Information Theory and Statistics: A Tutorial

Information Theory and Statistics: A Tutorial

Imre Csiszár

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

csiszar@renyi.hu

Paul C. Shields

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

paul.shields@utoledo.edu

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1 781 871 0245
www.nowpublishers.com
sales@nowpublishers.com

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

A Cataloging-in-Publication record is available from the Library of Congress

Printed on acid-free paper

ISBN: 1-933019-05-0; ISSNs: Paper version 1567-2190; Electronic version 1567-2328

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 Chapter 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 Chapters 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 Chapter 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. Chapters 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 Chapter 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 Chapters, 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 Chapter 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 [7] or Sections 1.1, 1.3, 1.4 of the Csiszár-Körner book [14], for information theory, and Chapters 1–4 and Sections 9.1–9.3 of the book by Cox and Hinckley [8], for statistical theory.

1

Preliminaries

The symbol $A = \{a_1, a_2, \ldots, a_{|A|}\}$ denotes a finite set of cardinality $|A|$; x_m^n denotes the sequence $x_m, x_{m+1}, \ldots, x_n$, 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 $C\colon A \mapsto B^*$. Its *length function* $L\colon A \mapsto N$ 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 $u = C(x_1)\ldots C(x_i)$ is a prefix need not look at the part of w following u in order to identify u as the code of the sequence $x_1 \ldots x_i$.

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, $t(a) = 0.b_1 \ldots b_{L(a)}$. 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 $t(i) = 0.b_1 \ldots b_{L(a_i)}$, 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 $C: A \mapsto B^*$ 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 $C_n: A^n \mapsto B^*$ with length functions $L_n: A^n \mapsto N$, 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 $\{J(x_1^n a):\ a \in A\}$ 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 $C: A^n \mapsto B^*$ 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.

References

[1] S. Amari, *Differential-Geometrical Methods in Statistics.* NewYork: Springer, 1985.

[2] N. L. A.P. Dempster and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Royal Stat. Soc., Ser.B*, vol. 39, pp. 1–38, 1977.

[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] 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.

[6] T. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inform. Theory*, vol. 30, pp. 369–373, 1984.

[7] T. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[8] D. Cox and D. Hinckley, *Theoretical Statistics.* London: Chapman and Hall, 1974.

[9] 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. pp 85–108, 1963.

[10] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[11] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals Probab.*, vol. 3, pp. 146–158., 1975.

[12] I. Csiszár, "A geometric interpretation of darroch and ratcliff's generalized iterative scaling," *Annals Statist.*, vol. 17, pp. 1409–1413, 1989.
[13] I. Csiszár, "Why least squares and maximum entropy? an axiomatic approach to linear inverse problems," *Annals Statist.*, vol. 19, pp. 2031–2066, 1991.
[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Akadémiai Kiadó, Budapest and Academic Press, 1981.
[15] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1474–1490, 2003.
[16] I. Csiszár and P. Shields, "The consistency of the bic markov order estimator," *Annals Statist.*, vol. 28, pp. pp.1601–1619, 2000.
[17] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions, Suppl.*, vol. 1, pp. 205–237., 1984.
[18] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Annals Math. Statist.*, vol. 43, pp. 1470–1480, 1972.
[19] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. 19, pp. 783–796, 1973.
[20] L. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. 26, pp. 166–174, 1980.
[21] L. Finesso, *Order estimation for functions of Markov chains.* PhD thesis, Univ. Maryland, College Park, 1990.
[22] B. Fitingof, "Coding in the case of unknown and and changing message statistics (in russian," *Probl. Inform. Transmission*, vol. 2, no. 2, pp. 3–11, 1966.
[23] Y. S. F.M.J. Willems and T. Tjalkens, "The context weighting method: basic properties," *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, 1995.
[24] D. Gokhale and S. Kullback, *The Information in Contingency Tables.* New York: Marcel Dekker, 1978.
[25] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Annals Math. Statist.*, vol. 36, pp. 369–400, 1965.
[26] C. Ireland and S. Kullback, "Contingency tables with given marginals," *Biometrica*, vol. 55, pp. 179–188, 1968.
[27] R. Krichevsky, *Lectures in Information Theory (in Russian.* Novosibirsk State University, 1970.
[28] R. Krichevsky and V. Trofimov, "The performance of universal coding," *IEEE Trans. Inform. Theory*, vol. 27, pp. 199–207, 1981.
[29] J. Kruithof, "Telefoonverkeersrekening," *De Ingenieur*, vol. 52, pp. E15–E25, 1937.
[30] S. Kullback, *Information Theory and Statistics.* New York: Wiley, 1959.
[31] S. Kullback and R. Leibler, "On information and sufficiency," *Annals Math. Statist.*, vol. 22, pp. 79–86, 1951.
[32] M. P. L.D. Davisson, R.J. McEliece and M. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. 27, pp. 269–279, 1981.
[33] F. Liese and I. Vajda, *Convex Statistical Distances.* Leipzig: Teubner, 1987.
[34] J. Rissanen, "Generalized kraft inequality and arithmetic coding," *IBM J. Res. Devel.*, vol. 20, pp. 198–203, 1976.

[35] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[36] J. Rissanen, "Tight lower bounds for optimum code length," *IEEE Trans. Inform. Theory*, vol. 28, pp. 348–349, 1982.

[37] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. 30, pp. 629–636, 1984.

[38] J. Rissanen, *Stochastic Complexity in Statistical Inquiry.* World Scientific, 1989.

[39] J. Rissanen and G. Langdon, "Arithmetic coding," *IBM J. Res. Devel.*, vol. 23, pp. 149–162, 1979.

[40] B. Ryabko, "Twice-universal coding (in russian," *Probl. Inform. Transmission*, vol. 20, no. 3, pp. 24–28, 1984.

[41] I. Sanov, "On the probability of large deviations of random variables (in russian," *Mat. Sbornik*, vol. 42, pp. 11–44, 1957.

[42] G. Schwarz, "Estimating the dimension of a model," *Annals Statist.*, vol. 6, pp. 461–464, 1978.

[43] C. Shannon, "A mathematical theory of communication," *Bell Syst. Techn. J.*, vol. 27, pp. 379–423 and 623–656, 1948.

[44] P. Shields, "The ergodic theory of discrete sample paths," *Amer. Math. Soc.*, vol. 13, 1996. Graduate Studies in Mathematics.

[45] Y. Shtarkov, "Coding of discrete sources with unknown statistics," *Colloquia Math. Soc. J. Bolyai*, vol. Vol. 23, pp. 175–186, 1977. In: Topics in Information Theory.

[46] S. S. S.M. Ali, "A general class of coefficients of divergence of one distribution from another," *J. Royal Stat. Soc. Ser.B*, vol. 28, pp. 131–142, 1996.

[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] N. Čencov, "Statistical decision rules and optimal inference," *Amer. Math. Soc., Providence*, 1982. Russian original: Nauka, Moscow, 1972.

[50] A. Wald, *Sequential Analysis.* New York: Wiley, 1947.

[51] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. 24, 1977.