Reprinted from INFORMATION AND CONTROL, Volume 8, No. 4, August 1965

INFORMATION AND CONTROL 8, 423–429 (1965)

A Coding Theorem and Rényi's Entropy*

L. L. CAMPBELL

Department of Mathematics, Queen's University, Kingston, Ontario, Canada

Let $L(t) = t^{-1} \log_D (\sum p_i D^{tn_i})$, where p_i is the probability of the ith input symbol to a noiseless channel, and n_i is the length of the code sequence for the ith symbol in some uniquely decipherable code. Limiting values of $L(t)$ are $\sum n_i p_i$ for $t = 0$ and max (n_i) for $t = \infty$. It is shown that $L(t)$ has some desirable properties as a measure of typical code length. A coding theorem for a noiseless channel is proved. The theorem states roughly that it is possible to encode so that $L(t)$ is close to H_α, where H_α is Rényi's entropy of order α and $\alpha = (1 + t)^{-1}$.

INTRODUCTION

In the usual discussion of the coding theorem for a noiseless channel (Feinstein, 1958) one chooses code lengths to minimize the average code length. The minimization is done subject to the constraint that the code be uniquely decipherable. The solution of this minimization problem is that the best code length for an input symbol of probability p is $-\log p$. This solution has the disadvantage that the code length is very great if the probability of the symbol is very small.

Implicit in the use of average code length as a criterion of performance is the assumption that cost varies linearly with code length. This is not always the case. In the present paper another measure of code length is introduced which implies that the cost is an exponential function of code length. Linear dependence is a limiting case of this measure.

A coding theorem analogous to the ordinary coding theorem for a noiseless channel will be proved. The theorem states that it is possible to encode so that the measure of length is arbitrarily close to the Rényi entropy of the input. The code lengths produced by this theorem are less than those produced by the ordinary theorem for improbable input

* This research was supported in part by the Defence Research Telecommunications Establishment (DRB) under Contract No. CD. DRB/313002 with Queen's University.

symbols. In compensation, longer codes are required for the most probable symbols.

The Rényi entropy was introduced by Rényi (1961, 1962) as a generalization of the usual notion of entropy. Since this generalized entropy has many properties in common with the ordinary Shannon entropy and includes it as a special case, it is not surprising that there should be a coding theorem associated with Rényi's entropy.

A MEASURE OF LENGTH

Let $p_1, p_2, \cdots, p_N$ be the probabilities of N input symbols $x_1, x_2, \cdots, x_N$ which we wish to encode. We assume that $p_i > 0$ for $i = 1, 2, \cdots, N$ and that $\sum p_i = 1$. Suppose there is an alphabet of D symbols into which the input symbols are to be encoded. Let x_i be represented by a sequence of n_i characters from the alphabet. It can be shown (Feinstein, 1958) that there is a uniquely decipherable code with lengths $n_1, \cdots, n_N$ if and only if

$$\sum_{i=1}^{N} D^{-n_i} \leqq 1. \tag{1}$$

There are many different codes whose lengths satisfy the constraint (1). To compare different codes and pick out an optimum code it is customary to examine the mean length, $\sum n_i p_i$, and to minimize this quantity. This is a good procedure if the cost of using a sequence of length n_i is directly proportional to n_i. However, there may be occasions when the cost is more nearly an exponential function of n_i. This could be the case, for example, if the cost of encoding and decoding equipment were an important factor. Thus, in some circumstances, it might be more appropriate to choose a code which minimizes the quantity

$$C = \sum_{i=1}^{N} p_i D^{tn_i},$$

where t is some parameter related to the cost. For reasons which will become evident later we prefer to minimize a monotonic function of C. Clearly this will also minimize C.

In order to make the result of this paper more directly comparable with the usual coding theorem we introduce a quantity which resembles the mean length. Let a code length of order t be defined by

$$L(t) = \frac{1}{t} \log_D \left(\sum_{i=1}^{N} p_i D^{tn_i} \right) \qquad (0 < t < \infty). \tag{2}$$

An application of l'Hospital's rule shows that

$$L(0) = \lim_{t \to 0} L(t) = \sum_{i=1}^{N} n_i p_i . \tag{3}$$

For large t,

$$\sum_{i=1}^{N} p_i D^{tn_i} \doteq p_j D^{tn_j},$$

where n_j is the largest of the numbers $n_1, \cdots, n_N$. Thus,

$$L(\infty) = \lim_{t \to \infty} L(t) = \max_{1 \leqq i \leqq N} n_i . \tag{4}$$

Moreover (Beckenbach and Bellman, 1961, see p. 16), $L(t)$ is a monotonic nondecreasing function of t. Thus $L(0)$ is the conventional measure of mean length and $L(\infty)$ is the measure which would be used if the maximum length were of prime importance. Intermediate values of t provide a measure of length which lies between these limits. The larger the value of t, the more is the weight which is given to the larger values of n_i.

Note also that if all n_i are the same, say $n_i = n$, then $L(t) = n$. This is a reasonable property for any measure of length to possess.

A CODING THEOREM

Before proceeding to the coding theorem we need a definition and a lemma. Rényi (1961, 1962) has introduced the entropy of order α, defined by

$$H_\alpha = \frac{1}{1 - \alpha} \log_D \left(\sum_{i=1}^{N} p_i^{\alpha} \right) \qquad (\alpha \neq 1). \tag{5}$$

L'Hospital's rule shows that

$$H_1 = \lim_{\alpha \to 1} H_\alpha = -\sum_{i=1}^{N} p_i \log_D p_i . \tag{6}$$

Thus H_1 is the ordinary Shannon entropy. The entropy of order α behaves in much the same way as H_1. For example, H_α is a continuous and symmetric function of $p_1, \cdots, p_N$. If $p_i = N^{-1}$ for each i, $H_\alpha = \log_D N$. In addition, if X and Y are two independent sources,

$$H_\alpha(X, Y) = H_\alpha(X) + H_\alpha(Y).$$

Properties of this sort can be used to give an axiomatic characterization of H_α in a fashion similar to the well known axiomatic charac-

terizations of H_1. Axiomatic characterizations of H_α have been studied by Rényi (1961, 1962), Aczél and Daróczy (1963a, 1963b), and Daróczy (1963). The theorem of this paper can be regarded as giving an alternative characterization of H_α in the same way that the noiseless coding theorem provides an alternative characterization of H_1.

An inequality relating H_α and $L(t)$ is provided by the

LEMMA. *Let* $n_1, \cdots, n_N$ *satisfy* (1). *Then*

$$L(t) \geqq H_\alpha, \tag{7}$$

where $\alpha = 1/(t+1)$.

Proof: If $t = 0$ and $\alpha = 1$ the result is given by Feinstein (1958) in his proof of the noiseless coding theorem.

If $t = \infty$ and $\alpha = 0$ we have $L(\infty) = \max n_i$ and $H_0 = \log_D N$. If the n_i satisfy (1) we must have

$$D^{-n_i} \leqq N^{-1}$$

for at least one value of i and hence for the maximum n_i. It follows easily that $\max n_i \geqq \log_D N$.

Now let $0 < t < \infty$. By Hölder's inequality,

$$\left(\sum_{i=1}^{N} x_i^{\,p}\right)^{1/p} \left(\sum_{i=1}^{N} y_i^{\,q}\right)^{1/q} \leqq \sum_{i=1}^{N} x_i\, y_i \tag{8}$$

where $p^{-1} + q^{-1} = 1$ and $p < 1$. Note that the direction of Hölder's inequality is the reverse of the usual one for $p < 1$ (Beckenbach and Bellman, 1961, see p. 19). In (8), let $p = -t$, $q = 1 - \alpha$, $x_i = p_i^{-1/t} D^{-n_i}$, and $y_i = p_i^{1/t}$. The equation $p^{-1} + q^{-1} = 1$ implies that $\alpha = (t+1)^{-1}$. With these substitutions (8) becomes

$$(\sum p_i D^{tn_i})^{-1/t} (\sum p_i^{\,\alpha})^{1/(1-\alpha)} \leqq \sum D^{-n_i}.$$

Therefore

$$(\sum p_i\, D^{tn_i})^{1/t} \geqq \frac{(\sum p_i^{\,\alpha})^{1/(1-\alpha)}}{\sum D^{-n_i}} \geqq (\sum p_i^{\,\alpha})^{1/(1-\alpha)}, \tag{9}$$

where the last inequality follows from the assumption that (1) is satisfied. If we take logarithms of the first and third members of (9) we have the statement of the Lemma.

An easy calculation shows that we have equality in (7) and (9) and (1) is satisfied if

$$D^{-n_i} = \frac{p_i^{\alpha}}{\sum_{j=1}^{N} p_j^{\alpha}},$$

or

$$n_i = -\alpha \log_D p_i + \log_D \left(\sum_{j=1}^{N} p_j^{\alpha} \right). \tag{10}$$

Thus, if we ignore the additional constraint that each n_i should be an integer, it is seen that the minimum possible value of $L(t)$ is H_α, where $\alpha = (t+1)^{-1}$. Moreover, as $p_i \to 0$, the optimum value of n_i is asymptotic to $(-\log_D p_i)/(t+1)$ so that the optimum length is less than $-\log_D p_i$ for $t > 0$ and sufficiently small p_i.

We can now proceed to prove a coding theorem for a noiseless channel with independent input symbols. Let a sequence of input symbols be generated independently, where each symbol is governed by the probability distribution $(p_1, \cdots, p_N)$. Consider a typical input sequence of length M, say $s = (a_1, a_2, \cdots, a_M)$. The probability of s is

$$P(s) = p_{i_1} p_{i_2} \cdots p_{i_M} \tag{11}$$

if $a_1 = x_{i_1}, \cdots, a_M = x_{i_M}$. Let $n(s)$ be the length of the code sequence for s in some uniquely decipherable code. Then the length of order t for the M-sequences is

$$L_M(t) = \frac{1}{t} \log_D \sum P(s) D^{tn(s)} \qquad (0 < t < \infty), \tag{12}$$

where the summation extends over the N^M sequences s. The entropy of order α of this product space is

$$H_\alpha(M) = \frac{1}{1-\alpha} \log_D Q, \tag{13}$$

where

$$Q = \sum [P(s)]^{\alpha}. \tag{14}$$

It follows directly from (11) that

$$Q = \left(\sum_{i=1}^{N} p_i^{\alpha} \right)^M$$

and hence that

$$H_\alpha(M) = M H_\alpha. \tag{15}$$

Now let $n(s)$ be the integer which satisfies

$$-\alpha \log_D P(s) + \log_D Q \leqq n(s) < 1 - \alpha \log_D P(s) + \log_D Q. \quad (16)$$

As we remarked earlier in connection with (10), if every $n(s)$ equals the left member of (16) then $L_M(t) = H_\alpha(M)$. If $n(s)$ satisfies (16) for each sequence s, the numbers $n(s)$ satisfy

$$\sum D^{-n(s)} \leqq 1,$$

so that there is a uniquely decipherable code with lengths $n(s)$. It also follows from (16) that

$$[P(s)]^{-\alpha t} Q^t \leqq D^{tn(s)} < D^t [P(s)]^{-\alpha t} Q^t. \quad (17)$$

If we multiply each member of (17) by $P(s)$, sum over all s, and use the fact that $\alpha t = 1 - \alpha$, we get

$$Q^{1+t} \leqq \sum P(s) D^{tn(s)} < D^t Q^{1+t}.$$

Now take logarithms, divide by t, and use the relations $1 + t = \alpha^{-1}$ and $\alpha t = 1 - \alpha$. From (12) and (13) we have

$$H_\alpha(M) \leqq L_M(t) < H_\alpha(M) + 1. \quad (18)$$

Finally, if we divide by M and use (15), we have

$$H_\alpha \leqq \frac{L_M(t)}{M} < H_\alpha + \frac{1}{M}. \quad (19)$$

The quantity $L_M(t)/M$ might be called the average code length of order t per input symbol. By choosing M sufficiently large the average length can be made as close to H_α as desired. Thus we have proved most of the

THEOREM. *Let $\alpha = (1 + t)^{-1}$. By encoding sufficiently long sequences of input symbols it is possible to make the average code length of order t per input symbol as close to H_α as desired. It is not possible to find a uniquely decipherable code whose average length of order t is less than H_α.*

The second half of the theorem follows directly from (7) and (15). If $t = 0$ this is just the ordinary coding theorem. If $t = \infty$ the above proof is not quite correct but the theorem is still true. In this case we choose each $n(s)$ to satisfy

$$\log_D N^M \leqq n(s) < 1 + \log_D N^M.$$

Then since $H_0 = \log_D N$ and $L_M(\infty) = \max n(s)$, we have

$$H_0 \leqq \frac{L_M(\infty)}{M} < H_0 + \frac{1}{M}.$$

Thus the theorem follows as before.

RECEIVED: August 12, 1964

REFERENCES

ACZÉL, J., AND DARÓCZY, Z. (1963a), Charakterisierung der Entropien positiver Ordnung und der Shannonschen Entropie. *Acta Math. Acad. Sci. Hung.* **14,** 95–121.

ACZÉL, J., AND DARÓCZY, Z. (1963b), Sur la caractérisation axiomatique des entropies d'ordre positif, y comprise l'entropie de Shannon. *C. R. Acad. Sci. Paris* **257,** 1581–1584.

BECKENBACH, E. F., AND BELLMAN, R. (1961), "Inequalities." Springer, Berlin.

DARÓCZY, Z. (1963), Über die gemeinsame Charakterisierung der zu den nicht völlstandigen Verteilungen gehörigen Entropien von Shannon und von Rényi. *Z. Wahrscheinlichkeitstheorie u. Verw. Gebiete* **1,** 381–388.

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

RÉNYI, A. (1961), On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1,** 547–561. Univ. of California Press, Berkeley.

RÉNYI, A. (1962), "Wahrscheinlichkeitsrechnung. Mit einem Anhang über Informationstheorie." VEB Deutscher Verlag der Wissenschaften, Berlin.