#3.

SEQUENTIAL TRANSMISSION OF DIGITAL INFORMATION WITH FEEDBACK

MICHAEL HORSTEIN

TECHNICAL REPORT 375

SEPTEMBER 8, 1960

LOAN COPY ONLY

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the U.S. Army (Signal Corps), the U.S. Navy (Office of Naval Research), and the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039-sc-78108, Department of the Army Task 3-99-20-001 and Project 3-99-00-000.

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology with the joint support of the U.S. Army, Navy, and Air Force under Air Force Contract AF19(604)-5200.

The work was also published as Report 34G-0006, Lincoln Laboratory, M.I.T., 8 August 1960.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 375 September 8, 1960

SEQUENTIAL TRANSMISSION OF DIGITAL INFORMATION WITH FEEDBACK

Michael Horstein

Submitted to the Department of Electrical Engineering, M.I.T., August 22, 1960, in partial fulfillment of the requirements for the degree of Doctor of Science.

Abstract

Most of the known error-correcting procedures for digital communication involve codes of fixed constraint length. In a two-way communication system, however, it is possible to base the inclusion of redundancy on information concerning the results of previous transmissions, which is received over a feedback channel. A wide variety of sequential transmission procedures is thereby made possible. This report is concerned with the asymptotic error-correcting capability of several types of sequential transmission and the examination of a particular sequential transmission scheme. Before the results can be summarized, however, several terms must be defined.

In sequential transmission schemes, as in one-way communication, there are two general approaches to the encoding of information for transmission over a noisy channel. The information or message sequence (which we assume is a binary sequence in which each digit carries one bit of information) can, on the one hand, be divided into blocks of fixed length, which are then encoded as separate units. Alternatively, the transmission process may be a continuous one, in which bits are introduced into the encoder, and later decoded, individually.

The term "decision feedback" is reserved for block-transmission systems in which a code word of indefinite length is assigned to each possible message block. As much of the code word is transmitted as is required for the intended message to be identified with the required degree of reliability. The term "information feedback" applies to any continuous or block-transmission system in which each transmitted symbol is a function of both the intended message and information available at the transmitter concerning the results of previous transmissions.

Abstract (continued)

Bounds on the error exponent characterizing the asymptotic relation between the specified probability of error and the average constraint length required to achieve it are derived for several types of sequential transmission. It is shown that the error exponent of a sequential sphere-packed code, which is the (decision-feedback) sequential analog of a sphere-packed block code, is greater, at all rates less than capacity, than the largest exponent obtainable with a fixed-constraint-length block code. It is conjectured that it is at least as great as that of any realizable decision-feedback system. It is also shown, however, that asymptotically better error correction is possible with an information-feedback block-transmission system. Within the realm of information feedback, even larger error exponents can be obtained with continuous systems.

The last point is illustrated by an example of an information-feedback continuous-transmission system which assumes the existence of a noiseless feedback channel. Its error exponent, as a function of the transmission rate, varies between the "sequential dichotomy" exponent at zero rate, and the "fixed-length dichotomy" exponent at capacity. (For all other systems mentioned, the exponent at capacity is zero.) Bounds are obtained on the over-all constraint length at capacity. A computer simulation of the system reveals that the average constraint length required to achieve an error probability of 10^{-15} at channel capacity is only approximately 1/7 of that required by the "optimum" fixed-constraint-length block code, when the latter operates at only half capacity.

TABLE OF CONTENTS

I. INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

It is a requirement of many communication systems that the information provided at the receiver be an extremely accurate reproduction of the transmitted information. For a given rate of transmission, an increase in either the signal power or its bandwidth makes possible a reduction in the frequency of transmission errors. If, regardless of the form of the transmitted signals and the type of detection used, the available power and bandwidth are not sufficient to provide the required degree of reliability, some form of redundancy must be added at the transmitter.

We shall confine ourselves to the consideration of communication through a discrete memoryless channel. Such a channel is defined by a set of transition probabilities, $p(y_j/x_i)$, which gives the probability of receiving symbol $y_j (j=1, 2, \ldots, m)$ when symbol $x_i (i=1, 2, \ldots, n)$ is transmitted. Shannon (1) has shown, for any discrete memoryless channel, that if sufficient redundancy is added so that the rate of transmission is less than the channel capacity, it is possible by proper encoding and decoding to make the average probability of error after decoding arbitrarily small. This is not possible for rates greater than capacity. Since the publication of this fundamental result, numerous schemes have been devised to implement what the theorem had shown to be possible. Most of these schemes envision a channel that allows transmission of information from transmitter to receiver only. They necessarily involve fairly complex coding at the transmitter and a decoding procedure of at least comparable (but usually considerably greater) complexity at the receiver. Specifically, each information digit must be involved in the determination of a number of transmitted symbols that increases as the required average error probability decreases. In a one-way transmission system, for fixed values of the channel capacity, transmission rate, and average probability of error, this number is a constant known as the block length or constraint length.

In communication systems that have provisions for transmission in both directions, it is possible to add redundancy in ways significantly different from the method used in the coding schemes referred to above, by making use of information received through the reverse channel. For example, the constraint length among the transmitted symbols can be a variable quantity; a particular information digit need only influence the selection of a sufficient number of transmitted symbols to allow it to be decoded with the required degree of reliability. Furthermore, there is the possibility of having available, when selecting a transmitted symbol, information concerning the effect of the channel on each previous transmission. Thus, the presence of a feedback channel affords the opportunity to use a wide variety of sequential transmission procedures.

A second theorem of Shannon (2) states that the capacity, in the forward direction, of a discrete memoryless channel is not increased by the availability of a channel operating in the reverse direction, even though the latter be noiseless and have unlimited capacity. We cannot hope, therefore, to use feedback to increase the maximum rate at which we can communicate reliably. Nevertheless, there may be considerable advantages to be gained by the use of feedback. It is the purpose of this report to explore these possibilities. In order that we may be in a position to assess the relative importance of these possible advantages, it will be helpful to review the extent to which the theoretical possibilities indicated by Shannon have been realized and the practical difficulties that would be encountered in trying to implement these schemes.

1.2 A BRIEF HISTORY OF THE CODING EFFORT

Shannon's original theorem showed that by using long blocks of channel symbols as code words, an arbitrarily small error probability could be achieved. Feinstein (3) provided an upper bound to the probability of error that decreases exponentially with the block length of the code. Elias (4) then found asymptotic upper and lower bounds on the average probability of error for the best possible binary codes of a given length, used in conjunction with the binary symmetric channel (BSC). (This channel is described by the transition diagram of Fig. 1. The probability of a correct transmission is q_o; that of an incorrect transmission, known as the crossover probability, is $p_o = 1 - q_o$.) The exponents in these bounds coincide for rates greater than a certain critical rate (4). Elias also showed that the average probability of error for the best possible code of the much smaller class of parity check-symbol (pcs) codes has the same exponential behavior as that for the class of all binary codes.

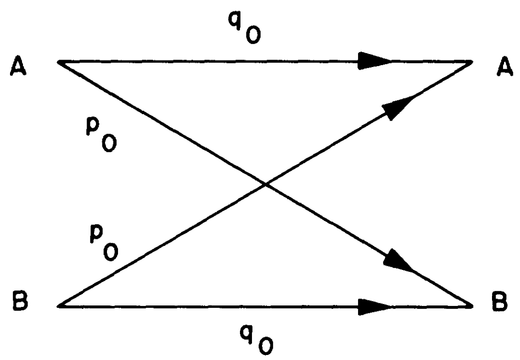

Fig. 1. The binary symmetric channel.

Mainly for reasons of mathematical tractability, most of the significant work in the construction of specific codes has been restricted to binary codes. The evaluation of these codes has been almost exclusively with respect to the binary symmetric channel.

Slepian (5) recognized that every group code is a pcs code, and vice versa. Because of the optimum properties of pcs codes found by Elias, the search for good codes with large block lengths can be restricted to the class of group codes. In Slepian's notation, (n, k) indicates a group code comprised of 2^k n-digit sequences. Slepian's decoding scheme for group codes has the advantage of being a maximum-likelihood detection scheme. However, the structure of group codes is not yet sufficiently well understood to enable us to evaluate the minimum probability of error for an arbitrarily chosen (n, k) pair or to construct the code that achieves it.

The second major direction taken by the coding effort has been the construction of codes designed to achieve an arbitrarily small probability of error. The block lengths required are typically 100 digits or more. Elias (6) introduced an iterative decoding scheme which can be used with an iterated Hamming (7) code, for example. This was the first example of a code that achieves an arbitrarily small probability of error while maintaining a nonzero transmission rate, although the rate is less than channel capacity.

The main problem with codes of large constraint length is that the decoding effort, as measured by the amount of computation or the storage space required, increases exponentially with the constraint length in maximum-likelihood detection. Wozencraft (8) proposed a convolutional code for the binary symmetric channel in conjunction with a sequential decoding scheme (which does not insist that the most likely sequence be chosen). He obtained a partial bound (which does not converge for rates too close to capacity) on the resulting average computational effort per decoded bit which grows almost linearly with the constraint length of the code. At the same time the asymptotic probability of error is equal to that of the best possible block code having the same constraint length. An experimental investigation of this scheme (9) indicates that during intervals when the number of transmission errors is much greater than the expected number, which is quite small, the required number of computations per bit is many times its average value. The number of consecutive bits for which this is true tends to be proportional to the constraint length. Thus, in a code designed to produce a very small probability of error, long delays in decoding are often encountered. These delays, in turn, are responsible for large storage requirements at the receiver.

It is evident, therefore, that a practical system of this sort should enable the transmitter to monitor closely the capacity of the channel, and provide the receiver with the opportunity of requesting retransmission if the decoding effort should become overwhelming. In fact, a knowledge of the channel capacity must be presumed by any successful coding procedure in order that the attempted rate of transmission not exceed the capacity, yet remain an appreciable fraction of the capacity, so that the transmitting facilities are used to the limit of their capabilities. If the channel capacity is not a static quantity, this implies the existence of transmitting facilities in the reverse direction. This leads us quite naturally to the consideration of two-way communication systems or, alternatively, communication with the aid of feedback.

The first instance of the use of feedback for error-correcting purposes is provided by Chang (10). In Chang's iterative discarding system, an example of what he terms an information-feedback system, the reverse channel is used to indicate each received symbol to the sender. If the symbol received through the feedback channel agrees with the symbol originally sent, it is confirmed by the transmission of the next symbol. If the two symbols disagree, the sender transmits a special erasure symbol followed by the correct symbol. When the feedback channel is error-free, the error probabilities decay to zero at an exponential rate with each subsequent confirmed symbol. Otherwise they converge to constants. This scheme can be used to transmit at capacity only under very special circumstance - namely, when there is no residual information (10) in an erased symbol.

Under the appropriate conditions, Chang's iterative discarding scheme is an example of a system in which information is transmitted essentially error-free, at channel capacity, without the use of coding. (Channel capacity was first shown to have significance in the absence of coding by Kelly (11), who showed that a gambler can increase his capital at an exponential rate governed by the capacity of the channel through which he receives his "inside information.") Elias (12) has shown that the presence of noiseless feedback makes possible communication at channel capacity, without coding or delay, over a continuous forward channel characterized by additive, white Gaussian noise.

In summary, then, although we cannot hope to increase the forward capacity of a discrete memoryless channel by using feedback, we might hope to reduce the average constraint length required to achieve a given probability of error. This would imply a reduction both in the average decoding complexity and in the average decoding delay. If, in addition, the variance of the decoding effort, which results primarily from short-term fluctuations in channel behavior, can be decreased, storage requirements at the receiver would be considerably reduced.

1.3 BLOCK-TRANSMISSION VERSUS CONTINUOUS-TRANSMISSION SYSTEMS

Whether or not we have available a feedback channel, there are two possible approaches to the encoding of information for transmission over a noisy channel. Suppose, for the sake of argument, that the information is in the form of a binary sequence of 0's and 1's, the two digits being equally likely in every position; that is, each digit of the sequence carries one bit of information. (If the information is not in this form, converting it to a binary sequence of independent digits is a useful intermediate step in matching the information source to the channel.) We may choose to divide the information sequence into blocks of fixed length and, by a prearranged encoding scheme, assign a code word composed of channel symbols to each information block. Each of the code words is then transmitted as an independent unit. If the received sequence of symbols is decoded as any information block other than the intended one, an error is said to have occurred.

On the other hand, we may use a continuous or homogeneous type of transmission system in which the information bits are introduced into the encoder one at a time. Since they are also decoded individually, we speak of a per-digit probability of error when evaluating such a scheme. Wozencraft's convolutional encoding and sequential decoding scheme is an example of a continuous transmission system with a fixed constraint length. Bits enter the encoder at a constant rate and have an opportunity to influence a fixed number of transmitted symbols. (The reader should not be misled into thinking that the transmission process is sequential in any way. Although the system can be improved through the use of a reverse channel (13), only one-way communication is implied by Wozencraft (8). The word "sequential" refers to the systematic search procedure used in decoding, which is much simpler to implement than maximum likelihood detection.) Attention will be given later to a continuous system employing a variable constraint length. In this case, bits are introduced as they are needed to maintain a fixed rate of transmission. Each bit affects as many transmitted symbols as is necessary to enable it to be decoded with the required degree of reliability.

The distinction between block transmission and continuous transmission has not been stressed in one-way systems, at least partly because the error exponent for Wozencraft's sequential decoding procedure coincides with the best that can be achieved with block coding. However, there is no a priori reason for expecting the best possible error exponents for the two types of transmission to be the same. In fact, it turns out that in the case of sequential transmission, larger exponents can be achieved with continuous systems.

II. THE NATURE OF SEQUENTIAL TRANSMISSION

2.1 CLASSIFICATION OF FEEDBACK SYSTEMS

We stated previously that the presence of a feedback channel makes possible a wide variety of sequential transmission schemes. Any sequential procedure can be described by listing, for each possible combination of transmitted message (or information sequence) and set of symbols received through the feedback channel, the symbol to be transmitted next. When the information provided by the feedback symbols about the results of previous transmissions is used to select the next transmitted symbol, we are dealing with an information-feedback system. In general, the feedback channel will be noisy; consequently, the sender may not have precise knowledge of the results of his previous transmissions. In the sequel, however, the specific schemes presented will assume that the feedback channel is noise-free, so that we may determine the greatest possible advantages that feedback can afford.

A decision-feedback system is defined by Chang (10) as one in which redundancy is added only at the request of the receiver, when he is not sufficiently certain about what has been transmitted. We shall use the term "decision feedback" in the following specific sense, in connection with block transmission. We assume that there is available at the transmitter an indefinitely long code word corresponding to each possible information block. Transmission of the appropriate code word continues until the received segment can be decoded with the required reliability. At this point the receiver indicates his satisfaction by the transmission of a prearranged symbol (or set of symbols). On receiving this symbol, the sender begins transmitting the code word corresponding to the next information block.

As the required reliability is increased, the average transmitted code-word length increases and a decoding decision is made more infrequently. Consequently, a high-reliability, decision-feedback system requires almost insignificant capacity in the reverse direction. However, the presence of the feedback channel is essential if the transmission is to be more than a one-shot operation.

It will be seen that information-feedback systems of a block type can provide exponentially better error correction than decision-feedback systems having the same average constraint length. The reason is that the former enable the sender to discriminate more effectively against those incorrect messages that have become most probable and simultaneously increase the probability of the correct message.

2.2 A DESCRIPTION OF SEQUENTIAL BLOCK-TRANSMISSION SYSTEMS

If we wish to compare the capabilities of various types of transmission systems, we must first find a general representation for each type of system from which its properties can be derived. We do this now for the class of sequential block type of transmission systems, which includes all decision-feedback systems as well as information-feedback systems using block transmission. This representation, which is an extension of one often used with block type of systems with fixed constraint lengths, will be particularly useful in connection with information-feedback systems.

Since the usefulness of the model depends on the ease with which the channel noise can be represented, we shall confine our efforts to systems which have a binary symmetric channel in the forward direction. The effect of the channel noise on a block of N transmissions is then represented by a sequence of N 0's and 1's, in which a 0 is used to represent a correct reception and a 1 indicates a crossover.

The reverse channel is not restricted in any way. It may be noisy or noiseless, and it may have any number of input and output symbols. In fact, the transmission of information about the symbols received may be only one of its functions.

A total of, let us say, M bits is to be transmitted over the forward channel by the choice of 1 of 2^M equally likely messages. The sender determines each transmitted symbol according to the message selected and his knowledge of the results of all previous transmissions. The receiver decodes the received sequence of binary symbols as the most probable message as soon as the probability of error in doing so has been reduced to a specified value, which we shall call P_e.

The receiver's decision process can be summarized by dividing the possible received sequences of length N, $N \geq 1$, into 2^M+1 disjoint sets. Corresponding to each message is the set comprised of those sequences that are decoded as that message. The remaining set consists of all those sequences for which no decision is made at length N. Only those sequences which are not extensions of decodable sequences are listed. Nevertheless, the total number of sequences listed can be kept finite only by an agreement to terminate the transmission process after a prearranged number of transmissions.

A somewhat more detailed picture of the transmission process is afforded by a set of tables, one for each value of N, like that in Fig. 2. In each table there are 2^M rows,

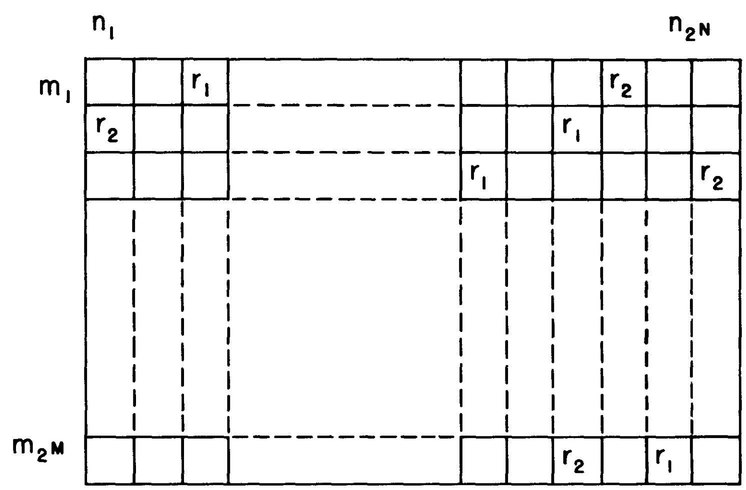

Fig. 2. Typical decoding table.

each one identified with a different message. The 2^N noise sequences are listed across the top of the table, from left to right, in order of decreasing probability. Sequences with the same number of crossovers, being equally probable, are listed according to the binary numbers they represent, in decreasing numerical order. In each square of the table is listed the received sequence that would result from the selection of the message corresponding to its row and the occurrence of the noise sequence corresponding to its column. A square is left blank if the corresponding message-noise pair would result in a decision prior to the N^{th} transmission.

Regardless of the message chosen, the appropriate noise sequence can cause any of the possible received sequences to occur. Therefore, each received sequence that is listed for a given value of N appears exactly once in each row. If a received sequence is decoded on the N^{th} transmission, it is associated with the message corresponding to the row in which it appears farthest to the left. If this happens to be the j^{th} row, and if we designate by P_{n_i} the probability of the noise sequence corresponding to the square in the i^{th} row in which the sequence appears, the resulting probability of error is

$$\sum_{\substack{i=1\\ i\neq j}}^{2^M} P_{n_i} \bigg/ \sum_{i=1}^{2^M} P_{n_i}$$

If this is no larger than P_e, the sequence under consideration will, in fact, be decoded on the N^{th} transmission.

2.3 THE SEQUENTIAL ERROR EXPONENT

The asymptotic average error probability, $\overline{P}_e$, of a transmission system of fixed constraint length can be expressed in the form

$$\overline{P}_e \doteq 2^{-E_1(R, C)N}$$

in which the relation $x \doteq y$ is used as an abbreviation for

$$\lim_{N\to\infty} \frac{\log x}{N} = \lim_{N\to\infty} \frac{\log y}{N}$$

and N is the constraint length, R is the transmission rate, and C is the channel capacity. Thus E_1 characterizes the asymptotic error behavior of the system.

In a sequential system, on the other hand, the probability of error, P_e, is a fixed, preset number, and the constraint length is a variable quantity. The sender uses information received through the feedback channel to insure that the number of transmitted symbols influenced by each information digit is no greater than that necessary to achieve the specified value of P_e. The average constraint length, $\overline{N}$, required to do this is asympotically related to P_e by an expression for the form

$$P_e \doteq 2^{-E_2(R,C)\overline{N}}$$

in which the symbol $\doteq$ has the same interpretation as before, except that N is replaced by $\overline{N}$.

Let us examine the simplest conceivable situation, for which E_1 and E_2 are well known (14). Suppose that a binary symmetric channel with crossover probability p_o is used to communicate the result of an event having two possible outcomes each of which is equally likely a priori. Label the two outcomes A and B, and let the channel inputs be 0 and 1.

If we are restricted to using the channel N times, it is clear that the smallest possible average probability of error is achieved by the following scheme: Send N 0's for outcome A, let us say, and N 1's for outcome B. The receiver chooses A or B according to whether a majority of the received symbols are 0's or 1's. Assuming for the sake of definiteness that N is odd, $\overline{P}_e$ is then given by the probability that more than half the transmissions have been received incorrectly.

$$\overline{P}_e = \sum_{k=\frac{N+1}{2}}^{N} \binom{N}{k} p_o^k q_o^{N-k}$$

Using approximations that will be justified later, we find that

$$\overline{P}_e \doteq \binom{N}{\frac{N+1}{2}} p_o^{\frac{N+1}{2}} q_o^{\frac{N+1}{2}} \doteq 2^{N\left(1+\frac{1}{2}\log p_o+\frac{1}{2}\log q_o\right)} = 2^{-\left(\frac{1}{2}\log\frac{1}{4p_oq_o}\right)N}$$

Therefore,

$$E_1 = \frac{1}{2}\log\frac{1}{4p_oq_o}$$

If, on the other hand, we are allowed to send as many symbols as are required to achieve a probability of error of, at most, P_e with each decision, the optimum strategy is again to send all 0's or all 1's, depending on whether A or B has occurred. In calculating E_2, we need only consider the case in which A has occurred because of the obvious symmetry of the situation. Zeros are continually transmitted until the probability of B has been reduced from its a priori value of 1/2 to P_e. At this point the receiver will have been provided with log $2P_e$ bits of information about outcome B. If we let $\overline{I}_B$ represent the limiting value, as the probability of B approaches zero, of the average information provided per transmission about B, it can be shown (14) that

$$\lim_{P_e\to 0} \frac{\overline{N}}{\log 2P_e/\overline{I}_B} = 1$$

(The use of $\overline{I}_B$ as a measure of the average information provided about B per transmission is justified heuristically by the fact that the asymptotic behavior, as $P_e \to 0$, dominates the entire process.) It follows that

$$P_e \doteq 2^{\bar{I}_B \bar{N}}$$

Consequently, $E_2 = -\bar{I}_B$. We must now calculate $\bar{I}_B$.

For this purpose we consider the situation in Fig. 3, in which the probability of outcome B has been reduced to ϵ, let us say. The input symbols 0 and 1 therefore

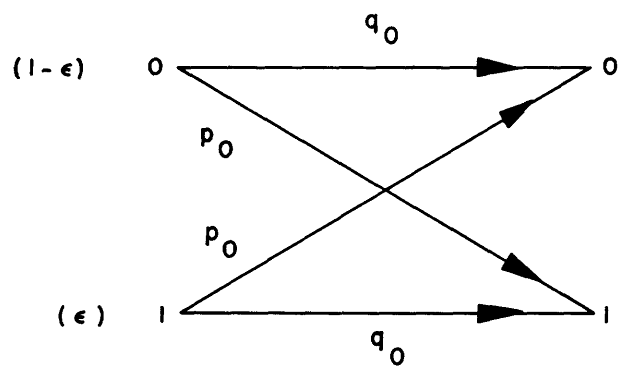

Fig. 3. Asymptotic behavior of sequential 1-bit transmission.

have as a priori probabilities, $1-\epsilon$ and ϵ, respectively. If we define $\bar{I}_B(\epsilon)$ to be the average information received about the symbol 1 when a 0 is sent,

$$\bar{I}_B(\epsilon) = q_o \log \frac{p_o}{(1-\epsilon)\, q_o + \epsilon p_o} + p_o \log \frac{q_o}{(1-\epsilon)\, p_o + \epsilon q_o}$$

Since

$$\bar{I}_B = \lim_{\epsilon \to 0} \bar{I}_B(\epsilon)$$

Then

$$\bar{E}_2 = -\bar{I}_B = (q_o - p_o) \log \frac{q_o}{p_o}$$

Curves of E_1 and E_2 are plotted in Fig. 4 as a function of the channel parameter p_o. The exponential superiority of the sequential procedure is quite clear; that is, the asymptotic values of $\bar{N}$ and N required to achieve specified values of P_e and $\bar{P}_e$, respectively, satisfy the relation, $\bar{N} < N$, if $P_e = \bar{P}_e$. E_1 and E_2 will be referred to as the fixed-length and sequential 1-bit error exponents, respectively.

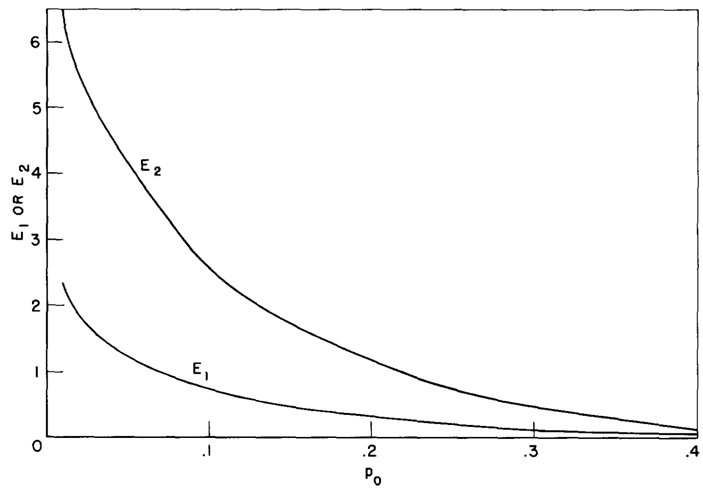

Fig. 4. Error exponents for fixed-length and sequential 1-bit transmission.

2.4 SUMMARY OF RESULTS

Information-feedback, block-transmission systems are discussed first, in section 3.1. An upper bound on the largest attainable error exponent for such a system is derived, based solely on the model of Fig. 2. The limiting values of this bound, as R approaches C and zero, respectively, are equal to the fixed-length and the sequential 1-bit exponents. It is shown in Section IV, where a specific information-feedback system (both continuous and block-type) is analyzed, that this bound cannot be achieved by a block-transmission system (section 4.5), although it can be realized by a continuous system (section 4.4). The largest block-transmission error exponent is a linear function of R, varying between zero at R = C to the sequential 1-bit exponent at R = 0. This asymptotic superiority of continuous (over block) transmission systems has neither been demonstrated nor disproved for one-way systems.

In section 3.2, where decision-feedback systems are discussed, it is conjectured that the asymptotic error behavior of a sequential sphere-packed code (one that assumes the existence of a sphere-packed array of code points for every value of the constraint length) is at least equal to that of any decision-feedback system. Although the error exponent for such a hypothetical code has the same limiting values, as R approaches C and zero, as the optimum information-feedback block-transmission system, it is smaller at all intermediate rates. However, it is greater, at all rates less than capacity, than the optimum error exponent for fixed-constraint-length block codes.

Except for section 4.5, Section IV is devoted to an evaluation of a continuous version of the information-feedback system referred to above. Bounds governing the over-all process are derived in section 4.2 for R = C. The operation of the entire transmission system, again for R = C, was simulated on an IBM 709 computer. The results are presented in section 4.3. Of particular significance is the fact that the average constraint length required to achieve an error probability of 10^{-15}, at a rate essentially equal to channel capacity, was only about one-seventh of the constraint length required by the optimum fixed-constraint-length block code, when the latter is designed for a rate of only half capacity. The extension of this scheme to other discrete memoryless channels is considered briefly in section 4.6.

III. BOUNDS ON SEQUENTIAL BLOCK-TRANSMISSION ERROR EXPONENTS

We consider a two-way communication system, in which we identify one direction as the forward direction and the other as the reverse direction, and ask the following question: To what extent can communication from a sender to a receiver over the forward channel be facilitated by means of information sent over the reverse channel? Specifically, we wish to find an asymptotic lower bound (as $P_e \to 0$) on the average number of transmissions required to achieve a probability of error P_e, at an average rate of transmission R, when the forward channel is a binary symmetric channel with crossover probability p_o. It is evident from the discussion of section 2.3 that the quantity we seek is a bound on the sequential error exponent. We shall treat both information-feedback and decision-feedback systems.

3.1 AN UPPER BOUND ON INFORMATION-FEEDBACK BLOCK-TRANSMISSION ERROR EXPONENTS

We wish to find a bound that will be valid for all possible feedback channels and valid regardless of the information fed back to the sender. We therefore assume that there is a noiseless binary channel in the reverse direction which the receiver uses to notify the sender of each received symbol. This information is available to the sender before the transmission of the next symbol.

We can summarize the behavior of any information-feedback block-transmission system by a set of tables like that in Fig. 2. Each of these tables can then be converted to a table of identical dimensions that indicates those combinations of message and noise that would result in a correct decision, an incorrect decision, or no decision at all. The letters A, C, and B, respectively, will be used to indicate the three possibilities. To convert from the first set of tables to the second, we replace each received sequence that is decoded on the N^{th} transmission by an A in its leftmost position and by a C everywhere else. A received sequence that is not decoded is replaced by a B wherever it appears. Note that there are $(2^M - 1)$ C's for each A.

Let us imagine that P_e represents an average allowed error probability at each value of N (the average being taken over all received sequences for which a decision is made), rather than a probability of error that must be met by each sequence that is decoded. We may then decode additional received sequences, for which the probability of error is greater than P_e, and still keep the average probability of error less than P_e. Since this can only decrease the average number of transmissions required to reach a decision, we may interpret P_e in this way for the purpose of calculating the desired bound.

Let P_A represent the total probability of the noise sequences associated with the A's, counting each sequence as many times as there are A's in that column; P_B and P_C are similarly defined. An arrangement of A's, B's, and C's is allowed only if

$$P_e \geqslant \frac{P_C}{P_A + P_C}$$

If we let N_A, N_B, and N_C stand for the number of A's, B's, and C's, respectively, we have the additional constraint

$$N_C = (2^M-1)\, N_A \tag{1}$$

We first find the arrangement of A's, B's, and C's, consistent with P_e, that maximizes the probability of decoding on the N^{th} transmission. (Note that a set of tables, so arranged for all $N \leqslant N_o$, let us say, does not necessarily imply that the probability of having reached a decision by the $N_o{}^{th}$ transmission has been maximized.) Consider the arrangement in Fig. 5, where Eq. 1 is satisfied and

$$P_e \doteq \frac{P_C}{P_A + P_C} \tag{2}$$

We shall show that any other arrangement yielding the same average error probability must result in a smaller probability of decoding on the N^{th} transmission.

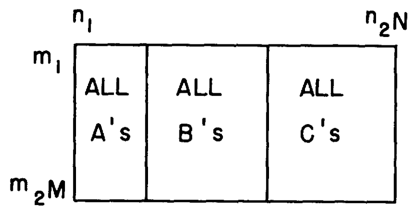

Fig. 5. Table maximizing probability of decoding.

Consider any other arrangement. We interchange the A's with the B's and/or C's until no B or C is farther to the left than any A. Each interchange must reduce the average probability of error and, in the case of the B's, it also increases the probability of decoding. We then associate with each letter of the altered table the probability of its occurrence, which is 2^{-M} times the probability of the corresponding noise sequence, and interchange subregions of B's and C's having (approximately) the same total probability until no B is farther to the right than any C. This does not affect the probability of error or the probability of decoding. However, each interchange increases the number of C's because each shifted block of C's has been moved to a region in which individual letters have smaller probabilities. In order to maintain the relation, $N_C = (2^M-1)\, N_A$, we must change some of the remaining B's to A's. This decreases the average probability of error still further. Finally, we can restore this probability to the value P_e by adding A's and C's in the correct relative numbers. This further increases the probability of decoding.

We shall find an asymptotic lower bound to the average number of transmissions required to achieve an average error probability of P_e by using an arrangement of A's, B's, and C's like that in Fig. 5 to bound the probability of decoding at each value of $N \geq 1$. The boundaries of the B region in Fig. 5 can be found by using constraints, Eqs. 1 and 2. Instead of using Eq. 2, we shall measure the average probability of error by the joint probability of decoding and making an error; that is, we shall replace Eq. 2 by

$$P_e \doteq P_C \tag{3}$$

Since $P_C \leq P_C/(P_A+P_C)$, this allows more A's and C's to be included in the table of Fig. 5 and thereby increases the probability of decoding on the N^{th} transmission. It is therefore a permissible substitution.

We shall need the following relation, which makes use of Stirling's approximation. Here $0 < \alpha < 1$ and $\beta = 1 - \alpha$.

$$\begin{aligned}\binom{N}{\alpha N} &= \frac{N!}{(\alpha N)!\,(\beta N)!} \doteq \frac{N^N}{(\alpha N)^{\alpha N}\,(\beta N)^{\beta N}} \\ &= 2^{N\log N - \alpha N\log(\alpha N) - \beta N\log(\beta N)} \\ &= 2^{N[\alpha\log N + \beta\log N - \alpha\log(\alpha N) - \beta\log(\beta N)]} \\ &= 2^{N[-\alpha\log\alpha - \beta\log\beta]} \\ &= 2^{NH(\alpha)}\end{aligned}$$

where $H(\alpha) = -\alpha \log \alpha - (1-\alpha) \log (1-\alpha)$.

In addition, we need the relations:

$$\sum_{j=0}^{L} \binom{N}{j} \doteq \binom{N}{L} \qquad 0 < \frac{L}{N} < \frac{1}{2}$$

$$\sum_{k=L}^{N} \binom{N}{k} p_o^k q_o^{N-k} \doteq \binom{N}{L} p_o^L q_o^{N-L} \qquad p_o < \frac{L}{N} < 1$$

Their validity follows from the fact that $\binom{N}{j}$ is an increasing function of j if $\frac{j}{N} < \frac{1}{2}$, and $\binom{N}{k} p_o^k q_o^{N-k}$ is a decreasing function of k if $\frac{k}{N} > p_o$. Consequently, we have

$$\binom{N}{L} < \sum_{j=0}^{L} \binom{N}{j} < (L+1) \binom{N}{L}$$

$$\binom{N}{L} p_o^L q_o^{N-L} < \sum_{k=L}^{N} \binom{N}{k} p_o^k q_o^{N-k} < (N-L-1) \binom{N}{L} p_o^L q_o^{N-L}$$

The relation, $N_C = (2^M-1)\, N_A$, is equivalent to

$$\sum_{k=k_2}^{N} \binom{N}{k} = (2^M-1) \sum_{j=0}^{j_1} \binom{N}{j}$$

where j_1 and k_2 are the number of crossovers in the noise sequences bounding the B region. (In general an exact equality is not possible using integral values for j_1 and k_2; however, the form of the succeeding approximations permits us to treat j_1 and k_2 as continuous parameters.) Since it will be seen that $k_2 > \frac{N}{2}$,

$$\binom{N}{k_2} \doteq 2^M \binom{N}{j_1}$$

If we define $q_2 = 1 - p_2 = \frac{k_2}{N}$ and $p_1 = 1 - q_1 = \frac{j_1}{N}$:

$$2^{NH(q_2)} \doteq 2^{M+NH(p_1)}$$

$$H(q_2) = \frac{M}{N} + H(p_1) \tag{4}$$

Eq. 3 becomes

$$P_e \doteq P_c = \sum_{k=k_2}^{N} \binom{N}{k} p_o^k q_o^{N-k}$$

$$\doteq \binom{N}{k_2} p_o^{k_2} q_o^{N-k_2}$$

$$\doteq 2^{N[H(q_2) + q_2 \log p_o + p_2 \log q_o]}$$

$$= 2^{N[H(q_2) - H(p_o) - p_o \log p_o - q_o \log q_o + q_2 \log p_o + p_2 \log q_o]}$$

$$= 2^{-N[(q_2 - p_o) \log q_o/p_o + H(p_o) - H(q_2)]} \tag{5}$$

Equations 4 and 5 yield the boundaries of the B region for any value of N.

For fixed values of P_e and M, specification of any one of the quantities j_1, k_2, and N fixes the other two. For each N, j_1 pair we have already defined p_1 by the relation $j_1 = p_1 N$. Let N_1, j_{11} be the N, j_1 pair for which $p_1 = p_o$, and let q_{21} be the corresponding value of q_2. By making the appropriate substitutions in Eqs. 4 and 5, we find that N_1 is given by

$$P_e \doteq 2^{-N_1[(q_{21} - p_o) \log q_o/p_o - M/N_1]} \tag{6}$$

where q_{21} is determined (as a function of N_1) by the relation,

$$H(q_{21}) = \frac{M}{N_1} + H(p_o) \qquad (7)$$

For fixed values of P_e and M, and any $\delta > 0$, let $N_2(\delta)$, j_{12} be the N, j_1 pair for which $j_1 = (p_o - \delta)$ N. Equation 5 shows that as N increases, q_2 decreases, and Eq. 4 indicates that this must be accompanied by an increase in p_1. It follows that $N_2(\delta) < N_1$. In Appendix A-1 it is shown that

$$\lim_{P_e \to 0} \Pr[N' < N_2(\delta)] = 0 \qquad (8)$$

for all $\delta > 0$, where N' is the number of transmissions required by the transmission scheme that actually minimizes the average number of transmissions. $\left(\frac{M}{N_1}\right.$ is held fixed as $\lim_{P_e \to 0}$ is taken. This implies, according to Eqs. 6 and 7, that $\frac{M}{\log P_e^{-1}}$ must be held constant during this limiting process.$\left.\right)$ Equation 8 implies that

$$\lim_{P_e \to 0} \frac{N'}{N_2(\delta)} \geqslant 1 \qquad (9)$$

However, it is also proved in Appendix A-1 that

$$\lim_{\delta \to 0} \lim_{P_e \to 0} \frac{N_2(\delta)}{N_1} = 1 \qquad (10)$$

Combining Eqs. 9 and 10, we obtain

$$\lim_{\delta \to 0} \lim_{P_e \to 0} \frac{\overline{N'}}{N_1} = \lim_{P_e \to 0} \frac{\overline{N'}}{N_1} \geqslant 1 \qquad (11)$$

With the aid of Eqs. 6, 7, and 11, we can summarize the results of this section by Theorem 1.

<u>Theorem 1</u>

We have given a binary symmetric channel with crossover probability p_o and a noiseless binary feedback channel which the receiver uses to inform the transmitter of each received symbol. An asymptotic lower bound on the average number of transmissions, $\overline{N}$, required to transmit a finite amount of information at rate R (by choosing 1 of 2^M equally probable messages) with an error probability of at most P_e is determined by

$$P_e = 2^{-\overline{N}[(q^* - p_o) \log q_o/p_o - R]} \qquad (12)$$

where q* is determined by

$$H(q^*) = R + H(p_o) \qquad (13)$$

and the rate R for the process is defined by $R = M/\overline{N}$.

Once the lower bound on $\overline{N}$ has been determined from Eq. 12, a lower bound on the value of M required to achieve a rate R is given by $M \geq R\overline{N}$. That M is an increasing function of R can be seen from the following argument. Consider an increased rate $R' = R + dR$, and indicate by primes all quantities corresponding to R'. Equation 13 shows that $q^{*\prime} < q^*$. Define $\overline{N}''$ by the relation, $R'\overline{N}'' = R\overline{N}$. Since $R' > R$, then $\overline{N}'' < \overline{N}$. It follows that

$$\overline{N}''\left[(q^{*\prime}-p_o)\log\frac{q_o}{p_o}-R'\right] < \overline{N}\left[(q^*-p_o)\log\frac{q_o}{p_o}-R\right] \tag{14}$$

It is clear from Eqs. 12 and 14 that in order to achieve the same probability of error, it must be true that $\overline{N}' > \overline{N}''$. Hence $M' = R'\overline{N}' > R\overline{N}'' = R\overline{N} = M$.

Now let us examine the exponent, E(R), characterizing the asymptotic relation between P_e and $\overline{N}$.

$$E(R) = (q^*-p_o)\log\frac{q_o}{p_o} - R$$

To find $\lim_{R\to 0} E(R)$, we first note from Eq. 13 that $\lim_{R\to 0} q^* = q_o$. Therefore

$$\lim_{R\to 0} E(R) = (q_o-p_o)\log\frac{q_o}{p_o} \tag{15}$$

Similarly, $\lim_{R\to C} q^* = \frac{1}{2}$ and

$$\lim_{R\to C} E(R) = \left(\frac{1}{2}-p_o\right)\log\frac{q_o}{p_o} - p_o\log 2p_o - q_o\log 2q_o = \frac{1}{2}\log\frac{1}{4p_oq_o} \tag{16}$$

The shape of the curve of E(R), $0 \leq R < C$, is revealed by an examination of $\frac{dE(R)}{dR}$:

$$\frac{dE(R)}{dR} = \left(\log\frac{q_o}{p_o}\right)\frac{dq^*}{dR} - 1 = \frac{\log(q_o/p_o)}{dR/dq^*} - 1$$

From Eq. 13 we find that $dR/dq^* = \log(p^*/q^*)$, where $p^* = 1 - q^*$, and hence

$$\frac{dE(R)}{dR} = -\frac{\log(q_o/p_o)}{\log(q^*/p^*)} - 1 \tag{17}$$

It follows that
$$\lim_{R\to 0}\frac{dE(R)}{dR} = -2 \tag{18}$$

$$\lim_{R\to C}\frac{dE(R)}{dR} = -\infty \tag{19}$$

Furthermore, Eqs. 13 and 17 imply that $dE(R)/dR$ is a strictly decreasing function of R. This fact, together with Eqs. 15, 16, 18, and 19, is sufficient to characterize E(R), which is plotted in Fig. 6 for $p_o = 0.1$.

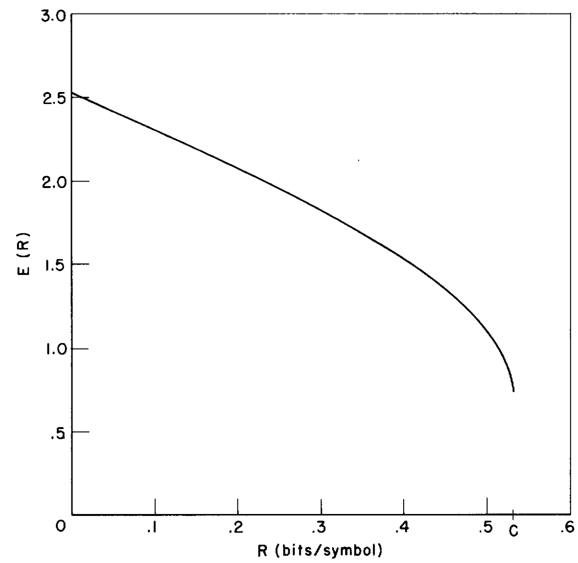

Fig. 6. E(R) versus R for $p_o = 0.1$.

3.2 AN UPPER BOUND ON THE SEQUENTIAL SPHERE-PACKING ERROR EXPONENT

We turn now to the consideration of decision-feedback systems. Once again M bits are transmitted by the choice of 1 of 2^M equally likely messages, and the entire discussion in relation to Fig. 2 is applicable here. However, it now makes sense to associate with each of the messages a variable-length binary (0's and 1's) code word. As much of the code word is transmitted as is needed for the receiver to decode it with the required degree of reliability.

The key to bounding the largest achievable error exponent lies in finding a set of decoding tables having an arrangement of A's, B's, and C's (such as that in Fig. 5 for the case of information feedback) that bounds the probability that decoding has occurred after a given number of transmissions. For this purpose we assume the existence of a code in which, for each received-sequence length N, the 2^M code words, considered as points in an N-dimensional space, form a sphere-packed array. In other words, if we define the distance between any two points in this space as the number of places in which the corresponding sequences differ, and if we consider a sphere about each code point which includes all those points whose distance from the code point is, let us say, $r_1 = p_1N$ or less, each of the 2^N sequences of length N will lie in one, and only one, of these spheres. (The word "sphere" will be used only in reference to one of these sets of points.) A received sequence is decoded only if its distance from the nearest code point is, let us say, $r_2 = p_2N$ or less, where $p_2 \leq p_1$ and r_2 is a function of p_o, P_e, N, and M (or, alternatively, p_o, P_e, N, and p_1). The decoding table for such a system is shown in Fig. 7. A set of like tables, one for each value of N, will be said to define a sequential sphere-packed code and its decoding scheme.

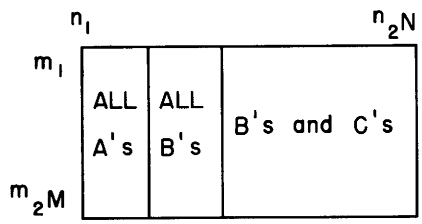

Fig. 7. Decoding table for sequential sphere-packed code.

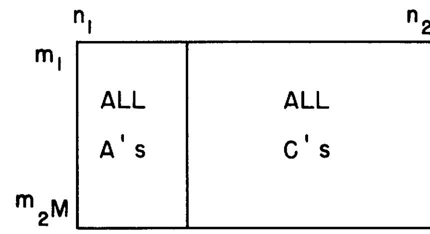

Fig. 8. Decoding table for sphere-packed code as $R \rightarrow C$.

Such codes do not exist, in general; however, it is reasonable to expect that for the situation depicted in Fig. 7, in which $P_C \doteq P_e$, $P_A + P_C$ bounds the probability that decoding has taken place after N transmissions, for any decision-feedback system. (As in the information-feedback case, it is permissible to treat P_e as an average allowed error probability.)

In support of this conjecture we observe that, as P_e approaches 0, the minimum value of N at which decoding can take place grows without limit. Consequently, the fractional number of crossovers that has occurred when decoding takes place can be set equal to p_o. It follows that, as R approaches C, the value of p_1 at which decoding takes place approaches p_o. The corresponding decoding table, shown in Fig. 8, is also the maximum-likelihood decoding table for a (fixed-constraint-length) sphere-packed block code. For the class of block codes of length N, it yields the smallest value of $\overline{P}_e$. If we assume, therefore, that decoding for the sequential sphere-packed code takes place at the value of N for which $r_2/N = p_o$, the arrangement of A's and C's in Fig. 8 requires the fewest transmissions for a given value of P_e.

We shall proceed under the assumption that a set of tables like that in Fig. 7 does, in fact, yield a probability of decoding which is at least as great as that of any realizable decision-feedback scheme. Subject to the truth of this conjecture, the bounding relations we obtain for a sequential sphere-packed code will apply equally well to any decision-feedback system.

The sequential sphere-packing results are interesting in their own right. In the case of fixed-length block codes it has been shown (4) that, at rates greater than the critical rate referred to earlier, codes exist for which maximum-likelihood detection yields an error exponent equal to the exponent that would result from a sphere-packed code. This gives us reason to expect that variable-length codes whose asymptotic error behavior is the same as that of a sequential sphere-packed code do exist.

In order to underbound the average constraint length (or overbound the error exponent) of a sequential sphere-packed code, we must first find the boundary of the decoding region for each value of N. We shall do this by expressing P_e as a function of p_o, p_1, N, and r_2. Because of the symmetry of a sphere-packed code, the probability of error is independent of the code point transmitted. We may therefore assume that the origin of coordinates (which represents a code word composed entirely of 0's) is the code point that is transmitted. A decoding error is made if the received sequence falls within a part of the decoding region surrounding one of the other code points before it is first included in the part of the decoding region centered on the origin (where it must eventually appear, with probability one).

We shall include in our calculation of P_e only the contributions of those points closest to the origin whose reception results in a decoding error. These points have the smallest number of 1's of all the points in the decoding region other than those in the sphere surrounding the origin. Underbounding P_e in this way increases the radius of each decoding sphere for a given value of N. The resulting asymptotic bound on $\overline{N}$ is therefore less than the value that would result from a sequential sphere-packed code. This question is discussed at greater length in Appendix B, where it is shown that, if P_e is actually to be treated as an average quantity, the difference between the true value of the error exponent and the value we get by using this approximation is extremely small. In any event, the value we obtain is an upper bound on the error exponent that would result from a sequential sphere-packed code.

The indicated lower bound on P_e can be expressed as the product of three factors: the number of code points nearest the origin, the number of points closest to the origin in the part of the decoding region surrounding one of these code points, and the probability with which one of these points is received.

We note first that the number of points at a distance $r_1 + 1$ from the origin is $\binom{N}{r_1+1}$. Each of these points is a distance r_1 from one (and only one) of the code points nearest the origin. This follows directly from the definition of a sphere-packed code. The number of these points in the sphere of the same code point is $\binom{2r_1+1}{r_1}$, which is the number of ways r_1 1's, of the $(2r_1+1)$ 1's in each code point, can be changed to 0's. The number of code points nearest the origin, n_1, is therefore given by

$$n_1 = \frac{\binom{N}{r_1+1}}{\binom{2r_1+1}{r_1}} \tag{20}$$

Each of the points closest to the origin whose reception results in a decoding error has $(2r_1+1-r_2)$ 1's. The number of these points belonging to the part of the decoding region about a given code point having $(2r_1+1)$ 1's is clearly $\binom{2r_1+1}{r_2}$. Furthermore, each of these points is received with probability $p_o^{2r_1+1-r_2}\, q_o^{N-2r_1-1+r_2}$. If we combine these two facts with Eq. 20, we find that

$$P_e \geqslant \frac{\binom{N}{r_1+1}}{\binom{2r_1+1}{r_1}} \binom{2r_1+1}{r_2} p_o^{\,2r_1+1-r_2}\, q_o^{\,N-2r_1-1+r_2}$$

$$= \frac{N!\, r_1!\, 2^{(2r_1+1-r_2)\log p_o}\, 2^{(N-2r_1-1+r_2)\log q_o}}{(N-r_1-1)!\, r_2!\,(2r_1+1-r_2)!}$$

$$\dot{\approx} \frac{N^N r_1^{\,r_1}\, 2^{N\log q_o + (r_2-2r_1)\log (q_o/p_o)}}{(N-r_1)^{N-r_1}\, r_2^{\,r_2} (2r_1-r_2)^{2r_1-r_2}} \tag{21}$$

Therefore,

$$\lim_{N\to\infty} \frac{\log P_e^{-1}}{N} \leqslant q_1 \log q_1 + p_2 \log p_2 + (2p_1-p_2)\log (2p_1-p_2)$$
$$- p_1 \log p_1 - \log q_o + (2p_1-p_2)\log (q_o/p_o) \tag{22}$$

We are particularly interested in this relation when N is such that $p_2 = p_o$. If we label this block length N_1 and add the subscript one to indicate the corresponding values of those parameters that vary with N, Eq. 22 becomes

$$\lim_{P_e\to 0} \frac{\log P_e^{-1}}{N_1} \leqslant q_{11} \log q_{11} - p_{11} \log p_{11} + (2p_{11}-p_o)\log (2p_{11}-p_o)$$
$$- (1+p_o-2p_{11})\log q_o - 2(p_{11}-p_o)\log p_o \tag{23}$$

We define $N_2(\delta)$ as the block length for which $p_2 = p_o - \delta$. It can be shown, just as it is for the case of information feedback in Appendix A-1, that for any $\delta > 0$,

$$\lim_{P_e\to 0} \Pr[N<N_2(\delta)] = 0 \tag{24}$$

Also, it is clear from Eqs. 22 and 23 that

$$\lim_{\delta\to 0} \lim_{P_e\to 0} \frac{N_2(\delta)}{N_1} = 1 \tag{25}$$

Combining Eqs. 24 and 25, we have

$$\lim_{P_e\to 0} \frac{\overline{N}}{N_1} \geqslant \lim_{\delta\to 0} \lim_{P_e\to 0} \frac{\overline{N}}{N_2(\delta)} \frac{N_2(\delta)}{N_1} \geqslant 1 \tag{26}$$

Substitution of Eq. 26 in Eq. 23 yields Theorem 2.

Theorem 2

We have given a binary symmetric channel with crossover probability p_o and a noiseless feedback channel. An asymptotic lower bound on the average number of transmissions, $\overline{N}$, needed to transmit a finite amount of information at rate R (by the choice of 1 of 2^M equally probable messages) by means of a sequential sphere-packed code, with an error probability of at most P_e, is given by

$$\overline{N} \geq \begin{bmatrix} q_1 \log q_1 - p_1 \log p_1 + (2p_1 - p_o) \log (2p_1 - p_o) \\ -(1+p_o-2p_1) \log q_o - 2(p_1 - p_o) \log p_o \end{bmatrix}^{-1} \log P_e^{-1} \tag{27}$$

where $H(p_1) = 1 - R(p_1)$ and a lower bound on the value of M required is given by $M \geq R\overline{N}$.

The upper bound on the error exponent, which is the term in brackets in Eq. 27, is plotted (curve B) in Fig. 9. It can be verified that the slope of the curve is infinite at R = 0, and that it is equal to -1 at R = C. The adjacent dashed curve is the lower bound on the sequential sphere-packing error exponent derived in Appendix B.

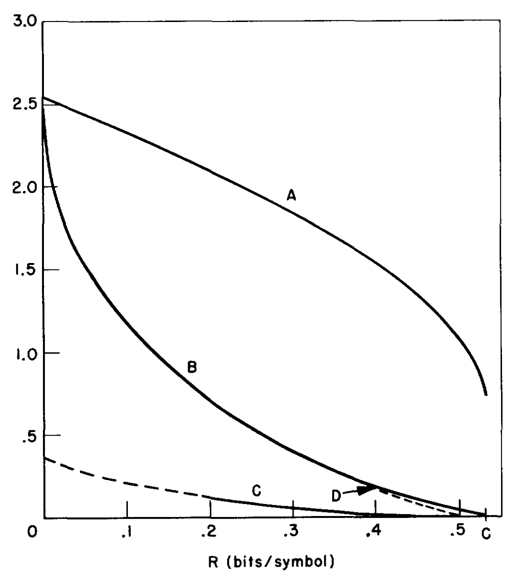

Fig. 9. Sequential and fixed-length error exponents for $p_o = .1$.

Also shown in Fig. 9 are the bounds on the information-feedback error exponent previously derived (curve A) and the optimum fixed-length block-code error exponent (solid part of curve C) for rates greater than the critical rate. Above the critical rate the latter curve coincides with the exponent that would result from a sphere-packed code. The dashed portion of this curve indicates that the value of the exponent is not known below the critical rate except at R = 0, where it is equal to half the fixed-length 1-bit exponent. This exponent would be a point on curve C if the sphere-packing exponent were applicable at all values of R < C.

IV. AN INFORMATION-FEEDBACK SYSTEM

The discussion thus far has been concerned with upper bounds on the error-correcting capabilities of sequential transmission systems. In an effort to determine what actually can be achieved by a sequential system, we turn now to the consideration of a particular type of information-feedback, continuous transmission system in which the feedback channel is assumed to be noiseless.

Following a description of the system when it operates essentially at channel capacity, bounds on the average constraint length are given. In particular, it is shown that the limiting error exponent as the rate approaches capacity is equal to $\frac{1}{2}\log\frac{1}{4p_o q_o}$, the exponent for fixed-length, 1-bit transmission. Furthermore, as a function of the rate of transmission, the error exponent coincides with the bound for information-feedback, block-transmission derived in section 3.1. Experimental results for transmission at capacity are also presented. Next, a block version of this system is evaluated. Its error exponent, which is greater than the sequential sphere-packing exponent for all rates less than capacity, is the best achievable with an information-feedback, block-transmission system. However, it is smaller than the bound previously derived for a system of this type. Finally, the application of this information-feedback, continuous transmission scheme to more general discrete channels is discussed briefly.

4.1 DESCRIPTION OF THE SYSTEM

It is assumed that the message to be transmitted is in the form of a sequence of independent binary digits in which 0's and 1's occur with equal probability. (It will become apparent that this is a suitable form for the data regardless of the number of channel symbols.) A binary symmetric channel with crossover probability $p_o < 1/2$ is assumed for the forward direction. The receiver uses the error-free feedback channel to indicate to the transmitter the symbol just received. This information is available at the transmitter when the next channel symbol is selected.

If we imagine that a binary point is placed to the left of the message sequence, which is written from left to right, we can regard the sequence as a binary fraction. Consider the interval (0, 1) on the line of real numbers. A sequence of infinite length is represented by a point in this interval; a sequence of length N, by the subinterval of length 2^{-N} whose left end point is represented by the sequence that is being considered. The receiver has no a priori knowledge of the message sequence. Since the density of infinite binary sequences on the interval (0, 1) is uniform, the receiver's initial probability distribution for the location of the point representing the message sequence (which we shall call the transmitted point) is also uniform. This distribution will be referred to as the receiver's distribution.

The interval (0, 1) is divided in half, as shown in Fig. 10a. If the transmitted point lies in the lower half, A is the first transmitted symbol; otherwise B is sent. Let us assume that P is the transmitted point, so that A is sent. With probability

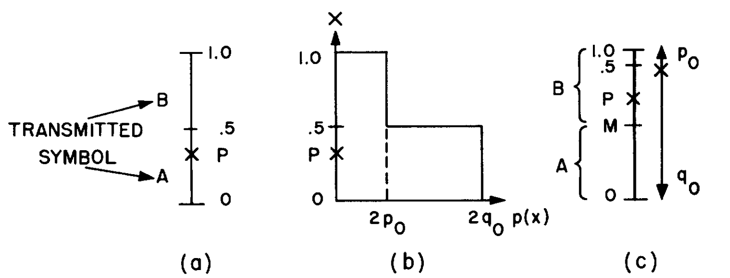

Fig. 10. Transmission process.

$q_o = 1 - p_o$, A is received, and Fig. 10b represents the receiver's new distribution, in accordance with the application of Bayes' rule to the binary symmetric channel. We choose to use, instead, the representation of Fig. 10c. The interval (0, .5) has been stretched uniformly by a factor of $2q_o$, and the interval (.5, 1) has been compressed uniformly by a factor of $2p_o$. (The end points of an interval are labeled according to their positions on the original interval of Fig. 10a.) As a result, the probability that any subinterval contains the transmitted point is now equal to its length. If B is received, Figs. 10b and 10c are inverted in an obvious way.

Since the sender is informed of the received symbol, he is aware of the receiver's new distribution. He again divides the interval (0, 1) at its mid-point (M in Fig. 10c) and sends A or B according to the new position of the transmitted point. In the case depicted it would be B. After the second symbol is received, one of the halves of the interval (0, 1), which in this case are $(0, .5/2q_o)$ and $(.5/2q_o, 1)$, is stretched by a factor of $2q_o$ and the other is compressed by a factor of $2p_o$. The receiver's distribution then consists of 3 subintervals that have been uniformly expanded or contracted, adjacent subintervals by different factors. The possible factors are $(2q_o)^2$, $(2q_o)(2p_o)$, and $(2p_o)^2$. As the transmission process is continually repeated, the number of such subintervals is always one more that the total number of received symbols.

By the process just described, any subinterval containing the transmitted point is gradually expanded so that its length approaches unity. Such an interval, however, is not necessarily expanded by each correctly received transmission. This is true only when the interval does not contain the mid-point of the interval (0, 1), or if the transmitted point lies in the larger of the two parts into which the interval is divided by the mid-point. Any subinterval that does not contain the transmitted point gradually vanishes.

The receiver becomes more nearly certain about successive bits as additional symbols are received. A decision about the first bit is made as soon as the size of either of the intervals (0, .5) or (.5, 1) exceeds $1-P_e$, where P_e is the specified probability of error. If the interval (0, .5) reaches this size first, its probability is immediately increased to 1, and the interval (.5, 1) is discarded, in accordance with the fact that the receiver's determination of the first bit at this time is final. The second bit is then determined the first time that either of the intervals (0, .25) or (.25, .5) exceeds $1-P_e$, and so on.

If the receiver should happen to decode a bit incorrectly, the sender immediately changes the bit in question to agree with the receiver's decision and transmits the point in the receiver's distribution corresponding to the altered message sequence. In this way the occurrence of a decoding error is prevented from disrupting the transmission process.

The proper interpretation of P_e is now seen to be the following. The probability that all of the first m bits will be decoded correctly is no smaller than $(1-P_e)^m$. We are not entitled to state, however, that in a decoded block of m bits the average number that will be in error is $P_e m$. The reason is that we are guaranteed that the probability of error in decoding a bit is no worse than P_e only when all preceding decoding decisions have been correct. However, when an error has occurred, P_e will once again be an accurate measure of the probability of error after an additional number of transmissions of the order of magnitude of the average constraint length.

Several other features of this transmission scheme are immediately apparent. Since the two symbols, A and B, are always equally likely in the eyes of the receiver, each transmitted symbol carries one bit of information about the message sequence. The average information received is just the capacity of the binary symmetric channel, $p_o \log 2p_o + q_o \log 2q_o$ bits per symbol. Thus we have another example in which reliable communication at channel capacity is realized and yet there is no mention of coding in the usual sense.

We now take into account the fact that the number of undecoded bits available to the sender is necessarily finite. Thus, with each transmission the sender has in mind an interval (which we shall call the transmitter's interval) rather than a point. In order to achieve channel capacity, it is necessary that each transmitted symbol carry one additional bit of information about the message bits defining the transmitter's interval. This implies that the transmitter's interval should not include the mid-point of the interval $(0, 1)'$. (When it does, the transmitted symbol is chosen to correspond to the larger of the two parts into which the mid-point divides the transmitter's interval.) While there is no upper limit to the number of bits that it might be necessary to know to prevent this from occurring on a particular transmission, communication at an average rate arbitrarily close to capacity is possible if the transmitter's interval is kept sufficiently small. (The symbol d_o, $d_o \ll 1$, will be used to denote its maximum allowed size.) We shall assume that additional bits are always available when needed for this purpose. We have introduced the notation $(0, 1)'$ to indicate the entire interval under consideration. The end points of this interval are actually 0 and 1 only until the first bit is decoded.)

The constraint length is measured by the number of transmissions, N, occurring from the time a bit is first introduced at the transmitter until it is decoded - that is, the number of transmissions required to expand an interval containing the transmitted point from roughly d_o to $1-P_e$. (In calculating $\overline{N}$, we shall refer to this interval as TI (for transmitted interval) to avoid confusing it with the transmitter's interval, which will always lie within TI.) The average amount of computation required to recalculate the receiver's distribution each time a symbol is received is a linearly increasing function of $\overline{N}$. When a symbol is received, the position of every point that has ever been the mid-point of the interval (0, 1)' and has not yet been in an interval discarded by a decoding decision is recomputed. The number, W, of such points is equal to the number of times, since the inception of transmission, that the interval which has grown to be (0, 1)' has included the mid-point of the interval (0, 1)'. It is clear from the definition of N that $\overline{W}$ increases linearly with $\overline{N}$. In fact, we might expect $\overline{W}$ to be given by $\overline{N} - K$, where K represents the average number of transmissions on which the interval which has grown to be (0, 1)' has not included the mid-point of the interval (0, 1)'. K is a constant essentially independent of the value of P_e, if $P_e \ll 1$.

Both the sender and the receiver need to know p_o in order to calculate the receiver's distribution. If the channel is slowly varying, the receiver can obtain a good estimate of p_o by using the fact that, over time intervals during which the channel is fairly constant, the channel capacity can be obtained by an iterative process in which it is approximated, during each iteration, by the average rate at which bits were decoded during the previous iteration. (In channels other than the binary symmetric channel, however, knowledge of the capacity is usually insufficient to specify the channel.) The sender can compute p_o more directly by comparing the symbols he receives through the feedback channel with those he has transmitted. However, he should compute p_o exactly as the receiver does so that he can construct an exact replica of the receiver's distribution.

It should be mentioned that if P_e has been set so low that there is no concern over the possible occurrence of an error, the sender need only recompute the location of the transmitter's interval in order to determine the next transmitted symbol.

4.2 A BOUND ON THE AVERAGE CONSTRAINT LENGTH

It has been pointed out that $\overline{N}$ is a measure of both the decoding delay and the computational severity of the decoding procedure. In calculating an upper bound on $\overline{N}$, we shall assume that no decoding errors are made during the time interval defined by N – that is, from the time the bit in question is introduced until it is decoded.

We consider a slightly modified transmission procedure, in which the interval (0, 1)' is randomly divided into two equal parts prior to the selection of each transmitted symbol (see Fig. 11). The parameter a of the division process is distributed uniformly over the range, $0 \leq a \leq 1/2$. We shall refer to P and Q as the cut positions. Such a random division can be simulated in practice by using a pseudo-random binary sequence to determine each pair of cut positions. The resulting transmission process differs significantly from that previously described only in that the number of cut positions that must be relocated following each transmission has been doubled. (It is interesting to note that when random cut positions are used, it is virtually impossible for decoding to take place when an A is received, for this causes both ends of the

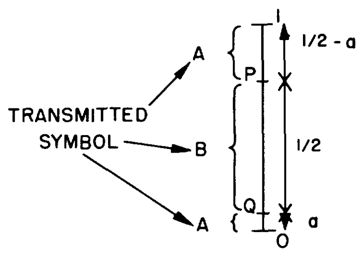

Fig. 11. Modified transmission process.

interval (0, 1)' to be expanded. On the other hand, if a single fixed cut position at the mid-point of the interval (0, 1)' is used (see Fig. 10a), only 0's can be decoded when an A is received, and only 1's when a B is received.)

We assume that the transmitter's interval never includes either of the cut positions. The probability that it does on any transmission is at most $2d_o$; this can be made negligibly small.

We have defined $\overline{N}$ as the average number of transmissions needed to expand TI from d_o to $1-P_e$. We place ourselves in the position of an observer who knows the initial location of TI (when its size is d_o) but who has no knowledge of the point within TI that is governing its expansion. The initial probability distribution for the location of the transmitted point must be uniform over TI because a 0 and a 1 are equally likely in each place of the message sequence. This initial distribution guarantees that the distribution remains uniform throughout the transmission process. This becomes evident when we recall that the construction in Fig. 10c maintains a uniform probability measure over the interval (0, 1)'.

We calculate $\overline{N}$ by simultaneously averaging over the possible positions of the transmitted point within TI, over all possible sequences of cut positions, and over all possible patterns of transmission errors.

In order to bound $\overline{N}$ we divide the expansion of TI (whose size we shall call d) into three ranges - from d_o to a value $d_a \leq 1/2$, from d_a to $d_b > 1/2$, and from d_b to $1-P_e$ - and then add the separate contributions ($\overline{N}_A$, $\overline{N}_B$, and $\overline{N}_C$, respectively) to get $\overline{N}$; that is, $\overline{N} = \overline{N}_A + \overline{N}_B + \overline{N}_C$. The following argument is used to bound $\overline{N}_A$.

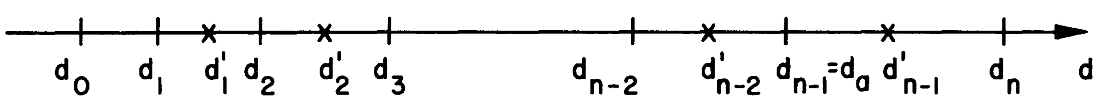

Fig. 12. Expansion of TI.

We let d_i, $1 \leq i \leq n$, be the maximum TI size attainable by a single transmission, starting from d_{i-1} (see Fig. 12). We assume for the moment that $d_{n-1} = d_a$. For $d_{i-1} \leq 1/2$, $d_i = 2q_o d_{i-1}$. We also define d'_i, $1 \leq i < n$, to be the value of d the first time $d \geq d_i$. Clearly, $d'_i < d_{i+1}$. Furthermore, we define $I_i = \log \frac{d_i}{d_{i-1}}$, $1 \leq i \leq n$. Also, $I'_i = \log \frac{d'_i}{d'_{i-1}}$, $I'_{ia} = \log \frac{d_i}{d'_{i-1}}$, and $I'_{ib} = \log \frac{d'_i}{d_i}$, $1 \leq i < n$, with $d'_o = d_o$. Finally, we set

N_i, $1 \leq i < n$, equal to the number of transmissions required to expand TI from d'_{i-1} to d'_i. Then N_A, the number of transmissions needed to expand TI from d_o to (at least) d_a, is given by

$$N_A = \sum_{i=1}^{n-1} N_i$$

Therefore

$$\overline{N}_A = \sum_{i=1}^{n-1} \overline{N}_i$$

In order to bound $\overline{N}_i$ we assume at first that d'_{i-1} takes on an arbitrary, but fixed value, d^*_{i-1}, in the range $d_{i-1} \leq d^*_{i-1} < d_i$. We make use of a theorem of Wald (15) which states that in a random walk with two absorbing barriers, in which each step size is chosen from the same distribution, the average number of steps to absorption is equal to the net average distance walked (including overshoot) divided by the (algebraic) average step size. Our situation differs from that described by the theorem in two respects. First, our walk, which results in an informational gain of $I'_i(d^*_{i-1})$, is limited in only one direction by an absorbing barrier; however, it eventually crosses this barrier with probability 1. Secondly, our walk is a first-order Markov process since the distribution governing each step size is a function of d. It is easily shown that the average step size, S(d), measured in units of informational gain, is a decreasing function of d. It is clear from this fact and the definition of d'_i that for each value of d encountered during the expansion of TI from d^*_{i-1} to d'_i, $S(d) > S(d_i)$. This last fact enables us to obtain the following, quite plausible result, which involves only a slight modification of Wald's derivation.

$$\overline{N}_i(d^*_{i-1}) \leq \frac{\overline{I}'_i(d^*_{i-1})}{S(d_i)}$$

If we let p^*_{i-1} represent the probability density for d^*_{i-1},

$$\overline{N}_i \leq \int_{d_{i-1}}^{d_i} p^*_{i-1}\overline{N}_i(d^*_{i-1})\, d(d^*_{i-1}) = \frac{1}{S(d_i)} \int_{d_{i-1}}^{d_i} p^*_{i-1}\overline{I}'_i(d^*_{i-1})\, d(d^*_{i-1}) = \frac{\overline{I}'_i}{S(d_i)}$$

Recalling the definitions of I_i, I'_{ia}, and I'_{ib}, we find that it follows that

$$\overline{N}_A = \sum_{i=1}^{n-1} \overline{N}_i \leq \sum_{i=1}^{n-1} \frac{\overline{I}'_i}{S(d_i)} = \sum_{i=1}^{n-1} \frac{(\overline{I}'_{ia}+\overline{I}'_{ib})}{S(d_i)} \leq \sum_{i=1}^{n-1} \frac{\overline{I}'_{ia}}{S(d_i)} + \sum_{i=1}^{n-2} \frac{\overline{I}'_{ib}}{S(d_{i+1})} + \frac{I_n}{S(d_{n-1})}$$

$$= \sum_{i=1}^{n-1} \left[\frac{\overline{I}'_{ia}}{S(d_i)} + \frac{\overline{I}'_{(i-1)b}}{S(d_i)} \right] + \frac{I_n}{S(d_{n-1})} = \sum_{i=1}^{n-1} \frac{I_i}{S(d_i)} + \frac{I_n}{S(d_{n-1})} \qquad (28)$$

Since $d_a \leq 1/2$, $I_i \equiv I = \log 2q_o$ for all $i \leq n$. Therefore

$$\overline{N}_A \leq \sum_{i=1}^{n-2} \frac{I}{S(d_i)} + \frac{2I}{S(d_{n-1})}$$

It is shown in Appendix C-1 that for $d \leq 1/2$

$$S(d) = C - \left(\log e - \frac{2p_o q_o}{q_o - p_o} \log \frac{q_o}{p_o}\right) d = C - k(p_o)\, d$$

where $C = q_o \log 2q_o + p_o \log 2p_o$ is the channel capacity. Therefore

$$\overline{N}_A \leq \sum_{i=1}^{n-2} \frac{I}{C - k(p_o)\, d_i} + \frac{2I}{C - k(p_o)\, d_{n-1}} \leq \frac{I}{C} \sum_{i=1}^{n-2} \left[1 + \frac{2k(p_o)\, d_i}{C}\right] + \frac{2I}{C - k(p_o)\, d_{n-1}} \tag{29}$$

provided that $\frac{k(p_o)\, d_{n-2}}{C} = \frac{k(p_o)\, d_a}{2q_o C} \leq \frac{1}{2}$. We should like to choose $d_a = 1/2$. It can be verified that the required inequality, $\frac{k(p_o)}{2q_o C} \leq 1$, is satisfied for $p_o < .32$. This includes all cases of interest, so we shall set $d_a = 1/2$, subject to this restriction. It then follows from Eq. 29 that

$$\overline{N}_A \leq \frac{\log (1/2d_o)}{C} + \frac{I}{C}\left[\frac{1 + [k(p_o)/2C]}{1 - [k(p_o)/2C]}\right] + \frac{k(p_o)\, I}{2q_o C^2} \sum_{i=0}^{n-3} \frac{1}{(2q_o)^i} \tag{30}$$

The first term on the right-hand side of Eq. 30 indicates that information is provided about TI at an asymptotic rate (as $d_o \to 0$) equal to C. The sum of the last two terms is a bound on the number of additional transmissions, $\overline{N}_a$, that is made necessary by those occasions when one of the cut positions lies within TI. However, $\overline{N}_a$ also includes the effect of the overshoot of d'_{n-1} beyond 1/2, since d'_{n-1} may be as large as q_o. We can find a bound on $\overline{N}_a$ that is independent of d_o by letting n approach ∞ in Eq. 30. Thus

$$\overline{N}_a \leq \frac{I}{C}\left[\frac{1 + [k(p_o)/2C]}{1 - [k(p_o)/2C]}\right] + \frac{k(p_o)\, I}{C^2(q_o - p_o)} \tag{31}$$

This bound is plotted as a function of p_o in Fig. 13.

Equation 28 can also be used to bound $\overline{N}_B$, which is the average number of transmissions required to expand TI from 1/2 to d_b. Since I_i is no longer independent of i, however, a step-by-step calculation is required. Furthermore, the bound obtained will be many times greater than the actual value of $\overline{N}_B$ because S(d) is a rapidly decreasing function of d for $d > 1/2$. (For the sake of completeness, S(d) for $d > 1/2$ is also derived in Appendix C-1.) Experimental evidence presented in section 4.3, in which $d_a = 1/32$ and $d_b = .99$, will shed light on this portion of the process. The important

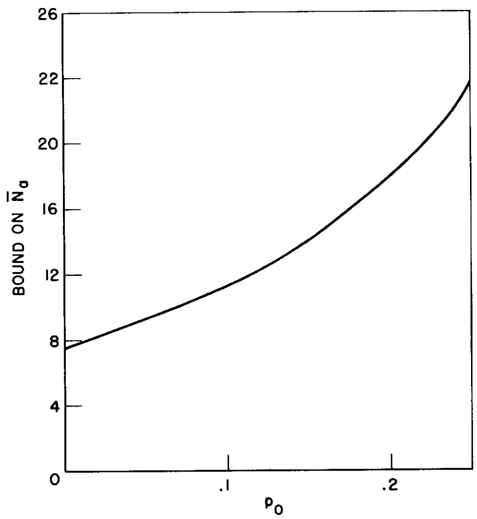

Fig. 13. An upper bound on N_a.

thing to note at this time is that $\overline{N}_B$ is a constant that depends only on p_o and the choice of d_b; it is independent of d_o and P_e. It therefore has no effect on the asymptotic functional dependence of $\overline{N}$ on P_e and/or d_o.

We now turn to the calculation of a bound on $\overline{N}_C$, the average number of transmissions needed to expand TI from d_b, which we shall choose later, to $1-P_e$. For this purpose it is more convenient to follow the contraction, from $1-d_b$ to P_e, of $[(0,1)'-\text{TI}]$, which we shall refer to as the wrong interval (WI).

We define $S_W(d)$ as the average information provided per transmission, about WI, as a function of the TI size d. It is shown in Appendix C-2 that, for $d \geq 1/2$, $S_W(d)$ is given by

$$S_W(d) = \frac{1}{d}\left[\frac{1}{2(q_o-p_o)}\log\frac{q_o}{p_o} - \log e\right] + d\left[\log e - \frac{2p_oq_o}{q_o-p_o}\log\frac{q_o}{p_o}\right] + \log 2 + p_o \log q_o + q_o \log p_o \tag{32}$$

An interesting property of $S_W(d)$ is that, for $p_o > .039$, it is most negative for some value of $d < 1$. (For smaller values of p_o, the minimum value occurs at $d = 1$.) Curves of $|S_W(d)|$ versus d are plotted in Fig. 14 for several values of p_o. The value of d at which $S_W(d)$ assumes its minimum value, d_m, is shown in Fig. 15. This property of $S_W(d)$ will be used to find a bound on $\overline{N}_C$. Although the derivation is valid only for $p_o > .039$, the asymptotic behavior arrived at is characteristic of the process for all values of p_o.

We observe that there is a number d_b', $1/2 < d_b' < d_m$, for which $S_W(d_b') = S_W(1)$. Consequently, $S_W(d) \leq S_W(1)$ for $d_b' \leq d \leq 1$. Since we shall later let d_b approach 1, we assume that $d_b > d_b'$ and ask for the average number of symbols, $\overline{N}_{C1}$, that must

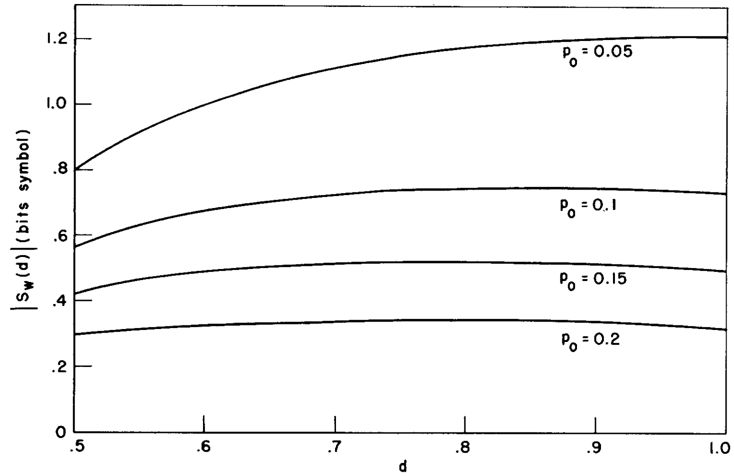

Fig. 14. $|S_W(d)|$ versus d.

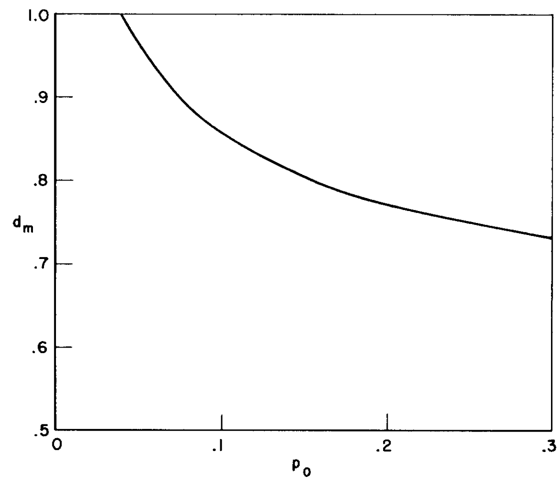

Fig. 15. TI size for which $S_W(d)$ is a minimum.

be transmitted before WI, starting from size $1-d_b$, is either expanded to $1-d'_b$ or contracted to P_e. Using Wald's formula (15) once again, we find

$$\overline{N}_{C1} \leqslant \frac{\log\left[(2p_oP_e)/(1-d_b)\right]}{S_W(1)}$$

If we define P_1 as the probability that WI is expanded to $1-d'_b$ before it is contracted to P_e, a recent result of Shannon (16) tells us that

$$P_1 \leqslant e^{-s_oA}\left[\frac{s_oA}{\mu(s_1)} + \frac{1}{1-e^{\mu(s_1)}}\right]$$

where

$$A = \log \frac{1-d'_b}{1-d_b}$$

$$\mu(s) = \log \frac{1}{2}\left[(2q_o)^s+(2p_o)^s\right]$$

$$\mu(s_o) = 0, \ s_o > 0$$

$$\mu'(s_1) = 0, \ s_1 > 0$$

A fuller explanation of this result is given in Appendix C-2. For the present it is only necessary to appreciate that P_1 is essentially exponentially decreasing in A. We can therefore make P_1 as small as we wish by choosing d_b sufficiently close to 1.

We define $\overline{N}_{C2}$ to be the average number of transmissions required to expand TI from $d'_b/2q_o$ to d_b. $\overline{N}_{C2}$ can be bounded by the method used to bound $\overline{N}_1$ and $\overline{N}_2$. It follows that

$$\overline{N}_C \leq \overline{N}_{C1} + P_1(\overline{N}_{C2}+\overline{N}_C)$$

$$\overline{N}_C \leq \frac{\overline{N}_{C1} + P_1\overline{N}_{C2}}{1 - P_1}$$

$$\leq \frac{1}{(1-P_1)} \frac{\log P_e^{-1}}{S_W(1)} + \frac{1}{(1-P_1)}\left(P_1\overline{N}_{C2} + \frac{\log (2p_o/1-d_b)}{S_W(1)}\right) \tag{33}$$

Once we choose d_b, the second term on the right-hand side of Eq. 33 is a constant, independent of P_e. Since P_1 can be made as small as desired, the asymptotic relation determining $\overline{N}_C$ is

$$P_e \doteq 2^{-|S_W(1)|\overline{N}_C}$$

where $S_W(1)$ is found by setting d = 1 in Eq. 32.

$$S_W(1) = \frac{1}{2} \log 4p_oq_o$$

$S_W(1)$ could have been found directly by the following heuristic argument. The probability distribution for the location of the transmitted point is uniform over TI on each transmission. Therefore, in the limit as $d \to 1$, the probability is exactly 1/2 that the transmitted point and WI will be in the same half of the interval $(0, 1)'$ on a given transmission, the two halves being determined by the random cut positions. The probability is then $\frac{1}{2}(q_o) + \frac{1}{2}(p_o) = \frac{1}{2}$ that WI will be expanded by a factor of $2q_o$ as the result of the transmission. The other half of the time it will be reduced by a factor of $2p_o$. Therefore,

$$S_W(1) = \frac{1}{2} \log 2q_o + \frac{1}{2} \log 2p_o = \frac{1}{2} \log 4p_o q_o$$

This same sort of reasoning will be used to find the asymptotic behavior of an extension of this transmission procedure in which we are free to select any transmission rate less than channel capacity.

4.3 EXPERIMENTAL RESULTS

In order to obtain a better picture of the operation of the transmission scheme described above, the entire system was simulated on an IBM 709 computer. Three independent pseudo-random number generators were used to generate the cut positions (when random cuts were used), the bits comprising the information sequence, and the sequence of 0's and 1's describing the channel noise. In the information sequence, 0's and 1's occurred with equal probability. In the noise sequence, 1's occurred with probability p_o.

It was not necessary to follow the receiver's decoding procedure in order to keep track of the significant points in the receiver's distribution. It will be recalled that each time a symbol is received, the receiver relocates each point remaining in his distribution that has ever been a cut position. A linear interpolation between an adjacent pair of these points allows him to locate any remaining (significant) point from the original interval (0, 1)'. In the simulation, however, it was possible to compute directly the position of each (significant) point whose subsequent arrival within P_e of either end of the interval (0, 1)' signified the decoding of a bit. In the absence of a decoding error, the significant points are revealed by the information bits. For example, if the first three information bits are 011, the first four significant points are .5, .25, .375, and .4375. This information, of course, is not normally available to the receiver.

The simulation was carried out for three different values of p_o: .04, .1, and .2. With $p_o = .1$, two values of P_e, 10^{-6} and 10^{-15}, were used; otherwise, P_e was always set equal to 10^{-15}. Transmission was initiated in each case by the selection of 24 information bits, which defined the initial transmitter's interval. Transmission continued (additional bits being generated as needed) until the 24th bit had been decoded. At this point the entire process was begun anew. The number of repetitions was 100 for all combinations of p_o and P_e except for $p_o = .2$, in which case the process was repeated 150 times. At no time was a bit ever decoded incorrectly.

The value of d_o used throughout the experiments was .0005; that is, new bits were generated whenever the size of the transmitter's interval exceeded .0005. As a result, the transmitter's interval included one of the cut positions, on the average, not more than once in 1000 transmissions. Thus, the transmission rate was not significantly different from channel capacity.

It should be noted, however, that d_o could have been increased to .01, for example, without decreasing R by more than .02 C. This follows from the fact that if $d_o = .01$, the transmitter's interval includes a cut position less than .02 of the time, if the cut positions are chosen in a random manner. (Random cut positions were used throughout the experiments except where otherwise stated.) The average value of the mutual information between the transmitted and received symbols on these occasions is clearly positive since the symbol corresponding to the larger portion of the transmitter's interval is always sent. An increase in d_o from .0005 to .01 would have reduced the average constraint length by more than 5 transmissions for $p_o = .04$, by more than 8 for $p_o = .1$, and by more than 15 for $p_o = .2$.

The fact that the system does, in fact, enable the receiver to decode bits at an average rate equal to the channel capacity was verified. The average number of symbols that should have been received between the decoding of the 1st and the 24th bits is $\overline{T}(C) = \frac{23}{C}$. $\overline{T}(C)$ is listed in Fig. 16 together with $\overline{T}'(C)$, the average number that was found experimentally.

p_o	C (bits/symbol)	$\overline{T}(C)$	$\overline{T}'(C)$
.04	.758	30.4	31.4
.10	.531	43.3	46.6
.20	.278	82.8	78.2

Fig. 16. Data pertaining to the decoding rate.

The asymptotic error behavior was checked by observing the added number of transmissions needed, on the average, to reduce P_e from 10^{-6} to 10^{-15}, with $p_o = .1$. This reduction in P_e corresponds to receiving an additional −9 log 10 bits about WI. We have seen that information is provided about WI at an average rate per transmission of $\frac{1}{2} \log 4p_o q_o = -.737$ bit transmission for $p_o = .1$. The expected number of additional transmissions was therefore $\frac{9 \log 10}{.737} = 40.6$. The number found experimentally was 37.3.

Figure 17 indicates the average constraint length that was found for $P_e = 10^{-15}$ in combination with each of the three different values of p_o. N was found by observing

p_o	$C\left(\frac{\text{bits}}{\text{symbol}}\right)$	$\overline{N}$	$R\left(\frac{\text{bits}}{\text{symbol}}\right)$	N_{min}
.04	.758	54.4	.379	340
.10	.531	85.2	.266	695
.20	.278	194.2	.139	1705

Fig. 17. Average constraint length for $d_o = .0005$ and $P_e = 10^{-15}$.

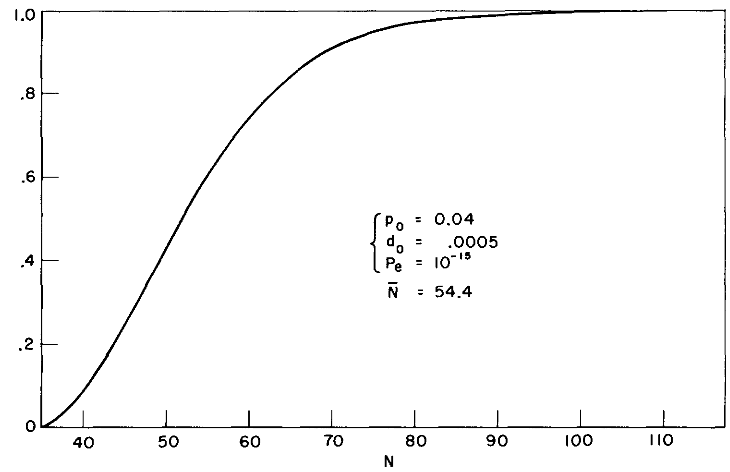

Fig. 18. Distribution function of the constraint length.

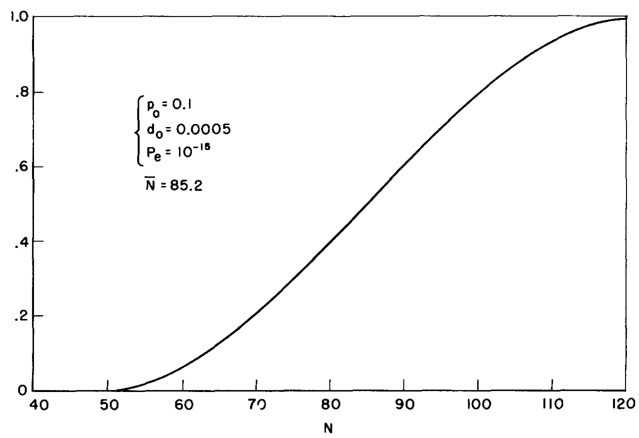

Fig. 19. Distribution function of the constraint length.

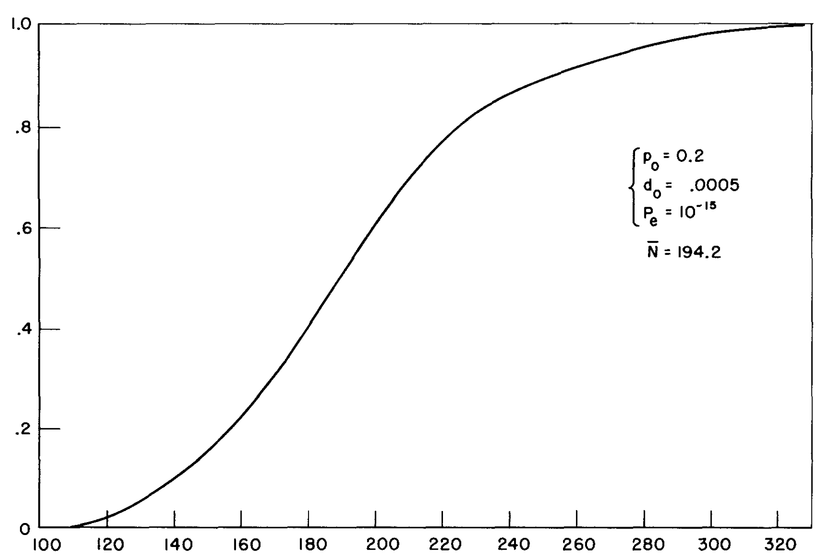

Fig. 20. Distribution function of the constraint length.

the number of transmissions needed to decode the 11^{th} of the 24 bits. The size of the interval defined by the first 11 bits is 2^{-11} = .000488, which is slightly less than .0005, the chosen value of d_o. A lower bound on the constraint length, N_{min}, of a fixed-constraint-length block code, which is required to achieve a value of $P_e = 10^{-15}$, with $p_o = 0.1$ and $R = C/2$, is also shown.

The relatively small value of N made possible by the use of feedback is quite striking. Of almost equal significance is the narrowness of the first-order probability distribution of N. The distribution function of N is shown in Figs. 18, 19, and 20 for the three values of p_o. The ratio of the largest to the smallest value of N encountered is less than 4:1 for $p_o = .04$, less than 2.5:1 for $p_o = .1$, and just over 3:1 for $p_o = .2$. If we exclude the largest 10% of the N values, these ratios are replaced by 2:1, 2:1, and 2.5:1, respectively.

We assume that the asymptotic behavior governs the process from the time that the size of TI first exceeds .99. The experimental results then indicate that the asymptotic part of the process accounts for about 2/3 of the average constraint length when $P_e = 10^{-15}$ and $d_o = .0005$. The exact fraction depends on p_o, varying from about .61 when $p_o = .04$ to about .69 when $p_o = .2$.

Information regarding the middle portion of the process was obtained by observing the expansion of the interval representing the first 5 bits from its initial size of 1/32 until it first exceeded .99. The average number of transmissions, N_5, required for this expansion was 12.8 for $p_o = .04$, 18.8 for $p_o = .1$, and 37.6 for $p_o = .2$. The distribution of N_5 for $p_o = .1$ is shown in Fig. 21.

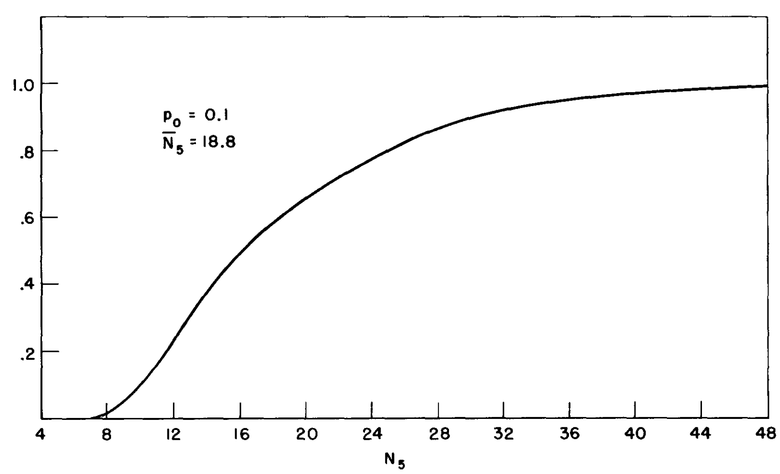

Fig. 21. Distribution function of N_5.

It is interesting to note that the extreme values of N_5 in Fig. 11 are in the ratio 7:1, whereas the extreme values of N are only in the ratio 2.5:1. The explanation lies in the fact that the randomness of the cut positions plays no part in the asymptotic behavior, which dominates the over-all process. This follows from the fact that, in the limit as d → 1, for any given pair of cut positions, the received symbol results in a contraction of WI exactly half the time. In the part of the process described by Fig. 21, on the other hand, the average information received about TI per transmission varies considerably with the cut position. The dependence on this extra parameter accounts for the large variance of N_5.

Figure 22 indicates the average size, after each transmission, of the interval representing the first 5 bits. 50 different runs, each with p_o = .1, contributed to the average. However, only those intervals that had not previously reached .99 were included in the average for a given transmission. The number of runs contributing to each average value is indicated by the stepped curve. One of these runs (see Fig. 23) has been selected to illustrate a typical step-by-step expansion of the interval representing the first 5 bits.

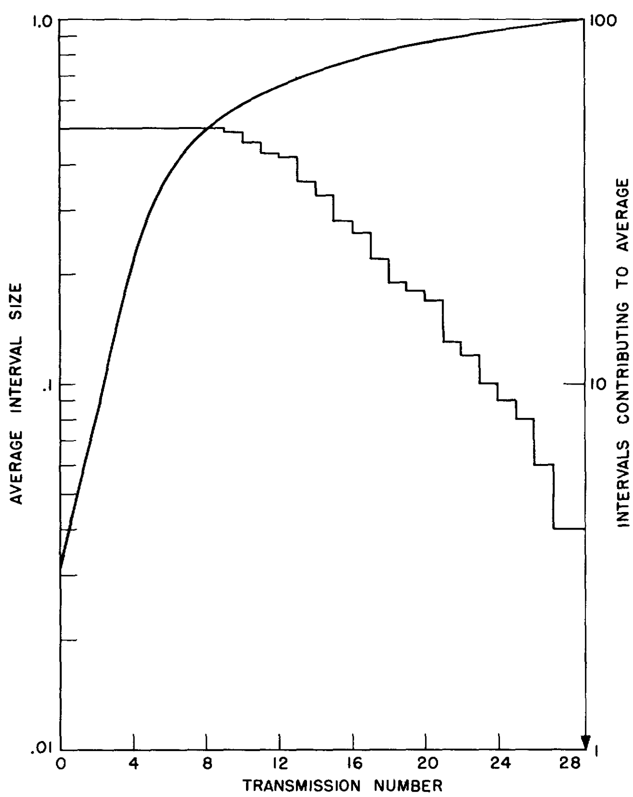

Fig. 22. Average size of interval representing first five bits, p_o = .1.

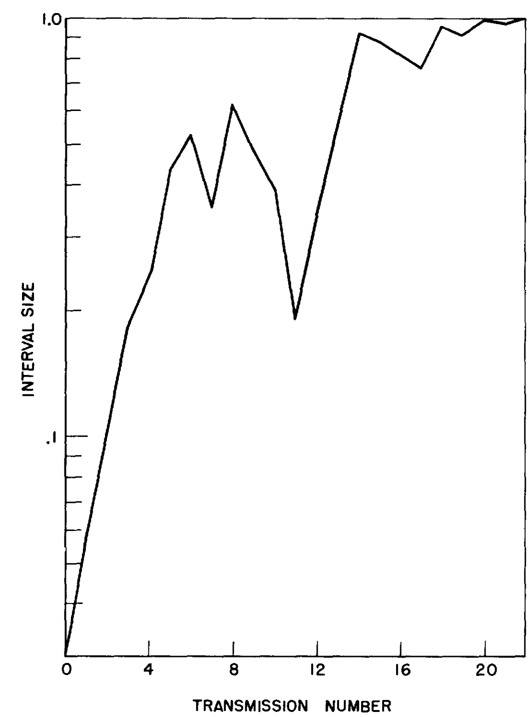

Fig. 23. Size of interval representing first five bits, p_o = .1.

In the transmission process as originally described, the interval (0, 1)' is divided into two equal parts at its mid-point prior to each transmission. The introduction of random cut positions was for reasons of mathematical tractability; it made the analysis independent of the bits comprising the information sequence – that is, independent of the location of the transmitted point on the original interval (0, 1)'. Experiments were also conducted with the use of a fixed cut position at the mid-point of the interval (0, 1)'. The average constraint length that resulted was slightly greater than that found when random cut positions were used.

This rather surprising result can be explained by noting, from Fig. 11, that when the interval (0, 1)' is divided in a random manner, both ends of the interval (0, 1)' are either expanded or contracted together. Thus, the asymptotic part of the process proceeds at the same pace for both 0's and 1's. We see from Fig. 10a that when a single fixed cut is used, and more A's than B's are received, 0's tend to be decoded more rapidly than 1's; when more B's are received, 1's tend to be decoded more rapidly. However, a bit is never decoded until all previous bits in the sequence have been decoded. Therefore, when many more A's than B's have recently been received, for example, the decoding of a bit as a 0 may be delayed to the point where its probability of being incorrect, conditional on all previously decoded bits being correct, is considerably less than the assigned value of P_e. Probabilities of error as low as 10^{-29} were observed when P_e was 10^{-6}. The average probability of error, however, remained within an order of magnitude of P_e.

It was not possible to use a single fixed cut with $P_e = 10^{-15}$ because the resulting probabilities of error were often too small for the IBM 709 to handle. The idea of using fixed cuts can be salvaged, however. It is only necessary to alternate between two pairs of fixed cuts, located at distances a and $a + \frac{1}{2}$ from either end of the interval (0, 1)'. The choice of a in the range, $0 < a < \frac{1}{2}$, is arbitrary, except that the two pairs of cut positions must be distinct; hence a cannot equal $\frac{1}{4}$. The reason why two pairs of cut positions are needed will become evident in section 4.4.

4.4 ERROR EXPONENT AS A FUNCTION OF RATE

With fixed-constraint-length systems, a reduction in the transmission rate reduces the constraint length required to achieve a desired average error probability. The feedback scheme considered thus far does not possess this degree of flexibility, for the transmission rate is fixed essentially at capacity. However, it is possible to operate at any rate less than capacity by choosing the a priori probabilities of the two input symbols to the channel appropriately. To see how this idea can be exploited to improve the asymptotic error behavior, consider the transmission scheme depicted in Fig. 24.

Prior to each transmission, the interval (0, 1)' is divided into two parts of size r and 1−r, $r \geq \frac{1}{2}$, by cut positions located at distances a and a + r from one end of the interval. Thus, a is restricted to the range, $0 < a < 1-r$. The cut positions are alternated between the solid and the dashed positions shown in Fig. 24a. Because of the uniform probability measure maintained on the interval (0, 1)', the probabilities of input symbols A and B are 1−r and r, respectively.

The transmission rate for the process is seen from Fig. 24b to be

$$R(r) = r\left[q_o \log \frac{q_o}{rq_o + (1-r)\,p_o} + p_o \log \frac{p_o}{rp_o + (1-r)\,q_o}\right]$$

$$+ (1-r)\left[q_o \log \frac{q_o}{rp_o + (1-r)\,q_o} + p_o \log \frac{p_o}{rq_o + (1-r)\,p_o}\right]$$

$$= q_o \log q_o + p_o \log p_o - [p_o+(q_o-p_o)r] \log [p_o+(q_o-p_o)r]$$

$$- [q_o-(q_o-p_o)r] \log [q_o-(q_o-p_o)r] \qquad (34)$$

The asymptotic error behavior is derived by considering the limiting situation as $d \to 1$. Consider Fig. 24c, for example, which indicates a possible mapping of the original interval $(0, 1)'$ just prior to the decoding of the first bit. (The ends of the interval have been greatly magnified.) The most likely information sequence at this point clearly

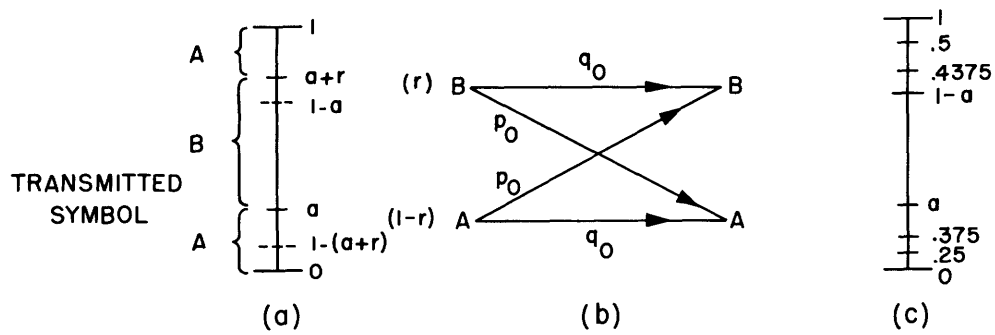

Fig. 24. Transmission at rate determined by r.

begins with the bits 0110. If we happen to be interested in the constraint length for the third bit (we assume that the first three bits are, in fact, 011), TI is (.375, .5), and WI is the union of (0, .375) and (.5, 1). Thus the two portions of WI lie at either end of the interval $(0, 1)'$. Consequently, they are contracted by the reception of a B and expanded by the reception of an A. This is true, in general, if transmission has progressed to the point (relative to the bit under consideration) where the asymptotic behavior obtains.

We are now in a position to calculate $R'(r) = \lim_{d \to 1} S_W(d,r)$, which is seen, in the light of the above discussion, to be equivalent to the average information provided about symbol A per transmission.

$$R'(r) = r\left[q_o \log \frac{p_o}{rq_o + (1-r)p_o} + p_o \log \frac{q_o}{rp_o + (1-r)q_o}\right]$$

$$+ (1-r)\left[q_o \log \frac{q_o}{rp_o + (1-r)q_o} + p_o \log \frac{p_o}{rq_o + (1-r)p_o}\right]$$

$$= [p_o + (q_o - p_o)r] \log \frac{p_o}{p_o + (q_o - p_o)r}$$

$$+ [q_o - (q_o - p_o)r] \log \frac{q_o}{q_o - (q_o - p_o)r} \tag{35}$$

If we define $p' = 1 - q' = q_o - (q_o - p_o) r$, it follows from Eqs. 34 and 35 that

$$R' = R - (q' - p_o) \log \frac{q_o}{p_o} \tag{36}$$

In addition, Eq. 34 becomes

$$H(q') = R + H(p_o) \tag{37}$$

By comparing Eq. 37 with Eq. 13, we can identify q' with q^*. It is now evident from Eq. 36 that R' is identical to the previously derived bound (curve A in Fig. 9) on the error exponent of an information-feedback, block type of system. Thus, the continuous system under consideration has the same asymptotic error-correcting capability as would a block-transmission system (if one existed) whose decoding table looks like that in Fig. 5 for every value of N.

Although the error exponent is improved when a smaller value of R is used, it is not necessarily true that the over-all constraint length will be reduced, for a given value of P_e. The reason is that the initial portion of the transmission process, which is concerned with the expansion of TI from d_o to 1/2 (say), will proceed at a slower average rate if R is reduced. However, as P_e is made smaller, the asymptotic improvement must eventually result in a decreased value of $\overline{N}$.

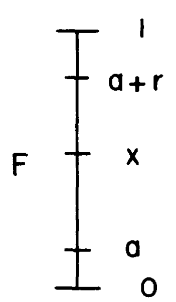

Fig. 25. Illustrating the fact that at least two pairs of cut positions must be used.

In order to understand the reason for alternating the cut positions, as in Fig. 24a, let us assume that only one pair, at distances a and a + r from the lower end of the interval (0, 1)', is used. There then exists a point (F in Fig. 25), other than the end points of the interval (0, 1)', whose position remains unchanged regardless of the received symbol. If we let x represent the distance of F from the lower end of the interval (0, 1)', it is clear from Fig. 25 that x is determined by the condition:

$$\frac{a}{1-(a+r)} = \frac{x-a}{(a+r)-x}$$

Because F is fixed, points that start out below F remain below F, and those that are originally above F remain above F. Thus, for example, if $x > .5$ and the first information bit is a 0, it will never be decoded correctly. By alternating the cut positions, we also alternate the fixed points on successive transmissions, thereby providing any point with access to the entire interval $(0,1)'$. Since the two pairs of cut positions must be distinct, we must not choose $a = \frac{1-r}{2}$. A good choice would be $a = \frac{1-r}{4}$.

4.5 A BLOCK-TRANSMISSION VERSION

The idea underlying the continuous information-feedback system considered thus far can be used as the basis for a block-transmission system. If M bits are to be transmitted by the choice of 1 of 2^M equally likely messages, the interval $(0,1)'$ is initially divided into 2^M subintervals of size 2^{-M}. One subinterval is assigned to each of the messages. The subinterval corresponding to the message chosen for transmission becomes the transmitter's interval. Since there are no further bits available to subdivide this interval as it is expanded, it remains the transmitter's interval until a decoding decision is made.

Initially the transmission scheme is similar to the one already described for continuous transmission. Label the messages $m_1, m_2, \ldots, m_{2^M}$, and let their probabilities after n transmissions be $P_1^n, P_2^n, \ldots, P_{2^M}^n$. Before the $n+1$ transmission the messages are divided into two disjoint sets, $M_1 = \{m_1, \ldots, m_j\}$ and $M_2 = \{m_{j+1}, \ldots, m_{2^M}\}$, where

$$\left|\frac{1}{2} - \sum_{i=1}^{j} P_i^n\right| = \min_k \left|\frac{1}{2} - \sum_{i=1}^{k} P_i^n\right|$$

If the message being transmitted is m_k, the symbol A is transmitted if $k \leq j$; otherwise, B is sent. The process continues in this fashion until $P_k^{n_o} \geq \frac{1}{2}$ for some n_o, k pair.

At this point the sender changes his rule for determining the next transmitted symbol. He sends an A if the intended interval is the one whose size is greater than $\frac{1}{2}$; otherwise he sends a B. If it should happen that for some $n_1 > n_o$, $P_i^{n_1} < \frac{1}{2}$ for all i, the sender reverts to his original rule for determining the transmitted symbols.

The constraint length, N, which is the number of transmissions required to reach a decoding decision, is bounded by the number of transmissions, N_c, needed to expand the transmitter's interval from 2^{-M} to $1-P_e$. This is immediately evident from the fact that, on those occasions when a decoding error is made, the incorrect decision is reached before the size of the correct interval ever reaches $1-P_e$. However, it is readily appreciated that $\lim_{P_e \to 0} \frac{\overline{N}}{\overline{N}_c} = 1$. In determining $\overline{N}_c$ we divide the expansion of TI into three ranges: from 2^{-M} to $d_a < \frac{1}{2}$, from d_a to $d_b > \frac{1}{2}$, and from d_b to $1 - P_e$, where d_a and d_b are independent of P_e. The principle difference between this situation and the continuous case already studied is that now two (rather than one) of the three ranges contribute to the asymptotic behavior. The reason is that M must increase as P_e is decreased in such a way that $R = M/\overline{N}$ remains constant.

If we again call the combination of all intervals other than the correct one WI, we note that the contraction of WI from $1 - d_b$ to P_e is characterized by the sequential 1-bit error exponent, $(q_o - p_o) \log \frac{q_o}{p_o}$. Since the expansion of TI from 2^{-M} to d_a proceeds at an asymptotic rate equal to C, in the limit as $P_e \to 0$,

$$\overline{N}_c \sim \frac{M}{C} + \frac{\log P_e^{-1}}{(q_o - p_o) \log \frac{q_o}{p_o}} = \frac{R\overline{N}_c}{C} + \frac{\log P_e^{-1}}{(q_o - p_o) \log \frac{q_o}{p_o}}$$

$$\left[(1 - \frac{R}{C}) (q_o - p_o) \log \frac{q_o}{p_o} \right] \overline{N}_c = \log P_e^{-1}$$

(This part of the process will be examined in greater detail later.)

Thus, the error exponent, E, for the process is

$$E = (1 - \frac{R}{C}) (q_o - p_o) \log \frac{q_o}{p_o}$$

We note, first of all, that E is a linear function of R, approaching zero as $R \to C$, and the sequential 1-bit error exponent as $R \to 0$. Thus, for all nonzero rates less than capacity, E is greater than the sequential sphere-packing error exponent. The exponential superiority of information-feedback, block transmission is thereby demonstrated.

E is the largest exponent that can be achieved by a system of this type. This becomes evident if we consider the two portions of the process that affect its asymptotic behavior. In one case information is provided about TI at the largest conceivable (positive) rate-channel capacity. In the other, information is provided about WI at the largest conceivable (negative) rate, in accordance with the sequential 1-bit error exponent. However, E is smaller than the bound derived in section 3.1 for information-feedback, block-type systems. We have already seen that the latter exponent can be achieved by a continuous information-feedback system. We conclude, therefore, that within the realm of information feedback, larger error exponents are attainable with continuous systems than with block-type systems.

Two minor points in connection with the expansion of TI from 2^{-M} to d_a require further explanation. First, the probabilities of the two sets, M_1 and M_2, into which the 2^M messages are divided are not, in general, exactly equal. Consequently, the two channel symbols are not equiprobable on each transmission, and the average information received about the intended message is somewhat less than channel capacity. We note, however, that as long as all the message probabilities are less than d (say), the a priori probabilities of the channel symbols will differ by less than d. It is easily shown that the average received information for any value of $d < \frac{1}{2}$ will be less than $C(p_o)$ by at most $C(p') = q' \log 2q' + p' \log 2p'$, where $p' = \frac{1}{2} - \frac{(q_o - p_o)\, d}{2}$ and $q' = 1 - p'$. For small values of d, $C(p') \doteq \frac{(q_o-p_o)^2\, d^2}{32}$. It follows that the slightly unequal symbol probabilities cannot have any effect on the asymptotic rate (as $P_e \to 0$) at which information is provided about TI.

Secondly, the expansion of TI proceeds at a rate determined by C provided the size of each of the $2^M - 1$ incorrect intervals is less than 1/2. If we designate by t $(t > \frac{1}{2})$ the size of the largest interval when this is not the case, the corresponding input probabilities to the channel are t and $1 - t$. The symbol of probability $1 - t$ is always transmitted, and the information received about it, I(t), is also the information received about the interval representing the message. The average value of this information is easily seen to be

$$\bar{I}(t) = q_o \log \frac{q_o}{tp_o + (1-t)\, q_o} + p_o \log \frac{p_o}{tq_o + (1-t)\, p_o}$$

$$= q_o \log q_o + p_o \log p_o - q_o \log \alpha - p_o \log (1-\alpha)$$

where $\alpha = q_o - (q_o - p_o)\, t$. Since $C = 1 + q_o \log q_o + p_o \log p_o$, $\bar{I}(t) > C$ for all $t > \frac{1}{2}$ for, as we shall show, $-q_o \log \alpha - p_o \log (1-\alpha) > 1$. The latter inequality follows from Fig. 26.

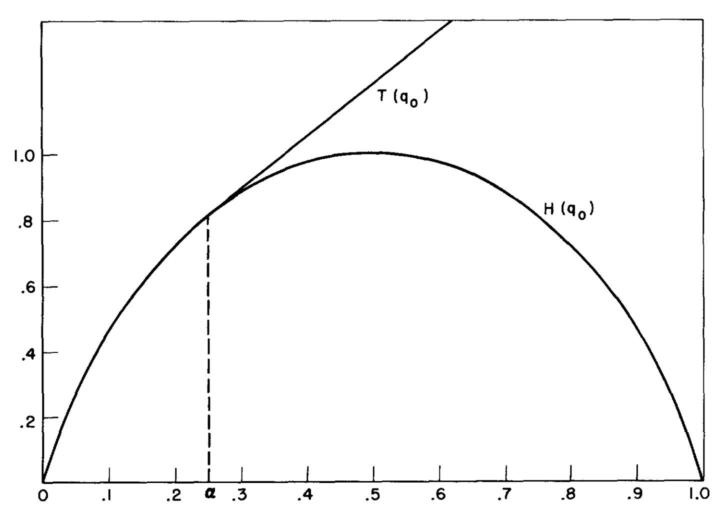

Fig. 26. Pertaining to the expansion of TI.

Since $t > \frac{1}{2}$ implies that $\alpha < \frac{1}{2}$, the line tangent to the entropy curve at the point whose abscissa is α is given by

$$T(q_o) = H(\alpha) + (q_o - \alpha) \log \frac{1-\alpha}{\alpha} = -q_o \log \alpha - p_o \log (1-\alpha)$$

Since $q_o > \frac{1}{2}$, it is clear from Fig. 26 that $T > 1$.

4.6 THE GENERAL DISCRETE MEMORYLESS CHANNEL

We consider now the extension of the continuous information-feedback system already discussed to the general discrete memoryless channel. The channel is defined by a set of transition probabilities, $p(y_j/x_i)$ (abbreviated p_{ij}), which represents the probability of receiving symbol y_j when symbol x_i has been transmitted. If u_i represents the probability with which x_i is transmitted, v_j, the probability with which y_j is received, is given by

$$v_j = \sum_i u_i p_{ij}$$

The information is again assumed to be in the form of a binary sequence, with each digit carrying 1 bit of information. Prior to each transmission the interval $(0, 1)'$ is divided into several subintervals, one for each channel input symbol. The size of each subinterval is made equal to the probability with which the corresponding channel symbol is to be used. The transmission process is as before. The location of the point representing the information sequence determines the symbol transmitted. The remapping of the interval $(0, 1)'$ after each new symbol is received is done in accordance with the input probabilities to the channel and the channel transition probabilities.

We shall assume that the input symbol probabilities have been so chosen that the channel capacity is realized. If we represent by $\{u\}$ any possible set of input probabilities ($u_i \geqslant 0$ for all i and $\sum_i u_i = 1$), it follows from the definition of channel capacity that

$$C = \max_{\{u\}} \sum_{i,j} p_{ij} u_i \log \frac{p_{ij}}{\sum_m p_{mj} u_m} \qquad (38)$$

We label the maximizing $\{u\}$ set $\{w\}$.

The asymptotic error behavior is again determined by the average compression of the ends of the interval $(0, 1)'$. If we assume that the subinterval representing x_k (say) is split into two parts, one at either end of the interval $(0, 1)'$, the error exponent, E_k, is equal to the magnitude of the average amount of information provided about symbol x_k per transmission. When symbol y_j is received, the information received about x_k is

$$I_{kj} = \log \frac{p_{kj}}{v_j}$$

Therefore,

$$E_k = -\sum_j v_j I_{kj} = -\sum_j \log \frac{p_{kj}}{\sum_m w_m p_{mj}} \sum_i w_i p_{ij} \tag{39}$$

In order to evaluate E_k we must know $\{w\}$. An explicit solution of Eq. 38 is not possible, in general. However, once $\{w\}$ is known, the asymptotic error behavior is optimized by letting the subinterval representing the symbol that maximizes E_k occupy the ends of the interval $(0, 1)'$. Precaution must be taken, as before, to avoid the creation of any fixed points in the interior of the interval $(0, 1)'$.

Although a trial-and-error method of solution is generally required by Eq. 38, we can solve explicitly for the error exponent, E, in cases where the symmetry of the channel enables us to determine the set $\{w\}$ by inspection. As a case in point, consider a channel having M input symbols and M output symbols, such that $p_{ii} = q_o$ and $p_{ij} = p_o = \frac{1-q_o}{1-M}$ for $i \neq j$. Clearly, channel capacity will be realized if each of the M symbols is used with probability $\frac{1}{M}$. In this case we see from Eq. 39 that

$$E = -\frac{1}{M} \sum_j \log Mp_{kj} \qquad \text{(for any k)}$$

$$= -\frac{1}{M}[\log q_o + (M-1) \log p_o] - \log M \tag{40}$$

Equation 40 is plotted against q_o in Fig. 27 for several values of M. (It is assumed that $q_o > \frac{1}{M}$.) It is interesting to observe that although C increases without limit as M is increased, E is bounded by the curve for $M = \infty$.

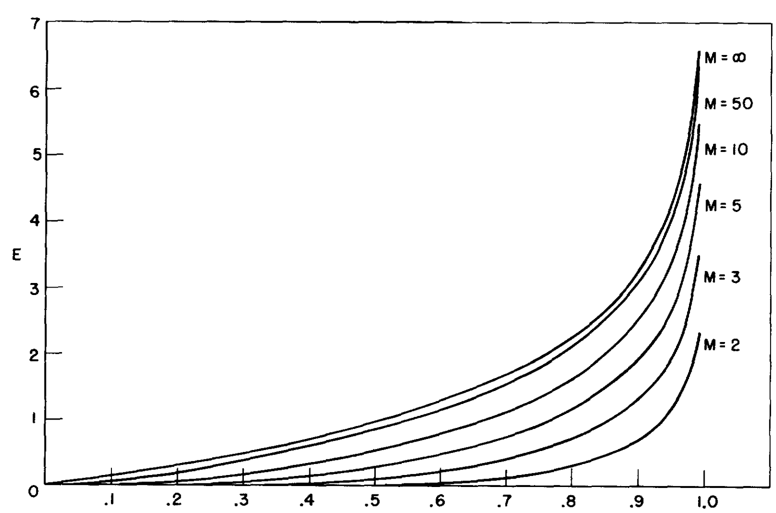

Fig. 27. Error exponent of M-ary symmetric channel.

V. CONCLUDING REMARKS

In our attempt to evaluate the advantages resulting from the inclusion of a feedback channel in a system designed to provide reliable communication between two points, we have glossed over several points that merit further attention. The most serious of these omissions concerns the fact that the specific transmission schemes presented assume the existence of a noiseless feedback channel.

In particular, the information-feedback scheme considered relies on precise knowledge of each received symbol. A wrong assumption about even a single received symbol is sufficient to disrupt the procedure. It is not known, therefore, whether the error exponent associated with this scheme can be obtained as the limiting (noiseless-feedback) value of the error exponent of a sequential transmission system in which there is a nonzero probability of a transmission error in the feedback channel. It is to be hoped that this possibility will spur the search for procedures that can be used with noisy feedback channels. Even if the ideal behavior associated with a noiseless feedback channel cannot be approached, substantial improvement over existing one-way systems may be expected.

The assumption of a noiseless feedback channel is somewhat less serious in the case of decision feedback. Since the sole role of the feedback channel is to indicate the occurrence of each decoding decision, a high degree of reliability can be injected into this operation without usurping a significant fraction of the reverse capacity, and without causing an appreciable delay in the forward direction.

A second omission has been the failure to take into account the capacity required of the feedback channel, especially in the case of information feedback. In the scheme presented, for example, the feedback channel, which is noiseless, must have a capacity of one bit for each symbol transmitted in the forward direction.

Obviously, the relative "cost" of the reverse capacity in a given communication system depends on the various functions that the system is required to perform. If the feedback channel would otherwise be idle and if power is cheap, we may be entitled to ignore the cost of operating the feedback channel. On the other hand, if information of equal volume and importance is to be transmitted in each direction, and if the two channels are identical, the operating cost of the entire system may reasonably be taken as twice that of either individual channel.

In connection with the latter type of situation, it may prove more feasible to set aside a fraction of the capacity in each direction as a feedback channel, to aid in the transmission of information in the opposite direction, than to operate the two channels independently. This has been done to a limited extent by Wozencraft (13), in order to increase the efficiency of his sequential decoding procedure. Although the convolutional code employed has a fixed constraint length, the philosophy involved is similar to that of decision feedback. At intervals equal to the constraint length, a symbol is inserted into the transmitted sequence in order to indicate whether the decoding is proceeding at a satisfactory rate. If not, the transmission is repeated from a point which guarantees the retransmission of the troublesome portion of the received sequence.

The primary purpose of the feedback is to eliminate the overwhelming decoding effort that follows an interval during which an unusually large number of crossovers has occurred. At the same time, the probability of error is greatly reduced, for most decoding errors follow the reception of a sequence that is not a good approximation to any of the possible transmitted sequences. The resulting decoding effort, which would be very great in the absence of feedback, is truncated before decoding takes place. The improvement in the error exponent resulting from the use of feedback is not known, however.

Although the assumption of a noiseless feedback channel is unrealistic for most applications, a model in which the channel capacities in the two directions are widely different may be entirely justified. In fact, in the case of communication between a facility on land and an orbiting satellite, for example, the relatively small power available at the satellite may cause the opposite channel to appear almost error-free by comparison.

A final point concerns the need for additional work on continuous transmission systems. Existing models for block-transmission systems have made it possible to determine quite thoroughly the capabilities of this type of system. However, lack of a suitable model has prevented a corresponding evaluation of continuous systems. Thus, it is not yet known, for one-way systems, whether asymptotically better error correction is possible with continuous systems. Although this question can be answered in the affirmative for feedback systems, an upper bound on the error exponent at non-zero rates remains to be determined. Since less complex decoding procedures appear to be possible with continuous systems, these questions are of more than passing interest.

APPENDIX A - DETAILS OF THE PROOF OF THEOREM 1

A.1 To prove: $\lim_{P_e \to 0} \Pr[N' < N_2(\delta)] = 0$

Let N'_m be the smallest value of N' at which a decoding decision can possibly be made. N'_m is given by

$$P_e \doteq 2^M \left(\frac{p_o}{q_o}\right)^{N'_m} \tag{A-1}$$

This is the probability of error when only 1 of the $2^{N'_m}$ possible received sequences is decoded.

Let $C_{N'}$ be the number of crossovers occurring in the first N' transmissions. We represent by j'_1 and p'_1 the values of j_1 and p_1 for which $j_1 = p_1 N'$. The probability of decoding and making a correct decision on the N'^{th} transmission, $N' < N_2(\delta)$, is bounded, according to Fig. 5, by $\Pr\left[\frac{C_{N'}}{N'} \leq p'_1\right]$. Since $N' < N_2(\delta)$, $p'_1 \leq p_o - \delta$. Consequently,

$$\Pr\left[\frac{C_{N'}}{N'} \leq p'_1\right] \leq \Pr\left[\frac{C_{N'}}{N'} \leq p_o - \delta\right]$$

$$= \sum_{j=0}^{(p_o-\delta)N'} \binom{N'}{j} p_o^j q_o^{N'-j} \doteq \binom{N'}{(p_o-\delta)N'} p_o^{(p_o-\delta)N'} q_o^{(1-p_o+\delta)N'}$$

$$\doteq 2^{N'[H(p_o-\delta)+(p_o-\delta)\log p_o+(1-p_o+\delta)\log q_o]} = 2^{N'T(p_o-\delta)}$$

Since $\frac{dT(p_1)}{dp_1} = 0$ and $\frac{d^2T(p_1)}{d^2p_1} = -\frac{1}{p_o(1-p_o)}$ at $p_1 = p_o$,

for small values of δ, we have

$$\Pr\left[\frac{C_{N'}}{N'} \leq p'_1\right] \dot{\lessapprox} 2^{\frac{-\delta^2}{2p_oq_o}N'} \leq 2^{\frac{-\delta^2}{2p_oq_o}N'_m} \qquad \text{for } N'_m \leq N' < N_2(\delta).$$

A decoding decision may also result in an error. However, this probability is bounded by P_e for every value of N'. Therefore, the probability, $P_{N'}$, of decoding on the N'^{th} transmission, $N'_m \leq N' < N_2(\delta)$, is bounded by

$$P_{N'} \dot{\lessapprox} 2^{\frac{-\delta^2}{2p_oq_o}N'_m} + P_e$$

It follows that

$$\Pr[N'<N_2(\delta)] \stackrel{\cdot}{\lessapprox} N_2(\delta)\left(2^{\frac{-\delta^2}{2p_o q_o}N'_m} + P_e\right) \leq N_1\left(2^{\frac{-\delta^2}{2p_o q_o}N'_m} + P_e\right) \tag{A-2}$$

From Eq. A-1 and the remark preceding Eq. 9 we see that the behavior of N'_m as P_e is varied can be described by

$$N'_m \doteq \frac{\log P_e^{-1} + M}{\log \frac{q_o}{p_o}} = K_1 \log P_e^{-1} \tag{A-3}$$

where K_1 is a constant independent of P_e. Equations 6 and 7 tell us that as P_e is varied with M/N_1 held constant, the behavior of N_1 is given by

$$N_1 = \frac{\log P_e^{-1}}{(q_{21}-p_o)\log\frac{q_o}{p_o} - R_1} \tag{A-4}$$

Equation A-2, together with Eqs. A-3 and A-4, yields the desired result, namely

$$\lim_{P_e\to 0} \Pr[N'<N_2(\delta)] = 0$$

A.2 To prove: $\lim_{\delta\to 0} \lim_{P_e\to 0} \frac{N_2(\delta)}{N_1} = 1$

From Eq. 5 and the fact that $H(q_2)$ is a continuous function of q_2, if follows that it is sufficient to show that

$$\lim_{\delta\to 0} \lim_{P_e\to 0} \frac{q_{22}}{q_{21}} = 1 \tag{A-5}$$

where q_{22} is the value of q_2 corresponding to j_{12}, q_{21} and q_{22} are related to j_{11} and j_{12}, respectively, by

$$H\left(\frac{j_{11}}{N_1}\right) + \frac{M}{N_1} = H(q_{21}) \tag{A-6}$$

$$H\left(\frac{j_{12}}{N_2}\right) + \frac{M}{N_2} = H(q_{22}) \tag{A-7}$$

Subtracting Eq. A-7 from Eq. A-6, we have

$$H(q_{21}) - H(q_{22}) = H\left(\frac{j_{11}}{N_1}\right) - H\left(\frac{j_{12}}{N_2}\right) + \frac{M}{N_1} - \frac{M}{N_2} \tag{A-8}$$

Here, j_{11} and j_{12} are determined by

$$j_{11} = p_o N_1$$

$$j_{12} = (p_o - \delta)\, N_2$$

Therefore, $\frac{j_{11}}{N_1} = p_o$ for all δ and P_e, while $\lim_{\delta \to 0} \frac{j_{12}}{N_2} = p_o$ for all P_e.

Consequently,

$$\lim_{\delta \to 0} \lim_{P_e \to 0} H\left(\frac{j_{11}}{N_1}\right) - H\left(\frac{j_{12}}{N_2}\right) = 0 \qquad \text{(A-9)}$$

Furthermore, as $P_e \to 0$, N_1 and N_2 grow without limit for any $\delta > 0$. Therefore,

$$\lim_{\delta \to 0} \lim_{P_e \to 0} \left(\frac{M}{N_1} - \frac{M}{N_2}\right) = 0 \qquad \text{(A-10)}$$

Using Eqs. A-9 and A-10 in conjunction with Eq. A-8,

$$\lim_{\delta \to 0} \lim_{P_e \to 0} [H(q_{21}) - H(q_{22})] = 0$$

This implies that $\lim_{\delta \to 0} \lim_{P_e \to 0} (q_{21} - q_{22}) = 0$. Thus Eq. A-5 is proved.

APPENDIX B – A LOWER BOUND ON THE SEQUENTIAL SPHERE-PACKING ERROR EXPONENT

An upper bound on the error exponent of a sequential sphere-packed code was derived in section 3.2 by considering only some of the points whose reception results in a decoding error. If we now assume that the decoding criterion is such that P_e does represent an average allowed error probability, we can obtain a lower bound on the error exponent by considering the points that were previously neglected. (The previously calculated upper bound is still applicable.) Our result will be a bound rather than the actual value of the error exponent because our method of adding the contributions to P_e of the various points in the decoding region involves some double counting for values of R close to C.

We need only consider the situation in which points on the boundaries of the decoding region are a distance $r_o = p_o N$ from the nearest code point, for it is a trivial matter to show, by using the results of Appendix A.1, that the value of p_2 at which decoding takes place satisfies the following condition:

$$\lim_{P_e \to 0} \Pr[|p_2 - p_o| > \delta] = 0 \qquad \text{for any } \delta > 0.$$

Assuming, once again, that the origin is the transmitted code point, we can classify each point in the decoding region according to the number of 1's in the point, its distance from the nearest code point, and the number of 1's in that code point. The number of points of each type in a sphere centered on a code point which is at a distance of $2r_1 + a$ from the origin is listed in Fig. B-1. (r_o is assumed to be even.) Points listed in the same row have the same number of 1's. This number increases by one with each successive row. The distance of a point from the nearest code point is $(r_o - k + 2b)$. Points at the same distance from the nearest code point are therefore listed along a diagonal line that reverses its direction at $k = r_o$.

In Fig. B-1a, which lists those points that are at least as close to the origin as is the nearest code point, a typical term has the form $\binom{2r_1 + a}{r_1 - k + b}\binom{N - 2r_1 - a}{b}$. This is the number of points which can be obtained by replacing $(r_1 - k + b)$ 1's in the nearest code point by 0's, and b of the 0's by 1's. The number of 1's in each such point is $(2r_1 + a - r_o + k)$.

In Fig. B-1b, which lists those points that are farther than the nearest code point from the origin, a typical term has the form $\binom{N - 2r_1 - a}{r_o - k + b}\binom{2r_1 + a}{b}$. This is the number of ways $(r_o - k + b)$ 0's in the nearest code point can be replaced by 1's, and b 1's by 0's. The number of 1's in each resulting point is $(2r_1 + a + r_o - k)$. It is assumed in Fig. B-1b that $r_o \leq N - 2r_1 - 1$. This is equivalent to the condition, $p_1 \leq \frac{1 - p_o}{2}$, which is certainly true in all cases of interest.

It should be clear that the approximation to P_e in section 3.2 was obtained by neglecting all terms except the one in Fig. B-1a for which $a = 1$ and $k = b = 0$.

We shall show that only the spheres surrounding the code points nearest the origin (those for which $a = 1$) need be considered. We first examine the relative contributions to P_e, for different values of a, of the points for which $k = b = 0$. Define n_a as the number of code points having $(2r_1+a)$ 1's. Since only some of the $\binom{N}{r_1+a}$ points with (r_1+a) 1's are included in spheres of these code points, and since the sphere about each of the n_a code points contains $\binom{2r_1+a}{r_1}$ points having (r_1+a) 1's.

$$\frac{n_a}{n_1} \leqslant \frac{\binom{N}{r_1+a}\Big/\binom{2r_1+a}{r_1}}{\binom{N}{r_1+1}\Big/\binom{2r_1+1}{r_1}} = \frac{(N-r_1-1)!\,(2r_1+1)!}{(N-r_1-a)!\,(2r_1+a)!}$$

$$= \frac{(N-r_1-1)\ \ldots\ (N-r_1-a+1)}{(2r_1+a)\ \ldots\ (2r_1+2)} \leqslant \left(\frac{N-r_1}{2r_1}\right)^{a-1} = \left(\frac{q_1}{2p_1}\right)^{a-1} \qquad \text{(B-1)}$$

For a given value of a, define n_a^o as the number of points in each sphere for which $k = b = 0$. Then

$$\frac{n_a^o}{n_1^o} = \frac{\binom{2r_1+a}{r_o}}{\binom{2r_1+1}{r_o}} = \frac{(2r_1+a)!\,(2r_1-r_o+1)!}{(2r_1+a-r_o)!\,(2r_1+1)!} = \frac{(2r_1+a)\ \ldots\ (2r_1+2)}{(2r_1-r_o+a)\ \ldots\ (2r_1-r_o+2)} \leqslant \left(\frac{2r_1+2}{2r_1-r_0+2}\right)^{a-1} \qquad \text{(B-2)}$$

The number of 1's in each such point is $(2r_1+a-r_o)$. If we define C_a^o as the contribution to P_e of the $k = b = 0$ points, combination of this last fact with Eqs. B-1 and B-2 yields

$$\frac{C_a^o}{C_1^o} \leqslant \left[\left(\frac{q_1}{2p_1}\right)\left(\frac{2r_1+2}{r_1+2}\right)\left(\frac{p_o}{q_o}\right)\right]^{a-1} \doteq \left(\frac{q_1/p_1}{q_o/p_o}\right)^{a-1} \qquad \text{(B-3)}$$

If $R < C$, then $p_1 > p_o$, and the successive contributions (for increasing values of a) to P_e of the points for which $k = b = 0$ are bounded by the terms of a geometric series.

We note further (from Fig. B-1) that, within each sphere, the contribution from a given k, b pair relative to that of the term for which $k = b = 0$, can be written in the form (Fig. B-1a)

$$\frac{\binom{2r_1+a}{r_o-k+b}\binom{N-2r_1-a}{b}}{\binom{2r_1+a}{r_o}} = \frac{r_o!(N-2r_1-a)\ \ldots\ (N-2r_1-a-b+1)}{b!(r_o-k+b)!\,(2r_1+a-r_o+k-b)\ \ldots\ (2r_1+a-r_o+1)} \qquad \text{(B-4a)}$$

k \ b	0	1	…	$\frac{r_o}{2}-1$	$\frac{r_o}{2}$
0	$\binom{2r_1+a}{r_o}$				
1	$\binom{2r_1+a}{r_o-1}$				
2	$\binom{2r_1+a}{r_o-2}$	$\binom{2r_1+a}{r_o-1}\binom{N-2r_1-a}{1}$			
⋮		⋮			
r_o-1	$\binom{2r_1+a}{1}$	$\binom{2r_1+a}{2}\binom{N-2r_1-a}{1}$	………	$\binom{2r_1+a}{\frac{r_o}{2}}\binom{N-2r_1-a}{\frac{r_o}{2}-1}$	
r_o	1	$\binom{2r_1+a}{1}\binom{N-2r_1-a}{1}$	………	………	$\binom{2r_1+a}{\frac{r_o}{2}}\binom{N-2r_1-a}{\frac{r_o}{2}}$

(a)

k \ b	0	1	. . . .	$\frac{r_o}{2}-1$
r_o-1	$\binom{N-2r_1-a}{1}$	$\binom{N-2r_1-a}{2}\binom{2r_1+a}{1}$	. . . .	$\binom{N-2r_1-a}{\frac{r_o}{2}}\binom{2r_1+a}{\frac{r_o}{2}-1}$
r_o-2	$\binom{N-2r_1-a}{2}$	$\binom{N-2r_1-a}{3}\binom{2r_1+a}{1}$	. . . .	$\binom{N-2r_1-a}{\frac{r_o}{2}+1}\binom{2r_1+a}{\frac{r_o}{2}-1}$
⋮		⋮		
2	$\binom{N-2r_1-a}{r_o-2}$	$\binom{N-2r_1-a}{r_o-1}\binom{2r_1+a}{1}$		
1	$\binom{N-2r_1-a}{r_o-1}$			
0	$\binom{N-2r_1-a}{r_o}$			

(b)

Fig. B-1. Classification of decodable points.

or in the form (Fig. B-1b)

$$\frac{\binom{2r_1+a}{b}\binom{N-2r_1-a}{r_o-k+b}}{\binom{2r_1+a}{r_o}} = \frac{r_o!(N-2r_1-a)\ .\ .\ .\ (N-2r_1-a-r_o+k+b)}{a!(r_o-k+b)!\,(2r_1+a-b)\ .\ .\ .\ (2r_1+a-r_o+1)} \tag{B-4b}$$

Ratios B-4a and B-4b are both decreasing functions of a. This fact, together with Eq. B-3, tells us that P_e is exponentially equivalent to the contribution from those spheres for which a = 1.

Among the spheres for which a = 1, we need only consider the points in the k, b category that makes the greatest contribution to P_e. This is obvious from the fact there are approximately $\frac{r_o^2}{2}$ terms in Fig. B-1, and a factor of $\frac{r_o^2}{2} = \frac{p_o^2N^2}{2}$ cannot influence the error exponent. It is easily shown that, if $p_1 \leq \frac{1}{2} - \frac{7p_o}{12}$ (a condition hardly more restrictive than the previous one), the largest term in each row in Fig. B-1 is the one on the extreme right. The largest contribution to P_e must therefore come from a term on the diagonal for which k = 2b. The general form of such a term is $\binom{2r_1+1}{r_o-b}\binom{N-2r_1-1}{b}$.

Since the probability of reception of a point is reduced by a factor of p_o/q_o with each succession row in Fig. B-1, b_M, the value of b which yields the maximum contribution to P_e, is determined by the following condition:

$$\frac{\binom{2r_1+1}{r_o-b_M-1}\binom{N-2r_1-1}{b_M+1}}{\binom{2r_1+1}{r_o-b_M}\binom{N-2r_1-1}{b_M}}\frac{p_o^2}{q_o^2} \doteq 1 \tag{B-5}$$

If we define $p_M = b_M/N$, a lengthy, but straightforward, calculation shows that

$$p_M = -\frac{1}{2}(2p_1-p_o) - \frac{p_o\alpha}{2} + \frac{1}{2}\sqrt{(2p_1-p_o)^2 + 4p_o^2p_1 + 2p_o^2\alpha + p_o^2\alpha^2} \tag{B-6}$$

where $\alpha = \frac{p_o}{1-2p_o}$.

In each of the $\frac{\binom{N}{r_1+1}}{\binom{2r_1+1}{r_1}}$ spheres nearest the origin, we must therefore consider

$\binom{2r_1+1}{r_o-b_M}\binom{N-2r_1-1}{b_M}$ points, each of which has $(2r_1+1-r_o+2b_M)$ 1's. As a result

$$P_e \dot{\gtrless} \frac{\binom{N}{r_1+1}}{\binom{2r_1+1}{r_1}} \binom{2r_1+1}{r_o-b_M}\binom{N-2r_1-1}{b_M} p_o^{(2r_1+1-r_o+2b_M)} q_o^{(N-2r_1-1+r_o-2b_M)} \tag{B-7}$$

Making use of Sterling's approximation, we find that a lower bound on the error exponent is given by

$$\begin{aligned}\lim_{P_e\to 0} \frac{\log P_e^{-1}}{N} \geqslant\ & (1-p_1)\log(1-p_1) + (p_o-p_M)\log(p_o-p_M) \\ & + (2p_1-p_o+p_M)\log(2p_1-p_o+p_M) + p_M \log p_M \\ & + (1-2p_1-p_M)\log(1-2p_1-p_M) - p_1 \log p_1 \\ & - (1-2p_1)\log(1-2p_1) - \log q_o + (2p_1-p_o+2p_M)\log\frac{q_o}{p_o}\end{aligned} \tag{B-8}$$

This bound is plotted in Fig. 9 (curve D) for $p_o = .1$. At rates very close to capacity, the decoding region occupies nearly all of the surrounding spheres. Because a sphere-packing array cannot be achieved, in general, some decoding regions near the origin must overlap. As the result of our having counted the points in these overlapping regions twice in calculating P_e, the lower bound of Eq. B-8 is negative for R close to C.

The closeness of the upper and lower bounds results from the fact that b_M, which distinguishes the bounds of Eqs. 21 and B-7 when r_2 is set equal to r_o in the former, is only a small fraction of r_o. This can be seen by an examination of Table B-1, which gives the values of p_M corresponding to several values of p_1, all for $p_o = .1$.

Table B-1. Values of p_M and p_1 for $p_o = .1$.

p_1	.10	.11	.12	.15	.20	.30
p_M	.0083	.0070	.0060	.0040	.0024	.001

APPENDIX C

C.1 Derivation of an expression for S(d). (Natural units of information (nats) will be used for analytical convenience.)

Case A: $d \leq \frac{1}{2}$

We define P_c to be the probability that TI includes one of the cut positions; $P_u = 1 - P_c$, the probability that it does not. The corresponding average values of the information provided about TI per transmission will be represented by $S_c(d)$ and $S_u(d)$. Then

$$S(d) = P_u S_u(d) + P_c S_c(d) \tag{C-1}$$

It is easily seen that

$$P_c = 2d \tag{C-2}$$

$$S_u(d) = q_o \ln 2q_o + p_o \ln 2p_o \tag{C-3}$$

It remains to calculate $S_c(d)$.

On those occasions when TI includes one of the cut positions, we let α represent the smaller of the two fractional parts into which TI is divided by the cut. The probability distribution for the location of the transmitted point is uniform over TI. If the transmitted point lies in the larger part of TI and the transmitted symbol is received correctly, or if it lies in the smaller part of TI and a crossover occurs, the TI size after reception is $2p_o(\alpha d) + 2q_o(1-\alpha)\, d$. Thus, with probability $\alpha p_o + (1-\alpha)q_o - (q_o - p_o)\alpha$, the information received about TI is $\ln 2[q_o - (q_o - p_o)\alpha]$. If either of the two remaining compound events occurs, the TI size after reception is $2q_o(\alpha d) + 2p_o(1-\alpha)\, d$. Thus, with probability $\alpha q_o + (1-\alpha)p_o = p_o + (q_o - p_o)\alpha$, the received information is $\ln 2[p_o + (q_o - p_o)\alpha]$. For a given value of α, the average information received per transmission about TI is

$$S_{c_\alpha}(d) = [q_o - (q_o - p_o)\alpha] \ln 2[q_o - (q_o - p_o)\alpha] + [p_o + (q_o - p_o)\alpha] \ln 2[p_o + (q_o - p_o)\alpha] \tag{C-4}$$

Since α is uniformly distributed over the range, $0 \leq \alpha \leq \frac{1}{2}$,

$$S_c(d) = 2 \int_0^{1/2} S_{c_\alpha}(d)\, d\alpha \tag{C-5}$$

By substituting Eqs. C-2 – C-5 in Eq. C-1, we find, after considerable manipulation,

$$S(d) = 4d \int_0^{1/2} S_{c_\alpha}(d)\, d\alpha + (1-2d)\,(q_o \ln 2q_o + p_o \ln 2p_o) = C - \left(1 - \frac{2p_o q_o}{q_o - p_o} \ln \frac{q_o}{p_o}\right) d \tag{C-6}$$

where $C = q_o \ln 2q_o + p_o \ln 2p_o$ is the channel capacity.

Case B: $d > \frac{1}{2}$

In this case TI includes either one or both of the cut positions. We define P_{1c} and

P_{2c}, respectively, to be the probabilities of these two events. We further define $S_{1c}(d)$ and $S_{2c}(d)$ to be the corresponding average values of the information provided about TI per transmission. Then

$$S(d) = P_{1c}S_{1c}(d) + P_{2c}S_{2c}(d) \tag{C-7}$$

It is evident that

$$P_{2c} = 2\left(d - \frac{1}{2}\right) \tag{C-8}$$

$$P_{1c} = 2(1-d) \tag{C-9}$$

By reasoning analogous to that for case A, it is clear that when TI includes both cut positions, the possible TI sizes after transmission are $2q_o\left(\frac{1}{2}\right) + 2p_o\left(d - \frac{1}{2}\right)$ and $2p_o\left(\frac{1}{2}\right) + 2q_o\left(d - \frac{1}{2}\right)$. The probabilities with which they occur are $q_o\left(\frac{1}{2d}\right) + p_o\left(1 - \frac{1}{2d}\right)$ and $p_o\left(\frac{1}{2d}\right) + q_o\left(1 - \frac{1}{2d}\right)$, respectively. It follows that

$$S_{1c}(d) = \left(p_o + \frac{q_o - p_o}{2d}\right) \ln \left(2p_o + \frac{q_o - p_o}{d}\right) + \left(q_o - \frac{q_o - p_o}{2d}\right) \ln \left(2q_o - \frac{q_o - p_o}{2d}\right)$$

If TI includes only one of the cut positions and the smaller fractional part of TI is again α, $S_{c_\alpha}(d)$ is still given by Eq. C-4. However, the range of α is now restricted to $1 - \frac{1}{2d} < \alpha \leqslant \frac{1}{2}$. Therefore,

$$S_c(d) = \frac{2d}{1-d} \int_{1-\frac{1}{2d}}^{1/2} S_{c_\alpha}(d)\, d\alpha \tag{C-10}$$

By substituting Eqs. C-4 and C-8 – C-10 in Eq. C-7, we finally get, after much additional manipulation,

$$S(d) = \left(q_o + \frac{2p_oq_od}{q_o - p_o}\right) \ln \left(2p_o + \frac{q_o - p_o}{d}\right) + \left(p_o - \frac{2p_oq_od}{q_o - p_o}\right) \ln \left(2q_o - \frac{q_o - p_o}{d}\right) + d - 1 \tag{C-11}$$

C.2 Derivation of an expression for $S_W(d)$

We assume that $d \geqslant \frac{1}{2}$. Then WI includes one or neither of the two cut positions. We represent the probabilities of these two events by P_c and P_u, respectively. The corresponding average values of the information received about WI per transmission are $S_{Wc}(d)$ and $S_{Wu}(d)$. Then

$$S_W(d) = P_cS_{Wc}(d) + P_uS_{Wu}(d) \tag{C-12}$$

$$P_c = 2(1-d) \tag{C-13}$$

$$P_u = 2d - 1 \tag{C-14}$$

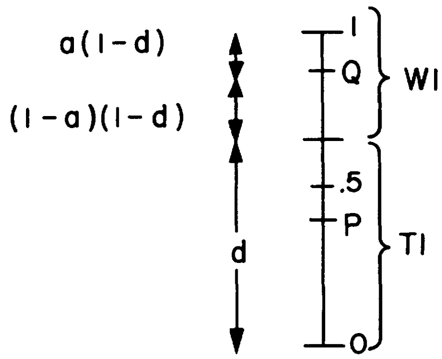

Fig. C-1. Pertaining to the asymptotic portion of the transmission process. (P and Q are cut positions.)

In the event that WI is not cut, the two possible values of received information are $\ln 2p_o$ and $\ln 2q_o$. The former results if the transmitted point lies in the opposite half of the interval (0, 1)' from WI and the channel symbol is received correctly, or if it lies in the same half and a crossover occurs. The sum of the probabilities of these two joint events is $q_o\left(\frac{1}{2d}\right) + p_o\left(1-\frac{1}{2d}\right)$. The two other possible joint events have a combined probability of $p_o\left(\frac{1}{2d}\right) + q_o\left(1-\frac{1}{2d}\right)$. Consequently,

$$S_{Wu}(d) = \left(p_o + \frac{q_o - p_o}{2d}\right) \ln 2p_o + \left(q_o - \frac{q_o - p_o}{2d}\right) \ln 2q_o \tag{C-15}$$

We let a represent the smaller fractional part of WI when it is cut. (See Fig. C-1, in which WI has been placed entirely at one end of the interval (0, 1)' for ease of visualization.) It is clear from the diagram that, with probability $\frac{q_o}{d}\left[\frac{1}{2} - a(1-d)\right] + \frac{p_o}{d}\left[\frac{1}{2} - (1-a)(1-d)\right]$, the WI size after transmission is $2q_o a(1-d) + 2p_o(1-a)(1-d)$. The other possible size, $2p_o a(1-d) + 2q_o(1-a)(1-d)$, occurs with probability

$$\frac{p_o}{d}\left[\frac{1}{2} - a(1-d)\right] + \frac{q_o}{d}\left[\frac{1}{2} - (1-a)(1-d)\right]$$

As a result, the average information received about WI per transmission is

$$S_{Wc_a}(d) = \left\{\frac{1}{2d} - \left(\frac{1}{d} - 1\right)\left[p_o + (q_o - p_o)a\right]\right\} \ln 2\left[p_o + (q_o - p_o)a\right]$$

$$+ \left\{\frac{1}{2d} - \left(\frac{1}{d} - 1\right)\left[q_o - q_o(q_o - p_o)a\right]\right\} \ln 2\left[q_o - (q_o - p_o)a\right] \tag{C-16}$$

Since a is distributed uniformly over the range, $0 < a \leq \frac{1}{2}$,

$$S_{Wc}(d) = 2\int_0^{1/2} S_{Wc_a}(d)\, da \tag{C-17}$$

The result of substituting Eqs. C-13 – C-17 in Eq. C-12 is

$$S_W(d) = \frac{1}{d}\left[\frac{1}{2(q_o - p_o)} \ln \frac{q_o}{p_o} - 1\right] + d\left[1 - \frac{2p_o q_o}{q_o - p_o} \ln \frac{q_o}{p_o}\right] + \ln 2 + p_o \ln q_o + q_o \ln p_o \tag{C-18}$$

APPENDIX D - AN UPPER BOUND ON P_1

The argument used to obtain the bound on P_1 will be outlined here. For further details, see Shannon (16).

P_1 is defined as the probability that WI will be expanded from $1-d_b$ to $1-d_b'$ before it is reduced to P_e. If we let x_i stand for the information provided about WI by the i^{th} transmission (starting from size $1-d_b$), and if we define $I_n = \sum_{i=1}^{n} x_i$, we have

$$P_1 \leqslant \Pr[\text{any } I_n \geqslant A] \leqslant \sum_{n=1}^{\infty} \Pr[I_n \geqslant A] \tag{D-1}$$

where $A = \ln \frac{1-d_b'}{1-d_b}$

Since the probability distribution for x_i involves I_{i-1} as a parameter, I_n is the sum of n dependent random variables. It can be shown that

$$\Pr I_n \geqslant \left[\sum_{i=1}^{n} \mu_i'(s)\right] \leqslant e^{\sum_{i=1}^{n} \mu_i(s) - s \sum_{i=1}^{n} \mu_i'(s)} \qquad s \geqslant 0 \tag{D-2}$$

where the $\mu_i(s)$ are (uniform) bounds on the semi-invariant generating functions (s.i.g.f.) $\mu_i(s/x_1, \ldots, x_{i-1})$ associated with the variables $x_i (i = 1, 2, \ldots n)$. Equation D-2 is identical in form to the Chernoff bound for sums of independent random variables.

We shall apply Eq. D-2 to our situation by finding a function $\mu(s)$ that is such that we can set $\mu_i(s) = \mu(s)$ for all $i \leqslant n$. The distribution of x_i is taken from an ensemble of distributions having d as a parameter, $d_b' < d < 1$. Let x_d stand for a random variable chosen from the member of this ensemble with parameter d. If we designate by $\mu_d(s)$ any corresponding bounding s.i.g.f., $\mu(s)$ must satisfy the relation, $\mu(s) \geqslant \mu_d(s)$, $d_b' < d < 1$.

Note that regardless of the value of d, $\ln 2p_o \leqslant x_d \leqslant \ln 2q_o$. We choose for $\mu_d(s)$ the s.i.g.f. of the probability distribution in Fig. D-1. c(d) is chosen so that $c(d) \ln 2q_o + [1-c(d)] \ln 2p_o = \overline{x}_d = S_W(d)$. We can think of constructing this distribution from the distribution for x_d by subdividing the open interval, $\ln 2p_o < x_d < \ln 2q_o$, into infinitesimal intervals, and then shifting the probability assigned to each infinitesimal

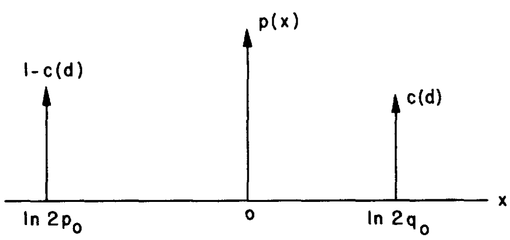

Fig. D-1. Probability distribution yielding $\mu_d(s)$.

interval to the extreme values of $\ln 2p_o$ and $\ln 2q_o$ in such a way that the average value of the distribution is unchanged after each transfer of probability. Since the s.i.g.f. of a probability distribution p(x) is given by $\ln \int_{-\infty}^{\infty} e^{sx} p(x)\, dx$, the convex downward shape of e^{sx} (as a function of x) tells us that the s.i.g.f. can only be increased by this process. Hence our choice for $\mu_d(s)$ is permissible.

We recall that d_b' is determined by the relation, $S_W(d_b') = S_W(1)$, and that $S_W(d) < S_W(1)$ for $d_b' < d < 1$. Therefore, $c(d) < c(1) = \frac{1}{2}$ for $d_b' < d < 1$. It follows, since $s \geq 0$, that

$$\mu_d(s) = \ln \left\{ (2q_o)^s c(d) + (2p_o)^s [1-c(d)] \right\} \leq \ln \frac{1}{2} \left[(2q_o)^s + (2p_o)^s \right]$$

We therefore choose

$$\mu(s) = \ln \frac{1}{2} \left[(2q_o)^s + (2p_o)^s \right] \tag{D-3}$$

Equation D-2 now becomes

$$P_r[I_n \geq A] \leq e^{n[\mu(s) - s\mu'(s)]} \qquad s \geq 0 \tag{D-4}$$

where s is chosen so that $n\mu'(s) = A$. At this point, a rather detailed argument (16) is needed to show that Eq. D-1, together with Eq. D-4, implies that

$$P_1 \leq e^{-s_o A} \left[\frac{s_o A}{\mu(s_1)} + \frac{1}{1 - e^{\mu(s_1)}} \right] \tag{D-5}$$

where $\mu(s_o) = \mu'(s_1) = 0$, $s_o > 0$, and $s_1 > 0$.

Acknowledgment

I wish to express my gratitude to Professor Peter Elias for providing the impetus for this study and the patient guidance that was vital to its successful completion. His continued encouragement made the writing of my thesis a most rewarding experience.

I also wish to thank the members of the Processing and Transmission of Information Group, at the Research Laboratory of Electronics, M.I.T., in particular, my fellow-student Robert G. Gallager, for many helpful suggestions.

I would also like to express my appreciation for the support of the Lincoln Laboratory, M.I.T., where I was a Staff Associate for two and one-half years.

References

1. C. E. Shannon and W. Weaver, The Mathematical Theory of Communication (University of Illinois Press, Urbana, Ill., 1949).
2. C. E. Shannon, The zero error capacity of a noisy channel, Trans. IRE, Vol. IT-2, No. 3, September 1956.
3. A. Feinstein, Error bounds in noisy channels without memory, Trans. IRE, Vol. IT-1, No. 2, September 1955.
4. P. Elias, Coding for noisy channels, IRE National Convention Record, Part 4, 1955.
5. D. Slepian, A class of binary signaling alphabets, Bell System Tech. J. 35, 203-234 (1956).
6. P. Elias, Error-free coding, Trans. IRE, Vol. PGIT-4, September 1954.
7. R. W. Hamming, Error detecting and error correcting codes, Bell System Tech. J. 29, 147-160 (1950).
8. J. M. Wozencraft, Sequential Decoding for Reliable Communication, Technical Report 325, Research Laboratory of Electronics, M.I.T., August 9, 1957.
9. M. Horstein, An Experimental Study of Sequential Decoding for the Binary Symmetric Channel, Group Report 34-74, Lincoln Laboratory, M.I.T., November 1958.
10. S. S. L. Chang, Theory of information feedback systems, Trans. IRE, Vol. IT-2, No. 3, September 1956.
11. J. L. Kelly, Jr., A new interpretation of information rate, Bell System Tech. J. 35, 917-926 (1956).
12. P. Elias, Channel capacity without coding, Quarterly Progress Report, Research Laboratory of Electronics, M.I.T., October 15, 1956, pp. 90-93.
13. J. M. Wozencraft and M. Horstein, Coding for two-way channels, Fourth London Symposium on Information Theory, London, 1960; Technical Report 383, Research Laboratory of Electronics, M.I.T. (to be published).
14. E. M. Hofstetter, Asymptotic behavior of optimum fixed-length and sequential dichotomies, Quarterly Progress Report No. 53, Research Laboratory of Electronics, M.I.T., April 15, 1959, pp. 117-121.
15. A. Wald, Sequential Analysis (John Wiley and Sons, Inc., New York, 1947), pp. 170-171.
16. C. E. Shannon, Bounds on the Tails of Martingales and Related Questions, Research Laboratory of Electronics, M.I.T. (unpublished memorandum).
17. W. Feller, An Introduction to Probability Theory and Its Applications (John Wiley and Sons, Inc., New York, 2d edition, 1957).