A NOTE ON THE ENTROPY OF A CONTINUOUS DISTRIBUTION

BY HIROHISA HATORI

1. The uniqueness of the expression $H = -\sum p_\nu \log p_\nu$ for the entropy of a discrete distribution has been discussed by Shannon [4] and Khintchin [2]. Goldman [1] has given an explanation of the entropy

$$H = -\int \cdots \int f(x_1, \cdots, x_n) \log f(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

of a continuous distribution on the basis of the discrete case. On the other hand Reich [3] has derived directly the expression for the information rate of a continuous distribution from some postulates. In this short paper we shall try to give another explanation of the entropy of a continuous distribution which is rather similar to the one in the discrete case.

2. Let $f(x_1, \cdots, x_n)$ denote the probability density function of the joint distribution of random variables $X_1, \cdots, X_n$. And we set

POSTULATE I. *The entropy $H(X_1, \cdots, X_n)$ of $(X_1, \cdots, X_n)$ is determined by f alone.*

Owing to this postulate we shall denote $H(X_1, \cdots, X_n)$ as $H(f)$.

Secondly, let $g_S(x_1, \cdots, x_n)$ denote the probability density function of the uniform distribution on a subset S with finite positive measure of the n-dimensional Euclidean space E_n.

POSTULATE II. *If f is the probability density function of an n-dimensional distribution where $f \fallingdotseq g_S$ (a.e.) and car. $(f) \subset S$, then $H(f) < H(g_S)$.*

Let $\phi(x_1, \cdots, x_k)$ be the probability density function of the radom variable $A \equiv (X_1, \cdots, X_k)$ and $\psi_{x_1, \ldots, x_k}(x_{k+1}, \cdots, x_n)$ be the conditional probability density function of the random variable $B \equiv (X_{k+1}, \cdots, X_n)$ under the condition $X_1 = x_1, \cdots, X_k = x_k$. We set

POSTULATE III. $$H(AB) = H(A) + H_A(B),$$

Received October 30, 1958.

i. e.

$$H(f)=H(\phi)+\int\cdots\int_{E_k} H(\psi_{x_1,\dots,x_k})\phi(x_1,\cdots,x_k)\,dx_1\cdots dx_n.$$

Lastly we make the following assumption:

POSTULATE IV. *If f takes the finitely many values $c_1, \cdots, c_s$ and*

$$\mu_\nu=\int\cdots\int_{A_\nu} dx_1\cdots dx_n \qquad \text{for} \quad \nu=1,\cdots,s,$$

where $A_\nu=\{(x_1,\cdots,x_n);\ f(x_1,\cdots,x_n)=c_\nu\}$, then $H(f)$ is the function of the variables $c_1,\cdots,c_s$ and $\mu_1,\cdots,\mu_s$ only, and does not depend on the dimension number n, where $\sum_{\nu=1}^{s} c_\nu\mu_\nu=1$.

This postulate shows that the entropy is invariant by relabelling the states or the transform preserving the probability measure. In the following, we shall use this postulate in the case $s=2$ only, i. e. where f is the probability density function of a uniform distribution. However the independence of the dimension number n will play an important role in the sequel.

3. THEOREM. *Under the postulates* I, II, III, IV, *we have*

$$H(f)=-\lambda\int\cdots\int_{E_n} f(x_1,\cdots,x_n)\log f(x_1,\cdots,x_n)\,dx_1\cdots dx_n \tag{1}$$

where λ is a positive constant.

Proof. In the first place we consider the case $f=g_S$ where g_S is the probability density function of the uniform distribution on the measurable subset S of E_n. Since

$$g_S(x_1,\cdots,x_n)=\begin{cases} p\equiv\left(\int\cdots\int_S dx_1\cdots dx_n\right)^{-1} & \text{on } S,\\ 0 & \text{otherwise,}\end{cases}$$

we find by Postulate IV that $H(g_S)$ is a function of p only and does not depend on the shape and the position of S and the dimension number n of the space in which S is lying. Let $L(p)$ denote this function $H(g_S)$ of p. From Postulate II we get easily that

$$L(p)<L(p') \qquad \text{for} \quad p>p'. \tag{2}$$

To investigate the character of $L(p)$, we consider the probability density function $g_D(x_1,\cdots,x_r)$ of the uniform distribution on the r-dimensional direct set $D\equiv[0,1/p]\times\cdots\times[0,1/p]$. Since

$$g_D(x_1,\cdots,x_r)=\begin{cases} p^r & \text{on } D,\\ 0 & \text{otherwise,}\end{cases}$$

we have $H(g_D)=L(p^r)$. On the other hand, it is easily verified, by Postulate III, that

$$H(ABC\cdots) = H(A) + H(B) + H(C) + \cdots,$$

where A, B, C, $\cdots$ are mutually independent random variables. Therefore, noting that every marginal distribution of the uniform distribution on D is the uniform distribution on the interval $[0, 1/p]$, we have

$$\begin{aligned} H(g_D) &= H(g_{[0,1/p]}) + H(g_{[0,1/p]}) + \cdots \\ &= rH(g_{[0,1/p]}) = rL(p). \end{aligned}$$

Consequently, we have

(3) $$L(p^r) = rL(p) \qquad \text{for} \quad r = 1, 2, \cdots.$$

Since $L(1) = rL(1)$ in the case $p = 1$, we get

(4) $$L(1) = 0.$$

Similarly we have

$$g_{D'}(x_1, x_2) = \begin{cases} 1 & \text{on } D' \\ 0 & \text{otherwise,} \end{cases}$$

where $g_{D'}(x_1, x_2)$ is the probability density function of the uniform distribution on the two-dimensional direct set $D' \equiv [0, p] \times [0, 1/p]$ and both of its marginal distributions are the uniform distributions on $[0, p]$ and $[0, 1/p]$, respectively. Therefore, we have $H(g_{D'}) = H(g_{[0,p]}) + H(g_{[0,1/p]})$ or $L(1) = L(p) + L(1/p)$. Since $L(p^{-1}) = -L(p)$ by this relation and (4), we have, by using (3),

(5) $$L(p^{-r}) = L((p^{-1})^r) = rL(p^{-1}) = -rL(p),$$

where r is a positive integer. The relations (4) and (5) show that (3) is sufficient also for $r = 0, -1, -2, \cdots$. For an arbitrary number $p > 0$, a fixed number $q > 1$ and an arbitrary positive integer r, we can find an integer s such that $q^s \leqq p^r < q^{s+1}$. Then we have

$$sL(q) \geqq rL(p) > (s+1)L(q).$$

Since $L(q) < L(1) = 0$, we have

(6) $$\frac{s}{r} \leqq \frac{L(p)}{L(q)} < \frac{s}{r} + \frac{1}{r}.$$

Similarly, by the property of logarithmic function, we get

(7) $$\frac{s}{r} \leqq \frac{\log p}{\log q} < \frac{s}{r} + \frac{1}{r}.$$

From (6) and (7), we have

$$\left| \frac{L(p)}{L(q)} - \frac{\log p}{\log q} \right| < \frac{1}{r}.$$

Since r can be chosen arbitrarily large, we get $L(p)/\log p = L(q)/\log q$, which means that

(8) $$L(p) = -\lambda \log p$$

where $\lambda = -L(q)/\log q$ is a constant. By (2), we have $\lambda > 0$.

Now we shall consider the more general case, where f is the probability density function of an n-dimensoinal distribution. Taking the random variables $X_1, \cdots, X_{n+1}$ whose joint distribution is the uniform distribution on the subset

$$E \equiv \{(x_1, \cdots, x_{n+1});\ 0 \leqq x_{n+1} \leqq f(x_1, \cdots, x_n)\}$$

of E_{n+1}, we find for its probability density function that

$$g_E(x_1, \cdots, x_{n+1}) = \begin{cases} 1 & \text{on } E, \\ 0 & \text{otherwise,} \end{cases}$$

because

$$\int \cdots \int_E dx_1 \cdots dx_{n+1} = \int \cdots \int_{E_n} f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n = 1.$$

Then we have

$$(9) \qquad H(X_1, \cdots, X_{n+1}) = H(g_E) = L(1) = 0.$$

Since the conditional probability distribution of X_{n+1} under the condition $X_1 = x_1, \cdots, X_n = x_n$ is the uniform distribution on the interval $[0, f(x_1, \cdots, x_n)]$, we get by (8) that

$$H_{X_1, \ldots, X_n}(X_{n+1}) = \int \cdots \int_{E_n} L\left(\frac{1}{f(x_1, \cdots, x_n)}\right) f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n$$

$$= \lambda \int \cdots \int_{E_n} f(x_1, \cdots, x_n) \log f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n.$$

By Postulate III, i. e.

$$H(X_1, \cdots, X_{n+1}) = H(X_1, \cdots, X_n) + H_{X_1, \ldots, X_n}(X_{n+1})$$

and (9), the relation

$$(10) \qquad H(X_1, \cdots, X_n) = -H_{X_1, \ldots, X_n}(X_{n+1})$$

is obtained and this means that

$$H(f) = -\lambda \int \cdots \int_{E_n} f(x_1, \cdots, x_n) \log f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n$$

which was to be proved, since the probability density function of the random variables $(X_1, \cdots, X_n)$ is clearly $f(x_1, \cdots, x_n)$.

In conclusion, the author expresses his sincerest thanks to Prof. K. Kunisawa who has suggested this investigation and given valuable advice.

References

[1] Goldman, S., Information theory. Prentice-Hall, New York (1954).
[2] Khinchin, A. I, Mathematical foundations of information theory. New York

(1957). (Translated into English)

[3] REICH, E., On the definition of informations. Journ. Math. Phys., M.I.T. **30** (1951), 156–162.

[4] SHANNON, C. E., The mathematical theory of communication. Bell Syst. Tech. Journ., **27** (1948), 379–423, 623–656.

DEPARTMENT OF MATHEMATICS,
TOKYO INSTITUTE OF TECHNOLOGY.