The Basis of Sequential Estimation Process from the Viewpoint of Information Theory

SUGURU ARIMOTO*

This paper attempts to establish a theoretic basis for sequential estimation problems from the viewpoint of information theory. The problem which is dealt with consists in that one has to make a decision which point of a set of categories an observed sequence of random variables belongs to. It is supposed that the distribution of the random variables depends on a point of the finite set of categories. Much attentions are paid to constructing equivocation quantities with respect to decision rules, which satisfy some axiomatically required conditions of goodness measure. A necessary and sufficient condition for an equivocation to be minimized by the Bayes decision rule is obtained when the number of categories is larger than two. In the simplest case where only two categories are possible the Bayes decision rule minimizes an infinite class of equivocation quantities.

1. *Definition of Decision Rules*

Let $X=\{x_1, \cdots, x_n\}$ be a set of categories and $\{Y_t\}$, $(t=1, 2, \cdots)$, be a sequense of random variables whose distribution depends on the point of X. It is assumed that the sample space of the random variable Y_t, denoted by $\mathcal{Y}_t$, consists of a fixed set of letters such that $\mathcal{Y}_t=\{a_1, \cdots, a_m\}$. It is further assumed that for every integer t the distributions of the random variable $Y^t=(Y_1, \cdots, Y_t)$ are known and given by $p(y^t/x_k)$, $(k=1, 2, \cdots, n)$, where y^t is a realization of Y^t and denotes a point of $\mathcal{Y}^t=\mathcal{Y}_1\times\cdots\times\mathcal{Y}_t$.

Definition 1. Let Θ be the set of all vectors θ such that $\theta=(\theta_1, \cdots, \theta_n)$, $1>\theta_k>0$, $\sum\theta_k=1$, and $\overline{\Theta}$ be the closure of Θ. A mapping that maps the sample space $\mathcal{Y}^t$ into $\overline{\Theta}$ is called a decision rule at time t and is denoted by $\varphi^t: \mathcal{Y}^t\to\overline{\Theta}$, or in a concrete form $\theta=\varphi^t(y^t)$. The set of all decision rules at time t is denoted by $\Phi^t=\{\varphi^t\}$.

Example. The Bayes decision rule is defined as follows:

$$\varphi_k{}^t(y^t)=p(y^t/x_k)\theta_k/\sum_j p(y^t/x_j)\theta_j. \tag{1}$$

Here θ is called a prior distribution.

This paper first appeared in Japanese in Joho-Shori (the Journal of the Information Processing Society of Japan), Vol. 10, No. 2 (1969), pp. 61–67.

* Faculty of Engineering Science, Osaka University, Toyonaka, Osaka.

2. *Generalized Equivocation*

The problem which will be discussed in this section is how one can construct a goodness or equivocation measure concerning such decision rules. Reasonably a goodness measure of a given decision rule φ^t, which is considered to be a random variable by definition, should be a certain expected value of φ^t. However, since the true value of categories is not known, one can not evaluate the true expected value of the decision rule φ^t. Therefore, as far as one takes the Bayes viewpoint and attributes a prior probability distribution $\theta\in\Theta$ to the set of categories X, the only meaningful expectation of a random variable in question should be taken with the joint distribution over $X\times{}^{Q}\mathcal{Y}^t$. Thus, according to this Bayes viewpoint, we shall henceforth confine our considerations to the following type of goodness measure.

$$F(\varphi^t:\theta)=\sum_k\sum_{y^t} p(y^t/x_k)\theta_k f(\varphi_k{}^t(y^t)). \tag{2}$$

Here the scalar function $f(\alpha)$ is assumed to be continuous and have a continuous derivative on $0<\alpha\leq 1$.

We can now describe some axiomatically required conditions that the quantity (2) has to satisfy.

Definition 2. When the quantity (2) satisfies the following four conditions, it is called a generalized equivocation with respect to the decision rule φ^t.

(i) For any $\varphi^t\in\Phi^t$ and an arbitrarily fixed $\theta\in\Theta$ $F(\varphi^t:\theta)$ is non-negative and there exists at least one decision rule $\varphi^t\in\Phi^t$ such that $F(\varphi^t:\theta)>0$.

(ii) If for any pair (j,k) such that $j\neq k$,

$$\sum_{y^t}\sqrt{p(y^t/x_j)p(y^t/x_k)}=0$$

then it follows that

$$\inf_{\varphi^t} F(\varphi^t:\theta)=0. \tag{3}$$

(iii) If for all $y^t\in{}^{Q}\mathcal{Y}^t$, $p(y^t/x_1)=\cdots=p(y^t/x_n)$, then it follows that

$$\inf_{\varphi^t} F(\varphi^t:\theta)=F(\theta) \tag{4}$$

where $F(\theta)$ is independent of t and continuous on $\bar{\Theta}$.

(iv) In general it holds that

$$\inf_{\varphi^{t+1}} F(\varphi^{t+1}:\theta)\leq\inf_{\varphi^t} F(\varphi^t:\theta)\leq F(\theta). \tag{5}$$

The requirement of (i) is clear. The "if" part of (ii) implies that the intersection of the sets of y^t such that $p(y^t/x_j)>0$ and $p(y^t/x_k)>0$ is void for any pair (j,k) such that $j\neq k$. This means that the observation $Y^t=y^t$ is sufficient with probability one to decide which category the observed data belong to. The form (3) is a version of this statement. On the other hand, when the assumption of (iii) is satisfied there occurs no difference between categories, that is, any observation of the random variable Y^t gives us no information to make a decision. This circumstance may be reduced to that the infimum of

equivocations remains equal to $F(\theta)$ which is considered to be an equivocation quantity with respect to the prior distribution θ. The inequality (5) is reasonably required since at least the infimum of equivocations should decrease with increasing observed data.

3. *Results*

In this section a few reults concerning the generalized equivocation are described.

Theorem 1. If a function $f(\alpha)$ in (2) satisfies the following two conditions, then the quantity $F(\varphi^t : \theta)$ has the properties (i)~(iv) of Definition 2.

(C1) $\inf_{0<\alpha\leq 1} f(\alpha)=0$.

(C2) There exists at least a number such that
$f(\alpha)>0$ and $0<\alpha\leq 1$.

Proof. The properties (i) and (ii) are almost clear. When the assumption of (iii) is satisfied,

$$\inf_{\varphi^t} F(\varphi^t : \theta)=\sum_{y^t} p(y^t/x_1)\inf_{\lambda}\left[\sum_k \theta_k f(\lambda_k)\right]=\inf_{\lambda}\left[\sum_k \theta_k f(\lambda_k)\right].$$

Thus we put

$$F(\theta)=\inf_{\lambda}\sum_k \theta_k f(\lambda_k). \tag{6}$$

Clearly from this, $F(\theta)$ is independent of t and continuous on $\bar{\Theta}$ because of the continuity of $f(\alpha)$. In order to prove the inequality (5), we first show the concavity of $F(\theta)$. Let θ^1, $\theta^2\in\bar{\Theta}$ and α be an arbitrary constant such that $0\leq\alpha\leq 1$. Then it follows immediatcly from (6) that

$$\alpha F(\theta^1)+(1-\alpha)F(\theta^2)\leq\inf_{\lambda}\sum_k(\alpha\theta_k{}^1+(1-\alpha)\theta_k{}^2)f(\lambda_k)=F(\alpha\theta^1+(1-\alpha)\theta^2). \tag{7}$$

We shall now show only the left half part of the inequality (5) since the remaining part is proven analogously. Let $y^{t+1}=y^t\times y^1$ and put

$$p(y^1/y^t, x_k)=p(y^t\times y^1/x_k)/(p^t/x_k).$$

Then, using the notations

$$p_k=p(y^1/y^t, x_k),\quad r=\sum_j p(y^t/x_j)\theta_j,\quad \theta_k{}^*=p(y^t/x_k)\theta_k/\sum_j p(y^t/x_j)\theta_j \tag{8}$$

we obtain

$$\inf_{\varphi^{t+1}} F(\varphi^{t+1} : \theta)=\sum_{y^{t+1}} r\cdot\inf_{\lambda}\left[\sum_j p_j\overset{*}{\theta}_j f(\lambda_j)\right]=r\cdot q\cdot F(q_1/q, \cdots, q_n/q). \tag{9}$$

Here we use

$$q_j=p_j\theta_j{}^*,\quad q=\sum_j q_j.$$

On the other hand, it follows from the concavity of $F(\theta)$ that

$$\sum_{y^1} q\cdot F(q_1/q, \cdots, q_n/q)\leq F\left(\sum_{y^1} q_1, \cdots, \sum_{y^1} q_n\right)=F(\theta_1{}^*, \cdots, \theta_n{}^*).$$

Substituting this inequality into (9), we obtain

$$\inf_{\varphi^{t+1}} F(\varphi^{t+1} : \theta)\leq\sum_{y^t} r\cdot F(\theta^*)=\sum_{y^t}\inf_{\lambda}\left[\sum_k p(y^t/x_k)\theta_k f(\lambda_k)\right]=\inf_{\varphi^t} F(\varphi^t : \theta).$$

Theorem 2. Suppose that $f(\alpha)$ satisfies (C1) and (C2) of Definition 2 and

let $n\geq 3$. Then, a necessary and sufficient condition for a generalized equivocation $F(\varphi^t:\theta)$ to be minimized by the Bayes decision rule (1) is that $f(\alpha)$ should have the form $f(\alpha)=-c\log\alpha$ where c is a positive constant.

Proof. The proof of the sufficiency follows immediately from noting the inequality

$$-\sum_k \theta_k \log\theta_k \leq -\sum_k \theta_k \log\lambda_k \tag{10}$$

where θ and λ are arbitrary vectors of Θ. Hence we shall prove the necessity. At first, note that

$$\inf_{\varphi^t} F(\varphi^t:\theta)=\sum_{y^t}\inf_{\lambda}\left[\sum_k p(y^t/x_k)\theta_k f(\lambda_k)\right]=\sum_{y^t} r\cdot\inf_{\lambda}\left[\sum_k \theta_k^* f(\lambda_k)\right] \tag{11}$$

where the notations r and θ^* are the same as (8). Since θ^* is equivalent to the Bayes decision rule defined by (1), the expression (11) implies that in general $f(\alpha)$ should satisfy

$$\inf_{\lambda}\left[\sum_k \theta_k f(\lambda_k)\right]=\sum_k \theta_k f(\theta_k) \tag{12}$$

in order that the equivocation $F(\varphi^t:\theta)$ is minimized by the Bayes decision rule. Therefore, from the theorem of Lagrange's multiplier rule for a nonlinear programming problem, (12) yields

$$\theta_k f'(\theta_k)-\xi=0 \quad \text{for} \quad k=1,\ 2,\ \cdots,\ n \tag{13}$$

where ξ is a constant multiplier and f' is the derivative of f. Taking into account the assumption that $n\geq 3$, we have from (13)

$$\alpha f'(\alpha)=\text{const.} \quad \text{for} \quad 0<\alpha<1.$$

This yields

$$f(\alpha)=-c\log\alpha+c_1.$$

Due to the conditions (C1) and (C2) of Definition 2, c_1 becomes zero and c should be positive. Thus the theorem has been proven.

On account of Theorem 2 it is very reasonable in case of $n\geq 3$ to consider the quantity

$$J(\varphi^t:\theta)=\sum_{y^t}\sum_k -p(y^t/x_k)\theta_k\log\varphi_k{}^t(y^t). \tag{14}$$

It may be worth noting that the conventional equivocation quantity denoted by $H(X/Q_Y{}^t)$ in information theory is obtained by minimizing $J(\varphi^t:\theta)$ with respect to φ^t, that is,

$$H(X/Q_Y{}^t)=\inf_{\varphi^t} J(\varphi^t:\theta)=\sum_{y^t}\sum_k -p(y^t/x_k)\theta_k\log\left\{\frac{p(y^t/x_k)\theta_k}{\sum_j p(y^t/x_j)\theta_j}\right\}.$$

It is also very interesting to note that in the simplest case of $n=2$ the Bayes decision rule minimizes an infinite class of equivocations. If the derivative of the function f has the form

$$f'(\alpha)=g(\alpha(1-\alpha))/\alpha \tag{15}$$

where $g(\beta)$ is an arbitrary function, then $f(\alpha)$ satisfies (13) if $n=2$. Namely, as far as considerations are confined only to the case where $n=2$ there is no positive reason to adopt a quantity with the form (14) as a measure of goodness of

decision rules. The case where $f(\alpha)=\sqrt{(1-\alpha)/\alpha}$, which clearly satisfies (14), was first examined by Rényi [1] without giving any implication to derive an upper bound for his "missing information". Here we remark that Rényi's result relies heavily on the fact that $\sqrt{(1-\alpha)/\alpha}\geq-\log\alpha$ for $0<\alpha\leq 1$, where the logarithm to base e is employed. In particular, as an illustration of applications of the above inequality and the idea of generalized equivocations, the following inequality is obtained.

$$H(X/\mathcal{Y}^t)=\inf_{\varphi^t}J(\varphi^t:\theta)\leq\inf_{\varphi^t}\sum_{y^t}\sum_{k=1,2}p(y^t/x_k)\theta_k\sqrt{(1-\varphi_k{}^t(y^t))/\varphi_k{}^t(y^t)}$$
$$=2\sqrt{\theta_1(1-\theta_1)}\sum_{y^t}\sqrt{p(y^t/x_1)p(y^t/x_2)}$$

where the last equality is derived from substituting the Bayes decision rule (1) into φ^t.

Finally, we comment that a similar discussion is carried out for the case where the sample space $\mathcal{Y}^t$ is infinite and continuous, and useful inequalities concerning the generalized equivocation and error probability of practical decision making are obtained [2].

References

[1] Rényi, A., On Some Basic Problems of Statistics from the Point of View of Information Theory, *Proceedings of the Fifth Berkeley Symposium, 1*, University of California Press(1967), 531-543.

[2] Arimoto, S., Bayesian Decision Rule and Quantity of Equivocation (in Japanese), *the Transactions of the Institute of Electronics and Communication Engineers of Japan*, 53-*C*, 1(1970), 16-22.