THE CODING OF MESSAGES SUBJECT TO CHANCE ERRORS[1]

BY J. WOLFOWITZ

1. The transmission of messages

Throughout this paper we assume that all "alphabets" involved contain exactly two symbols, say 0 and 1. What this means will be apparent in a moment. This assumption is made only in the interest of simplicity of exposition, and the changes needed when this assumption is not fulfilled will be obvious.

Suppose that a person has a vocabulary of S words (or messages), any or all of which he may want to transmit, in any frequency and in any order, over a "noisy channel". For example, S could be the number of words in the dictionary of a language, provided that it is forbidden to coin words not in the dictionary. What a "noisy channel" is will be described in a moment. Here we want to emphasize that we do not assume anything about the frequency with which particular words are transmitted, nor do we assume that the words to be transmitted are selected by any random process (let alone that the distribution function of the random process is known). Let the words be numbered in some fixed manner. Thus transmitting a word is equivalent to transmitting one of the integers $1, 2, \cdots, S$.

We shall now explain what is meant by a "noisy channel" of memory m. A sequence of $(m + 1)$ elements, each zero or one, will be called an α-sequence. A function p, defined on the set of all α-sequences, and such that always $0 \leqq p \leqq 1$, is associated with the channel and called the channel probability function. A sequence of n elements, each of which is zero or one, will be call an x-sequence. To describe the channel, it will be sufficient to describe how it transmits any given x-sequence, say x_1. Let α_1 be the α-sequence of the first $(m + 1)$ elements of x_1. The channel "performs" a chance experiment with possible outcomes 1 and 0 and respective probabilities $p(\alpha_1)$ and $(1 - p(\alpha_1))$, and transmits the outcome of this chance experiment. It then performs another chance experiment, independently of the first, with possible outcomes 1 and 0 and respective probabilities $p(\alpha_2)$ and $(1 - p(\alpha_2))$, where α_2 is the α-sequence of the $2^{\text{nd}}, 3^{\text{rd}}, \cdots, (m + 2)^{\text{nd}}$ elements of the sequence x_1. This is repeated until $(n - m)$ independent experiments have been performed. The probability of the outcome one in the i^{th} experiment is $p(\alpha_i)$, where α_i is the α-sequence of the $i^{\text{th}}, (i + 1)^{\text{st}}, \cdots, (i + m)^{\text{th}}$ elements of x_1. The x-sequence x_1 is called the transmitted sequence. The chance sequence $Y(x_1)$ of outcomes of the experiments in consecutive order is called the received sequence. Any sequence of $(n - m)$ elements, each zero or one, will be called a y-sequence. Let y_1 be any y-sequence. If $P\{Y(x_1) = y_1\} > 0$ (the symbol

Received January 26, 1957; received in revised form May 15, 1957.

[1] Research under contract with the Office of Naval Research.

$P\{\ \}$ denotes the probability of the relation in braces), we shall say that y_1 is a possible received sequence when x_1 is the transmitted sequence.

Let λ be a positive number which it will usually be desired to have small. A "code" of length t is a set $\{(x_i, A_i)\}$, $i = 1, \cdots, t$, where (a) each x_i is an x-sequence, (b) each A_i is a set of y-sequences, (c) for each i

$$P\{Y(x_i) \in A_i\} \geqq 1 - \lambda,$$

(d) $A_1, \cdots, A_t$ are disjoint sets. The coding problem which is a central concern of the theory of transmission of messages may be described as follows: For given S, to find an n and then a code of length S. The practical applications of this will be as follows: When one wishes to transmit the i^{th} word, one transmits the x-sequence x_i. Whenever the receiver receives a y-sequence which is in A_j, he always concludes that the j^{th} word has been sent. When the receiver receives a y-sequence not in $A_1 \cup A_2 \cup \cdots \cup A_S$, he may draw any conclusion he wishes about the word that has been sent. The probability that any word transmitted will be correctly received is $\geqq 1 - \lambda$.

When such a code is used, s/n is called the "rate of transmission," where $s = \log S$. (All logarithms which occur in the present paper are to the base 2.) Except for certain special[2] functions p, one can find a code for any s, provided that one is willing to transmit at a sufficiently small rate; for the law of large numbers obviously applies, and by sufficient repetition of the word to be transmitted, one can insure that the probability of its correct reception exceeds $1 - \lambda$. The practical advantages of a high rate of transmission are obvious. If there were no "noise" (error in transmission) and signals were received exactly as sent, then s symbols zero or one would suffice to transmit any word in the vocabulary, and one could transmit at the rate one. The existence of an error of transmission means that the sequences to be sent must not be too similar in some reasonable sense, lest they be confused as a result of transmission errors. When n is sufficiently large, we can find $S = 2^s$ sufficiently dissimilar sequences. The highest possible rate of transmission obviously depends on the channel probability function.

2. The contents of this paper

The fundamental ideas of the present subject and paper are due to the fundamental and already classical paper [1] of Shannon. Theorem 1 below was stated and proved by Shannon. However, the latter permits the use of what are called "random codes," and indeed proves Theorem 1 by demonstrating the existence of a random code with the desired property. It seems to the present writer questionable whether random codes are properly codes at all. The definition of a code given in Section 1 of the present paper does not admit random codes as codes; what we have called a code is called in the literature of communication theory an "error correcting" code. In any case,

[2] For example, if $p(\alpha_1) = p(\alpha_2)$, then α_1 and α_2 are indistinguishable in transmission.

the desirability of proving the existence of an error correcting code which would satisfy the conclusion of Shannon's Theorem 1 has always been recognized and well understood (see, for example, [8], Section 3).

The achievement of such a proof is due to Feinstein [2] and Khintchine [4]. The latter utilized an idea from the earlier, not entirely rigorous and without gaps, work of Feinstein, to prove, in full rigor, the general Theorem 3 below. In the present paper, starting from first principles in Section 3, we give already in Section 5 a short and simple proof of Theorem 1. We then return to the subject in Section 8 to prove Theorem 3. Even after allowance is made for the fact that Lemmas 8.2 and 8.3 are not proved here, it seems that our proofs have something to offer in simplicity and brevity.

Theorem 2 for general memory m was stated by Shannon in [1]. Khintchine in [4] pointed out that neither the argument of [1] nor any of the arguments to be found in the literature constitute a proof or even the outline of a proof; he also pointed out the desirability of proving the result and mentioned some of the difficulties. In the present paper we give what seems to be the first proof of Theorem 2. We have reason to believe that it is possible to treat the case of general finite memory along the same lines.[3]

The notion of extending the result for stationary Markov chains (Theorem 1) to stationary, not necessarily Markovian processes (Theorem 3) is due to McMillan [5]. The difficult achievement of carrying out this program correctly and without gaps is due to Khintchine [4]. The theorem we cite below as Lemma 8.3 is due to McMillan. Lemma 8.2 is due to Khintchine.

In [4] Khintchine acknowledges his debt to the paper [2] of Feinstein, although he states that its argument is not exact and that it deals largely with the case of zero memory (and only with Theorem 1 of course). The main idea of [2] seems, to the present writer, to be the ingenious one of proving an inequality like (5.4) below. This pretty idea is employed in the present paper; we find it possible to dispense with many of the details which occur in this connection in [2] and [4].

Shannon and all other writers cited above employ the law of large numbers. The simple notion of δx-sequences and the sequences they generate, which so simplifies our proof below and makes the proof of Theorem 2 possible, also enables us to use Chebyshev's inequality instead of the law of large numbers in Theorems 1 and 2. This has the incidental effect of slightly improving Theorems 1 and 2 over Shannon's original formulation by replacing $o(n)$ terms by $O(n^{1/2})$ terms.

This entire paper is self-contained except for the following incidental remark which we make here in passing: The quantity called $e(n)$ in [2] (the maximum

[3] *Added in proof.* A sequel to the present paper, which has been accepted for publication by this Journal, gives an upper bound on the length of a code for any memory m. When $m = 0$ this bound is the same as that given by Theorem 2. The proof of this result is different from that of Theorem 2.

probability of incorrectly receiving any word) is shown there, for $m = 0$, to approach zero "faster than $1/n$". Using the arguments of the present paper and the inequality (96) of page 288 of [9], one can prove easily for any m that

$$e(n) < c_1 n^{-1/2} e^{-c_2 n},$$

where c_1 and c_2 are positive constants.

3. Combinatorial preliminaries

Let x be any x-sequence and α be any α-sequence. Let $N(\alpha \mid x)$ be the number of elements in x such that each, together with the m elements of x which follow it, constitute the sequence α. Let δ and δ_2 be fixed positive numbers. Let π be any nonnegative function defined on the set of all α-sequences such that

$$\sum_\alpha \pi(\alpha) = 1.$$

We shall say that an x-sequence x is a $\delta\pi x$-sequence if

$$(3.1) \qquad | N(\alpha \mid x) - n\pi(\alpha) | \leqq \delta n^{1/2}$$

for every α-sequence.

A y-sequence y will be said to be generated by the x-sequence x if (1) y is a possible received sequence when x is the transmitted sequence, (2) for any α-sequence α_1 the following is satisfied: Let $j(1), \cdots, j(N(\alpha_1 \mid x))$ be the serial numbers of the elements of x which begin the sequence α_1 (e.g., the elements in the places with serial numbers $j(1), j(1) + 1, \cdots, j(1) + m$, constitute the sequence α_1). Then the number $N(\alpha_1, y \mid x)$ of elements one among the elements of y with serial numbers $j(1), \cdots, j(N(\alpha_1 \mid x))$ satisfies

$$(3.2) \quad | N(\alpha_1, y \mid x) - N(\alpha_1 \mid x)p(\alpha_1) | \leqq \delta_2[N(\alpha_1 \mid x)(p(\alpha_1))(1 - p(\alpha_1))]^{1/2}.$$

Let $M(x)$ denote the number of y-sequences generated by x.

Whenever in this paper the expression 0 log 0 occurs, it is always to be understood as equal to zero. We remind the reader that all logarithms occurring in this paper are to the base 2. For any x-sequence x we define $H_x(Y)$, the conditional entropy of $Y(x)$, by

$$(3.3) \quad \begin{aligned} H_x(Y) = & -(1/n)\textstyle\sum_\alpha N(\alpha \mid x)p(\alpha) \log p(\alpha) \\ & -(1/n)\textstyle\sum_\alpha N(\alpha \mid x)(1 - p(\alpha)) \log (1 - p(\alpha)). \end{aligned}$$

LEMMA 3.1. *For any δ_2 there exists a $K_1 > 0$ such that, for any n and any x-sequence x,*

$$(3.4) \qquad M(x) < 2^{nH_x(Y) + K_1 n^{1/2}}.$$

Proof. Let θ_2 be a generic real number $\leqq \delta_2$ in absolute value. Let y be any y-sequence generated by x. Then

$$
\begin{aligned}
\log P\{Y(x) = y\} &= \textstyle\sum_\alpha N(\alpha \mid x)p(\alpha) \log p(\alpha) \\
&\quad + \textstyle\sum_\alpha N(\alpha \mid x)(1 - p(\alpha)) \log (1 - p(\alpha)) \\
&\quad + \textstyle\sum_\alpha \theta_2 (\alpha)[N(\alpha \mid x)p(\alpha)(1 - p(\alpha))]^{1/2} \log p(\alpha) \\
&\quad - \textstyle\sum_\alpha \theta_2 (\alpha) [N(\alpha \mid x)p(\alpha)(1 - p(\alpha))]^{1/2} \log(1 - p(\alpha)) \\
&\quad > -nH_x(Y) + \delta_2 n^{1/2} \textstyle\sum_\alpha \log p(\alpha) \\
&\quad + \delta_2 n^{1/2} \textstyle\sum_\alpha \log (1 - p(\alpha)) = -nH_x(Y) - K_1 n^{1/2},
\end{aligned} \tag{3.5}
$$

with

$$K_1 = -\delta_2 \textstyle\sum_\alpha \log p(\alpha) - \delta_2 \textstyle\sum_\alpha \log (1 - p(\alpha)).$$

(Here the first summation is over all α such that $p(\alpha) > 0$, and the second summation is over all α such that $p(\alpha) < 1$.) The lemma follows at once from (3.5).

LEMMA 3.2. *Let $\lambda > 0$ be any number. Then, for δ_2 larger than a bound which depends only upon λ, we have, for any n and any x-sequence x,*

$$P\{Y(x) \text{ is a sequence generated by } x\} > 1 - \tfrac{1}{2}\lambda. \tag{3.6}$$

There then exists a $K_2 > 0$ which depends only on δ_2 such that, for any n and any x-sequence x,

$$M(x) > 2^{nH_x(Y) - K_2 n^{1/2}}. \tag{3.7}$$

Proof. (3.6) follows at once from Chebyshev's inequality. As in (3.5) we have, for any y-sequence y generated by x,

$$\log P\{Y(x) = y\} < -nH_x(Y) + K_1 n^{1/2}. \tag{3.8}$$

From (3.6) and (3.8) we have at once that

$$M(x) > (1 - \tfrac{1}{2}\lambda) 2^{nH_x(Y) - K_1 n^{1/2}}. \tag{3.9}$$

Then (3.7) follows at once from (3.9).

4. Preliminaries on Markov chains[4]

Let $X_1, X_2, \cdots$ be a stationary, metrically transitive Markov chain with two possible states, 0 and 1; we shall call this the X process, for short. Suppose

[4] Since we do not assume that the words to be transmitted are chosen by any random process or sent with any particular frequency, the introduction of the X process is not a necessity. The lemmas which involve the X process are of purely combinatorial character (e.g., Lemmas 4.1 and 6.1). The X process serves merely as a device for stating or proving certain combinatorial facts. The reader is invited to verify that this entire paper could be written without the introduction of the X process. In that case the $Y(x)$ process would take the place of the Y process. Only the entropies $H(Y)$ and $H_X(Y)$ need be introduced, and this can be done by means of the $Y(x)$ process and δQx-sequences.

(4.1) $$Q_i = P\{X_1 = i\}, \qquad i = 0, 1,$$
and
(4.2) $$q_{ij} = P\{X_{k+1} = j \mid X_k = i\}$$

is the probability of a transition from state i to state j; $i, j = 0, 1$. For any α-sequence α define

(4.3) $$Q(\alpha) = P\{(X_1, \cdots, X_{m+1}) = \alpha\}.$$

The function Q is a function which satisfies the requirements on the function π of Section 3. Let $\gamma < 1$ be any number. It follows at once from Chebyshev's inequality that, for any n and any δ greater than a lower bound which is a function only of γ and the q_{ij},

(4.4) $$P\{(X_1, \cdots, X_n) \text{ is a } \delta Qx\text{-sequence}\} > \gamma.$$

By the Y process we shall mean the sequence $Y_1, Y_2, \cdots$, where Y_i is a chance variable which assumes only the values zero and one, and the conditional probability that $Y_i = 1$, given the values of $X_1, X_2, \cdots, Y_1, \cdots, Y_{i-1}$, is $p(X_i, \cdots, X_{i+m})$. Henceforth we write for short $X = (X_1, \cdots, X_n)$. Then the conditional distribution of $Y = (Y_1, \cdots, Y_{n-m})$, given X, is the same as that of the sequence received when X is the sequence transmitted. The Y process is obviously stationary, and, by Lemma 8.2 below (proof in [4], page 53), metrically transitive. The conditional entropy $H_X(Y)$ of the Y process relative to the X process is defined by

$$H_X(Y) = -\textstyle\sum_\alpha Q(\alpha)p(\alpha)\log p(\alpha) - \sum_\alpha Q(\alpha)(1 - p(\alpha))\log(1 - p(\alpha)).$$

One verifies easily that there exists a $K_3 > 0$ such that, for any δQx-sequence x,
(4.5) $$| H_x(Y) - H_X(Y) | < K_3\, \delta/n^{1/2}.$$
We at once obtain

LEMMA 4.1. *For any n and any δQx-sequence x, the inequalities* (3.4) *and* (3.7) *hold with $H_x(Y)$ replaced by $H_X(Y)$ and K_1 and K_2 replaced by K_1' and K_2', where K_1' and K_2' are positive numbers which depend only upon δ and δ_2.*

We define the chance variable (function of X) $P\{X\}$ as follows: when $X = x$, $P\{X\} = P\{X = x\}$. Similarly we define the chance variable (function of Y) $P\{Y\}$ as follows: when $Y = y$, $P\{Y\} = P\{Y = y\}$. We define the entropy $H(X)$ of the X process by

$$H(X) = \lim_{n\to\infty} -\frac{1}{n} E\,[\log P\{X\}].$$

This limit obviously exists.

Let the symbol σ^2 ( ) denote the variance of the chance variable in parentheses. We now prove

LEMMA 4.2. *We have*

$$E[\log P\{Y\}] = -Dn + D_0, \tag{4.6}$$

where D is a nonnegative constant and D_0 is a bounded function of n. Also,

$$\sigma^2(\log P\{Y\}) = O(n). \tag{4.7}$$

The quantity

$$D = \lim_{n\to\infty} -\frac{1}{n} E\,[\log P\{Y\}]$$

is called the entropy $H(Y)$ of the Y process.

Proof. We have

$$\log P\{Y\} = \sum_{i=1}^{n} \log P\{Y_i \mid Y_1, \cdots, Y_{i-1}\}. \tag{4.8}$$

Let α^* be some fixed α-sequence such that $Q(\alpha^*) > 0$. In the sequence $X_1, X_2, \cdots$, let $j(1), j(2), \cdots$ be the indices such that

$$(X_{j(i)}, X_{j(i)+1}, \cdots, X_{j(i)+m}) = \alpha^*, \qquad i = 1, 2, \cdots .$$

(These exist with probability one.) Let l^* be the smallest integer such that $j(l^*) \geqq n + 1$. (Again l^* is defined with probability one.) Define symbols such as

$$C_1 = \log P\{Y_1, \cdots, Y_{j(1)-1}\}$$

in the obvious manner analogous to that in which $\log P\{Y\}$ was defined. Since C_1 is a sum of quantities which enter into (4.8) and which are all zero or negative, it follows that $B_1 = EC_1$ could fail to exist only if it were $-\infty$. It will be seen that the latter cannot be. Define

$$C_2 = \sum_{i=n-m+1}^{j(l^*)-1} \log P\{Y_i \mid Y_1, \cdots, Y_{i-1}\}.$$

As before, $B_2 = EC_2$ either exists or is $-\infty$. It will be seen that $B_2 \neq -\infty$.

It is easy to see that

$$Ej(1) \leqq \text{a constant, independent of } n \tag{4.9}$$

$$El^* = 1 + nQ(\alpha^*) \tag{4.10}$$

$$E\big(j(i) - j(i-1)\big) = \text{a constant, independent of } n \text{ and } i. \tag{4.11}$$

$$E\big(j(l^*) - n\big) \leqq \text{a constant, independent of } n. \tag{4.12}$$

From the construction of $j(1), j(2), \cdots$ it follows that the chance variables

$$W_i = \log P\{Y_{j(i)}, \cdots, Y_{j(i+1)-1} | Y_1, \cdots, Y_{j(i)-1}\}, \qquad i = 1, 2, \cdots \tag{4.13}$$

are independently and identically distributed. Actually

$$W_i = \log P\{Y_{j(i)}, \cdots, Y_{j(i+1)-1}\}.$$

From Wald's equation ([10], Theorems 7.1 and 7.4), (4.9), (4.11), (4.12), and the fact that the chance variables W_i, C_1, and C_2 are sums of always non-positive and bounded chance variables which appear in the right member of (4.8), it follows that B_1, B_2, and $EW_i = w$ are all finite, and that B_1 and B_2 are bounded uniformly in n. Applying Wald's equation again we obtain, using (4.10), that

$$E \log P\{Y_1, \cdots, Y_{j(l^*)-1}\} = B_1 + nwQ(\alpha^*). \tag{4.14}$$

Hence

$$E \log P\{Y\} = B_1 + nwQ(\alpha^*) - B_2, \tag{4.15}$$

which proves (4.6).

Now

$$\begin{aligned} \log P\{Y\} - E \log P\{Y\} &= (C_1 - B_1) \\ &\quad + \left(\sum_{i=1}^{l^*-1} W_i - nwQ(\alpha^*)\right) - (C_2 - B_2) \\ &= (C_1 - B_1) + \left(\sum_{i=1}^{l^*-1} W_i - (l^* - 1)w\right) \\ &\quad + ((l^* - 1)w - nwQ(\alpha^*)) - (C_2 - B_2). \end{aligned} \tag{4.16}$$

Now we note that

$$\sigma^2(j(i) - j(i-1)) = \text{a constant, independent of } n. \tag{4.17}$$

Applying an argument like that which leads to Theorem 7.2 of [10], together with (4.17) and Schwarz's inequality, we obtain first that

$$\sigma^2(W_i) = \text{(a finite) constant,} \tag{4.18}$$

and then that

$$E\left(\sum_{i=1}^{l^*-1} W_i - (l^* - 1)w\right)^2 = \sigma^2(W_1)nQ(\alpha^*) \tag{4.19}$$

by (4.10) and

$$E([l^* -1]w - E[l^* -1]w)^2 = O(n) \tag{4.20}$$

(see [6], p. 263, equation (8.10)). Obviously

$$\sigma^2(C_1) \leqq \text{a constant independent of } n, \tag{4.21}$$

$$\sigma^2(C_2) \leqq \text{a constant independent of } n. \tag{4.22}$$

Now take the expected value of the squares of the first and third members of (4.16). Using (4.19), (4.20), (4.21), and (4.22) we obtain that the sum of the expected values of the squares which occur after squaring the third member of (4.16) is $O(n)$. The cross products have expected value $O(n)$ by the Schwarz inequality. This proves (4.7) and completes the proof of the lemma.

Another proof of the fact that the variance of $\log P\{Y\}$ is $O(n)$ can be based on the following: It is known from the theory of Markov chains ([7], page 173, equation (2.2)) that there exists a number h, $0 < h < 1$, such that the absolute value of the correlation coefficient between X_i and X_j is less than $h^{|i-j|}$. From the distribution of the Y_i it follows that a similar statement is true of the correlation coefficient between Y_i and Y_j and also of the correlation coefficient between

$$\log P\{Y_i \mid Y_1, \cdots, Y_{i-1}\}$$

and

$$\log P\{Y_j \mid Y_1, \cdots, Y_{j-1}\}.$$

Since $\log P\{Y\}$ can be written in the form (4.8), the desired conclusion can be deduced from the above.

An immediate consequence of Lemma 4.2 and Chebyshev's inequality is that, for any $\varepsilon' > 0$, there exists a $K_4 > 0$ such that, for any n,

$$(4.23) \quad P\{-nH(Y)-K_4 n^{1/2} < \log P\{Y\} < -nH(Y) + K_4 n^{1/2}\} > 1 - \varepsilon'.$$

The following lemma is now an immediate consequence of (4.23):

LEMMA 4.3. *Let $\varepsilon' > 0$ be any number, and $K_4 > 0$ be a number which, for any n, satisfies* (4.23). *For any n let B be any set of y-sequences such that*

$$P\{Y \epsilon B\} > \gamma_1 > \varepsilon'.$$

Then the set B must contain at least

$$(\gamma_1 - \varepsilon')2^{nH(Y)-K_4 n^{1/2}}$$

y-sequences.

Proof. From (4.23) it follows that the y-sequences in B which satisfy the relationship in braces in (4.23) have probability greater than $\gamma_1 - \varepsilon'$. Since the probability of each such sequence is bounded above by $2^{-nH(Y)+K_4 n^{1/2}}$, the desired result follows.

5. The coding theorem

THEOREM 1. *Let $X_1, X_2, \cdots$ be a stationary, metrically transitive Markov chain with states* 0 *and* 1 *and notation as in Section* 4. *Let the Y process be as defined in Section* 4. *Let λ be an arbitrary positive number. There exists a $K > 0$ such that, for any n, there is a code of length at least*[5]

$$(5.1) \quad 2^{n(H(Y)-H_X(Y))-Kn^{1/2}}.$$

The probability that any word transmitted according to this code will be incorrectly received is less than λ.

[5] An alternate and perhaps more graphic way to state Theorem 1 is to replace $(H(Y)—H_X(Y))$ in (5.1) by $C_1 = \max (H(Y)—H_X(Y))$, where the maximum is over all Markov processes X and their associated Y processes as defined in the statement of Theorem 1. It is obvious that this is an equivalent way of stating Theorem 1.

Proof. We may take $\lambda < \frac{1}{2}$. Let $\gamma < 1$ be any positive number. Let δ be sufficiently large so that (4.4) holds, and choose δ_2 sufficiently large so that (3.6) holds.

Let x_1 be any δQx-sequence, and A_1 any set of y-sequences generated by x_1 such that the following is satisfied for $i = 1$:

$$P\{Y(x_i) \text{ is a sequence generated by } x_i \text{ and not in } A_i\} < \tfrac{1}{2}\lambda. \tag{5.2}$$

Let x_2 be any other δQx-sequence for which we can find a set A_2 of y-sequences generated by x_2 such that A_1 and A_2 are disjoint and (5.2) is satisfied for $i = 2$. Continue in this manner as long as possible, i.e., as long as there exists another δQx-sequence, say x_i, and a set A_i of y-sequences generated by x_i such that $A_1, A_2, \cdots, A_i$ are all disjoint and A_i satisfies (5.2). Let

$$(x_1, A_1), \cdots\cdots, (x_N, A_N)$$

be the resulting code. We have to show that N is large enough.

Let x^* be any δQx-sequence (if one exists) not in the set $x_1, \cdots, x_N$. Then

$$P\{Y(x^*) \text{ is a sequence generated by } x^* \text{ and belongs to } (A_1 \cup A_2 \cup \cdots \cup A_N)\} \geqq \tfrac{1}{2}\lambda. \tag{5.3}$$

If this were not so, we could prolong the code by adding (x^*, A^*), where A^* is the totality of y-sequences generated by x^* and not in $A_1 \cup A_2 \cup \cdots \cup A_N$; this would violate the definition of N. From (4.4), (3.6), (5.2), and (5.3) it follows that

$$P\{Y \epsilon (A_1 \cup A_2 \cup \cdots \cup A_N)\} > \tfrac{1}{2}\gamma\lambda. \tag{5.4}$$

Let the ε' of (4.23) and Lemma 4.3 be equal to $\frac{1}{4}\gamma\lambda$ and let $K_4 > 0$ be any number for which (4.23) is satisfied. It follows from Lemma 4.3 that the set $A_1 \cup A_2 \cup \cdots \cup A_N$ contains at least

$$\tfrac{1}{4}\gamma\lambda \cdot 2^{nH(Y) - K_4 n^{1/2}} \tag{5.5}$$

y-sequences. By Lemma 4.1 the number of y-sequences in $A_1 \cup A_2 \cup \cdots \cup A_N$ is at most

$$N \cdot 2^{nH_X(Y) + K_1' n^{1/2}}. \tag{5.6}$$

The desired result follows at once from (5.5) and (5.6), with

$$K = K_1' + K_4 - \log(\tfrac{1}{4}\gamma\lambda).$$

6. Further preliminaries[6]

The essential part of the present section is the second part of the inequality (6.13) below, which is basic in the proof of Theorem 2 of Section 7. Neither

[6] All the lemmas of this section are of purely combinatorial character. Lemma 6.3 could be easily proved by a purely combinatorial argument without any use of the Y process. This entire section is a concession to the conventional treatment of the subject. All that is needed for the statement and proof of Theorem 2 is the second part of (6.13) and a formal analytic definition of capacity. See also footnote 4.

Lemma 6.1 nor Lemma 6.2 is used in the sequel, and both are given only for completeness. The proof of Lemma 6.1 is omitted because it is very simple, and the proof of Lemma 6.2 is omitted because it involves some computation.

Let the X and Y processes be as defined in Section 4. Obviously

$$H(X) = -\textstyle\sum_{i,j} Q_i \, q_{ij} \log q_{ij} \, . \tag{6.1}$$

We define the chance variables (functions of X and Y) $P\{Y \mid X\}$ and $P\{X \mid Y\}$ as follows: when $X = x$ and $Y = y$, $P\{Y \mid X\} = P\{Y = y \mid X = x\}$, and $P\{X \mid Y\} = P\{X = x \mid Y = y\}$. We verify easily that

$$H_X(Y) = \lim_{n\to\infty} -\frac{1}{n} E\,[\log P\{Y \mid X\}]. \tag{6.2}$$

We define $H_Y(X)$, the conditional entropy of the X process relative to the Y process, by

$$H_Y(X) = \lim_{n\to\infty} -\frac{1}{n} E\,[\log P\{X \mid Y\}]. \tag{6.3}$$

(We shall see in a moment $\big((6.5)\big)$ that this limit exists.) From the obvious relation

$$\log P\{X\} + \log P\{Y \mid X\} = \log P\{Y\} + \log P\{X \mid Y\}, \tag{6.4}$$

we obtain

$$H(X) + H_X(Y) = H(Y) + H_Y(X). \tag{6.5}$$

Throughout the rest of this section we assume that the memory $m = 0$, and that the X_i are independent, identically distributed chance variables. Hence

$$Q_i = q_{ji}\,, \qquad i, j = 0, 1. \tag{6.6}$$

Since $m = 0$, there are only two α-sequences, namely, (0) and (1), and $Q(i) = Q_i$, $i = 0, 1$. Write for short $Q(1) = q$. We assume that $0 < q < 1$. It seems reasonable in this case to denote what was called in Section 4 a "δQx-sequence" by the term "δqx-sequence", and we shall employ this usage (when $m = 0$ and the chance variables X_i are independent). We now give the values of the various entropies, inserting a zero in the symbol for entropy to indicate that $m = 0$ and the X_i are independently (and identically) distributed.

$$H(X_0) = -\,q \log q - (1 - q) \log (1 - q). \tag{6.7}$$

$$\begin{aligned} H(Y_0) = &- [qp(1) + (1 - q)p(0)] \log [qp(1) + (1 - q)p(0)] \\ &- [(1 - q)\big(1 - p(0)\big) + q\big(1 - p(1)\big)]\cdot \\ &\cdot\log[(1 - q)\big(1 - p(0)\big) + q\big(1 - p(1)\big)]. \end{aligned} \tag{6.8}$$

$$
\begin{aligned}
H_X(Y_0) = &- qp(1) \log p(1) - (1 - q)p(0) \log p(0) \\
&- q(1 - p(1)) \log (1 - p(1)) \\
&- (1 - q)(1 - p(0)) \log (1 - p(0)).
\end{aligned} \tag{6.9}
$$

From (6.5) we obtain

$$
\begin{aligned}
H_Y(X_0) = &-(1 - q)p(0) \log \frac{(1 - q)p(0)}{[qp(1) + (1 - q)p(0)]} \\
&- qp(1) \log \frac{qp(1)}{[qp(1) + (1 - q)p(0)]} \\
&- q(1 - p(1)) \log \frac{q(1 - p(1))}{[q(1 - p(1)) + (1 - q)(1 - p(0))]} \\
&- (1 - q)(1 - p(0)) \log \frac{(1 - q)(1 - p(0))}{[q(1 - p(1)) + (1 - q)(1 - p(0))]}.
\end{aligned} \tag{6.10}
$$

The maximum, with respect to q, of

$$H(X_0) - H_Y(X_0) = H(Y_0) - H_X(Y_0)$$

is called the capacity (when $m = 0$) C_0 of the channel.

LEMMA 6.1. *There exists a $K_5 > 0$ such that, for any n, the number $M(\delta q)$ of δqx-sequences satisfies*

$$2^{nH(X_0)-K_5 n^{1/2}} < M(\delta q) < 2^{nH(X_0)+K_5 n^{1/2}}. \tag{6.11}$$

Let δ' be some fixed positive number such that any y-sequence which is generated by a δqx-sequence, cannot be generated by an x-sequence which is not a $\delta' qx$-sequence. Such a δ' exists; we have only to take δ' larger than a lower bound which is a function of q, δ, δ_2, $p(0)$, and $p(1)$. We have

LEMMA 6.2. *There exists a $K_6 > 0$ with the following property: Let y be any y-sequence which is generated by some δqx-sequence. Then the number $M'(y)$ of $\delta' qx$-sequences which generate y satisfies*

$$2^{nH_Y(X_0)-K_6 n^{1/2}} < M'(y) < 2^{nH_Y(X_0)+K_6 n^{1/2}}. \tag{6.12}$$

We now prove

LEMMA 6.3. *There exists a $K_7 > 0$ such that, for any n, the number $M''(\delta q)$ of different y-sequences generated by all δqx-sequences satisfies*

$$2^{nH(Y_0)-K_7 n^{1/2}} < M''(\delta q) < 2^{nH(Y_0)+K_7 n^{1/2}}. \tag{6.13}$$

Proof. Let θ, with any subscript, denote a number not greater than one in absolute value. The chance variables $Y_1, Y_2, \cdots$ are independently and identically distributed. We have

$$P\{Y_1 = 1\} = qp(1) + (1 - q)p(0) = u, \text{ say.}$$

If y is generated by a δqx-sequence, then the number V_1 of elements one in y is given by

$$V_1 = n(qp(1) + (1 - q)p(0)) + n^{1/2}(\theta_1\delta + 2\theta_2\delta_2) + 0(1).$$

Since

$$P\{Y = y\} = u^{V_1}(1 - u)^{n-V_1} > 2^{-nH(Y_0)-K_7 n^{1/2}}$$

for a suitable $K_7 > 0$, the second part of (6.13) follows at once.

The first part of (6.13) follows from Lemma 4.3. It may be necessary to increase the above K_7.

When $p(1) = p(0)$, $C_0 = 0$. One can verify that otherwise $C_0 > 0$. Incidentally, it follows from Lemmas 6.1 and 6.2 that $H(X_0) \geqq H_Y(X_0)$.

7. Impossibility of a rate of transmission greater than the capacity when $m = 0$

In this section we prove the following

THEOREM 2. *Let $m = 0$, and let λ, $1 > \lambda > 0$, be any given number. There exists a $K' > 0$ such that, for any n, any code with the property that the probability of transmitting any word incorrectly is $< \lambda$, cannot have a length greater than*

$$2^{nC_0+K'n^{1/2}} \tag{7.1}$$

If $p(1) = p(0)$ and therefore $C_0 = 0$, the theorem is trivial. For then it makes no difference whether one transmits a zero or a one, and it is impossible to infer from the sequence received what sequence has been transmitted. We therefore assume henceforth that $C_0 > 0$.

It will be convenient to divide the proof into several steps. Let q_0 be the value of q which maximizes $H(Y_0) - H_X(Y_0)$. We shall have occasion to consider the various entropies as functions of q, which, in this section, we shall always exhibit explicitly, e.g., $H(Y; q)$.[7] Let δ and δ_2 be positive constants. Throughout this section it is to be understood that by the word "code" we always mean a code with the property that the probability of transmitting any word incorrectly is $< \lambda$.

LEMMA 7.1. *There exists a $K_8 > 0$ with the following property: Let n be any integer. Let $(x_1, A_1), \cdots, (x_N, A_N)$ be any code such that $x_1, \cdots, x_N$ are $\delta q_0 x$-sequences, and A_i, $i = 1, \cdots, N$, contains only y-sequences generated by x_i. Then*

$$N < 2^{nC_0+K_8 n^{1/2}}. \tag{7.2}$$

Proof. It follows from (3.8) and (4.5) that there exists a $K_8' > 0$ such that the set $(A_1 \cup A_2 \cup \cdots \cup A_N)$ contains at least

[7] Naturally, this is the entropy of Y when the X's are independently distributed and $m = 0$.

$$N \cdot 2^{nH_X(Y;q_0)-K_8' n^{1/2}} \tag{7.3}$$

sequences. By Lemma 6.3 it cannot contain more than

$$2^{nH(Y;q_0)+K_7 n^{1/2}} \tag{7.4}$$

sequences. The lemma follows at once with $K_8 = K_7 + K_8'$.

LEMMA 7.2. *There exists a* $K_9 > 0$ *with the following property: Let* n *be any integer. Let* $(x_1, A_1), \cdots, (x_N, A_N)$ *be any code such that* $x_1, \cdots, x_N$ *are* $\delta q_0 x$*-sequences. Then*

$$N < 2^{nC_0+K_9 n^{1/2}}. \tag{7.5}$$

(*In other words, the conclusion of Lemma* 7.1 *holds even if the* A_i, $i = 1, \cdots, N$, *are not required to consist only of sequences generated by* x_i.)

Proof. Let δ_2 be so large that (3.6) holds. From A_i, $i = 1, \cdots, N$, delete the y-sequences not generated by x_i; call the resulting set A_1'. The A_i', $i = 1, \cdots, N$, are of course disjoint. The set $(x_1, A_i'), \cdots, (x_N, A_N')$ fulfills all the requirements of a code except perhaps the one that the probability of correctly transmitting any word is $> 1 - \lambda$. However, from (3.6) it follows that the probability of correctly transmitting any word when this latter set is used is $> 1 - 3\lambda/2$. But now the result of Lemma 7.1 applies,[8] and the present lemma follows. (Of course the constant K_8 of Lemma 7.1 depends on λ, but this does not affect our conclusion.)

LEMMA 7.3. *There exists a constant* $K_{10} > 0$ *with the following property: Let* q *be any point in the closed interval* [0, 1], let n *be any integer, and let* $(x_1, A_1), \cdots, (x_N, A_N)$ *be any code such that* $x_1, \cdots, x_N$ *are* δqx*-sequences. Then*

$$N < 2^{nC_0+K_{10} n^{1/2}}. \tag{7.6}$$

Proof. Let q', $0 < q' < \frac{1}{2}$, be such that $H(X; q) < \frac{1}{2} C_0$ if $q < q'$ or $q > 1 - q'$. If $q < q'$ or $q > 1 - q'$, the total number of all δqx-sequences is less than the right member of (7.6) for suitable $K_{10} > 0$. Then (7.6) holds a fortiori.

It remains to consider the case $q' \leqq q \leqq 1 - q'$. If now one applies the argument of Lemma 7.2 and considers how K_9 depends upon q, one obtains that there exists a positive continuous function $K_9(q)$ of q, $q' \leqq q \leqq 1 - q'$, such that, for any n,

$$\begin{aligned} N &< 2^{n(H(Y;q)-H_X(Y;q))+K_9(q)n^{1/2}} \\ &\leqq 2^{nC_0+K_9(q)n^{1/2}}. \end{aligned}$$

We now increase, if necessary, the constant K_{10} of the previous paragraph so that it is not less than the maximum of $K_9(q)$ in the closed interval $[q', 1 - q']$, and obtain the desired result (7.6).

[8] Except when $3\lambda/2 \geqq 1$. In that case we choose δ_2 so large that the right member of (3.6) is $1 - \lambda/a$, where $a > 0$ is such that $\lambda + \lambda/a < 1$.

Proof of Theorem 2. Divide the interval [0, 1] into $J = n^{1/2}/2\delta$ intervals of length $2\delta/n^{1/2}$ and let $t_1, \cdots, t_J$ be the midpoints of these intervals. Let $(x_1, A_1), \cdots, (x_N, A_N)$ be any code. Then this code is the union of J codes $W_1, \cdots, W_J$ as follows: For $i = 1, \cdots, J$, W_i is that subset of the original code all of whose x-sequences are δt_i x-sequences. By Lemma 7.3 the length of W_i, $i = 1, \cdots, J$, is less than $2^{nC_0+K_{10}n^{1/2}}$. Hence the length N of the original code is less than

$$J \cdot 2^{nC_0+K_{10}n^{1/2}}.$$

The theorem follows at once if K' is sufficiently large.

8. Extension to stationary processes

Throughout this section let $X_1, X_2, \cdots$ be a stationary, metrically transitive stochastic process such that X_i, $i = 1, 2, \cdots$, takes only the values one and zero. Define the Y process, $Q(\alpha)$, and $H_X(Y)$ exactly as in Section 4. Let $\varepsilon^* > 0$ be any number, no matter how small, and write $\delta^* = \varepsilon^* n^{1/2}$. Let $\gamma < 1$ be any positive number. From the ergodic theorem we obtain at once the following analogue of (4.4): For n sufficiently large,

$$(8.1) \qquad P\{(X_1, \cdots, X_n) \text{ is a } \delta^*Qx\text{-sequence}\} > \gamma.$$

For δ_2 sufficiently large the inequalities (3.6), (3.4), and (3.7) hold exactly as before, and we obtain the following analogue of Lemma 4.1:

LEMMA 8.1. *For any $\varepsilon > 0$, ε^* sufficiently small, and δ_2 sufficiently large, we have, for n sufficiently large and any δ^*Qx-sequence x,*

$$(8.2) \qquad 2^{n(H_X(Y)-\varepsilon)} < M(x) < 2^{n(H_X(Y)+\varepsilon)}.$$

The following lemmas are proved in [4]:

LEMMA 8.2. *The process $Y_1, Y_2, \cdots$ is metrically transitive.*

LEMMA 8.3. *Let $Z_1, Z_2, \cdots$ be any stationary, metrically transitive stochastic process such that Z_i can take only finitely many values. Let $Z = (Z_1, \cdots, Z_n)$. Define the chance variable (function of Z) $P\{Z\}$ as follows: When Z is the sequence z, $P\{Z\} = P\{Z = z\}$. Then $-(1/n) \log P\{Z\}$ converges stochastically to a constant.*

For our Y process the constant limit of Lemma 8.3 is called the entropy $H(Y)$ of the Y process. This definition of $H(Y)$ is easily verified to be consistent with that of Section 4.

Lemma 8.3 implies the following analogue of (4.23) for our Y process: Let $\varepsilon' > 0$ be any number. Then, for n sufficiently large,

$$(8.3) \qquad P\{-n(H(Y)+\varepsilon') < \log P\{Y\} < -n(H(Y)-\varepsilon')\} > 1 - \varepsilon'.$$

Exactly as (4.23) easily implies Lemma 4.3, so (8.3) implies

LEMMA 8.4. *Let $\varepsilon' > 0$ be any number and let n be sufficiently large for* (8.3) *to hold. Let B be any set of y-sequences such that*

$$P\{Y \in B\} > \gamma_1 > \varepsilon'.$$

Then the set B must contain at least

$$(\gamma_1 - \varepsilon')2^{n(H(Y)-\varepsilon')}$$

y-sequences.

Now the analogues of all the preliminaries needed to prove Theorem 1 have been established, and we have, by exactly the same proof,

THEOREM 3. *Let $X_1, X_2, \cdots$ be a stationary, metrically transitive stochastic process with states 0 and 1. Let the Y process be as defined in Section 4. Let λ and ε be arbitrary positive numbers. For any n sufficiently large there exists a code of length at least*

$$2^{n(H(Y)-H_X(Y)-\varepsilon)}. \tag{8.4}$$

The probability that any word transmitted according to this code will be incorrectly received is less than λ.[9]

The author is grateful to Professor K. L. Chung and Professor J. Kiefer for their kindness in reading the manuscript and for interesting comments.

REFERENCES

1. C. E. SHANNON, *A mathematical theory of communication*, Bell System Tech. J., vol. 27 (1948), pp. 379–423, 623–656.
2. A. FEINSTEIN, *A new basic theorem of information theory*, Transactions of the Institute of Radio Engineers, Professional Group on Information Theory, 1954 Symposium on Information Theory, pp. 2–22.
3. A. KHINTCHINE, *The concept of entropy in the theory of probability*, Uspehi Matem. Nauk (N.S.), vol. 8 no. 3 (55), (1953), pp. 3–20.
4. ———, *On the fundamental theorems of the theory of information*, Uspehi Matem. Nauk (N.S.), vol. 11 no. 1 (67), (1956), pp. 17–75.
5. B. MCMILLAN, *The basic theorems of information theory*, Ann. Math. Statistics, vol. 24 (1953), pp. 196–219.
6. W. FELLER, *An introduction to probability theory and its applications*, New York, John Wiley and Sons, 1950.
7. J. L. DOOB, *Stochastic processes*, New York, John Wiley and Sons, 1953.
8. E. N. GILBERT, *A Comparison of signaling alphabets*, Bell System Tech. J., vol. 31 (1952), pp. 504–522.
9. PAUL LÉVY, *Théorie de l'addition des variables aléatoires*, Paris, Gauthier-Villars, 1937.
10. J. WOLFOWITZ, *The efficiency of sequential estimates and Wald's equation for sequential processes*, Ann. Math. Statistics, vol. 18 (1947), pp. 215–230.

CORNELL UNIVERSITY
ITHACA, NEW YORK

[9] An equivalent way of stating Theorem 3 is to replace $(H(Y) - H_X(Y))$ in (8.4) by $C_2 = \sup (H(Y) - H_X(Y))$, where the supremum operation is over all X processes as described in the theorem, each X process with its associated Y process. Obviously $C_0 \leqq C_1 \leqq C_2$ (C_1 is defined in footnote 5). When $m = 0$ it follows from Theorem 2 that $C_0 = C_1 = C_2$. Hence, when $m = 0$, Theorem 3 is actually weaker than Theorem 1.