A NOTE ON THE CHARACTERIZATION OF SHANNON-WIENER'S MEASURE OF INFORMATION

By Sadao Ikeda

(Received Jan. 26, 1962)

Summary

In [1] the author treated a characterization problem of the Shannon-Wiener measure of information for continuous probability distributions defined over an abstract measure space (R, S, m), where m is a σ-finite measure over a σ-field S of subsets of R, whose range $M(S)$ is such that $M(S)=[0, \infty]$. This condition on the range of the basic measure, however, can slightly be altered such that $M(S)=[0, 1]$, and this modification is useful for characterization of the Kullback-Leibler mean information.

In the present paper, it is shown that the characterization procedure of [1] can be applicable to continuous probability distributions defined on a finite measure space.

1. Characterization

In the first place, we shall introduce, for the sake of precision of our discussion, some notations about the basic measure space and the family of probability distributions, to which our procedure of characterizing the information will be applied. These notations were not so clear in [1].

Let (R_0, S_0, m_0) be a finite measure space satisfying the condition

$$M(S_0)=[0, 1]\,, \tag{1}$$

where $M(S_0)=\{m_0(E)\,;\, E\in S_0\}$. Then it is clear that the product measure space $(R_0^n, S_0^n, m_0^n)=(R_0\times R_0\times\cdots\times R_0, S_0\times S_0\times\cdots\times S_0, m_0\times m_0\times\cdots\times m_0)$ is also a finite measure space satisfying the condition (1), i.e., $M(S_0^n)=[0, 1]$. Denote by $\mathscr{T}$ the set of all (finite) products of the space (R_0, S_0, m_0), which is assumed to be fixed, that is,

$$\mathscr{T}=\{(R_0^n, S_0^n, m_0^n)\,;\, M(S_0)=[0, 1] \text{ and } n \text{ is finite}\}\,.$$

It will easily be noticed that a product of any two spaces belonging to $\mathscr{T}$ is also a member of $\mathscr{T}$.

For any space (R, S, m) belonging to $\mathscr{T}$, denote by $V(R, S, m)$ the family of all probability distributions defined on (R, S, m) which are absolutely continuous with respect to m, while $\bar{V}(R, S, m)$ designates sub-family of $V(R, S, m)$ consisting of all simple probability distributions, i.e., of those having generalized probability density functions (gpdf., in short) which are simple up to an equivalence (m). In both of these definitions, of course, the carriers of gpdf.'s are not necessarily the same to the whole space R. Hereafter throughout the present paper, $D(X)$ designates the carrier of a determination of gpdf., p(x), of a probability distribution (or, equally, a random variable) X.

Let $V(\mathscr{T})$ be the totality of members of $V(R, S, m)$ for all (R, S, m)'s belonging to $\mathscr{T}$, i.e.,

$$V(\mathscr{T}) = \cup\{V(R, S, m)\,;\ (R, S, m) \in \mathscr{T}\}\,,$$

and analogously we shall define

$$\bar{V}(\mathscr{T}) = \cup\{\bar{V}(R, S, m)\,;\ (R, S, m) \in \mathscr{T}\}\,.$$

Under this situation of basic measure space, we shall prove a continuity property of Shannon-Wiener's information measure.

THEOREM 1. *Let (R, S, m) be a finite measure space, and let X and $\{X_i\}$, $(i=1, 2, \cdots)$, be the probability distributions belonging to $V(R, S, m)$ with respective gpdf.'s $p(x)$ and $\{p_i(x)\}$, $(i=1, 2, \cdots)$. If the conditions*

(2) $D(X_i) \subseteqq D(X)\ (m),\ (i=1, 2, \cdots)$ *and* $m(D(X)-D(X_i)) \to 0,\ (i\to\infty)$,

and

$$d_i(X, X_i) \equiv \operatorname*{ess.\,sup}_{x\in D(X_i)} |p(x)-p_i(x)| \to 0,\ \ (i\to\infty)\,, \tag{3}$$

are satisfied, then it holds that

$$\int_R p_i(x)\ \log\ p_i(x)\ \mathrm{dm} \to \int_R p(x)\ \log\ p(x)\ \mathrm{dm},\ (i\to\infty)\,. \tag{4}$$

PROOF. Without any loss of generality we can assume that $D(X)=R$. Put, as in the proof of Theorem 2.1 of [1],

$$f(x) = p(x)\ \log\ p(x)\ \text{ and }\ f_i(x) = p_i(x)\ \log\ p_i(x), \qquad (i=1, 2, \cdots)\,.$$

Then, by Lemma 2.1 (ii) of [1] and the condition (3) above, it holds that, for any $\varepsilon>0$ there exists a positive integer $N=N(\varepsilon)$ such that $i\geqq N$ implies

$$|f(x)-f_i(x)|\leqq\varepsilon(p(x)+1+\varepsilon)\ (m),\ \text{on}\ D(X_i)\,. \tag{5}$$

First we shall consider the case when the Shannon-Wiener information for X, i.e., the second member of the expression (4) is definite. The inequality (5) implies that

$$f_i(x)\leqq f(x)+\varepsilon(p(x)+1+\varepsilon)\ (m),\ \text{on}\ D(X_i)\,,$$

for $i\geqq N$. Hence, defining a non-negative, integrable (m) function $K(x)$ by

$$K(x)=|f(x)|+p(x)+2,\ \text{on}\ R\,,$$

we get

$$f_i(x)\leqq K(x)\ (m),\ \text{on}\ R\,,$$

for all $i\ (\geqq N)$. Since the conditions (2) and (3) imply the almost everywhere (m) convergence of $f_i(x)$ to $f(x)$ as $i\to\infty$, it follows from Lemma 2.2 (a) (Lebesgue's convergence theorem) that

$$\int_R f_i(x)\,\mathrm{dm}\to\int_R f(x)\,\mathrm{dm},\quad (i\to\infty)\,,$$

which is (4).

Secondly, we shall examine the case when the second member of (4) is infinite (positive). Since the function of t, $\varphi(t)=t\log t\ (t>0)$ is bounded from the below, it will easily be seen that Lemma 2.2 (b) of [1], and the a.e.(m) convergence of $f_i(x)$ to $f(x)$ guaranteed by (2) and (3) imply that

$$\int_R f_i(x)\,\mathrm{dm}\to\infty,\quad (i\to\infty)\,,$$

which is identical with (4) in this case. Thus the proof of the theorem is completed.

It will be noticed that the condition (3) of the above theorem is weaker than the condition (2.7) of Theorem 2.1 of [1], and the present result has wider applicability.

Now, we proceed to a characterization problem. Consider a function $H(X)$ defined over $V(\mathscr{T})$, which ranges in the space $W_0=W\cup\{\infty\}$, where W is the whole real line, and depends only on a determination of gpdf. of X. In addition to the conditions stated in the beginning of this section, we assume that for any members X and Y of $V(\mathscr{T})$ the conditional probability distribution of Y exists under the condition that X is fixed.

We shall postulate some assumptions on $H(X)$, among which the last one is slightly weaker than that of [1]. For any E in S such that $m(E)>0$, a probability distribution X_E will be called a uniform distribution if it has a gpdf. given by

$$p_E(x)=\begin{cases}1/m(E), & \text{on } E\,,\\ 0, & \text{elsewhere, } (m)\,.\end{cases}$$

Assumption I. For the uniform distributions

(i) for any E in S of any space (R, S, m) belonging to $\mathscr{T}$, it holds that $0\leqq H(X_E)<\infty$, and

(ii) for any X_E in $\bar{V}(R, S, m)$ and any $X_{E'}$ in $\bar{V}(R', S', m')$, the relations $m(E)>$and$=m'(E')$ imply that $H(X_E)<$and$=H(X_{E'})$, respectively.

Assumption II. Let X be in $\bar{V}(R, S, m)$ and X' be in $\bar{V}(R', S', m')$. If at least one of the three members, $H(X)$, $H_X(X')$ and $H(X, X')$ is finite, then it holds that

$$H(X, X')=H(X)+H_X(X')\,, \tag{6}$$

where $H(X, X')$ designates the value of H for the joint distribution (X, X') which is a member of $\bar{V}(R\times R', S\times S', m\times m')$, and

$$H_X(X')=\mathscr{E}^X[H(X'\mid X)]=\int_R H(X'\mid x)p(x)\,\mathrm{dm}\,,$$

in which $H(X'\mid x)$ is the value of H corresponding to the conditional probability distribution of X' given $X=x$.

Assumption III. Let X be a member of $V(R, S, m)$ with gpdf. $p(x)$, and let $\{X_i\}$, $(i=1, 2, \cdots)$, be a sequence of members of $V(R, S, m)$, with gpdf.'s $\{p_i(x)\}$, $(i=1, 2, \cdots)$. Then the conditions

$$\begin{cases}D(X_i)\subseteqq D(X),\ (i=1, 2, \cdots) \text{ and } m(D(X)-D(X_i))\rightarrow 0,\ (i\rightarrow\infty)\,,\\ d_i(X, X_i)\equiv \operatorname*{ess.\,sup}_{x\in D(X_i)} |p(x)-p_i(x)|\rightarrow 0,\quad (i\rightarrow\infty)\end{cases} \tag{7}$$

imply that

$$H(X_i)\rightarrow H(X),\quad (i\rightarrow\infty)\,. \tag{8}$$

Under these assumptions we can show the following

THEOREM 2. *If a function $H(X)$ satisfies assumptions I to III, then it is expressed as*

$$H(X)=\int_{E} p(x) \log p(x) \, dm , \tag{9}$$

up to a multiplicative positive constant depending only on $\mathscr{T}$, where $p(x)$ designates any determination of gpdf. of X.

A proof of this theorem can be given in a similar manner as that of [1]. Firstly, for a uniform distribution X_E with $m(E)=v(>0)$, it holds that

$$H(X_E)=H(v)=c \log\frac{1}{v} , \quad (c>0) , \tag{10}$$

and hence, if $v=1$, then $H(v)=0$, as was described in Lemma 3.1 of [1].

Secondly, it will be noted that it is sufficient to prove Lemma 3.2 of [1] only in the case when $Z=\{A_i\}$, $(i=1, 2, \cdots, k)$, is a finite m-partition of E. Let X_Z be a probability distribution in $\bar{V}(R, S, m)$ with simple gpdf. defined by

$$P_Z(x)=\begin{cases} p_i/v_i, & \text{on } A_i \quad (i=1, 2, \cdots , k) , \\ 0, & \text{elsewhere,} \end{cases} \tag{11}$$

where $p_i>0$, $v_i=m(A_i)$ and $\sum_{i=1}^{k} p_i=1$. Then under the Assumptions I and II, it is shown that

$$H(X_Z)=c\sum_{i=1}^{k} p_i \log \frac{p_i}{v_i} , \quad (c>0) , \tag{12}$$

as will be seen in the following.

Since Z is a finite m-partition, we can take a corresponding set $Z'=\{A_i'\}$, $(i=1, 2, \cdots , k)$, of disjoint subsets of R such that

$$m(A_i')=p_i/\lambda v_i, \quad (i=1, 2, \cdots , k) \tag{13}$$

where $\lambda=\sum_{i=1}^{k} p_i/v_i$, for which it is easy to see that $1\leqq\lambda<\infty$. Denote by $X_{Z'}$ the probability distribution such that, if X_Z falls in A_i then $X_{Z'}$ is distributed uniformly on A_i' for $i=1, 2, \cdots, k$. The joint distribution of X_Z and $X_{Z'}$, $(X_Z, X_{Z'})$, has a gpdf. on the product measure space $(R\times R,\ S\times S,\ m\times m)$ given by

$$p(x, x')=\begin{cases} \lambda, & \text{on } A_i\times A_i', \quad (i=1, 2, \cdots , k) , \\ 0, & \text{elsewhere,} \end{cases} \tag{14}$$

that is, $(X_z, X_{z'})$ is a uniform distribution on the set $F=\sum_{i=1}^{k} A_i \times A_i'$ with $m\times m(F)=1/\lambda$. Hence, it follows from (10) that

$$H(X_z, X_{z'})=c \log \lambda , \tag{15}$$

where c is a positive constant.

On the other hand, by (10) and (13) we obtain

$$H(X_{z'}|x)=c' \log (\lambda v_i/p_i), \text{ if } x\in A_i, \quad (i=1, 2, \cdots , k) ,$$

from which it follows that

$$\begin{aligned} H_{X_z}(X_{z'}) &= c'\sum_{i=1}^{k} p_i \log \frac{\lambda v_i}{p_i} \\ &= c' \log \lambda - c'\sum_{i=1}^{k} p_i \log \frac{p_i}{v_i} . \end{aligned} \tag{16}$$

By Assumption II we have

$$H(X_z, X_{z'})=H(X_z)+H_{X_z}(X_{z'}) ,$$

which is written, by (15) and (16) as

$$c \log \lambda = H(X_z)+c' \log \lambda - c'\sum_{i=1}^{k} p_i \log \frac{p_i}{v_i} . \tag{17}$$

If, in particular, we take X_z as a simple distribution on R, for which $Z=\{A_i\}$ $(i=1, 2, \cdots , k ;\ k>1)$ constitutes an m-partition of the whole space R and $v_i=p_i$, $(i=1, 2, \cdots, k)$ in (11), then $\lambda=k$ (>1) and it holds, by (17), that $c=c'$.

Hence (12) follows from (17), which shows that the expression (9) is valid for finite-simple distributions.

Finally, corresponding to the proof of Theorem 3.1 of [1], we shall prove (9) for any probability distribution X belonging to $V(R, S, m)$ with a determination $p(x)$ of gpdf.. We can choose a sequence of probability distributions in $V(R, S, m)$, $\{X^{(n)}\}$ $(n=1, 2, \cdots)$, whose gpdf.'s are defined by

$$p^{(n)}(x)=f_n(x)/\mu_n \quad (n=1, 2, \cdots) ,$$

where

$$f_n(x)=\begin{cases} 0 , & \text{if } p(x)<1/n , \\ p(x) , & \text{if } 1/n\leqq p(x)\leqq n , \\ 0 , & \text{if } n<p(x) , \end{cases}$$

and

$$\mu_n=\int_R f_n(x)\mathrm{dm}$$

for $n=1, 2, \cdots$.

From the definition of $p^{(n)}(x)$, it is clear that

$$d_n(X, X^{(n)})\equiv \operatorname*{ess.\,sup}_{x\in D(X^{(n)})} |p(x)-p^{(n)}(x)|\to 0, \quad (n\to\infty) . \tag{18}$$

For each $X^{(n)}$, since $p^{(n)}(x)\log p^{(n)}(x)$ is bounded, there exists a sequence of finite-simple distributions $\{X_i^{(n)}\}$, $(i=1, 2,\cdots)$, with respective gpdf.'s $\{p_i^{(n)}(x)\}$, $(i=1, 2, \cdots)$, defined by

$$p_i^{(n)}(x)=\begin{cases} p_{ij}^{(n)}/v_{ij}^{(n)}, & \text{on } A_{ij}^{(n)} \quad (j=1, 2, \cdots, k_{in}) , \\ 0, & \text{elsewhere,}\end{cases}$$

where $Z_i^{(n)}=A_{ij}^{(n)}$, $(j=1, 2, \cdots, k_{in})$, is a special m-partition of R as was constructed in the proof of Corollary 2.1 of [1], and

$$p_{ij}^{(n)}=\int_{A_{ij}^{(n)}} p^{(n)}(x)\mathrm{dm}, \quad v_{ij}^{(n)}=m(A_{ij}^{(n)}) ,$$

such that

$$d(X^{(n)}, X_i^{(n)})\equiv \operatorname*{ess.\,sup}_{x\in R} |p^{(n)}(x)-p_i^{(n)}(x)|\to 0, \quad (i\to\infty) . \tag{19}$$

Note that, by (12), we have

$$H(X_i^{(n)})=c\sum_{j=1}^{k_{in}} p_{ij}^{(n)} \log \frac{p_{ij}^{(n)}}{v_{ij}^{(n)}} \quad (i=1, 2, \cdots)$$

for every n.

Since the right-hand member of the above expression converges, by (19) and Theorem 1, to the right-hand member of the following expression (20), Assumption III and (19) imply that

$$H(X^{(n)})=c\int_R p^{(n)}(x)\log p^{(n)}(x)\mathrm{dm} \tag{20}$$

for $n=1, 2, \cdots$.

Therefore, it follows from (18), Theorem 1 and Assumption III that

$$H(X)=c\int_R p(x) \log p(x)\mathrm{dm}, \quad (c>0)$$

as is expected, and our characterization procedure is complete, if we remark that Assumption I implies the last statement of the theorem concerning the multiplicative constant.

DEPT. MATH., COLLEGE OF SCI. AND ENG., NIHON UNIVERSITY

REFERENCE

[1] S. Ikeda, "Continuity and characterization of Shannon-Wiener information measure for continuous probability distributions," *Ann. Inst. Stat. Math.*, Vol. XI (1959), pp. 131-144.