TWO-WAY COMMUNICATION CHANNELS

CLAUDE E. SHANNON
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS

1. Introduction

A two-way communication channel is shown schematically in figure 1. Here x_1 is an input letter to the channel at terminal 1 and y_1 an output while x_2 is an

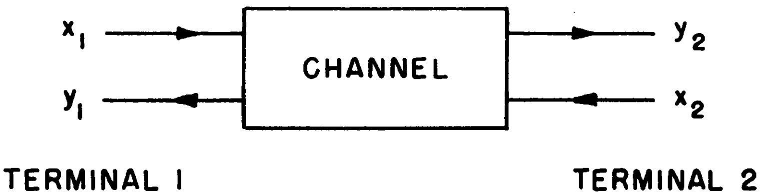

FIGURE 1

input at terminal 2 and y_2 the corresponding output. Once each second, say, new inputs x_1 and x_2 may be chosen from corresponding input alphabets and put into the channel; outputs y_1 and y_2 may then be observed. These outputs will be related statistically to the inputs and perhaps historically to previous inputs and outputs if the channel has memory. The problem is to communicate in both directions through the channel as effectively as possible. Particularly, we wish to determine what pairs of signalling rates R_1 and R_2 for the two directions can be approached with arbitrarily small error probabilities.

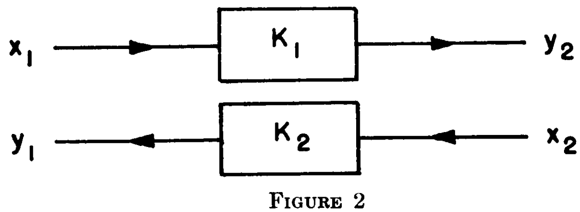

FIGURE 2

Before making these notions precise, we give some simple examples. In figure 2 the two-way channel decomposes into two independent one-way noiseless binary

This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research).

channels K_1 and K_2. Thus x_1, x_2, y_1 and y_2 are all binary variables and the operation of the channel is defined by $y_2 = x_1$ and $y_1 = x_2$. We can here transmit in each direction at rates up to one bit per second. Thus we can find codes whose

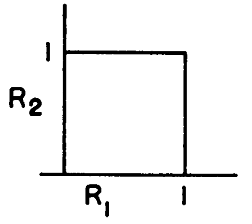

FIGURE 3

rates (R_1, R_2) approximate as closely as desired any point in the square, figure 3, with arbitrarily small (in this case, zero) error probability.

In figure 4 all inputs and outputs are again binary and the operation is defined

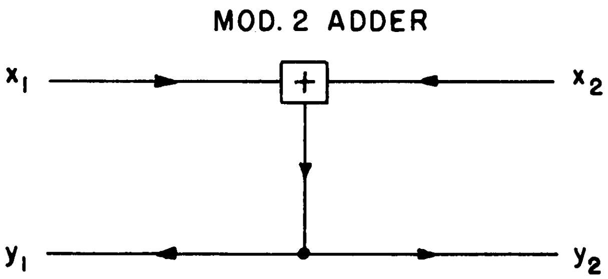

FIGURE 4

by $y_1 = y_2 = x_1 + x_2$ (mod 2). Here again it is possible to transmit one bit per second in each direction simultaneously, but the method is a bit more sophisticated. Arbitrary binary digits may be fed in at x_1 and x_2 but, to decode, the observed y must be corrected to compensate for the influence of the transmitted x. Thus an observed y_1 should be added to the just transmitted x_1 (mod 2) to determine the transmitted x_2. Of course here, too, one may obtain lower rates than the (1, 1) pair and again approximate any point in the square, figure 3.

A third example has inputs x_1 and x_2 each from a *ternary* alphabet and outputs y_1 and y_2 each from a binary alphabet. Suppose that the probabilities of different output pairs (y_1, y_2), conditional on various input pairs (x_1, x_2), are given by table I. It may be seen that by using only $x_1 = 0$ at terminal 1 it is possible to send one bit per second in the $2 - 1$ direction using only the input letters 1 and 2 at terminal 2, which then result with certainty in a and b respectively at terminal 1. Similarly, if x_2 is held at 0, transmission in the $1 - 2$ direction is possible at one bit per second. By dividing the time for use of these two strategies in the ratio λ to $1 - \lambda$ it is possible to transmit in the two directions with

TABLE I

	x_1x_2 \ y_1y_2	Output Pair aa	ab	ba	bb
Input Pair	00	1/4	1/4	1/4	1/4
	01	1/2	1/2	0	0
	02	0	0	1/2	1/2
	10	1/2	0	1/2	0
	11	1/4	1/4	1/4	1/4
	12	1/4	1/4	1/4	1/4
	20	0	1/2	0	1/2
	21	1/4	1/4	1/4	1/4
	22	1/4	1/4	1/4	1/4

average rates $R_1 = 1 - \lambda$, $R_2 = \lambda$. Thus we can find codes approaching any point in the triangular region, figure 5. It is not difficult to see, and will follow

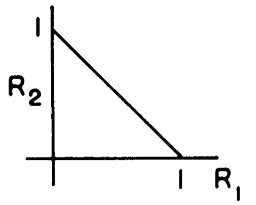

FIGURE 5

from later results, that no point outside this triangle can be approached with codes of arbitrarily low error probability.

In this channel, communication in the two directions might be called incompatible. Forward communication is possible only if x_2 is held at zero. Otherwise, all x_1 letters are completely noisy. Conversely, backward communication is possible only if x_1 is held at zero. The situation is a kind of discrete analogue to a common physical two-way system; a pair of radio telephone stations with "push-to-talk" buttons so arranged that when the button is pushed the local receiver is turned off.

A fourth simple example of a two-way channel, suggested by Blackwell, is the binary multiplying channel. Here all inputs and outputs are binary and the operation is defined $y_1 = y_2 = x_1x_2$. The region of approachable rate pairs for this channel is not known exactly, but we shall later find bounds on it.

In this paper we will study the coding properties of two-way channels. In particular, inner and outer bounds on the region of approachable rate pairs (R_1, R_2) will be found, together with bounds relating to the rate at which zero error probability can be approached. Certain topological properties of these bounds will be discussed and, finally, we will develop an expression describing the region of approachable rates in terms of a limiting process.

2. Summary of results

We will summarize here, briefly and somewhat roughly, the main results of the paper. It will be shown that for a memoryless discrete channel there exists a convex region G of approachable rates. For any point in G, say (R_1, R_2), there exist codes signalling with rates arbitrarily close to the point and with arbitrarily small error probability. This region is of the form shown typically in figure 6,

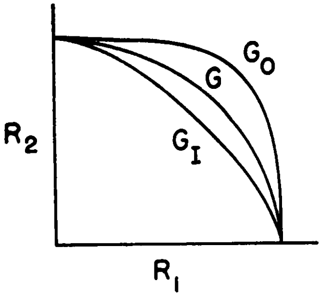

FIGURE 6

bounded by the middle curve G and the two axis segments. This curve can be described by a limiting expression involving mutual informations for long sequences of inputs and outputs.

In addition, we find an inner and outer bound, G_I and G_O, which are more easily evaluated, involving, as they do, only a maximizing process over single letters in the channel. G_O is the set of points (R_{12}, R_{21}) that may be obtained by assigning probabilities $P\{x_1, x_2\}$ to the input letters of the channel (an arbitrary joint distribution) and then evaluating

$$
\begin{aligned}
R_{12} &= E\left(\log \frac{P\{x_1|x_2, y_2\}}{P\{x_1|x_2\}}\right) = \sum_{x_1x_2y_2} P\{x_1x_2y_2\} \log \frac{P\{x_1|x_2, y_2\}}{P\{x_1|x_2\}} \\
R_{21} &= E\left(\log \frac{P\{x_2|x_1, y_1\}}{P\{x_2|x_1\}}\right),
\end{aligned} \tag{1}
$$

where $E(\mu)$ means expectation of μ. The inner bound G_I is found in a similar way but restricting the distribution to an independent one $P\{x_1, x_2\} = P\{x_1\}P\{x_2\}$. Then G_I is the *convex hull* of (R_{12}, R_{21}) points found under this restriction.

It is shown that in certain important cases these bounds are identical so the capacity region is then completely determined from the bounds. An example is also given (the binary multiplying channel) where there is a discrepancy between the bounds.

The three regions G_I, G and G_O are all convex and have the same intercepts on the axes. These intercepts are the capacities in the two directions when the other input letter is fixed at its best value [for example, x_1 is held at the value which maximizes R_{21} under variation of $P\{x_2\}$]. For any point inside G the error probabilities approach zero exponentially with the block length n. For any point outside G at least one of the error probabilities for the two codes will be bounded away from zero by a bound independent of the block length.

Finally, these results may be partially generalized to channels with certain types of memory. If there exists an internal state of the channel such that it is possible to return to this state in a bounded number of steps (regardless of previous transmission) then there will exist again a capacity region G with similar properties. A limiting expression is given determining this region.

3. Basic definitions

A *discrete memoryless two-way channel* consists of a set of transition probabilities $P\{y_1, y_2|x_1, x_2\}$ where x_1, x_2, y_1, y_2 all range over finite alphabets (not necessarily the same).

A *block code pair* of length n for such a channel with M_1 messages in the forward direction and M_2 in the reverse direction consists of two sets of n functions

$$\begin{aligned} &f_0(m_1), f_1(m_1, y_{11}), f_2(m_1, y_{11}, y_{12}), \cdots, f_{n-1}(m_1, y_{11}, \cdots, y_{1,n-1}) \\ &g_0(m_2), g_1(m_2, y_{21}), g_2(m_2, y_{21}, y_{22}), \cdots, g_{n-1}(m_2, y_{21}, \cdots, y_{2,n-1}). \end{aligned} \tag{2}$$

Here the f functions all take values in the x_1 alphabet and the g functions in the x_2 alphabet, while m_1 takes values from 1 to M_1 (the forward messages) and m_2 takes values from 1 to M_2 (the backward messages). Finally y_{1i}, for $i = 1, 2, \cdots, n-1$, takes values from the y_1 alphabet and similarly for y_{2i}. The f functions specify how the next input letter at terminal 1 should be chosen as determined by the message m_1 to be transmitted and the observed outputs $y_{11}, y_{12}, \cdots$ at terminal 1 up to the current time. Similarly the g functions determine how message m_2 is encoded as a function of the information available at each time in the process.

A *decoding system* for a block code pair of length n consists of a pair of functions $\phi(m_1, y_{11}, y_{12}, \cdots, y_{1n})$ and $\psi(m_2, y_{21}, y_{22}, \cdots, y_{2n})$. These functions take values from 1 to M_2 and 1 to M_1 respectively.

The decoding function φ represents a way of deciding on the original transmitted message from terminal 2 given the information available at terminal 1 at the end of a block of n received letters, namely, $y_{11}, y_{12}, \cdots, y_{1n}$ together with the transmitted message m_1 at terminal 1. Notice that the transmitted sequence $x_{11}, x_{12}, \cdots, x_{1n}$ although known at terminal 1 need not enter as an argument in the decoding function since it is determined (via the encoding functions) by m_1 and the received sequence.

We will assume, except when the contrary is stated, that all messages m_1 are equiprobable (probability $1/M_1$), that all messages m_2 are equiprobable (probability $1/M_2$), and that these events are statistically independent. We also assume that the successive operations of the channel are independent,

$$P\{y_{11}, y_{12}, \cdots, y_{1n}, y_{21}, y_{22}, \cdots, y_{2n}|x_{11}, x_{12}, \cdots, x_{1n}, x_{21}, x_{22}, \cdots, x_{2n}\} = \prod_{i=1}^{n} P\{y_{1i}, y_{2i}|x_{1i}, x_{2i}\}. \tag{3}$$

This is the meaning of the memoryless condition. This implies that the probability of a set of outputs from the channel, conditional on the corresponding inputs, is the same as this probability conditional on these inputs and any previous inputs.

The *signalling rates* R_1 and R_2 for a block code pair with M_1 and M_2 messages for the two directions are defined by

$$
\begin{aligned}
R_1 &= \frac{1}{n} \log M_1 \\
R_2 &= \frac{1}{n} \log M_2.
\end{aligned}
\tag{4}
$$

Given a code pair and a decoding system, together with the conditional probabilities defining a channel and our assumptions concerning message probability, it is possible, in principle, to compute error probabilities for a code. Thus one could compute for each message pair the probabilities of the various possible received sequences, if these messages were transmitted by the given coding functions. Applying the decoding functions, the probability of an incorrect decoding could be computed. This could be averaged over all messages for each direction to arrive at final error probabilities P_{e1} and P_{e2} for the two directions.

We will say that a point (R_1, R_2) belongs to the *capacity region* G of a given memoryless channel K if, given any $\epsilon > 0$, there exists a block code and decoding system for the channel with signalling rates R_1^* and R_2^* satisfying $|R_1 - R_1^*| < \epsilon$ and $|R_2 - R_2^*| < \epsilon$ and such that the error probabilities satisfy $P_{e1} < \epsilon$ and $P_{e2} < \epsilon$.

4. Average mutual information rates

The two-way discrete memoryless channel with finite alphabets has been defined by a set of transition probabilities $P\{y_1, y_2|x_1, x_2\}$. Here x_1 and x_2 are the input letters at terminals 1 and 2 and y_1 and y_2 are the output letters. Each of these ranges over its corresponding finite alphabet.

If a set of probabilities $P\{x_1\}$ is assigned (arbitrarily) to the different letters of the input alphabet for x_1 and another set of probabilities $P\{x_2\}$ to the alphabet for x_2 (these two taken statistically independent) then there will be definite corresponding probabilities for y_1 and y_2 and, in fact, for the set of four random variables x_1, x_2, y_1, y_2, namely,

$$
\begin{aligned}
P\{x_1, x_2, y_1, y_2\} &= P\{x_1\}P\{x_2\}P\{y_1, y_2|x_1, x_2\} \\
P\{y_1\} &= \sum_{x_1,x_2,y_2} P\{x_1, x_2, y_1, y_2\},
\end{aligned}
\tag{5}
$$

and so forth.

Thinking first intuitively, and in analogue to the one-way channel, we might think of the rate of transmission from x_1 to the terminal 2 as given by $H(x_1) - H(x_1|x_2, y_2)$, that is, the uncertainty or entropy of x_1 less its entropy conditional on what is available at terminal 2, namely, y_2 and x_2. Thus, we might write

$$R_{12} = H(x_1) - H(x_1|x_2, y_2) = E\left[\log \frac{P\{x_1, x_2, y_2\}}{P\{x_1\}P\{x_2, y_2\}}\right] = E\left[\log \frac{P\{x_1|x_2, y_2\}}{P\{x_1\}}\right] \tag{6}$$

$$R_{21} = H(x_2) - H(x_2|x_1, y_1) = E\left[\log \frac{P\{x_1, x_2, y_1\}}{P\{x_2\}P\{x_1, y_1\}}\right] = E\left[\log \frac{P\{x_2|x_1, y_1\}}{P\{x_2\}}\right]. \tag{7}$$

These are the average mutual informations with the assigned input probabilities between the input at one terminal and the input-output pair at the other terminal. We might expect, then, that by suitable coding it should be possible to send in the two directions *simultaneously* with arbitrarily small error probabilities and at rates arbitrarily close to R_{12} and R_{21}. The codes would be based on these probabilities $P\{x_1\}$ and $P\{x_2\}$ in generalization of the one-way channel. We will show that in fact it is possible to find codes based on the probabilities $P\{x_1\}$ and $P\{x_2\}$ which do this.

However the capacity region may be larger than the set of rates available by this means. Roughly speaking, the difference comes about because of the probability of having x_1 and x_2 dependent random variables. In this case the appropriate mutual informations are given by $H(x_2|x_1) - H(x_2|x_1, y_1)$ and $H(x_1|x_2) - H(x_1|x_2, y_2)$. The above expressions for R_{21} and R_{12} of course reduce to these when x_1 and x_2 are independent.

5. The distribution of information

The method we follow is based on random codes using techniques similar to those used in [1] for the one-way channel. Consider a sequence of n uses of the channel or, mathematically, the product probability space. The inputs are $X_1 = (x_{11}, x_{12}, \cdots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \cdots, x_{2n})$ and the outputs $Y_1 = (y_{11}, y_{12}, \cdots, y_{1n})$ and $Y_2 = (y_{21}, y_{22}, \cdots, y_{2n})$, that is, sequences of n choices from the corresponding alphabets.

The conditional probabilities for these blocks are given by

$$P\{Y_1, Y_2|X_1, X_2\} = \prod_k P\{y_{1k}, y_{2k}|x_{1k}, x_{2k}\}. \tag{8}$$

This uses the assumption that the channel is memoryless, or successive operations independent. We also associate a probability measure with input blocks X_1 and X_2 given by the product measure of that taken for x_1, x_2. Thus

$$P\{X_1\} = \prod_k P\{x_{1k}\}$$
$$P\{X_2\} = \prod_k P\{x_{2k}\}. \tag{9}$$

It then follows that other probabilities are also the products of those for the individual letters. Thus, for example,

$$P\{X_1, X_2, Y_1, Y_2\} = \prod_k P\{x_{1k}, x_{2k}, y_{1k}, y_{2k}\}$$
$$P\{X_2|X_1, Y_1\} = \prod_k P\{x_{2k}|x_{1k}, y_{1k}\}. \tag{10}$$

The (unaveraged) mutual information between, say, X_1 and the pair X_2, Y_2 may be written as a sum, as follows:

$$\begin{aligned} I(X_1; X_2, Y_2) &= \log \frac{P\{X_1, X_2, Y_2\}}{P\{X_1\}P\{X_2, Y_2\}} = \log \frac{\prod_k P\{x_{1k}, x_{2k}, y_{2k}\}}{\prod_k P\{x_{1k}\} \prod_k P\{x_{2k}, y_{2k}\}} \\ &= \sum_k \log \frac{P\{x_{1k}, x_{2k}, y_{2k}\}}{P\{x_{1k}\}P\{x_{2k}, y_{2k}\}} \\ I(X_1; X_2, Y_2) &= \sum_k I(x_{1k}; x_{2k}, y_{2k}). \end{aligned} \tag{11}$$

Thus, the mutual information is, as usual in such independent situations, the sum of the individual mutual informations. Also, as usual, we may think of the mutual information as a random variable. Here $I(X_1; X_2, Y_2)$ takes on different values with probabilities given by $P\{X_1, X_2, Y_2\}$. The *distribution function* for $I(X_1; X_2, Y_2)$ will be denoted by $\rho_{12}(Z)$ and similarly for $I(X_2; X_1, Y_1)$

$$\begin{aligned} \rho_{12}(Z) &= P\{I(X_1; X_2, Y_2) \leqq Z\} \\ \rho_{21}(Z) &= P\{I(X_2; X_1, Y_1) \leqq Z\}. \end{aligned} \tag{12}$$

Since each of the random variables $I(X_1; X_2, Y_2)$ and $I(X_2; X_1, Y_1)$ is the sum of n independent random variables, each with the same distribution, we have the familiar statistical situation to which one may apply various central limit theorems and laws of large numbers. The mean of the distributions ρ_{12} and ρ_{21} will be nR_{12} and nR_{21} respectively and the variances n times the corresponding variances for one letter. As $n \to \infty$, $\rho_{12}[n(R_{12} - \epsilon)] \to 0$ for any fixed $\epsilon > 0$, and similarly for ρ_{21}. In fact, this approach is exponential in n; $\rho_{12}[n(R_{12} - \epsilon)] \leqq \exp[-A(\epsilon)n]$.

6. Random codes for the two-way channel

After these preliminaries we now wish to prove the existence of codes with certain error probabilities bounded by expressions involving the distribution functions ρ_{12} and ρ_{21}.

We will construct an ensemble of codes or, more precisely, of *code pairs*, one code for the $1 - 2$ direction and another for the $2 - 1$ direction. Bounds will be established on the error probabilities P_{e1} and P_{e2} *averaged over the ensemble*, and from these will be shown the existence of *particular* codes in the ensemble with related bounds on their error probabilities.

The random ensemble of code pairs for such a two-way channel with M_1 words in the $1 - 2$ code and M_2 words in the $2 - 1$ code is constructed as follows. The M_1 integers $1, 2, \cdots, M_1$ (the messages of the first code) are mapped in all possible ways into the set of input words X_1 of length n. Similarly the integers $1, 2, \cdots, M_2$ (the messages of the second code) are mapped in all possible ways into the set of input words X_2 of length n.

If there were a_1 possible input *letters* at terminal 1 and a_2 input *letters* at terminal 2, there will be a_1^n and a_2^n input *words* of length n and $a_1^{nM_1}$ *mappings* in the first code and $a_2^{nM_2}$ in the second code. We consider all pairs of these codes, a total of $a_1^{nM_1}a_2^{nM_2}$ pairs.

Each code pair is given a weighting, or probability, equal to the probability of occurrence of that pair if the two mappings were done independently and an integer is mapped into a word with the assigned probability of that word. Thus, a code pair is given a weighting equal to the product of the probabilities associated with all the input words that the integers are mapped into for both codes. This set of code pairs with these associated probabilities we call the *random ensemble of code pairs* based on the assigned probabilities $P\{X_1\}$ and $P\{X_2\}$.

Any particular code pair of the ensemble could be used to transmit information, if we agreed upon a method of decoding. The method of decoding will here consist of two functions $\phi(X_1, Y_1)$ and $\psi(X_2, Y_2)$, a special case of that defined above. Here X_1 varies over the input words of length n at terminal 1, and Y_1 over the possible received blocks of length n. The function ϕ takes values from 1 to M_2 and represents the decoded message for a received Y_1 if X_1 was transmitted. (Of course, X_1 is used in the decoding procedure in general since it may influence Y_1 and is, therefore, pertinent information for best decoding.)

Similarly, $\psi(X_2, Y_2)$ takes values from 1 to M_1 and is a way of deciding on the transmitted message m_1 on the basis of information available at terminal 2. It should be noted here that the decoding functions, ϕ and ψ, need not be the same for all code pairs in the ensemble.

We also point out that the encoding functions for our random ensemble are more specialized than the general case described above. The sequence of input letters X_1 for a given message m_1 do not depend on the received letters at terminal 1. In any particular code of the ensemble there is a strict mapping from messages to input sequences.

Given an ensemble of code pairs as described above and decoding functions, one could compute for each particular code pair two error probabilities for the two codes: P_{e1}, the probability of error in decoding the first code, and P_{e2} that for the second. Here we are assuming that the different messages in the first code occur with equal probability $1/M_1$, and similarly for the second.

By the *average error probabilities for the ensemble* of code pairs we mean the averages $E(P_{e1})$ and $E(P_{e2})$ where each probability of error for a particular code is weighted according to the weighting factor or probability associated with the code pair. We wish to describe a particular method of decoding, that is, a choice of ϕ and ψ, and then place upper bounds on these average error probabilities for the ensemble.

7. Error probability for the ensemble of codes

THEOREM 1. *Suppose probability assignments $P\{X_1\}$ and $P\{X_2\}$ in a discrete memoryless two-way channel produce information distribution functions $\rho_{12}(Z)$ and $\rho_{21}(Z)$. Let $M_1 = \exp(R_1 n)$ and $M_2 = \exp(R_2 n)$ be arbitrary integers and θ_1 and θ_2 be arbitrary positive numbers. Then the random ensemble of code pairs with M_1 and M_2 messages has (with appropriate decoding functions) average error probabilities bounded as follows:*

$$\begin{aligned} E(P_{e1}) &\leqq \rho_{12}[n(R_1 + \theta_1)] + e^{-n\theta_1} \\ E(P_{e2}) &\leqq \rho_{21}[n(R_2 + \theta_2)] + e^{-n\theta_2}. \end{aligned} \tag{13}$$

There will exist in the ensemble at least one code pair whose individual *error probabilities are bounded by* two times *these expressions, that is, satisfying*

$$\begin{aligned} P_{e1} &\leqq 2\rho_{12}[n(R_1 + \theta_1)] + 2e^{-n\theta_1} \\ P_{e2} &\leqq 2\rho_{21}[n(R_2 + \theta_2)] + 2e^{-n\theta_2}. \end{aligned} \tag{14}$$

This theorem is a generalization of theorem 1 in [1] which gives a similar bound on P_e for a one-way channel. The proof for the two-way channel is a generalization of that proof.

The statistical situation here is quite complex. There are several statistical events involved: the choice of messages m_1 and m_2, the choice of code pair in the ensemble of code pairs, and finally the statistics of the channel itself which produces the output words Y_1 and Y_2 according to $P\{Y_1, Y_2|X_1, X_2\}$. The ensemble error probabilities we are calculating are averages over *all* these statistical events.

We first define decoding systems for the various codes in the ensemble. For a given θ_2, define for each pair X_1, Y_1 a corresponding set of words in the X_2 space denoted by $S(X_1, Y_1)$ as follows:

$$S(X_1, Y_1) = \left\{X_2 \middle| \log \frac{P\{X_1, X_2, Y_2\}}{P\{X_2\}P\{X_1, Y_1\}} > n(R_2 + \theta_2)\right\}. \tag{15}$$

That is, $S(X_1, Y_1)$ is the set of X_2 words whose mutual information with the particular pair (X_1, Y_1) exceeds a certain level, $n(R_2 + \theta_2)$. In a similar way, we define a set $S'(X_2, Y_2)$ of X_1 words for each X_2, Y_2 pair as follows:

$$S'(X_2, Y_2) = \left\{X_1 \middle| \log \frac{P\{X_1, X_2, Y_1\}}{P\{X_1\}P\{X_2, Y_2\}} > n(R_1 + \theta_1)\right\}. \tag{16}$$

We will use these sets S and S' to define the decoding procedure and to aid in overbounding the error probabilities. The decoding process will be as follows. In any particular code pair in the random ensemble, suppose message m_1 is sent and this is mapped into input word X_1. Suppose that Y_1 is received at terminal 1 in the corresponding block of n letters. Consider the subset of X_2 words, $S(X_1, Y_1)$. Several situations may occur. (1) There is no message m_2 mapped into the subset $S(X_1, Y_1)$ for the code pair in question. In this case, X_1, Y_1 is decoded (conventionally) as message number one. (2) There is exactly one message mapped into the subset. In this case, we decode as this particular message. (3) There are more than one such messages. In this case, we decode as the smallest numbered such message.

The error probabilities that we are estimating would normally be thought of as calculated in the following manner. For each code pair one would calculate the error probabilities for all messages m_1 and m_2, and from their averages get the error probabilities for that code pair. Then these error probabilities are averaged over the ensemble of code pairs, using the appropriate weights or probabilities. We may, however, interchange this order of averaging. We may consider the cases where a particular $\overline{m}_1$ and $\overline{m}_2$ are the messages and these are mapped into particular $\overline{X}_1$ and $\overline{X}_2$, and the received words are $\overline{Y}_1$ and $\overline{Y}_2$. There is still, in the statistical picture, the range of possible code pairs, that is, mappings of the other $M_1 - 1$ messages for one code and $M_2 - 1$ for the other. We wish to show that, averaged over this subset of codes, the probabilities of any of these messages being mapped into subsets $S'(\overline{X}_2, \overline{Y}_2)$ and $S(\overline{X}_1, \overline{Y}_1)$ respectively do not exceed $\exp(-n\theta_1)$ and $\exp(-n\theta_2)$.

Note first that if X_1 belongs to the set $S'(\overline{X}_2, \overline{Y}_2)$ then by the definition of this set

$$\log \frac{P\{X_1, \overline{X}_2, \overline{Y}_2\}}{P\{X_1\}P\{\overline{X}_2, \overline{Y}_2\}} > n(R_1 + \theta_1) \tag{17}$$
$$P\{X_1|\overline{X}_2, \overline{Y}_2\} > P\{X_1\}e^{n(R_1+\theta_1)}.$$

Now sum each side over the set of X_1 belonging to $S'(\overline{X}_2, \overline{Y}_2)$ to obtain

$$1 \geqq \sum_{X_1 \in S'(\overline{X}_2, \overline{Y}_2)} P\{X_1|\overline{X}_2, \overline{Y}_2\} > e^{n(R_1+\theta_1)} \sum_{X_1 \in S'(\overline{X}_2, \overline{Y}_2)} P\{X_1\}. \tag{18}$$

The left inequality here holds since a sum of disjoint probabilities cannot exceed one. The sum on the right we may denote by $P\{S'(\overline{X}_2, \overline{Y}_2)\}$. Combining the first and last members of this relation

$$P\{S'(\overline{X}_2, \overline{Y}_2)\} < e^{-n(R_1+\theta_1)}. \tag{19}$$

That is, the total probability associated with any set $S'(\overline{X}_2, \overline{Y}_2)$ is bounded by an expression involving n, R_1 and θ_1 but *independent* of the particular $\overline{X}_2$, $\overline{Y}_2$.

Now recall that the messages were mapped independently into the input words using the probabilities $P\{X_1\}$ and $P\{X_2\}$. The probability of a particular message being mapped into $S'(\overline{X}_2, \overline{Y}_2)$ in the ensemble of code pairs is just $P\{S'(\overline{X}_2, \overline{Y}_2)\}$. The probability of being in the complementary set is $1 -$

$P\{S'(\overline{X}_2, \overline{Y}_2)\}$. The probability that *all* messages other than $\overline{m}_1$ will be mapped into this complementary set is

$$(20)\qquad \begin{aligned}[1 - P\{S'(\overline{X}_2, \overline{Y}_2)\}]^{M_1-1} &\geqq 1 - (M_1 - 1)P\{S'(\overline{X}_2, \overline{Y}_2)\}\\ &\geqq 1 - M_1P\{S'(\overline{X}_2, \overline{Y}_2)\}\\ &\geqq 1 - M_1e^{-n(R_1+\theta_1)}\\ &= 1 - e^{-n\theta_1}.\end{aligned}$$

Here we used the inequality $(1 - x)^p \geqq 1 - px$, the relation (19) and finally the fact that $M_1 = \exp(nR_1)$.

We have established, then, that in the subset of cases being considered ($\overline{m}_1$ and $\overline{m}_2$ mapped into $\overline{X}_1$ and $\overline{X}_2$ and received as $\overline{Y}_1$ and $\overline{Y}_2$), with probability at least $1 - \exp(-n\theta_1)$, there will be no other messages mapped into $S'(\overline{X}_2, \overline{Y}_2)$. A similar calculation shows that with probability exceeding $1 - \exp(-n\theta_2)$ there will be no other messages mapped into $S(\overline{X}_1, \overline{Y}_1)$. These bounds, as noted, are independent of the particular $\overline{X}_1$, $\overline{Y}_1$ and $\overline{X}_2$, $\overline{Y}_2$.

We now bound the probability of the actual message $\overline{m}_1$ being within the subset $S'(\overline{X}_2, \overline{Y}_2)$. Recall that from the definition of $\rho_{12}(Z)$

$$(21)\qquad \rho_{12}[n(R_1 + \theta_1)] = P\left\{\log \frac{P\{X_1, X_2, Y_2\}}{P\{X_1\}P\{X_2, Y_2\}} \leqq n(R_1 + \theta_1)\right\}.$$

In the ensemble of code pairs a message $\overline{m}_1$, say, is mapped into words X_1 with probabilities just equal to $P\{X_1\}$. Consequently, the probability in the full ensemble of code pairs, message choices and channel statistics, that the actual message is mapped into $S'(\overline{X}_2, \overline{Y}_2)$ is precisely $1 - \rho_{12}[n(R_1 + \theta_1)]$.

The probability that the actual message is mapped *outside* $S'(\overline{X}_2, \overline{Y}_2)$ is therefore given by $\rho_{12}[n(R_1 + \theta_1)]$ and the probability that there are any other messages mapped into $S'(\overline{X}_2, \overline{Y}_2)$ is bounded as shown before by $\exp(-n\theta_1)$. The probability that *either* of these events is true is then certainly bounded by $\rho_{12}[n(R_1 + \theta_1)] + \exp(-n\theta_1)$; but this is then a bound on $E(P_{e1})$, since if neither event occurs the decoding process will correctly decode.

Of course, the same argument with interchanged indices gives the corresponding bound for $E(P_{e2})$. This proves the first part of the theorem.

With regard to the last statement of the theorem, we will first prove a simple combinatorial lemma which is useful not only here but in other situations in coding theory.

LEMMA. *Suppose we have a set of objects $B_1, B_2, \cdots, B_n$ with associated probabilities $P_1, P_2, \cdots, P_n$, and a number of numerically valued properties (functions) of the objects $f_1, f_2, \cdots, f_d$. These are all nonnegative, $f_i(B_j) \geqq 0$, and we know the averages A_i of these properties over the objects,*

$$(22)\qquad \sum_j P_jf_i(B_j) = A_i, \qquad i = 1, 2, \cdots, d.$$

Then there exists an object B_p for which

$$(23)\qquad f_i(B_p) \leqq dA_i, \qquad i = 1, 2, \cdots, d.$$

More generally, given any set of $K_i > 0$ *satisfying* $\sum_{i=1}^{d} (1/K_i) \leqq 1$, *then there exists an object* B_p *with*

$$f_i(B_p) \leqq K_i A_i, \qquad i = 1, 2, \cdots, d. \tag{24}$$

PROOF. The second part implies the first by taking $K_i = d$. To prove the second part let Q_i be the total probability of objects B for which $f_i(B) > K_i A_i$. Now the average $A_i > Q_i K_i A_i$ since $Q_i K_i A_i$ is contributed by the B_i with $f(B) > K_i A_i$ and all the remaining B have f_i values $\geqq 0$. Hence

$$Q_i < \frac{1}{K_i}, \qquad i = 1, 2, \cdots, d. \tag{25}$$

The total probability Q of objects violating *any* of the conditions is less than or equal to the sum of the individual Q_i, so that

$$Q < \sum_{i=1}^{d} \frac{1}{K_i} \leqq 1. \tag{26}$$

Hence there is at least one object not violating any of the conditions, concluding the proof.

For example, suppose we know that a room is occupied by a number of people whose average age is 40 and average height 5 feet. Here $d = 2$, and using the simpler form of the theorem we can assert that there is someone in the room not over 80 years old and not over ten feet tall, even though the room might contain aged midgets and youthful basketball players. Again, using $K_1 = 8/3$, $K_2 = 8/5$, we can assert the existence of an individual not over 8 feet tall and not over 106 2/3 years old.

Returning to the proof of theorem 1, we can now establish the last sentence. We have a set of objects, the code pairs, and two properties of each object, its error probability P_{e1} for the code from 1 to 2 and its error probability P_{e2} for the code from 2 to 1. These are nonnegative and their averages are bounded as in the first part of theorem 1. It follows from the combinatorial result that there exists at least one particular code pair for which simultaneously

$$\begin{aligned} P_{e1} &\leqq 2\{\rho_{12}[n(R_1 + \theta_1)] + e^{-n\theta_1}\} \\ P_{e2} &\leqq 2\{\rho_{21}[n(R_2 + \theta_2)] + e^{-n\theta_2}\}. \end{aligned} \tag{27}$$

This concludes the proof of theorem 1.

It is easily seen that this theorem proves the possibility of code pairs arbitrarily close in rates R_1 and R_2 to the mean mutual information per letter R_{12} and R_{21} for any assigned $P\{x_1\}$ and $P\{x_2\}$ and with arbitrarily small probability of error. In fact, let $R_{12} - R_1 = R_{21} - R_2 = \epsilon > 0$ and in the theorem take $\theta_1 = \theta_2 = \epsilon/2$. Since $\rho_{12}[n(R_{12} - \epsilon/2)] \to 0$ and, in fact, exponentially fast with n (the distribution function $\epsilon n/2$ to the left of the mean, of a sum of n random variables) the bound on P_{e1} approaches zero with increasing n exponentially fast. In a similar way, so does the bound on P_{e2}. By choosing, then, a sequence of the M_1 and M_2 for increasing n which approach the desired rates R_1 and R_2 from below, we obtain the desired result, which may be stated as follows.

THEOREM 2. *Suppose in a two-way memoryless channel K an assignment of probabilities to the input letters* $P\{x_1\}$ *and* $P\{x_2\}$ *gives average mutual informations in the two directions*

$$
\begin{aligned}
R_{12} &= E\left(\log \frac{P\{x_1|x_2, y_2\}}{P\{x_1\}}\right) \\
R_{21} &= E\left(\log \frac{P\{x_2|x_1, y_1\}}{P\{x_2\}}\right).
\end{aligned}
\tag{28}
$$

Then given $\epsilon > 0$ *there exists a code pair for all sufficiently large block length* n *with signalling rates in the two directions greater than* $R_{12} - \epsilon$ *and* $R_{21} - \epsilon$ *respectively, and with error probabilities* $P_{e1} \leqq \exp[-A(\epsilon)n]$, $P_{e2} \leqq \exp[-A(\epsilon)n]$ *where* $A(\epsilon)$ *is positive and independent of* n.

By trying different assignments of letter probabilities and using this result, one obtains various points in the capacity region. Of course, to obtain the best rates available from this theorem we should seek to maximize these rates. This is most naturally done à la Lagrange by maximizing $R_{12} + \lambda R_{21}$ for various positive λ.

8. The convex hull G_1 as an inner bound of the capacity region

In addition to the rates obtained this way we may construct codes which are *mixtures* of codes obtained by this process. Suppose one assignment $P\{x_1\}$, $P\{x_2\}$ gives mean mutual informations R_{12}, R_{21} and a second assignment $P'\{x_1\}$, $P'\{x_2\}$ gives R'_{12}, R'_{21}. Then we may find a code of (sufficiently large) length n for the first assignment with error probabilities $< \delta$ and rate discrepancy less than or equal to ϵ and a second code of length n' based on $P'\{x_1\}$, $P'\{x_2\}$ with the same δ and ϵ. We now consider the code of length $n + n'$ with $M_1M'_1$ words in the forward direction, and $M_2M'_2$ in the reverse, consisting of all words of the first code followed by all words for the same direction in the second code. This has signalling rates R^*_1 and R^*_2 equal to the weighted average of rates for the original codes $[R^*_1 = nR_1/(n + n') + n'R'_1/(n + n'); R^*_2 = nR_2/(n + n') + n'R'_2/(n + n')]$ and consequently its rates are within ϵ of the weighted averages, $|R^*_1 - nR_{12}/(n + n') - n'R'_{12}/(n + n')| < \epsilon$ and similarly. Furthermore, its error probability is bounded by 2δ, since the probability of either of two events (an error in either of the two parts of the code) is bounded by the sum of the original probabilities. We can construct such a mixed code for *any* sufficiently large n and n'. Hence by taking these large enough we can approach any weighted average of the given rates and simultaneously approach zero error probability exponentially fast. It follows that *we can annex to the set of points found by the assignment of letter probabilities all points in the convex hull of this set.* This actually does add new points in some cases as our example, of a channel (table I) with incompatible transmission in the two directions, shows. By mixing the codes for assignments which give the points (0, 1) and (1, 0) in equal proportions, we obtain the point $(1/2, 1/2)$. There is no single letter assignment giving this pair of rates. We may summarize as follows.

THEOREM 3. *Let G_I be the convex hull of points (R_{12}, R_{21})*

$$(29)\qquad \begin{aligned} R_{12} &= E\left(\log \frac{P\{x_1|x_2, y_2\}}{P\{x_1\}}\right) \\ R_{21} &= E\left(\log \frac{P\{x_2|x_1, y_1\}}{P\{x_2\}}\right) \end{aligned}$$

when $P\{x_1\}$ and $P\{x_2\}$ are given various probability assignments. All points of G_I are in the capacity region. For any point (R_1, R_2) in G_I and any $\epsilon > 0$ we can find codes whose signalling rates are within ϵ of R_1 and R_2 and whose error probabilities in both directions are less than $\exp[-A(\epsilon)n]$ for all sufficiently large n, and some positive $A(\epsilon)$.

It may be noted that the convex hull G_I in this theorem is a closed set (contains all its limit points). This follows from the continuity of R_{12} and R_{21} as functions of the probability assignments $P\{x_1\}$ and $P\{x_2\}$. Furthermore if G_I contains a point (R_1, R_2) it contains the projections $(R_1, 0)$ and $(0, R_2)$. This will now be proved.

It will clearly follow if we can show that the projection of any point obtained by a letter probability assignment is also in G_I. To show this, suppose $P\{x_1\}$ and $P\{x_2\}$ give the point (R_{12}, R_{21}). Now R_{12} is the average of the various particular R_{12} when x_2 is given various particular values. Thus

$$(30)\qquad R_{12} = \sum_{x_2} P\{x_2\} \sum_{x_1, y_2} P\{x_1, y_2|x_2\} \log \frac{P\{x_1|x_2, y_2\}}{P\{x_1\}}.$$

There must exist, then, a particular x_2, say x_2^*, for which the inner sum is at least as great as the average, that is, for which

$$(31)\qquad \sum_{x_1, y_2} P\{x_1, y_2|x_2^*\} \log \frac{P\{x_1|x_2^*, y_2\}}{P\{x_1\}} \geqq \sum_{x_2} P\{x_2\} \sum_{x_1, y_2} P\{x_1, y_2|x_2\} \log \frac{P\{x_1|x_2, y_2\}}{P\{x_1\}}.$$

The assignment $P\{x_1|x_2^*\}$ for letter probabilities x_1 and the assignment $P\{x_2\} = 1$ if $x_2 = x_2^*$ and 0 otherwise, now gives a point on the horizontal axis below or to the right of the projection of the given point R_{12}, R_{21}. Similarly, we can find an x_1^* such that the assignment $P\{x_2|x_1^*\}$ for x_2 and $P\{x_1^*\} = 1$ gives a point on the vertical axis equal to or above the projection of R_{12}, R_{21}. Note also that the assignment $P\{x_1^*\} = 1, P\{x_2^*\} = 1$ gives the point $(0, 0)$. By suitable mixing of codes obtained for these four assignments one can approach any point of the quadrilateral defined by the corresponding pairs of rates, and in particular any point in the rectangle subtended by R_{12}, R_{21}. It follows from these remarks that the convex hull G_I is a region of the form shown typically in figure 7 bounded by a horizontal segment, a convex curve, a vertical segment, and two segments of the axes. Of course, any of these parts may be of zero length.

The convex hull G_I is, as we have seen, inside the capacity region and we will refer to it as the *inner bound.*

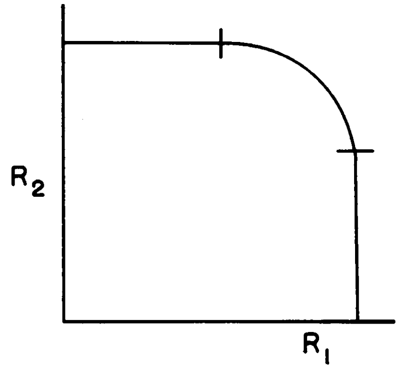

FIGURE 7

It is of some interest to attempt a sharper evaluation of the rate of improvement of error probability with increasing code length n. This is done in the appendix and leads to a generalization of theorem 2 in [1]. The bound we arrive at is based on logarithms of moment generating functions.

9. An outer bound on the capacity region

While in some cases the convex hull G_I, the inner bound defined above, is actually the capacity region this is not always the case. By an involved calculation R. G. Gallager has shown that in the binary multiplying channel the inner bound is strictly interior to the capacity region. However a *partial* converse to theorem 3 and an *outer* bound on the capacity region can be given. Suppose we have a code starting at time zero with messages m_1 and m_2 at the two terminals. After n operations of the channel, let Y_1 and Y_2 be the received blocks at the two terminals (sequences of n letters), and let x_1, x_2, y_1, y_2 be the next transmitted and received letters. Consider the change in "equivocation" of message at the two terminals due to the next received letter. At terminal 2, for example, this change is (making some obvious reductions)

$$\begin{aligned}(32)\qquad \Delta &= H(m_1|m_2, Y_2) - H(m_1|m_2, Y_2, y_2)\\ &= E\left[\log \frac{P\{m_2, Y_2\}}{P\{m_1, m_2, Y_2\}}\right] - E\left[\log \frac{P\{m_2, Y_2, y_2\}}{P\{m_1, m_2, Y_2, y_2\}}\right]\\ &= E\left[\log \frac{P\{y_2|m_1, m_2, Y_2\}}{P\{y_2|x_2\}} \frac{P\{y_2|x_2\}}{P\{y_2|Y_2, m_2\}}\right].\end{aligned}$$

Now $H(y_2|m_1, m_2, Y_2) \geqq H(y_2|m_1, m_2, Y_1, Y_2) = H(y_2|x_1, x_2)$ since adding a conditioning variable cannot increase an entropy and since $P\{y_2|m_1, m_2, Y_1, Y_2\} = P\{y_2|x_1, x_2\}$.

Also $H(y_2|x_2) \geqq H(y_2|Y_2, m_2)$ since x_2 is a function of Y_2 and m_2 by the coding function. Therefore

$$\Delta \leqq E\left(\log \frac{P\{y_2|x_1, x_2\}}{P\{y_2|x_2\}}\right) + H(y_2|Y_2, m_2) - H(y_2|x_2) \tag{33}$$

$$\Delta \leqq E\left(\log \frac{P\{y_2|x_1, x_2\}}{P\{y_2|x_2\}}\right) = E\left(\log \frac{P\{y_2, x_1, x_2\}P\{x_2\}}{P\{x_2, y_2\}P\{x_1, x_2\}}\right) \tag{34}$$

$$= E\left(\log \frac{P\{x_1|x_2, y_2\}}{P\{x_1|x_2\}}\right).$$

This would actually lead to a converse of theorem 1 if we had independence of the random variables x_1 and x_2. This last expression would then reduce to $E[\log (P\{x_1|x_2, y_2\}/P\{x_1\})]$. Unfortunately in a general code they are not necessarily independent. In fact, the next x_1 and x_2 may be functionally related to received X and Y and hence dependent.

We may, however, at least obtain an outer bound on the capacity surface. Namely, the above inequality together with the similar inequality for the second terminal imply that the vector change in equivocation due to receiving another letter must be a vector with components bounded by

$$E\left(\log \frac{P\{x_1|x_2, y_2\}}{P\{x_1|x_2\}}\right), E\left(\log \frac{P\{x_2|x_1, y_1\}}{P\{x_2|x_1\}}\right) \tag{35}$$

for some $P\{x_1, x_2\}$. Thus the vector change is included in the convex hull of all such vectors G_O (when $P\{x_1, x_2\}$ is varied).

In a code of length n, the *total* change in equivocation from beginning to end of the block cannot exceed the sum of n vectors from this convex hull. Thus this sum will lie in the convex hull nG_O, that is, G_O expanded by a factor n.

Suppose now our given code has signalling rates $R_1 = (1/n) \log M_1$ and $R_2 = (1/n) \log M_2$. Then the initial equivocations of message are nR_1 and nR_2. Suppose the point (nR_1, nR_2) is outside the convex hull nG_O with nearest distance $n\epsilon_1$, figure 8. Construct a line L passing through the nearest point of nG_O and

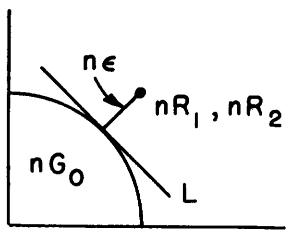

FIGURE 8

perpendicular to the nearest approach segment with nG_O on one side (using the fact that nG_O is a convex region). It is clear that for any point (nR_1^*, nR_2^*) on the nG_O side of L and particularly for any point of nG_O, that we have $|nR_1 - nR_1^*| + |nR_2 - nR_2^*| \geqq n\epsilon$ (since the shortest distance is $n\epsilon$) and furthermore at least one of the $nR_1 - nR_1^*$ and $nR_2 - nR_2^*$ is at least $n\epsilon/\sqrt{2}$. (In a right triangle at least one leg is as great as the hypotenuse divided by $\sqrt{2}$.)

Thus after n uses of the channel, if the signalling rate pair R_1, R_2 is distance ϵ outside the convex hull G_O, at least one of the two final equivocations is at least $\epsilon/\sqrt{2}$, where all equivocations are on a per second basis. Thus for signalling rates ϵ outside of G_O the equivocations per second are bounded from below independent of the code length n. This implies that the error probability is also bounded from below, that is, at least one of the two codes will have error probability $\geqq f(\epsilon) > 0$ independent of n, as shown in [2], appendix.

To summarize, the capacity region G is included in the convex hull G_O of all points R_{12}, R_{21}

$$(36)\qquad \begin{aligned} R_{12} &= E\left[\log \frac{P\{x_1|x_2, y_2\}}{P\{x_1|x_2\}}\right] \\ R_{21} &= E\left[\log \frac{P\{x_2|x_1, y_1\}}{P\{x_2|x_1\}}\right] \end{aligned}$$

when arbitrary joint probability assignments $P\{x_1, x_2\}$ are made.

Thus the inner bound G_I and the outer bound G_O are both found by the same process, assigning input probabilities, calculating the resulting average mutual informations R_{12} and R_{21} and then taking the convex hull. The only difference is that for the outer bound a general joint assignment $P\{x_1, x_2\}$ is made, while for the inner bound the assignments are restricted to independent $P\{x_1\}P\{x_2\}$.

We now develop some properties of the outer bound.

10. The concavity of R_{12} and R_{21} as functions of $P(x_1,x_2)$

THEOREM 4. *Given the transition probabilities $P\{y_1, y_2|x_1, x_2\}$ for a channel K, the rates*

$$(37)\qquad \begin{aligned} R_{12} &= E\left[\log \frac{P\{y_2|x_1, x_2\}}{P\{y_2|x_2\}}\right] \\ R_{21} &= E\left[\log \frac{P\{y_1|x_1, x_2\}}{P\{y_1|x_1\}}\right] \end{aligned}$$

are concave downward functions of the assigned input probabilities $P\{x_1, x_2\}$. For example, $R_{12}(P_1\{x_1, x_2\}/2 + P_2\{x_1, x_2\}/2) \geqq R_{12}(P_1\{x_1, x_2\})/2 + R_{12}(P_2\{x_1, x_2\})/2$.

This concave property is a generalization of that given in [3] for a one-way channel. To prove the theorem it suffices, by known results in convex functions, to show that

(38)

$$R_{12}\left(\frac{1}{2} P_1\{x_1, x_2\} + \frac{1}{2} P_2\{x_1, x_2\}\right) \geqq \frac{1}{2} R_{12}(P_1\{x_1, x_2\}) + \frac{1}{2} R_{12}(P_2\{x_1, x_2\}).$$

But $R_{12}(P_1\{x_1, x_2\})$ and $R_{12}(P_2\{x_1, x_2\})$ may be written

$$R_{12}(P_1\{x_1, x_2\}) = \sum_{x_2} P_1\{x_2\} \sum_{x_1,y_2} P_1\{x_1, y_2|x_2\} \log \frac{P_2\{y_2|x_1, x_2\}}{P_1\{y_2|x_2\}} \tag{39}$$

$$R_{12}(P_2\{x_1, x_2\}) = \sum_{x_2} P_2\{x_2\} \sum_{x_1,y_2} P_2\{x_1, y_2|x_2\} \log \frac{P_2\{y_2|x_1, x_2\}}{P_2\{y_2|x_2\}}. \tag{40}$$

Here the subscripts 1 on probabilities correspond to those produced with the probability assignment $P_1\{x_1, x_2\}$ to the inputs, and similarly for the subscript 2. The inner sum $\sum_{x_1,y_2} P_1\{x_1, y_2|x_2\} \log (P_1\{y_2|x_1, x_2\}/P_1\{y_2|x_2\})$ may be recognized as the rate for the channel from x_1 to y_2 conditional on x_2 having a particular value and with the x_1 assigned probabilities corresponding to its conditional probability according to $P_1\{x_1, x_2\}$.

The corresponding inner sum with assigned probabilities $P_2\{x_1, x_2\}$ is $\sum_{x_1,y_2} P_2\{x_1, y_2|x_2\} \log (P_2\{y_2|x_1, x_2\}/P_2\{y_2|x_2\})$, which may be viewed as the rate conditional on x_2 for the same one-way channel but with the assignment $P_2\{x_1|x_2\}$ for the input letters.

Viewed this way, we may apply the concavity result of [2]. In particular, the weighted average of these rates with weight assignments $P_1\{x_2\}/(P_1\{x_2\} + P_2\{x_2\})$ and $P_2\{x_2\}/(P_1\{x_2\} + P_2\{x_2\})$ is dominated by the rate for this one-way channel when the probability assignments are the weighted average of the two given assignments. This weighted average of the given assignment is

$$\begin{aligned} P_3\{x_1, x_2\} &= \frac{P_1\{x_2\}}{P_1\{x_2\} + P_2\{x_2\}} P_1\{x_1|x_2\} + \frac{P_2\{x_2\}}{P_1\{x_2\} + P_2\{x_2\}} P_2\{x_1|x_2\} \\ &= \frac{1}{2} \frac{1}{(P_1\{x_2\} + P_2\{x_2\})} 2\,(P_1\{x_1, x_2\} + P_2\{x_1, x_2\}). \end{aligned} \tag{41}$$

Thus the sum of two corresponding terms (the same x_2) from (38) above is dominated by $P_1\{x_2\} + P_2\{x_2\}$ multiplied by the rate for this one-way channel with these averaged probabilities. This latter rate, on substituting the averaged probabilities, is seen to be

$$\sum_{x_1,y_2} P_3\{x_1, y_2|x_2\} \log \frac{P_3\{y_2|x_1, x_2\}}{P_3\{y_2|x_2\}} \tag{42}$$

where the subscript 3 corresponds to probabilities produced by using $P_3\{x_1, x_2\} = (P_1\{x_1, x_2\} + P_2\{x_1, x_2\}/2)/2$. In other words, the sum of (39) and (40) (including the first summation on x_2) is dominated by

$$\begin{aligned} &\sum_{x_2} (P_1\{x_2\} + P_2\{x_2\}) \sum_{x_1,y_2} P_3\{x_1, y_2|x_2\} \log \frac{P_3\{y_2|x_1, x_2\}}{P_3\{y_2|x_2\}} \\ &\qquad = 2 \sum_{x_1,x_2,y_2} P_3\{x_1, y_2, x_2\} \log \frac{P_3\{y_2|x_1, x_2\}}{P_3\{y_2|x_2\}}. \end{aligned} \tag{43}$$

This is the desired result for the theorem.

11. Applications of the concavity property; channels with symmetric structure

Theorem 4 is useful in a number of ways in evaluating the outer bound for particular channels. In the first place, we note that $R_{12} + \lambda R_{21}$ as a function of $P\{x_1, x_2\}$ and for positive λ is also a concave downward function. Consequently any local maximum is the absolute maximum and numerical investigation in locating such maxima by the Lagrange multiplier method is thereby simplified.

In addition, this concavity result is very powerful in helping locate the maxima when "symmetries" exist in a channel. Suppose, for example, that in a given channel the transition probability array $P\{y_1, y_2|x_1, x_2\}$ has the following property. There exists a relabelling of the input letters x_1 and of the output letters y_1 and y_2 which interchanges, say, the first two letters of the x_1 alphabet but leaves the set of probabilities $P\{y_1, y_2|x_1, x_2\}$ the same. Now if some particular assignment $P\{x_1, x_2\}$ gives outer bound rates R_{12} and R_{21}, then if we apply the same permutation to the x alphabet in $P\{x_1, x_2\}$ we obtain a new probability assignment which, however, will give exactly the same outer bound rates R_{12} and R_{21}. By our concavity property, if we average these two probability assignments we obtain a new probability assignment which will give at least as large values of R_{12} and R_{21}. In this averaged assignment for any particular x_2 the first two letters in the x_1 alphabet are assigned equal probability. In other words, in such a case an assignment for maximizing $R_{12} + \lambda R_{21}$, say $P\{x_1, x_2\}$ viewed as a matrix, will have its first two rows identical.

If the channel had sufficiently symmetric structure that *any* pair of x_1 letters might be interchanged by relabelling the x_1 alphabet and the y_1 and y_2 alphabets while preserving $P\{y_1, y_2|x_1, x_2\}$, then a maximizing assignment $P\{x_1, x_2\}$ would exist in which *all* rows are identical. In this case the entries are functions of x_2 only: $P\{x_1, x_2\} = P\{x_2\}/\alpha$ where α is the number of letters in the x_1 alphabet. Thus the maximum for a dependent assignment of $P\{x_1, x_2\}$ is actually obtained with x_1 and x_2 independent. *In other words, in this case of a full set of symmetric interchanges on the x_1 alphabet, the inner and outer bounds are identical.* This gives an important class of channels for which the capacity region can be determined with comparative ease.

An example of this type is the channel with transition probabilities as follows. All inputs and outputs are binary, $y_1 = x_2$ (that is, there is a noiseless binary channel from terminal 2 to terminal 1). If $x_2 = 0$, then $y_2 = x_1$, while if $x_2 = 1$, y_2 has probability .5 of being 0 and .5 of being 1. In other words, if x_2 is 0 the binary channel in the forward direction is noiseless, while if x_2 is 1 it is completely noisy. We note here that if the labels on the x_1 alphabet are interchanged while we simultaneously interchange the y_2 labels, the channel remains unaltered, all conditional probabilities being unaffected. Following the analysis above, then, the inner and outer bounds will be the same and give the capacity region. Furthermore, the surface will be attained with equal rows in the $P\{x_1, x_2\}$ matrix as shown in table II.

TABLE II

		x_2: 0	1
x_1	0	$p/2$	$q/2$
	1	$p/2$	$q/2$

For a particular p this assignment gives the rates

$$R_{12} = p, \qquad R_{21} = -(p \log p + q \log q). \tag{44}$$

These come from substituting in the formulas or by noting that in the $1 - 2$ direction the channel is acting like an erasure channel, while in the $2 - 1$ direction it is operating like a binary noiseless channel with unequal probabilities assigned to the letters. This gives the capacity region of figure 9.

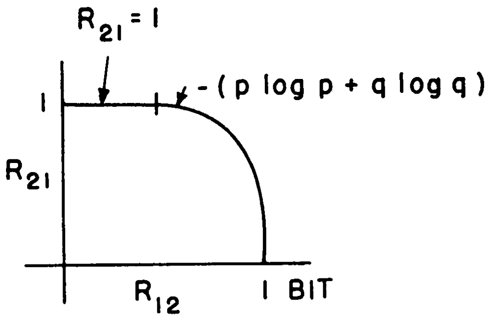

FIGURE 9

There are many variants and applications of these interchange and symmetry tricks for aid in the evaluation of capacity surfaces. For example, if *both* the x_1 and x_2 alphabets have a full set of interchanges leaving the transition probabilities the same, then the maximizing distribution must be identical both in rows and columns and hence all entries are the same, $P\{x_1, x_2\} = 1/\alpha c$ where α and c are the number of letters in the x_1 and x_2 alphabets. In this case, then, all attainable $R_{12}R_{21}$ points are dominated by the particular point obtained from this uniform probability assignment. *In other words, the capacity region is a rectangle in the case of a full set of symmetric interchanges for both x_1 and x_2.*

An example of this type is the channel of figure 2 defined by $y_1 = y_2 = x_1 \oplus x_2$ where $\oplus$ means mod 2 addition.

12. Nature of the region attainable in the outer bound

We now will use the concavity property to establish some results concerning the set Γ of points (R_{12}, R_{21}) that can be obtained by all possible assignments of probabilities $P\{x_1, x_2\}$ in a given channel K, and whose convex hull is G_O. We will show that the set Γ is in fact already convex and therefore identical with G_O and that it consists of all points in or on the boundary of a region of the type shown in figure 10 bounded by a horizontal segment L_1, an outward convex seg-

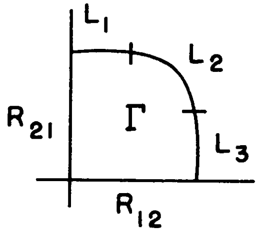

FIGURE 10

ment L_2, a vertical segment L_3 and two segments of the coordinate axes. Thus G_O has a structure similar to G_I.

Suppose some $P\{x_1, x_2\}$ gives a point (R_{12}, R_{21}). Here R_{12} is, as we have observed previously, an average of the different R_{12} which would be obtained by fixing x_2 at different values, that is, using these with probability 1 and applying the conditional probabilities $P\{x_1|x_2\}$ to the x_1 letters. The weighting is according to factors $P\{x_2\}$. It follows that some particular x_2 will do as well at least as this weighted average. If this particular x_2 is x_2^*, the set of probabilities $P\{x_1|x_2^*\}$ gives at least as large a value of R_{12} and simultaneously makes $R_{21} = 0$. In

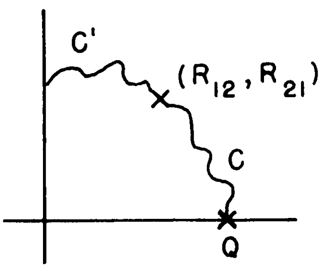

FIGURE 11

figure 11 this means we can find a point in Γ below or to the right of the projection of the given point as indicated (point Q).

Now consider mixtures of these two probability assignments, that is, assignments of the form $\lambda P\{x_1x_2\} + (1 - \lambda)P\{x_1|x_2^*\}$. Here λ is to vary continuously from 0 to 1. Since R_{12} and R_{21} are continuous functions of the assigned probability, this produces a continuous curve C running from the given point to the point Q. Furthermore, this curve lies entirely to the upper right of the connecting line segment. This is because of the concavity property for the R_{12} and R_{21} expressions. In a similar way, we construct a curve C', as indicated, of points belonging to Γ and lying on or above the horizontal straight line through the given point.

Now take all points on the curves C and C' and consider mixing the corresponding probability assignments with the assignment $P\{x_1^*, x_2^*\} = 1$ (all other pairs given zero probability). This last assignment gives the point $(0, 0)$. The fraction of this $(0, 0)$ assignment is gradually increased for 0 up to 1. As this is done the curve of resulting points changes continuously starting at the CC' curve and collapsing into the point $(0, 0)$. The end points stay on the axes during this operation. Consequently by known topological results the curve sweeps through the entire area bounded by C, C' and the axes and in particular *covers the rectangle subtended by the original point* (R_{12}, R_{21}).

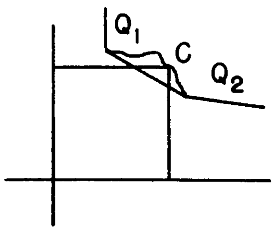

FIGURE 12

We will show that the set of points Γ is a convex set. Suppose Q_1 and Q_2, figure 12, are two points which can be obtained by assignments $P_1\{x_1, x_2\}$ and $P_2\{x_1, x_2\}$.

By taking mixtures of varying proportions one obtains a continuous curve C connecting them, lying, by the concavity property, to the upper right of the connecting line segment. Since these are points of Γ all of their subtended rectangles are, as just shown, points of Γ. It follows that all points of the connecting line segment are points of Γ. Note that if Q_1 and Q_2 are in the first and third quadrants relative to each other the result is trivially true, since then the connecting line segment lies in the rectangle of one of the points.

These results are sufficient to imply the statements at the beginning of this section, namely the set Γ is convex, identical with G_O, and if we take the largest attainable R_{12} and for this R_{12} the largest R_{21}, then points in the subtended rectangle are attainable. Similarly for the largest R_{21}.

It may be recalled here that the set of points attainable by *independent* assignments, $P\{x_1, x_2\} = P\{x_1\}P\{x_2\}$, is not necessarily a convex set. This is shown by the example of table I.

It follows also from the results of this section that *the end points of the outer bound curve* (where it reaches the coordinate axes) *are the same as the end points of the inner bound curve.* This is because, as we have seen, the largest R_{12} can be achieved using only one particular x_2 with probability 1. When this is done, $P\{x_1, x_2\}$ reduces to a product of independent probabilities.

13. An example where the inner and outer bounds differ

The inner and outer bounds on the capacity surface that we have derived above are not always the same. This was shown by David Blackwell for the binary multiplying channel defined by $y_1 = y_2 = x_1x_2$. The inner and outer bounds for this channel have been computed numerically and are plotted in figure 13. It may be seen that they differ considerably, particularly in the middle

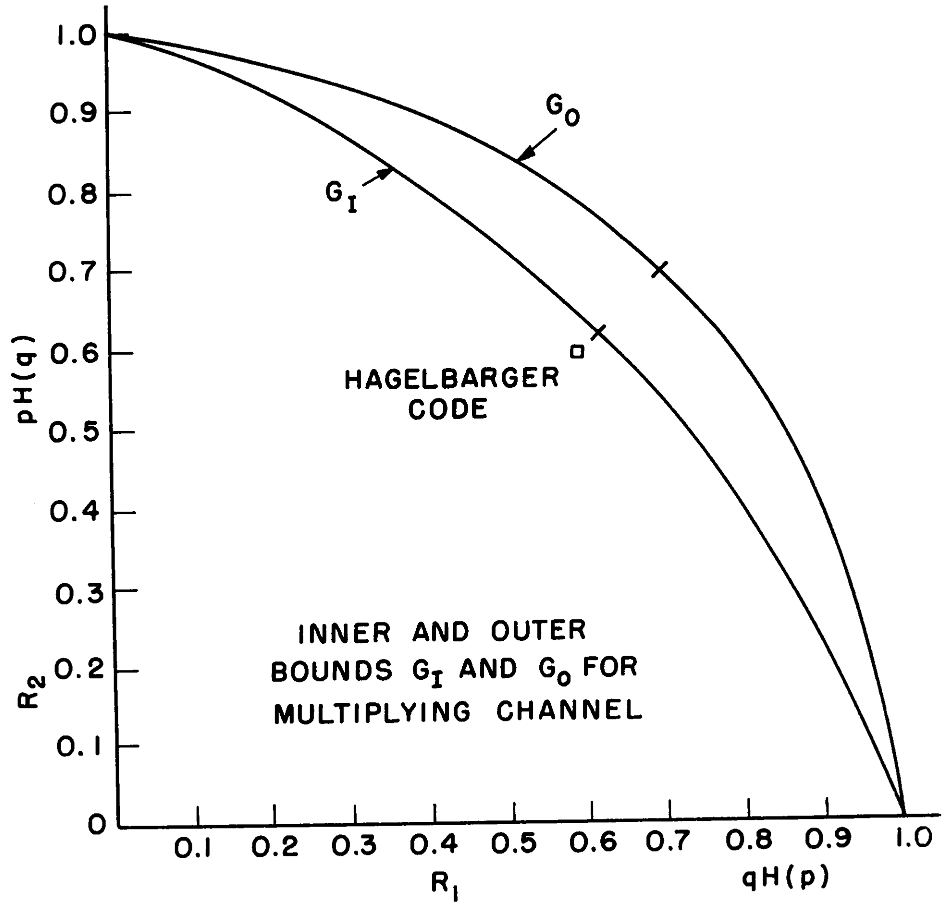

FIGURE 13

of the range. The calculation of the inner bound, in this case, amounts to finding the envelope of points

$$
\begin{aligned}
R_{12} &= -p_2[p_1 \log p_1 + (1 - p_1) \log (1 - p_1)] \\
R_{21} &= -p_1[p_2 \log p_2 + (1 - p_2) \log (1 - p_2)].
\end{aligned}
\tag{45}
$$

These are the rates with independent probability assignments at the two ends:

probability p_1 for using letter 1 at terminal 1 and probability p_2 for using letter 1 at terminal 2. By evaluating these rates for different p_1 and p_2 the envelope shown in the figure was obtained.

For the outer bounds, the envelope of rates for a general dependent assignment of probabilities is required. However it is easily seen that any assignment in which $P\{0, 0\}$ is positive can be improved by transferring this probability to one of the other possible pairs. Hence we again have a two parameter family of points (since the sum of the three other probabilities must be unity). If the probabilities are denoted by $p_1 = P\{1, 0\}$, $p_2 = P\{0, 1\}$, $1 - p_1 - p_2 = P\{1, 1\}$, we find the rates are

$$\begin{aligned} R_{12} &= -(1-p_1)\left[\frac{p_2}{1-p_1}\log\frac{p_2}{1-p_1} + \left(1-\frac{p_2}{1-p_1}\right)\log\left(1-\frac{p_2}{1-p_1}\right)\right] \\ R_{21} &= -(1-p_2)\left[\frac{p_1}{1-p_2}\log\frac{p_1}{1-p_2} + \left(1-\frac{p_1}{1-p_2}\right)\log\left(1-\frac{p_1}{1-p_2}\right)\right]. \end{aligned} \tag{46}$$

Here again a numerical evaluation for various values of p_1 and p_2 led to the envelope shown in the figure.

In connection with this channel, D. W. Hagelbarger has devised an interesting and simple code (not a block code however) which is error free and transmits at average rates $R_{12} = R_{21} = .571$, slightly less than our lower bound. His code operates as follows. A 0 or 1 is sent from each end with independent probabilities 1/2, 1/2. If a 0 is received then the next digit transmitted is the complement of what was just sent. This procedure is followed at both ends. If a 1 is received, both ends progress to the next binary digit of the message. It may be seen that three-fourths of the time on the average the complement procedure is followed and one-fourth of the time a new digit is sent. Thus the average number of channel uses per message digit is $(3/4)(2) + (1/4)(1) = 7/4$. The average rate is $4/7 = .571$ in both directions. Furthermore it is readily seen that the message digits can be calculated without error for each communication direction.

By using message sources at each end with biased probabilities it is possible to improve the Hagelbarger scheme slightly. Thus, if 1's occur as message digits with probability .63 and 0's with probability .37, we obtain rates in both directions

$$R_{12} = R_{21} = \frac{-.63\log .63 - .37\log .37}{1-(.63)^2} = .593. \tag{47}$$

We will, in a later section, develop a result which in principle gives for any channel the exact capacity region. However, the result involves a limiting process over words of increasing length and consequently is difficult to evaluate in most cases. In contrast, the upper and lower bounds involve only maximizing operations relating to a single transmitted letter in each direction. Although sometimes involving considerable calculation, it is possible to actually evaluate them when the channel is not too complex.

14. Attainment of the outer bound with dependent sources

With regard to the outer bound there is an interesting interpretation relating to a somewhat more general communication system. Suppose that the message sources at the two ends of our channel are not independent but statistically dependent. Thus, one might be sending weather information from Boston to New York and from New York to Boston. The weather at these cities is of course not statistically independent. If the dependence were of just the right type for the channel or if the messages could be transformed so that this were the case, then it may be possible to attain transmission at the rates given by the outer bound. For example, in the multiplying channel just discussed, suppose that the messages at the two ends consist of streams of binary digits which occur with the dependent probabilities given by table III. Successive x_1, x_2 pairs

TABLE III

		x_2 0	1
x_1	0	0	.275
	1	.275	.45

are assumed independent. Then by merely sending these streams into the channel (without processing) the outer bound curve is achieved at its midpoint.

It is not known whether this is possible in general. Does there always exist a suitable pair of dependent sources that can be coded to give rates R_1, R_2 within ϵ of any point in the outer bound? This is at least often possible in the noiseless, memoryless case, that is, when y_1 and y_2 are strict functions of x_1 and x_2 (no channel noise). The source pair defined by the assignment $P\{x_1, x_2\}$ that produces the point in question is often suitable in such a case without coding as in the above example.

The inner bound also has an interesting interpretation. If we artificially limit the codes to those where the transmitted sequence at each terminal depends only on the message and not on the received sequence at that terminal, then the inner bound is indeed the capacity region. This results since in this case we have at each stage of the transmission (that is, given the index of the letter being transmitted) independence between the two next transmitted letters. It follows that the total vector change in equivocation is bounded by the sum of n vectors, each corresponding to an independent probability assignment. Details of this proof are left to the reader. The independence required would also occur if the transmission and reception points at each end were at different places with no direct cross communication.

15. General solution for the capacity region in the two-way channel

For a given memoryless two-way channel K we define a series of *derived channels* $K_1, K_2, \cdots$. These will also be memoryless channels and the capacity region for K will be evaluated as a limit in terms of the inner bounds for the series K_n.

The channel K_1 is identical with K. The derived channel K_2 is one whose input letters are actually strategies for working with K for a block of two input letters. Thus the input letters at terminal 1 for K_2 consist of pairs $[x_1^1, f(x_1^1, y_1^1)]$. Here x_1^1 is the first transmitted letter of the pair and ranges therefore over the a possible input letters of K. Now $f(x_1^1, y_1^1)$ represents any function from the first input letter x_1^1 and output letter y_1^1 to the second input letter x_1^2. Thus this function may be thought of as a rule for choosing a second input letter at terminal 1 depending on the first input letter and the observed first output letter. If x_1^1 can assume a values and y_1^1 can assume b values, then the (x_1^1, y_1^1) pair can assume ab values, and since the function f takes values from a possibilities there are a^{ab} possible functions. Hence there are $a \cdot a^{ab}$ possible pairs $[x_1^1, f(x_1^1, y_1^1)]$, or possible input letters to K_2 at terminal 1.

In a similar way, at terminal 2 consider pairs $[x_2^1, g(x_2^1, y_2^1)]$. Here g ranges over functions from the first received and transmitted letters at terminal 2 and takes values from the x_2 alphabet. Thus these pairs have $c \cdot c^{cd}$ values, where c and d are the sizes of the input and output alphabets at terminal 2.

The pairs $[x_1^1, f(x_1^1, y_1^1)]$ and $[x_2^1, g(x_2^1, y_2^1)]$ may be thought of as strategies for using the channel K in two letter sequences, the second letter to be dependent on the first letter sent and the first letter received. The technique here is very similar to that occurring in the theory of games. There one replaces a sequence of moves by a player (whose available information for making a choice is increasing through the series) by a single move in which he chooses a strategy. The strategy describes what the player will do at each stage in each possible contingency. Thus a game with many moves is reduced to a game with a single move chosen from a larger set.

The *output* letters for K_2 are, at terminal 1, pairs (y_1^1, y_1^2) and, at terminal 2, pairs (y_2^1, y_2^2); that is, the pairs of received letters at the two terminals. The transition probabilities for K_2 are the probabilities, if these strategies for introducing a particular pair of letters were used in K, that the output pairs would occur. Thus

$$
\begin{aligned}
(48) \qquad & P_{K_2}\{(y_1^1, y_1^2), (y_2^1, y_2^2) | [x_1^1, f(x_1^1, y_1^1)], [x_2^1, g(x_2^1, y_2^1)]\} \\
& \qquad = P_K\{y_1^1, y_2^1 | x_1^1, x_2^1\} P_K\{y_1^2, y_2^2 | f(x_1^1, y_1^1), g(x_2^1, y_2^1)\}.
\end{aligned}
$$

In a similar way the channels $K_3, K_4, \cdots$ are defined. Thus K_n may be thought of as a channel corresponding to n uses of K with successive input letters at a terminal functions of previous input and output letters at that terminal. Therefore the input letters at terminal 1 are n-tuples

$$(49)\qquad [x_1^1, f(x_1^1, y_1^1), \cdots, f_{n-1}(x_1^1, x_1^2, \cdots, x_1^{n-1}, y_1^1, y_1^2, \cdots, y_1^{n-1})],$$

a possible alphabet of

$$(50)\qquad a a^{ab} a^{(ab)^2} \cdots a^{(ab)^{n-1}} = a^{[(ab)^n - 1]/(ab-1)}$$

possibilities. The output letters at terminal 1 consist of n-tuples

$$(51)\qquad (y_1^1, y_1^2, \cdots, y_1^n)$$

and range therefore over an alphabet of b^n generalized letters. The transition probabilities are defined for K_n in terms of those for K by the generalization of equation (39)

$$(52)\qquad P_{K_n}\{y_1^1, y_1^2, \cdots, y_1^n | (x_1^1, f_1, f_2, \cdots, f_{n-1}), (x_2^1, g_1, g_2, \cdots, g_{n-1})\} = \prod_{i=1}^{n} P_K\{y_1^i | f_{i-1}, g_{i-1}\}.$$

The channel K_n may be thought of, then, as a memoryless channel whose properties are identical with using channel K in blocks of n, allowing transmitted and received letters within a block to influence succeeding choices.

For each of the channels K_n one could, in principle, calculate the lower bound on its capacity region. The lower bound for K_n is to be multiplied by a factor $1/n$ to compare with K, since K_n corresponds to n uses of K.

THEOREM 5. *Let B_n be the lower bound of the capacity region for the derived channel K_n reduced in scale by a factor $1/n$. Then as $n \to \infty$ the regions B_n approach a limit B which includes all the particular regions and is the capacity region of K.*

PROOF. We first show the positive assertion that if (R_{12}, R_{21}) is any point in some B_n and ϵ is any positive number, then we can construct block codes with error probabilities $P_e < \epsilon$ and rates in the two directions at least $R_{12} - \epsilon$ and $R_{21} - \epsilon$. This follows readily from previous results if the derived channel K_n and its associated inner bound B_n are properly understood. K_n is a memoryless channel, and by theorem 3 we can find codes for it transmitting arbitrarily close to the rates R_{12}, R_{21} in B_n with arbitrarily small error probability. These codes are sequences of letters from the K_n alphabet. They correspond, then, to sequences of *strategies* for blocks of n for the original channel K.

Thus these codes can be directly translated into codes for K n times as long, preserving all statistical properties, in particular the error probability. These codes, then, can be interpreted as codes signalling at rates $1/n$ as large for the K channel with the same error probability. In fact, from theorem 3, it follows that for any pair of rates strictly inside B_n we can find codes whose error probability decreases at least exponentially with the code length.

We will now show that the regions B_n approach a limit B as n increases and that B includes all the individual B_n. By a limiting region we mean a set of points B such that for any point P of B, and $\epsilon > 0$, there exists n_0 such that for $n > n_0$ there are points of B_n within ϵ of P, while for any P not in B there exist ϵ and n_0 such that for $n > n_0$ no points of B_n are within ϵ of P. In the first place B_n is included in B_{kn} for any integer k. This is because the strategies for B_{kn} include as special cases strategies where the functional influence only involves subblocks of n. Hence all points obtainable by independent probability assignments with K_n are also obtainable with K_{kn} and the convex hull of the latter set must include the convex hull of the former set.

It follows that the set B_{kn} approaches a limit B, the union of all the B_{kn} plus limit points of this set. Also B includes B_{n_1} for any n_1. For n and n_1 have a common multiple, for example nn_1, and B includes B_{nn_1} while B_{nn_1} includes B_{n_1}.

Furthermore, any point obtainable with K_{kn} can be obtained with $K_{kn+\alpha}$, for $0 \leqq \alpha \leqq n$, reduced in both coordinates by a factor of not more than $k/(k+1)$. This is because we may use the strategies for K_{kn} followed by a series of α of the first letters in the x_1 and x_2 alphabets. (That is, fill out the assignments to the length $kn + \alpha$ with essentially dummy transmitted letters.) The only difference then will be in the normalizing factor, 1/(block length). By making k sufficiently large, this discrepancy from a factor of 1, namely $1/(k+1)$, can be made as small as desired. Thus for any $\epsilon > 0$ and any point P of B there is a point of B_{n_1} within ϵ of P for *all* sufficiently large n_1.

With regard to the converse part of the theorem, suppose we have a block code of length n with signalling rates (R_1, R_2) corresponding to a point outside B, closest distance to B equal to ϵ. Then since B includes B_n, the closest distance to B_n is at least ϵ. We may think of this code as a block code of length 1 for the channel K_n. As such, the messages m_1 and m_2 are mapped directly into "input letters" of K_n without functional dependence on the received letters. We have then since m_1 and m_2 are independent the independence of probabilities associated with these input letters sufficient to make the inner bound and outer bound the same. Hence the code in question has error probability bounded away from zero by a quantity dependent on ϵ but not on n.

16. Two-way channels with memory

The general discrete two-way channel with memory is defined by a set of conditional probabilities

$$(53) \qquad P\{y_{1n}, y_{2n} | x_{11}, x_{12}, \cdots, x_{1n}; x_{21}, x_{22}, \cdots, x_{2n}; \\ y_{11}, y_{12}, \cdots, y_{1n-1}; y_{21}, y_{22}, \cdots, y_{2n-1}\}.$$

This is the probability of the nth output pair y_{1n}, y_{2n} conditional on the preceding history from time $t = 0$, that is, the input and output sequences from the starting time in using the channel. In such a general case, the probabilities might change in completely arbitrary fashion as n increases. Without further limitation, it is too general to be either useful or interesting. What is needed is some condition of reasonable generality which, however, ensures a certain stability in behavior and allows, thereby, significant coding theorems. For example, one might require finite historical influence so that probabilities of letters depend only on a bounded past history. (Knowing the past d inputs and outputs, earlier inputs and outputs do not influence the conditional probabilities.) We shall, however, use a condition which is, by and large, more general and also more realistic for actual applications.

We will say that a two-way channel has the *recoverable state property* if it satisfies the following condition. There exists an integer d such that for any input and output sequences of length n, X_{1n}, X_{2n}, Y_{1n}, Y_{2n}, there exist two functions $f(X_{1n}, Y_{1n})$, $g(X_{2n}, Y_{2n})$ whose values are sequences of input letters of the same length less than d and such that if these sequences f and g are now sent over the channel it is returned to its original state. Thus, conditional probabilities after this are the same as if the channel were started again at time zero.

The recoverable state property is common in actual physical communication systems where there is often a "zero" input which, if applied for a sufficient period, allows historical influences to die out. Note also that the recoverable state property may hold even in channels with an infinite set of internal states, provided it is possible to return to a "ground" state in a bounded number of steps.

The point of the recoverable state condition is that if we have a block code for such a channel, we may annex to the input words of this code the functions f and g at the two terminals and then repeat the use of the code. Thus, if such a code is of length n and has, for one use of the code, signalling rates R_1 and R_2 and error probabilities P_{e1} and P_{e2}, we may *continuously* signal at rates $R_1' \geqq nR_1/(n + d)$ and $R_2' \geqq nR_2/(n + d)$ with error probabilities $P_{e1}' \leqq P_{e1}$ and $P_{e2}' \leqq P_{e2}$.

For a recoverable state channel we may consider strategies for the first n letters just as we did in the memoryless case, and find the corresponding inner bound B_n on the capacity region (with scale reduced by $1/n$). We define the region B which might be called the limit supremum of the regions B_n. Namely, B consists of all points which belong to an infinite number of B_n together with limit points of this set.

THEOREM 6. *Let (R_1, R_2) be any point in the region B. Let n_0 be any integer and let ϵ_1 and ϵ_2 be any positive numbers. Then there exists a block code of length $n > n_0$ with signalling rates R_1', R_2' satisfying $|R_1 - R_1'| < \epsilon_1$, $|R_2 - R_2'| < \epsilon_1$ and error probabilities satisfying $P_{e1} < \epsilon_2$, $P_{e2} < \epsilon_2$. Conversely, if (R_1, R_2) is not in B then there exist n_0 and $\delta > 0$ such that any block code of length exceeding n_0 has either $P_{e1} > \delta$ or $P_{e2} > \delta$ (or both).*

PROOF. To show the first part of the theorem choose an $n_1 > n_0$ and also large enough to make both $dR_1/(d + n)$ and $dR_2/(d + n)$ less than $\epsilon_1/2$. Since the point (R_1, R_2) is in an infinite sequence of B_n, this is possible. Now construct a block code based on n_1 uses of the channel as individual "letters," within $\epsilon_1/2$ of the rate pair (R_1, R_2) and with error probabilities less than ϵ_2. To each of the "letters" of this code annex the functions which return the channel to its original state. We thus obtain codes with arbitrarily small error probability $< \epsilon_2$ approaching the rates R_1, R_2 and with arbitrarily large block length.

To show the converse statement, suppose (R_1, R_2) is *not* in B. Then for some

n_0 every B_n, where $n > n_0$, is outside a circle of some radius, say ϵ_2, centered on (R_1, R_2). Otherwise (R_1, R_2) would be in a limit point of the B_n. Suppose we have a code of length $n_1 > n_0$. Then its error probability is bounded away from zero since we again have a situation where the independence of "letters" obtains.

The region B may be called the capacity region for such a recoverable state channel. It is readily shown that B has the same convexity properties as had the capacity region G for a memoryless channel. Of course, the actual evaluation of B in specific channels is even more impractical than in the memoryless case.

17. Generalization to T-terminal channels

Many of the tricks and techniques used above may be generalized to channels with three or more terminals. However, some definitely new phenomena appear in these more complex cases. In another paper we will discuss the case of a channel with two or more terminals having inputs only and one terminal with an output only, a case for which a complete and simple solution of the capacity region has been found.

◇ ◇ ◇ ◇ ◇

APPENDIX. ERROR PROBABILITY BOUNDS IN TERMS OF MOMENT GENERATING FUNCTIONS

Suppose we assign probabilities $P\{x_1\}$ to input letters at terminal 1 and $P\{x_2\}$ to input letters at terminal 2. (Notice that we are here working with letters, not with words as in theorem 2.) We can then calculate the log of the moment generating functions of the mutual information between input letters at terminal 1 and input letter-output letter pairs at terminal 2. (This is the log of the moment generating function of the distribution ρ_{12} when $n = 1$.) The expressions for this and the similar quantity in the other direction are

$$\begin{aligned}(54)\qquad \mu_1(s) &= \log \sum_{x_1,x_2,y_2} P\{x_1, x_2, y_2\} \exp\left(s \log \frac{P\{x_1, x_2, y_2\}}{P\{x_1\}P\{x_2, y_2\}}\right)\\ &= \log \sum_{x_1,x_2,y_2} \frac{P\{x_1, x_2, y_2\}^{s+1}}{P\{x_1\}^s P\{x_2, y_2\}^s},\\ (55)\qquad \mu_2(s) &= \log \sum_{x_1,x_2,y_1} \frac{P\{x_1, x_2, y_1\}^{s+1}}{P\{x_2\}^s P\{x_1, y_1\}^s}.\end{aligned}$$

These functions μ_1 and μ_2 may be used to bound the tails on the distributions ρ_{12} and ρ_{21} obtained by adding n identically distributed samples together. In fact, Chernoff [4] has shown that the tail to the left of a mean may be bounded as follows:

$$(56)\qquad \begin{aligned}\rho_{12}[n\mu_1'(s_1)] &\leqq \exp\{n[\mu_1(s_1) - s_1\mu_1'(s_1)]\}, & s_1 &\leqq 0,\\ \rho_{21}[n\mu_2'(s_2)] &\leqq \exp\{n[\mu_2(s_2) - s_2\mu_2'(s_2)]\}, & s_2 &\leqq 0.\end{aligned}$$

Thus, choosing an arbitrary negative s_1, this gives a bound on the distribution function at the value $n\mu_1'(s_1)$. It can be shown that $\mu'(s)$ is a monotone increasing function and that $\mu'(0)$ is the mean of the distribution. The minimum $\mu'(s)$ corresponds to the minimum possible value of the random variable in question, in this case, the minimum $I(x_1; x_2, y_2)$. Thus, an s_1 may be found to place $\mu_1(s_1)$ anywhere between $I_{\min}(x_1; x_2, y_2)$ and $E(I)$. Of course, to the left of $I_{\min}$ the distribution is identically zero and to the right of $E(I)$ the distribution approaches one with increasing n.

We wish to use these results to obtain more explicit bounds on P_{e1} and P_{e2}, using theorem 2. Recalling that in that theorem θ_1 and θ_2 are arbitrary, we attempt to choose them so that the exponentials bounding the two terms are equal. This is a good choice of θ_1 and θ_2 to keep the total bound as small as possible. The first term is bounded by $\exp\{n[\mu_1(s_1) - s_1\mu_1'(s_1)]\}$ where s_1 is such that $\mu_1'(s_1) = R_1 + \theta_1$, and the second term is equal to $\exp(-n\theta_1)$. Setting these equal, we have

$$\mu_1(s_1) - s_1\mu_1'(s_1) = -\theta_1, \qquad R_1 + \theta_1 = \mu_1'(s_1). \tag{57}$$

Eliminating θ_1, we have

$$R_1 = \mu_1(s_1) - (s_1 - 1)\mu_1'(s_1) \tag{58}$$

and

$$E(P_{e1}) \leqq 2\exp\{n[\mu_1(s_1) - s_1\mu_1'(s_1)]\}. \tag{59}$$

This is because the two terms are now equal and each dominated by $\exp\{n[\mu_1(s_1) - s_1\mu_1'(s_1)]\}$. Similarly, for

$$R_2 = \mu_2(s_2) - (s_2 - 1)\mu_2'(s_2) \tag{60}$$

we have

$$E(P_{e2}) \leqq 2\exp\{n[\mu_2(s_2) - s_2\mu_2'(s_2)]\}. \tag{61}$$

These might be called parametric bounds in terms of the parameters s_1 and s_2. One must choose s_1 and s_2 such as to make the rates R_1 and R_2 have the desired values. These s_1 and s_2 values, when substituted in the other formulas, give bounds on the error probabilities.

The derivative of R_1 with respect to s_1 is $-(s_1 - 1)\mu_1''(s_1)$, a quantity always positive when s_1 is negative except for the special case where $\mu''(0) = 0$. Thus, R_1 is a monotone increasing function of s_1 as s_1 goes from $-\infty$ to 0, with R_1 going from $-I_{\min} - \log P\{I_{\min}\}$ to $E(I)$. The bracketed term in the exponent of $E(P_{e1})$, namely $\mu_1(s_1) - s_1\mu_1(s_1)$, meanwhile varies from $\log P\{I_{\min}\}$ up to zero. The rate corresponding to $s_1 = -\infty$, that is, $-I_{\min} - \log P\{I_{\min}\}$, may be positive or negative. If negative (or zero) the entire range of rates is covered from zero up to $E(I)$. However, if it is positive, there is a gap from rate $R_1 = 0$ up to this end point. This means that there is no way to solve the equation for rates in this interval to make the exponents of the two terms equal. The best course here to give a good bound is to choose θ_1 in such a way that $n(R_1 + \theta_1)$ is just smaller than $I_{\min}$, say $I_{\min} - \epsilon$. Then $\rho_{12}[n(R_1 + \theta_1)] = 0$ and only the second term, $\exp(\theta_1 n)$, is left in the bound. Thus $\exp[-n(I_{\min} - R_1 - \epsilon)]$ is a bound on P_e. This is true for any $\epsilon > 0$. Since we can construct such codes for any positive ϵ and since there are only a finite number of codes, this implies that we can construct a code satisfying this inequality with $\epsilon = 0$. Thus, we may say that

$$E(P_{e1}) \leqq \exp[-n(I_{\min} - R_1)], \qquad R_1 \leqq I_{\min}. \tag{62}$$

Of course, exactly similar statements hold for the second code working in the reverse direction. Combining and summarizing these results we have the following.

THEOREM 7. *In a two-way memoryless channel K with finite alphabets, let $P\{x_1\}$ and $P\{x_2\}$ be assignments of probabilities to the input alphabets, and suppose these lead to the logarithms of moment generating functions for mutual information $\mu_1(s_1)$ and $\mu_2(s_2)$,*

$$\begin{aligned} \mu_1(s_1) &= \log \sum_{x_1,x_2,y_2} \frac{P\{x_1, x_2, y_2\}^{s+1}}{P\{x_1\}^s P\{x_2, y_2\}^s} \\ \mu_2(s_2) &= \log \sum_{x_1,x_2,y_2} \frac{P\{x_1, x_2, y_1\}^{s+1}}{P\{x_2\}^s P\{x_1, y_1\}^s}. \end{aligned} \tag{63}$$

Let $M_1 = \exp(R_1 n)$, $M_2 = \exp(R_2 n)$ be integers, and let s_1, s_2 be the solutions (when they exist) of

$$\begin{aligned} R_1 &= \mu_1(s_1) - (s_1 + 1)\mu_1'(s_1) \\ R_2 &= \mu_2(s_2) - (s_2 + 1)\mu_2'(s_2). \end{aligned} \tag{64}$$

The solution s_1 will exist if

$$-I_{\min}(x_1; x_2, y_2) - \log P\{I_{\min}(x_1; x_2, y_2)\} \leqq R_1 \leqq E[I(x_1; x_2, y_2)], \tag{65}$$

and similarly for s_2. If both s_1 and s_2 exist, then there is a code pair for the channel K of length n with M_1 and M_2 messages and error probabilities satisfying

$$\begin{aligned} P_{e1} &\leqq 4 \exp\{+n[\mu_1(s_1) - s_1\mu_1'(s_1)]\} \\ P_{e2} &\leqq 4 \exp\{+n[\mu_2(s_2) - s_2\mu_2'(s_2)]\}. \end{aligned} \tag{66}$$

If either (or both) of the R is so small that the corresponding s does not exist, a code pair exists with the corresponding error probability bounded by

$$P_{e1} \leqq 2 \exp\{-n[I(x_1; x_2, y_2) - R_1]\} \tag{67}$$

or

$$P_{e2} \leqq 2 \exp\{-n[I(x_2; x_1, y_1) - R_2]\}. \tag{68}$$

Thus, if s_1 exists and not s_2, then inequalities (66) *would be used. If neither exists,* (67) *and* (68) *hold.*

REFERENCES

[1] C. E. SHANNON, "Certain results in coding theory for noisy channels," *Information and Control*, Vol. 1 (1957), pp. 6–25.

[2] ———, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, Vol. 2 (1958), pp. 289–293.

[3] ———, "Geometrische Deutung einiger Ergebnisse bei der Berechnung der Kanalkapazitat," *Nachrtech. Z.*, Vol. 10 (1957).

[4] H. CHERNOFF, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, Vol. 23 (1952), pp. 493–507.