ON THE BASIC THEOREMS OF INFORMATION THEORY

By KINSAKU TAKANO

(Received Oct. 25, 1957)

Introduction and summary

The theorems of information theory on the relation between the capacity of a noisy channel and entropy of a source have interested many mathematicians, since Shannon's pioneering work: Shannon [5]. Particularly we must cite the works of McMillan [4], Feinstein [1] and Hinčin [3]. McMillan [4] refined Shannon's concepts and generalized some of his results. Feinstein [1] reformulated the relationship in question and introduced a very excellent idea. Recently, Hinčin has written a full treatment (Hinčin [3]) on the basis of these works. Regretfully, however, this treatment have some accounts which are very long and are hard to read. The purpose of this paper is to refine some proofs of Hinčin [3], to note a careless mistake there (see § 1), and to reformulate the second theorem of Shannon given by Hinčin [3], § 16 (see § 4, Theorem 4.2). We have intended to write as completely as possible, because Hinčin's paper [3] has been written in Russian and information theory will interest most of probabilists and statisticians.

§ 1. Preliminaries

1.1. Information sourre. By definition, an alphabet is a finite set, any element of which is called its letter or symbol. Let A be an alphabet, and let I denote the sequence of all integers: $I=(\cdots, -1, 0, 1, 2, \cdots)$. Then A^I denotes the class of infinite sequences

$$x=(\cdots, x_{-1}, x_0, x_1, x_2, \cdots), \qquad x_i \in A, \quad i \in I.$$

Let F_A be the Borel field of subsets of A^I determined by all cylinder sets. If μ is a probability measure defined over the Borel field F_A, the probability space $[A^I, F_A, \mu]$*, or, the stochastic process

$$(1) \qquad \cdots, x_{-1}, x_0, x_1, x_2, \cdots$$

which is constructed by the coordinate variables in this space, is called

* Let us suppose that F_A is enlarged so that μ is completed if necessary.

an information source, and is denoted simply by $[A, \mu]$. Occasionally, the stochastic process (1) is said to represent the source $[A, \mu]$. An information source is said to be stationary or ergodic, if the stochastic process representing it is strictly stationary or ergodic.

We shall denote by T the coordinate-shift transformation in A^I:

$$(Tx)_n=(x)_{n+1}\,, \qquad x\in A^I\,, \quad -\infty<n<\infty\,,$$

where $(x)_n$ denotes the nth coordinate of x, that is,

$$(x)_n=x_n$$

if $x=(\cdots, x_{-1}, x_0, x_1, \cdots)$. Let χ_S be the indicator of a set $S\in F_A$, that is, the function such that $\chi_S(x)=1$ or 0 according as $x\in S$ or not. If for almost all $x\in A^I$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\chi_S(T^kx)=\mu(S)\,,$$

then it is said that the source $[A, \mu]$ "reflects" the set S. Let $a_i\in A$ and let ν_i be integers for $i=1, 2, \cdots, n$. We shall denote by $[a_1, a_2, \cdots, a_n; \nu_1, \nu_2, \cdots, \nu_n]$ the cylinder set over the coordinate numbers $\nu_1, \nu_2, \cdots, \nu_n$

$$\{x;\ x_{\nu_1}=a_1, x_{\nu_2}=a_2, \cdots, x_{\nu_n}=a_n\}$$

where

$$x=(\cdots, x_{-1}, x_0, x_1, x_2, \cdots)\,.$$

We shall use, in the sequel, the following

LEMMA 1.1. *In order that a stationary information source $[A, \mu]$ is ergodic, it is necessary and sufficient that the source reflects all cylinder sets of the form $[a_1, a_2, \cdots, a_n; 1, 2, \cdots, n]$, where $n=1, 2, \cdots, a_i\in A$.* (see Hinčin [3], § 4).

The entropy of a discrete probability distribution $\{p_1, p_2, \cdots, p_n\}$, $p_i\geqq 0$, $\sum p_i=1$, is defined by

$$H(p_1, p_2, \cdots, p_n)=-\sum_i p_i\log p_i\,,$$

where $p_i\log p_i=0$ if $p_i=0$. Let us take 2 as the base of the system of logarithms. $H(p_1, p_2, \cdots, p_n)$ is also denoted by $H(p_i; i=1, 2, \cdots, n)$. Occasionally the entropy of the distribution of a random element x is denoted by $H(x)$. Let x, y be random elements with a discrete joint probability

distribution. The average value of the entropy of the conditional distribution of x when y is given is called the average conditional entropy of x relative to y, and denoted by $H_y(x)$. As is easily verified, we have

$$H_y(x)=H(x, y)-H(y)\ , \tag{2}$$

and

$$H_y(x)\leqq H(x) \tag{3}$$

(see, for instance, Shannon [5], Hinčin [2].)

Let

$$\cdots, x_0, x_1, x_2, \cdots$$

be the stochastic process representing an information source $[A, \mu]$. Put

$$X_n=(x_1, x_2, \cdots, x_n)\ ,$$

and denote the entropy of random sequence X_n of n letters by H_n, that is, put

$$H_n=-E\log \mu(X_n)$$

where E denotes " mean value of ". The H_n can be interpreted as 1) the measure of uncertainty concerning an unknown realization of random sequence X_n, when only its distribution is known, or as 2) the average amount of information when this unknown realization is known. (see Snannon [5] and Hinčin [2]).

We use the following facts.

THEOREM 1.1. *If the information source is stationary, then H_n/n converges to a finite limit $H\geqq 0$.*

$$H=\lim_{n\to\infty}\frac{H_n}{n}\ . \tag{4}$$

This limiting value is called the entropy per symbol of the information source. (4) is rewritten as $\lim_{n\to\infty}E\left[-\frac{1}{n}\log \mu(X_n)\right]=H$. As a stronger proposition we have the following

THEOREM 1.2. (McMillan [4]) *If the information source $[A, \mu]$ is ergodic, then $-\frac{1}{n}\log \mu(X_n)$ converges to H in probability, as $n\to\infty$.*

1.2. Channel. Let A and B be two alphabets. If there exists a

family of information source $[B, \nu_x]$, one for each $x \in A^I$, a triple of two alphabets A and B and the list of probability measures $\{\nu_x;\ x \in A^I\}$ is called an channel, and is denoted by $[A, \nu_x, B]$, and A and B are called input and output alphabets of the channel, respectively.

The concept of stationarity extends to channels. A channel $[A, \nu_x, B]$ is called stationary if, for any $x \in A^I$ and for any $S \in F_B$,

$$\nu_x(S) = \nu_{Tx}(TS) ,$$

where T is the coordinate-shift transformation:

$$(Tx)_n = (x)_{n+1} , \qquad (Ty)_n = (y)_{n+1} , \qquad \text{for} \quad x \in A^I , \quad y \in B^I .$$

Let $[A, \nu_x, B]$ be a channel. Let

$$y^{(j)} = (\cdots, y_{-1}^{(j)}, y_0^{(j)}, y_1^{(j)}, y_2^{(j)}, \cdots) , \qquad j = 1, 2 ,$$

be the stochastic processes representing the source $[B, \nu_{x^{(j)}}]$, corresponding to sequences

$$x^{(j)} = (\cdots, x_{-1}^{(j)}, x_0^{(j)}, x_1^{(j)}, x_2^{(j)}, \cdots) , \qquad j = 1, 2 .$$

Suppose that, for each t, if

$$x_i^{(1)} = x_i^{(2)} , \qquad \text{for all} \quad i \leqq t ,$$

then $(\cdots, y_{t-1}^{(1)}, y_t^{(1)})$ and $(\cdots, y_{t-1}^{(2)}, y_t^{(2)})$ have the same probability distribution. In this case the channel is said to have no foresight. If, moreover, there exists a constant nonnegative integer m such that

$$x_i^{(1)} = x_i^{(2)} , \qquad \text{for} \quad s - m \leqq i \leqq t$$

implies that $(y_s^{(1)}, y_{s+1}^{(1)}, \cdots, y_t^{(1)})$ and $(y_s^{(2)}, y_{s+1}^{(2)}, \cdots, y_t^{(2)})$ are identically distributed, for each pair of integers (s, t) with $s \leqq t$, then the channel is said to have finite memory m. If, for each $x \in A^I$, there corresponds an integer $m_1(x) \geqq 0$ such that the stochastic process

$$\cdots, y_{-1}, y_0, y_1, y_2, \cdots$$

representing the source $[B, \nu_x]$ is $m_1(x)$-step dependent, that is, $(y_p, y_{p+1}, \cdots, y_q)$ and $(y_r, y_{r+1}, \cdots, y_s)$ are independent if $p \leqq q$, $r \leqq s$, $r - q > m_1(x)$, then the channel $[A, \nu_x, B]$ is said to be $m_1(\cdot)$-step dependent.

The proof of Hinčin [3] goes on as if a stationary channel with finite memory m and no foresight be m-step dependent. However, this is not true. For example, let $A = \{1, 2, \cdots, n\}$, and let $\{Y(a_0, a_1, \cdots, a_m);\ a_0, \cdots, a_m \in A\}$ be a family of random elements taking letters of an

alphabet B as values, one for each sequence $(a_0, a_1, \cdots, a_m) \in A^{m+1}$. For each sequence

$$x=(\cdots, x_{-1}, x_0, x_1, \cdots) \in A^I$$

define ν_x by the probability measure over F_B determined by the sequence

$$\cdots, y_{-1}, y_0, y_1, y_2, \cdots$$

where

$$y_i=Y(x_{i-m}, x_{i-m+1}, \cdots, x_i) \ . \tag{5}$$

Thus we have a stationary channel $[A, \nu_x, B]$ with no foresight and with finite memory m, which is not m-step dependent. Next, we shall give an example of a stationary m_1-step dependent channel with no foresight and finite memory m, by a modification of ν_x. For a sequence

$$x=(\cdots, x_{-1}, x_0, x_1, \cdots) \in A^I \ ,$$

let

$$\cdots, y'_{-1}, y'_0, y'_1, y'_2, \cdots \tag{6}$$

be a sequence of random elements taking letters of B as values, such that

i) $$(y'_{(t-1)(m_1+1)+1}, y'_{(t-1)(m_1+1)+2}, \cdots, y'_{t(m_1+1)}) \ ,$$
$$t=\cdots, -1, 0, 1, 2, \cdots$$

is an independent sequence with t as index, and

ii) $$(y'_{(t-1)(m_1+1)+1}, y'_{(t-1)(m_1+1)+2}, \cdots, y'_{t(m_1+1)})$$

has the same distribution with

$$(y_{(t-1)(m_1+1)+1}, y_{(t-1)(m_1+1)+2}, \cdots, y_{t(m_1+1)})$$

for each $t=\cdots, -1, 0, 1, \cdots$, where y_i are defined by (5). Denote by ν'_x the probability measure over F_B determined by the sequence (6). Then we have a new channel $[A, \nu'_x, B]$, which has clearly desired properties.

Interpretation for the above mathematical model of an channel. $i \in I$ denotes time. Let x_i be a letter of A transmitted at time i from the input of a channel, $i=1, 2, \cdots$. Corresponding to the sequence or message transmitted from the input

$$x=(x_1, x_2, \cdots) \ , \qquad x_i \in A \ ,$$

we have a message

$$y=(y_1, y_2, \cdots), \qquad y_t \in B,$$

received by the output. If the transmitted message x is not always able to be uniquely determined by the received message y, then we say that there exists essential interference or noise in the channel. In this case the received message y is considered as a realization of a stochastic process with x as parameter.

Suppose that the durations of all letters are equal each other (quantumization), and suppose that the sequence $x_1, x_2, \cdots, x_n$ yields the sequence $y_1, y_2, \cdots, y_n$ for each $n=1, 2, \cdots$. Then, it is natural to suppose that the distribution of y_n does not depend on $x_{n+1}, x_{n+2}, \cdots$. The distribution of y_n depends not only on the letter x_n, but also, in general, on the past letters $x_1, x_2, \cdots, x_{n-1}$, by the after effect of the channel. If the sphere of influence of the after effect is bounded, then we have the concept of finite memory. Since we shall consider only channels of finite memory in this paper, we can proceed taking $I_0=(1, 2, \cdots)$ instead of $I=(\cdots, -1, 0, 1, 2, \cdots)$ in the definitions of an information source and a channel, but still we use I in order to simplify the mathematical descriptions.

1.3. Composite source of a source and a channel. We wish now to examine a stationary channel "driven" by a stationary source. Consider a source $[A, \mu]$ and a channel $[A, \nu_x, B]$ whose input alphabet coincides with the alphabet of the source. Suppose that the channel $[A, \nu_x, B]$ has finite memory and no foresight. Denote the product space $A\times B=\{(a, b);\ a\in A, b\in B\}$ by C. We can define a measure over F_C by means of

$$\omega(M\times N)=\int_M \nu_x(N)d\mu(x), \qquad \text{for} \quad M\in F_A, \quad N\in F_B,$$

where N is a subset of a finite product space $B\times B\times\cdots\times B$. Note that, from our hypothesis, for such a set N $\nu_x(N)$ depends on only a finite number of coordinates of x and hence is a simple function of x. Thus, we have a new information source $[C, \omega]$, which is called the composite source of the source $[A, \mu]$ and the channel $[A, \nu_x, B]$. Each element of C^I, a sequence of pairs (x_t, y_t), $x_t\in A$, $y_t\in B$, can be considered as a pair (x, y) of sequences $x=(\cdots, x_{-1}, x_0, x_1, \cdots)\in A^I$ and $y=(\cdots, y_{-1}, y_0, y_1, \cdots)\in B^I$. In the composite source $[C, \omega]$, ν_x is interpreted as the

conditional probability measure of y when x is given, ω is the joint distribution of x, y, and μ is the marginal distribution of x. Denoty by η the marginal distribution of y. The sources $[A, \mu]$ and $[B, \eta]$ are called the input and output sources of the composite source $[C, \omega]$, respectively. Let

$$\cdots(x_{-1}, y_{-1}),\ (x_0, y_0),\ (x_1, y_1),\ (x_2, y_2),\ \cdots$$

be the stochastic process representing the product source $[C, \omega]$. Then,

$$\cdots, x_{-1}, x_0, x_1, \cdots \quad \text{and} \quad \cdots, y_{-1}, y_0, y_1, \cdots$$

are the stochastic processes representing the sources $[A, \mu]$ and $[B, \eta]$, respectively. Put

$$X_n=(x_1, x_2, \cdots, x_n)\ , \qquad Y_n=(y_1, y_2, \cdots, y_n)\ .$$

The entropy $H(X_n)$ of random message X_n represents the measure of uncertainty concerning an unknown transmitted message, and the average conditional entropy $H_{Y_n}(X_n)$ of X_n relative to Y_n represents the average measure of uncertainty concerning a transmitted message when a received message is given. If the amount of decrease of the measure of uncertainty is considered as the average amount of increase of information, the difference $H(X_n)-H_{Y_n}(X_n)$ represents the average amount of information obtained by the transmission of random message of n symbols through the channel. Note that $H(X_n)-H_{Y_n}(X_n)$ is always non-negative by help of (3). Now, by the formula (2) we have

$$H_{Y_n}(X_n)=H(X_n, Y_n)-H(Y_n)\ ,$$

and hence,

$$(7) \qquad H(X_n)-H_{Y_n}(X_n)=H(X_n)+H(Y_n)-H(X_n, Y_n)\ .$$

Now, suppose that both the source $[A, \mu]$ and the channel $[A, \nu_x, B]$ are stationary. Then it is easily proved that the composite source $[C, \omega]$ is also stationary (see Hinčin [3], § 10). This stationarity implies the stationarity of the output source $[B, \eta]$. By Theorem 1.1 a stationary source has the entropy per symbol. Denote by H_I, H_O, and H_C the entropies per symbol of the sources $[A, \mu]$, $[B, \eta]$ and $[C, \omega]$ respectively. Then we have

$$\lim \frac{H(X_n)}{n}=H_I\ , \qquad \lim \frac{H(Y_n)}{n}=H_O \qquad \text{and} \qquad \lim \frac{H(X_n, Y_n)}{n}=H_C\ .$$

Hence we have from (7)

$$(8) \qquad \lim \frac{1}{n}\{H(X_n)-H_{Y_n}(X_n)\}=H_I+H_0-H_G=R \qquad \text{(say)}.$$

The limit $R\geqq 0$, which is considered as the average amount of information per symbol, is called the speed of transmission when the channel $[A, \nu_x, B]$ is driven by the source $[A, \mu]$.

The relation (8) can be rewritten as

$$(9) \qquad \lim E\left[\frac{1}{n}\log\frac{\omega(X_n, Y_n)}{\mu(X_n)\eta(Y_n)}\right]=R.$$

If the composite source $[C, \omega]$ is ergodic, then by Theorem 1.2 (9) is strengthened as follows:

$$(10) \qquad \frac{1}{n}\log\frac{\omega(X_n, Y_n)}{\mu(X_n)\eta(X_n)} \text{ converges to } R \text{ in probability}.$$

This property (10) will play an important role in the later discussion. In the next section, we will give a sufficient condition in order that the composite source become ergodic.

§ 2. Ergodicity of the composite source.

Let $[A, \mu]$ be a stationary source, let $[A, \nu_x, B]$ be a stationary channel with no foresight and with finite memory m, and denote by $[C, \omega]$ the composite source of $[A, \mu]$ and $[A, \nu_x, B]$. Let $m_1(\cdot)$ be a non-negative integer-valued function defined on A^I. The purpose of this section is to prove the following

THEOREM 2.1. *If a stationary source $[A, \mu]$ is ergodic, and if a stationary channel $[A, \nu_x, B]$ with no foresight is $m_1(\cdot)$-step dependent and has finite memory m, then the composite source $[C, \omega]$ is ergodic.*

To prove this theorem, we use the following lemmas.

LEMMA 2.1. *Let $[A, \mu]$ be an ergodic source, let u be a set of F_A with $\mu(u)>0$, and for a point $x\in A^I$, denote the values of $k\geqq 0$ such that $T^k x\in u$ by $k_1, k_2, \cdots$ in increasing order. Then, for any positive integer t, and for almost all x, this sequence $k_1, k_2, \cdots$ has an asymptotic distribution modulo t. This means the following: let $\lambda(\tau, n)$ be the number of k_i such that $k_i\equiv\tau \bmod (t)$ and that $k_i<n$ for each $\tau=0, 1, \cdots, t-1$ and*

put $\lambda(n)=\sum_{\tau=0}^{t-1}\lambda(\tau, n)$, *then there exist limits*

$$\lim_{n\to\infty}\frac{\lambda(\tau, n)}{\lambda(n)}=\lambda_\tau\ , \qquad \tau=0, 1, 2, \cdots, t-1\ ,$$

for almost all x.

PROOF. Denote by χ_u the indicator of the set u. By the ergodicity of $[A, \mu]$ we have

$$(1) \qquad \lim_{n\to\infty}\frac{\lambda(n)}{n}=\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\chi_u(T^k x)=\mu(u)>0$$

for almost all x. And, by the Birkhoff's indivisual ergodic theorem, there exist limits

$$(2) \quad \lim_{s\to\infty}\frac{\lambda(\tau, st)}{s}=\lim_{s\to\infty}\frac{1}{s}\sum_{\sigma=0}^{s-1}\chi_u(T^{\tau+\sigma t}x)=\psi_\tau(x)\ , \qquad \tau=0, 1, 2, \cdots, t-1\ ,$$

for almost all x. Let M be the set of x which satisfy (1) and for which the limits (2) exist for all $\tau=0, 1, \cdots, t-1$. Then $\mu(M)=1$. Fix $x\in M$, and put $s=[n/t]$ for each positive integer n, then we have

$$\lambda(\mathrm{st})\leqq\lambda(n)\leqq\lambda(st)+t\ ,$$

$$\lambda(\tau, st)\leqq\lambda(\tau, n)\leqq\lambda(\tau, st)+1$$

which imply with (1) and (2) that

$$\lambda(n)=st(\mu(u)+o(1))\ , \qquad (s\to\infty)\ ,$$

$$\lambda(\tau, n)=s(\psi_\tau(x)+o(1))\ , \quad (s\to\infty)\ .$$

Hence we have

$$\lim_{n\to\infty}\frac{\lambda(\tau, n)}{\lambda(n)}=\frac{\psi_\tau(x)}{t\mu(u)}$$

which completes the proof.

LEMMA 2.2. *Let* $[B, \nu]$ *be an* m_1*-step dependent information source. If for some cylinder set* $v=[b_1, b_2, \cdots, b_l; 1, 2, \cdots, l]$, *and for some infinite sequence of integers* $0\leqq k_1<k_2<\cdots$ *which has asymptotic distribution modulo* m_1+l,

$$\nu(T^{-k_i}v)=c \quad (\text{const.})\ , \qquad i=1, 2, \cdots,$$

then we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\chi_v(T^{k_i}y)=c$$

for almost all $y\in B^I$.

PROOF. Let us note that

$$T^{-k_i}v=[b_1, b_2, \cdots, b_i;\ k_i+1, k_i+2, \cdots, k_i+l]$$

and that $k_i-k_j\geqq m_1+l$ implies that $(k_i+1)-(k_j+l)>m_1$. Put $t=m_1+l$ and write $k\equiv\tau$, if $k\equiv\tau \mod(t)$. From the hypotheses it follows that $\{\chi_v(T^{k_i}y);\ k_i\equiv\tau\}$ is a sequence of independent identically distributed random variables with mean c, for each $\tau=0, 1, \cdots, t-1$. Hence if $\lambda(\tau, n)$, the number of k_i such that $k_i\equiv\tau$ and $k_i<n$, infinitely increases with n, he have

$$\lim_{n\to\infty}\frac{1}{\lambda(\tau, n)}\sum_{k_i<n,\,k_i\equiv\tau}\chi_v(T^{k_i}y)=c \tag{3}$$

for almost all $y\in B^I$, by the strong law of large numbers. Put

$$\lambda(n)=\sum_{\tau=0}^{t-1}\lambda(\tau, n)\ .$$

By the hypothesis there exist limits

$$\lim_{n\to\infty}\frac{\lambda(\tau, n)}{\lambda(n)}=\lambda_\tau \tag{4}$$

for all $\tau=0, 1, \cdots, t-1$, and

$$\sum_{\tau=0}^{t-1}\lambda_\tau=1\ . \tag{5}$$

From (3) and (4) it follows that

$$\lim_{n\to\infty}\frac{1}{\lambda(n)}\sum_{k_i<n,\,k_i\equiv\tau}\chi_v(T^{k_i}y)=\lambda_\tau c \tag{6}$$

for almost all $y\in B^I$. This holds even for τ such that $\lambda(\tau, n)$ is hounded. Hence, neglecting y belonging to a ν-null set, (6) holds for all $\tau=0, 1, \cdots, t-1$. From (5) and (6) we have, with probability one,

$$\lim_{n\to\infty}\frac{1}{\lambda(n)}\sum_{k_i<n}\chi_v(T^{k_i}y)=c$$

which completes the proof.

PROOF OF THEOREM 2.1. By Lemmma 1.1 it is sufficient to prove

that the source $[C, \omega]$ reflects the cylinder set of the form

$$S=[(a_1, b_1), (a_2, b_2), \cdots, (a_l, b_l); 1, 2, \cdots, l]$$

for arbitrary positive integer l, arbitrary $a_1, a_2, \cdots, a_l \in A$ and arbitrary $b_1, b_2, \cdots, b_l \in B$. For this purpose, it suffices to prove that the source $[C, \omega]$ reflects the cylinder set

$$S'=[a_{-m+1}, \cdots, a_0, a_1, \cdots, a_l; -m+1, -m+2, \cdots, l] \times [b_1, b_2, \cdots, b_l; 1, 2, \cdots, l]$$

for arbitrary $a_{-m+1}, \cdots, a_0 \in A$, because S is the finite sum of disjoint S'. Hence, fixing cylinder sets

$$u=[a_{-m+1}, a_{-m+2}, \cdots, a_l; -m+1, -m+2, \cdots, l]$$

and

$$v=[b_1, b_2, \cdots, b_l; 1, 2, \cdots, l],$$

we shall prove that the source $[C, \omega]$ reflects the cylinder set $u \times v$, that is, that we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_{u\times v}(T^k x, T^k y)=\omega(u \times v) \tag{7}$$

for almost all (x, y). Now, the channel $[A, \nu_x, B]$ has finite memory m, and hence the values of $\nu_x(v)$ are equal for all $x \in u$. Denoting this value by $\nu_u(v)$ we have

$$\omega(u \times v)=\int_u \nu_x(v) d\mu(x)=\mu(u)\nu_u(v),$$

and (7) becomes

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_u(T^k x)\chi_v(T^k y)=\mu(u)\nu_u(v). \tag{8}$$

For a point $x \in A^I$, denote the values of $k \geqq 0$ satisfying $T^k x \in u$, by $k_1, k_2, \cdots$ in increasing order, and denote by $\lambda(n)$ the number of k_i which are smaller than n, then the ergodicity of $[A, \mu]$ assures that

$$\lim_{n\to\infty} \frac{\lambda(n)}{n}=\mu(u) \tag{9}$$

for almost all x. If $\mu(u)=0$ then (8) is trivial by (9). Suppose that $\mu(u)>0$. Then Lemma 2.1 assures that the sequence $k_1, k_2, \cdots$ has asymptotic distribution modulo $t=m_1(x)+l$, for almost all x. Let M be

the set of x for each of which (9) holds and the sequence $k_1, k_2, \cdots$ has asymptotic distribution modulo t. Then we have $\mu(M)=1$. It suffices to prove that for each point $x \in M$.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k_i<n} \chi_v(T^{k_i}y)=\mu(u)\nu_u(v) \tag{10}$$

for almost all $y(\nu_x)$. Now fix a point $x \in M$ and consider on the source $[B, \nu_x]$. From the stationarity of the channel and the definition of k_i it follows that

$$\nu_x(T^{-k_i}v)=\nu_{T^{k_i}x}(v)=\nu_u(v).$$

Hence, by Lemma 2.2 we have

$$\lim_{n\to\infty} \frac{1}{\lambda(n)} \sum_{k_i<n} \chi_v(T^{k_i}y)=\nu_u(v) \tag{11}$$

for almost all $y(\nu_x)$. Then (10) follows from (9) and (11). This completes the proof.

§ 3. Feinstein-Hinčin's fundamental lemma

Suppose we are given a stationary m_1-step dependent channel $[A, \nu_x, B]$ with finite memory m and no foresight. Let $[A, \mu]$ be an ergodic information source, the alphabet of which coincides with the input alphabet A of the channel, and let R be the speed of transmission when the channel $[A, \nu_x, B]$ is driven by this source $[A, \mu]$. The supremum of these R for all ergodic sources $[A, \mu]$ is defined to be the ergodic transmission capacity of the channel $[A, \nu_x, B]$.

Now we shall call an cylinder set $[a_1, a_2, \cdots, a_n; 1, 2, \cdots, n] \in F_A$ a u-set, an cylinder set $[b_{m+1}, b_{m+2}, \cdots, b_n; m+1, m+2, \cdots, n] \in F_B$ a v-set, and do sum of some v-sets a V-set, where n is a positive integer greater than m. Suppose that there exist u-sets $u_1, u_2, \cdots, u_N$ and V-sets $V_1, V_2, \cdots, V_N$ such that

i) $V_1, V_2, \cdots, V_N$ are pairwise disjoint,

ii) $\nu_{u_i}(V_i)>1-\varepsilon$, $i=1, 2, \cdots, N$,

where ε is a small positive number, and suppose that N is taken as large as possible. Let us choose a message u among $u_1, u_2, \cdots$, and u_N with equal probabilities $1/N$, and transmitt it from the input of the channel, then the received message v will belong to either V_i and we

will be able to determine u excepting the case with probability smaller than ε. Thus we can obtain, by receiving n symbols through the channel, the amount of information which is equivalent to the selection of one thing from N things with equal probabilities, and hence which is equal to $\log N$, with a probability of error smaller than ε. On the other hand, denoting by C the ergodic transmission capacity of the channel, there exists an ergodic source $[A, \mu]$ such that $R>C-\varepsilon$. If the channel is driven by this source, the theoretical amount of information obtained by receiving n symbols is equal to $H(X_n)-H_{Y_n}(X_n)$ (see § 1.3), which is between $n(C-\varepsilon)$ and $n(C+\varepsilon)$ for sufficiently large n. Hence it is natural to expect that

$$\log N>n(C-\varepsilon)\,, \qquad \text{or} \qquad N>2^{n(C-\varepsilon)}\,.$$

In fact, this holds:

FEINSTEIN-HINČIN'S FUNDAMENTAL LEMMA. *Let $[A, \nu_x, B]$ be a channel satisfying the above conditions. Then, for each $\varepsilon>0$, there corresponds a positive integer $n(\varepsilon)$, in such a way that if $n\geqq n(\varepsilon)$ there exist a positive number N, u sets over $(1, 2, \cdots, n)$ $u_1, u_2, \cdots, u_N$ and V-sets over $(m+1, m+2, \cdots, n)$ $V_1, V_2, \cdots, V_N$ such that*

i) $V_1, V_2, \cdots, V_N$ *are pairwise disjoint*,

ii) $\nu_{u_i}(V_i)>1-\varepsilon$, $\quad i=1, 2, \cdots, N$,

iii) $N>2^{n(C-\varepsilon)}$.

PROOF. Take an ergodic source $[A, \mu]$ such that $R>C-\varepsilon/2$, where R is the speed of transmission when the channel $[A, \nu_x, B]$ is driven by the source $[A, \mu]$. Denote by $[A\times B, \omega]$ the product source of this $[A, \mu]$ and the channel $[A, \nu_x, B]$, and denote by $[B, \eta]$ the output source of $[A\times B, \omega]$. Then, by Theorem 2.1 the composite source $[A\times B, \omega]$ and hence the output source $[B, \eta]$ are ergodic. Let H_I, H_O and H_G be entropies per symbol of sources $[A, \mu]$, $[B, \eta]$ and $[A\times B, \omega]$, respectively. Let

$$\cdots, (x_{-1}, y_{-1}), (x_0, y_0), (x_1, y_1), \cdots$$

be the stochastic process representing the source $[A\times B, \omega]$, and put

$$u=(x_1, x_2, \cdots, x_n)\,,$$

$$v=(y_{m+1}, \cdots, y_n)\,,$$

where $n \geqq n(\varepsilon)$, and $n(\varepsilon)$ will be determined afterwards. u, v and $u \times v$ are considered as random cylinder sets in A^I, B^I and $(A \times B)^I$, respectively. By McMillan's theorem (Theorem 1.2)

$$-\frac{\log \mu(u)}{n}, \qquad -\frac{\log \eta(v)}{n} \qquad \text{and} \qquad -\frac{\log \omega(u \times v)}{n}$$

converge, respectively, to H_I, H_O, and H_O, in probability (ω)*. Hence,

$$\frac{1}{n} \log \frac{\nu_u(v)}{\eta(v)} = \frac{1}{n} \log \frac{\omega(u \times v)}{\mu(u)\eta(v)}$$

converges to $H_I + H_O - H_O = R > C - \frac{\varepsilon}{2}$ in probability. Therefore, if $n(\varepsilon)$ is sufficiently large, we have

$$(1) \qquad \omega\left(\frac{1}{n} \log \frac{\nu_u(v)}{\eta(v)} > C - \frac{\varepsilon}{2}\right) > 1 - \frac{\varepsilon}{2} .$$

Now, for each u-set u, let V_u be the sum of v-sets v such that the event in the blacket holds for (u, v). Then the left hand side of (1) is equal to

$$\omega(\bigcup_u (u \times V_u)) = \sum_u \omega(u \times V_u) = \sum_u \mu(u)\nu_u(V_u)$$

hence we have

$$(2) \qquad \sum_u \mu(u)\nu_u(V_u) > 1 - \frac{\varepsilon}{2} .$$

On the other hand, if (u, v) satisfies the event in the blacket in (1)

$$\nu_u(v) > 2^{n(\sigma - \frac{\varepsilon}{2})}\eta(v) .$$

Adding over v, we have

$$\nu_u(V_u) > 2^{n(\sigma - \frac{\varepsilon}{2})}\eta(V_u) .$$

As the left hand side does not exceed 1, it holds that

$$(3) \qquad \eta(V_u) < 2^{-n(\sigma - \frac{\varepsilon}{2})} .$$

Now, we shall prove that from (2) and (3) the lemma follows, if $n(\varepsilon)$ is so large that

---

* For the verification of the last, use the inequalities

$$\log \omega(u \times v') \leqq \log \omega(u \times v) \leqq \log \omega(u' \times v)$$

where $u' = (x_{m+1}, x_{m+2}, \cdots, x_n)$ and $v' = (y_1, y_2, \cdots, y_n)$.

(4) $$n(\varepsilon)\geqq\frac{2}{\varepsilon}\log\frac{2}{\varepsilon}\,.$$

Let u_1 be a u-set such that

$$\nu_u(V_u)>1-\varepsilon\,.$$

This is possible by (2). Put $V_1=V_{u_1}$. Next, let u_2 be a u-set such that

$$\nu_u(V_u-V_1)>1-\varepsilon\,,$$

if such a u exists, and put $V_2=V_{u_2}-V_1$. Next, let u_3 be a u-set such that

$$\nu_u(V_u-V_1-V_2)>1-\varepsilon\,,$$

if such a u exists, and put $V_3=V_{u_3}-V_1-V_2$. This process ends at finite steps. Let

(5) $$u_1,\ u_2,\ \cdots,\ u_N \qquad \text{and} \qquad V_1,\ V_2,\ \cdots,\ V_N$$

be the sequences obtained in this way. From the construction, (5) satisfies i) and ii), and for all u

(6) $$\nu_u(V_u-\bigcup_i V_i)\leqq 1-\varepsilon\,.$$

From this it follows that

$$\nu_u(V_u)\leqq\nu_u(\bigcup V_i)+\nu_u(V_u-\bigcup V_i)\leqq\nu_u(\bigcup V_i)+1-\varepsilon\,.$$

Averaging over u, we have

$$\sum_u \mu(u)\nu_u(V_u)\leqq\eta(\bigcup V_i)+1-\varepsilon\,,$$

which implies with (2) that

(7) $$\eta(\bigcup V_i)>\frac{\varepsilon}{2}\,.$$

On the other hand we have

$$V_i\subset V_{u_i}\,.$$

Hence, from (3) it follows that

$$\eta(\bigcup_{i=1}^{N} V_i)\leqq\sum_{i=1}^{N}\eta(V_{u_i})<N\cdot 2^{-n(\sigma-\frac{\varepsilon}{2})}\,,$$

which implies with (7) that

(8) $$N>\frac{\varepsilon}{2}\cdot 2^{n(\sigma-\frac{\varepsilon}{2})}\,.$$

iii) follows from (8) and (4). This completes the proof.

In the next section we shall use the following

COROLLARY. *Suppose that there is given a stationary m_1-step dependent channel with finite memory m, no foresight, and ergodic transmition capacity C. Then for each positive integer $n>m$, there exist a positive integer $N=N_n$, positive number ε_n, u-sets over* $(1, 2, \cdots, n)$ $u_1, u_2, \cdots, u_N$, *and V-sets over* $(m+1, m+2, \cdots, n)$ $V_1, V_2, \cdots, V_N$ *such that*

i) $V_1, V_2, \cdots, V_N$ *are pairwise disjoint*,

ii) $\nu_{u_i}(V_i)>1-\varepsilon_n$, $i=1, 2, \cdots, N$,

iii) $N>2^{n(C-\varepsilon_n)}$,

iv) $\lim_{n\to\infty}\varepsilon_n=0$.

PROOF. Let $[A, \mu]$ be a stationary ergodic source, let $[A\times B, \omega]$ be the composite source of the source $[A, \mu]$ and the channel $[A, \nu_x, B]$, and let $[B, \eta]$ be the output source of $[A\times B, \omega]$. Of source, ω and η depend on μ. Take a small positive number e, and put

$$\varepsilon_n=\max\{(1+e)\varepsilon_n', \varepsilon_n''\}$$

where

$$\varepsilon_n'=\inf\left\{\varepsilon;\ \sup_{\mu}\omega\left(\frac{1}{n}\log\frac{\nu_u(v)}{\eta(v)}>C-\frac{\varepsilon}{2}\right)>1-\frac{\varepsilon}{2}\right\},$$

and ε_n'' is the root of the equation

$$\frac{2}{\varepsilon}\log\frac{2}{\varepsilon}=n.$$

Then we have iv), and

$$\sup_{\mu}\omega\left(\frac{1}{n}\log\frac{\nu_u(v)}{\eta(v)}>C-\frac{\varepsilon_n}{2}\right)>1-\frac{\varepsilon_n}{2}, \tag{9}$$

$$\frac{2}{\varepsilon_n}\log\frac{2}{\varepsilon_n}\leqq n. \tag{10}$$

From (9) there exists $\mu=\mu_n$ such that

$$\omega\left(\frac{1}{n}\log\frac{\nu_u(v)}{\eta(v)}>C-\frac{\varepsilon_n}{2}\right)>1-\frac{\varepsilon_n}{2}. \tag{11}$$

From (10) and (11) it follows that there exist $u_1, \cdots, u_N$, $V_1, \cdots, V_N$

such that i), ii) and iii) hold as is proved in the above lemma.

§ 4. Shannon's fundamental theorems

Suppose we are given a stationary ergodic source $[A_0, \mu]$ with entropy H per symbol and a stationary m_1-step dependent channel $[A, \nu_x, B]$ with finite memory m, no foresight, and ergodic transmission capacity C, and suppose that A_0 and A do not coincide. Then we have to insert an encoder, that is, a one-valued transformation φ from A_0^I into A^I. A message $\theta \in A_0^I$ is coded into a message $\varphi(\theta) \in A^I$, and yields a stochastic process $[B, \nu_{\varphi(\theta)}]$. Thus we have a new channel $[A_0, \nu_{\varphi(\theta)}, B]$, which will be called the composite channel of the encoder φ and the channel $[A, \nu_x, B]$. Let n be a positive integer, and let φ_n be a one-valued transformation from A_0^n into A^n. For an element

$$\theta=(\cdots, \theta_{-1}, \theta_0, \theta_1, \theta_2, \cdots) \in A_0^I$$

define $\varphi(\theta) \in A^I$ by

$$\varphi(\theta)=(\cdots, x_{-1}, x_0, x_1, x_2, \cdots)$$

where

$$(x_{(t-1)n+1}, x_{(t-1)n+2}, \cdots, x_{tn})=\varphi_n(\theta_{(t-1)n+1}, \theta_{(t-1)n+2}, \cdots, \theta_{tn}) ,$$
$$t=\cdots, -1, 0, 1, \cdots$$

Then φ is an encoder from A_0^I into A^I. In this section, we shall consider only such encoders with $n>m$. Let $[A_0 \times B, \omega]$ be the composite source of the given source $[A_0, \mu]$ and the composite channel $[A_0, \nu_{\varphi(\theta)}, B]$, where φ is generated by φ_n, and let $[B, \eta]$ be the output source of $[A_0 \times B, \omega]$. Let $\alpha_1, \alpha_2, \cdots, \alpha_{a^n}$ be the cylinder sets of the form:

$$\alpha=[a_1, a_2, \cdots, a_n; 1, 2, \cdots, n] , \qquad a_i \in A_0 ,$$

and let $\beta_1, \beta_2, \cdots, \beta_{b^{n-m}}$ be the cylinder sets of the form

$$\beta=[b_{m+1}, b_{m+2}, \cdots, b_n; m+1, m+2, \cdots, n] , \qquad b_i \in B ,$$

where a and b are the numbers of letters of A_0 and B respectively. Suppose $\mu(\alpha_1) \geqq \mu(\alpha_2) \geqq \cdots$ for the later discussion. For each $k=1, 2, \cdots, b^{n-m}$ define i_k by

$$\omega(\alpha_{i_k} \times \beta_k)=\max_{1 \leqq i \leqq a^n} \omega(\alpha_i \times \beta_k) ,$$

and put

$$M=\bigcup_k \alpha_{i_k}\times\beta_k .$$

Now, the famous Shannon's fundamental theorem can be formulated as follows.

THEOREM 4.1. (Shannon's fundamental theorem). *If $H<C$ and if $\varepsilon>0$, then, for sufficiently large n, the coding φ_n can be done in such a way that*

$$\omega(M)>1-\varepsilon .$$

INTERPRETATION: Let α be a sequence of n letters of A_0. α is coded into $\varphi_n(\alpha)$, a sequence of n letters of A. This $\varphi_n(\alpha)$ is transmitted from the input through the channel, and then will be received a sequence of n letters of the output alphabet B, the sequence constructed by the last $n-m$ of which will be called β. If α is considered as a cylinder set belonging to the class $\{\alpha_1, \alpha_2, \cdots, \alpha_{a^n}\}$, then β must be considered as a cylinder set belonging to the class $\{\beta_1, \beta_2, \cdots, \beta_{b^{n-m}}\}$. If α is chosen in accordance with source probabilities μ, then we have

$$\Pr(\alpha=\alpha_i,\ \beta=\beta_k)=\omega(\alpha_i\times\beta_k) .$$

The above theorem says that if α is identified by α_{i_k} when $\beta=\beta_k$, then the probability of identification error is at most ε.

PROOF OF THEOREM 4.1. Let

$$\cdots, \theta_{-1}, \theta_0, \theta_1, \cdots$$

be the stochastic process representing the source $[A_0, \mu]$, and put

$$\alpha=(\theta_1, \theta_2, \cdots, \theta_n) .$$

Then by McMillan's theorem $-\dfrac{\log\mu(\alpha)}{n}$ converges to entropy H per symbol of $[A_0, \mu]$ in probability. Hence for sufficiently large n, we have

$$\mu\left(\frac{\log\mu(\alpha)}{n}+H\geqq-\varepsilon\right)\geqq 1-\varepsilon .$$

Let $\alpha_1, \alpha_2, \cdots, \alpha_N$ be the values of α which satisfy the event in the above blacket. Then, we have

$$(1)\qquad \sum_1^N \mu(\alpha_i)\geqq 1-\varepsilon ,$$

$$(2)\qquad \mu(\alpha_i)\geqq 2^{-n(H+\varepsilon)} ,\qquad i=1, 2, \cdots, N .$$

From (2) it follows that

$$N \leqq 2^{n(H+\varepsilon)} . \tag{3}$$

On the other hand, by the application of Feinstein-Hinčin's fundamental lemma to the given channel $[A, \nu_x, B]$, we have, for sufficiently large n, u-sets over $(1, 2, \cdots, n)$ $u_1, u_2, \cdots, u_{N'}$ and V-sets over $(m+1, \cdots, n)$ $B_1, B_2, \cdots, B_{N'}$ such that

$$B_1, B_2, \cdots, B_{N'} \quad \text{are pairwise disjoint} , \tag{4}$$

$$\nu_{u_i}(B_i) > 1-\varepsilon , \qquad 1 \leqq i \leqq N' , \tag{5}$$

$$N' > 2^{n(C-\varepsilon)} . \tag{6}$$

Suppose that

$$2\varepsilon < C-H$$

without loss of generality. Then we have by (3) and (6)

$$N \leqq 2^{n(H+\varepsilon)} < 2^{n(C-\varepsilon)} < N' ,$$

and we can define φ_n by

$$\varphi_n(\alpha_i) = u_i , \qquad \text{for} \quad i=1, 2, \cdots, N ,$$

$$\varphi_n(\alpha_i) = u_{N+1} , \qquad \text{for} \quad i=N+1, N+2, \cdots, a^n .$$

Now, by definition,

$$\omega(\alpha_i \times \beta_k) = \int_{\alpha_i} \nu_{\varphi(\theta)}(\beta_k) d\mu(\theta) .$$

Let $i \leqq N$. If θ varies in α_i, then $\varphi(\theta) \in u_i$, and $\nu_{\varphi(\theta)}(\beta_k) = \nu_{u_i}(\beta_k)$. Hence we have

$$\omega(\alpha_i \times \beta_k) = \mu(\alpha_i)\nu_{u_i}(\beta_k) , \qquad i=1, 2, \cdots, N .$$

Adding over $\beta_k \subset B_i$, we have, noting (5)

$$\omega(\alpha_i \times B_i) = \mu(\alpha_i)\nu_{u_i}(B_i) > (1-\varepsilon)\mu(\alpha_i) .$$

Adding over $i=1, 2, \cdots, N$, we have, making use of (1)

$$\sum_{i=1}^{N} \omega(\alpha_i \times B_i) > (1-\varepsilon) \sum_{1}^{N} \mu(\alpha_i) \geqq (1-\varepsilon)^2 > 1-2\varepsilon . \tag{7}$$

Now,

$$\begin{aligned} \omega(M) = \sum_k \omega(\alpha_{i_k} \times \beta_k) &\geqq \sum_{i=1}^{N} \sum_{\beta_k \subset B_i} \omega(\alpha_{i_k} \times \beta_k) \\ &\geqq \sum_{i=1}^{N} \sum_{\beta_k \subset B_i} \omega(\alpha_i \times \beta_k) = \sum_{i=1}^{N} \omega(\alpha_i \times B_i) . \end{aligned} \tag{8}$$

From (7) and (8) it follows that

$$\omega(M)>1-2\varepsilon\ .$$

This completes the proof.

Notice: We may replace M by M' defined as follows: For each $k=1, 2, \cdots, b^{n-m}$ define i'_k by

$$\omega(\alpha_{i'_k}\times\beta_k)=\max_{1\leq i\leq N}\omega(\alpha_i\times\beta_k)$$

and put

$$M'=\bigcup_k \alpha_{i'_k}\times\beta_k\ .$$

Let

$$\cdots, (x_{-1}, y_{-1}),\ (x_0, y_0),\ (x_1, y_1), \cdots$$

be the stochastic process representing the composite source $[A_0\times B, \omega]$. Put

$$X_s=(x_1, x_2, \cdots, x_s)\ , \qquad \text{and} \qquad Y_s=(y_1, y_2, \cdots, y_s)\ .$$

Hinčin has considered the asymptotic behaviour of the ratio

$$\frac{H(X_s)-H_{Y_s}(X_s)}{s}$$

when $s\to\infty$, on the fixed encoder φ_n. However, this seems meaningless at least to me, from the practical point of view, because identification is carried out whenever a sequence of n letters of the output alphabet is received, and the theoretical amount of information $H(X_s)-H_{Y_s}(X_s)$ is not used by our method of decoding. We shall reformulate the second theorem of Shannon given by Hinčin [3], § 16 as follows.

THEOREM 4.2. *Suppose that $H<C$. Then, for each sufficiently large n, there corresponds a coding φ_n from A_0^n to A^n in such a way that*

i) $$\lim_{n\to\infty}\omega(M)=1\ ,$$

ii) $$\lim_{n\to\infty}\frac{H(\alpha)-H_\beta(\alpha)}{n}=H\ .$$

Theorem 4.1 is a direct corollary to this theorem.

PROOF. For each n, define δ_n by

$$\delta_n=\inf\left\{\varepsilon\,;\,\mu\left(\frac{\log\mu(\alpha)}{n}+H\geqq-\varepsilon\right)\geqq 1-\varepsilon\right\}\,.$$

Then, we have

$$\mu\left(\frac{\log\mu(\alpha)}{n}+H\geqq-\delta_n\right)\geqq 1-\delta_n\,, \tag{9}$$

$$\lim_{n\to\infty}\delta_n=0\,, \tag{10}$$

the latter of which follows from the convergence of $\dfrac{\log\mu(\alpha)}{n}+H$ to 0 in probability. Fix n. Let $\alpha_1, \alpha_2, \cdots, \alpha_N$ be the values of α which statisfy the event in the blacket of (9). Then we have

$$\sum_1^N \mu(\alpha_i)\geqq 1-\delta_n\,, \tag{11}$$

$$N\leqq 2^{n(H+\delta_n)}\,. \tag{12}$$

On the other hand, by applying Corollary to Feinstein-Hinčin's fundamental lemma to the given channel $[A, \nu_x, B]$, we have, for each $n>m$, u-sets over $(1, 2, \cdots, n)$ $u_1, u_2, \cdots, u_{N'}$ and V-sets over $(m+1, \cdots, n)$ $B_1, B_2, \cdots, B_{N'}$ such that

$$B_1, B_2, \cdots, B_{N'} \quad \text{are pairwise disjoint}\,, \tag{13}$$

$$\nu_{u_i}(B_i)>1-\varepsilon_n\,, \qquad 1\leqq i\leqq N'\,, \tag{14}$$

$$N'>2^{n(C-\varepsilon_n)}\,, \tag{15}$$

$$\lim_{n\to\infty}\varepsilon_n=0\,. \tag{16}$$

From (10) and (16) we have

$$\lim(\varepsilon_n+\delta_n)=0 \tag{17}$$

By the hypothesis $H<C$, there exists a positive integer n_0, such that for all $n\geqq n_0$

$$\varepsilon_n+\delta_n<C-H\,.$$

For each $n\geqq n_0$, we have by (12) and (15)

$$N<N'\,,$$

and we can define φ_n by

$$\varphi_n(\alpha_i)=u_i\,, \qquad \text{for} \quad i=1, 2, \cdots, N\,,$$

$$\varphi_n(\alpha_i)=u_{N+1}\,, \qquad \text{for} \quad i=N+1, N+2, \cdots, a^n\,.$$

If ω and M are defined by using this coding φ_n, then we can prove that

$$\omega(M)>1-\varepsilon_n-\delta_n \tag{18}$$

in the same way as in the proof of Theorem 4.1. This with (17) proves i).

Now, let us turn to the proof of ii). As

$$\lim_{n\to\infty}\frac{H(\alpha)}{n}=H$$

by the definition of H, it is sufficient to prove that

$$\lim_{n\to\infty}\frac{H_\beta(\alpha)}{n}=0\ . \tag{19}$$

To prove this, we shall use the following

LEMMA 4.1. *For arbitrary non-negative numbers* $x_1, x_2, \cdots, x_n$,

$$\sum_{i=1}^{n} x_i \log x_i \geqq \sum_{i=1}^{n} x_i \log\left(\sum_{i=1}^{n} x_i \Big/ n\right).$$

This follows from the convexity of the function $f(x)=x\log x$.

Take any β_k with $\gamma(\beta_k)>0$. Denote by $\omega(\alpha_i|\beta_k)$ the conditional probability that $\alpha=\alpha_i$ if $\beta=\beta_k$ and denote by $H(\alpha|\beta=\beta_k)$ the entropy of the conditional distribution of α if $\beta=\beta_k$. Then we have

$$\begin{aligned}H(\alpha|\beta=\beta_k)&=-\sum_i \omega(\alpha_i|\beta_k)\log\omega(\alpha_i|\beta_k)\\&=-\omega(\alpha_{i_k}|\beta_k)\log\omega(\alpha_{i_k}|\beta_k)-\sum_{i\neq i_k}\omega(\alpha_i|\beta_k)\log\omega(\alpha_{i_k}|\beta_k)\ .\end{aligned}$$

Put $p_k=\omega(\alpha_{i_k}|\beta_k)$. Then by making use of Lemma 4.1 twice we have

$$\begin{aligned}H(\alpha|\beta=\beta_k)&\leqq -p_k\log p_k-(1-p_k)\log\{(1-p_k)/(a^n-1)\}\\&=-p_k\log p_k-(1-p_k)\log(1-p_k)+(1-p_k)\log(a^n-1)\\&\leqq 1+(1-p_k)\log(a^n-1)\ .\end{aligned}$$

Averaging over β_k we have

$$\begin{aligned}H_\beta(\alpha)&\leqq 1+\{1-\omega(M)\}\log(a^n-1)\\&\leqq 1+(\varepsilon_n+\delta_n)n\log a\ .\end{aligned}$$

The last inequality follows from (18). Finally, we have

$$0\leqq\frac{H_\beta(\alpha)}{n}<(\varepsilon_n+\delta_n)\log a+\frac{1}{n}$$

which yields (19), by (17). This completes the proof.

§ 5. Case of a source without letter-durations

Let $[A_0, \mu]$ be a stationary ergodic source with entropy H per symbol and let $[A, \nu_x, B]$ be a stationary m_1-step dependent channel with finite memory m, no foresight, and ergodic transmission capacity C. In the previous section, any sequence of n letters of A_0 was coded into a sequence of n letters of A, and the length of a sequence was not altered by encoding. This means that it is supposed implicity that each letter of the source $[A_0, \mu]$ has the same duration as each letter of the imput alphabet of the channel. However, since our definition of an information source does not contain the concept of durations of letters, it is possible to consider encoders of the following type. Let r and s be positive integers, and let $\varphi_{r,s}$ be a one-valued transformation from A_0^r into A^s. For an element

$$\theta=(\cdots, \theta_{-1}, \theta_0, \theta_1, \theta_2, \cdots) \in A_0^I$$

define $\varphi(\theta) \in A^I$ by

$$\varphi(\theta)=(\cdots, x_{-1}, x_0, x_1, x_2, \cdots) ,$$

where

$$(x_{(t-1)s+1}, x_{(t-1)s+2}, \cdots, x_{ts})=\varphi_{r,s}(\theta_{(t-1)r+1}, \theta_{(t-1)r+2}, \cdots, \theta_{tr})$$
$$t=\cdots, -1, 0, 1, \cdots$$

Then φ is an encoder from A_0^I into A^I. In this section, we shall consider such encoders with $s>m$. Let $[A_0 \times B, \omega]$ be the composite source of the given source $[A_0, \mu]$ and the composite channel $[A_0, \nu_{\varphi(\theta)}, B]$, where φ is generated by $\varphi_{r,s}$. Let $\alpha_1, \alpha_2, \cdots, \alpha_{a^r}$ be the cylinder sets over $(1, 2, \cdots, r)$ in A_0^I, and let $\beta_1, \beta_2, \cdots \beta_{b^{s-m}}$ be the cylinder sets over $(m+1, m+2, \cdots, s)$ in B^I, where a and b are the numbers of letters of A_0 and B, respectively. Suppose that $\mu(\alpha_1) \geqq \mu(\alpha_2) \geqq \cdots$ for the later discussion. For each $k=1, 2, \cdots, b^{s-m}$ define i_k by

$$\omega(\alpha_{i_k} \times \beta_k)=\max_{1 \leqq i \leqq a^r} \omega(\alpha_i \times \beta_k) ,$$

and put

$$M=\bigcup_k \alpha_{i_k} \times \beta_k .$$

Next, let

$$\cdots, (x_{-1}, y_{-1}), (x_0, y_0), (x_1, y_1), \cdots$$

be the stochastic process representing the composite source $[A_0 \times B, \omega]$, and put

$$\alpha=(x_1, x_2, \cdots, x_r) \quad \text{and} \quad \beta=(y_{m+1}, y_{m+2}, \cdots, y_s) .$$

Then we have the following

THEOREM 5.1. *Suppose that $H>0$ and $C>0$. Then for each positive integer r, there correspond a positive integer s and a coding $\varphi_{r,s}$ from A_0^r into A^s in snch a way that*

i) $$\lim_{r\to\infty} \omega(M)=1 ,$$

ii) $$\lim_{r\to\infty} \frac{H(\alpha)-H_\beta(\alpha)}{s}=C.$$

PROOF. As our source $[A_0, \mu]$ has entropy H per symbol, for each r there correspond $N=N_r$ and δ_r such that

(1) $$\sum_{i=1}^{N} \mu(\alpha_i)\geqq 1-\delta_r ,$$

(2) $$N\leqq 2^{r(H+\delta_r)} ,$$

(3) $$\lim_{r\to\infty} \delta_r=0$$

(See the proof of Theorem 4.2.) Next, as our channel $[A, \nu_x, B]$ has ergodic transmission capacity C, for each $s>m$, there correspond $N'=N'_s$, ε_s, u-sets over $(1, 2, \cdots, s)$ $u_1, u_2, \cdots, u_{N'}$ and V-sets over $(m+1, m+2, \cdots, s)$ $B_1, B_2, \cdots, B_{N'}$ such that

(4) $B_1, B_2, \cdots, B_{N'}$ are pairwise disjoint ,

(5) $$\nu_{u_i}(B_i)>1-\varepsilon_s , \qquad 1\leqq i\leqq N' ,$$

(6) $$N'>2^{s(C-\varepsilon_s)} ,$$

(7) $$\lim_{s\to\infty} \varepsilon_s=0 .$$

(Corollary to Feinstein-Hinčin's lemma). Now, for each r define $s=s(r)$ by the smallest positive integer s such that

(8) $$s(C-\varepsilon_s)>r(H+\delta_r) , \qquad s>m .$$

Then for sufficiently large r we have

$$s(C-\varepsilon_s)>r(H+\delta_r)\geqq(s-1)(C-\varepsilon_{s-1}) ,$$

from which it follows that

$$\lim_{r\to\infty}\frac{s}{r}=\frac{H}{C} . \tag{9}$$

By (2), (6) and (8) we have

$$N<N'$$

and we can define $\varphi_{r,s}$ as follows:

$$\varphi_{r,s}(\alpha_i)=u_i , \qquad 1\leqq i\leqq N ,$$

$$\varphi_{r,s}(\alpha_i)=u_{N+1} , \qquad N+1\leqq i\leqq a^r .$$

Then Theorem 5.1 follows from the above in the same way as in the proof of Theorem 4.2.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] A. Feinstein, A new basic theorem of information theory, *Trans. I.R.E. Professional Group on Information theory*, No. PGIT-4 (1954), 2-22.

[2] A. Ya. Hinčin, The concept of entropy in the theory of probability, *Uspehi Mat. Nauk* (N. S.) Vol. 8 (1953), No. 3 (55), 3-20, (Russian).

[3] ———, On the basic theorems of information theory, *Uspehi Mat. Nauk* (N. S.) Vol. 11 (1956), No. 1 (67), 17-75, (Russian).

[4] B. McMillan, The basic theorems of information theory, *Ann, Math. Stat.* Vol. 24 (1953), No. 2, 196-219.

[5] C. E. Shannon, A mathematical theory of communication, *Bell System Tech. Journ.* Vol. 27 (1948), 379-423, 623-656.