Z. Wahrscheinlichkeitstheorie verw. Geb. 6, 113—118 (1966)

Definition of Entropy by Means of a Coding Problem

L. L. CAMPBELL*

Received August 2, 1965

1. Introduction

There have been several axiomatic characterizations [*1, 2, 3, 4*] of Shannon's entropy and more recently [*5, 6, 7, 8*] of Rényi's entropy of order α. These axioms have been chosen to provide a convenient measure of the uncertainty of the outcome of a random experiment.

However, these axiomatic characterizations have little apparent connection with the coding theorems of information theory. In this note a generalized entropy is defined in a way that connects entropy directly with a coding problem. If additional hypotheses are imposed this generalized entropy becomes the Rényi or Shannon entropy. In fact, the final theorem of this paper is a characterization of the entropy of order α which is different from those mentioned above.

Let $X = \{x_1, x_2, \ldots, x_m\}$ be a finite set of events and let $P = (p_1, p_2, \ldots, p_m)$ be an associated distribution of probabilities, so that the probability of x_i is p_i and $\sum p_i = 1$. Suppose that we wish to represent the events in X by finite sequences of elements of the set $\{0, 1, \ldots, D - 1\}$ where $D > 1$. There is a uniquely decipherable code [*9*] which represents x_i by a sequence of n_i elements if and only if the integers n_i satisfy the inequality

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

Now suppose that there is a "cost function" φ associated with the length so that the "cost" of using a sequence of length n is $\varphi(n)$. Then the average cost of encoding X by a distribution of lengths $N = (n_1, n_2, \ldots, n_m)$ is

$$C = \sum p_i \varphi(n_i).$$

It will be assumed that φ is a continuous strictly monotonic increasing function on the non-negative real numbers. Then φ has an inverse, φ^{-1}. We can now introduce a mean length for the cost function φ by

$$L(P, N, \varphi) = \varphi^{-1}(C) = \varphi^{-1}\left(\sum p_i \varphi(n_i)\right). \tag{2}$$

The reason for calling L a mean length is that when $n_1 = n_2 = \cdots = n_m = n$, then $L = n$. Moreover, if $\varphi(x) = x$,

$$L(P, N, \varphi) = \sum p_i n_i,$$

the ordinary mean length.

*This research was supported in part by the Defence Research Telecommunications Establishment (DRB) under Contract No. CD. 513003 with Queen's University.

A coding problem of some interest is to minimize the cost by an appropriate choice of the distribution of lengths, subject to the constraint (1). Since L is a monotonic increasing function of C, an equivalent problem is to minimize L, subject to (1). This minimum will be called the generalized entropy and denoted by $H(P, \varphi)$. Thus,

$$H(P, \varphi) = \inf_{N \in S} L(P, N, \varphi), \tag{3}$$

where S is the set of all real distributions $N = (n_1, n_2, \ldots, n_m)$ for which (1) is satisfied. Since the numbers n_i are not restricted to integer values in (3), $H(P, \varphi)$ merely provides a lower bound on the values of L which are possible in a real coding problem. However, when we consider coding sequences of elements of X, $H(P, \varphi)$ sometimes takes on added significance [*9*, *10*].

2. Properties of generalized entropy

In view of the general nature of $H(P, \varphi)$ it is not to be expected that many results can be derived. However there are some properties which follow from (1)—(3). For example:

(i) If $P = (p_1, p_2, \ldots, p_m)$, $H(P, \varphi)$ is a symmetric function of $p_1, p_2, \ldots, p_m$.

(ii) If $P_1 = (p_1, p_2, \ldots, p_m)$ and $P_2 = (p_1, p_2, \ldots, p_m, 0)$, $H(P_1, \varphi) = H(P_2, \varphi)$.

(iii) If $P = (1, 0, 0, \ldots, 0)$, $H(P, \varphi) = 0$.

(iv) If $P = (p_1, p_2, \ldots, p_m)$, $H(P, \varphi) \leqq \log_D m$.

Property (i) is obvious from the definitions. Property (ii) follows easily from the observation that $D^{-n_{m+1}}$ can be made arbitrarily small without affecting the value of L (where n_{m+1} is the length corresponding to the event of probability 0). Property (iii) follows similarly, by letting $n_1 \to 0$ and the other $n_i \to \infty$, since in this case $L = n_1$. Property (iv) is a consequence of the fact that (1) is satisfied if $n_i = \log_D m$ for $i = 1, 2, \ldots, m$.

Other bounds on $H(P, \varphi)$ can be derived if φ is a convex or concave function. The function φ will be called convex if, for any probability distribution $P = (p_1, p_2, \ldots, p_m)$,

$$\sum p_i \varphi(x_i) \leqq \varphi\left(\sum p_i x_i\right),$$

and concave if the direction of inequality is reversed. If φ is convex, φ^{-1} is concave and *vice versa*.

If φ is concave, the fact that φ^{-1} is an increasing function shows that

$$L(P, N, \varphi) \geqq \sum p_i n_i.$$

Since $n_i = -\log_D p_i$ minimizes $\sum p_i n_i$ under the constraint (1),

$$H(P, \varphi) \geqq -\sum p_i \log_D p_i. \tag{4}$$

Thus, if φ is concave and $p_i = m^{-1}$ for $i = 1, 2, \ldots, m$, it follows from the last inequality and Property (iv) above that

$$H(P, \varphi) = \log_D m.$$

In fact, a somewhat stronger result can be proved:

Theorem 1. *Let $\psi(x) = \varphi(-\log_D x)$ and let P_0 be the uniform distribution $(m^{-1}, m^{-1}, \ldots, m^{-1})$. If ψ is concave*

$$H(P_0, \varphi) = \log_D m.$$

Proof. If we put $n_i = -\log_D x_i$ (3) becomes

$$H(P_0, \varphi) = \inf_{\Sigma x_i \leqq 1} \varphi^{-1}\left(\frac{1}{m} \sum \psi(x_i)\right).$$

Since ψ is concave,

$$\frac{1}{m} \sum \psi(x_i) \geqq \psi\left(\frac{1}{m} \sum x_i\right).$$

Now φ is an increasing function and so ψ is a decreasing function. Therefore, in the region $\sum x_i \leqq 1$,

$$\psi\left(\frac{1}{m} \sum x_i\right) \geqq \psi\left(\frac{1}{m}\right) = \varphi(\log_D m).$$

Hence, since φ^{-1} is increasing,

$$\varphi^{-1}\left(\frac{1}{m} \sum \psi(x_i)\right) \geqq \log_D m.$$

Therefore

$$H(P_0, \varphi) \geqq \log_D m.$$

This inequality and Property (iv) above prove the theorem.

The example

$$\varphi(x) = 1 - D^{tx} \qquad (t < -1)$$

shows that Theorem 1 is not always true if ψ is convex. For this example,

$$\psi(x) = 1 - x^{-t}$$

is convex. Also,

$$L = \varphi^{-1}\left(\frac{1}{m} \sum \psi(x_i)\right) = \frac{1}{t} \log_D \left(\frac{1}{m} \sum x_i^{-t}\right).$$

If we put $x_i = m^{-1}$ for $i = 1, \ldots, m$, then $L = \log_D m$. If we put $x_1 = 1$, $x_2 = x_3 = \cdots = x_m = 0$, then $L = -t^{-1} \log_D m$. Since $t < -1$ the second value of L is less than the first and so $\log_D m$ is not the infimum of L. The same example, for $-1 < t < 0$, provides an example of a convex φ for which ψ is concave.

When φ is convex the direction of the inequality (4) is reversed. Since φ^{-1} is now concave, it follows from (2) that

$$L(P, N, \varphi) \leqq \sum p_i n_i.$$

If we put $n_i = -\log_D p_i$ we have

$$H(P, \varphi) \leqq -\sum p_i \log_D p_i.$$

3. Characterization of Rényi and Shannon Entropy

If it is required that $H(P_0, \varphi) = \log_D m$ when P_0 is the uniform distribution $(m^{-1}, m^{-1}, \ldots, m^{-1})$ and that $L(P, N, \varphi)$ satisfy an additivity condition, then the generalized entropy becomes either Rényi's entropy of order α or Shannon's entropy. The first requirement is made because it is unnatural that an optimum code should assign unequal code lengths to equiprobable events. The additivity condition which is about to be formulated is natural in itself and also makes it possible to derive more precise coding theorems [*9*, *10*].

Consider two independent sets of events $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_k\}$ with associated probability distributions $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_k)$. Since X and Y are independent the probability of the pair (x_i, y_j) is $p_i q_j$. We denote by $P\,Q$ the set of probabilities $(p_i q_j)$.

Let x_i be represented by a sequence of length n_i and let y_j be represented by a sequence of length m_j. Moreover, suppose that the pair (x_i, y_j) is represented by the sequences for x_i and y_j put side by side. Then the length of the sequence for (x_i, y_j) is $n_i + m_j$. Let us denote these three distributions of lengths by N, M, and $N+M$ respectively. Then, if L is to be a measure of mean length, it is natural to require that

$$L(P\,Q, N + M, \varphi) = L(P, N, \varphi) + L(Q, M, \varphi),$$

or that

$$\varphi^{-1}\left(\sum\sum p_i q_j \varphi(n_i + m_j)\right) = \varphi^{-1}\left(\sum p_i \varphi(n_i)\right) + \varphi^{-1}\left(\sum q_j \varphi(m_j)\right). \qquad (5)$$

This is the additivity condition mentioned earlier.

The solution of the functional equation (5) is easily found. Let $m = k = 2$ and let $M = (a, a)$. For this case (5) becomes

$$\varphi^{-1}(p_1 \varphi(n_1 + a) + p_2 \varphi(n_2 + a)) = \varphi^{-1}(p_1 \varphi(n_1) + p_2 \varphi(n_2)) + a. \qquad (6)$$

But it is known [*11*, p. 122] that the only strictly monotonic increasing solutions of (6) are

$$\varphi_0(x) = c\,x + b$$

and

$$\varphi_t(x) = \gamma D^{tx} + b,$$

where $c > 0$ and $\gamma t > 0$. It is easy to verify that (5) is satisfied by φ_0 and φ_t.

The mean lengths associated with these functions are

$$L(P, N, \varphi_0) = \sum p_i n_i$$

and

$$L(P,N,\varphi_t) = \frac{1}{t} \log_D \sum p_i D^{tn_i}.$$

It is easily shown that

$$\lim_{t \to 0} L(P, N, \varphi_t) = \sum p_i n_i.$$

It is well known [*9*] that the minimum of $L(P, N, \varphi_0)$ is

$$H(P, \varphi_0) = -\sum p_i \log_D p_i,$$

which is just Shannon's entropy.

As the example following Theorem 1 shows, if $t < -1$ then $H(P_0, \varphi_t) < < \log_D m$, where P_0 is the uniform distribution. Since we wish to exclude this possibility, we shall consider only the case $t \geqq -1$.

The result of minimizing $L(P, N, \varphi_t)$ is Rényi's entropy of positive order, as is shown by the following:

Lemma. *Let* $-1 < t < \infty$, $t \neq 0$ *and let* $\alpha = (1 + t)^{-1}$. *Then*

$$\inf_{N \in S} L(P, N, \varphi_t) = H_\alpha ,$$

where

$$H_\alpha = \frac{1}{1-\alpha} \log_D \left(\sum_{i=1}^{m} p_i^\alpha \right). \tag{7}$$

Proof. By Hölder's inequality,

$$\left(\sum \xi_i^p\right)^{1/p} \left(\sum \eta_i^q\right)^{1/q} \leqq \sum \xi_i \eta_i , \tag{8}$$

where $p^{-1} + q^{-1} = 1$ and $p < 1$. In (8), let

$$p = -t, \quad q = 1 - \alpha, \quad \xi_i = p_i^{-1/t} D^{-n_i}, \quad \eta_i = p_i^{1/t} .$$

The relation $\alpha = (1 + t)^{-1}$ is equivalent to $p^{-1} + q^{-1} = 1$. Then

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

where the last inequality follows from (1). Taking logarithms, we have that

$$L(P, N, \varphi_t) \geqq H_\alpha \tag{9}$$

for any N which satisfies (1). But the choice

$$n_i = -\alpha \log_D p_i + \log_D \left(\sum p_j^\alpha\right)$$

satisfies (1) and produces equality in (9). This completes the proof.

A result similar to this was proved in [*10*]. The Lemma shows that $H(P, \varphi_t) = H_\alpha$. It was pointed out above that $L(P, N, \varphi_0)$ is the limit as $t \to 0$ of $L(P, N, \varphi_t)$. It is also easy to prove from (7) that

$$H_1 = \lim_{\alpha \to 1} H_\alpha = -\sum_{i=1}^{m} p_i \log_D p_i , \tag{10}$$

so that the Lemma, suitably interpreted, also holds for $t = 0$. There remains the case $t = -1$, $\alpha = \infty$. Now

$$H_\infty = \lim_{\alpha \to \infty} H_\alpha = -\log_D p^* ,$$

where p^* is the largest of $p_1, \ldots, p_m$. Also, for $t = -1$,

$$L(P, N, \varphi_{-1}) = -\log_D \sum p_i D^{-n_i} \geqq -\log_D p^* \left(\sum D^{-n_i}\right) \geqq -\log_D p^* ,$$

when (1) is satisfied. But equality is approached in this inequality if $n_j \to 0$ and all other $n_i \to \infty$, where j is such that $p_j = p^*$. Thus, for $t = -1$,

$$H(P, \varphi_{-1}) = H_\infty ,$$

and hence

$$H(P, \varphi_t) = H_\alpha , \quad \alpha = (1 + t)^{-1}, \quad -1 \leqq t < \infty . \tag{11}$$

The results of this section can be summarized as follows:

Theorem 2. *Let φ be a continuous strictly monotonic increasing function which satisfies the additivity condition (5) and let $H(P_0, \varphi) = \log_D m$ when $P_0 = (m^{-1}, m^{-1}, \ldots, m^{-1})$. Then the generalized entropy $H(P, \varphi)$ must be the entropy of positive order, H_α, for some $\alpha > 0$.*

It is even possible to extend (11) to the case $t = \infty$, $\alpha = 0$ [*10*], although φ is no longer defined, since

$$\lim_{t \to \infty} L(P, N, \varphi_t) = \max_{1 \leqq i \leqq m} n_i$$

and

$$H_0 = \log_D m .$$

Clearly,

$$\inf_{N \in S} \left(\max_{1 \leqq i \leqq m} n_i \right) = \log_D m = H_0 .$$

References

[*1*] SHANNON, C. E.: A mathematical theory of communication. Bell System techn. J. **27**, 379—423 and 623—656 (1948).

[*2*] FADIEV, D. A.: On the notion of entropy of a finite probability space. Uspechi mat. Nauk **11**, 227—231 (1956).

[*3*] KENDALL, D. G.: Functional equations in information theory. Z. Wahrscheinlichkeitstheorie verw. Geb. **2**, 225—229 (1964).

[*4*] LEE, P. M.: On the axioms of information theory. Ann. math. Statistics **35**, 415—418 (1964).

[*5*] RÉNYI, A.: On measures of entropy and information. Proc. Fourth Berkeley Sympos. math. Statist. Probability I, 547—561 (1961).

[*6*] ACZÉL, J., and Z. DARÓCZY: Charakterisierung der Entropien positiver Ordnung und der Shannonschen Entropie. Acta math. Acad. Sci. Hungar. **14**, 95—121 (1963).

[*7*] DARÓCZY, Z.: Über die gemeinsame Charakterisierung der zu den nicht vollständigen Verteilungen gehörigen Entropien von SHANNON und von RÉNYI. Z. Wahrscheinlichkeitstheorie verw. Geb. **1**, 381—388 (1963).

[*8*] ACZÉL, J.: Zur gemeinsamen Charakterisierung der Entropien α-ter Ordnung und der Shannonschen Entropie bei nicht unbedingt vollständigen Verteilungen. Z. Wahrscheinlichkeitstheorie verw. Geb. **3**, 177—183 (1964).

[*9*] FEINSTEIN, A.: Foundations of Information Theory. New York: McGraw-Hill 1958.

[*10*] CAMPBELL, L. L.: A coding theorem and Rényi's entropy. Inform. and Control 8, 423—429 (1965).

[*11*] ACZÉL, J.: Vorlesungen über Funktionalgleichungen und ihre Anwendungen. Basel-Stuttgart: Birkhäuser 1961.

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