A REMARK ON THE CONVERGENCE OF KULLBACK-LEIBLER'S MEAN INFORMATION

By SADAO IKEDA

(Received June 10, 1960)

Summary

Under fairly general assumptions two sufficient conditions for the convergence of the Kullback-Leibler mean information are obtained, which are generalizations of the conditions given in Lemma 2.1 and Theorem 2.2 of Chapter 4 in S. Kullback [1].

1. A convergence theorem

Let (R, S, m) be a σ-finite measure space, and $\{\mu, \nu\}$ a set of two measures on the measurable space (R, S) dominated by the measure m. Let $f(x)$ and $g(x)$ be the Radon-Nikodym derivatives, i.e., the density functions (with respect to measure m) of measures μ and ν, respectively. $D(f)$ and $D(g)$ denote the carriers of $f(x)$ and $g(x)$.

Assume measures μ and ν are absolutely continuous with respect to each other. Then, it follows that $D(f)=D(g)(m)$, and vice versa. Under this assumption, the "mean information for discrimination in favor of μ against ν" is defined by

$$I(f:g)=\int_R f(x)\log\frac{f(x)}{g(x)}dm(x)\ . \tag{1.1}$$

For brevity, we call the above expression the "Kullback-Leibler mean information".

If both measures μ and ν are finite, then the inequality

$$\mu(R)\log\frac{\mu(R)}{\nu(R)}\leqq I(f:g) \tag{1.2}$$

will be obtained, where the equality holds when and only when the ratio of $f(x)$ to $g(x)$ is a constant for almost all (m) x in R. In particular, if both μ and ν are the probability measures, then (1.2) becomes

$$0\leqq I(f:g)\ , \tag{1.3}$$

where the equality holds if and only if $f(x)=g(x)(m)$ on R.

The following theorem gives a sufficient condition for the convergence of the mean information (1.1) for finite measures.

THEOREM 1.1. *Assume that*

(i) *$f(x)$ and $g(x)$ are the density functions with respect to m, of two finite measures μ and ν dominated by each other, i.e., $f(x)$ and $g(x)$ are nonnegative almost everywhere (m) on R, $\mu(R)=\int_R f dm<\infty$, and $\nu(R)=\int_R g dm<\infty$, and $D(f)=D(g)=E(m)$,*

(ii) *$\{f_i, g_i\}$ $(i=1, 2, \cdots)$ is a sequence of pairs of the density functions of finite measures μ_i and ν_i dominated by each other, or more precisely, $f_i(x)$ and $g_i(x)$ are nonnegative almost everywhere (m) on R, $\mu_i(R)=\int_R f_i dm<K$, and $\nu_i(R)=\int_R g_i dm<K$, and $D(f_i)=D(g_i)=E_i(m)$ for $i=1, 2, \cdots$, where K is a positive constant independent of i, and* (iii) *$E_i \subset E(m)$ $(i=1, 2, \cdots)$ and $\nu(E-E_i)\to 0$ as $i\to\infty$.*

Under these assumptions, it holds that

$$I(f_i : g_i)\to I(f : g) \qquad (i\to\infty), \tag{1.4}$$

if either of the following conditions is satisfied:

$$\text{(A)} \qquad \begin{cases} \text{(a)} \quad h_i(g, g_i)=\operatorname{ess\,sup}_{E_i}\left|1-\dfrac{g_i}{g}\right|\to 0 \quad (i\to\infty), \text{ and} \\ \text{(b)} \quad d_i(p, p_i)=\operatorname{ess\,sup}_{E_i}|p-p_i|\to 0 \quad (i\to\infty) \end{cases}$$

and

$$\text{(B)} \qquad \begin{cases} \text{(a)}' \quad h_i(g, g_i)=\operatorname{ess\,sup}_{E_i}\left|1-\dfrac{g_i}{g}\right|\to 0 \quad (i\to\infty), \text{ and} \\ \text{(b)}' \quad h_i(p, p_i)=\operatorname{ess\,sup}_{E_i}\left|1-\dfrac{p_i}{p}\right|\to 0 \quad (i\to\infty) \end{cases}$$

where

$$p(x)=\frac{f(x)}{g(x)}(m), \quad \text{and} \quad p_i(x)=\frac{f_i(x)}{g_i(x)}(m) \qquad (i=1, 2, \cdots),$$

respectively on E and E_i, and ess sup *is taken with respect to measure m.*

PROOF. 1°) In the first place, we consider the case when

$$|I(f : g)|<\infty . \tag{1.5}$$

From definition (1.1) we can write as

$$\begin{aligned} I(f : g)&=\int_E p \log p d\nu , \\ \text{and} \quad I(f_i : g_i)&=\int_{E_i} p_i \log p_i d\nu_i \qquad (i=1, 2, \cdots), \end{aligned} \tag{1.6}$$

therefore, we have

$$(1.7) \qquad |I(f:g)-I(f_i:g_i)| \leqq \left|\int_{E_i} p\log p d\nu - \int_{E_i} p_i \log p_i d\nu_i\right| + \left|\int_{E-E_i} p \log p d\nu\right| \qquad (i=1, 2, \cdots).$$

The second term of the right-hand side of (1.7) converges to zero as $i\to\infty$, by virtue of assumption (iii) and the finiteness of $I(f:g)$. The first term of the right-hand side of (1.7) will be evaluated by the following inequality:

$$(1.8) \quad \left|\int_{E_i} p\log p d\nu - \int_{E_i} p_i \log p_i d\nu_i\right| \leqq \left|\int_{E_i} p\log p d\nu - \int_{E_i} p \log p d\nu_i\right| + \left|\int_{E_i} p\log p d\nu_i - \int_{E_i} p_i \log p_i d\nu_i\right| .$$

First, we consider the first part of the right-hand side of (1.8). Since

$$(1.9) \qquad \left|\int_{E_i} p\log p d\nu - \int_{E_i} p \log p d\nu_i\right| \leqq \int_{E_i} |p\log p|\,|g-g_i|\,dm \leqq h_i(g, g_i)\int_{E_i} |p\log p|\,d\nu ,$$

it follows from assumption (iii), (1.5) and condition (a) of (A) or (B) that

$$(1.10) \qquad \left|\int_{E_i} p\log p d\nu - \int_{E_i} p \log p d\nu_i\right| \to 0 \qquad (i\to\infty).$$

Next, we consider the second part of the right-hand side of (1.8). From Lemma 2.1 (ii) of the author's paper [2], and condition (A), it will easily be seen that, for any $\varepsilon>0$, there exists a positive integer N such that $N\leqq i$ implies $h_i(g, g_i)<\varepsilon$ and $d_i(p, p_i)<\varepsilon$, and

$$(1.11) \quad \left|\int_{E_i} p\log p d\nu_i - \int_{E_i} p_i \log p_i d\nu_i\right| \leqq \int_{E_i} |p\log p - p_i \log p_i|\,d\nu_i \leqq \int_{E_i} \varepsilon(p+1+\varepsilon) d\nu_i = \varepsilon\left\{\int_{E_i} pg_i dm + (1+\varepsilon)\int_{E_i} g_i dm\right\} .$$

Since $g_i(x)<(1+\varepsilon)g(x)(m)$ on E_i, it holds that

$$\int_{E_i} pg_i dm \leqq (1+\varepsilon)\int_{E_i} f dm .$$

Therefore, from assumptions (i) and (ii) it will easily be seen that the values of the members within the bracket in the last expression of (1.11) are bounded. Hence, we obtain

$$\left|\int_{E_i} p\log p\,d\nu_i-\int_{E_i} p_i\log p_i\,d\nu_i\right|\to 0 \qquad (i\to\infty), \tag{1.12}$$

under condition (A). When condition (B) is satisfied instead of (A), the above convergence (1.12) will be shown as follows: from Lemma 2.1 (i) of the author's paper [2], we have

$$\begin{aligned}\left|\int_{E_i} p\log p\,d\nu_i-\int_{E_i} p_i\log p_i\,d\nu_i\right| &\leqq \int_{E_i} |p\log p-p_i\log p_i|\,d\nu_i\\ &\leqq \int_{E_i}\left|1-\frac{p_i}{p}\right|(|p\log p|+p+p_i)\,d\nu_i\\ &\leqq h_i(p,p_i)\left\{\int_{E_i}|p\log p|\,g_i\,dm+\int_{E_i} pg_i\,dm+\int_{E_i} f_i\,dm\right\}.\end{aligned} \tag{1.13}$$

By the investigation analogous to that of the case of condition (A) above, the values of the members in the bracket of the last expression (1.13) are bounded for sufficiently large i. Hence, (1.12) follows from (1.13) and condition (B).

Thus our theorem is proved in the case when $|I(f:g)|<\infty$.

2°) Secondly, we shall prove the theorem in the case when

$$I(f:g)=\infty\ . \tag{1.14}$$

For this case it will be shown that

$$I(f_i:g_i)\to\infty \qquad (i\to\infty). \tag{1.15}$$

For each positive integer N, we define the function such as

$$f^N(x)=\begin{cases} f(x)\,, & \text{on}\quad E\cap\{x:p(x)\leqq N\}\,,\\ Ng(x)\,, & \text{on}\quad E\cap\{x:p(x)\leqq N\}\,,\\ 0\,, & \text{otherwise,}\end{cases}$$

(1.16) and

$$f_i^N(x)=\begin{cases} f_i(x)\,, & \text{on}\quad E_i\cap\{x:p_i(x)\leqq N\}\,,\\ Ng_i(x)\,, & \text{on}\quad E_i\cap\{x:p_i(x)>N\}\,,\\ 0\,, & \text{otherwise}\end{cases} \qquad (i=1,2,\cdots).$$

Put $p^N(x)=f^N(x)/g(x)$ on E and $p_i^N(x)=f_i^N(x)/g_i(x)$ on E_i $(i=1,2,\cdots)$. Then, these functions will be definite except for the set of m-measure zero on each of the sets E and E_i's, respectively.

For these functions we consider the mean information such as

$$I(f^N:g)=\int_E p^N\log p^N\,d\nu\ ,$$

(1.17) and

$$I(f_i^N:g_i)=\int_{E_i} p_i^N\log p_i^N\,d\nu_i \qquad (i=1,2,\cdots).$$

Since the integrands of the expressions of the right-hand sides of (1.17) are all bounded from above and measures ν and ν_i's are all finite measures, it follows from (1.2) that

$$|I(f^N:g)|<\infty\,,\quad\text{and}\quad |I(f_i^N:g_i)|<\infty \qquad (i=1,2,\cdots), \tag{1.18}$$

for any fixed N. Moreover, since $p^N(x)$ and $p_i^N(x)$'s coincide with $p(x)$ and $p_i(x)$'s respectively on the domains where they are less than or equal to N ($\geqq 1$), and p^N and p_i^N's are all monotone nondecreasing functions of N, it holds that

(1.19) and
$$\begin{aligned}&I(f^N:g)\uparrow I(f:g) \quad (N\to\infty),\\ &I(f_i^N:g_i)\uparrow I(f_i:g_i) \quad (N\to\infty) \text{ for any fixed } i \qquad (i=1,2,\cdots).\end{aligned}$$

First, we shall show that, for any fixed N,

$$I(f_i^N:g_i)\to I(f^N:g) \qquad (i\to\infty). \tag{1.20}$$

Since condition (1.5) in the proof of case 1° is fulfilled for $I(f^N:g)$ by virtue of (1.18), in order to show the convergence (1.20) it will be sufficient to confirm that functions f^N, g and $\{f_i^N, g_i^N\}$ $(i=1,2,\cdots)$ satisfy all assumptions (i)-(iii) and condition (A) or (B) of the present theorem. It will be evident that they satisfy all assumptions (i)-(iii). Since the definitions of functions p^N and p_i^N's in (1.16) do not change the values of functions g and g_i's, and it holds that

$$\begin{aligned}&d_i(p^N, p_i^N)\leqq d_i(p, p_i)\,,\quad\text{and}\\ &h_i(p^N, p_i^N)\leqq h_i(p, p_i) \qquad (i=1,2,\cdots;\ N=1,2,\cdots),\end{aligned} \tag{1.21}$$

it is seen that condition (A) or (B) is fulfilled for our present case if (A) or (B) is satisfied for f, g and $\{f_i, g_i\}$ $(i=1,2,\cdots)$, respectively. Hence, (1.20) holds true for any fixed N.

From (1.14) and (1.19) it follows that

$$I(f^N:g)\to\infty \qquad (N\to\infty). \tag{1.22}$$

Therefore, for any M (>0), there exists a positive integer N' such that

$$M+1<I(f^{N'}:g)\,. \tag{1.23}$$

It will be seen from (1.20) that, for this N' there exists a positive integer N'' such that $N''\leqq i$ implies

$$|I(f_i^{N'}:g_i)-I(f^{N'}:g)|<1\,. \tag{1.24}$$

Hence, it follows from (1.19), (1.23) and (1.24) that

$$M<I(f_i:g_i) \qquad \text{for } i\geqq N'', \tag{1.25}$$

which implies (1.15).

This completes the proof of the theorem.

2. Corollaries

The result of Lemma 2.1 in Chapter 4 of S. Kullback [1] states that a necessary and sufficient condition for the convergence of the mean information $I(f_i:f)$ to zero for $i\to\infty$, where f and f_i's are generalized probability density functions, is given by

$$h(f,f_i)=\operatorname*{ess\,sup}_E \left|1-\frac{f_i}{f}\right|\to 0 \qquad (i\to\infty), \tag{2.1}$$

where the set E is the carrier of f and f_i's. In general, however, this condition is not a necessary condition, as will be seen by some simple examples. Corollary 2.1 below shows that our theorem 1.1 gives the same condition as (2.1) which is sufficient for the convergence $I(f_i:f)\to 0$ for $i\to\infty$.

Theorem 2.2 in Chapter 4 of S. Kullback [1] is concerned with the convergence of $I(f_i:g)$ to $I(f:g)$ for $i\to\infty$, when the functions f, g and f_i's are generalized probability density functions with the same carrier. A necessary condition was given by the same one as (2.1) above under the assumption that $I(f:g)$ is finite, but Corollary 2.2 below does not require the finiteness of $I(f:g)$.

The notation $(f_i, g_i; f, g)\Rightarrow(f_i, f; f, f)$, for example, in Corollary 2.1 below means that functions f_i, f, f and f are taken instead of f_i, g_i, f and g in Theorem 1.1.

COROLLARY 2.1. $(f_i, g_i; f, g)\Rightarrow(f_i, f; f, f)$.

Assume that

(i) f and $\{f_i\}$ $(i=1, 2, \cdots)$ are generalized probability density functions with respect to m, with $D(f)=D(f_i)=E(m)$ $(i=1, 2, \cdots)$. Then, if the condition

$$h(f,f_i)=\operatorname*{ess\,sup}_E \left|1-\frac{f_i}{f}\right|\to 0 \qquad (i\to\infty) \tag{2.2}$$

is satisfied, then it holds that

$$I(f_i:f)\to 0 \qquad (i\to\infty). \tag{2.3}$$

COROLLARY 2.2. $(f_i, g_i; f, g) \Rightarrow (f_i, g; f, g)$.

Assume that

(i) f, g and $\{f_i\}$ $(i=1, 2, \cdots)$ are generalized probability density functions with respect to m, with $D(f)=D(g)=D(f_i)=E(m)$ $(i=1, 2, \cdots)$.

Under this assumption, if the condition

$$h(f, f_i)=\operatorname*{ess\,sup}_E \left|1-\frac{f_i}{f}\right| \to 0 \qquad (i\to\infty) \tag{2.4}$$

is satisfied, then it holds that

$$I(f_i : g) \to I(f : g) \qquad (i\to\infty). \tag{2.5}$$

COROLLARY 2.3. $(f_i, g_i : f, g) \Rightarrow (f, f_i : f, f)$.

Under the assumption of Corollary 2.1, if the condition (2.2) is satisfied, then it holds that

$$I(f : f_i) \to 0 \qquad (i\to\infty). \tag{2.6}$$

COROLLARY 2.4. $(f_i, g_i; f, g) \Rightarrow (f, g_i; f, g)$.

Assume that

(i) f, g and $\{g_i\}$ $(i=1, 2, \cdots)$ are generalized probability density functions with respect to m, with $D(f)=D(g)=D(g_i)=E(m)$ $(i=1, 2, \cdots)$.

Then, the condition

$$h(g, g_i)=\operatorname*{ess\,sup}_E \left|1-\frac{g_i}{g}\right| \to 0 \qquad (i\to\infty) \tag{2.7}$$

implies that

$$I(f : g_i) \to I(f : g) \qquad (i\to\infty). \tag{2.8}$$

Finally, we consider two types of truncation of the generalized probability density function; suppose μ and ν are two probability measures on the measurable space (R, S) which are absolutely continuous with respect to m, with densities $f(x)$ and $g(x)$ and with $D(f)=D(g)=E(m)$. Let $\{E_i\}$ $(i=1, 2, \cdots)$ be a sequence of sets in S such that $E_i \subset E(m)$ $(i=1, 2, \cdots)$ and $\nu(E-E_i)\to 0$ as $i\to\infty$. Define

$$f^i(x)=\begin{cases} f(x) & \text{on } E_i \\ 0 & \text{otherwise} \end{cases}$$

(2.9) and

$$g^i(x)=\begin{cases} g(x) & \text{on } E_i \\ 0 & \text{otherwise} \end{cases} \qquad (i=1, 2, \cdots),$$

and

$$f^{(i)}(x)=\begin{cases} f(x)/\mu(E_i) & \text{on } E_i \\ 0 & \text{otherwise} \end{cases}$$

(2.10) and

$$g^{(i)}(x)=\begin{cases} g(x)/\nu(E_i) & \text{on } E_i \\ 0 & \text{otherwise} \end{cases} \qquad (i=1, 2, \cdots).$$

For these truncated probability density functions, Theorem 1.1 shows also that

$$I(f^i : g^i) \rightarrow I(f : g) \qquad (i\rightarrow\infty), \tag{2.11}$$

and

$$I(f^{(i)} : g^{(i)}) \rightarrow I(f : g) \qquad (i\rightarrow\infty). \tag{2.12}$$

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

REFERENCES

[1] S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, 1959.
[2] S. Ikeda, "Continuity and characterization of Shannon-Wiener information measure for continuous probability distributions," *Ann. Inst. Stat. Math.*, Vol. XI (1959).