processes, with $\Theta \subseteq R^k$ of positive Lebesgue measure, such that for some estimators $\hat{\theta}_n \colon A^n \mapsto \Theta$

$$E_\theta \|\theta - \hat{\theta}_n\|^2 \le \frac{c(\theta)}{n}, \quad \theta \in \Theta \;, n = 1, 2, \ldots .$$

Then, for a suitable constant K,

$$\overline{R}_n \ge \frac{k}{2} \log n - K, \quad n = 1, 2, \ldots .$$

Moreover, if $\lambda(\Theta) < +\infty$, then to any $\delta > 0$ there exists a constant K such that for each n and distribution Q_n on A^n

$$\lambda(\{\theta \in \Theta \colon D(P_{\theta,n} \| Q_n) < \frac{k}{2} \log n - K\}) < \delta.$$

Proof: It suffices to prove the second assertion. Fixing $0 < \delta \le \lambda(\Theta)$, take C so large that $\Theta' = \{\theta \in \Theta, c(\theta) > C\}$ has $\lambda(\Theta') \le \delta/2$. Then, for arbitrary $\Theta_1 \subseteq \Theta$ with $\lambda(\Theta_1) \ge \delta$, Theorem 7.3 applied to $\{P_{\theta,n} \colon \theta \in \Theta_1 \setminus \Theta'\}$ with $\varepsilon = C/n$ gives

$$I(\nu) \ge \frac{k}{2} \log \frac{kn}{2\pi e C} + \log \lambda(\Theta_1 \setminus \Theta')$$

where ν is the uniform distribution on $\Theta_1 \setminus \Theta'$.

Since here $\lambda(\Theta_1 \setminus \Theta') \ge \delta/2$, this and Theorem 7.1 applied to $\{P_{\theta,n} \colon \theta \in \Theta_1\}$ yield

$$\sup_{\theta \in \Theta_1} D(P_\theta \| Q_n) \quad \ge \quad I(\nu) \ge \frac{k}{2} \log \frac{kn}{2\pi e C} + \log \frac{\delta}{2}$$

$$= \quad \frac{k}{2}\log n - K; \quad K = \frac{k}{2}\log\frac{2\pi eC}{k} + \log\frac{2}{\delta}\ ,$$

whenever $\lambda(\Theta_1) \geq \delta$. This proves that the set $\{\theta \in \Theta\colon D(P_{\theta,n}\|Q_n) < \frac{k}{2}\log n - K\}$ cannot have Lebesgue measure $\geq \delta$, as claimed. □

Corollary 7.2. For $\mathcal{P}$ as above, if the expected redundancy of a prefix code $\{C_n\colon n = 1, 2, \ldots\}$ satisfies

$$E_P(R_{P,C_n}) - \frac{k}{2}\log n \to -\infty,\ P = P_\theta, \theta \in \Theta_0$$

for some subset Θ_0 of Θ then $\lambda(\Theta_0) = 0$.

Proof. Note that $E_P(R_{P,C_n}) - \frac{k}{2}\log n \to -\infty$ implies $D(P\|Q_n) - \frac{k}{2}\log n \to -\infty$ for the distributions Q_n associated with C_n by $Q_n(x_1^n) = c2^{-L(x_1^n)}$. Hence it suffices to show that for no $\Theta_0 \subseteq \Theta$ with $\lambda(\Theta_0) > 0$ can the latter limit relation hold for each $P = P_\theta$ with $\theta \in \Theta_0$.

Now, if such Θ_0 existed, with $\lambda(\Theta_0) = 2\delta$, say, Theorem 7.4 applied to Θ_0 in the role of Θ would give $\lambda(\{\theta \in \Theta_0, D(P_{\theta,n}\|Q_n) \geq \frac{k}{2}\log n - K\}) > \delta,\ n = 1, 2, \ldots$, contradicting $D(P_{\theta,n}\|Q_n) - \frac{k}{2}\log n \to -\infty, \theta \in \Theta_0$. □

Theorem 7.5. (i) For the class of i.i.d. processes with alphabet $A = \{1, \ldots, k\}$,

$$\frac{k-1}{2}\log n - K_1 \leq \overline{R}_n \leq R_n^* \leq \frac{k-1}{2}\log n + K_2,$$

where K_1 and K_2 are constants. The worst case maximum and expected redundancies $R_{C_n}^*$ and $\overline{R}_{C_n}$ of the arithmetic code determined by the coding process Q given by (6.5) are the best possible for any prefix code, up to a constant.

(ii) For the class of m'th order Markov chains with alphabet $A = \{1, \ldots, k\}$,

$$\frac{(k-1)k^m}{2}\log n - K_1 \leq \overline{R}_n \leq R_n^* \leq \frac{(k-1)k^m}{2}\log n + K_2$$

with suitable constants K_1 and K_2. The arithmetic code determined by the coding process Q given by (6.14) is nearly optimal in the sense of (i).

Proof. (i) The class $\mathcal{P}$ of i.i.d. processes satisfies the hypothesis of Theorem 7.4, with k replaced by $k-1$. Suitable estimators $\hat{\theta}_n$ are the natural ones: for $x_1^n \in A^n$ with empirical distribution $\hat{P} = (\hat{p}_1, \dots, \hat{p}_k)$, set $\hat{\theta}_n(x_1^n) = (\hat{p}_1, \dots, \hat{p}_{k-1})$. Thus the lower bound to $\overline{R}_n$ follows from Theorem 7.4. Combining this with the bound in (6.6) completes the proof.

(ii) To prove the lower bound to $\overline{R}_n$, consider the m'th order Markov chains with uniform initial distribution, say, restricting attention to the irreducible ones. The role of θ is now played by the $(k-1)k^m$-tuple of transition probabilities $P(j|a_1^m)$, $a_1^m \in A^m$, $j = 1, \dots, k-1$. It is not hard to see that estimating $P(j|a_1^m)$ from $x_1^m \in A^n$ by the ratio $n_{a_1^m j}/n_{a_1^m}$ (with the notation in equation (6.14)) gives rise to estimators $\hat{\theta}_n$ of θ that satisfy the hypothesis of Theorem 7.4, with $(k-1)k^m$ in the role of k. Then the claimed lower bound follows, and combining it with the bound in (6.16) completes the proof. □

Remark 7.1. Analogous results hold, with similar proofs, for any subclass of the m'th order Markov chains determined by a context function, see Subsection 6.2.

8

Redundancy and the MDL principle

Further results about redundancy for processes are discussed in this Section, with applications to statistical inference via the *minimum description length* (MDL) principle.

As in the last Sections, the term code means either an n-code $C_n\colon A^n \mapsto \{0,1\}^*$, or a sequence of n-codes $\{C_n\colon n = 1,2,\ldots\}$. Codes $\{C_n\colon n = 1,2,\ldots\}$ determined by a "coding process" Q will play a distinguished role. For convenience, we will use the term *Q-code* for an "ideal code" determined by Q, with length function $L(x_1^n) = -\log Q(x_1^n)$, whose redundancy function relative to a process P is

$$R(x_1^n) \;=\; \log \frac{P(x_1^n)}{Q(x_1^n)}.$$

The results below stated for such ideal codes are equally valid for real (Shannon or arithmetic) codes whose length and redundancy functions differ from those of the ideal Q-codes by less than 2 bits.

Theorem 8.1. If P and Q are mutually singular probability measures on A^∞, the P-redundancy of a Q-code goes to infinity, with P-probability 1.

Proof: Let $\mathcal{F}_n$ be the σ-algebra generated by the cylinder sets $[x_1^n]$,

$x_1^n \in A^n$. Then $\left\{ Z_n = \frac{Q(x_1^n)}{P(x_1^n)},\ n = 1, 2, \dots \right\}$ is a non-negative martingale with respect to the filtration $\{\mathcal{F}_n\}$, with underlying probability measure P, hence the almost sure limit

$$\lim_{n\to\infty} Z_n = Z \geq 0$$

exists. We have to show that $Z = 0$ (a.s.), or equivalently that $E(Z) = 0$.

By the singularity hypothesis, there exists a set $\tilde{A} \in \mathcal{F} = \sigma(\cup \mathcal{F}_n)$ such that $P(\tilde{A}) = 1$, $Q(\tilde{A}) = 0$. Define a measure μ by

$$\mu(B) = Q(B) + \int\limits_B Z dP, \qquad B \in \mathcal{F} .$$

Since $\mathcal{F} = \sigma(\cup \mathcal{F}_n)$, to every $\varepsilon > 0$ and sufficiently large m there exists $\tilde{A}_m \in \mathcal{F}_m$ such that the symmetric difference of $\tilde{A}$ and $\tilde{A}_m$ has μ-measure less than ε; thus,

$$Q(\tilde{A}_m) + \int_{\tilde{A}\setminus\tilde{A}_m} Z dP < \varepsilon .$$

Since the definition of Z_n implies $\int_{\tilde{A}_m} Z_n dP = Q(\tilde{A}_m)$ for $n \geq m$, Fatou's lemma gives

$$\int_{\tilde{A}_m} Z dP \leq \liminf_{n\to\infty} \int_{\tilde{A}_m} Z_n dP = Q(\tilde{A}_m) .$$

Combining these two bounds, we obtain

$$E(Z) = \int_{\tilde{A}_m} Z dP + \int_{\tilde{A}\setminus\tilde{A}_m} Z dP < \varepsilon .$$

Since $\varepsilon > 0$ was arbitrary, $E(Z) = 0$ follows. □

8.1 Codes with sublinear redundancy growth

While by Theorem 8.1 the redundancy of a Q-code relative to a process P typically goes to infinity, the next theorem gives a sufficient condition for a sublinear growth of this redundancy, that is, for the per letter redundancy to go to zero.

For this, we need the concept of *divergence rate*, defined for processes P and Q by

$$\overline{D}(P\|Q) = \lim_{n\to\infty} \frac{1}{n} D(P_n\|Q_n),$$

provided that the limit exists. The following lemma gives a sufficient condition for the existence of divergence rate, and a divergence analogue of the entropy theorem. For ergodicity, and other concepts used below, see the Appendix.

Lemma 8.1. Let P be an ergodic process and Q a Markov chain of order m with $D(P_{m+1}\|Q_{m+1}) < +\infty$. Then

$$\begin{aligned}\frac{1}{n}\log\frac{P(x_1^n)}{Q(x_1^n)} &\to \overline{D}(P\|Q)\\ &= -\overline{H}(P) - \textstyle\sum_{x_1^{m+1}\in A^{m+1}} P(x_1^{m+1})\log Q(x_{m+1} \mid x_1^m),\end{aligned}$$

both P-almost surely and in $L_1(P)$, with $Q(x_{m+1} \mid x_1^m)$ denoting the transition probabilities of the Markov chain Q.

Proof: Since Q is Markov of order m,

$$\log\frac{P(x_1^n)}{Q(x_1^n)} = \log P(x_1^n) - \log Q(x_1^m) - \sum_{i=1}^{n-m} \log Q(x_{m+i} \mid x_i^{m+i-1}),\ n \geq m;$$

here $\log Q(x_1^m)$ is finite with P-probability 1, and so is $\log Q(x_{m+1} \mid x_1^m)$, since $D(P_{m+1}\|Q_{m+1}) < +\infty$.

By the entropy theorem, and the ergodic theorem applied to $f(x_1^\infty) = \log Q(x_{m+1} \mid x_1^m)$, we have

$$\begin{gathered}\tfrac{1}{n}\log P(x_1^n) \to -\overline{H}(P)\\ \tfrac{1}{n}\log \textstyle\sum_{i=1}^{n-m} \log Q(x_{m+i} \mid x_i^{m+i-1}) \to E_P(\log Q(x_{m+1} \mid x_1^m)),\end{gathered}$$

both P-almost surely and in $L_1(P)$. The lemma follows. □

Theorem 8.2. Let P be an ergodic process, and let $Q = \int U_\vartheta \nu(d\vartheta)$ be a mixture of processes $\{U_\vartheta, \vartheta \in \Theta\}$ such that for every $\varepsilon > 0$ there exist an m and a set $\Theta' \subseteq \Theta$ of positive ν-measure with

$$U_\vartheta \text{ Markov of order } m \text{ and } \overline{D}(P\|U_\vartheta) < \varepsilon\ , \quad \text{if } \ \vartheta \in \Theta'.$$

Then for the process P, both the pointwise redundancy per symbol and the expected redundancy per symbol of the Q-code go to zero as $n \to \infty$, the former with probability 1.

Remark 8.1. Here Q need not be the mixture of a parametric class of processes, that is, unlike in Section 7, the index set Θ need not be a subset of an Euclidean space. It may be any set, endowed with a σ-algebra Σ such that $U_\vartheta(a_1^n)$ is a measurable function of ϑ for each $a_1^n \in A^n,\ n = 1, 2, \ldots$, and ν is any probability measure on (Θ, Σ). All subsets of Θ we consider are supposed to belong to Σ.

Proof of Theorem 8.2. We first prove for the pointwise redundancy per symbol that

$$\frac{1}{n} R(x_1^n) = \frac{1}{n} \log \frac{P(x_1^n)}{Q(x_1^n)} \to 0 \ , \quad P\text{-a.s.} \tag{8.1}$$

To establish this, on account of Theorem 6.1, it suffices to show that for every $\varepsilon > 0$

$$\limsup_{n\to\infty} \frac{1}{n} R(x_1^n) \le \varepsilon, \qquad P\text{-a.s..}$$

This will be established by showing that

$$\frac{2^{\varepsilon n} Q(x_1^n)}{P(x_1^n)} \to +\infty, \quad P\text{-a.s.}$$

Since

$$Q(x_1^n) = \int\limits_{\Theta} U_\vartheta(x_1^n)\nu(d\vartheta) \ge \int\limits_{\Theta'} U_\vartheta(x_1^n)\nu(d\vartheta), \tag{8.2}$$

we have

$$\frac{2^{\varepsilon n} Q(x_1^n)}{P(x_1^n)} \ge \int\limits_{\Theta'} \frac{2^{\varepsilon n} U_\vartheta(x_1^n)}{P(x_1^n)} \nu(d\vartheta) = \int\limits_{\Theta'} 2^{n\left(\varepsilon - \frac{1}{n}\log\frac{P(x_1^n)}{U_\vartheta(x_1^n)}\right)} \nu(d\vartheta).$$

If $\vartheta \in \Theta'$, Lemma 8.1 implies

$$\frac{1}{n} \log \frac{P(x_1^n)}{U_\vartheta(x_1^n)} \to \overline{D}(P \| U_\vartheta) < \varepsilon$$

for P-almost all $x_1^\infty \in A^\infty$ (the exceptional set may depend on ϑ). It follows that the set of pairs $(x_1^\infty, \vartheta) \in A^\infty \times \Theta'$ for which the last limit relation does not hold, has $P \times \nu$ -measure 0, and consequently for P-almost all $x_1^\infty \in A^\infty$ the set of those $\vartheta \in \Theta'$ for which that limit relation does not hold, has ν-measure 0 (both by Fubini's theorem).

Thus, for P-almost all x_1^∞, the integrand in the above lower bound to $2^{\varepsilon n} Q(x_1^n)/P(x_1^n)$ goes to infinity for ν-almost all $\vartheta \in \Theta'$. Hence, by Fatou's lemma, the integral itself goes to $+\infty$, completing the proof of (8.1).

To prove that also the expected redundancy per symbol $\frac{1}{n} E_P(R(x_1^n))$ goes to zero, we have to show that

$$\frac{1}{n} E_P(\log Q(x_1^n)) \to -H(\overline{P}).$$

On account of the entropy theorem, the result (8.1) is equivalent to

$$\frac{1}{n} \log Q(x_1^n) \to -H(\overline{P}) \qquad P\text{-a.s.},$$

hence it suffices to show that $\frac{1}{n} \log Q(x_1^n)$ is uniformly bounded (P-a.s.).

Since for $\vartheta \in \Theta'$ the Markov chains U_ϑ of order m satisfy $\overline{D}(P \| U_\vartheta) < \varepsilon$, their transition probabilities $U_\vartheta(x_{m+1} \mid x_1^m)$ are bounded below by some $\gamma > 0$ whenever $P(x_1^{m+1}) > 0$, see the expression of $\overline{D}$ in Lemma 8.1. This implies by (8.2) that $Q(x_1^n)$ is bounded below by a constant times γ^n, P-a.s. The proof of Theorem 8.2 is complete. □

Example 8.1. Let $Q = \sum_{m=0}^{\infty} \alpha_m Q^{(m)}$, where $\alpha_0, \alpha_1, \ldots$ are positive numbers with sum 1, and $Q^{(m)}$ denotes the process defined by equation (6.14) (in particular, $Q^{(0)}$ and $Q^{(1)}$ are defined by (6.5) and (6.13)). This Q satisfies the hypothesis of Theorem 8.2, for each ergodic process P, on account of the mixture representation of the processes $Q^{(m)}$ established in Subsection 7.2. Indeed, the divergence rate formula in Lemma 8.1 implies that $\overline{D}(P \| U_\vartheta) < \varepsilon$ always holds if U_ϑ is a Markov chain of order m whose transition probabilities $U(x_{m+1} \mid x_1^m)$ are sufficiently close to the conditional probabilities $\text{Prob}\{X_{m+1} = x_{m+1} \mid X_1^m = x_1^m\}$ for a representation $\{X_n\}$ of the process P, with m so large that $H(X_{m+1} \mid X_1^m) < \overline{H}(P) + \varepsilon/2$, say. It follows by Theorem 8.2 that the Q-code with

$Q = \sum_{m=0}^{\infty} \alpha_m Q^{(m)}$ is weakly universal for the class of ergodic processes, in the sense of (6.1), and also its pointwise redundancy per symbol goes to zero P-a.s., for each ergodic process P.

Recall that the weak universality of this Q-code has already been established in Subsection 6.2, even for the class of all stationary processes.

Example 8.2. Let $\{U_\gamma : \gamma \in \Gamma\}$ be a countable family of Markov processes (of arbitrary orders), such that for each ergodic process P,

$$\inf_{\gamma \in \Gamma} D(P \| U_\gamma) = 0 \; . \tag{8.3}$$

Then for arbitrary numbers $\alpha_\gamma > 0$ with $\sum \alpha_\gamma = 1$, the process $Q = \sum_{\gamma \in \Gamma} \alpha_\gamma U_\gamma$ satisfies the conditions of Theorem 8.2, for every ergodic process P. Hence the Q-code is weakly universal for the class of ergodic processes. Note that the condition (8.3) is satisfied, for example, if the family $\{U_\gamma : \gamma \in \Gamma\}$ consists of all those Markov processes, of all orders, whose transition probabilities are rational numbers.

While the last examples give various weakly universal codes for the class of ergodic processes, the next example shows the non-existence of strongly universal codes for this class.

Example 8.3. Associate with each $a_1^m \in A^m$ a process P, the probability measure on A^∞ that assigns weights $1/m$ to the infinite sequences $a_i^m a_1^m a_1^m \ldots, i = 1, \ldots, m$. Clearly, this P is an ergodic process. Let $\mathcal{P}^{(m)}$ denote the class of these processes for all $a_1^m \in A^m$. We claim that for the class $\mathcal{P}$ equal to the union of the classes $\mathcal{P}^{(m)}, m = 1, 2, \ldots$

$$\overline{R}_n = \inf_{Q_n} \sup_{P \in \mathcal{P}} D(P_n \| Q_n),$$

see equation (7.1), is bounded below by $n \log |A| - \log n$.

Denote by $P_{a_1^m}$ the marginal on A^m of the process P associated with a_1^m as above, and by ν_m the uniform distribution on A^m. Since $\mathcal{P}^{(n)}$ is a subset of $\mathcal{P}$, Theorem 7.1 implies

$$\overline{R}_n \geq \inf_{Q_n} \sup_{P \in \mathcal{P}^{(n)}} D(P_n \| Q_n) \geq I(\nu_n)$$

$$= H\left(\tfrac{1}{|A|^n}\textstyle\sum_{a_1^n\in A^n} P_{a_1^n}\right) - \tfrac{1}{|A|^n}\textstyle\sum_{a_1^n\in A^n} H(P_{a_1^n}).$$

As $P_{a_1^n}$ is concentrated on the cyclic shifts of a_1^n, implying $H(P_{a_1^n}) \le \log n$, and the "output distribution" $|A|^{-n}\sum_{a_1^n\in A^n} P_{a_1^n}$ equals the uniform distribution on A^n, this establishes our claim. In particular, no strongly universal codes exist for the class $\mathcal{P}$, let alone for the larger class of all ergodic processes.

Next we consider a simple construction of a new code from a given (finite or) countable family of codes $\{C_\gamma, \gamma \in \Gamma\}$, where $C_\gamma = \{C_n^\gamma\colon A^n \mapsto B^*,\ n = 1, 2, \ldots\},\ B = \{0, 1\}$. Let the new code assign to each $x_1^n \in A^n$ one of the codewords $C_n^\gamma(x_1^n)$, with $\gamma \in \Gamma$ chosen depending on x_1^n, preambled by a code $C(\gamma)$ of the chosen $\gamma \in \Gamma$. Here $C\colon \Gamma \mapsto B^*$ can be any prefix code; the preamble $C(\gamma)$ is needed to make the new code decodable. We assume that γ above is chosen optimally, that is, to minimize $L(\gamma) + L_\gamma(x_1^n)$, where $L_\gamma(x_1^n)$ and $L(\gamma)$ denote the length functions of the codes C_γ and C. Then the new code has length function

$$L(x_1^n) = \min_{\gamma\in\Gamma}[L(\gamma) + L_\gamma(x_1^n)].$$

If the family $\{C_\gamma, \gamma \in \Gamma\}$ consists of Q_γ-codes for a list of processes $\{Q_\gamma, \gamma \in \Gamma\}$, the code constructed above will be referred to as generated by that list.

Lemma 8.2. A code generated by a list of processes $\{Q_\gamma, \gamma \in \Gamma\}$ is effectively as good as a Q-code for a mixture Q of these processes, namely its length function satisfies

$$-\log Q^{(1)}(x_1^n) \le L(x_1^n) \le -\log Q^{(2)}(x_1^n) + \log c_2,$$

where

$$Q^{(1)} = c_1 \sum_{\gamma\in\Gamma} 2^{-L(\gamma)} Q_\gamma\ , \qquad Q^{(2)} = c_2 \textstyle\sum_{\gamma\in\Gamma} 2^{-2L(\gamma)} Q_\gamma,$$

$$c_1 = \left(\sum_{\gamma\in\Gamma} 2^{-L(\gamma)}\right)^{-1}, \qquad c_2 = \left(\textstyle\sum_{\gamma\in\Gamma} 2^{-2L(\gamma)}\right)^{-1}.$$

Proof. The Q_γ-code C_γ has length function $L_\gamma(x_1^n) = -\log Q_\vartheta(x_1^n)$, hence

$$L(x_1^n) = \min_{\gamma\in L}[L(\gamma) - \log Q_\gamma(x_1^n)] = -\log \max_{\gamma\in L} 2^{-L(\gamma)} Q_\gamma(x_1^n).$$

Since

$$\begin{aligned} Q^{(1)}(x_1^n) \geq \sum_{\gamma\in\Gamma} 2^{-L(\gamma)} Q_\gamma(x_1^n) &\geq \max_{\gamma\in L} 2^{-L(\gamma)} Q_\gamma(x_1^n) \\ &\geq \sum_{\gamma\in\Gamma} 2^{-2L(\gamma)} Q_\gamma(x_1^n) = \frac{Q^{(2)}(x_1^n)}{c_2} , \end{aligned}$$

where the first and third inequalities are implied by Kraft's inequality $\sum_{\gamma\in\Gamma} 2^{-L(\gamma)} \leq 1$, the assertion follows. □

Recalling Examples 8.1 and 8.2, it follows by Lemma 8.2 that the list of processes $\{Q^{(m)},\ m = 0, 1, \ldots\}$, with $Q^{(m)}$ defined by equation (6.14), as well as any list of Markov processes $\{U_\gamma, \gamma \in \Gamma\}$ with the property (8.3), generates a code such that for every ergodic process P, the redundancy per symbol goes to 0 P-a.s., and also the mean redundancy per symbol goes to 0.

8.2 The minimum description length principle

The idea of the above construction of a new code from a given (finite or countable) family of codes underlies also the *minimum description length* (MDL) *principle* of statistical inference that we discuss next.

MDL principle. *The statistical information in data is best extracted when a possibly short description of the data is found. The statistical model best fitting to the data is the one that leads to the shortest description, taking into account that the model itself must also be described.*

Formally, in order to select a statistical model that best fits the data x_1^n, from a list of models indexed with elements γ of a (finite or) countable set Γ, one associates with each candidate model a code C_γ, with length function $L_\gamma(x_1^n)$, and takes a code $C\colon \Gamma \mapsto B^*$ with length function $L(\gamma)$ to describe the model. Then, according to the MDL principle, one adopts that model for which $L(\gamma) + L_\gamma(x_1^n)$ is minimum.

For a simple model stipulating that the data are coming from a specified process Q_γ, the associated code C_γ is a Q_γ-code with length function $L_\gamma(x_1^n) = -\log Q_\gamma(x_1^n)$. For a composite model stipulating that the data are coming from a process in a certain class, the associated code C_γ should be universal for that class, but the principle admits a freedom in its choice. There is also a freedom in choosing the code $C\colon \Gamma \mapsto B^*$.

To relate the MDL to other statistical principles, suppose that the candidate models are parametric classes $\mathcal{P}_\gamma = \{P_\vartheta, \vartheta \in \Theta_\gamma\}$ of processes, with γ ranging over a (finite or) countable set Γ. Suppose first that the code C_γ is chosen as a Q_γ-code with

$$Q_\gamma = \int\limits_{\Theta_\gamma} P_\vartheta \nu_\gamma(d\vartheta), \tag{8.4}$$

where ν_γ is a suitable probability measure on Θ_γ, see Subsection 7.2. Then MDL inference by minimizing $L(\gamma) + L_\gamma(x_1^n) = L(\gamma) - \log Q_\gamma(x_1^n)$ is equivalent to Bayesian inference by maximizing the *posterior probability* (conditional probability given the data x_1^n) of γ, if one assigns to each $\gamma \in \Gamma$ a prior probability proportional to $2^{-L(\gamma)}$, and regards ν_γ as a prior probability distribution on Θ_γ. Indeed, with this choice of the priors, the posterior probability of γ is proportional to $2^{-L(\gamma)} Q_\gamma(x_1^n)$.

Suppose next that the codes C_γ associated with the models $\mathcal{P}_\gamma$ as above are chosen to be NML codes, see Theorem 6.2, with length functions

$$L_\gamma(x_1^n) = -\log NML_\gamma(x_1^n) = -\log P_{\mathrm{ML}}^{(\gamma)}(x_1^n) + \log \sum_{a_1^n \in A^n} P_{\mathrm{ML}}^{(\gamma)}(a_1^n),$$

where

$$P_{\mathrm{ML}}^{(\gamma)}(x_1^n) = \sup_{\vartheta \in \Theta_\gamma} P_\vartheta(x_1^n)\ .$$

Then the MDL principle requires minimizing

$$L(\gamma) + L_\gamma(x_1^n) = -\log P_{\mathrm{ML}}^{(\gamma)}(x_1^n) + R_n(\gamma)$$

where

$$R_n(\gamma) = L(\gamma) + \log \sum_{a_1^n \in A^n} P_{\mathrm{ML}}^{(\gamma)}(a_1^n) \ .$$

In statistical terminology, this is an instance of penalized maximum likelihood methods, that utilize maximization of $\log P_{\mathrm{ML}}^{(\gamma)}(x_1^n) - R_n(\gamma)$, where $R_n(\gamma)$ is a suitable "penalty term".

Remark 8.2. We note without proof that, under suitable regularity conditions, $L(\gamma) + L_\gamma(x_1^n)$ is asymptotically equal (as $n \to \infty$) to $-\log P_{\mathrm{ML}}^{(\gamma)}(x_1^n) + \frac{1}{2}k_\gamma \log n$, for both of the above choices of the codes C_γ, where k_γ is the dimension of the model $\mathcal{P}_\gamma$ (meaning that Θ_γ is a subset of positive Lebesgue measure of R^{k_γ}). When Γ is finite, this admits the conclusion that MDL is asymptotically equivalent to penalized maximum likelihood with the so-called BIC (Bayesian information criterion) penalty term, $R_n(\gamma) = \frac{1}{2}k_\gamma \log n$. This equivalence, however, need not hold when Γ is infinite, as we see later.

The next theorems address the consistency of MDL inference, namely, whether the true model is always recovered, eventually almost surely, whenever one of the candidate models is true.

Theorem 8.3. Let $\{Q_\gamma, \gamma \in \Gamma\}$ be a (finite or) countable list of mutually singular processes, and let $L(\gamma)$ be the length function of a prefix code $C \colon \Gamma \mapsto B^*$. If the true process P is on the list, say $P = Q_{\gamma^*}$, the unique minimizer of $L(\gamma) - \log Q_\gamma(x_1^n)$ is γ^*, eventually almost surely as $n \to \infty$.

Remark 8.3. The singularity hypothesis is always satisfied if the processes $Q_\gamma, \gamma \in \Gamma$, are (distinct and) ergodic.

Proof. Consider the mixture process

$$Q = c \sum_{\gamma \in \Gamma \setminus \{\gamma^*\}} 2^{-L(\gamma)} Q_\gamma$$

where $c > 1$ (due to Kraft's inequality). Then

$$Q(x_1^n) \geq \sum_{\gamma \in \Gamma \setminus \{\gamma^*\}} 2^{-L(\gamma)} Q_\gamma \geq \max_{\gamma \in \Gamma \setminus \{\gamma^*\}} 2^{-L(\gamma)} Q_\gamma(x_1^n) \ .$$

The hypothesis implies that Q and Q_{γ^*} are mutually singular, hence by Theorem 8.1

$$\log Q_{\gamma^*}(x_1^n) - \log Q(x_1^n) \to +\infty \qquad Q_{\gamma^*} \ - a.s.$$

This and the previous inequality complete the proof. □

Theorem 8.4. Let $\{\mathcal{P}_\gamma, \gamma \in \Gamma\}$ be a (finite or) countable list of parametric classes $\mathcal{P}_\gamma = \{P_\vartheta, \vartheta \in \Theta_\gamma\}$ of processes, let $Q_\gamma, \gamma \in \Gamma$, be mixture processes as in equation (8.4), supposed to be mutually singular, and let $L(\gamma)$ be the length function of a prefix code $C\colon \Gamma \mapsto B^*$. Then, with possible exceptional sets $N_\gamma \subset \Theta_\gamma$ of ν_γ-measure 0, if the true process is a non-exceptional member of either class $\mathcal{P}_\gamma$, say $P = Q_\vartheta, \vartheta \in \Theta_{\gamma^*} \setminus N_{\gamma^*}$, the unique minimizer of $L(\gamma) - \log Q_\gamma(x_1^n)$ is γ^*, eventually almost surely as $n \to \infty$.

Remark 8.4. A necessary condition for the singularity hypothesis is the essential disjointness of the classes $\mathcal{P}_\gamma, \gamma \in \Gamma$, that is, that for no $\gamma \neq \gamma'$ can $\Theta_\gamma \cap \Theta_{\gamma'}$ be of positive measure for both ν_γ and $\nu_{\gamma'}$. This condition is also sufficient if all processes P_ϑ are ergodic, and processes with different indices are different.

Proof of Theorem 8.4. By Theorem 8.3, the set of those $x_1^\infty \in A^\infty$ for which there exist infinitely many n with

$$L(\gamma^*) - \log Q_{\gamma^*}(x_1^n) \geq \inf_{\gamma \in \Gamma \setminus \{\gamma^*\}} [L(\gamma) - \log Q_\gamma(x_1^n)]$$

has Q_{γ^*}-measure 0, for any $\gamma^* \in \Gamma$. By the definition of Q_{γ^*}, see (8.4), this implies that the above set has P_ϑ-measure 0 for all $\vartheta \in \Theta_{\gamma^*}$ except possibly for ϑ in a set N_{γ^*} of μ_{γ^*}-measure 0. □

As an application of Theorem 8.4, consider the estimation of the order of a Markov chain, with alphabet $A = \{1, \ldots, k\}$. As in Example 8.1, denote by $Q^{(m)}$ the coding process tailored to the class of Markov chains of order m. According to the MDL principle, given a sample $x_1^n \in A^n$ from a Markov chain P of unknown order m^*, take the minimizer $\widehat{m} = \widehat{m}(x_1^n)$ of $L(m) - \log Q^{(m)}(x_1^n)$ as an estimate of m^*, where $L(\cdot)$ is the length function of some prefix code $C\colon N \mapsto B^*$.

Recall that $Q^{(m)}$ equals the mixture of m'th order Markov chains with uniform initial distribution, with respect to a probability distribution which is mutually absolutely continuous with the Lebesgue measure on the parameter set Θ_m, the subset of $k^m(k-1)$ dimensional Euclidean space that represents all possible transition probability matrices of m-th order Markov chains. It is not hard to see that the processes $Q^{(m)}$, $m = 0, 1\dots$ are mutually singular, hence Theorem 8.4 implies that

$$\widehat{m}(x_1^n) = m^* \quad \text{eventually almost surely,} \tag{8.5}$$

unless the transition probability matrix of the true P corresponds to some $\vartheta \in N_{m^*}$ where $N_{m^*} \subset \Theta_{m^*}$ has Lebesgue measure 0. (Formally, this follows for Markov chains P with uniform initial distribution, but events of probability 1 for a Markov chain P with uniform initial distribution clearly have probability 1 for all Markov chains with the same transition probabilities as P.)

Intuitively, the exceptional sets $N_m \subset \Theta_m$ ought to contain all parameters that do not represent irreducible chains, or represent chains of smaller order than m. It might appear a plausible conjecture that the exceptional sets N_m are thereby exhausted, and the consistency assertion (8.5) actually holds for every irreducible Markov chain of order exactly m^*. Two results (stated without proof) that support this conjecture are that for Markov chains as above, the MDL order estimator with a prior bound to the true order, as well as the BIC order estimator with no prior order bound, are consistent. In other words, equation (8.5) will always hold if $\widehat{m}(x_1^n)$ is replaced either by the minimizer of $L(m) - \log Q^{(m)}(x_1^n)$ subject to $m \le m_0$, where m_0 is a known upper bound to the unknown m^*, or by the minimizer of

$$-\log P_{\mathrm{ML}}^{(m)}(x_1^n) + \frac{1}{2}k^m(k-1)\log n \ .$$

Nevertheless, the conjecture is false, and we conclude this subsection by a counterexample. It is unknown whether other counterexamples also exist.

Example 8.4. Let P be the i.i.d. process with uniform distribution, that is,

$$P(x_1^n) = k^{-n}, \quad x_1^n \in A^n, \quad A = \{1, \dots, k\}.$$

Then $m^* = 0$, and as we will show,

$$L(0) - \log Q^{(0)}(x_1^n) > \inf_{m>0}[L(m) - \log Q^{(m)}(x_1^n)], \quad \text{eventually a.s.,} \tag{8.6}$$

provided that $L(m)$ grows sublinearly with m, $L(m) = o(m)$. This means that (8.5) is false in this case. Actually, using the consistency result with a prior bound to the true order, stated above, it follows that $\widehat{m}(x_1^n) \to +\infty$, almost surely.

To establish equation (8.6), note first that

$$-\log Q^{(0)}(x_1^n) = -\log P_{\mathrm{ML}}^{(0)}(x_1^n) + \frac{k-1}{2}\log n + O(1),$$

where the $O(1)$ term is uniformly bounded for all $x_1^n \in A^*$. Here

$$P_{\mathrm{ML}}^{(0)}(x_1^n) = \sup_{\{p_1,\dots,p_k\}} \prod_{i=1}^{k} p_i^{n_i} = \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i}$$

is the largest probability given to x_1^n by i.i.d. processes, with n_i denoting the number of times the symbol i occurs in x_1^n, and the stated equality holds since $P_{\mathrm{ML}}^{(0)}(x_1^n)/Q^{(0)}(x_1^n)$ is bounded both above and below by a constant times $n^{\frac{k-1}{2}}$, see Remark 6.2, after Theorem 6.3.

Next, since P is i.i.d. with uniform distribution, the numbers n_i above satisfy, as $n \to \infty$,

$$n_i = \frac{n}{k} + O(\sqrt{n \log\log n}\ ), \quad \text{eventually a.s.,}$$

by the law of iterated logarithm. This implies

$$-\log P_{\mathrm{ML}}^{(0)}(x_1^n) = \sum_{i=1}^{k} n_i \log\left(\frac{n}{n_i}\right) = n\log k + O(\log\log n),$$

since

$$\begin{aligned} \log\frac{n}{n_i} &= \log k + \log\left(1 + \frac{n}{kn_i} - 1\right) = \\ &= \log k + \left(\frac{n}{kn_i} - 1\right)\log e + O\left(\frac{n}{kn_i} - 1\right)^2. \end{aligned}$$

It follows that the left hand side of equation (8.6) equals $n\log k + \frac{k-1}{2}\log n + O(\log\log n)$, eventually almost surely as $n \to \infty$.

Turning to the right hand side of equation (8.6), observe that if no m-block $a_1^m \in A^m$ occurs in x_1^{n-1} more than once then $Q^{(m)}(x_1^n) = k^{-n}$. Indeed, then $n_{a_1^m}$ is non-zero for exactly $n-m$ blocks $a_1^m \in A^m$ in the definition (6.14) of $Q^{(m)}$; for these, $n_{a_1^m} = 1$ and there is exactly one $j \in A$ with $n_{a_1^m j}$ nonzero, necessarily with $n_{a_1^m j} = 1$. Hence equation (6.14) gives $Q^{(m)}(x_1^n) = k^{-n}$ as claimed.

The probability that there is an m-block occurring in x_1^{n-1} more than once is less than $n^2 k^{-m}$. To see this, note that for any $1 \leq j < \ell < n-m+1$, the conditional probability of $x_j^{j+m-1} = x_\ell^{\ell+m-1}$, when $x_1^{\ell-1} \in A^{\ell-1}$ is fixed, is k^{-m}, as for exactly one of the k^m equiprobable choices of $x_\ell^{\ell+m-1} \in A^m$ will $x_\ell^{\ell+m-1} = x_j^{j+m-1}$ hold. Hence also the unconditional probability of this event is k^{-m}, and the claim follows. In particular, taking $m_n = \frac{4}{\log k} \log n$, the probability that some m_n-block occurs in x_1^{n-1} more than once is less than n^{-2}. By Borel–Cantelli, and the previous observation, it follows that

$$-\log Q^{(m_n)}(x_1^n) = n \log k, \quad \text{eventually a.s.}$$

This, and the assumption $L(m) = o(m)$, imply that the right hand side of (8.6) is $\leq n \log k + o(\log n)$, eventually almost surely, completing the proof of equation (8.6).

APPENDIX

A

Summary of process concepts

A (stochastic) process is frequently defined as a sequence of random variables $\{X_n\}$; unless stated otherwise, we assume that each X_n takes values in a fixed finite set A called the alphabet. The n-fold joint distribution of the process is the distribution P_n on A^n defined by the formula

$$P_n(x_1^n) = \mathrm{Prob}(X_i = x_i,\ 1 \le i \le n), \qquad x_1^n \in A^n.$$

For these distributions, the *consistency conditions*

$$P_n(x_1^n) = \sum_{a \in A} P_{n+1}(x_1^n a)$$

must hold. The process $\{X_n\}$, indeed, any sequence of distributions P_n on A^n, $n = 1, 2, \ldots$ that satisfies the consistency conditions, determines a unique Borel probability measure P on the set A^∞ of infinite sequences drawn from A such that each cylinder set $[a_1^n] = \{x_1^\infty : x_1^n = a_1^n\}$ has P-measure $P_n(a_1^n)$; a Borel probability measure on A^∞ is a probability measure defined on the σ-algebra $\mathcal{F}$ of Borel subsets of A^∞, the smallest σ-algebra containing all cylinder sets.

The probability space on which the random variables X_n are defined is not important, all that matters is the sequence of joint distributions

P_n. For this reason, a process can also be defined as a sequence of distributions P_n on A^n, $n = 1, 2, \ldots$, satisfying the consistency conditions, or as a probability measure P on $(A^\infty, \mathcal{F})$. In this tutorial we adopt the last definition: by a process P we mean a Borel probability measure on A^∞. The probabilities $P_n(a_1^n) = P([a_1^n])$ will be usually denoted briefly by $P(a_1^n)$.

A sequence of random variables $\{X_n\}$ whose n-dimensional joint distributions equal the n-dimensional marginals P_n of P, will be referred to as a representation of the process P. Such a representation always exists, for example the Kolmogorov representation, with X_n defined on the probability space $(A^\infty, \mathcal{F}, P)$ by $X_n(x_1^\infty) = x_i$, $i = 1, 2, \ldots$.

A process P is stationary if P is invariant under the shift T, the transformation on A^∞ defined by the formula $Tx_1^\infty = x_2^\infty$. Thus P is stationary if and only if $P(T^{-1}A) = P(A)$, $A \in \mathcal{F}$.

The *entropy rate* of a process P is defined as

$$\overline{H}(P) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n),$$

provided that the limit exists, where $\{X_n\}$ is a representation of the process P. A stationary process P has entropy rate

$$\overline{H}(P) = \lim_{n\to\infty} H(X_n|X_1, \ldots, X_{n-1});$$

here the limit exists since stationarity implies that

$$H(X_n|X_1, \ldots, X_{n-1}) = H(X_{n+1}|X_2, \ldots, X_n) \geq H(X_{n+1}|X_1, \ldots, X_n),$$

and the claimed equality follows by the additivity of entropy,

$$H(X_1, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_1, \ldots, X_{i-1}).$$

If $\{P_\vartheta \in \Theta\}$ is a family processes, with ϑ ranging over an arbitrary index set Θ endowed with a σ-algebra Σ such that $P_\vartheta(a_1^n) = P_\vartheta([a_1^n])$ is a measurable function of ϑ for each $a_1^n \in A^n$, $n = 1, 2, \ldots$, the *mixture* of the processes P_ϑ with respect to a probability measure μ on (Θ, Σ) is the process $P = \int P_\vartheta \mu(d\vartheta)$ defined by the formula

$$P(a_1^n) = \int P_\vartheta(a_1^n)\mu(d\vartheta), \qquad a_1^n \in A^n, \; n = 1, 2, \ldots \; .$$

A process P is called *ergodic* if it is stationary and, in addition, no non-trivial shift-invariant sets exist (that is, if $A \in \mathcal{F}$, $T^{-1}A = A$, then $P(A) = 0$ or 1), or equivalently, P cannot be represented as the mixture $P = \alpha P_1 + (1 - \alpha)P_2$ of two stationary processes $P_1 \neq P_2$ (with $0 < \alpha < 1$). Each stationary process is representable as a mixture of ergodic processes (by the so-called *ergodic decomposition theorem*). Other key facts about ergodic processes, needed in Section 8, are the following:

Ergodic theorem. For an ergodic process P, and P-integrable function f on A^∞,

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i^\infty) \to \int f dP,$$

both P-almost surely and in $L_1(P)$.

Entropy theorem. (Shannon–McMillan–Breiman theorem) For an ergodic process P,

$$-\frac{1}{n}\log P(x_1^n) \to \overline{H}(P),$$

both P-almost surely and in $L_1(P)$.

For an ergodic process P, almost all infinite sequences $x_1^\infty \in A^\infty$ are P-typical, that is, the "empirical probabilities"

$$\hat{P}(a_1^k|x_1^n) = \frac{1}{n-k+1}|\{i : x_{i+1}^{i+k} = a_1^k,\ 0 \leq i \leq n-k\}|$$

of k-blocks $a_1^k \in A^k$ in x_1^n approach the true probabilities $P(a_1^k)$ as $n \to \infty$, for each $k \geq 1$ and $a_1^k \in A^k$. This follows applying the ergodic theorem to the indicator functions of the cylinder sets $[a_1^k]$ in the role of f. Finally, we note that also conversely, if P-almost all $x_1^\infty \in A^\infty$ are P-typical then the process P is ergodic.

Historical Notes

Section 1. Information theory was created by Shannon [44]. The information measures entropy, conditional entropy and mutual information were introduced by him. A formula similar to Shannon's for entropy in the sense of statistical physics dates back to Boltzmann [4]. Information divergence was used as a key tool but had not been given a name in Wald [49]; it was introduced as an information measure in Kullback and Leibler [33]. Theorem 1.1 is essentially due to Shannon [44], Theorem 1.2 is drawn from Rissanen [37]. Arithmetic coding, whose origins are commonly attributed to unpublished work of P. Elias, was developed to a powerful data compression technique primarily by Rissanen, see [35], [40].

Section 2. The combinatorial approach to large deviations and hypothesis testing originates in Sanov [42] and Hoeffding [27]. A similar approach in statistical physics goes back to Boltzmann [4]. The method of types emerged as a major technique of information theory in Csiszár and Körner [15]. "Stein's lemma" appeared in Chernoff [6], attributed to C. Stein. The theory of sequential tests has been developed by Wald

[49]; the error bounding idea in Remark 2.2 appears there somewhat implicitly.

Section 3. Kullback [32] suggested I-divergence minimization as a principle of statistical inference, and proved special cases of several results in this Section. Information projections were systematically studied in Čencov [5] , see also Csiszár [12], Csiszár and Matúš [16]. In these references, distributions on general alphabets were considered; our finite alphabet assumption admits a simplified treatment. The characterization of the closure of an exponential family mentioned in Remark 3.1 is a consequence of a general result in [16] for exponential families whose domain of parameters is the whole R^k; the last hypothesis is trivially satisfied in the finite alphabet case.

The remarkable analogy of certain information theoretic concepts and results to geometric ones, instrumental in this Section and later on, has a profound background in a differential geometric structure of probability distributions, beyond the scope of this tutorial, see Čencov [5], Amari [2].

Section 4. f-Divergences were introduced by Csiszár [10], [11], and independently by Ali and Silvey [1]; see also the book Liese and Vajda [34]. A proof that the validity of Lemma 4.2 characterizes I-divergence within the class of f-divergences appears in Csiszár [14]. Theorem 4.2 can be regarded as a special case of general results about likelihood ratio tests, see Cox and Hinkley, [9, Section 9.3]; this special case, however, has admitted a simple proof. For the information theoretic approach to the analysis of contingency tables see Kullback [32], Gokhale and Kullback [26].

Section 5. Iterative scaling has long been used in various fields, primarily in the two-dimensional case as an intuitive method to find a non-negative matrix with prescribed row and column sums, "most similar" to a previously given non-negative matrix; the first reference known to us is Kruithof [31]. Its I-divergence minimizing feature was pointed out in Ireland and Kullback [28], though with an incomplete convergence proof. The proof here, via Theorem 5.1, is by Csiszár [12].

Generalized iterative scaling is due to Darroch and Ratcliff [19]. Its geometric interpretation admitting the convergence proof via Theorem 5.1 is by Csiszár [13]. Most results in Subsection 3.2 are from Csiszár and Tusnády [18], where the basic framework is applied also to other problems such as capacity and reliability function computation for noisy channels. The portfolio optimizing algorithm in Remark 4.3 is due to Cover [7]. The EM algorithm has been introduced by Dempster, Laird and Rubin [23].

Section 6. Universal coding was first addressed by Fitingof [25], who attributed the idea to Kolmogorov. An early theoretical development is Davisson [20]. Theorem 6.1 is by Barron [3], and Theorem 6.2 is by Shtarkov [46]. The universal code for the i.i.d class with coding process defined by equation (6.3) appears in Krichevsky and Trofimov [30] and in Davisson, McEliece, Pursley and Wallace [22]. Our proof of Theorem 6.3 follows [22]. Theorem 6.6 is due to Shtarkov [46]. The construction of "twice universal" codes via mixing (or "weighting") as in Remark 6.2 was suggested by Ryabko [41]. The context weighting algorithm mentioned in Remark 6.2 was developed by Willems, Shtarkov and Tjalkens [50].

Section 7. The approach here follows, though not in the details, Davisson and Leon–Garcia [21]. Lemma 7.1 dates back to Topsœ[47]. The first assertion of Theorem 7.2 appears in [21] (crediting R. Gallager for an unpublished prior proof), with a proof using the minimax theorem; see also (for Θ finite) Csiszár and Körner [15], p.147, and the references there. Theorem 7.4 and Corollary 7.2 are based on ideas of Davisson, McEliece, Pursley and Wallace [22] and of Rissanen [38]. For early asymptotic results on worst case redundancy as in Theorem 7.5, see Krichevski [29] (i.i.d.case) and Trofimov [48] (Markov case); the latter reference attributes the upper bound to Shtarkov.

Section 8. The main results Theorems 8.1–8.4 are due to Barron [3]. While Examples 8.1 and 8.2 give various weakly universal codes for the class of ergodic processes, those most frequently used in practice (the Lempel–Ziv codes, see [51]) are not covered here. The MDL principle of statistical inference has been proposed by Rissanen, see [36], [39].

The BIC criterion was introduced by Schwarz [43]. The consistency of the BIC Markov order estimator was proved, assuming a known upper bound to the order, by Finesso [24], and without that assumption by Csiszár and Shields [17]. The counterexample to the conjecture on MDL consistency suggested by Theorem 8.4 is taken from [17].

Appendix. For details on the material summarized here see, for example, the first Section of the book Shields [45].

References

[1] S. M. Ali and S. D. Silvey, “A general class of coefficients of divergence of one distribution from another”, *J. Royal Stat. Soc., Ser.B*, vol. 28, pp. 131-142, 1966.

[2] S. Amari, *Differential-Geometrical Methods in Statistics.* NewYork: Springer, 1985.

[3] A. Barron, *Logically smooth density estimation.* PhD thesis, Stanford Univ., 1985.

[4] L. Boltzmann, “Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht,” *Wien. Ber.*, vol. 76, pp. 373–435, 1877.

[5] N. Čencov, *Statistical Decision Rules and Optimal Inference.* Providence: Amer.Math.Soc. Russian original: Moscow: Nauka, 1972.

[6] H. Chernoff, “A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations,” *Annals Math. Statist.*, vol. 23, pp. 493–507, 1952.

[7] T. Cover, “An algorithm for maximizing expected log investment return,” *IEEE Trans. Inform. Theory*, vol. 30, pp. 369–373, 1984.

[8] T. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[9] D. Cox and D. Hinckley, *Theoretical Statistics.* London: Chapman and Hall, 1974.

[10] I. Csiszár, “Eine informationtheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten,” *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, 1963.

[11] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[12] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals Probab.*, vol. 3, pp. 146–158, 1975.

[13] I. Csiszár, "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling," *Annals Statist.*, vol. 17, pp. 1409–1413, 1989.

[14] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to linear inverse problems," *Annals Statist.*, vol. 19, pp. 2031–2066, 1991.

[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Budapest: Akadémiai Kiadó, and New York: Academic Press, 1981.

[16] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1474–1490, 2003.

[17] I. Csiszár and P. Shields, "The consistency of the BIC Markov order estimator," *Annals Statist.*, vol. 28, pp. 1601–1619, 2000.

[18] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions,* Suppl. Issue 1, pp. 205–237, 1984.

[19] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Annals Math. Statist.*, vol. 43, pp. 1470–1480, 1972.

[20] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. 19, pp. 783–796, 1973.

[21] L. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. 26, pp. 166–174, 1980.

[22] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. 27, pp. 269–279, 1981.

[23] A. P. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc., Ser.B*, vol. 39, pp. 1–38, 1977.

[24] L. Finesso, *Order estimation for functions of Markov chains.* PhD thesis, Univ. Maryland, College Park, 1990.

[25] B. Fitingof, "Coding in the case of unknown and changing message statistics" (in Russian), *Probl. Inform. Transmission*, vol. 2, no. 2, pp. 3–11, 1966.

[26] D. Gokhale and S. Kullback, *The Information in Contingency Tables.* New York: Marcel Dekker, 1978.

[27] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Annals Math. Statist.*, vol. 36, pp. 369–400, 1965.

[28] C. Ireland and S. Kullback, "Contingency tables with given marginals," *Biometrica*, vol. 55, pp. 179–188, 1968.

[29] R. Krichevsky, *Lectures in Information Theory* (in Russian). Novosibirsk State University, 1970.

[30] R. Krichevsky and V. Trofimov, "The performance of universal coding," *IEEE Trans. Inform. Theory*, vol. 27, pp. 199–207, 1981.

[31] J. Kruithof, "Telefoonverkeersrekening," *De Ingenieur*, vol. 52, pp. E15–E25, 1937.

[32] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959.

[33] S. Kullback and R. Leibler, "On information and sufficiency," *Annals Math. Statist.*, vol. 22, pp. 79–86, 1951.

[34] F. Liese and I. Vajda, *Convex Statistical Distances.* Leipzig: Teubner, 1987.

[35] J. Rissanen, "Generalized Kraft inequality and arithmetic coding," *IBM J. Res. Devel.*, vol. 20, pp. 198–203, 1976.

[36] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[37] J. Rissanen, "Tight lower bounds for optimum code length," *IEEE Trans. Inform. Theory*, vol. 28, pp. 348–349, 1982.

[38] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. 30, pp. 629–636, 1984.

[39] J. Rissanen, *Stochastic Complexity in Statistical Inquiry.* World Scientific, 1989.

[40] J. Rissanen and G. Langdon, "Arithmetic coding," *IBM J. Res. Devel.*, vol. 23, pp. 149–162, 1979.

[41] B. Ryabko, "Twice-universal coding" (in Russian), *Probl. Inform. Transmission*, vol. 20, no. 3, pp. 24–28, 1984.

[42] I. Sanov, "On the probability of large deviations of random variables" (in Russian), *Mat. Sbornik*, vol. 42, pp. 11–44, 1957.

[43] G. Schwarz, "Estimating the dimension of a model," *Annals Statist.*, vol. 6, pp. 461–464, 1978.

[44] C. Shannon, "A mathematical theory of communication," *Bell Syst. Techn. J.*, vol. 27, pp. 379–423 and 623–656, 1948.

[45] P. Shields, *The ergodic theory of discrete sample paths*, Amer. Math. Soc., Graduate Studies in Mathematics, vol. 13, 1996.

[46] Y. Shtarkov, "Coding of discrete sources with unknown statistics." In: *Topics in Information Theory.* Colloquia Math. Soc. J. Bolyai, vol. Vol. 23, pp. 175–186, 1977.

[47] F. Topsœ, "An information theoretic identity and a problem involving capacity," *Studia Sci. Math. Hungar*, vol. 2, pp. 291–292, 1967.

[48] V. Trofimov, "Redundancy of universal coding of arbitrary Markov sources" (in Russian), *Probl. Inform. Transmission*, vol. 10, pp. 16–24, 1974.

[49] A. Wald, *Sequential Analysis.* New York: Wiley, 1947.

[50] F.M.J. Willems, Y. Shtarkov, and T. Tjalkens, "The context weighting method: basic properties," *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, 1995.

[51] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. 24, 1977.