INFORMATION AND CONTROL 11, 317–322 (1967)

Two Remarks to Noiseless Coding

IMRE CSISZÁR

Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary

An inequality concerning Kullback's I-divergence is applied to obtain a necessary condition for the possibility of encoding symbols of the alphabet of a discrete memoryless source of entropy H by sequences of symbols of another alphabet of size D in such a way that the average code length be close to the optimum $H/\log D$. The same idea is applied to the problem of maximizing entropy per second for unequal symbol lengths, too.

LIST OF SYMBOLS

$\mathcal{P}$, $\mathcal{Q}$	Finite probability distributions
H	Entropy
$H(\mathcal{P})$	Entropy of the distribution $\mathcal{P}$
$I(\mathcal{P} \parallel \mathcal{Q})$	I-divergence of $\mathcal{P}$ and $\mathcal{Q}$
ℓ_i	Code word length
L	Average code word length
t_i, t_{ij}^s	Symbol length
T	Average symbol length
δ_{ij}	Kronecker's delta
$(\cdot)_{ij}$	Matrix
$\| \cdot \|$	Determinant

1. Given a discrete source with alphabet $(a_1, \cdots, a_d)$, the symbols a_i having probabilities p_i $(i = 1, \cdots, d)$ the well-known "noiseless coding theorem" (see Feinstein, 1958) asserts that for any encoding of the symbols a_i by sequences of symbols from an alphabet of size D in a uniquely decipherable way, the average code length

$$L = \sum_{i=1}^{d} p_i \ell_i, \tag{1}$$

satisfies the inequality

$$L \geqq H/\log D, \tag{2}$$

where ℓ_i stands for the length of the code word corresponding to a_i (the code letters are assumed to be of unit length) and $H = H(\mathcal{P}) = -\sum_{i=1}^{d} p_i \log p_i$ is the entropy of the distribution $\mathcal{P} = (p_1, \cdots, p_d)$. (We do not specify the base of the logarithms; of course, it has to be the same in all formulas.) There arises the question, in what case an average code length L, near to its lower bound $H/\log D$, can be attained. In this note a necessary condition is presented.[1]

Recall that (2) follows from the Kraft-McMillan inequality

$$\sum_{i=1}^{d} D^{-\ell_i} \leqq 1, \tag{3}$$

valid for every uniquely decipherable code (see Feinstein 1958) simply by applying the inequality

$$I(\mathcal{P} \,\|\, \mathcal{Q}) = \sum_{i=1}^{d} p_i \log (p_i/q_i) \geqq 0, \tag{4}$$

(valid for arbitrary probability distributions $\mathcal{P}$ and $\mathcal{Q}$) to the source distribution $\mathcal{P}$ and the auxiliary distribution $\mathcal{Q} = (q_1, \cdots, q_d)$ where

$$q_i = \delta D^{-\ell_i}, \qquad \delta = (\sum_{i=1}^{d} D^{-\ell_i})^{-1} \geqq 1. \tag{5}$$

In fact, (4) and (5) give rise to

$$-H + \log D \sum_{i=1}^{d} p_i \ell_i - \log \delta \geqq 0, \tag{6}$$

whence (2) directly follows.

Observe that $I(\mathcal{P} \,\|\, \mathcal{Q})$ in (4) is the I-divergence of the distributions $\mathcal{P}$ and $\mathcal{Q}$ in the sense of Kullback (1959). The I-divergence is a measure of how different the distributions $\mathcal{P}$ and $\mathcal{Q}$ are; there holds

$$\sum_{i=1}^{d} |p_i - q_i| \leqq C[I(\mathcal{P} \,\|\, \mathcal{Q})]^{1/2} \tag{7}$$

where C is a constant.[2] An inequality of type (7) has first been given by

[1] An interesting sharpening of the "noiseless coding theorem" in another direction has recently been obtained by Gy. Katona and G. Tusnády (1967).

[2] A similar result holds for arbitrary distributions; in this note, however, we have to do only with discrete ones. A somewhat sharper inequality than (7) with (8) has been proved by S. Kullback, IEEE Transactions on Information Theory IT-13, 126–127 (1967) (Added in proof).]

Pinsker (1960). As it has been shown by the author (Csiszár, 1966 and 1967) such an estimation holds for a wide class of information-type measures of difference of probability distributions, including Rényi's "information gain of order α" (Rényi, 1961), (7) being a consequence of the convexity of the function $f(u) = u \log u$ alone. The smallest constant C, for which (7) is valid, turns out to be

$$C_{\min} = (2/\log e)^{1/2}. \tag{8}$$

(See Csiszár, 1967 where logarithms to the base e are used, thus the constant $C_{\min} = \sqrt{2}$ is obtained).

On applying (7) to the distributions $\mathcal{P}$ and $\mathcal{Q}$ above we obtain

$$\sum_{i=1}^{d} | p_i - \delta D^{-l_i} | \leqq C(-H + L \log D - \log \delta)^{1/2}. \tag{9}$$

Hence it follows that $L \leqq H/\log D + \epsilon$ implies

$$\sum_{i=1}^{d} | p_i - D^{\epsilon_1 - l_i} | \leqq C[(\epsilon - \epsilon_1) \log D]^{1/2} \tag{10}$$

where ϵ_1 is defined by

$$D^{\epsilon_1} = \delta = (\sum_{i=1}^{d} D^{-l_i})^{-1}, \qquad (0 \leqq \epsilon_1 \leqq \epsilon). \tag{11}$$

As a simple consequence of (10) and (11) we also have

$$\begin{aligned} &\sum_{i=1}^{d} | p_i - D^{-l_i} | \\ &\leqq C[(\epsilon - \epsilon_1) \log D]^{1/2} + 1 - D^{-\epsilon_1} < C(\epsilon \log D)^{1/2} + 1 - D^{-\epsilon}. \end{aligned} \tag{12}$$

This inequality can be considered as a rigorous version of the loose statement that almost-optimal average code length can be attained only if the symbol probabilities p_i are approximately of the form D^{-l_i}.

2. Inequality (7) and its continuous analogon can also be used to arrive at useful estimates in other problems. For instance, the difference of the entropy and its maximal possible value (under certain constraints) is often expressible as $-I(\mathcal{P} \| \mathcal{P}_0)$ where $\mathcal{P}$ is the distribution in question and $\mathcal{P}_0$ the one having maximum entropy; then (7) gives rise to a necessary condition of nearly maximum entropy in terms of the closeness of $\mathcal{P}$ to $\mathcal{P}_0$.

Let us give an example related to alphabets with symbols of different lengths. Let the symbol lengths be $t_i > 0$ $(i = 1, \cdots, d)$ and set $q_i = e^{-\beta_0 t_i}$ $(i = 1, \cdots, d)$, where β_0 is the (unique) positive root of the equation

$$\sum_{i=1}^{d} e^{-\beta t_i} = 1. \tag{13}$$

Then inequality (4) gives rise to

$$\frac{-\sum_{i=1}^{d} p_i \log p_i}{\sum_{i=1}^{d} p_i t_i} \leqq \beta_0 \log e \tag{14}$$

with equality if and only if $p_i = q_i$ $(i = 1, \cdots, d)$. (See Krause, 1962.) Making use of (7) and (8) we may even write

$$\frac{-\sum_{i=1}^{d} p_i \log p_i}{\sum_{i=1}^{d} p_i t_i} \leqq \left(\beta_0 + \frac{\left(\sum_{i=1}^{d} | p_i - e^{-\beta_0 t_i} | \right)^2}{2 \sum_{i=1}^{d} p_i t_i} \right) \log e. \tag{15}$$

This inequality can be interpreted as a necessary condition (in terms of the closeness of the p_i to the optimal probabilities $q_i = e^{-\beta_0 t_i}$) in order that the "entropy per second" (left side of (14) and (15)) be close to its optimal value $\beta_0 \log e$.

In the theory of finite-state noiseless channels (Shannon and Weaver, 1949) the problem of maximizing

$$\frac{H}{T} = \frac{-\sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{s=1}^{n_{ij}} p_i \, p_{ij}^s \log p_{ij}^s}{\sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{s=1}^{n_{ij}} p_i \, p_{ij}^s \, t_{ij}^s} \tag{16}$$

is also considered[3], where t_{ij}^s $(i = 1, \cdots, d; j = 1, \cdots, d; s = 1, \cdots, n_{ij})$ are given positive numbers (n_{ij} may be zero for some pairs (i, j) but the matrix $N = (n_{ij})_{ij}$ is assumed to be indecomposable) and the p_i's and p_{ij}^s have to be nonnegative and satisfy

[3] A recent contribution to this problem is Radke (1966). Our aim here is to show that by the aid of (4) the problem can be solved in a straightforward way and no need for Lagrange multipliers ever arises.

$$\sum_{i=1}^{d} p_i = 1; \qquad \sum_{j=1}^{d} \sum_{s=1}^{n_{ij}} p_{ij}^{s} = 1; \qquad \sum_{i=1}^{d} \sum_{s=1}^{n_{ij}} p_i \, p_{ij}^{s} = p_j . \tag{17}$$

Let β_0 be the greatest positive root of the equation

$$\| \sum_{s=1}^{n_{ij}} e^{-\beta t_{ij}^{s}} - \delta_{ij} \| = 0 \tag{18}$$

and let A_i $(i = 1, \cdots, d)$ be positive numbers satisfying

$$\sum_{j=1}^{d} A_j \Big(\sum_{s=1}^{n_{ij}} e^{-\beta_0 t_{ij}^{s}} \Big) = A_i \quad (i = 1, \cdots, d); \qquad \sum_{i=1}^{d} A_i = 1. \tag{19}$$

Then, applying (4) to p_{ij}^{s} and $q_{ij}^{s} = (A_j/A_i) e^{-\beta_0 t_{ij}^{s}}$ and utilizing (17) we obtain for the "entropy per second" (16)

$$H/T \leqq \beta_0 \log e \tag{20}$$

with equality if and only if $p_{ij}^{s} = q_{ij}^{s}$, $p_i = A_i$. By the aid of (7) and (8) an estimate of type (15) might also be given. To make the above argument rigorous, we have only to show that β_0 and A_i's with the required properties do exist.

By well-known properties of indecomposable matrices with nonnegative elements (Gantmacher, 1959) the greatest positive eigenvalue $r(\beta)$ of the matrix $(\sum_{s=1}^{n_{ij}} e^{-\beta t_{ij}^{s}})_{ij}$ is a decreasing function of β,

$$r(0) > 1, \qquad \lim_{\beta \to \infty} r(\beta) = 0. \tag{21}$$

Hence, by continuity, there exists a (unique) positive β_0 with $r(\beta_0) = 1$; this β_0 is obviously the greatest positive root of (18) and then, by Frobenius' theorem, (19) can be solved with positive A_i. The proof is complete.

RECEIVED: May 16, 1967

REFERENCES

CSISZÁR, I. (1966), A note on Jensen's inequality. *Studia Sci. Math. Hung.* **1**, 185-188.

CSISZÁR, I. (1967), Information-type measures of difference of probability distributions. *Studia Sci. Math. Hung.* **2**, 299-318.

FEINSTEIN, A. (1958), "Foundations of Information Theory." McGraw-Hill, New York.

GANTMACHER, F. R. (1959), "Applications of the Theory of Matrices." Wiley (Interscience), New York.

KATONA, GY. AND TUSNÁDY, G. (1967), The principle of conservation of entropy in a noiseless channel. *Studia Sci. Math. Hung.* **2**, 29–35.

KRAUSE, R. M. (1962), Channels which transmit letters of inequal duration. *Inform. Control*, **5**, 13–24.

KULLBACK, S. (1959), "Information Theory and Statistics." Wiley, New York.

PINSKER, M. S. (1960), Information and information stability of random variables and processes. *Prob. Peredači Inform.*, **7**, Izd. AN SSSR, Moscow.

RADKE, C. (1966), Necessary and sufficient conditions on conditional probabilities to maximize entropy. *Inform. Control*, **9**, 279–284.

RÉNYI, A. (1961), On measures of entropy and information, *In Proc. 4th Berkeley Symp. Math. Stat. Prob.*, Ed. J. Neyman, **1**, 541–561, University of California Press, Berkeley, California.

SHANNON, C. E. AND WEAVER, W. (1949), "The Mathematical Theory of Communication," University of Illinois Press, Urbana, Illinois.