A. SUGAWARA
KODAI MATH. J.
8 (1985), 103—111

ON CONVERGENCE OF CONDITIONAL PROBABILITY MEASURES

By Akihiro Sugawara

1. Introduction.

In this paper we are concerned with a relation between convergence of conditional probability measures given sample means and minimization of I-divergence (Kullback-Leibler information quantity) under some constraint. For statistical and information-theoretical meanings of this problem, we refer to Vincze [15]. Similar problem was investigated by Bártfai [3] and Vasicek [14], but our approach is slightly different from theirs. We apply Sanov-type theorems which were obtained by Groeneboom, Oosterhoff and Ruymgaart [6]. They were working in the intrest of rates of convergence for probabilities of large deviations given sample means. We recognize their result as a limit of average information quantity gained by measurement of sample means.

The basic definitions and results which will be used in the sequal are provided in Section 2. The result of Section 3 is not so difficult, but it would be helpful for understanding the following work. Section 4 is our main one. At first we rewrite a large deviation theorem obtained in Groeneboom, Oosterhoff and Ruymgaart [6] employing I-divergence of conditional probability measures given sample means. From this point of view we can show convergence of conditional probability measures in the total variation metric. We also consider a problem of convergence in τ-topology.

2. Preliminaries.

The purpose of this section is to state the basic definitions and the principal results which will be used in what follows.

Let X be a Hausdorff space of points x, $\mathcal{B}$ the σ-field of Borel subsets of X. Let Π be the set of all probability measures on $(X, \mathcal{B})$, which is considered as a convex set in the usual sense: $(a\lambda+(1-a)\mu)(\cdot)=a\lambda(\cdot)+(1-a)\mu(\cdot)$, $0\leqq a\leqq 1$, $\lambda, \mu\in\Pi$. The *I-divergence or Kullback-Leibler information quantity* $I(\lambda|\mu)$ for λ, μ in Π is defined by

$$I(\lambda|\mu)=\int\log\frac{d\lambda}{d\mu}d\lambda \qquad \text{if} \quad \lambda\ll\mu,$$

$$=+\infty \qquad \text{otherwise.}$$

Received April 26, 1984

The information quantity $I(\lambda|\mu)$ is always non-negative and vanishes only for $\lambda=\mu$. If Λ is a subset of Π, we define for μ in Π

$$I(\Lambda|\mu)=\inf\{I(\lambda|\mu):\lambda\in\Lambda\}.$$

If Λ is empty, we put $I(\Lambda|\mu)=+\infty$.

A topology on Π, which we will consider in this paper, is defined by setwise convergence on all Borel sets. In Groeneboom et al. [6], this topology is called τ-topology. A sequence of probability measures $\{\lambda_n\}$ in Π converges to a probability measure $\lambda\in\Pi$ in τ-topology, if and only if,

$$\lim_{n\to\infty}\int g\,d\lambda_n=\int g\,d\lambda$$

for any bounded measurable function $g:X\to R$. For a fixed μ in Π, the function $\lambda\to I(\lambda|\mu)$, $\lambda\in\Pi$, is lower τ-semicontinuous.

Let $\mu\in\Pi$ be fixed. For each positive integer n, let $(X^n, \mathcal{B}^n, \mu^n)$ be the n-th product of a probability space $(X, \mathcal{B}, \mu)$. We define the empirical probability measure $\delta_n(\cdot|x_1, \cdots, x_n)$ on $(X, \mathcal{B})$ for a sample $(x_1, \cdots, x_n)\in X^n$ by

$$\delta_n(B|x_1, \cdots, x_n)=\frac{1}{n}\sum_{i=1}^{n}1_B(x_i), \qquad B\in\mathcal{B}.$$

For a subset Λ of Π, $\hat{\mu}_n[\Lambda]$ which means the probability that the empirical probability measure belongs to Λ is defined by

$$\hat{\mu}_n[\Lambda]=\mu^n\{(x_1, \cdots, x_n):\delta_n(\cdot|x_1, \cdots, x_n)\in\Lambda\}.$$

If Λ is an arbitrary subset of Π, the event $\{(x_i)\in X^n:\delta_n(\cdot|x_1, \cdots, x_n)\in\Lambda\}$ is not necessarily $\mathcal{B}^n$-measurable. Henceforth it will be assumed without explicit reference that $\hat{\mu}_n[\Lambda]$ is well defined for all positive integer n (c.f. Remark 3.1 in Groeneboom et al. [6]).

We are now in position to describe a few theorems which play a very important role in the sequel. They were all obtained in Groeneboom et al. [6].

THEOREM A. *Let $\mu\in\Pi$ and Λ be a nonempty τ-closed subset in Π. Then there exists a probability measure $\mu^*\in\Lambda$ such that*

$$I(\mu^*|\mu)=I(\Lambda|\mu).$$

We remark that if Λ is also a convex subset in Π and $I(\Lambda|\mu)<+\infty$, then μ^* is unique since $I(\lambda|\mu)$ is strictly convex in λ.

THEOREM B. *Let $\mu\in\Pi$ and Λ be a τ-closed subset in Π. Then*

$$\limsup_{n\to\infty}\frac{1}{n}\log\hat{\mu}_n[\Lambda]\leqq-I(\Lambda|\mu).$$

Let $\mu\in\Pi$, let V be a real Hausdorff topological vector space and let $\{B_i:i=1, 2, 3, \cdots\}$ be an increasing sequence of Borel subsets of X such that

$\lim_{m\to\infty}\mu(B_m)=1$. Let $\Pi_m=\{\lambda\in\Pi : \lambda(B_m)=1\}$ for any positive integer m and let $\Pi^*=\bigcup_{m=1}^{\infty}\Pi_m$. Let $T:\Pi^*\to V$ be a transformation whose restriction $T|\Pi_m$ is affine and τ-continuous at each $\lambda\in\Pi_m$ such that $I(\lambda|\mu)<+\infty$, for each positive integer m. Then the following theorem holds.

THEOREM C. *For a convex subset C of V with nonempty interior C° satisfying $I(T^{-1}(C^\circ)|\mu)<+\infty$, then*

$$\lim_{n\to\infty}\frac{1}{n}\log\hat{\mu}_n[T^{-1}(C)]=-I(T^{-1}(C)|\mu).$$

3. Conditional Probabilities.

Let $\mu\in\Pi$ and let A be an event in $\mathcal{B}$ such that $\mu(A)\neq 0$. The conditional probability measure $\mu_A(\cdot)$ is defined as

$$\mu_A(B)=\frac{\mu(B\cap A)}{\mu(A)}, \qquad B\in\mathcal{B}.$$

Under this notion we have the following result. It is not so difficult, but it seems worthwhile to us for understanding intuitively a relation between I-divergence and conditional probability measures.

THEOREM 3.1. *Let $\mu\in\Pi$ and let A be an event in $\mathcal{B}$ satisfying $\mu(A)\neq 0$. Let $\Pi_A=\{\lambda\in\Pi : \lambda(A)=1\}$. Then,*

$$I(\mu_A|\mu)=I(\Pi_A|\mu).$$

Proof. We first recall $I(\mu_A|\mu)=-\log\mu(A)<+\infty$. Therefore we may assume that $\lambda\in\Pi_A$ and $I(\lambda|\mu)<+\infty$. Denote by f the Radon-Nikodym derivative $d\lambda/d\mu$. We also write g to denote $d\mu_A/d\mu=1_A/\mu(A)$. Then,

$$\int f(x)\log g(x)\,\mu(dx)=\int\log g(x)\,\lambda(dx)=-\log\mu(A)$$

$$=\int g(x)\log g(x)\,\mu(dx).$$

Therefore

$$\int g(x)\log g(x)\,\mu(dx)-\int f(x)\log f(x)\,\mu(dx)$$

$$=\int f(x)\log\frac{g(x)}{f(x)}\,\mu(dx)$$

$$\leqq\int f(x)\left(\frac{g(x)}{f(x)}-1\right)\mu(dx)=0,$$

where we use the inequality $\log a\leqq a-1$, $a\geqq 0$. Thus we obtain the following,

$$I(\mu_A|\mu)\leqq I(\lambda|\mu)\,,\qquad \lambda\in\Pi_A\,.$$

This completes the proof.

We can afford another proof of this theorem employing Theorem 2.1 in Csiszár [5]. However we have a preference for a proof which does not require an extra knowledge.

4. Convergences.

Throughout this section we fix a probability measure $\mu\in\Pi$ and a bounded measurable function $f: X\to R^k$, where R^k is k-dimensional Euclidean space.

We consider a transformation $T: \Pi\to R^k$, defined by

$$T(\lambda)=\int f\,d\lambda\,,\qquad \lambda\in\Pi.$$

Clearly $T(\cdot)$ is τ-continuous. Hereafter we write

$$\Pi_f[C]=\left\{\lambda\in\Pi: \int f\,d\lambda\in C\right\}=T^{-1}(C)\,,$$

for a measurable subset C of R^k. We note that the empirical probability measure $\delta_n(\cdot|x_1,\cdots,x_n)$ belongs to $\Pi_f[C]$ if and only if $\frac{1}{n}\sum_{i=1}^{n}f(x_i)\in C$. Thus

$$\hat{\mu}_n[\Pi_f[C]]=\mu^n\left\{(x_i)\in X^n: \frac{1}{n}\sum_{i=1}^{n}f(x_i)\in C\right\}.$$

For convention, we put $f_n(x_1,\cdots,x_n)=\frac{1}{n}\sum_{i=1}^{n}f(x_i)$. Then we write $\mu^n_{[f_n\in C]}(\cdot)$ to denote the conditional probability measure $\mu^n(\cdot|f_n\in C)$.

Now we can rewrite Theorem C employing conditional probability measures in the above case.

THEOREM 4.1. *Let $\mu\in\Pi$ and f be a bounded measurable function with values in R^k. If C is a measurable convex subset of R^k with nonempty interior C° such that $I(\Pi_f[C^\circ]|\mu)<+\infty$, then*

$$\lim_{n\to\infty}\frac{1}{n}I(\mu^n_{[f_n\in C]}|\mu^n)=I(\Pi_f[C]|\mu)\,.$$

For $i=1,2,\cdots,n$, the i-th marginal measure of $\mu^n_{[f_n\in C]}$ is denoted by $\mu^{(i)}_{[f_n\in C]}$. By the symmetry of the event $\{(x_i)\in X^n: f_n\in C\}$,

$$\mu^{(i)}_{[f_n\in C]}=\mu^{(j)}_{[f_n\in C]}\,,\qquad i\neq j\,.$$

Hence we simply write $\mu_{[f_n\in C]}$ to denote the i-th marginal measure. If C is a closed convex set and $I(\Pi_f[C]|\mu)<+\infty$, then there exists the unique probability measure $\mu^*\in\Pi_f[C]$ such that

$$I(\mu^*|\mu)=I(\Pi_f[C]|\mu),$$

by Theorem A.

Then it seems that Theorem 4.1 suggests some relation between $\mu_{(f_n\in C)}$ and μ^*. Really we obtain the following theorems.

THEOREM 4.2. *Let $\mu\in\Pi$ and f be a bounded measurable function with values in R^k. Let C be a measurable convex subset of R^k with interior C° such that $\int f\,d\mu\in C^\circ$. Then,*

$$\mu_{(f_n\in C)}\to\mu \qquad (n\to\infty)$$

in the total variation metric.

Proof. It is easy to see that

$$I(\mu^n_{(f_n\in C)}|\mu^n)=I(\mu^n_{(f_n\in C)}|\otimes_{i=1}^n\mu_{(f_n\in C)})+nI(\mu_{(f_n\in C)}|\mu).$$

The assumption $\int f\,d\mu\in C^\circ$ implies $I(\Pi_f[C]|\mu)=0$. Then, by Theorem 4.1,

$$\lim_{n\to\infty}\frac{1}{n}I(\mu^n_{(f_n\in C)}|\mu^n)=0.$$

Henceforth it follows that

$$\lim_{n\to\infty}I(\mu_{(f_n\in C)}|\mu)=0.$$

Now we recall the inequality,

$$\|\lambda-\eta\|\leqq(2I(\lambda|\eta))^{1/2},$$

for all λ, $\eta\in\Pi$. This completes the proof.

LEMMA 4.1. *Let $\{a_n: n=1, 2, \cdots\}$ be a sequence of real numbers. If $\limsup_{n\to\infty}\frac{1}{n}a_n<0$, then $\lim_{n\to\infty}a_n=-\infty$.*

THEOREM 4.3. *Let $\mu\in\Pi$ and f be a bounded measurable function with values in R^k. Let C be a closed convex subset of R^k with nonempty interior C° such that $I(\Pi_f[C^\circ]|\mu)<+\infty$. Then*

$$\mu_{(f_n\in C)}\to\mu^* \qquad (n\to\infty)$$

in τ-topology, where $\mu^*\in\Pi_f[C]$ satisfies

$$I(\mu^*|\mu)=I(\Pi_f[C]|\mu).$$

Proof. Let g be a real-valued bounded measurable function. To prove this theorem, we observe

$$\lim_{n\to\infty}\int g\, d\mu_{(f_n\in C)}=\int g\, d\mu^*.$$

We use the following notations:

$$\bar{g}^*=\int g\, d\mu^*,\ g_n(x_1, \cdots, x_n)=\frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

For an arbitrary $\varepsilon>0$, we write

$$\Pi_{(f,g)}[C, \varepsilon]=\left\{\lambda\in\Pi : \lambda\in\Pi_f[C],\ \left|\int g\, d\lambda-\bar{g}^*\right|\geqq\varepsilon\right\}.$$

Since μ^* does not belong to $\Pi_{(f,g)}[C, \varepsilon]$, it is easy to see that

$$I(\Pi_f[C]\,|\,\mu)<I(\Pi_{(f,g)}[C, \varepsilon]\,|\,\mu).$$

By Theorem 4.1,

$$\lim_{n\to\infty}\frac{1}{n}\log\mu^n\{f_n\in C\}=-I(\Pi_f[C]\,|\,\mu).$$

From τ-closedness of $\Pi_{(f,g)}[C, \varepsilon]$, it follows by Theorem B that

$$\limsup_{n\to\infty}\ \frac{1}{n}\log\mu^n\{f_n\in C,\ |g_n-\bar{g}^*|\geqq\varepsilon\}\leqq-I(\Pi_{(f,g)}[C, \varepsilon]\,|\,\mu).$$

Therefore we obtain

$$\limsup_{n\to\infty}\ \frac{1}{n}\log\mu^n\{f_n\in C,\ |g_n-\bar{g}^*|\geqq\varepsilon\}<\lim_{n\to\infty}\frac{1}{n}\log\mu^n\{f_n\in C\}.$$

This implies

$$\limsup_{n\to\infty}\ \frac{1}{n}\log\frac{\mu^n\{f_n\in C,\ |g_n-\bar{g}^*|\geqq\varepsilon\}}{\mu^n\{f_n\in C\}}<0.$$

By Lemma 4.1, it follows that

$$\lim_{n\to\infty}\mu^n\{|g_n-\bar{g}^*|\geqq\varepsilon\,|\,f_n\in C\}=0.$$

Then it is easily verified that

$$\left|\int g_n d\mu^n_{(f_n\in C)}-\bar{g}^*\right|\leqq 2\varepsilon,$$

for all sufficiently large n. Here we note that

$$\int g_n d\mu^n_{(f_n\in C)}=\frac{1}{n}\sum_{i=1}^{n}\int g(x_i)d\mu^n_{(f_n\in C)}(x_1, \cdots, x_n)$$

$$=\int g\, d\mu_{(f_n\in C)}.$$

Since ε is arbitrary, we obtain

$$\lim_{n\to\infty}\int g\, d\mu_{(f_n\in C)}=\bar{g}^*.$$

This completes the proof.

In the previous proof, the following corollary is already shown. A similar result is proved in Vasisek [15], where he required that X was a finite set.

COROLLARY 4.1. *Let g be a real-valued bounded measurable function. Then for an arbitrary $\varepsilon>0$*

$$\mathrm{Prob}\left\{\left|\int g\, d\delta_n-\int g\, d\mu^*\right|<\varepsilon \mid \delta_n\in\Pi_f[C]\right\}\to 1$$

as $n\to\infty$.

Next we consider a condition such that $I(\Pi_f[C]\,|\,\mu)<+\infty$ for an open set C. In Bahadur and Zabell [2], they provide a condition on which

$$\lim_{n\to\infty}\frac{1}{n}\log\mu^n\{f_n\in C\}$$

exists and is finite for an open convex set C. Inspired by their result we obtain the following proposition.

PROPOSITION 4.1. *Let $\mu\in\Pi$ and f be a bounded measurable function with values in R^k. Let $S(\mu f^{-1})$ be the support of a probability measure μf^{-1} on R^k. If C is an open subset of R^k, then $C\cap\overline{\mathrm{co}}\, S(\mu f^{-1})\neq\emptyset$ if and only if $I(\Pi_f[C]\,|\,\mu)<+\infty$.*

Proof. If $C\cap\overline{\mathrm{co}}\, S(\mu f^{-1})\neq\emptyset$, we can choose an element v in $C\cap\overline{\mathrm{co}}\, S(\mu f^{-1})$. Then there exists an open neighborhood U of v such that

$$v\in U\subset\bar{U}\subset C.$$

Since v is also belonging to $\overline{\mathrm{co}}\, S(\mu f^{-1})$, there exist non-negative numbers $a_1,\cdots,a_m$ and open convex sets $U_1,\cdots,U_m$ in R^k which satisfy that

$$\sum_{i=1}^m a_i=1,\quad \sum_{i=1}^m a_iU_i\subset U$$

and

$$U_i\cap S(\mu f^{-1})\neq\emptyset\qquad (i=1,\cdots,m).$$

Since $\mu f^{-1}(U_i)\neq 0$ $(i=1,\cdots,m)$, we can define a probability measure λ as

$$\lambda(\cdot)=\sum_{i=1}^m a_i\mu(\cdot\,|\,f^{-1}(U_i)).$$

Then it follows that

$$\int f\, d\lambda=\sum_{i=1}^m a_i\int f(x)\mu(dx\,|\,f^{-1}(U_i))$$

$$=\sum_{i=1}^m a_i\int w\ \mu f^{-1}(dw\,|\,U_i)\in\sum_{i=1}^m a_i\bar{U}_i.$$

From $\Sigma_i a_i \bar{U}_i \subset \bar{U}$, it is clear that $\lambda \in \Pi_f[C]$. And it is easy to see that $I(\lambda|\mu) < +\infty$.

Conversely if $I(\Pi_f[C]|\mu) < +\infty$, then there exists a probability measure $\lambda \in \Pi_f[C]$ such that $I(\lambda|\mu) < +\infty$. Hence it follows that

$$C \cap \overline{\text{co}}\, S(\lambda f^{-1}) \neq \emptyset$$

and

$$S(\lambda f^{-1}) \subset S(\mu f^{-1}).$$

This completes the proof.

Our next object is to give additional information about the probability measure $\mu^* \in \Pi_f[C]$ assuming that f is real-valued. For convention we write $\Pi_f[a]$ instead of $\Pi_f[\{a\}]$ for a real number a.

LEMMA 4.2. *Let $\mu \in \Pi$ and f be a real-valued measurable function. Assume that $\bar{f} = \int f\, d\mu$ exists. If $\bar{f} < a < b$ or $\bar{f} > a > b$, then*

$$I(\Pi_f[a]|\mu) \geqq I(\Pi_f[b]|\mu).$$

Proof. We may assume that $I(\Pi_f[b]|\mu) < +\infty$. Then, for $\lambda \in \Pi_f[b]$, it follows that $(1-t)\mu + t\lambda \in \Pi_f[a]$, where t is the positive constant satisfying that $a = (1-t)\bar{f} + tb$ $(0 < t < 1)$. Therefore we obtain that

$$I(\Pi_f[a]|\mu) \geqq I((1-t)\mu + t\lambda|\mu) \geqq I(\lambda|\mu)$$

for all $\lambda \in \Pi_f[b]$. This completes the proof.

As an easy consequence of this lemma, we can see the following proposition.

PROPOSITION 4.2. *Let $\mu \in \Pi$ and f be a real-valued bounded measurable function. Let $\bar{f} = \int f\, d\mu$ and $C = [a, b]$. If $\bar{f} \leqq a$ (resp., $\bar{f} \geqq b$), then there exists $\mu^* \in \Pi_f[a]$ (resp., $\mu^* \in \Pi_f[b]$) which satisfies that*

$$I(\mu^*|\mu) = I(\Pi_f[C]|\mu).$$

We remark that if $\mu^* \in \Pi_f[a]$ satisfies that $I(\mu^*|\mu) = I(\Pi_f[a]|\mu) < +\infty$, then the Radon-Nikodym derivative $d\mu^*/d\mu$ is of form

$$\frac{d\mu^*}{d\mu}(x) = \alpha \exp \beta f(x) \quad \text{if} \quad x \notin N,$$
$$= 0 \quad \text{if} \quad x \in N,$$

where N has $\lambda(N) = 0$ for every $\lambda \in \Pi_f[a]$ such that $I(\lambda|\mu) < +\infty$ (c.f. Theorem 3.1 in Csiszár [5]).

Acknowledgement. The auther would like to express his hearty thanks to Professor Hisaharu UMEGAKI for his useful comments and encouragement in the course of preparing this paper.

REFERENCES

[1] BAHADUR, R. R., 'Some Limit Theorems in Statistics', SIAM, Philadelphia, 1971.
[2] BAHADUR, R. R. AND S. L. ZABELL, Large deviations of sample mean in general vector spaces, Ann. Probability, 7 (1979), 587-621.
[3] BÁRTFAI, P., On a conditional limit theorem, Colloquia Math. Soc. J. Bolyai, No. 9, European Meeting of Statisticians, Budapest, 1972, 81-91.
[4] CHERNOFF, H., A measure of asymptotic efficiency for tests of a hypothesis based on sums of observations, Ann. Math. Statist., 23 (1952), 493-507.
[5] CSISZÁR, I., I-divergence geometry of probability distributions and minimization problems, Ann. Probability, 3 (1975), 146-158.
[6] GROENEBOOM, R., OOSTERHOFF, J. AND F. H. RUYMGAART, Large deviation theorems for empirical probability measures, Ann. Probability, 7 (1979), 553-586.
[7] KAWAMURA, K., The asymptotic distribution of information per unit cost concerning a linear hypothesis for means of given two normal populations, Kodai Math. Sem. Rep., 22 (1970), 251-271.
[8] KULLBACK, S., 'Information and Statistics', John Wiley and Sons, New York, 1959.
[9] KULLBACK, S. AND R. A. LEIBLER, On information and sufficiency, Ann. Math. Statist. 22 (1951), 79-86.
[10] SANOV, I. N., On the probability of large deviations of random variables, Sel. Transl. Math. Statist. Prob., 1 (1967), 213-244.
[11] STEINEBACH, J., 'Large Deviation Probabilities and Some Related Topics', Carleton Math. Lecture Notes, No. 28, 1980.
[12] UMEGAKI, H., Conditional expectation in an operator algebra III, Kodai Math. Sem. Rep., 11 (1959), 51-64.
[13] UMEGAKI, H., Conditional expectation in an operator algebra IV, Kodai Math. Sem. Rep., 14 (1962), 59-85.
[14] UMEGAKI, H. AND M. OHYA, 'Probability Theoretic Entropy' (in Japanese), Kyoritsu, Tokyo, 1983.
[15] VASICEK, O. A., A conditional law of large numbers, Ann. Probability, 8 (1980), 142-147.
[16] VINCZE, I., On the maximum probability principle in statistical physics, Colloquia Math. Soc. J. Bolyai, No. 9, European Meeting of Statisticians, Budapest, 1972, 869-893.

DEPARTMENT OF INFORMATION SCIENCES
TOKYO INSTITUTE OF TECHNOLOGY
OH-OKAYAMA, MEGURO-KU, TOKYO 152