PROOF OF SHANNON'S TRANSMISSION THEOREM FOR FINITE-STATE INDECOMPOSABLE CHANNELS[1]

By David Blackwell, Leo Breiman, and A. J. Thomasian

University of California, Berkeley

1. Summary. For finite-state indecomposable channels, Shannon's basic theorem, that transmission is possible at any rate less than channel capacity but not at any greater rate, is proved. A necessary and sufficient condition for indecomposability, from which it follows that every channel with finite memory is indecomposable, is given. An important tool is a modification, for some processes which are not quite stationary, of theorems of McMillan and Breiman on probabilities of long sequences in ergodic processes.

2. Notation, definitions. For any positive integer N, we denote by $I(N)$ the set of integers $1, 2, \cdots, N$ and for any set S we denote by $S^{(N)}$ the set of N-tuples $(s_1, \cdots, s_N)$ with $s_i \varepsilon S$, $i \varepsilon I(N)$.

Let A be a fixed positive integer. A *source* is a pair (M, ϕ), where M is a finite, say $D \times D$, indecomposable Markov matrix and ϕ is a function from $I(D)$ to $I(A)$. A *channel* is a sequence of A Markov matrices $C(1), \cdots, C(A)$ of the same size, say $R \times R$, and a function ψ from $I(R)$ to $I(B)$, where B is some positive integer.

The elements of $I(D)$ and $I(R)$ will be considered as states of the source and channel respectively. The source will be considered as driving the channel as follows. If $d \varepsilon I(D)$, $r \varepsilon I(R)$ are the states of the source and channel at the beginning of a cycle, the source moves from d to a state $e \varepsilon I(D)$, selected according to the Markov transition matrix M, so that $M(d, e)$ is the probability that the new state is e, given that the initial state is d. The source then emits the number $\phi(e) \varepsilon I(A)$, which is fed into the channel. The channel then moves into a state $s \varepsilon I(R)$, selected according to the matrix $C(\phi(e))$, and emits the number $\psi(s)$, completing the cycle. A new cycle then begins, with e, s as the initial states of the source and channel. The joint motion of the source and channel is thus described by the *source-channel matrix*, which is a $DR \times DR$ Markov matrix L, with elements $L((d, r), (e, s)) = M(d, e)C(\phi(e), r, s)$. A channel will be called *indecomposable* if for every source the source-channel matrix L is indecomposable. Thus, for any source and any indecomposable channel, there is a sequence of random variables $\{(d_n, r_n), -\infty < n < \infty\}$, which is an ergodic Markov process with transition matrix L. Moreover, the joint distribution of $\{(d_n, r_n)\}$ depends only on L. McMillan [4], extending the work of Shannon [5], has shown that associated with any stationary ergodic process $\{z_k\}$ with a finite set F of

Received January 6, 1958; revised May 22, 1958.

[1] This research was supported in part by the Office of Naval Research under Contract Nonr-222(53) and in part by a research grant (No. G-3666) from the National Institutes of Health, Public Health Service.

states, is a number h, called the *entropy* of the process, such that for large N it is practically certain that the sequence of states of length N which occurs is one whose probability is about 2^{-Nh}; more precisely, for any sequence $f \varepsilon F^{(N)}$ let

$$Q_N(f) = \text{Prob } \{(z_1, \cdots, z_N) = f\}.$$

Then

$$N^{-1} \log Q_N(z_1, \cdots, z_N) \to -h \text{ in } L_1 \text{ as } N \to \infty, \tag{1}$$

where the log above and throughout this paper has base 2. Breiman [1] has shown that convergence with probability 1 also occurs in (1). For the ergodic process $\{(d_k, r_k)\}$, the processes $\{x_k = \phi(d_k)\}$, $\{y_k = \psi(r_k)\}$, $\{(x_k, y_k)\}$ are of course also ergodic; we denote their entropies by $H(X)$, $H(Y)$, $H(X, Y)$ respectively.

For a fixed indecomposable channel, the upper bound H over all sources of the number $H(X) + H(Y) - H(X, Y)$ is called, following Shannon, the *capacity* of the channel. Shannon [5] and, subsequently, McMillan [4], Feinstein [2], Hincin [3], and Wolfowitz [6] have shown that, under various hypotheses on the channel, it is possible to transmit over the channel at any rate less than its capacity, but not at any rate greater than its capacity. For a channel as defined above, this means, as in [6], the following. For a given channel, to say that it is possible to transmit at rate G means that for every $\epsilon > 0$ there is an N_0 such that for any $N \geqq N_0$ there are $2^{NG} = J$ distinct sequences $u_1, \cdots, u_J$, where each $u_j \varepsilon I(A)^{(N)}$, and J disjoint subsets $E_1, \cdots, E_J$ of $I(B)^N$ such that

$$Q(r, u_j, E_j) > 1 - \epsilon \text{ for all } j \text{ and all } r \varepsilon I(R), \tag{2}$$

where for any $r \varepsilon I(R)$, $u = (u(1), \cdots, u(N)) \varepsilon I(A)^{(N)}$, $E \subset I(B)^{(N)}$

$$Q(r, u, E) = \sum C(u(1), r, r_1) \cdots C(u(N), r_{N-1}, r_N),$$

where the sum is over those sequences $(r_1, \cdots, r_N)$ for which

$$(\psi(r_1), \cdots, \psi(r_N)) \varepsilon E.$$

Thus $Q(r, u, E)$ is the probability that the output sequence from the channel is an element of E, when the channel is initially in state r and u is the input sequence.

For a given channel, denote by H^* the upper bound of the rates G at which it is possible to transmit. We shall show that, for indecomposable channels of the type considered here, $H^* = H$, that is, it is possible to transmit at any rate less than the channel capacity, but not at a rate greater than channel capacity. Shannon and McMillan seem to have regarded $H^* \leqq H$ as more or less obvious, and devoted most of their attention to showing, under certain hypotheses, that $H \leqq H^*$. The other writers have given some attention to the inequality $H^* \leqq H$. In particular, Wolfowitz [6] obtained $H^* \leqq H$ for channels of zero memory. Our result, that $H^* = H$ for indecomposable channels, extends those obtained previously.

3. A necessary and sufficient condition for indecomposability. To verify that the results to be proved in Sections 5 and 6 are valid for a given channel, we must show that the channel is indecomposable. The following criterion is helpful.

THEOREM 1. *A channel* $(C(1), \cdots, C(A))$ *is indecomposable if and only if every finite product* $C(a_1) \cdots C(a_k)$ *is an indecomposable Markov matrix,* $k = 1, 2, \cdots$, $a_i \varepsilon I(A)$.

PROOF. Suppose the channel is indecomposable and let $a_1, \cdots, a_k$ be any finite sequence of elements of $I(A)$. Consider the source with k states $1, \cdots, k$ with $M(i, i+1) = 1$ for $i < k$, $M(k, 1) = 1$, and $\phi(i) = a_i$. Let

$$F = C(a_1) \cdots C(a_k)$$

and let $r_1, r_2 \varepsilon I(R)$. To show that F is indecomposable it is sufficient to find integers T_1, T_2 and a state $r_3 \varepsilon I(R)$ such that $F^{T_1}(r_1, r_3) > 0$ and $F^{T_2}(r_2, r_3) > 0$, that is, such that r_3 is reachable from either r_1 or r_2 under transition matrix F. Since the source-channel matrix L is indecomposable, the two states (k, r_1), (k, r_2) have a common possible successor (i, r) which itself has a possible successor of the form (k, r_3). Thus (k, r_3) is a possible successor of either (k, r_1) or (k, r_2). Since the source has period k, the times after which (k, r_3) can be reached from (k, r_1) or (k, r_2) are multiples of k, that is, there are integers T_1, T_2 such that $L^{T_i k}((k, r_i), (k, r_3)) > 0$ for $i = 1, 2$. But $L^{Tk}((k, r), (k, s)) = F^T(r, s)$. Consequently $F^{T_i}(r_i, r_3) > 0$ for $i = 1, 2$ and F is indecomposable.

Now suppose that every finite product $C(a_1) \cdots C(a_k)$ is indecomposable, and let (M, ϕ) be any source. Let (d, r), (e, s) be any two source-channel states; we must find a common possible successor (f, t). Since M is indecomposable, d and e have a common possible successor f which is recurrent. There are then numbers r', s', such that (f, r') is a successor of (d, r) and (f, s') is a successor of (e, s), so that any common successor of (f, r') and (f, s') is also a common successor of (d, r) and (e, s). Thus we may suppose $d = e = f$, and must find a common successor (f, t) of (f, r'), (f, s'), where f is recurrent. Let $f_0 = f, f_1, \cdots, f_{k-1}, f_k = f$ be a possible path from f to itself, and let $F = C(\phi(f_1)) \cdots C(\phi(f_k))$. We assert that if t is a possible successor of r' with respect to F, then (f, t) is a possible successor of (f, r') in the source-channel matrix L. For $L^{Tk}((f, r'), (f, t)) \geqq [M(f_0, f_1) \cdots M(f_{k-1}, f_k)]^T F^T(r', t)$, and since the first factor on the right is positive, the left side is positive whenever $F^T(r', t)$ is. But since F is recurrent, r' and s' have a common possible successor t with respect to F, so that (f, t) is a common possible successor of (f, r'), (f, s) in L, completing the proof.

We shall say that a channel has memory m if every product $C(a_0) \cdots C(a_m)$ has identical rows. Thus a channel has memory m if and only if the conditional distribution of the present state of the channel, given the present input a_m, the m previous inputs $a_0, \cdots, a_{m-1}$ and the state r of the channel just prior to input a_0, is independent of r for every $a_0, \cdots, a_m$. A channel is said to have finite memory if for some m it has memory m. Every channel with finite memory is clearly indecomposable, for if $F = C(a_1) \cdots C(a_k)$, some power of F has

identical rows so that F is indecomposable. From Theorem 1, the channel is then indecomposable. That this includes, as a special case, the finite memory channels as defined by Feinstein [2] and Wolfowitz [6] can be seen from the following considerations: let the inputs to a channel be denoted by $\cdots, X_{-1}, X_0, X_1, \cdots$ and the outputs by $\cdots, Y_{-1}, Y_0, Y_1, \cdots$ and let the probability structure at the channel be defined, following McMillan [4], by specifying the conditional probabilities of the various output messages, given the input signals. That is, we are given the conditional probabilities $p(Y_n, \cdots, Y_k \mid X_n, X_{n-1}, \cdots)$ where we are now assuming that the channel is nonanticipatory and stationary. We assume, in addition, that there is an integer m such that

$$p(Y_n \mid X_n, Y_{n-1}, X_{n-1}, Y_{n-2}, X_{n-2}, \cdots) = p(Y_n \mid X_n, Y_{n-1}, X_{n-1}, \cdots, Y_{n-m}, X_{n-m}).$$

Now if we consider the finite state channel whose states consist of m-tuples of pairs, one member of the pairs being from the input alphabet and the other from the output alphabet, then the above assumption implies that this finite state channel is finitary in the sense described above, that is, it has the required Markov property. If we add the additional restriction that there is an integer M such that if two output messages m long, say y_1, y_2, are separated by a distance M, that

$$p(y_1, y_2 \mid \cdots X_1, X_0, X_{-1}, \cdots) = p(y_1 \mid \cdots, X_1, X_0, X_{-1}, \cdots)p(Y_2 \mid \cdots X_1, X_0, X_{-1})$$

then this finite state channel has finite memory M.

4. A modification of McMillan's theorem. In proving our main result, we shall need the following extension of a special case of McMillan's theorem.

THEOREM 2. *Let $d_1, d_2, \cdots$ be a Markov process with finite indecomposable transition matrix M, say $D \times D$, let ϕ be a function from $I(D)$ to $I(A)$, and let $y_n = \phi(d_n)$. For any sequence $s \in I(A)^{(N)}$ let $p(s) = P\{(y_1, \cdots, y_N) = s\}$, and let $z_N = p(y_1, \cdots, y_N)$. There is a constant h, depending only on M and ϕ, such that*

$$N^{-1} \log z_N \to -h \tag{3}$$

in L_1 and with probability 1 as $N \to \infty$.

PROOF. If the distribution of d_1 is the (unique) stationary distribution for M, the $\{y_n\}$ process is ergodic, and the theorems of McMillan [4] and Breiman [1] yield (3), with h as the entropy of the process.

For any $d \in I(D)$ and any event E, write $P_d(E)$ for $P(E \mid d_1 = d)$. Let $\lambda = (\lambda_1, \cdots, \lambda_D)$ be the stationary distribution for M, and let $Q(E) = \sum \lambda_d P_d(E)$. The theorems of McMillan and Breiman assert that

$$\frac{\log \sum_d \lambda_d z_{dN}}{N} \to -h \text{ a.e. and } L_1(Q), \tag{4}$$

where $p_d(s) = P_d\{(y_1, \cdots, y_N) = s\}$ and $z_{dN} = p_d(y_1, \cdots, y_N)$. For any d for which $\lambda_d > 0$, we have

$$\lambda_d z_{dN} = (\sum_e \lambda_e z_{eN})\, Q(d_1 = d \mid y_1, \cdots, y_N).$$

Taking logs, dividing by N, letting $N \to \infty$ and using (4) and the fact that $Q(d_1 = d \mid y_1, \cdots, y_N)$ converges a.e. (Q) to a limit which is positive a.e. (P_d) yields

$$\frac{\log z_{dN}}{N} \to -h \text{ a.e. } P_d \text{ for } \lambda_d > 0. \tag{5}$$

Now let d be a state for which $\lambda_d = 0$, let e be any state for which $\lambda_e > 0$, let k be any integer $\leqq N$ and let G denote the event $\{d_k = e\}$. We have

$$z_{dN}P_d(G \mid y_1, \cdots, y_N) = z_{dk}P_d(G \mid y_1, \cdots, y_k)p_e(y_k, \cdots, y_N). \tag{6}$$

Since the P_d conditional distribution of $y_k, y_{k+1}, \cdots$, given that G occurs, is the same as the unconditional P_e distribution of $y_1, y_2, \cdots$, we conclude from (5) that on G, a.e. P_d, $N^{-1} \log p_e(y_k, \cdots, y_N) \to -h$. Also, on G, a.e. P_d, $P_d(G \mid y_1, \cdots, y_N)$ has a positive limit as $N \to \infty$ and $z_{dk}P_d(G \mid y_1, \cdots, y_k)$ is positive. Taking logs in (6), dividing by N, letting $N \to \infty$ yields

$$N^{-1} \log z_{dN} \to -h \text{ a.e. } P_d \text{ on } G. \tag{7}$$

Since the union of the sets G obtained by varying k and e has P_d measure 1, we conclude

$$N^{-1} \log z_{dN} \to -h \text{ a.e. } P_d \text{ for all } d. \tag{8}$$

Next let $\mu = (\mu_1, \cdots, \mu_D)$ be any initial distribution and let $P = \sum \mu_d P_d$. For any d for which $\mu_d > 0$, we have

$$\mu_d z_{dN} = (\sum_d \mu_d z_{dN})\, P(d_1 = d \mid y_1, \cdots, y_N). \tag{9}$$

Taking logs, dividing by N, letting $N \to \infty$ and using (8) yields

$$N^{-1} \log (\sum \mu_d z_{dN}) \to -h \text{ a.e. } P_d, \tag{10}$$

from which we obtain

$$N^{-1} \log (\sum \mu_d z_{dN}) \to -h \text{ a.e. } P. \tag{11}$$

Thus the probability 1 convergence in (2) is established. Finally, to obtain L_1 convergence we note, following McMillan, that the sequence $\{N^{-1} \log z_N\}$ is uniformly integrable. We have

$$J(N,k) = \int_{B_k} | N^{-1} \log z_N | \, dP = N^{-1} \sum p(s) \, | \log p(s) | \tag{12}$$

$$\leqq (k+1)2^{-kN}A^N,$$

where B_k is the event $\{k \leqq |N^{-1} \log z_N| < k+1\}$ and the sum is extended over those $s \in I(A)^{(N)}$ for which $k \leqq |N^{-1} \log p(s)| < k+1$. Choose k_1 so that $2^{-k_1} A < 1$. For $k \geqq k_1$ we have

$$J(N, k) \leqq (k+1)2^{-k}2^{k_1}. \tag{13}$$

Thus $\sum_{k_0}^{\infty} J(N, k)$ goes to zero as $k_0 \to \infty$ uniformly in N, and uniform integrability is established, completing the proof.

5. The direct half of Shannon's theorem (possibility of transmission at every rate less than capacity). We shall need the following lemma.

LEMMA. *Let p be a probability distribution on a finite product space $X \times Y$. Write $a(x) = \sum_y p(x, y)$, $b(y) = \sum_x p(x, y)$, $p(y \mid x) = p(x, y) / a(x)$. For any numbers δ, λ such that $0 < \delta \leqq \lambda < 1$, let*

$$A = \{y : b(y) > \delta\}, \qquad B = \{(x, y) : p(y \mid x) < \lambda\}.$$

For any integer M there are M points $x_1, \cdots, x_M \in X$ and M disjoint subsets $E_1, \cdots, E_M$ of Y such that

$$\sum_{y \notin E_i} p(y \mid x_i) \leqq 4M(\delta/\lambda) + 2\sum_{y \in A} b(y) + 2\sum_{(x,y) \in B} p(x, y) \tag{14}$$

for $i = 1, \cdots, M$.

PROOF. Let $X_1, \cdots, X_{2M}$ be independent random variables with distribution $a(x)$. For each $i \in I(2M)$, $y \in Y$, we define the random variable $Z(i, y) = 1$ if $p(y \mid X_i) \leqq \max_{j \neq i} p(y \mid X_j)$, $Z(i, y) = 0$ otherwise, and define

$$f_i = \sum_y p(y \mid X_i) Z(i, y).$$

Then

$$\begin{aligned} Ef_i &= \sum_x a(x) E(f_i \mid X_i = x) = \sum_{x,y} p(x, y) E(Z(i, y) \mid X_i = x) \\ &\leqq \sum_{y \in A} b(y) + \sum_{(x,y) \in B} p(x, y) + \sum\nolimits^{*} p(x, y) E(Z(i, y) \mid X_i = x), \end{aligned} \tag{15}$$

where $\sum^*$ indicates summation over pairs (x, y) for which $b(y) \leqq \delta$ and $p(y \mid x) \geqq \lambda$. Now $E(Z(i, y) \mid X_i = x) = 1 - (1 - u(x, y))^{2M-1}$, where

$$u(x,y) = \sum_{v : p(y|v) \geqq p(y|x)} a(v).$$

For pairs (x, y) in $\sum^*$,

$$\delta \geqq b(y) = \sum_v a(v) p(y \mid v) \geqq \lambda \sum_{p(y|v) \geqq \lambda} a(v) \geqq \lambda u(x, y),$$

so that

$$E(Z(i, y) \mid X_i = x) \leqq 1 - (1 - (\delta/\lambda))^{2M-1} \leqq 2M\delta/\lambda.$$

Using this inequality in (15) yields

$$Ef_i \leqq \sum_{y \in A} b(y) + \sum_{x,y \in B} p(x,y) + 2M\delta/\lambda = \alpha. \tag{16}$$

It follows that $E(\sum_1^{2M} f_i / 2M) \leqq \alpha$. Thus there are values of $X_1, \cdots, X_{2M}$, say $x_1^*, \cdots, x_{2M}^*$, for which $\sum_1^{2M} f_i^* / 2M \leqq \alpha$, where $f_i^* = f_i(x_1^*, \cdots, x_{2M}^*)$. Since all f_i^* are $\geqq 0$, at least M of them, say $f_{i_1}^*, \cdots, f_{i_M}^*$, are $\leqq 2\alpha$. Then $2\alpha \geqq \sum' p(y \mid x_{i_j}^*)$ where the sum is over y for which

$$p(y \mid x_{i_j}^*) \leqq \max_{i \neq i_j} p(y \mid x_i^*).$$

Denoting $x_{i_j}^*$ by x_j and the set of y for which

$$p(y \mid x_{i_j}^*) > \max_{i \neq i_j} p(y \mid x_i^*)$$

by E_j yields (14), and the lemma is proved.

THEOREM 3. *For any indecomposable channel, $H^* \geqq H$, that is, it is possible to transmit at any rate less than the capacity of the channel.*

PROOF. Let (M, ϕ) be any source and let $\{(d_n, r_n), n = 0, 1, 2, \cdots\}$ be a Markov process whose transition matrix is the source-channel matrix and with d_0, r_0 having a uniform distribution on the DR states. Let $x_n = \phi(d_n)$, $y_n = \psi(r_n)$. For any $s \varepsilon I(A)^{(N)}, t \varepsilon I(B)^{(N)}, r \varepsilon I(R)$, write

$$a(s) = P((x_1, \cdots, x_N) = s), \qquad b(t) = P((y_1, \cdots, y_N) = t),$$

$$Q(r, s, t) = P((x_1, \cdots, x_N) = s, \quad (y_1, \cdots, y_N) = t, \quad r_0 = r)R/a(s),$$

$$p(s, t) = P((x_1, \cdots, x_N) = s, \quad (y_1, \cdots, y_N) = t) = a(s) \textstyle\sum_r Q(r, s, t)/R.$$

According to Theorem 2, as $N \to \infty$

$$N^{-1} \log a(x_1, \cdots, x_N) \to -H(X)$$

$$N^{-1} \log b(y_1, \cdots, y_N) \to -H(Y)$$

$$N^{-1} \log p(x_1, \cdots, x_N, y_1, \cdots, y_N) \to -H(X, Y).$$

Given $\epsilon > 0$, choose N so large that, with probability $\geqq 1 - \epsilon$,

$$\frac{\log p(x_1, \cdots, x_N, y_1, \cdots, x_N) - \log a(x_1, \cdots, x_N)}{N} \geqq H(X) - H(X,Y) - \epsilon$$

and

$$\frac{\log b(y_1, \cdots, y_N)}{N} \leqq -H(Y) + \epsilon.$$

We apply the lemma to the product space $U \times V$, where $U = I(A)^{(N)}$, $V = I(B)^{(N)}$, with $p(u, v)$ as defined above and $\delta = 2^{-N(H(Y)-\epsilon)}$, $\lambda = 2^{-N(H(X,Y)-H(X)-\epsilon)}$, and conclude the existence of $M = 2^{NG}$, say, points $u_1, \cdots, u_M \varepsilon X$ and M disjoint subsets $E_1, \cdots, E_M$ of V such that

$$\sum_{u \notin E_i} p(u \mid v_i) \leqq 4 \cdot 2^{-N[H(Y)+H(X)-H(X,Y)-G-2\epsilon]} + 8\epsilon.$$

Thus for any $G < H(X) + H(Y) - H(X, Y)$ we can, for any $\beta > 0$, by first choosing ϵ sufficiently small (less than min $(\beta/9, (H(X) + H(Y) - H(X, Y) - G)/2)$ and then choosing N sufficiently large, find $M = 2^{NG}$ X-sequences $u_1, \cdots, u_M$ of length N and M disjoint subsets $E_1, \cdots, E_M$ of $I(B)^{(N)}$ such that

$$\sum_{v \in E_i} p(v \mid u_i) > 1 - \beta. \tag{17}$$

This does not quite prove that it is possible to transmit at rate G as defined above, since (2) requires that

$$\sum_{y \in E_i} Q(r, u_i, v) > 1 - \epsilon \quad \text{for all} \quad r \in R,$$

that is, that for each initial state of the channel, each of the M messages can be correctly recovered, with large probability. This is an immediate consequence of (17), however, since (17) yields

$$R^{-1} \sum_r \left(\sum_{v \in E_i} Q(r, u_i, v) \right) > 1 - \beta,$$

so that, since $Q(r, u_i, E_i) \leqq 1$ for all r, i,

$$\sum_{v \in E_i} Q(r, u_i, v) > 1 - R\beta$$

for each r. Since β can be made arbitrarily small and R is a fixed number, the number of states of the channel, the proof is complete.

6. The converse half of Shannon's theorem (impossibility of transmission at a rate greater than capacity).

THEOREM 4. *For any indecomposable channel, $H^* \leqq H$, that is, it is not possible to transmit at a rate greater than the capacity of the channel.*

PROOF. Suppose that it is possible to transmit over a given channel at rate G, let ϵ be given, $0 < \epsilon < \frac{1}{2}$ and let $N, u_1, \cdots, u_J, J = 2^{NG}, E_1, \cdots, E_J$ denote the quantities whose existence is implied by the possibility of transmission at rate G. We may suppose that $\cup E_j = I(B)^{(N)}$, since if (2) is satisfied for E_j it is also satisfied if E_j is replaced by a superset. We must exhibit a source (M, ϕ) for which $H(X) + H(Y) - H(X, Y)$ is nearly G. Our source produces inputs in blocks of N by selecting one of the u_j at random, successive choices being independent. The entropy $H(X)$ will then be precisely G. Since observing a long y sequence nearly identifies the corresponding x sequence, the conditional entropy $H(X, Y) - H(Y)$ is small, so that $H(X) + H(Y) - H(X, Y)$ is nearly G.

More precisely, the input source will have NJ states (n, j), with $M((n, j), (n + 1, j)) = 1$ for $n < N$, $M((N, j), (1, i)) = 1/J$ for $i \in I(J)$. We define $\phi(n, j) = u_{jn}$, the nth symbol in the sequence u_j. Let (d_k, r_k) be a Markov process whose transition matrix is the source-channel matrix, and whose initial distribution is such that $d_1 = (1, i)$ with probability $1/J$, $i \in I(J)$ and write $x_k = \phi(d_k)$, $y_k = \psi(r_k)$. Then every x sequence of length NT which is possible has probability $J^{-T} = 2^{-NTG}$ (since $\epsilon < \frac{1}{2}$, $u_i \neq u_j$ for $i \neq j$). From Theorem 2, $H(X) = G$.

To estimate $H(X, Y) - H(Y)$, we recall some results of Shannon [5]. If x is any random variable assuming T distinct values with probabilities $p_1, \cdots, p_T$, the number $-\sum p_t \log p_t$ is called the *entropy* of x and will be denoted by $h(x)$. Always $h(x) \leqq \log T$. If (x, y) are two random variables, each with a finite set of values, the number $h(x, y) - h(y)$ is called the conditional entropy of x given y and is denoted by $h(x \mid y)$. It equals the expected value of the entropy of the conditional distribution of x given y. For any function ϕ defined on the range of y, $h(\phi(y)) \leqq h(y)$ and $h(x \mid \phi(y)) \geqq h(x \mid y)$.

Notice that, in the notation of Theorem 2, $E \log z_N = -h(y_1, \cdots, y_N)$, so that the L_1 convergence in (2) implies that $h(y_1, \cdots, y_N)/N \to H$ as $N \to \infty$. Thus, in our present notation,

$$h(x_1, \cdots, x_{NT} \mid y_1, \cdots, y_{NT})/NT \to H(X, Y) - H(Y)$$

as $T \to \infty$. We have

$$\begin{aligned} h(x_1, \cdots, x_{NT} \mid y_1, \cdots, y_{NT}) &\leqq \sum_{t=0}^{T-1} h(x_{Nt+1}, \cdots, x_{Nt+N} \mid y_{Nt+1}, \cdots, y_{Nt+N}) \\ &\leqq \sum_{t=1}^{T-1} h(a_t \mid b_t), \end{aligned}$$

where $a_t = (x_{Nt+1}, \cdots, x_{Nt+N})$ and $b_t = u_j$ if $(y_{Nt+1}, \cdots, y_{Nt+N}) \varepsilon E_j$ (we may suppose that $\cup E_j = I(B)^{(N)}$). We estimate $h(a_t \mid b_t)$ by the following lemma.

LEMMA. *For any distribution α on a product space $U \times U$ of pairs (a, b) such that $\sum_j \alpha(a, a) \geqq 1 - \epsilon > \frac{1}{2}$ we have*

$$h(a \mid b) \leqq -g(\epsilon) + \epsilon \log (J - 1),$$

where $g(t) = t \log t + (1 - t) \log (1 - t)$, $0 \leqq t \leqq 1$, and J is the number of elements of U.

PROOF OF THE LEMMA. Let $\beta(b) = \sum_a \alpha(a, b)$. Then

$$-h(a \mid b) = \sum_b \beta(b) \sum_a \frac{\alpha(a, b)}{\beta(b)} \log \frac{\alpha(a, b)}{\beta(b)}.$$

Now

$$\begin{aligned} &\sum_i \frac{\alpha(a, b)}{\beta(b)} \log \frac{\alpha(a, b)}{\beta(b)} \\ &\qquad = \frac{\alpha(b, b)}{\beta(b)} \log \frac{\alpha(b, b)}{\beta(b)} + \frac{\beta(b) - \alpha(b, b)}{\beta(b)} \\ &\qquad \cdot \sum_{a \neq b} \frac{\alpha(a, b)}{\beta(b) - \alpha(b, b)} \log \frac{\alpha(a, b)}{\beta(b) - \alpha(b, b)} + \frac{\beta(b) - \alpha(b, b)}{\beta(b)} \log \frac{\beta(b) - \alpha(b, b)}{\beta(b)} \\ &\qquad = g\left(\frac{\alpha(b, b)}{\beta(b)}\right) - \frac{\beta(b) - \alpha(b, b)}{\beta(b)} \log (J - 1). \end{aligned}$$

Consequently,

$$-h(a \mid b) \geqq \sum_b \beta(b) g\left(\frac{\alpha(b, b)}{\beta(b)}\right) - \epsilon \log (J - 1).$$

Since $g(t)$ is convex and $\sum_b \beta(b) = 1$,

$$\sum_b \beta(b) g\left[\frac{\alpha(b, b)}{\beta(b)}\right] \geqq g[\sum_b \alpha(b, b)] \geqq g(1 - \epsilon) = g(\epsilon).$$

The hypotheses of the lemma are satisfied for (a_t, b_t), so that

$$h(a_t \mid b_t) \leqq -g(\epsilon) + \epsilon \log J = -g(\epsilon) + \epsilon NG.$$

Thus

$$h(x_1, \cdots, x_{NT} \mid y_1, \cdots, y_{NT}) \leqq T(-g(\epsilon) + \epsilon NG).$$

Dividing by NT and letting $T \to \infty$ yields

$$H(X, Y) - H(Y) \leqq -\frac{g(\epsilon)}{N} + \epsilon G.$$

Thus, assuming that transmission at rate G is possible we have for every $\epsilon > 0$ and arbitrarily large N, exhibited a source for which

$$H(X) + H(Y) - H(X, Y) \geqq G(1 - \epsilon) + g(\epsilon)/N.$$

It follows that $H \geqq H^*$ and the proof is complete.

7. Another form of Shannon's Theorem. Let $\{w_n, n = 1, 2, \ldots\}$ be any stationary ergodic process whose variables have a finite set of values, say $I(W)$, and consider a given indecomposable channel as defined above. Shannon enquires whether the channel is adequate for transmitting the information produced by the source, with large probability of correct reception. To say that the channel is adequate means that, for every $\epsilon > 0$, there is an integer N_0 such that for any $N \geqq N_0$ there are (1) a function f (the encoder) from $I(W)^{(N)}$ to $I(A)^{(N)}$ and (2) a function g (the decoder) from $I(B)^{(N)}$ to $I(W)^N$ such that, for every initial state r of the channel,

$$\pi_r\{\alpha = \beta\} > 1 - \epsilon,$$

where α and β are random variables (the first N symbols produced by the source and the decoded estimate for these symbols respectively) whose joint distribution π_r is defined by

$$\pi_r\{\alpha = v, \beta = v'\} = \text{Prob } \{(w_1, \cdots, w_N) = v\} \sum_{g(\delta) = v'} Q(r, f(v), \delta)$$

where $Q(r, u, \delta)$, as defined earlier, is the probability that the channel, when initially in state r, on receiving an input u, will produce output δ. The form in which Shannon describes his result is the following.

THEOREM 5. *An indecomposable channel of capacity H is adequate for the stationary ergodic source $\{w_n\}$ if the entropy h of $\{w_n\}$ is less than H, and not if $h > H$.*

The idea of the proof of this result, based on McMillan's theorem and Theorems 3 and 4 above, is extremely simple. According to McMillan's theorem, the source w_n is very likely to produce one of about 2^{hN} sequences of length N, each of which has probability about 2^{-hN}. Accordingly, to have a large probability of transmitting the actual sequence accurately, it is necessary and sufficient that the channel be able to distinguish among about 2^{hN} different input sequences of length N which, by Theorem 3, it is if $h < H$ and is not if $h > H$. The proof below simply makes this idea precise.

PROOF. From (1), for any $\epsilon > 0$ there is an N_1 such that for any $N \geqq N_1$ there is a set $F \subset I(W)^{(N)}$ with not more than $2^{(h+\epsilon)N}$ elements such that

$$\text{Prob } \{(w_1, \cdots, w_N) \varepsilon F\} > 1 - \epsilon.$$

From Theorem 3 there is an $N_0 \geqq N_1$ such that for any $N \geqq N_0$ there are $2^{(H-\epsilon)N} = J$ distinct sequences $u_1, \cdots, u_J$ in $I(A)^{(N)}$ and J disjoint subsets $E_1, \cdots, E_J$ of $I(B)^{(N)}$ such that

$$Q(r, u_j, E_j) > 1 - \epsilon \text{ for all } j \text{ and } r.$$

If $H - \epsilon \geqq h + \epsilon$ there are at most J elements in F, so that there is a function f from $I(W)^{(N)}$ to $I(A)^N$ such that f maps distinct elements of F onto distinct u_j. With this f and with g chosen so that $g(\delta) \varepsilon F$, $f[g(\delta)] = u_j$ for all $\delta \varepsilon E_j$, we have

$$\pi_r\{\alpha = \beta\} > (1 - \epsilon)^2,$$

since the probability that $\alpha \varepsilon F$ is greater than $1 - \epsilon$ and the conditional probability, given that $\alpha = \alpha_0 \varepsilon F$, that $\beta = \alpha_0$ is at least $Q(r, u_j, E_j) > 1 - \epsilon$, where $f(\alpha_0) = u_j$. Thus if $h < H$, the channel is adequate.

Conversely, suppose the channel is adequate. From (1), for any $\epsilon > 0$ there is an N_2 such that for any $N \geqq N_2$ there is a set $F_1 \subset I(W)^{(N)}$ such that

$$\text{Prob } \{(w_1, \cdots, w_N) \varepsilon F_1\} > 1 - \epsilon$$

and Prob $(w_1, \cdots, w_N) = \alpha_0 < 2^{-(h-\epsilon)N}$ for all $\alpha_0 \varepsilon F_1$. Also, there is an $N_3 \geqq N_2$ such that for every $N \geqq N_3$ there are functions f, g satisfying the definition of adequacy. Since $\pi_r\{\alpha = \beta\} > 1 - \epsilon$, there is a subset F_2 of $I(W)^{(N)}$ such that $\pi_r\{\alpha \varepsilon F_2\} > 1 - \sqrt{\epsilon}$ the conditional probability

$$\pi_r\{\alpha = \beta \mid \alpha = \alpha_0\} > 1 - \sqrt{\epsilon} \text{ for } \alpha_0 \varepsilon F_2.$$

Then $\pi_r\{\alpha \varepsilon F_1 \cap F_2\} > 1 - \epsilon - \sqrt{\epsilon}$, so that $F_1 \cap F_2$, and hence F_2 has at least $2^{(h-\epsilon)N}(1 - \epsilon - \sqrt{\epsilon}) = J_1$ elements. For $\alpha_0 \varepsilon F_2$, define $E(\alpha_0)$ as the set of all $\delta \varepsilon I(B)^{(N)}$ such that $g(\delta) = \alpha_0$. The assertion $\pi_r\{\alpha = \beta \mid \alpha = \alpha_0\} > 1 - \sqrt{\epsilon}$ is equivalent to

$$Q(r, f(\alpha_0), E(\alpha_0) > 1 - \sqrt{\epsilon}. \tag{17}$$

Note that, since the sets $E(\alpha_0)$ are disjoint, so are the elements of (α_0), provided $\epsilon < .707$, which we may assume. In summary, for every $\epsilon > 0$ we have found an

N_3 such that for any $N \geqq N_3$ there are at least $J_1(N, \epsilon)$ distinct elements of $I(A)^{(N)}$ (namely the $f(\alpha_0)$, $\alpha_0 \varepsilon F_2$ and $J_1(N, \epsilon)$ corresponding subsets of $I(B)^{(N)}$ (namely the $E(\alpha_0)$) such that (17) holds. Thus if $g < h$, it is possible to transmit at rate G, since, for sufficiently small ϵ ($< h - G$), $J_1(N, \epsilon) > 2^{NG}$ for all sufficiently large N. It now follows from Theorem 4 that $h \leqq H$, and the proof is complete.

REFERENCES

[1] L. BREIMAN, "The individual ergodic theorem of information theory," *Ann. Math. Stat.*, Vol. 28, No. 3 (1957), pp. 809–811.

[2] A. FEINSTEIN, "A new basic theorem of information theory," *Trans. IRE PGIT* (1954), pp. 2–22.

[3] A. I. KHINCHIN, *Mathematical Foundations of Information Theory*, Dover Publications, Inc., 1957.

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

[5] C. E. SHANNON, "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27 (1948), pp. 379–423, and 623–656.

[6] J. WOLFOWITZ, "The coding of messages subject to chance errors," *Illinois Journal of Mathematics*, Vol. 1 (1957), pp. 591–606.