KYBERNETIKA ČÍSLO 5, ROČNÍK 2/1966

Minimum of Average Conditional Entropy for Given Minimum Probability of Error

LIBUŠE BALADOVÁ

Minimum of average conditional entropy, for given minimum probability of error in a statistical decision problem, is derived.

The average conditional entropy is a measure of uncertainty e.g. of messages after their transmission; the minimum probability of error is a different measure from the former. Therefore it is interesting and useful to be able to compare these two quantities. The maximum of the average conditional entropy for given minimum probability of error is well known; our task is to find the greatest lower bound of it. This problem is a generalization of the problem I dealt in [4] with.*

First we are introducing a notation and definitions. Let

$$L(\xi) = \xi \log \xi\,, \quad \xi > 0\,, \quad L(0) = 0\,.$$

(1) Let X, Y be spaces of n and m elements respectively, where $n > 1$, $m > 1$.

(2) Let $\bar{\mu}$ be the set of all probability distributions on X, i.e.

$$\sum_{x \in X} \mu(x) = 1\,, \quad \mu(x) \geqq 0 \quad \text{for} \quad \mu \in \bar{\mu}\,;$$

let $\bar{\nu}$ be the set of all conditional probability distributions on Y, for given $x \in X$, i.e.:

$$\sum_{y \in Y} \nu(y/x) = 1 \quad \text{for} \quad x \in X,\ \nu \in \bar{\nu}\,.$$

(3) Let $\nu(x, y)$ be the probability distribution on $X \otimes Y$, for a fixed $\mu \in \bar{\mu}$, $\nu \in \bar{\nu}$, defined by

* During the preparation of this paper for the print, I found out that V. A. Kovalevskij in [3] had come to the solution of the same problem without restricting to $f \leqq \frac{1}{2}$. As our method of solution essentially differs from his, we publish it.

$$\nu(x, y) = \mu(x)\, \nu(y/x)\,;$$

let $\nu(x/y)$ be the probability distribution on X, for given y, defined by

$$\nu(x/y) = \frac{\nu(x, y)}{\sum\limits_{x \in X} \nu(x, y)}\,.$$

(4) Let $H_{\mu\nu}(y)$ be the conditional entropy on X, for given element $y \in Y$, i.e.

$$H_{\mu\nu}(y) = -\sum_{x \in X} L(\nu(x/y))\,,$$

(5) Let $H_{\mu\nu}$ be the average conditional entropy on X, for given Y, i.e.

$$H_{\mu\nu} = \sum_{y \in Y} \nu(y) \left(-\sum_{x \in X} L(\nu(x/y)) \right),$$

where $\nu(y)$ is the probability on Y given by

$$\nu(y) = \sum_{x \in X} \mu(x)\, \nu(y/x) = \sum_{x \in X} \nu(x, y)\,,$$

(6) Let $f_{\mu\nu}$ be the minimum probability of error, i.e.

$$f_{\mu\nu} = \sum_{y \in Y} \nu(x_y^c, y)\,,$$

where

$$x_y^c = X - \{x_y\}\,,$$

and x_y is defined by means of

$$\nu(x_y, y) = \max_{x \in X} \nu(x, y)\,.$$

(7) Let $\bar{f}$ be, for every $f \in \langle 0, 1 \rangle$, the set of all

$$(\mu, \nu) \in \bar{\mu} \otimes \bar{\nu}$$

such that

$$f_{\mu\nu} = f\,.$$

A. Ya. Khinchin stated in [2] that $\bar{f} \neq 0$, $0 \leqq f \leqq (n-1)/n$, and $\max\limits_{\bar{f}} H_{\mu\nu}$ for every $0 \leqq f \leqq (n-1)/n$ exist and that the following equality is valid:

$$\max_{\bar{f}} H_{\mu\nu} = -L(f) - L(1-f) + f \log(n-1)\,.$$

The problem we shall deal with, is the existence and value of $\min\limits_{\bar{f}} H_{\mu\nu}$. Knowledge of this value together with the Khinchin's assertion enables us to estimate the value of the minimum probability of error by knowing the average conditional entropy and vice versa.

Lemma 1. *$L(\xi + 1)/\xi$ is increasing for $\xi \geqq 0$.*

This simple lemma may be proved by means of differentiation.

Lemma 2. *Let k, l be positive integers, let p_i, $i = 1, \ldots, k$, q_i, $i = 1, \ldots, l$ be non-negative numbers and let*

$$p_{i_0} = \max_i p_i \,, \quad q_{i_0} = \max_i q_i \,,$$

$$0 < p_{i_0} < \sum_i p_i \,, \quad 0 < q_{i_0} < \sum_i q_i \,,$$

$$B = -\sum_i p_i \frac{\log p_i}{\sum p_i} - \sum_i q_i \frac{\log q_i}{\sum q_i} \,.$$

Then there exist non-negative integers r_i, $i = 1, 2$ and real numbers $\varepsilon_i > 0$, $i = 1, 2$ so that $\varepsilon_1 + \varepsilon_2 = p_{i_0} + q_{i_0}$ as well as

1. $$B \geqq -\varepsilon_1 L(r_1 + 1) - \varepsilon_2 L(r_2 + 1) \,,$$

2. $$\sum_{i \neq i_0} p_i + \sum_{i \neq i_0} q_i = \varepsilon_1 r_1 + \varepsilon_2 r_2 \,.$$

Proof. [4] contains the following assertion: the minimum of the entropy $H = -\sum_i L(\mu_i)$, for given $\mu_{i_0} = \max_i \mu_i$, is equal to:

$$-(M + 1)\, L(\mu_{i_0}) - L(\mu_0) \,,$$

where

$$M = \left[(1 - \mu_{i_0})\, \mu_{i_0}^{-1}\right], \quad \mu_0 = 1 - (M + 1)\, \mu_{i_0} \,.$$

The proof of this assertion, being independent on the fact that $\sum_i \mu_i$ is equal to 1 or to another positive constant, can be employed as proof of the inequality

$$B \geqq B_1(c_1, \varepsilon_1) + B_2(c_1, \varepsilon_1) \,,$$

where:

$$B_1(c_1, \varepsilon_1) = -(r_1' + 1)\, L(\varepsilon_1) - L(c_1 - r_1'\varepsilon_1) + L(c_1 + \varepsilon_1) \,,$$

$$B_2(c_1, \varepsilon_1) = -(r_2' + 1)\, L(d - \varepsilon_1) - L(c - c_1 - r_2'(d - \varepsilon_1)) + L(c - c_1 + d - \varepsilon_1) \,,$$

and

$$c_1 = \sum_{i \neq i_0} p_i \,, \quad c = c_1 + \sum_{i \neq i_0} q_i \,,$$

$$\varepsilon_1 = p_{i_0} \,, \quad \varepsilon_2 = q_{i_0} \,, \quad d = \varepsilon_1 + \varepsilon_2 \,,$$

(*) $$r_1' = \left[\frac{c_1}{\varepsilon_1}\right], \quad r_2' = \left[\frac{c - c_1}{\varepsilon_2}\right].$$

By means of differentiation we easily find out, that both the functions $B_i(c_1, \varepsilon_1)$, $i = 1, 2$, are concave in the whole domain of ε_1, c_1 under consideration and therefore $\tilde{B} = B_1 + B_2$ is concave, too. [1] contains the following assertion: the minimum of a concave function, defined on a closed interval, is equal to the value of this function in one of the boundary points. We can easily extend this result of functions of more variables and therefore

$$\tilde{B}(c_1, \varepsilon_1) \geqq \min \{\tilde{B}(c_{11}, \varepsilon_{11}), \tilde{B}(c_{12}, \varepsilon_{11}), \tilde{B}(c_{11}, \varepsilon_{12}) \tilde{B}(c_{12}, \varepsilon_{12})\},$$

where

$$c_{11} = \inf c_1, \quad c_{12} = \sup c_1, \quad \varepsilon_{11} = \inf \varepsilon_1, \quad \varepsilon_{12} = \sup \varepsilon_1$$

holds.

By the calculation*) of c_{1i}, ε_{1j}, $i, j = 1, 2$, from (*) we easily obtain numbers r_{1ij}, r_{2ij}, such that

$$\tilde{B}(c_{1i}, \varepsilon_{1j}) = \varepsilon_1 L(r_{1ij} + 1) + \varepsilon_2 L(r_{2ij} + 1),$$

$$c = \varepsilon_1 r_{1ij} + \varepsilon_2 r_{2ij},$$

q.e.d.

Theorem. *For every $f \in \langle 0, \frac{1}{2} \rangle$ the following inequality holds:* $\inf H_{\mu\nu} \geqq f\, L(2)$.

Proof. (I) Let there be $\bar{y} \in Y$ such that:

$$\nu(x^c_{\bar{y}}, \bar{y}) = f.$$

Then we obtain:

$$\text{if } y \neq \bar{y} \quad \text{and} \quad x \neq x_y, \quad \text{then} \quad \nu(x/y) = 0;$$

$$\text{if } y \neq \bar{y}, \quad \text{then} \quad \nu(x_y/y) = 1.$$

Therefore, in accordance with the definition,

$$H_{\mu\nu} = -\sum_{x \in X} L(\nu(x, \bar{y})) + L(\nu(x_{\bar{y}}, \bar{y}) + f).$$

If we denote $\xi = \nu(x_{\bar{y}}, \bar{y})$, then clearly $\xi \leqq 1 - f$.

According to [4], for every ξ such that

(*)
$$r\xi \leqq f < (r + 1)\,\xi, \text{ i.e. } r = \left[\frac{f}{\xi}\right],$$

the following inequality takes place:

$$H_{\mu\nu} \geqq B_r(\xi) = -(r + 1)\, L(\xi) - L(f - r\xi) + L(f + \xi).$$

* E.g. if we calculate from (*) $c_{11} = r_1' \varepsilon_1$ we become $B_1(c_{11}, \varepsilon_1) = \varepsilon_1 L(r_1' + 1)$. If, for c_{11}, we calculate $\varepsilon_{11} = (r_2' d - c)/(r_2' - r_1')$ we become $B_2 = \varepsilon_2 L(r_2' + 1)$, q . e . d. At the same time we see that the case $r_1' = r_2'$ must be calçulated separately and this will be easily done by means of differentiation.

It remains to prove that, for every $r = 0, 1, 2, \ldots$, $B_r(\xi) \geqq f\,L(2)$ holds. By means of differentiation it can be proved that $B_r(\xi)$ is a concave function for every r and therefore in accordance with [1]

$$B_r(\xi) \geqq \min\{B_r(\xi_{\inf}), B_r(\xi_{\sup})\},$$

where $\xi_{\inf}$, $\xi_{\sup}$ are defined by (*).

If $r = 0$, then $f < \xi$, $\xi \leqq 1 - f$; therefore

$$B_0(\xi) \geqq \min\{B_0(f), B_0(1-f)\} \geqq f\,L(2),$$

where the validity of the second inequality is evident for $B_0(f)$, and, for $B_0(1-f)$ it may be verified by means of differentiation and application of Hardy's theorem.

If $r \geqq 1$, then

$$B_r(\xi) \geqq \min\left\{B_r\left(\frac{f}{r}\right), \quad B_r\left(\frac{f}{r+1}\right)\right\} = \min\left\{f\frac{L(r+1)}{r}, \quad f\frac{L(r+2)}{r+1}\right\}$$

and this expression is not, according to lemma 1, smaller than $f\,L(2)$.

(II) Suppose that $\nu(x_y^c, y) < f$ for every $y \in Y$. Let Y_1, Y_2 be a disjoint decomposition of the space Y defined by $y \in Y_1 \Leftrightarrow \nu(x_y^c, y) > 0$, and let's denote the number of elements in Y_1 by n_1. Then, according to the definition, $H_{\mu\nu}(y) = 0$ for every $y \in Y_2$ and

$$H_{\mu\nu} = -\sum_{y\in Y_1}\sum_{x\in X}\nu(x, y)\log\left(\frac{\nu(x, y)}{\sum_{x\in X}\nu(x, y)}\right).$$

If we put $k, l = n$, $p_i = \nu(x_i, y_1)$, $q_i = \nu(x_i, y_2)$ for $x_i \in X$, $i = 1, 2, \ldots, n$, then the assumptions of lemma 2 are satisfied for every $y_1, y_2 \in Y_1$. Therefore, for even n_1 from lemma 2 follows the existence of numbers $\varepsilon_i > 0$, $i = 1, \ldots, n_1$ and non-negative integers r_i, $i = 1, \ldots, n_1$ such that

(*)
$$H_{\mu\nu} \geqq \sum_{i=1}^{n_1}\varepsilon_i L(r_i + 1),$$

$$\sum_{i=1}^{n_1}\varepsilon_i \leqq 1 - f,$$

$$\sum_{i=1}^{n_1}\varepsilon_i r_i = f.$$

If n_1 is odd, then applying lemma 2 to the set $Y_1' = Y_1 - \{\bar{y}\}$, $\bar{y} \in Y_1$, we obtain $\varepsilon_i' > 0$, $r_i' \geqq 0$ such that:

$$H_{\mu\nu} \geqq \sum_{i=1}^{n-1}\varepsilon_i' L(r_i' + 1) + \sum_{x\in X}\nu(x, \bar{y})\,H_{\mu\nu}(\bar{y}),$$

$$\sum_{i=1}^{n_1=1} \varepsilon_i' r_i' = f - \nu(x_{\bar{y}}^c, \bar{y}) .$$

Acording to (II), there exists r_i' such that $r_i' > 0$. Let e.g. be $r_1' > 0$. If we apply lemma 2 for $k = n$, $p_i = \nu(x_i, \bar{y})$, $x_i \in X$, $i = 1, 2, \ldots, n$, $q_i' = \varepsilon_1'$, $i = 1, 2, \ldots, r_1' + 1$, we obtain ε_1'', ε_2'', r_1'' and r_2'' such that (*) holds for $\varepsilon_1 = \varepsilon_1''$, $\varepsilon_i = \varepsilon_i'$, $i = 1, 2, \ldots, \ldots, n_1 - 1$, $\varepsilon_{n_1} = \varepsilon_2''$, $r_1 = r_1''$, $r_i = r_i'$, $i = 2, \ldots, n_1 - 1$, $r_{n_1} = r_2''$.

Therefore (*) holds for every n_1. If we put: $\varepsilon_i^* = \varepsilon_i r_i$, $i = 1, \ldots, n_1$, be obtain:

$$H_{\mu\nu} \geqq \sum_{i=1}^{n_1} \varepsilon_i^* \frac{L(r_i + 1)}{r_i} ,$$

$$\sum_{i=1}^{n_1} \varepsilon_i^* = f .$$

Let $r = \min_{1 \leqq i \leqq n_1} r_i \geqq 1$; then according to lemma 1, the following inequality holds:

$$H_{\mu\nu} \geqq f \frac{L(r + 1)}{r} .$$

If we apply again lemma 1, we obtain:

$$H_{\mu\nu} \geqq f\, L(2) .$$

As for every average conditional entropy on X, Y being given, (I) or (II) is valid, then $f\, L(2)$ is the lower boundary of it. The method applied by us cannot be directly used for $f > \frac{1}{2}$.

Corolary. $\operatorname{Min}_f H_{\mu\nu}$, $f \in \langle 0, \frac{1}{2} \rangle$ *exists and is equal to* $f\, L(2)$.

Proof.: According to the previous assumptions, spaces X and Y contain at least two elements, say x_1, x_2 resp. y_1, y_2. We put:

$$\mu(x_1) = 1 - f,\ \mu(x_2) = f ;$$

$$\nu(y_1/x_1) = \frac{f}{1 - f} , \quad \nu(y_2/x_1) = \frac{1 - 2f}{1 - f} ,$$

$$\nu(y_1/x_2) = 1 .$$

It is easily proved, that the values of $H_{\mu\nu}$ and $f_{\mu\nu}$ satisfy the required equality.

(Received January 27th, 1966.)

[1] Hardy G. H., Littelwod J. E., Pólya G.: Inequalities. London 1952.
[2] Хинчин А. Я.: Об основных теоремах теории информации. Усп. Мат. Наук, *XI* (1956).
[3] Ковалевский В. А.: Задача распознавания образов с точки зрения математической статистики. In: Читающие автоматы и распознавание образцов. Киев 1965.
[4] Votavová L.: Ein Satz von Extremen der Entropie. Trans. of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, 1956. Praha 1957.

VÝTAH

Minimum střední podmíněné entropie pro danou nejmenší pravděpodobnost chyby

LIBUŠE BALADOVÁ

V práci se odvozuje minimální hodnota střední podmíněné entropie při dané nejmenší pravděpodobnosti chyby (Bayesovském riziku). Střední podmíněná entropie je mírou nejistoty např. zpráv po průchodu kanálem, to jest po jejich přenosu. Podobně nejmenší pravděpodobnost chyby je jinou mírou této neurčitosti. Protože střední podmíněná entropie není monotónní funkcí nejmenší pravděpodobnosti chyby, nemůžeme tyto dvě veličiny přímo porovnat. Nalezne-li se však minimální a maximální hodnota střední podmíněné entropie pro danou nejmenší pravděpodobnost chyby, bude srovnání obou veličin možné, např. též graficky.

Maximální hodnota střední podmíněné entropie při dané nejmenší pravděpodobnosti chyby je již delší dobu dobře známa; v této práci je na intervalu $0 \leqq f \leqq \frac{1}{2}$ nalezena nezáporná funkce argumentu f, jejíž hodnota je dolní dosažitelnou mezí všech podmíněných entropií, které odpovídají dané hodnotě f.

Dr. Libuše Baladová, Ústav teorie informace a automatizace ČSAV, Praha 2 - Nové Město, Vyšehradská 49.