INFORMATION AND CONTROL 1, 6–25 (1957)

Certain Results in Coding Theory for Noisy Channels*

CLAUDE E. SHANNON

Massachusetts Institute of Technology, Cambridge, Massachusetts

In this paper we will develop certain extensions and refinements of coding theory for noisy communication channels. First, a refinement of the argument based on "random" coding will be used to obtain an upper bound on the probability of error for an optimal code in the memoryless finite discrete channel. Next, an equation is obtained for the capacity of a finite state channel when the state can be calculated at both transmitting and receiving terminals. An analysis is also made of the more complex case where the state is calculable at the transmitting point but not necessarily at the receiving point.

PROBABILITY OF ERROR BOUND FOR THE DISCRETE FINITE MEMORYLESS CHANNEL

A discrete finite memoryless channel with finite input and output alphabets is defined by a set of transition probabilities $p_i(j)$,

$$i = 1, 2, \cdots, a; \qquad j = 1, 2, \cdots, b,$$

with $\sum_j p_i(j) = 1$ $(i = 1, 2, \cdots, a)$ and all $p_i(j) \geqq 0$. Here $p_i(j)$ is the probability, if input letter i is used, that output letter j will be received. A *code word* of length n is a sequence of n input letters (that is, n integers each chosen from $1, 2, \cdots, a$). A *block code of length n* with M words is a mapping of the integers from 1 to M (messages) into a set of code words each of length n. A *decoding system* for such a code is a mapping of all sequences of output words of length n into the integers from 1 to M (that is, a procedure for deciding on an original integer or message when any particular output word is received). We will be considering situa-

* This work was carried out at the Research Laboratory of Electronics, Massachusetts Institute of Technology, and was supported in part by the United States Army (Signal Corps), the United States Air Force (Office of Scientific Research, Air Research and Development Command), and the United States Navy (Office of Naval Research); and in part by Bell Telephone Laboratories, Inc.

tions in which all integers from 1 to M are used with the same probability $1/M$. The probability of error P_e for a code and decoding system is the probability of an integer being transmitted and received as a word which is mapped into a different integer (that is, decoded as another message).

Thus:

$$P_e = \sum_u \sum_{v \in S_u} \frac{1}{M} Pr(v \mid u)$$

where u ranges over all input integers 1, 2, $\cdots$, M; v ranges over the received words of length n; and S_u is the set of received words that are not decoded as u. $Pr\ (v \mid u)$ is of course the probability of receiving v if the message is u. Thus if u is mapped into input word $(i_1, i_2, \cdots, i_n)$ and v is word $(j_1, j_2, \cdots, j_n)$, then

$$Pr\ (v \mid u) = p_{i_1}(j_1)\ p_{i_2}(j_2) \cdots p_{i_n}(j_n).$$

While we assume all messages in a code to be used with equal probabilities $1/M$, it is useful, in studying a channel, to consider the assignment of different probabilities to input words. Suppose, in fact, that in a given channel we assign arbitrary probabilities to the different input words u of length n, probability $P(u)$ for word u. We then have probabilities for all input-output word pairs of length n,

$$Pr(u, v) = P(u)\ Pr(v \mid u),$$

where u and v are input and output words of length n and $Pr(v \mid u)$ is the probability of output word v if input word u is used. (This is the product of the transition probabilities for corresponding letters of u and v). Given $P(u)$ then, any numerical function of u and v becomes a random variable. In particular, the mutual information (per letter), $I(u, v)$ is a random variable

$$I(u, v) = \frac{1}{n} \log \frac{Pr(u, v)}{P(u)Pr(v)} = \frac{1}{n} \log \frac{Pr(v \mid u)}{\sum_u P(u)Pr(v \mid u)}$$

The distribution function for this random variable will be denoted by $\rho(x)$. Thus

$$\rho(x) = Pr[I(u, v) \leqq x]$$

The function $\rho(x)$ of course depends on the arbitrary assignment of

probabilities $P(u)$. We will now prove a theorem bounding the probability of error for a possible code in terms of the function $\rho(x)$.

THEOREM 1: *Suppose some $P(u)$ for input words u of length n gives rise to a distribution of information per letter $\rho(I)$. Then given any integer M and any $\theta > 0$ there exists a block code with M messages and a decoding system such that if these messages are used with equal probability, the probability of error P_e is bounded by*

$$P_e \leqq \rho(R + \theta) + e^{-n\theta}$$

where $R = (1/n)\log M$.

PROOF: For a given M and θ consider the pairs (u, v) of input and output words and define the set T to consist of those pairs for which $\log Pr(u, v)/P(u)Pr(v) > n(R + \theta)$. When the u's are chosen with probabilities $P(u)$, then the probability that the (u, v) pair will belong to the set T is, by definition of ρ, equal to $1 - \rho(R + \theta)$.

Now consider the ensemble of codes obtained in the following manner. The integers $1, 2, 3, \cdots, M = e^{nR}$ are associated independently with the different possible input words $u_1, u_2, \cdots, u_B$ with probabilities $P(u_1), P(u_2), \cdots P(u_B)$. This produces an ensemble of codes each using M (or less) input words. If there are B different input words u_i, there will be exactly B^M different codes in this ensemble corresponding to the B^M different ways we can associate M integers with B input words. These codes have different probabilities. Thus the (highly degenerate) code in which all integers are mapped into input word u_1 has probability $P(u_1)^M$. A code in which d_k of the integers are mapped into u_k has probability $\prod_k P(u_k)^{d_k}$. We will be concerned with the average probability of error for this ensemble of codes. By this we mean the average probability of error when these codes are weighted according to the probabilities we have just defined. We imagine that in using any one of these codes, each integer is used with probability $1/M$. Note that, for some particular selections, several integers may fall on the same input word. This input word is then used with higher probability than the others.

In any particular code of the ensemble, our decoding procedure will be defined as follows. Any received v is decoded as the integer with greatest probability conditional on the received v. If several integers have the same conditional probability we decode (conventionally) as the smallest such integer. Since all integers have unconditional probability $1/M$, this decoding procedure chooses one of those having the greatest probability of causing the received v.

We now wish to compute the average probability of error or "ambiguity" P_a in this ensemble of codes where we pessimistically include with the errors all cases where there are several equally probable causes of the received v.

In any particular code of the ensemble an input word u or a pair (u, v) will not, in general, occur with the probabilities $P(u)$ or $Pr(u, v)$. In the ensemble average, however, each word u has probability $P(u)$ and each (u, v) pair probability $Pr(u, v)$, since integers are mapped into u with just this probability. Indeed, a particular message, say the integer 1, will be mapped into u with probability $P(u)$. A particular case of integer 1, say, mapped into u and resulting in received v will result in an error or ambiguity if there are, in the code in question, one or more integers mapped into the set $S_v(u)$ of input of words which have a probability of causing v higher than are qual to that of u. Because of the independence in placing the other integers, it is easy to calculate the fraction of codes in which this occurs. In fact, let

$$Q_v(u) = \sum_{u' \epsilon S_v(u)} P(u')$$

Thus $Q_v(u)$ is the probability associated with all words more probable or as probable conditioned on the received word v as u is. The fraction of codes in which integer 2 is not in $S_v(u)$ is (because of the independence of placing of the integers) equal to $1 - Q_v(u)$. The fraction of codes in which $S_v(u)$ is free of *all* other integers is $(1 - Q_v(u))^{M-1}$. A similar argument applies to any other integer as well as 1. Thus, in the ensemble, the probability of error or ambiguity due to cases where the message is mapped into input word u and received as v is given exactly by

$$Pr(u, v)[1 - (1 - Q_v(u))^{M-1}].$$

The average probability of error or ambiguity, then, is given by

$$P_a = \sum_{u,v} Pr(u, v)[1 - (1 - Q_v(u))^{M-1}]. \tag{1}$$

We now wish to place a bound on this in terms of the information distribution ρ. First, break the sum into two parts, a sum over the (u, v) set T defined above where $\log Pr(u, v)/P(u)Pr(v) > n(R + \theta)$ and over the complementary set $\overline{T}$.

$$P_a = \sum_{\bar{T}} Pr(u, v)[1 - (1 - Qv(u))^{M-1}] + \sum_{T} Pr(u, v)[1 - (1 - Q_v(u))^{M-1}].$$

Since $[1 - (1 - Q_v(u))]^{M-1}$ is a probability, we may replace it by 1 in

the first sum, increasing the quantity. This term becomes, then, $\sum_{\bar{T}} Pr(u, v)$ which by definition is $\rho(R + \theta)$. In the second sum, note first that $(1 - Q_v(u))^{M-1} \geqq 1 - (M - 1)Q_v(u)$ by a well-known inequality. Hence, the second sum is increased by replacing

$$[1 - (1 - Q_v(u))^{M-1}]$$

by $(M - 1)Q_v(u)$ and even more so by $MQ_v(u)$.

$$P_e \leqq P_a \leqq \rho(R + \theta) + M \sum_T Pr(u, v)Q_v(u).$$

We now show that for u, v in T, $Q_v(u) \leqq e^{-n(R+\theta)}$. In fact, with u, v in T

$$\log \frac{Pr(v \mid u)}{Pr(v)} > n(R + \theta),$$

$$Pr(v \mid u) > Pr(v)e^{n(R+\theta)}.$$

If $u' \in S_v(u)$,

$$Pr(v \mid u') \geqq Pr(v \mid u) > Pr(v)e^{n(R+\theta)}$$

$$Pr(u', v) > Pr(u')Pr(v)e^{n(R+\theta)}$$

$$Pr(u' \mid v) > Pr(u')e^{n(R+\theta)}$$

Summing each side over $u' \in S_v(u)$ gives

$$1 \geq \sum_{u' \in S_v(u)} P_r(u' \mid v) > e^{n(R+\theta)}\, Q_v(u)$$

The left inequality holds because the sum of a set of disjoint probabilities cannot exceed 1. We obtain

$$Q_v(u) < e^{-n(R+\theta)} \qquad (u, v) \in T$$

Using this in our estimate of P_e we have

$$\begin{aligned} P_e &< \rho(R + \theta) + e^{nR}e^{-n(R+\theta)} \sum_T Pr(u, v) \\ &\leqq \rho(R + \theta) + e^{-n\theta} \end{aligned}$$

using again the fact that the sum of a set of disjoint probabilities cannot exceed one. Since the average P_e over the ensemble of codes satisfies $P_e \leqq \rho(R + \theta) + e^{-n\theta}$, there must exist a particular code satisfying the same inequality. This concludes the proof.

Theorem 1 is one of a number of results which show a close relation between the probability of error in codes for noisy channels and the

distribution of mutal information $\rho(x)$. Theorem 1 shows that if, by associating probabilities $P(u)$ with input words, a certain $\rho(x)$ can be obtained, then codes can be constructed with a probability of error bounded in terms of this $\rho(x)$. We now develop a kind of converse relation: given a code, there will be a related $\rho(x)$. It will be shown that the probability of error for the code (with optimal decoding) is closely related to this $\rho(x)$.

THEOREM 2: *Suppose a particular code has $M = e^{nR}$ messages and the distribution function for the mutual information I (per letter) between messages and received words is $\rho(x)$ (the messages being used with equal probability). Then the optimal detection system for this code gives a probability of error P_e satisfying the inequalities*

$$\tfrac{1}{2}\rho\left(R - \frac{1}{n}\log 2\right) \leqq P_e \leqq \rho\left(R - \frac{1}{n}\log 2\right)$$

It should be noted that ρ has a slightly different meaning here than in Theorem 1. Here it relates to mutual information between messages and received words—in Theorem 1, between *input words* and received words. If, as would usually be the case, all messages of a code are mapped into distinct input words, these reduce to the same quantity.

PROOF: We first prove the lower bound. By definition of the function ρ, the probability is equal to $\rho(R - (1/n)\log 2)$, that

$$\frac{1}{n}\log\frac{Pr(u, v)}{Pr(u)Pr(v)} \leqq R - \frac{1}{n}\log 2,$$

where u is a message and v a received word. Equivalently,

$$Pr(u \mid v) \leqq Pr(u)e^{Rn}\tfrac{1}{2}$$

or (using the fact that $Pr(u) = e^{-nR}$)

$$Pr(u \mid v) \leqq \tfrac{1}{2}$$

Now fix attention on these pairs (u, v) for which this inequality

$$Pr(u \mid v) \leqq \tfrac{1}{2}$$

is true, and imagine the corresponding (u, v) lines to be marked in black and all other (u, v) connecting lines marked in red. We divide the v points into two classes: C_1 consists of those v's which are decoded into u's connected by a red line (and also any v's which are decoded into u's not connected to the v's): C_2 consists of v's which are decoded into u's

connected by a black line. We have established that with probability $\rho(R - (1/n) \log 2)$ the (u, v) pair will be connected by a black line. The v's involved will fall into the two classes C_1 and C_2 with probability ρ_1, say and $\rho_2 = \rho(R - (1/n) \log 2) - \rho_1$. Whenever the v is in C_1 an error is produced since the actual u was one connected by a black line and the decoding is to a u connected by a red line (or to a disconnected u). Thus these cases give rise to a probability ρ_1 of error. When the v in question is in class C_2, we have $Pr(u \mid v) \leqq \frac{1}{2}$. This means that with at least an equal probability these v's can be obtained through other u's than the one in question. If we sum for these v's the probabilities of all pairs $Pr(u, v)$ except that corresponding to the decoding system, then we will have a probability at least $\rho_2/2$ and all of these cases correspond to incorrect decoding. In total, then, we have a probability of error given by

$$\rho_e \geqq \rho_1 + \rho_2/2 \geqq \tfrac{1}{2}\rho(R - (1/n) \log 2)$$

We now prove the upper bound. Consider the decoding system defined as follows. If for any received v there exists a u such that $Pr(u \mid v) > \frac{1}{2}$, then the v is decoded into that u. Obviously there cannot be more than one such u for a given v, since, if there were, the sum of these would imply a probability greater than one. If there is no such u for a given v, the decoding is irrelevant to our argument. We may, for example, let such u's all be decoded into the first message in the input code. The probability of error, with this decoding, is then less than or equal to the probability of all (u, v) pairs for which $Pr(u \mid v) \leqq \frac{1}{2}$. That is,

$$P_e \leqq \sum_S Pr(u, v) \quad \text{(where } S \text{ is the set of pairs } (u, v) \text{ with } Pr(u \mid v) \leqq \tfrac{1}{2}).$$

The condition $Pr(u \mid v) \leqq \frac{1}{2}$ *is equivalent to* $Pr(u, v)/Pr(v) \leqq \frac{1}{2}$, or, again, to $Pr(u, v)/Pr(u)\ Pr(v) \leqq \frac{1}{2}\ P_r(u)^{-1} = \frac{1}{2}\ e^{nR}$. This is equivalent to the condition

$$(1/n) \log Pr(u, v)/Pr(u)Pr(v) \leqq R - (1/n) \log 2.$$

The sum $\sum_S Pr(u, v)$ where this is true is, by definition, the distribution function of $(1/n) \log Pr(u, v)/Pr(u)Pr(v)$ evaluated at $R - (1/n) \log 2$, that is,

$$P_e \leqq \sum_S Pr(u, v) = \rho(R - (1/n) \log 2).$$

PROBABILITY OF ERROR BOUND IN TERMS OF MOMENT GENERATING FUNCTION

We will now develop from the bound of Theorem 1 another expression that can be more easily evaluated in terms of the channel parameters. Suppose first that the probabilities $P(u)$ assigned to words in Theorem 1 are equal to the product of probabilities for letters making up the words. Thus, suppose u consists of the sequence of letters $i_1, i_2, \cdots, i_n$ and $P(u)$ is then $P_{i_1} \cdot P_{i_2} \cdot P_{i_3} \cdots P_{i_n}$. If v consists of letters $j_1, j_2, \cdots, j_n$ then $Pr(v) = Pr(j_1) \cdot Pr(j_2) \cdots Pr(j_n)$ and $Pr(u, v) = Pr(i_1, j_1) \cdot Pr(i_2, j_2) \cdots Pr(i_n, j_n)$. Also

$$I(u, v) = \frac{1}{n}\left[\log \frac{Pr(i_1 j_2)}{Pr(i_1)Pr(j_2)} + \log \frac{Pr(i_2 j_2)}{Pr(i_2)Pr(j_2)} + \cdots\right]$$

$$= \frac{1}{n}[I_1 + I_2 + \cdots + I_n]$$

where I_k is the mutual information between the kth letters of u and v.

The different I's are here independent random variables all with the same distribution. We therefore have a central limit theorem type of situation; $nI(u, v)$ is the sum of n independent random variables with identical distributions. $\rho(x)$ can be bounded by any of the inequalities which are known for the distribution of such a sum. In particular, we may use an inequality due to Chernov on the "tail" of such a distribution (Chernov, 1952). He has shown, by a simple argument using the generalized Chebycheff inequality, that the distribution of such sums can be bounded in terms of the moment generating function for a single one of the random variables, say $\varphi(s)$. Thus let

$$\varphi(s) = E[e^{sI}]$$

$$= \sum_{ij} P_i p_i(j) \exp\left[s \log \frac{p_i(j)}{\sum_k P_k p_k(j)}\right]$$

$$= \sum_{ij} P_i p_i(j) \left[\frac{p_i(j)}{\sum_k P_k p_k(j)}\right]^s$$

It is convenient for our purposes to use the log of the moment generating function $\mu(s) = \log \varphi(s)$, (sometimes called the semi-invariant generating function). Chernov's result translated into our notation states that

$$\rho(\mu'(s)) \leqq e^{[\mu(s) - s\mu'(s)]n} \qquad s \leqq 0$$

Thus by choosing the parameter s at any negative value we obtain a

bound on the information distribution ρ of exponential form in n. It is easily shown, also, that if the variance of the original distribution is positive then $\mu'(s)$ is a strictly monotone increasing function of s and so also is the coefficient of n in the exponent, $\mu(s) - s\mu'(s)$ (for negative s). Indeed the derivatives of these quantities exist and are $\mu''(s)$ and $-s\mu''(s)$, respectively. $\mu''(s)$ is readily shown to be positive by a Schwartz inequality.

THEOREM 3: *In a memoryless channel with finite input and output alphabets, let $\mu(s)$ be the semi-invariant generating function for mutual information with some assignment of input letter probabilities, P_i for letter i, and with channel transition probabilities $p_i(j)$, that is:*

$$\mu(s) = \log \sum_{i,j} P_i p_i(j) \left[\frac{p_i(j)}{\sum_i P_i p_i(j)} \right]^s$$

Then there exists a code and decoding system of length n, rate R and probability of error P_e satisfying the inequalities

$$R \geqq \mu(s) - (s-1)\mu'(s)$$

$$P_e \leqq 2e^{(\mu(s)-s\mu'(s))n} \qquad s \leqq 0$$

If as $s \to -\infty$, $\mu(s) - (s-1)\mu'(s) \to R^* > 0$ then for $R \leqq R^*$

$$P_e \leqq e^{(E^*+R^*-R)n}$$

where $E^* = \lim (\mu(s) - s\mu'(s))$ as $s \to -\infty$.

PROOF: We have, from *Theorem* 1, that

$$P_e \leqq \rho(R+\theta) + e^{-n\theta}$$

$$\leqq e^{[\mu(s)-s\mu'(s)]n} + e^{-n\theta} \qquad s \leqq 0$$

where s is chosen so that $\mu'(s) = R + \theta$. This will hold when θ is such that the resulting s is negative. We choose θ (which is otherwise arbitrary) to make the coefficients of n in the exponents equal. (Since the first term is monotone increasing in θ and the second monotone decreasing, it is easily seen that this choice of θ is quite good to minimize the bound. In fact, the bound can never be less than half its value for this particular θ.) This relation requires that

$$\mu(s) - s\mu'(s) = -\theta$$

$$= R - \mu'(s)$$

$$R = \mu(s) + (1-s)\mu'(s)$$

Since the exponents are now equal, the probability of error is bounded by twice the first term:

$$P_e \leqq 2e^{[\mu(s)-s\mu'(s)]n}$$

These relations are true for all negative s and give the first results of the theorem.

However, in ome cases, as $s \to -\infty$ the rate R approaches a positive limiting value. In fact, $R \to I_{\min} + \log Pr[I_{\min}]$ and the exponent in the P_e bound approaches $\log Pr[I_{\min}]$. For rates R lower than this limiting value the exponents cannot be made equal by any choice of s. We may, however, now choose θ in such a way that $R + \theta$ is just smaller than $I_{\min}$, say $I_{\min} - \epsilon$. Since $\rho(I_{\min} - \epsilon) = 0$ the probability of error is now bounded by $P_e \leqq e^{-n\theta} = e^{-n(I_{\min}-R-\epsilon)}$. This being true for any ϵ we can construct codes for which it is true with $\epsilon = 0$. That is

$$P_e \leqq e^{-n(I_{\min}-R)}$$

for $R < I_{\min}$. Notice that as R approaches its limiting value in the first bound, $I_{\min} + \log Pr[I_{\min}]$, the exponents in both bounds approach the same value, namely $\log Pr[I_{\min}]$. The coefficient, however, improves from 2 to 1.

These bounds can be written in another form that is perhaps more revealing. Define a set of "tilted" probabilities $Q_s(I)$ for different values of information I by the following:

$$Q_s(I) = \frac{Pr(I)e^{sI}}{\sum_I Pr(I)e^{sI}}$$

In other words the original probability of a value I is increased or decreased by a factor e^{sI} and the resulting values normalized to sum to unity. For large positive values of s, this tilted set of probabilities $Q_s(I)$ tend to emphasize the probabilities $Pr(I)$ for positive I and reduce those for negative I. At $s = 0$ $Q_0(I) = Pr(I)$. At negative s the negative I values have enhanced probabilities at the expense of positive I values. As $s \to \infty$, $Q_s(I) \to 0$ except for $I = I_{\max}$ the largest value of I with positive probability (since the set of u, v pairs is finite, $I_{\max}$ exists), and $Q_s(I_{\max}) \to 1$. These tilted probabilities are convenient in evaluating the "tails" of distribution that are sums of other distributions. In terms

of $Q_s(I)$ we may write

$$\begin{aligned}
\mu(s) &= \log \sum Pr(I)e^{sI} \\
&= \textstyle\sum_{I'} Q_s(I') \log \sum_I Pr(I)e^{sI} \\
\mu'(s) &= \textstyle\sum_I Pr(I)e^{sI}I / \sum_I Pr(I)e^{sI} \\
&= \textstyle\sum_I Q_s(I)I \\
\mu(s) - s\mu'(s) &= \textstyle\sum_I Q_s(I) \log (Pr(I)/Q_s(I)) \\
\mu - (s-1)\mu'(s) &= \textstyle\sum_I Q_s(I)[I + \log Pr(I)/Q_s(I)]
\end{aligned}$$

The coefficients of n in these exponents are of some interest. They relate to the rapidity of approach of P_e to zero as n increases. Plotted as a function of R, the behavior is typically as shown in Fig. 1. Here we have assumed the P_i for the letters to be the P_i which give channel capacity. The coefficient E of n for the first bound in the theorem is a curve tangent to the axis at C (here $s = 0$), convex downward and ending ($s = -\infty$) at $R = I_{\min} + \log Pr[I_{\min}]$ and $E = \log Pr[I_{\min}]$. The second bound in the theorem gives an E curve which is a straight line of slope -1 passing through this point and intersecting the axes at $I_{\min}, 0$ and $0, I_{\min}$. In the neighborhood of $R = C$ the curve behaves as

$$E \doteq \frac{(C-R)^2}{2\mu''(0)}$$

Here $\mu''(0)$ is the variance of I. These properties all follow directly from the formulas for the curves.

The limiting exponent (as $n \to \infty$) satisfies $E = \mu(s) - (s-1)\mu'(s)$. We have

$$\begin{aligned}
\frac{dE}{dR} &= \frac{dE}{ds} \Big/ \frac{dR}{ds} \\
&= \frac{-s}{1-s}
\end{aligned}$$

so the slope of the ER curve is monotone decreasing as s ranges from 0 to $-\infty$, the slope going from 0 to -1. Since the second bound corresponds to a straight line of slope -1 in the ER plot, the two bounds not only join in value but have the same slope as shown in Fig. 1.

The curve would be as indicated if the P_i are those which maximize the rate at the channel capacity, for then

$$R(0) = \mu(0) - (0 - 1)\mu'(0) = \mu'(0) = C.$$

The bound, however, of the theorem applies for any set of P_i when the corresponding $\mu(s)$ is used. To obtain the strongest result the bound should be optimized for each value of R under variation of P_i. The same applies to the straight line portion where we maximize $I_{\min}$. If this were done a curve would be obtained which is the envelope of all possible curves of this type with different values of P_i. Since each individual curve is convex downward the envelope is also convex downward. The equations for this envelope may be found by the Lagrange method maximizing $R + \lambda E + \eta \sum_i P_i$. It must be remembered, of course, that the P_i must be non-negative. The problem is similar to that involved in calculating the channel capacity. The equations for the envelope will be

$$E = \mu(s) - s\mu'(s)$$

$$R = \mu(s) - (s - 1)\mu'(s)$$

$$(1 + \lambda)\frac{\partial \mu}{\partial P_i} - (1 + \lambda)s\frac{\partial \mu'}{\partial P_i} + \frac{\partial \mu'}{\partial P_i} + \eta = 0 \quad \text{for all } i \text{ except a set for which } P_i = 0$$

and subject to:

$$\sum P_i = 1$$

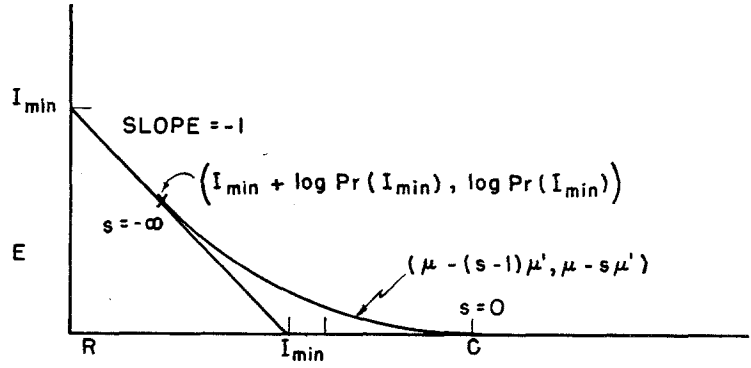

FIG. 1

The bound here should be maximized by choosing different subsets of the P_i for the nonvanishing set.

The upper bound obtained in Theorem 3 is by no means the strongest that can be found. As $n \to \infty$ even the coefficients of n in the exponent can be, in general, improved by more refined arguments. We hope in another paper to develop these further results, and also to give corresponding *lower* bounds on the probability of error of the same exponential type. The upper bound in Theorem 3 is, however, both simple and useful. It has a universality lacking in some of the stronger results (which only assume simple form when n is large).

CAPACITY OF THE FINITE STATE CHANNEL WITH STATE CALCULABLE AT BOTH TERMINALS

In certain channels with memory, the internal state of the channel can be calculated from the initial state (assumed known) at the beginning of transmission and the sequence of transmitted letters. It may also be possible to determine the state at any time at the receiving terminal from the initial state and the sequence of received letters. For such channels we shall say the state is *calculable at both terminals*.

To satisfy the first requirement it is clearly necessary that for any (attainable) internal state s, the next state t must be a function of s and x, $t = f(s, x)$, where x is the transmitted letter.

For the state to be calculable at the receiving point it is necessary that, for all attainable states s, the next stage t must be a function of s and the received letter y, $t = g(s, y)$.

For each possible s, t pair we may find the subset $A(s, t)$ of x's leading from s to t and the subset $B(s, t)$ of y's which correspond to a state transition from s to t. For each input letter x in the set $A(s, t)$ the output letter y will necessarily be in the set $B(s, t)$ and there will be a transition probability, the probability (in state s), if x is transmitted, that y will be received. For a particular s, t pair, the sets of letters $A(s, t)$ and $B(s, t)$ and the corresponding transition probabilities can be thought of as defining a memoryless discrete channel corresponding to the s, t pair. Namely, we consider the memoryless channel with input alphabet the letters from $A(s, t)$, output letters from $B(s, t)$ and the corresponding transition probabilities.

This channel would be physically realized from the given channel as follows. The given channel is first placed in state s, one letter is transmitted from set $A(s, t)$ (resulting in state t), the channel is then returned

to state s and a second letter from set $A(s, t)$ transmitted, etc. The capacity of such a discrete memoryless channel can be found by the standard methods. Let the capacity from state s to state t be C_{st} (in natural units) and let $N_{st} = e^{C_{st}}$. Thus N_{st} is the number of equivalent noiseless letters for the s, t sub-channel. If the set $A(s, t)$ is empty, we set $N_{st} = 0$.

The states of such a channel can be grouped into equivalence classes as follows. States s and s' are in the same class if there is a sequence of input letters which, starting with state s, ends in s', and conversely a sequence leading from s' to s. The equivalence classes can be partially ordered as follows. If there is a sequence leading from a member of one class to a member of a second class, the first class is higher in the ordering than the second class.

Within an equivalence class one may consider various possible closed sequences of states; various possible ways, starting with a state, to choose a sequence of input letters which return to this state. The number of states around such a cycle will be called the cycle length. The greatest common divisor of all cycle lengths in a particular equivalence class will be called the basic period of that class. These structural properties are analogous to those of finite state markoff processes, in which "transition with positive probability" takes the place of a "possible transition for some input letter."

We shall consider only channels in which there is just one equivalence class. That is, it is possible to go from any state s to any state t by some sequence of input letters (i.e., any state is accessible from any other). The more general case of several equivalence classes is more complex without being significantly more difficult.

THEOREM 4: *Let K be a finite state channel with finite alphabets, with state calculable at both terminals, and any state accessible from any other state. Let N_{st} be the number of equivalent letters for the sub-channel relating to transitions from state s to state t. Let N be the (unique) positive real eigenvalue of the matrix N_{st}, that is, the positive real root of*

$$| N_{st} - N\delta_s | = 0.$$

Then N is the equivalent number of letters for the given channel K; its capacity is $C = \log N$.

PROOF: We will first show that there exist block codes which transmit at any rate $R < C$ and with probability of error arbitrarily small. Consider the matrix N_{st}. If this is raised to the nth power we obtain a matrix with elements, say, $N_{st}^{(n)}$. The element $N_{st}^{(n)}$ can be thought of as

a sum of products, each product corresponding to some path n steps long from state s to state t, the product being the product of the original matrix elements along this path, and the sum being the sum of such products for all such possible paths. This follows immediately by mathematical induction and the definition of matrix multiplication.

Furthermore, $N_{st}^{(n)}$ can be interpreted as the equivalent number of letters for the memoryless channel defined as follows. Imagine starting the original channel in state s and using as input "letters" sequences of length n of the original letters allowing just those sequences which will end in state t after the sequence of n. The output "letters" are sequences of received letters of length n that could be produced under these conditions. This channel can be thought of as a "sum" of channels (corresponding to the different state sequences from s to t in n steps) each of which is a "product" of channels (corresponding to simple transitions from one state to another). (The sum of two channels is a channel in which a letter from either of the two channels may be used; the product is the channel in which a letter from both given channels is used, this ordered pair being an input letter of the product channel). The equivalent number of noise free letters for the sum of channels is additive, and for the product, multiplicative. Consequently the channel we have just described, corresponding to sequences from state s to state t in n steps, has an equivalent number of letters equal to the matrix element $N_{st}^{(n)}$.

The original matrix N_{st} is a matrix with non-negative elements. Consequently it has a positive real eigenvalue which is greater than or equal to all other eigenvalues in absolute value. Furthermore, under our assumption that it be possible to pass from any state to any other state by some sequence of letters, there is only one positive real eigenvalue. If d is the greatest common divisor of closed path lengths (through sequences of states), then there will be d eigenvalues equal to the positive real root multiplied by the different dth roots of unity. When the matrix N_{st} is raised to the nth power, a term $N_{st}^{(n)}$ is either zero (if it is impossible to go from s to t in exactly n steps) or is asymptotic to a constant times $N^{(n)}$.

In particular, for n congruent to zero, mod d, the diagonal terms $N_{tt}^{(n)}$ are asymptotic to a constant times N^n, while if this congruence is not satisfied the terms are zero. These statements are all well known results in the Frobenius theory of matrices with non-negative elements, and will not be justified here (Frobenius, 1912).

If we take n a sufficiently large multiple of d we will have, then,

$N_{11}^{(n)} > k\,N^n$ with k positive. By taking n sufficiently large, then, the capacity of the channel whose input "letters" are from state 1 to state 1 in n steps can be made greater than $(1/n)\log kN^n = \log N + (1/n)\log k$. Since the latter term can be made arbitrarily small we obtain a capacity as close as we wish to $\log N$. Since we may certainly use the original channel in this restricted way (going from state 1 to state 1 in blocks of n) the original channel has a capacity at least equal to $\log N$.

To show that this capacity cannot be exceeded, consider the channel K_n defined as follows for sequences of length n. At the beginning of a block of length n the channel K_n can be put into an arbitrary state chosen from a set of states corresponding to the states of K. This is done by choice of a "state letter" at the transmitting point and this "state letter" is transmitted noiselessly to the receiving point. For the next n symbols the channel behaves as the given channel K with the same constraints and probabilities. At the end of this block a new state can be freely chosen at the transmitter for the next block. Considering a block of length n (including its initial state information) as a single letter and the corresponding y block including the received "state letter," as a received letter we have a memoryless channel K_n.

For any particular initial-final state pair s, t, the corresponding capacity is equal to $\log N_{st}^{(n)}$. Since we have the "sum" of these channels available, the capacity of K_n is equal to $\log \sum_{s,t} N_{st}^{(n)}$. Each term in this sum is bounded by a constant times N^n, and since there are only a finite number of terms (because there are only a finite number of states) we may assume one constant for all the terms, that is $N_{st}^{(n)} < kN^n$ (all n, s, t). By taking n sufficiently large we clearly have the capacity of K_n per letter, bounded by $\log N + \epsilon$ for any positive ϵ. But now any code that can be used in the original channel can also be used in the K_n channel for any n since the latter has identical constraints except at the ends of n blocks at which point all constraints are eliminated. Consequently the capacity of the original channel is less than or equal to that of K_n for all n and therefore is less than or equal to $\log N$. This completes the proof of the theorem.

This result can be generalized in a number of directions. In the first place, the finiteness of the alphabets is not essential to the argument. In effect, the channel from state s to t can be a general memoryless channel rather than a discrete finite alphabet channel.

A second slight generalization is that it is not necessary that the state be calculable at the receiver after each received letter, provided it is

eventually possible at the receiver to determine all previous states. Thus, in place of requiring that the next state be a function of the preceding state and the received letter, we need only require that there should not be two different sequences of states from any state s to any state t compatible with the same sequence of received letters.

THE CAPACITY OF A FINITE STATE CHANNEL WITH STATE CALCULABLE AT TRANSMITTER BUT NOT NECESSARILY AT RECEIVER

Consider now a channel with a finite input alphabet, a finite output alphabet, and a finite number of internal states with the further property that the state is known at the beginning and can be calculated at the transmitter for each possible sequence of input letters. That is, the next state is a function of the current state and the current input letter. Such a channel is defined by this state transition function $s_{n+1} = f(s_n, x_n)$, (the $n + 1$ state as a function of state s_n and nth input symbol), and the conditional probabilities in state s, if letter x is transmitted, that the output letter will be y, $p_{sx}(y)$. We do not assume that the state is calculable at the receiving point.

As before, the states of such a channel can be grouped into a partially ordered set of equivalence classes. We shall consider again only channels in which there is just one equivalence class. That is, it is possible to go from any state s to any state t by some sequence of input letters.

We first define a capacity for a particular state s. Let the channel be in state s and let $X_1 = (x_1, x_2, \cdots, x_n)$ be a sequence of n input letters which cause the channel to end in the same state s. If the channel is in state s and the sequence X_1 is used, we can calculate the conditional probabilities of the various possible output sequences Y of length n. Thus, if the sequence X_1 leads through states $s, s_2, s_3, \cdots, s_n, s$ the conditional probability of $Y_1 = (y_1, y_2, \cdots, y_n)$ will be $Pr(Y_1/X_1) = Ps_{x_1}(y_1)Ps_2x_2(y_2) \cdots Ps_nx_n(y_n)$. Consider the X's (leading from s to s in n steps) as individual input letters in a memoryless channel with the y sequences Y as output letters and the conditional probabilities as the transition probabilities. Let $C(n, s)$ be the capacity of this channel. Let $C(s)$ be the least upper bound of $(1/n)C(n, s)$ when n varies over the positive integers. We note the following properties:

1. $C(kn, s) \geqq kC(n, s)$. This follows since in choosing probabilities to assign the X letters of length kn to achieve channel capacity one may at least do as well as the product probabilities for a sequence of kX's each of length n. It follows that if we approximate to $C(s)$

within ϵ at some particular n (i.e. $| C(s) - C(n, s) | < \epsilon$) we will approximate equally well along the infinite sequence $2n$, $3n$, $4n$, $\cdots$.

2. $C(s) = C$ is independent of the state s. This is proved as follows. Select a sequence of input letters U leading from state s' to state s and a second sequence V leading from s to s'. Neither of these need contain more than m letters where m is the (finite) number of states in the channel. Select an n_1 for which $C(n_1, s) > C(s) - \epsilon/2$ and with n_1 large enough so that:

$$(C(s) - \epsilon/2) \frac{n_1}{n_1 + 2m} \geq C(s) - \epsilon$$

This is possible since by the remark 1 above $C(s)$ is approximated as closely as desired with arbitrarily large n_1. A set of X sequences for the s' state is constructed by using the sequences for the s state and annexing the U sequence at the beginning and the V sequence at the end. If each of these is given a probability equal to that used for the X sequences in the s state to achieve $C(n, s)$, then this gives a rate for the s' sequences of exactly $C(n, s)$ but with sequences of length at most $n_1 + 2m$ rather than n_1. It follows that $C(s') \geqq (C(s) - \epsilon/2)(n_1/n_1 + 2m) \geqq C(s) - \epsilon$. Of course, interchanging s and s' gives the reverse result $C(s) \geqq C(s') - \epsilon$ and consequently $C(s) = C(s')$. (Note that, if there were several equivalence classes, we would have a C for each class, not necessarily equal).

3. Let $C(n, s, s')$ be the capacity calculated for sequences starting at s and ending at s' after n steps. Let $C(s, s') = \lim_{n\to\infty}(1/n)C(n, s, s')$. Then $C(s, s') = C(s) = C$. This is true since we can change sequences from s to s' into sequences from s to s by a sequence of length at most m added at the end. By taking n sufficiently large in the lim the effect of an added m can be made arbitrarily small, (as in the above remark 2) so that $C(s, s') \geqq C(s) - \epsilon$. Likewise, the s to s sequences which approximate $C(s)$ and can be made arbitrarily long can be translated into s to s' sequences with at most m added letters. This implies $C(s) \geqq C(s, s') - \epsilon$. Hence $C(s) = C(s, s')$.

We wish to show first that starting in state s_1 it is possible to signal with arbitrarily small probability of error at any rate $R < C$ where C is the quantity above in remark 3. More strongly, we will prove the following.

THEOREM 5: *Given any $R < C$ there exists $E(R) > 0$ such that for any*

$n = k\,d$ (an integer multiple of d, the basic cycle length) there are block codes of length n having M words with $(1/n) \log M \geqq R$ and with probability of error $P_e \leqq e^{-E(R)n}$. There does not exist a sequence of codes of increasing block length with probability of error approaching zero and rate greater than C.

PROOF: The affirmative part of the result is proved as follows. Let $R_1 = (R + C)/2$. Let s_1 be the initial state of the channel and consider sequences of letters which take the state from s_1 to s_1 in n_1 steps. Choose n_1 so that $C(n_1, s_1) > (3C + R)/4$. Use these sequences as input letters and construct codes for the rate R_1. By Theorem 2 the probability of error will go down exponentially in the length of the code. The codes here are of length $n_1, 2n_1, 3n_1, \cdots$ in terms of the original letters, but this merely changes the coefficient of n by a factor $1/n_1$. Thus, for multiples of n_1 the affirmative part of the theorem is proved. To prove it for all multiples of d, first note that it is true for all sufficiently large multiples of d, since by going out to a sufficiently large multiple of n_1 the effect of a suffix on the code words bringing the state back to s_1 after multiples of d, can be made small (so that the rate is not substantially altered). But now for smaller multiples of d one may use any desired code with a probability of error less than 1 (e.g., interpret any received word as message 1, with $P_e = 1 - 1/M < 1$). We have then a finite set of codes up to some multiple of d at which a uniform exponential bound takes over. Thus, one may choose a coefficient $E(R)$ such that $P_e < e^{-E(R)n}$ for n *any* integer multiple of d.

The negative part of our result, that the capacity C cannot be exceeded, is proved by an argument similar to that used for the case where the state was calculable at the receiver. Namely, consider the channel K_n defined as follows. The given channel K may be put at the beginning into any state and the name of this state transmitted noiselessly to the receiving point. Then n letters are transmitted with the constraints and probabilities of the given channel K. The final state is then also transmitted to the receiver point. This process is then repeated in blocks of n. We have here a memoryless channel which for any n "includes" the given channel. Any code for the given channel K could be used if desired in K_n with equally good probability of error. Hence the capacity of the given channel K must be less than or equal to that of K_n for every n. On the other hand K_n is actually the "sum" of a set of channels corresponding to sequences from state s to state t in n steps; channels with capacities previously denoted by $C(n, s, t)$. For all sufficiently large n, and

for all s, t, we have $(1/n)C(n, s, t) < C + \epsilon$ as we have seen above. Hence for all $n > n_0$, say, the capacity of K_n is bounded by $C + \epsilon + (1/n) \log m^2$ where m is the number of states. It follows that the capacity of K is not greater than C.

It is interesting to compare the results of this section where the state is calculable at the transmitter only with those of the preceding section where the state is calculable at both terminals. In the latter case, a fairly explicit formula is given for the capacity, involving only the calculation of capacities of memoryless channels and the solution of an algebraic equation. In the former case, the solution is far less explicit, involving as it does the evaluation of certain limits of a rather complex type.

RECEIVED: April 22, 1957.

REFERENCES

CHERNOV, H., (1952). A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. *Ann. Math. Stat.* **23,** 493–507.

ELIAS, P. (1956). *In* "Information Theory" (C. Cherry, ed.). Academic Press, New York.

FEINSTEIN, A. (1955). Error Bounds in Noisy Channels Without Memory. *IRE Trans. on Inform. Theory* **IT-1,** 13–14 (Sept.).

FROBENIUS, G. (1912). Über Matrizen aus nichtnegativen Elementen. *Akad. Wiss. Sitzber. Berlin*, pp. 456–477.

SHANNON, C. E. (1948). Mathematical Theory of Communication. *Bell System Tech. J.* **27,** 379–423.

SHANNON, C. E. (1956). The Zero Error Capacity of a Noisy Channel. *IRE Trans. on Inform. Theory* **IT-2,** 8–19 (Sept.).