SOME REMARKS ON THE CAPACITY OF A COMMUNICATION CHANNEL

MINORU SAKAGUCHI

University of Electro-communications, Chōfu, Tokyo

(Received Aug. 17. 1960)

One of the most important and well-known theorems in information theory is the capacity theorem (Theorem 1) of a discrete, memoryless and regular channel. Another is the matching theorem (Theorem 3) about the information systems with symbols possessing time-durations. Although the two are seemingly irrelevant to each other, a close connection between them will be shown in Section 2 by presenting a general theorem which includes the two theorems as the two special cases. In Section 1 an interpretation of the capacity theorem is given by introduing the cost consideration into the information system.

Information theory in its communication-engineering sense, the author thinks, consists of the two main branches, coding theory and information transmission theory. As to the former we do not know any example of applications to operations research problems. But as to the latter, Professor Kunisawa of Tokyo Institute of Technology has devoted his efforts in these several years to the exploration of both the theory and case studies [2]. The present note will be a little and further contribution along this line.

1. THE MATCHING OUTPUT-PROBABILITIES TO A REGULAR CHANNEL

The transmission of information requires the presence of a source of information coupled with an appropriate channel; the two together form what it is proposed to call an information system, or briefly a system. Hence an information system is described in terms of joint probabilities of inputs and outputs, and a channel is defined by its transition probabilities, we shall confine ourselves to channels with finite alphabets and zero memory; moreover it will be assumed that the successive inputs to the channel are independent.

Let the distinct inputs to the channel, that is, the symbols of the input alphabet, be numbered as i=1$\cdots$, n, and the outputs as j=1,$\cdots$, n.

Let p_{ij} (i, j=1, $\cdots$, n) be the conditional probabilities of an output number j given that the input had number i. The probability n-vector

$(p_{i1}, \cdots, p_{in})$ will be simply denoted by $p_{i\cdot}$. The set of n vectors $\{p_{i\cdot}\}_{i=1}^{n}$ characterizes a channel.

If ξ_i $(i=1, \cdots, n)$ is the probability of an input number i, then the amount of information transmitted per symbol is given by

$$T(\xi|p_{1\cdot}, \cdots, p_{n\cdot}) \equiv \sum_{i,j=1}^{n} \xi_i p_{ij} \log \frac{p_{ij}}{\sum_i \xi_i p_{ij}}$$

where $\xi=(\xi_1, \cdots, \xi_n)$ is the probability n-vector representing the input probability distribution. The capacity of the channel $\{p_{i\cdot}\}_{i=1}^{n}$ is defined by Shannon (Shannon and Weaver, 1949) as

$$C(p_{1\cdot}, \cdots, p_{n\cdot}) \equiv \max_{\xi} T(\xi|p_{1\cdot}, \cdots, p_{n\cdot}),$$

that is, the maximum rate of information transmation for all choices of the source probabilities. One of the most important theorems about the capacity of a memory-less channel with a finite alphabet is the following one (Muroga, 1953; Sakaguchi, 1959):

THEOREM 1. Let $p_{i\cdot}=(p_{i1}, \cdots, p_{in})$ $(i=1, \cdots, n)$ be n probability n-vectors which are linearly independent. Suppose that there exists a probability n-vector ξ^* such that

$$\sum_{i=1} \xi_i^* p_{ij} = e^{-X_j} (\sum_j e^{-X_j})^{-1}, \quad j=1, \cdots, n \qquad (1)$$

and $\xi_i^*>0$ $(i=1, \cdots, n)$, where the vector $X=(X_1, \cdots, X_n)$ is defined by

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = (p_{ij})^{-1} \begin{bmatrix} H_1 \\ \vdots \\ H_n \end{bmatrix} \qquad (2)$$

in which we have set

$$H_i \equiv -\sum_j p_{ij} \log p_{ij} \qquad (i=1, \cdots, n).$$

Then ξ^* maximizes the transmisson rate $T(\xi|p_{i\cdot}, \cdots, p_{n\cdot})$ and the capacity $C(p_{1\cdot}, \cdots, p_{n\cdot})$ is equal to $\log\left(\sum_1^n e^{-X_j}\right)$.

The proof is found in the literature (Sakaguchi, 1959). We shall present here an interesting interpretation of this capacity theorem.

Consider an information system and a person who observes the output number j.

He does not know the input probabilities ξ_i. Hence, of cource, he does not know the output probability distribution $p(j)$, even if he has full knowledge of the channel characteristic, *i. e.*, the set of the conditional

probabilities $\{p_{i\cdot}\}$.

Given the channel characteristic, he can compute the corresponding output probabilities $p(j)$ if he assumes an input probability distribution ξ_i. Suppose that if he observes the output number j there occurs a gain for him which is equal to the amount of information $-\log p(j)$ and a "cost" which is given by a real number X_j. Thus, when the output probabilities are $p(j)$, of which he is ignorant, his expected net gain will be

$$\sum_{j=1}^{n} p(j)\{-\log p(j)-X_j\}. \tag{3}$$

It must be noted that we shall not set the restriction that the cost must be a positive quantity. We allow the cases in which some X_j are non-positive.

Now we shall present an easy theorem which is as follows.

THEOREM 2. *If $X_j (j=1, \cdots, n)$ are given real numbers, the maximum of* (3) *for all choices of probabilities*

$$p(j)\geqq 0,\ (j=1, \cdots\cdots, n);\quad \sum_{1}^{n} p(j)=1,$$

is attained by

$$p^*(j)=e^{-X_j}(\sum_j e^{-X_j})^{-1},\qquad j=1, \cdots, n \tag{4}$$

and the maximum value is $\log(\sum_j e^{-X_j})$.

PROOF. By the strict concavity of $-\sum p(j) \log p(j)$ there evidently exists a unique maximal point. If the Lagrange equations (in the calculus of variations) yield a solution which is a probability n-vector, it provides the maximum. Partially differentiating

$$\sum_j p(j)\{-\log p(j)-X_j\}-\lambda\sum_j p(j)$$

with respect to $p(j)$ and equating to zero, we get the stated result.

Let us call the ξ^* and the $\sum_{i=1}^{n} \xi_i^*\, p_{i\cdot}$ stated in Theorem 1 as the *matching input*-probabilities and the *matching output*-probabilities to the channel $\{p_{i\cdot}\}$ respectively. Similarly we shall call the $p^*(j)$ stated in Theorem 2 as the *matching probabilities* to the costs $\{X_j\}$.

And let us finally set a definition as follows: a channel $\{p_{i\cdot}\}_{i=1}^{n}$ will be said to be *regular* if n probability n-vectors $p_{i\cdot}$ $(i=1, \cdots, n)$ are

linearly independent and if there exists a probability n-vecotor ξ^* which has positive components and satisfies equations (1) and (2).

Then from theorems 1 and 2, the following corollary follows at once.

COROLLARY. *The matching output-probabilities to a regular channel* $\{p_{i\cdot}\}_{i=1}^n$ *equal the matching probabilities to the costs* $\{X_j\}$ *satisfying the equation*

$$(p_{ij})\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}=\begin{bmatrix} H_1 \\ \vdots \\ H_n \end{bmatrix} \tag{2'}$$

where $H_i \equiv -\sum_j p_{ij} \log p_{ij}$ $(i=1, \cdots, n)$. *Moreover, the channel capacity equals the maximum expected net gain in the case of having these costs.*

The interpretation of the condition (2') is that the "matching costs" must be set at the level where the expected cost when given the input number i is just balanced by the conditional entropy (uncertainty) H_i.

The unique solution-vector of (2') is not necessarily positive. Clearly if some $H_i=0$ then some $X_j \leqq 0$. And even when all H_i are positive, we may have negative X_j, as is shown in the following example:

$$(p_{ij})=\begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$H_1=H_2=\log 2$, $H_3=\log 3$,

$X_2=X_3=3\log 3-2\log 2>0$, $X_1=4\log 2-3\log 3<0$.

The connections between non-positive "matching cost" and the matching output-probabilities and input-probabilities or regularity of the channel are open problems for further research.

2. A GENERAL CAPACITY THEOREM

Consider an information system with output probabilities $p(j)$ $(j=1, \cdots, n)$, and suppose that the output symbol j requires a time-duration (or a cost) t_j. The average amount of information per unit time-duration is

$$-\sum_j p_j \log p_j / \sum_j p_j t_j. \tag{5}$$

One of the most important and well-known theorems in applications of information theory is the following.

THEOREM 3. *Let* t_j $(j=1, \cdots, n)$ *be givin positive numbers. The*

average information (5) *is a maximum when the* p_j*'s are taken as*

$$p_j^* = e^{-Ct_j}, \qquad j=1, \cdots, n \tag{6}$$

were C is the unique positive root of the equation

$$\sum_{j=1}^{n} e^{-Ct_j} = 1. \tag{7}$$

The maximum value of the average information equals C.

Although the two important theorems 1 and 3 are seemingly unrelated there exists a connection between them. We shall show this point by proving a general theorem of which theorems 1 and 3 are two special cases.

Now let us consider an information channel $\{p_{i\cdot}\}_{i=1}^{n}$ which converts the input symbol i to the output symbol j with probability p_{ij} and with a cost (or time-duration) t_{ij} $(i, j=1, \cdots, n)$.

For an information system with this information channel the averge amount of transmitted information per unit cost is

$$\frac{\sum_{i,j} \xi_i p_{ij} \log(p_{ij}/\sum_i \xi_i p_{ij})}{\sum_{i,j} \xi_i p_{ij} t_{ij}} \tag{8}$$

where ξ is the probability n-vector describing the input probabilities of the system.

We note here that we can rewrite (8) as

$$\frac{\sum_i \xi_i I(p_{i\cdot}\,;\, \sum_i \xi_i p_{i\cdot})}{\sum_i \xi_i s_i} \tag{8'}$$

where we have set

$$s_i \equiv \sum_j p_{ij} t_{ij} \qquad (i=1, \cdots, n)$$

and

$$I(p_{i\cdot}\,;\, \sum_i \xi_i p_{i\cdot}) \equiv \sum_j p_{ij} \log (p_{ij}/\sum_i \xi_i p_{ij})$$

is the Kullback-Leibler information amount (Kullback, 1959).

LEMMA. *If there exists a convex linear combination* $\sum_i \xi_i p_{i\cdot}$ *of the probability vectors* $p_{i\cdot}$ $(i=1, \cdots, n)$ *with positive coefficients* $\xi_i^{*\prime}s$

$$I(p_{i\cdot}\,;\, \sum_i \xi_i^* p_{i\cdot})/s_i = \text{indep. cf i}, \tag{9}$$

then the probability n-vector ξ^* *maximizes the average transmitted information* (8').

PROOF. Partially differentiating the Lagrangian function

$$\log \sum_i \xi_i I(p_{i\cdot}\,;\ \sum_i \xi_i p_{i\cdot}) - \log \sum \xi_i s_i + \lambda \sum \xi_i$$

with respect to ξ_i $(i=1, \cdots, n)$ and equating to zero, we get

$$H_i + \sum_j p_{ij} \log \sum_i \xi_i p_{ij} + 1 = (\lambda - s_i / \sum_i \xi_i s_i) T(\xi) \qquad (i=1, \cdots, n),$$

where

$$T(\xi) \equiv \sum_i \xi_i I(p_{i\cdot}\,;\ \sum \xi_i p_{i\cdot}).$$

By multiplying both sides of the equality by ξ_i and summing up, we obtain

$$\lambda = 1/T(\xi).$$

Hence we have

$$I(p_{i\cdot}\,;\ \sum \xi_i p_{i\cdot})/s_i = T(\xi)/\sum \xi_i s_i = \text{indep. of } i$$

which is the desired result.

The following theorem is an immediate consequence of this lemma.

THEOREM 4. *Let $p_{i\cdot}$ $(i=1, \cdots, n)$ be n probability n-vectors which are linearly independent. Suppose that there exist a unique positive number C and a probability n-vector ξ^* with $\xi_i^* > 0$ $(i=1, \cdots, n)$ such that*

$$\sum_i \xi_i^* p_{ij} = e^{-X_j}, \qquad j=1, \cdots, n \tag{10}$$

and

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = (p_{ij})^{-1} \begin{bmatrix} H_1 + s_1 C \\ \vdots \\ H_n + s_n C \end{bmatrix}. \tag{11}$$

Then the probability n-vector ξ^* maximizes the average trasmitted information (8′), and the maximum value is C.

PROOF. From the lemma and (10) and (11) we have

$$s_i^{-1} I(p_{i\cdot}\,;\ \sum_i \xi_i^* p_{i\cdot}) = s_i^{-1} \sum_j p_{ij} \log(e^{-X_j} / \sum_i \xi_i^* p_{ij}) + s_i^{-1}(\sum_j p_{ij} X_j - H_i) = C$$

completing the proof.

As a special case of this theorem consider the case where all the t_{ij} are equal to t, say. The solution-vector X of (11) can be represented by

$$X_j = \bar{X}_j + tC \qquad (j=1, \cdots, n),$$

where $\bar{X}$ is determined by the equation (2). Thus by (10) it follows that

$$\sum_i \xi_i^* p_{ij}=e^{-\bar{X}_j}(\sum_j e^{-\bar{X}_j})^{-1}$$

and

$$C=t^{-2}\log(\sum_j e^{-\bar{X}_j}),$$

reducing to Theorem 1.

As another special case of our Theorem 4 we consider the case when $p_{ij}=\delta_{ij}$ (Kronecker's δ). The expression (8) reduces to

$$-\sum_i \xi_i \log \xi_i/\sum_i \xi_i t_i$$

since $H_i=0$ and $s_i\equiv\sum_j p_{ij}t_{ij}=t_{ii}(\equiv t_i$, say), $(i=1, \cdots, n)$. From (10) and (11) we get

$$\xi_j^*=e^{-X_j} \qquad (j=1, \cdots, n)$$

and

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}=(p_{ij})^{-1}\begin{bmatrix} t_1 & C \\ \vdots & \\ t_n & C \end{bmatrix}=C\begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}$$

respectively. We have thus arrived at the equations (6) and (7) in Theorem 3.

At the end of this section we shall remark that a sufficient condition for the existence of a unique positive C satisfying (10) and (11) is given by $\min_i \sum_j a_{ij}s_j>0$, where $(a_{ij})=(p_{ij})^{-1}$. In fact, we have by (10) and (11) $\sum_i A_i e^{BiC}=1$ where $A_i\equiv e^{-\sum_j a_{ij}H_j}$ and $B_i\equiv-\sum_i a_{ij}s_j$. It is easily found that the function $f(x)\equiv\sum A_i e^{Bix}$ is convex for $x>0$, and $f(0)>1$ and $f'(x)<0$. Hence we get the stated sufficiency.

ADDENDUM. Professor Peter Elias of Massachusetts Institute of Technology read carefully the manuscript and made the following comments on some technical points concerning our Theorem 4.

"It is impossible for me to see any channel of interest in a communications problem for which this is a useful mathematical model, except for the two known special cases to which you refer.

"If t_{ij} is not a constant, but varies with i and with j, and if the channel is really noisy, then it seems that any interpretation of t_{ij} as a time duration, as you suggest, would lead the receiver into hopeless confusion. That is, how the transmitter know how long to wait before sending the next symbol ? If the transmitter always waits long enough for

the longest t_{ij} for a given i, then the actual t_{ij} is a function of i alone, and not of j. Then what is going over the channel when symbols of shorter duration are received?

"That is, it may be impossible in a truly noisy channel, using symbols of different durations, for the receiver to keep track of where the transmitter is, and to know whether he is currently in the middle of a symbol or is just starting a new one.

"On the other hand if t_{ij} represents another kind of cost, then the only kind which I have been able to think of, such as transmitter power required, also depends only on i and not on j.

I was glad to receive these comments, and to include them here, even though I have my interpretation of the model somewhat differently:

An office, especially business management, is composed of some departments or sections, where various documents flow from one to the other. The flow pattern of documents inside the office is most easily described by saying that a document, after passing through a given department, flows to some specified section, its particular course being governed by a fixed probability distribution associated with the particular department that it is leaving.

Let t_{ij} be the cost required during the transformation from department i to department j. If the office handles N documents per unit time, the amount of information managed in this office per unit time and unit cost will be represented by the expression (8) multiplied by N.

At any rate it seems to the author that we may be able to interprete our Theorem 4 in some appropriate way from operations research viewpoints and some case studies in this field of management science are highly wanted.

REFERENCES

[1] Kullback, S. (1959). "Information Theory and Statistics." John Wiley, New York.

[2] Kunisawa, K. (1959). "Introduction to Information Theory for OR Workers." (in Japanese) JUSE Publishing Company, Tokyo.

[3] Muroga, S. (1953). On the capacity of a discrete channel I. *Journ. Phys. Soc. Japan.*

[4] Sakaguchi, M. (1959). Notes on statistical applications of information theory IV. *Rep. Stat. Appl. Res. JUSE.* **6**, 54—57.
[5] Shannon, C. E. and Weaver, W. (1949). "The Mathematical Theory of Communication". Univ.of Illinois Press, Urbana, Illinois.