Generalized Information Measure and Finite-Parameter Estimation Problems

SUGURU ARIMOTO*

1. *Introduction*

Recently, Rényi and his collaborators [1][2] have proposed an information-theoretic treatment of the parameter estimation problem and pointed out that an important role is played by the amount of information concerning an unknown parameter contained in a sequence of observations. This type of information measure is the same as the one previously proposed by Lindley [3] and has the same form as the so-called transinformation or mutual information in information theory.

This paper deals with the same problem as Rényi did but gives more basic considerations from the viewpoint of information theory. After a brief exposition of generalized entropy and equivocation measures previously proposed by the author [4]~[6], some basic relations between the generalized equivocation and the decision rule are investigated. It is shown that a few of basic inequalities, which are useful for evaluating the error-probability of decoding in information theory [7], can be derived directly from the definition of generalized equivocation measure.

2. *Definition of Generalized Entropy*

Let $f(t)$ be a real-valued function defined and nonnegative on $(0, 1)$ with a continuous derivative and $f(1)=0$. Let $p=(p_1, \cdots, p_n)$ be a probability vector such that $\sum_k p_k=1$ and $p_k \geq 0$. Then, in accordance with Arimoto[6], the following theorem holds:

Theorem 1. Let

$$F_f(p_1, \cdots, p_n)=\inf_q \sum_{k=1}^{n} p_k f(q_k), \tag{1}$$

where the operation inf is taken over all probability vectors such as $q=(q_1, \cdots, q_n)$, $\sum_k q_k=1$, and $q_k>0$. Then,

(i) $F_f(p_1, \cdots, p_n)$ is continuous and symmetric with respect to its arguments $p_1, \cdots, p_n$.

(ii) $F_f(p_1, \cdots, p_n)=F_f(p_1, \cdots, p_n, 0)$.

(iii) $F_f(p)\equiv F_f(p_1, \cdots, p_n)$ is a concave function with respect to the vector p.

(iv) $0 \leq F_f(p_1, \cdots, p_n) \leq f(1/n)$.

(v) If $f(t)$ is convex, then $F_f(p_1, \cdots, p_n) \leq F_f(1/n, \cdots, 1/n)=f(1/n)$.

This paper first appeared in Japanese in Joho-Shorl (Journal of the Information Processing Society of Japan), Vol. 12, No. 9 (1971), pp. 549~555.

* Faculty of Engineering Science, Osaka University

(vi) In general,

$$F_f(p_1, \cdots, p_n) \leq \sum_{k=1}^{n} p_k f(p_k). \tag{2}$$

When $n \geq 3$, the equality sign holds for any probability vector p if and only if the scalar function f has the form

$$f(t) = c \log t, \tag{3}$$

where c is a nonpositive constant. When $n=2$, the equality holds for an infinite family of f including the form (3).

Thus, the quantity (1) is called, in this paper, a generalized entropy.

3. *Generalized Equivocation and Parameter Estimation Problems*

Let $X = \{x_1, \cdots, x_n\}$ be a finite set of parameters and $(Y, \mathcal{Y})$ be a measurable space. Let (ξ, η) be a random variable defined by

$$P[\xi = x_k, \ \eta \in E] = p_k P_k(E), \tag{4}$$

where $E \in \mathcal{Y}$, $p = (p_1, \cdots, p_n)$ is a probability distribution on X, and P_k are probability distributions on $\mathcal{Y}$.

The finite-parameter estimation problem consists in that a decision maker has to select only one true value of the parameter X after observing a sample $\eta = y \in Y$. Hence, p may be interpreted as an a priori distribution of ξ. Given a sample $\eta = y$, the a posteriori distribution $p^* = (p_1^*(y), \cdots, p_n^*(y))$ is calculated by using Bayes' theorem in the following manner:

$$p_k^*(y) = p_k dP_k(y) / d\bar{P}(y) \tag{5}$$

where the derivative dP_k denotes the Radon-Nikodym density of the probability measure P_k with respect to the probability measure

$$\bar{P}(E) = \sum_{i=1}^{n} p_i P_i(E). \tag{6}$$

Then, the quantity

$$H_f(\xi) = F_f(p_1, \cdots, p_n) = \inf_{q} \sum_{k=1}^{n} p_k f(q_k) \tag{7}$$

is called a generalized entropy of the random variable ξ. Similarly, the quantity

$$H_f(\xi \mid \eta) = \varepsilon F_f(p^*) = \int_Y F_f(p^*) d\bar{P}(y) \tag{8}$$

is called a generalized equivocation.

In case of $f(t) = -\log t$, it is well known that the quantity

$$I = H(\xi) - H(\xi \mid \eta), \tag{9}$$

where

$$H(\xi) = H(p_1, \cdots, p_n) = -\sum_{k=1}^{n} p_k \log p_k$$

$$H(\xi \mid \eta) = \varepsilon H(p_1^*, \cdots, p_n^*),$$

is always nonnegative and is therefore called the amount of information contained in the observation η concerning the parameter ξ. In other words, the

amount of equivocation does not increase through observations. More generally, it holds

Theorem 2.

$$H_f(\xi) \geq H_f(\xi|\eta). \tag{10}$$

The proof of this theorem follows immediately from the property (iii) in Theorem 1.

It is now necessary to introduce the definitions:

Definition 1. A vector-valued measurable function $\varphi(y)=(\varphi_1(y), \cdots, \varphi_n(y))$ such that $\varphi_k(y) \geq 0$ and $\sum_k \varphi_k(y)=1$ for almost everywhere is called a decision rule.

Definition 2. A vector-valued measurable function $\delta(y)=(\delta_1(y), \cdots, \delta_n(y))$ such that $\delta_k(y)=0$ or 1 and $\sum_k \delta_k(y)=1$ for almost everywhere is called a decision.

Definition 3. A sequence $A=\{A_1, \cdots, A_m\}$ of disjoint subsets of Y such that $A_j \in \mathcal{Y}$ and $\sum_j A_j = Y$ is called a partition of Y. A partition $A=\{A_1, \cdots, A_m\}$ produces a set of conditional probabilities $P_k(A_j)$. Hence, such a partitioning with probability measures $P[A_j|\xi=x_k]=P_k(A_j)$ is called a data processing.

Definition 4. Let $\varphi(y)$ be a decision rule. The quantity

$$H_f(\xi|\eta;\ \varphi)=\int_Y \sum_{k=1}^{n} f(\varphi_k(y))p_k dP_k(y) \tag{11}$$

is called a generalized equivocation for the decision rule φ. Let $\delta(y)$ be a decision. Then, the quantity

$$P_e(\delta)=\int_Y \sum_{k=1}^{n} (1-\delta_k(y))p_k dP_k(y) \tag{12}$$

is called the error-probability of the decision $\delta(y)$. It should be noted that the error-probability is a special case of (11) in which $f(t)=1-t$ is used.

It is now possible to present the following theorems:

Theorem 3.

$$H_f(\xi|\eta)=\inf_{\varphi} H_f(\xi|\eta;\ \varphi) \leq H_f(\xi|\eta;\ \varphi). \tag{13}$$

Theorem 4. Any data processing does not reduce the quantity of equivocation in the sense that

$$H_f(\xi|\eta) \leq H_f(\xi|A;\ \psi) \equiv \sum_{j=1}^{m} \sum_{k=1}^{n} f(\psi_k(A_j))p_k P_k(A_j) \tag{14}$$

for any decision rule $\psi(A_j)=(\psi_1(A_j), \cdots, \psi_n(A_j))$ such that $\psi_k(A_j) \geq 0$ and $\sum_k \psi_k(A_j) = 1$. In particular,

$$H_f(\xi|\eta) \leq H_f(\xi|A) \equiv \sum_{j=1}^{m} P(A_j) F_f(p_1^*(A_j), \cdots, p_n^*(A_j)) \tag{15}$$

where

$$p_k^*(A_j)=p_k P_k(A_j)/P(A_j), \qquad P(A_j)=\sum_{i=1}^{n} p_i P_i(A_j). \tag{16}$$

Theorem 5. When f has the form $f(t)=1-t$, then

$$H_f(\xi|\eta)=\inf_{\delta} P_e(\delta) \leq P_e(\delta) \tag{17}$$

for any decision δ.

Finally, it is necessary to introduce the definition:

Definition 5. Let $\varphi(y)$ be a decision rule. Then, the decision defined by the form $\delta^{\varphi}(y)=(\delta_1{}^{\varphi}(y), \cdots, \delta_n{}^{\varphi}(y))$, where

$$\delta_k{}^{\varphi}(y)=\begin{cases}0 & \text{if there exists an integer } j \text{ such that } \delta_k(y)<\delta_j(y) \\ & \text{or } \delta_k(y)=\delta_j(y) \text{ for } k>j. \\ 1 & \text{otherwise,}\end{cases}$$

is called a maximum likelihood decision induced by the decision rule φ.

Theorem 6. Let f be a scalar function with a continuous and negative derivative on $(0, 1]$ with $f(1)=0$. If the decision rule $\varphi^*(y)$ is optimal in the sense that it minimizes the equivocation $H_f(\xi|\eta; \varphi)$, that is, $H_f(\xi|\eta)=H_f(\xi|\eta; \varphi^*)$, then the maximum likelihood decision δ^{φ^*} induced by φ^* minimizes the error-probability of decision.

Proof. To prove this, it is necessary to note that an arbitrary probability vector $p=(p_1, \cdots, p_n)$ with the property

$$p_{i_1}>p_{i_2}>\cdots>p_{i_n} \tag{18}$$

implies that the corresponding vector $q=(q_1, \cdots, q_n)$ such that

$$\inf_{\tilde{p}} \sum_{k=1}^{n} p_k f(\tilde{P}_k)=\sum_{k=1}^{n} p_k f(q_k) \tag{19}$$

has the property

$$q_{i_1}>q_{i_2}\geq\cdots\geq q_{i_n} \tag{20}$$

(see Lemma 1 [6]). In view of this, the optimality of the decision rule φ^* implies that if Bayes' form $p^*(y)$ has the property

$$p^*{}_{i_1}(y)>p^*{}_{i_2}(y)>\cdots>p^*{}_{i_n}(y) \tag{21}$$

then φ^* also has the following descending order:

$$\varphi^*{}_{i_1}(y)>\varphi^*{}_{i_2}(y)\geq\cdots\geq\varphi^*{}_{i_n}(y). \tag{22}$$

Since the function $f(t)=1-t$ itself satisfies the assumption of the theorem, the optimal decision should have the same order as in (22) if Bayes' form has the property (21). This proves the theorem.

4. *Inequalities and Bounds on Error-Probability*

Theorem 7. Let $H_f(\xi|\eta)$ be a generalized equivocation and $\delta(y)$ be an arbitrary decision. Then, it holds that

$$H_f(\xi|\eta)\leq[1-P_e(\delta)]f(1-\varepsilon)+P_e(\delta)f(\varepsilon/(n-1)), \tag{23}$$

where ε is an arbitrary number such that $0<\varepsilon<1$.

This inequality is an extension of Fano's inequality (see Fano [8] and Arimoto [6]) and plays an important role in coding theory. In fact, when $f(t)=(1-t)^{1-\beta}/(1-\beta)$ and Y is a discrete set, the inequality (23) becomes

$$[1-P_e(\delta)]\varepsilon^{1-\beta}/(1-\beta)+P_e(\delta)[1-\varepsilon/(n-1)]^{1-\beta}/(1-\beta)$$
$$\geq\frac{1}{1-\beta}\left[1-\sum_{y\in Y}\left\{\sum_{k=1}^{n} p_k{}^{1/\beta}P_k(y)^{1/\beta}\right\}^{\beta}\right]. \tag{24}$$

In particular, if $0<\beta<1$ and $p_k=1/n$ for all k, it follows from letting $\varepsilon\to 0$ in (24)

that

$$P_e(\delta)\geq 1-\frac{1}{n}\sum_{y\in Y}\{\sum_{k=1}^{n}P_k(y)^{1/\beta}\}^{\beta}, \tag{25}$$

which is essential in proof of the strong converse to the coding theorem (see Arimoto[7]).

Conversely, it is interesting to see that

Theorem 8. Assume that $p_k=1/n$ for all k and let P_e be the minimum of error-probability. Then,

$$P_e\leq\frac{1}{n}e^{-H(\gamma)}\sum_{y\in Y}\sum_{k=1}^{n}P_k(y)^{1-\varepsilon\gamma}\{\sum_{i\neq k}P_i(y)^{\varepsilon}\}^{\gamma} \tag{26}$$

for any γ and ε such that $0\leq\gamma\leq 1$ and $0\leq\varepsilon\gamma\leq$, where

$$H(\gamma)=-\gamma\log\gamma-(1-\gamma)\log(1-\gamma). \tag{27}$$

The inequality (26) follows from the relation

$$1-\max_{i=1\sim n}p_i\leq e^{-H(\gamma)}\sum_{j=1}^{n}p_j^{1-\varepsilon\gamma}(\sum_{i\neq j}p_i^{\varepsilon})^{\gamma} \tag{28}$$

which can be easily derived by using the inequality $1-t\leq e^{-H(\gamma)}[(1-t)/t]^{\gamma}$ for $t\in(0, 1]$ and noting the relation

$$1-\max_{i=1\sim n}p_i=\inf_{q}\sum_{j=1}^{n}p_j(1-q_j)\leq e^{-H(\gamma)}\inf_{q}\sum_{j=1}^{n}p_j\left[\frac{1-q_j}{q_j}\right]^{\gamma}. \tag{29}$$

In fact, letting q be

$$q_j=p_j^{\varepsilon}/(\sum_{i=1}^{n}p_i^{\varepsilon})\qquad\text{for } j=1,\cdots,n \tag{30}$$

and substituting this into (29) leads to (28).

It is easy to see that the inequality (26) yields a slight modification of Gallager's upper bound on the probability of decoding error in the coding theorem (see Gallager[9]).

References

[1] Rènyi, 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] Rènyi, A. (ed.), *Proceedings of the Colloquium on Information Theory, Vol. 1* and 2, Janos Bolyai Mathematical Society, Budapest, Hungary (1967).

[3] Lindley, D. V., On a Measure of the Information Provided by an Experiment, *Ann. Math. Stat.*, 27 (1956), 986-1005.

[4] Arimoto, S., The Basis of Sequential Estimation Process from the Viewpoint of Information Theory, *Information Processing in Japan*, 9 (1969), 51-55.

[5] Arimoto, S., Bayesian Decision Rule and Quantity of Equivocation, *The Transactions of the Institute of Electronics and Communication Engineers of Japan*, 53-C, 1 (1970), 16-22 (in Japanese). English version is available in "Systems Computers Controls, Scripta Electronica Japonica III, Vol. 1 (1970), 17-23".

[6] Arimoto, S., Information-Theoretical Considerations on Estimation Problems, *Information and Control*, 19 (1971), 181-194.

[7] Arimoto, S., On the Converse to the Coding Theorem for Discrete Memoryless Channels, *IEEE Trans. Information Theory, IT-19* (to appear).

[8] Fano, R. M., *Transmission of Information*, MIT Press, Cambridge, Mass., USA (1961).

[9] Gallager, R. G., A Simple Derivation of the Coding Theorem and Some Applications, *IEEE Trans. Information Theory, IT-11* (1965), 3-18.