6. On the Écart between Two "Amounts of Information"

By Kōmei SUZUKI

(Comm. by K. KUNUGI, M.J.A., Jan. 12, 1957)

§ 1. $d(\lambda_1, \lambda_2; \Lambda)=\sum_{i=0}^{\infty} \Delta P_i \log\left(1+\frac{\Delta P_i}{P_i}\right)$

As was shown in the preceding paper the "amount of information"[2)-4)] has been defined by a specified probability space (or distribution), $(R, \mathfrak{X}, \lambda)$, and the partition,[1)] Λ, imposed on the space R. And we have conventionally denoted it by $H(\lambda; \Lambda)$. As usual

$$\Lambda : R=\bigcup_{i=0}^{\infty} A_i,\ A_i \in \mathfrak{X},\ A_i \cap A_j=0 \quad (i \neq j).$$

For any two distributions $(R, \mathfrak{X}, \lambda_1)$ and $(R, \mathfrak{X}, \lambda_2)$, providing

(a) $\lambda_1(A_i)=P_i \geq 0,\ \lambda_2(A_i)=P_i+\Delta P_i \geq 0,\ \sum_i P_i=\sum_i (P_i+\Delta P_i)=1$

(b) the series $H(\lambda_1; \Lambda)=\sum_i P_i \log 1/P_i$ and $H(\lambda_2; \Lambda)=\sum_i (P_i+\Delta P_i) \log 1/(P_i+\Delta P_i)$ to converge

(c) $-1+\alpha \leq \Delta P_i/P_i \leq k;\ k>0,\ 1>\alpha>0$ for all i,

we have directly from the result obtained in the preceding paper

$$0 \leq \Delta H-\sum_{i=0}^{\infty} \Delta P_i \log \frac{1}{P_i+\Delta P_i} \leq \sum_{i=0}^{\infty} \Delta P_i \log\left(1+\frac{\Delta P_i}{P_i}\right)$$

where $\Delta H=H(\lambda_2; \Lambda)-H(\lambda_1; \Lambda)$.

Denoting $\sum_{i=0}^{\infty} \Delta P_i \log\left(1+\frac{\Delta P_i}{P_i}\right)$ by $d(\lambda_1, \lambda_2; \Lambda)$, we have easily

$$(1.1)\quad \begin{array}{ll} \text{(a)} & d(\lambda, \lambda; \Lambda)=0 \\ \text{(b)} & d(\lambda_1, \lambda_2; \Lambda) \geq 0 \quad \text{for } \lambda_1 \neq \lambda_2 \\ \text{(c)} & d(\lambda_1, \lambda_2; \Lambda)=d(\lambda_2, \lambda_1; \Lambda). \end{array}$$

It must be noted that we could not avoid the sign of equality in (b) of (1.1); because even though $\lambda_1 \neq \lambda_2$, we would often have that $\lambda_1(A_i)=\lambda_2(A_i)$, $i=0, 1, 2, \cdots$, for some partitions imposed on R.

To appreciate more fully we consider a distribution $(R, \mathfrak{X}, \lambda_3)$ together with the above $(R, \mathfrak{X}, \lambda_1)$ and $(R, \mathfrak{X}, \lambda_2)$.

Providing again the following

(d) $\lambda_1(A_i)=P_i^{(1)},\ \lambda_2(A_i)=P_i^{(2)}=P_i^{(1)}+\Delta P_i^{(1)},\ \lambda_3(A_i)=P_i^{(3)}=P_i^{(2)}+\Delta P_i^{(2)}$

(e) $-1+\alpha \leq \Delta P_i^{(\nu)}/P_i^{(\nu)} \leq k;\ 1>\alpha>0,\ k>0,\ i=0, 1, 2, \cdots,\ \nu=1, 2$

we have

$$d(\lambda_3, \lambda_1; \Lambda)-\{d(\lambda_1, \lambda_2; \Lambda)+d(\lambda_2, \lambda_3; \Lambda)\}$$
$$=\sum_{i=0}^{\infty}(\Delta P_i^{(1)} \log P_i^{(3)}/P_i^{(2)}+\Delta P_i^{(2)} \log P_i^{(2)}/P_i^{(1)})$$

(1.2) and

$$(\Delta P_i^{(1)} \log P_i^{(3)}/P_i^{(2)}+\Delta P_i^{(2)} \log P_i^{(2)}/P_i^{(1)}) \begin{cases} >0 \leftrightarrow \Delta P_i^{(1)} \cdot \Delta P_i^{(2)}>0 \\ =0 \leftrightarrow \Delta P_i^{(1)} \cdot \Delta P_i^{(2)}=0 \\ <0 \leftrightarrow \Delta P_i^{(1)} \cdot \Delta P_i^{(2)}<0. \end{cases}$$

These relations show that the quantity $d(\lambda_1, \lambda_2; \Lambda)$ does not satisfy the triangle law of distance; it could however well describe the degree of the discrepancy two distributions $(R, \mathfrak{X}, \lambda_1)$ and $(R, \mathfrak{X}, \lambda_2)$ under the partition imposed, Λ.

Thus, remembering its origin, we take $d(\lambda_1, \lambda_2; \Lambda)$ into consideration as the "écart" between two "amounts of information" about the capability of the source due to the distributions $(R, \mathfrak{X}, \lambda_1)$ and $(R, \mathfrak{X}, \lambda_2)$ with the partition, Λ, which is imposed on R;
(Cf. § 2 in the preceding paper.)

§ 2. $\int (f_2(x)-f_1(x)) \log \frac{f_2(x)}{f_1(x)} dx$

We consider, henceforth, the random variable X, with the probability density $f(x)$, attached to the probability space $(R, \mathfrak{X}, \lambda)$ while taking up conveniently the set of whole real numbers $\{x\}$ as the space R; and the components (A_i) of the partition (Λ) are considered to be reduced to the half open intervals $I_i = a_i < x \leq b_i$, $i=0, 1, 2, \cdots$, hence we can put

$$P_i = \lambda(I_i) = \int_{I_i} f(x)dx.$$

Then the following may be easily extended to the discussion in an n-dimensions Euclidean space $R=R_n$.

Let us provisionally attach the probability densities $f_1(x)$ and $f_2(x)$ to the measure λ_1, λ_2 respectively.

For any number ε, we may have an integer N such as

$$0 \leq \sum_{i=N}^{\infty} P_i \log \frac{1}{P_i}, \quad \sum_{i=N}^{\infty} (P_i + \Delta P_i) \log \frac{1}{P_i + \Delta P_i} \leq \varepsilon.$$

Then if we take a domain A such as $A = \bigcup_{\nu=0}^{n} I_{i_\nu} \supseteq \bigcup_{i=0}^{N} I_i$ we get

$$0 \leq H(\lambda_1; \Lambda) - \sum_{\nu=0}^{n} P_{i_\nu} \log \frac{1}{P_{i_\nu}} \leq \varepsilon$$

$$0 \leq H(\lambda_2; \Lambda) - \sum_{\nu=0}^{n} (P_{i_\nu} + \Delta P_{i_\nu}) \log \frac{1}{P_{i_\nu} + \Delta P_{i_\nu}} \leq \varepsilon.$$

Thus, referring to (e) of § 1, we can define a positive number g such as $0 \leq d(\lambda_1, \lambda_2; \Lambda) - \sum_{\nu=0}^{\infty} \Delta P_{i_\nu} \log\left(1 + \frac{\Delta P_{i_\nu}}{P_{i_\nu}}\right) \leq \varepsilon/g.$

If the integral $\int (f_2(x)-f_1(x)) \log \frac{f_2(x)}{f_1(x)} dx$ might be obtained over A, we could set

$$\sum_{\nu=0}^{n} \Delta P_{i_\nu} \log\left(1 + \frac{\Delta P_{i_\nu}}{P_{i_\nu}}\right) = \int_A (f_2(x)-f_1(x)) \log \frac{f_2(x)}{f_1(x)} dx + \epsilon(A, \Lambda)$$

and when $\max_\nu (b_{i_\nu} - a_{i_\nu})$ tends to zero, $\epsilon(A, \Lambda)$ also tends to zero.

Then

$$\epsilon(A, \Lambda) \leq d(\lambda_1, \lambda_2; \Lambda) - \int_\Lambda (f_2(x) - f_1(x)) \log \frac{f_2(x)}{f_1(x)} dx \leq \varepsilon/g + \epsilon(A, \Lambda).$$

Thus we can reach the formula

$$d(\lambda_1, \lambda_2) = \int_R (f_2(x) - f_1(x)) \log \frac{f_2(x)}{f_1(x)} dx.$$

And we have also

$$\text{(2.1)} \quad \begin{array}{ll} \text{(a)} & d(\lambda, \lambda) = 0 \\ \text{(b)} & d(\lambda_1, \lambda_2) > 0 \quad \text{for } \lambda_1 \neq \lambda_2 \\ \text{(c)} & d(\lambda_1, \lambda_2) = d(\lambda_2, \lambda_1) \end{array}$$

and corresponding to the relations (1.2), we have

$$d(\lambda_1, \lambda_3) - \{d(\lambda_3, \lambda_2) + d(\lambda_2, \lambda_1)\}$$
$$= \int_R \left\{ (f_2(x) - f_1(x)) \log \frac{f_3(x)}{f_2(x)} + (f_3(x) - f_2(x)) \log \frac{f_2(x)}{f_1(x)} \right\} dx$$

(2.2) and

$$\left\{ (f_2(x) - f_1(x)) \log \frac{f_3(x)}{f_2(x)} + (f_3(x) - f_2(x)) \log \frac{f_2(x)}{f_1(x)} \right\}$$
$$> 0 \leftrightarrow (f_2(x) - f_1(x))(f_3(x) - f_2(x)) > 0$$
$$= 0 \leftrightarrow (f_2(x) - f_1(x))(f_3(x) - f_2(x)) = 0$$
$$< 0 \leftrightarrow (f_2(x) - f_1(x))(f_3(x) - f_2(x)) < 0.$$

Thus, the proposition described in the probability mass (P_i) has been rewritten in the corresponding probability density $(f(x))$.
And we may call $d(\lambda_1, \lambda_2)$, the écart[5] between two amounts of information due to the probability distributions $(R, \mathfrak{X}, \lambda_1)$ and $(R, \mathfrak{X}, \lambda_2)$.

In the preceding paper we have had

$$\Delta H = \sum_i \Delta P_i \log \frac{1}{P_i} - \sum_i (P_i + \Delta P_i) \log \frac{P_i + \Delta P_i}{P_i}$$
$$= \sum_i \Delta P_i \log \frac{1}{P_i + \Delta P_i} + \sum_i P_i \log \frac{P_i}{P_i + \Delta P_i}$$

then

$$\sum_i \Delta P_i \log \frac{1}{P_i} - \sum_i \Delta P_i \log \frac{1}{P_i + \Delta P_i}$$
$$= \sum_i (P_i + \Delta P_i) \log \frac{P_i + \Delta P_i}{P_i} - \sum_i P_i \log \frac{P_i + \Delta P_i}{P_i}$$

S. Kullback and A. Leibler[6] were regardless about this, though it may be said that they have derived the formula $\int (f_2(x) - f_1(x)) \log \frac{f_2(x)}{f_1(x)} dx$ from the latter half of the above relation.

References

1) Darrow, C. K.,: Statistical theories of matter radiation and electricity, Phys. Rev. (1929).
2) Wiener, N.,: Cybernetics (1948).
3) Shannon, C. E.,: The Mathematical Theory of Communication (1949).
4) Weaver, W.,: Recent Contribution to the Mathematical Theory of Communication (1949).
5) Kunugi, K.,: Kaisekigaku Yōron (Essentials of Analysis) (1951).
6) Kullback, S., and Leibler, A.,: On information and sufficiency, Ann. Math. Stat., **22** (1951).