Copy 2

SEQUENTIAL ENCODING AND DECODING FOR THE DISCRETE MEMORYLESS CHANNEL

BARNEY REIFFEN

Lincoln Copy

AUGUST 12, 1960

RESEARCH LABORATORY OF ELECTRONICS
TECHNICAL REPORT 374

LINCOLN LABORATORY
TECHNICAL REPORT 231

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
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.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 12, 1960

Research Laboratory of Electronics Technical Report 374

Lincoln Laboratory Technical Report 231

SEQUENTIAL ENCODING AND DECODING FOR THE DISCRETE MEMORYLESS CHANNEL

Barney Reiffen

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 Philosophy.

Abstract

A scheme is described which sequentially encodes the output of a discrete letter source into the input symbols of a discrete memoryless channel, with adjacent channel symbols mutually constrained over a length, n. The encoder permits desired channel input symbol probabilities to be approximated closely. Decoding at the receiver is accomplished with delay n by means of sequential tests on potential transmitted sequences with reject criteria set so that incorrect sequences are likely to be rejected at short lengths and the correct sequence is likely to be accepted. Averaged over a suitably defined ensemble of encoders, the decoding scheme has an average probability of error, with an upper limit whose logarithm approaches $-nE(R)$ for large n. $E(R)$ is dependent only on the data rate, R. For a channel symmetric at its output with equally likely inputs, the exponent $E(R)$ is optimum for rates greater than a rate called R_{crit}. For such symmetric channels, a computation cutoff rate, R_{cutoff} is defined. For $R < R_{cutoff}$, the average number of decoding computations per symbol does not grow exponentially with n; it is upper bounded by a quantity proportional to n raised to the power $2[1+R/R_{cutoff}]$. A procedure for reducing an arbitrary discrete memoryless channel to a symmetric channel is given.

Acknowledgment

The environment provided by the community of scholars associated with the Data Processing Group of the Research Laboratory of Electronics made this research possible. I want especially to thank Professor P. Elias, my supervisor, and Professor C. E. Shannon who were always available for guidance and encouragement. The original work of Professor J. M. Wozencraft on sequential decoding provided the foundation for this work. My debt to him should be obvious to all who read this work. I also want to acknowledge helpful discussions with R. G. Gallager. A. I. Grayzel of Lincoln Laboratory suggested certain simplifications in the proofs of Theorems 1 and 2.

I want also to thank Professor R. M. Fano and Dr. H. Sherman of Lincoln Laboratory who encouraged me to undertake the graduate study which led to this research. Furthermore, I want to thank Lincoln Laboratory for permitting me to pursue this study as a Staff Associate.

TABLE OF CONTENTS

CONTENTS

CHAPTER I

INTRODUCTION

Shannon[1], [2] demonstrated the existence of block codes of length n, which signal over a discrete memoryless channel at a rate R less than channel capacity C with a probability of error* less than $2e^{-nE(R)}$, where E(R) is a positive exponent which is a function of R but independent of n. Shannon defines an ensemble of codes in a manner which facilitates the bounding of the probability of error averaged over the ensemble, P_e. If we randomly select a code from the ensemble with its corresponding probability, then, with probability at least 1-f, the code's probability of error will be no more than P_e/f. Thus we may expect almost any code of length n selected at random to have a probability of error bounded by a quantity proportional to $e^{-nE(R)}$.

The optimum decoding procedure for block codes requires that for each possible received sequence of length n, the input message most likely to have caused it be stored (codebook decoding). Since the number of possible output messages increases exponentially with n, the amount of storage required for decoding increases exponentially. Alternately, one can dispense with this large storage requirement and, from a knowledge of the code and channel statistics, calculate the a posteriori probability of each possible input sequence conditional upon the particular output sequence received and thereby accomplish a maximum likelihood decision. Since the number of possible input sequences $M = e^{nR}$ increases exponentially with n if R is to be held fixed, the amount of computation required for decoding increases exponentially with n.

---

(1), (2) All numbered references are listed on page 98. The results of (2) are stronger than the corresponding results of (1).

* In this discussion, by probability of error of a code, we mean average probability of error when the admissible inputs are equally likely.

The encoder at the transmitter must map one of $M = e^{nR}$ events into one of e^{nR} possible transmitted sequences associated with the code used. Similarly, if the receiver is to compute the set of a posteriori probabilities, it must duplicate the encoder. It is clear that in general an arbitrary code will require an encoder of complexity proportional to e^{nR}. In the special cases of the binary symmetric channel (BSC) and the binary erasure channel (BEC), the ensemble of parity check codes yields an upper bound to probability of error with the same exponential behavior as the set of all codes.[3]

If a binary sequence of length k is viewed as a vector in a k dimensional vector space V_k, parity check codes may be viewed as the set of linear transformations of the vectors in V_k to vectors in an n dimensional vector space V_n. The vector spaces are defined on the field $Z_2 = \{ 0, 1 \}$. The sequence of binary digits of length n corresponding to the vector in V_n is transmitted over the channel. The rate $R = \frac{k}{n}$. Clearly, a parity check code is defined by the k x n matrix that corresponds to the transformation from V_k to V_n.

Let the k x n matrix of a parity check code be partitioned as in Figure 1. Elias[3] showed that if $M_1 = I$, the k x k identity matrix, and M_2 is filled in by letting its first row be the first k digits of some generator sequence G of length n, its second row be the second to $(k + 1)^{st}$ digits of G, etc., then the set of all such codes have a probability of error with optimum exponent for rates near capacity. These are called "sliding parity check" codes. In this case, encoding may be accomplished with an amount of computation per check digit that is proportional to the code length n.

In the same paper, Elias defines convolutional "parity check symbol encoding" where information digits are interspersed with check digits, each of which is determined as the mod 2 sum of a fixed pattern of the preceding

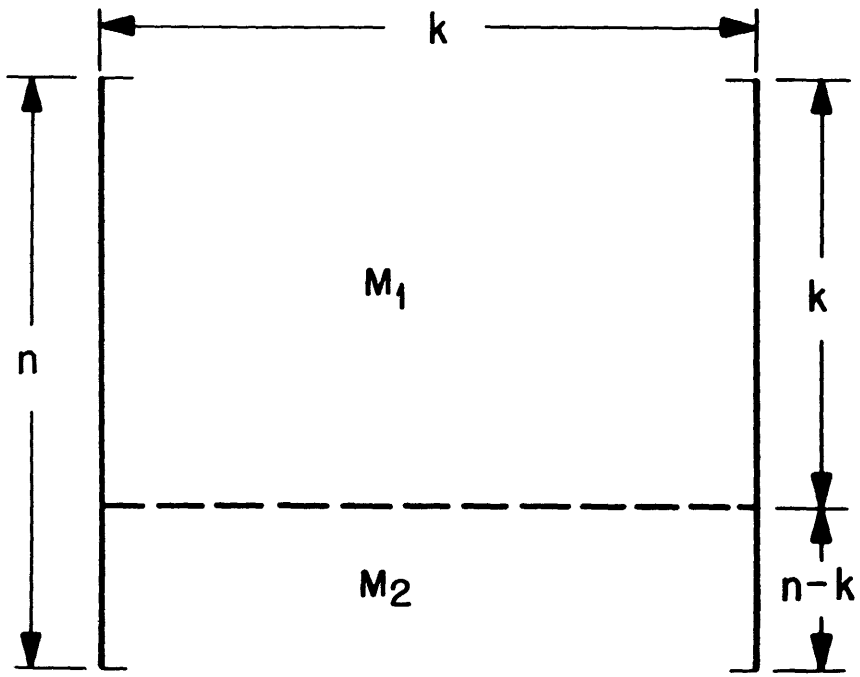


FIG. 1 PARITY CHECK MATRIX

$k = nR$ information digits.* This permits the check digits to be determined sequentially for information sequences of arbitrary length. Again, encoding is accomplished with an amount of computation per digit which is proportional to n, the effective constraint length. Individual information digits may be decoded after a delay of n digits with an error proportional to $\exp [-nE(R)]$, where $E(R)$ is optimum for rates near capacity.

These special means of encoding for channels with binary inputs which are to be used with equal frequency did not solve the decoding requirement of exponential storage or computation.

Wozencraft[4] devised a procedure which permits decoding the BSC with a computational complexity which increases only algebraically with n, and has probability of error which decreases exponentially with n, and with the optimum exponent for rates near capacity. He modified Elias' convolutional parity check encoding by convolving a fixed generator sequence G of length n with the last k information places which are expanded to length n by the interspersal of redundancy places set equal to 0. For example, if $R = \frac{k}{n} = \frac{1}{2}$, a message sequence would read...IOIOIO... where I is an information digit which may be 0 or 1. The encoded sequence is the output of the convolution operation. The message sequence steps once for each encoded digit.

Wozencraft specified a sequential procedure at the receiver which decodes individual information digits. The procedure is to compare the received sequence against branches of the tree generated by the convolutional generator starting from a particular point. All decisions up to this point are assumed correct. A path is followed until the a posteriori probability that it caused

---

* This description is applicable for $R = 1 - \frac{1}{m}$, where m is a positive integer.

the actual received sequence is less than some value 2^{-K_1}. A path may be eliminated before n digits are considered. If so, all branches originating from the end of this path are no longer considered. The process continues until either of the following occur:

1. A path is found which satisfies the criterion (is more likely than 2^{-K_1}). In this case a decision is made on the first branch of this path. This is equivalent to deciding on the information digit that produced this path.
2. No path satisfies the criterion. In this case the procedure is repeated for criterion $K_2 = K_1 + \Delta K$. In general, $K_j = K_{j-1} + \Delta K$.

Wozencraft indicated that for $R < R_{cutoff} \leqq C$, where R_{cutoff} is a function of C, sequential decoding may be accomplished with the average number of computations required per digit increasing no faster than An log n where A is a constant independent of n, but dependent on R and C. The probability that a digit is decoded incorrectly is bounded by $P_e < Bn \exp [-nE(R)]$ where E(R) is the optimum exponent for rates near capacity, and B is independent of n, but dependent on R and C. These results are conditional on the assumption that no prior decoding errors have been made.

Wozencraft bounded the computations required to eliminate all sequences starting with the digit to be decoded when it is incorrect. He did not, however, bound the computations required to accept a sequence starting with the digit to be decoded when it is correct. In unpublished work, R. Gallager has modified the decoding procedure in such a manner that the average number of computations may be firmly bounded. His result is that the average number of computations is less than $An^4 \log n$. The modification in the decoding procedure causes the probability of error bound to be deteriorated to $Bn^2 e^{-nE(R)}$.

Epstein[5] obtained some interesting results for the Binary Erasure Channel (BEC). Due to the especially simple nature of this channel (we can never confuse a zero for a one or a one for a zero; the channel can only introduce an ambiguity (erasure)), block decoding need not require a complexity growing more quickly than n^3. However, if convolutional encoding and sequential decoding is used, the average number of computations required per digit is independent of n, and dependent only on R and C. This result is applicable for all $R < C$.

In this research, we consider the problem of generalizing Wozencraft's work to the general discrete memoryless channel. There are four main aspects to this generalization: (1) encoding and decoding, (2) decoding complexity, (3) error behavior, and (4) channel inclusion. Our results for each aspect are summarized below.

1. Encoding and Decoding: A scheme is described in Chapter II which sequentially encodes the output of an arbitrary discrete letter source into the input symbols of a discrete memoryless channel, with adjacent channel symbols mutually constrained over a length, n. The encoder permits desired channel input symbol probabilities to be approximated arbitrarily closely in order to optimize the error exponent. Two closely related decoding schemes are described in Chapter III. These operate by performing sequential tests on potential transmitted sequences with reject criteria so set that incorrect sequences are likely to be rejected at short lengths, and the correct sequence is likely to be accepted. The decoding delay equals the effective constraint length, n.

2. Decoding Complexity: We discuss in Chapter IV the decoding complexity averaged over a suitably defined ensemble of encoding shcemes of Chapter II. The discussion is restricted to channels symmetric at their output* with equally likely inputs. The simpler of the two decoding schemes suggests, but does not guarantee, decoding with an average number of computations upper-bounded by a quantity in the form An^B, where n is the decoding delay, A is independent of n, but dependent on the source rate R, and $B \leqq R/R_{cutoff}$, where R_{cutoff} is a quantity defined in Chapter IV. This result holds for $R < R_{cutoff}$ ($B < 1$). The more complex decoding scheme guarantees decoding with an average number of computations upper-bounded by $An^{2(1+B)}$. This, too, holds for $R < R_{cutoff}$. The expression for R_{cutoff} reported here is particularized to the BSC in Chapter VI. It is greater than or equal to the previously reported estimate of it[4], with equality only in the trivial cases of crossover probability equal to 0 and 1/2.

3. Error Behavior: It is shown in Chapter V that both decoding schemes applied to the general channel give rise to an average probability of error over the ensemble less than a quantity proportional to $n^2 \exp[-nE_1(R)]$, where the exponent $E_1(R)$ is positive for all $R < C$, and independent of n. For large n, $E_1(R)$ approaches the exponent reported by Shannon[2]. In the special case of a channel symmetric at its output with equally likely inputs, the average probability of error for both schemes is less than a

---

* A channel symmetric at its output is defined on page 46 .

quantity proportional to $n^3 \exp[-nE_2(R)]$. For large n, $E_2(R)$ approaches the optimum error exponent for R greater than a rate, R_{crit}, defined in Chapter V. In all non-trivial cases, $R_{crit} < R_{cutoff}$.

4. Channel Inclusion: We show in Chapter VI that with reasonably simple switching logic at the input and output of an asymmetric channel, it can be made to look like a symmetric channel insofar as the sequential encoding and decoding procedures are concerned. Thus the decoding complexity results of Chapter IV can be made to apply to a general discrete memoryless channel.

If, over the encoding ensemble, the average number of decoding computations is $\bar{N}$, and the average probability of error is P_e, then a random selection from the encoding ensemble has a probability no more than f_1 of requiring an average number of decoding computations greater than $\bar{N}/f_1$. Similarly, the random selection has a probability no more than f_2 of giving rise to a probability of error greater than P_e/f_2. Thus a random selection has probability at least $1 - f_1 - f_2$ of <u>simultaneously</u> requiring an average number of computations no greater than $\bar{N}/f_1$ and producing a probability of error no greater than P_e/f_2. Thus we may expect almost any random selection from the encoding ensemble to exhibit satisfactory decoding complexity and probability of error characteristics simultaneously. Thus we have demonstrated that data can be sent at a finite rate over a discrete memoryless channel with an error that varies exponentially with decoding delay, n, and with a decoding complexity that varies, <u>not</u> exponentially with n, but as n raised to some power.

CHAPTER II

SEQUENTIAL ENCODING

Suppose a discrete memoryless channel has input symbols 1, . . . , a, output symbols 1, . . . , b, and transition probabilities $q_i(j)$ of receiving output symbol j when input symbol i is transmitted. This channel has a capacity, C. If we desire to transmit over this channel at a rate R, $R < C$, then we know from Shannon's work[1] that if block codes of length n are used, there is a choice of input symbol a priori probabilities which maximizes the exponent in the probability of error expression. Thus we may expect the performance of any coding scheme, block or otherwise, to be dependent on the frequency with which the various channel symbols are used.

Our object in this section is to describe an encoding scheme for mapping output sequences from an independent letter source into sequences of channel input symbols which are used with a desired probability. We desire to do this encoding in a sequential manner so that sequential decoding may be attempted at the receiver. By sequential encoding we mean that the channel symbol to be transmitted at any time is uniquely determined by the sequence of output letters from the message source up to that time. Sequential decoding will be discussed in the next chapter.

We will ignore the diaphantine problem for the moment and assume that an independent letter source delivers one of m equally likely letters each second, where log m = R, the desired rate. We let s_t^k be the event that at time t, the independent letter source output is the k^{th} letter, k = 1, . . . , m. Further, we let s_t be the output at time t, whichever letter it is, and s^k be the k^{th} output letter, whenever it occurs. Also we let $S = (\ldots, s_{t-1}, s_t, s_{t+1}, \ldots)$

be any sequence of source output letters, either finite or infinite. The set of all S sequences is denoted by $\{S\}$.*

The encoding problem is to map a source sequence S into a sequence of channel input symbols $U = (\ldots u_{t-1}, u_t, u_{t+1}, \ldots)$ such that adjacent u symbols are interdependent. It is precisely this dependency which is to be employed in decoding with low probability of error. We desire to encode source letters sequentially, i.e., one by one. This is so that the decoding at the receiver can proceed sequentially. A natural way to encode sequentially and build in the intersymbol constraints is to let u_t be some function, ψ, of the n previous source letters:

$$u_t = \psi (s_t, s_{t-1}, \ldots, s_{t-n+1})$$

where n is the effective constraint length.

The function ψ should be simple enough to instrument, preferably with digital equipment. Further, a fixed function, ψ, should give rise to a probability measure on $\{U\}$ which on the one hand matches the frequency with which the various u symbols should be used for maximum error exponent at rate R, and on the other hand, permits quantitative evaluation of the decoding procedure. We now describe an encoding scheme that meets these objectives.

Suppose by some means or another, we know the desired probabilities of the channel input symbols corresponding to source rate R, and we approximate these probabilities by the rational fractions $p_i = N_i/D$ where the N_i are

---

* Throughout this research we have occasion to refer to sequences of events of various length and to the particular events occurring at particular times. We will consistently use the following notation: Lower case letters will denote events at particular times; upper case letters will denote sequences of these events. Subscripts on lower case letters will denote the time at which the event occurs. Superscripts on lower case letters denote to which of the severan possible events we are referring. Variations from this notation will be pointed out as they occur.

positive integers, i = 1, . . . , a and D is a prime number raised to a positive integral power. It is clear that we may approximate the desired probabilities arbitrarily closely with D chosen sufficiently large. Clearly $\sum_{i=1}^{a} N_i = D$. Our discussion of the encoding scheme will proceed with the assumption that D is a prime number. This is done so that the reader not familiar with the theory of finite fields will not be confused. In Appendix I, the discussion is generalized to $D = p^k$ where p is prime and k is a positive integer greater than one.

At this point we define the one-one mapping ϕ of s into the integers, mod D. Thus

$$\phi(s^k) = x^k, \quad k = 1, \ldots, m.$$

Since $\log m = R < C \leqq -\sum_{i=1}^{a} p_i \log p_i \leqq \log a \leqq \log D$, $m \leqq D$; thus one-one mappings are possible.

A map ϕ acting on a sequence of s symbols induces a sequence of x symbols:

$$\phi(S) = X$$

For transmission over the channel at rate R we desire to generate one symbol from the channel input alphabet each time the message source delivers a letter. Coding implies that adjacent channel alphabet symbols be statistically dependent. Suppose we desire the channel alphabet symbol occurring at integral time t to be dependent on the channel alphabet symbols occurring at times t - j, where j = 1, 2, . . . , n-1. This can be accomplished in the following manner: Let

$$y_t = \sum_{j=0}^{n-1} g_j x_{t-j} \pmod{D}$$

where the g_j are as yet unspecified elements of $Z_D = \{ 0, 1, \ldots, D-1 \}$, the set of integers mod D. Define G as the ordered set $(g_0, g_1, \ldots, g_{n-1})$. In other words, we have convolved a G sequence of length n with an **X** sequence of length n, where all multiplication and addition operations have been performed mod D. This may be visualized by referring to Figure 2.

Once each second the **X** sequence steps one box to the right and a new y is computed. The generator sequence G remains fixed. Clearly y_t is also an element of Z_D. Thus the convolution of **X** with G induces a sequence of y integers which we call Y. We shall use the notation

$$\mathbf{X} * G = Y$$

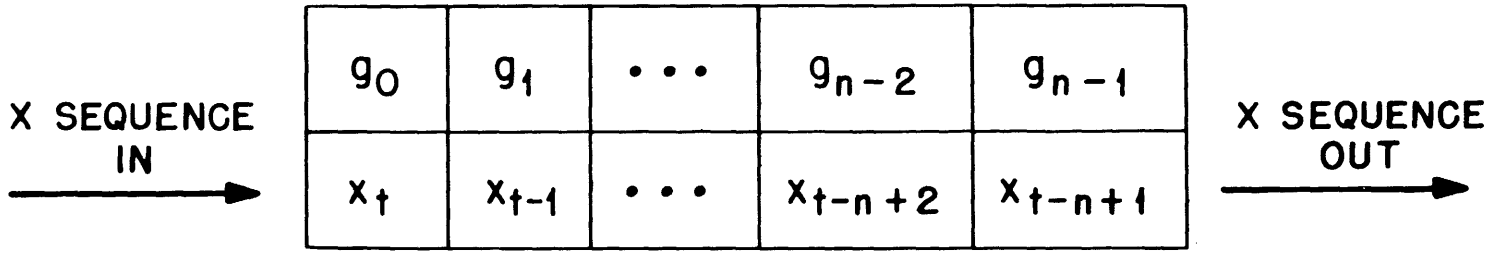
where the symbol * denotes convolution. The set of all possible G sequences will be denoted by $\{ G \}$. Clearly, there are D^n elements in $\{G\}$.

To each y_t we will add mod D an integer f_j as shown in Figure 3, where f_j is an element of Z_D.

Once each second the Y sequence steps one position to the right and the F wheel rotates one position counterclockwise. Here we have identified F with the ordered set $(f_1, \ldots, f_n)$. The output of this operation, x_t, is the mod D sum of y_t and the particular f_j immediately below it in Figure 3. Thus the addition of Y and F induces a sequence of z integers which we shall call Z. We shall use the notation

$$Y \oplus F = Z$$

where the symbol $\oplus$ denotes term by term addition mod D as shown in Figure 3. The set of all possible F sequences will be denoted by $\{ F \}$.



THE X SEQUENCE IS SHOWN IN POSITION TO COMPUTE y_t

FIG. 2 CONVOLUTION OPERATION

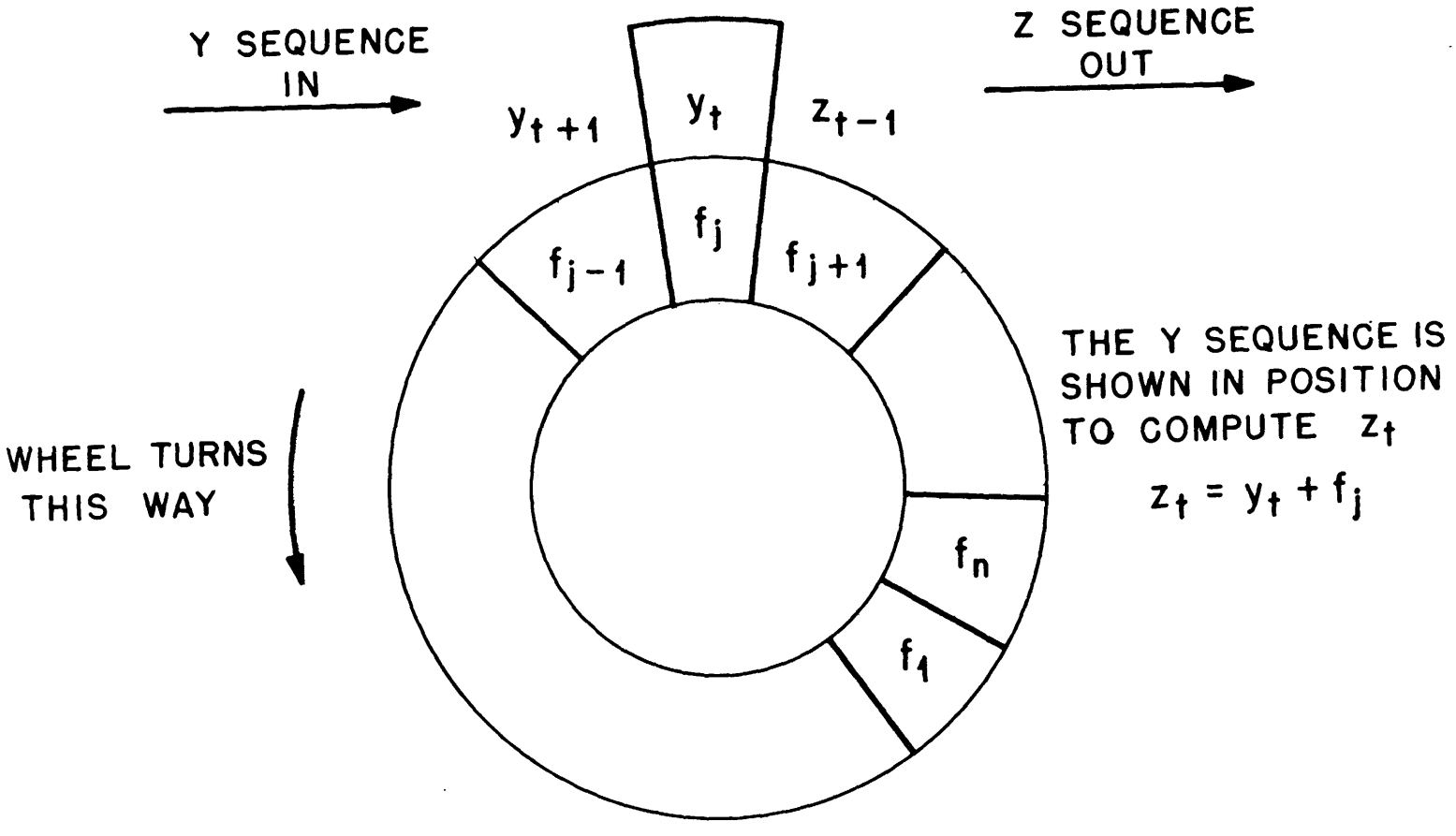


FIG. 3 ADDITION OPERATION

Clearly there are D^n elements in $\{F\}$. We now define a map Θ of the z integers onto* the channel input symbols. The domain of the map Θ is Z_D. We only consider a map which has the property that N_i elements of the domain have the channel input symbol i as their image, for i = 1, . . . , a. This condition is clearly consistent with the fact that

$$\sum_{i=1}^{a} N_i = D$$

We have discussed the case where we send one channel symbol for each output letter from the letter source. Since there are m possible output letters from the letter source, our rate R = log m. However, we must consider the situation where the desired rate $R \neq \log m$. There are two cases to consider:

1. $R < \log m$. Select R such that $\frac{R}{\log m}$ is a rational fraction less than 1, say α/β such that β divides n. We will transmit at the rate R if we modify the manner in which the X sequence to be convolved with G is formed. Suppose X is formed β places at a time. These β places will be made to correspond to a sequence of α symbols from the source. Thus, the source sequence $(s_1, s_2, \ldots s_\alpha)$ is to be mapped into the X sequence $(x_1, x_2, \ldots x_\beta)$. Let the i^{th} place of the S sequence map into the j^{th} place of the X sequence such such that $\phi(s_i^k) = x_j^k$, k = 1, 2, . . . m. The remaining $\beta - \alpha$ places of the X sequence are set identically equal to zero.

An example will make this clear. Suppose m = 3, α = 3, β = 5, $x^1 = 1$, $x^2 = 2$, $x^3 = 5$, and s_1, s_2, s_3 map into x_1, x_3, and x_5 respectively. Then the sequence $S = (s_1^2, s_2^3, s_3^1)$ is mapped into the sequence X = (2, 0, 5, 0, 1).

---

* By "onto" we mean the range of the map Θ is all of the channel input symbols.

2. Suppose $R > \log m$. Then define a new message source by taking γ letters at a time such that $R < \gamma \log m$. At this point the problem reduces to case 1.

We recapitulate the above discussion in the diagram of Figure 4.

The behavior of the encoding scheme for fixed ϕ, G, F, and Θ is difficult to evaluate. However, we shall define an ensemble of encoding schemes whose average behavior over the ensemble of codes can be evaluated. We proceed to show that if F and G are randomly selected, no further restrictions are necessary to obtain meaningful results.

First of all, we desire to find conditions such that $\Pr [u_t^i] = p_i$ for all t and $i = 1, \ldots, a$. We call this condition 1.

Theorem 1: Condition 1 is satisifed if the elements of $\{ F\}$ are equally likely.

The proofs of this and the following theorems are in Appendix A.

Secondly we desire to find conditions such that, if $U = (u_1, \ldots, u_w), w \leqq n$ $\Pr [U] = \prod_{r=1}^{w} p_{u_r}$. We call this condition 2.

Theorem 2: Condition 2 is satisfied if the elements of $\{ F\}$ are equally likely.

Finally, let S and S' be two sequences of output letters from the message source that are mapped into X and X' respectively by ϕ. Let $x_t = x_t'$ for $t \leqq 0$ and $x_1 \neq x_1'$. Suppose X and X' induce channel input sequences U and U' respectively. We will restrict our attention to U sequences of length n, i.e., $U = (u_1, \ldots, u_n)$ and $U' = (u_1', \ldots, u_n')$. We desire to find conditions for which U and U' are statistically independent. We call this condition 3.

Theorem 3: Condition 3 is satisfied for any pair (X, X') where $x_t = x_t'$, $t \leqq 0$ and $x_1 \neq x_1'$ if (a) the elements of $\{ F\}$ are equally likely, and (b) the elements of $\{ G\}$ are equally likely.

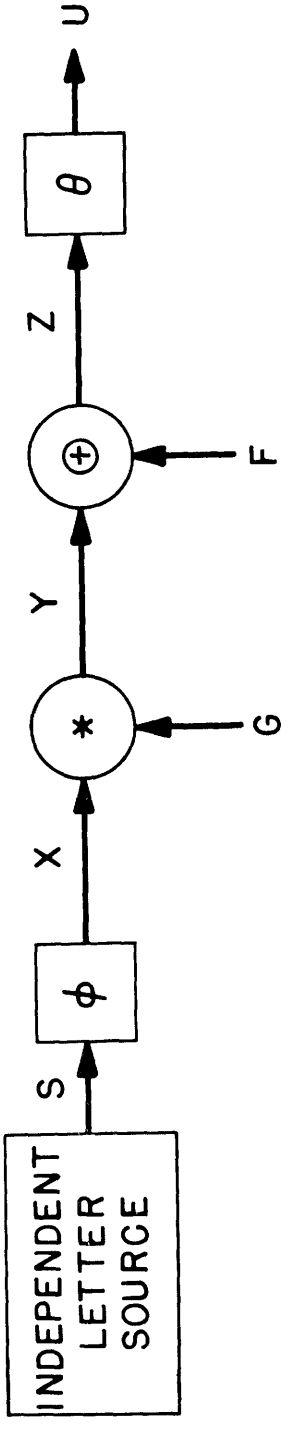


FIG. 4 SEQUENTIAL ENCODING

With a slight restriction on ϕ^*, condition 1 may be shown to hold if the Θ map operates directly on the Y sequence, if the elements of $\{G\}$ are equally likely. Thus the F sequence need not be introduced insofar as condition 1 is concerned. However, the F-ensemble is necessary in order to obtain condition 2. This can be seen by an example. Suppose $n = 2$, $D = 3$ and $m = 2$. Assume $x^1 = 1$, $x^2 = 2$. For the elements of $\{G\}$ equally likely, the probability measure on Y sequences is a function of the X sequence under consideration. For all possible X sequences equally likely, the probability measure on the Y sequences are as tabulated below:

Y	Pr[Y]
00	2/9
11	5/36
22	5/36
12	5/36
21	5/36
10	1/18
01	1/18
20	1/18
02	1/18

---

* The restriction is $\phi(s^k) \neq 0$ for $k = 1, \ldots, m$.

From the table, it can be seen that $\Pr [y_t = i] = 1/3$ for all t and i = 0, 1, 2. However, $\Pr [y_t = i | y_{t-1} = j] \neq \Pr [y_t = i]$ in general. For example $\Pr [y_t = 0 | y_{t-1} = 0] = 2/3 \neq 1/3$. Thus condition 2 is not satisfied.

From theorems 1, 2, and 3 we conclude that an encoding scheme selected at random from the ensemble of encoding schemes defined by equally likely elements of $\{ F \}$ and $\{ G \}$ will give rise to U sequences satisfying conditions 1, 2, and 3. There are no further restrictions on ϕ, Θ, or S. In order words, conditions 1, 2, and 3 will hold for a letter source with non-equally likely letters and with adjacent letters mutually dependent, i.e., for the most general discrete source. This research is primarily concerned with coding for the channel. Thus only the simplest of sources will be assumed: an independent source with equally likely letters. However, the results to be shown for number of computations for decoding and probability of error will hold for the most general discrete source, with a suitable adjustment in the rate R to reflect the source entropy.

CHAPTER III

SEQUENTIAL DECODING

A. Introduction

Sequential decoding, as is any effective decoding scheme, is closely related to maximum likelihood decoding. To explain this relationship, it is convenient to first describe maximum likelihood decoding for a block code.

B. Block Decoding

Suppose we construct a block code for use over a discrete memoryless channel. We will send a sequence of input symbols of length n and receive a sequence of output symbols of length n which we will call U and V respectively. Let $U = (u_1, \ldots u_n)$ and $V = (v_1, \ldots v_n)$ where subscript r corresponds to the r^{th} symbol in the sequence of length n. The channel is defined by the set of transition probabilities $q_i(j)$ from input symbol i to output symbol j. From the fact that the channel is memoryless, the transition probability from input sequence U to output sequence V is:

$$Q_U(V) = \prod_{k=1}^{n} q_{u_k}(v_k) \tag{3.1}$$

A code, c, is the mapping of the integers 1, 2, . . , M onto a subset of the U sequences. Let the map of integer ℓ corresponding to code c be U_ℓ^c. The decoding problem consists of deciding which integer ℓ caused a particular received output sequence V. The decision is most likely to be correct if that integer with the largest a posteriori probability is selected. In other words, we select that integer ℓ which maximizes the probability of ℓ conditional on V and c. By Baye's rule

$$\Pr[\ell \mid V, c] = \frac{\Pr[\ell \mid c]\ \Pr[V \mid \ell, c]}{\Pr[V \mid c]} \tag{3.2}$$

The event ℓ is independent of the code c. Also $\Pr[V \mid \ell, c] = \Pr[V \mid U_\ell^c]$ is only a function of the channel and is given by Eq. (3.1). Thus to achieve the maximum likelihood decision, it is sufficient to maximize the quantity $\Pr[\ell]\ Q_{U_\ell^c}(V)$ over ℓ.

In the special case where all input integers ℓ are equally likely, it is sufficient to maximize $Q_{U_\ell^c}(V)$. In the BSC, the U's and V's are sequences of zeroes and ones, and $Q_U(V)$ is monotonically related to the number of places in which the U and V sequences under consideration differ.

At this point, it is convenient to consider code c as a member of the set of codes $\{c\}$ constructed as follows. Suppose over the set $\{c\}$ we desire to use the individual input symbols with probabilities p_i. Then we may define a probability measure P_U on the set of input sequences $\{U\}$ by viewing the individual letters of U as independently selected with probability p_i. Thus:

$$P_U = \prod_{k=1}^{n} p_{u_k} \tag{3.3}$$

A particular code c will correspond to randomly selecting a U sequence to correspond to letter ℓ where the selection is done with probability P_U. This is done for all ℓ with sequences corresponding to ℓ and ℓ' selected independently for $\ell \neq \ell'$. All possible block codes make up the ensemble $\{c\}$.

The specification of p_i and $q_i(j)$ induce a probability measure r_j on output symbol j where

$$r_j = \sum_{i=1}^{a} p_i q_i(j) \tag{3.4}$$

Averaged over $\{ c \}$, the set of output sequences $\{ V \}$ will occur with a probability measure R_V given by

$$R_V = \prod_{k=1}^{n} r_{v_k} \tag{3.5}$$

It is convenient to define the mutual information function:

$$I(U; V) = \log \frac{Q_U(V)}{R_V} \tag{3.6}$$

Substituting Eqs. (3.1) and (3.4) into Eq. (3.6), we obtain

$$I(U; V) = \sum_{k=1}^{n} I_k \tag{3.6a}$$

where

$$I_k = \log \frac{q_{u_k}(v_k)}{r_{v_k}} \tag{3.7}$$

We note that there are ab values of I_k corresponding to all combinations of possible input symbols (a in number) and output symbols (b in number). These ab values of I_k need not all be different.

Shannon[2] obtained an upper bound to probability of error for block coding by constructing the ensemble $\{ c \}$ as described above. He then observed that over $\{ c \}$, if the received sequence V occurs when U is sent, the joint event (U, V) occurs with probability measure $Pr_1[U, V] = P_U Q_U(V)$. However, if the received sequence V occurs and a U is selected at random, the joint event (U, V) occurs with probability measure $Pr_2[U, V] = P_U R_V$.

Shannon then observed that a maximum likelihood decoding procedure is at least as good as a procedure where a threshold I_0 for the random variable I (U; V) is set, and an error is said to occur when either I (U; V) $\geqq I_0$ for V received and U = U_ℓ^c where ℓ is selected, or I (U, V) $\geqq I_0$ for V received and U = $U_{\ell'}^c$, where ℓ is selected and $\ell' \neq \ell$. In the former case (U; V) has the measure Pr_2. These simple forms permit a bounding of the probability of error which is minimized by an optimization of I_0.

C. Sequential Decoding (After Wozencraft)

We now introduce the concept of sequential decoding for the channel. Recognizing that the S sequence may not be synchronous with the U and V sequences (see discussion of R < log m, Chapter II), we will confine our discussion to the sequential decoding of the X sequence. Since the ϕ map is one-one, a decision on an X sequence is equivalent to a decision on an S sequence.

Sequential decoding implies that we decode one x symbol at a time. The symbol x_t is to be decoded immediately following the decoding of the symbol x_{t-1}. Thus the receiver has available the decoded set $(\ldots x_{-1}, x_0)$ when it is about to decode x_1. We shall assume that these symbols have been decoded without error. This assumption, although crucial to the decoding procedure, is not as restrictive as it may appear. It will be discussed in Chapter VI. In any case, the assumption permits the generation at the receiver of a potential transmitted sequence $(u_1, u_2, \ldots)$ generated by a potential X sequence $(x_1, \ldots x_n)$.

The symbol x_t enters into the determination of n input symbols $u_t, u_{t+1}, \ldots u_{t+n-1}$. But it is related to all output symbols v_k, $k \geqq t$, as will be shown below. We desire the probability

$\Pr [x_1 | \ldots, x_{-1}, x_0, v_1, v_2, \ldots]$. If position 1 is a redundancy position introduced because $R < \log m$, $x_1 \equiv 0$. Thus we only consider the non-trivial case where x_1 is the ϕ-map of some s^k, $k = 1, \ldots, m$. The desired probability equals $\sum_{\overline{X}} \Pr [x_1, x_2, \ldots | \ldots x_{-1}, x_0, v_1, v_2 \ldots]$, where the summation is over all sequences $\overline{X}$ of the form $(x_2, x_3, \ldots)$. By Baye's rule, we may express each of the probabilities in the sum as

$$\Pr [x_1, x_2 \ldots | \ldots, x_{-1}, x_0, v_1, v_2, \ldots] =$$

$$\frac{\Pr [v_1, v_2 \ldots | \ldots x_{-1}, x_0, x_1 \ldots] \Pr [x_1, x_2, \ldots]}{\Pr [v_1, v_2, \ldots | \ldots x_{-1}, x_0]}$$

In this expression use has been made of the fact that we are dealing with an independent letter source; thus the x_t's are independent.

Thus the maximum likelihood decision on x_1 conditional on a particular received sequence $(v_1, v_2, \ldots)$ and certainty of the previous X sequence $(\ldots x_{-1}, x_0)$ is that x_1 for which the sum

$$\sum_{\overline{X}} \Pr [v_1, v_2, \ldots | \ldots x_{-1}, x_0, x_1 \ldots] \Pr [x_1, x_2, \ldots]$$

is maximized. With the outputs of the message source equally likely, it is sufficient to maximize the sum:

$$\sum_{\overline{X}} \Pr [v_1, v_2, \ldots | \ldots x_{-1}, x_0, x_1, \ldots]$$

Since $\Pr [v_t | \ldots, x_t] = \Pr [v_t | x_{t-n+1}, \ldots, x_t]$, we may rewrite the last expression as:

$$\sum_{\overline{X}} \Pr [v_1, v_2, \ldots | x_{-n+2}, x_{-n+3}, \ldots x_{-1}, x_0, x_1, \ldots]$$

This is an awkward expression inasmuch as it is the sum of probabilities of semi-infinite sequences. It is intuitively clearer to interpret this sum as the limit of an expression which deals with finite sequences:

$$\lim_{w \to \infty} \sum_{\{x_2, \ldots, x_w\}} \Pr[v_1, \ldots v_w | x_{-n+2}, \ldots x_w]$$

For w finite, the expression is the finite sum of discrete probabilities. As $w \to \infty$, the expression becomes the infinite sum of probabilities each of which approaches zero.

Thus the exact maximum likelihood determination of x_1 requires an infinite delay. This is a consequence of the convolutional encoding. However, our approach to sequential decoding will not be maximum likelihood. It shall deviate in two major respects:

1. We shall make our decision at a fixed w = n. This is done because we have demonstrated a sequential encoding procedure with the properties implied by theorems 1 - 3 of Chapter II for U sequences of length less than or equal to n. We shall interpret n as the decoding delay. We note that for block codes of length n, the decoding delay is also n; thus this parameter will serve as a link for comparison purposes.

2. Instead of selecting that x_1 for which the sum

$$\sum_{\{x_2, \ldots x_n\}} \Pr[v_1, \ldots v_n | x_{-n+2}, \ldots, x_n]$$

is maximized, we will select the x_1 belonging to any sequence $(x_1, \ldots, x_n)$ which meets a criterion to be described below. This is done so that a favorable decoding complexity will result.

At this point we introduce some additional notation. Since we are concerned with the decoding of x_1, we restrict our attention to those X consistent with the previously decoded places (x_t, $t \leqq 0$). An X sequence ending with position w we will call X_w. With our encoding scheme, X_w induces transmitted sequence $U_w = (u_1, \ldots, u_w)$ and sequence $V_w = (v_1, \ldots, v_w)$ is received. (Fixed G and F sequences are assumed.) If x_w is not a redundancy place, each element of $\{ X_{w-1} \}$ and $\{ U_{w-1} \}$ can be viewed as generating m elements of $\{ X_w \}$ and $\{ U_n \}$ respectively.

Thus the sets $\{ X_n \}$ and $\{U_n\}$ may be viewed topologically as in Figure 5 . The set elements are all the directed paths from the input node to the output nodes of the tree. (There are no closed paths.) From each non-trivial intermediate node there emerges m directed links, one to each of m nodes. All paths pass through n + 1 nodes. All paths starting at the input and passing through w + 1 nodes are the elements of $\{X_w\}$ or $\{U_w\}$ depending upon whether it is the $\{X_n\}$ tree or $\{ U_n \}$ tree we are discussing.

As mentioned above, we will test sequences of the set $\{ X_n \}$ which are one to one related to sequences of $\{U_n\}$. The test we perform is to compute a quantity which varies monotonically with the quantity $Q_{U_n}(V_n)$. We may divide by the constant R_{V_n} as defined in Eqs. (3.4) and (3.5) and take the logarithm. We call this quantity $d_n(X_n)$. Substituting in Eq. (3.1), (3.4), (3.5), (3.6), and (3.7) we obtain

$$d_n(X_n) = \sum_{k=1}^{n} I_k \tag{3.8}$$

---

* The choice of d_n as defined rather than some other monotonic function of $Q_{U_n}(V_n)$ is motivated by the tractibility it offers in the bounding of the number of decoding computations and the probability of error. See Appendix B.

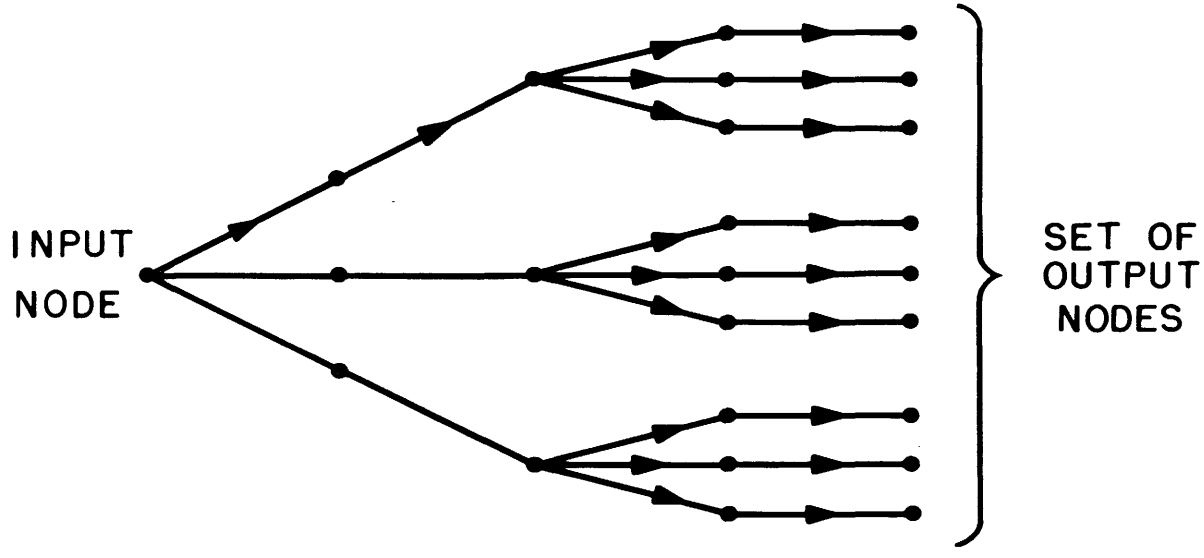


FIG. 5 TOPOLOGICAL REPRESENTATION OF SETS $|X_n|$ OR $|U_n|$ DRAWN FOR $m = 3, n = 4$, $R = 1/2$ LOG 3 (every other position in the X sequence is identically equal to zero).

At this point it is conveneient to define the partial sums, $d_w(X_w)$

$$d_w(X_w) = \sum_{k=1}^{w} I_k \tag{3.9}$$

For the properly chosen elements of $\{X_w\}$ and $\{ X_{w-1}\}$,

$$d_w = d_{w-1} + I_w \tag{3.10}$$

Our test will be that we will search for an X_n for which $d_n \geqq D_n$, where D_n is some preassigned value. Our decision will be the first element of the first encountered sequence X_n that meets the criterion $d_n(X_n) \geqq D_n$.

In computing d_n for a particular X_n we will successively form the truncated sums d_w for $1 \leqq w \leqq n$. To reduce the number of X_n sequences that must be considered, we have a set of preassigned values D_w, and we eliminate from further consideration any X_n which has a prefix in $\{ X_w\}$ for which $d_w(X_w) \leqq D_w$. We expect the incorrect sequences to be eliminated at lengths w much less than n; thus we expect to be able to make a decision on x_1 without having to search through a number of sequences which grows exponentially with n.

There exists the possibility that no sequence X_n meets criterion D_n. In this case no decision on x_1 is possible by the procedure as stated so far. However, instead of a single D_n, we will have a set of criteria $D_n^{(j)}$, $j = 1, 2, \ldots$, where $D_n^{(j+1)} < D_n^{(j)}$. We first try to find a sequence X_n which satisfies $D_n^{(1)}$. If all of the set $\{ X_n\}$ fail to meet $D_n^{(1)}$ we try $D_n^{(2)}$. Similarly if all of $\{ X_n\}$ fail to meet $D_n^{(j)}$, we try $D_n^{(j+1)}$ until some X_n satisfies some criterion $D_n^{(j)}$.

Associated with each $D_n^{(j)}$ is a set of values $D_w^{(j)}$ for $1 \leqq w \leqq n$ which are compared against the truncated sums d_w. When testing against criterion $D_n^{(j)}$, we do not consider sequences X_n which have prefixes X_w for which $d_w(X_w) \leqq D_w^{(j)}$.

The choice of the magnitudes of $D_w^{(j)}$ and the average amount of computation required to make a decision on x_1 are discussed in Chapter IV.

This decoding scheme is conceptually identical to that described by Wozencraft[4] with the random variable d_w replacing Wozencraft's "number of places in which a potential transmitted sequence and the received sequence, both of length w, differ".

D. Sequential Decoding (After Gallager)

We have not been able to meaningfully bound the average number of decoding computations for the decoding procedure described in section C. However, a modification of the decoding procedure adapted from one suggested by Gallager* for the BSC permits such bounding.

Suppose that for each set $\{ D_1^{(j)}, \ldots, D_n^{(j)} \}$ and X_w being tested, besides comparing $d_w(X_w)$ to $D_w^{(j)}$, we compare each of the quantities $d_w - d_1$, $d_w - d_2, \ldots, d_w - d_{w-1}$ to $D_{w-1}^{(j)}$, $D_{w-2}^{(j)}$, $\ldots$, $D_1^{(j)}$ respectively. We eliminate from consideration with respect to index j the sequence X_w and all sequences branching from it, unless all the comparisons satisfy their respective criteria.

If no sequence X_n is found which satisfies this modified test, we repeat the process for the set $\{D_1^{(j+1)}, \ldots, D_n^{(j+1)}\}$.

---

* Gallager, R. G., Unpublished work.

CHAPTER IV

NUMBER OF DECODING COMPUTATIONS AND THE COMPUTATION CUTOFF RATE

A. Introduction

We shall use the encoding and decoding procedure described in Chapters II and III respectively. We assume a criterion K, i.e., the correct sequence of length w will be rejected if the probability that it caused the received sequence of length w is less than or equal to e^{-K}* when averaged over the ensemble implied by the equally likely use of the elements of $\{F\}$ and $\{G\}$. We desire to find conditions under which the average number of computations required to make a decision does not increase exponentially with n.

B. Definition of D_w

Over the ensemble as defined, conditions 1-3 hold. Thus, when U_w is transmitted and V_w is received, the probability measure on the pair (U_w, V_w) is $Pr_1[U_w, V_w] = P_{U_w} Q_{U_w}(V_w)$. On the other hand, when a sequence U_w generated by an S starting with the incorrect s_1 is paired with a received V_w, the probability measure on the pair (U_w, V_w) is $Pr_2[U_w, V_w] = P_{U_w} R_{V_w}$.

---

* We will derive the cutoff values $D_w^{(j)}$ described in Chapter III in terms of the parameter K_j introduced here. There will be a sequence of K values: K_1, K_2, ... used in the analysis starting in section E of this chapter. To simplify the notation we dispense with subscripts on K and the corresponding superscripts on D_w until then.

We want to select a value of D_w such that

$$Pr_1[d_w \leqq D_w] \leqq e^{-K} \tag{4.1}$$

The maximum value of D_w which satisfies Eq. (4.1), $(D_w)_{max}$, can be found if one wants to work hard enough with a known probability measure Pr_1. However, we week a general result for relating D_w to w and K, with D_w closely approximating $(D_w)_{max}$. To achieve this we make use of the Chernoff bound:*

$$Pr_1[d_w \leqq w\mu_1'(\sigma)] \leqq e^{w[\mu_1(\sigma) - \sigma\mu_1'(\sigma)]},\ \sigma \leqq 0 \tag{4.2}$$

We set $E_1(\sigma) = \sigma\mu_1'(\sigma) - \mu_1(\sigma)$ equal to $\frac{K}{w}$, thereby determine a value of σ which we call σ_w, and set

$$D_w = w\mu_1'(\sigma_w) \tag{4.3}$$

This may be interpreted graphically as in Figure 6 . From the figure it is clear that K may be so large or w so small that the equation $E_1(\sigma) = \frac{K}{w}$ has no solution. Physically this means that no matter what sequence U_w is transmitted over the channel, the corresponding received sequence V_w will be such that $Pr_1[U_w, V_w] > e^{-K}$. Clearly, there is in general some maximum value of w which we call n_o which does not permit solution of the equation $E_1(\sigma) = \frac{K}{w}$. For $w \leqq n_o$, we can set $\frac{D_w}{w} = I_{min} - \epsilon$ where ϵ is an arbitrary positive number: $0 < \epsilon < \infty$.

---

* The various Chernoff bounds used in this research, together with notation, are described in Appendix B.

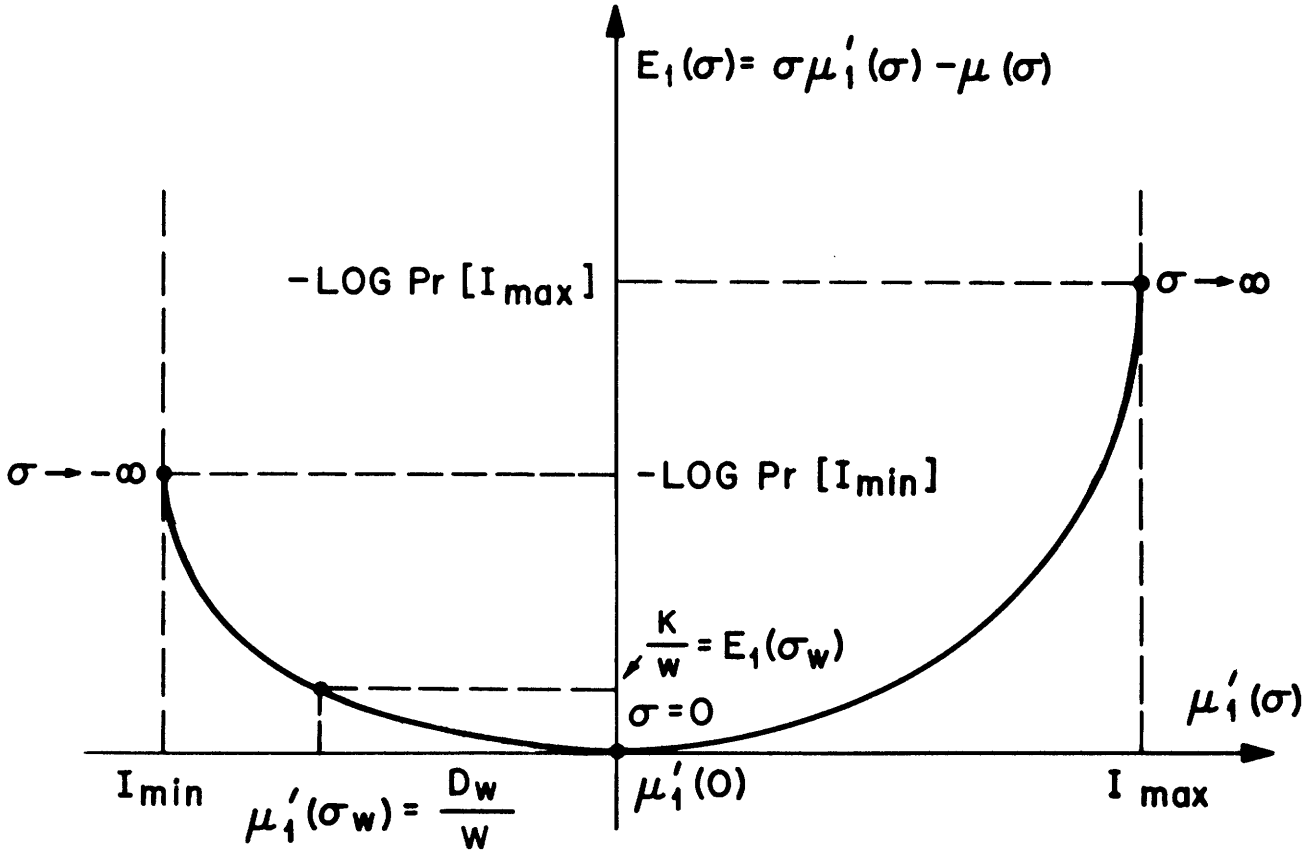


FIG. 6 PLOT OF $E_1(\sigma)$ vs. $\mu_1'(\sigma)$ AND DETERMINATION OF σ_w

To recapitulate: we define n_o such that

$$\frac{K}{n_o+1} \leqq -\log \Pr [\, I_{min}] < \frac{K}{n_o}$$

Then we define

$$D_w = \begin{cases} w[\, I_{min} - \epsilon\,], & 1 \leqq w \leqq n_o \\ w\mu_1'(\sigma_w), & n_o < w \leqq n \end{cases} \tag{4.4}$$

C. The Correct and Incorrect Subsets

Let $X_w = (x_1, \ldots x_w)$ and suppose x_1' is the correct x_1. Then we define $\{ X_w^c\} = \{ X_w \mid x_1 = x_1'\}^*$ and $\{X_w^i\} = \{ X_w | x_1 \neq x_1'\}$ as the correct and incorrect subsets of length w. Note that for a fixed F and G, each element of $\{ X_w\}$ uniquely specifies an element of $\{U_w\}$. We call $\{ U_w^c\}$ and $\{ U_w^i\}$ the maps of $\{ X_w^c\}$ and $\{ X_w^i\}$ respectively. With the decoding scheme described in Chapter III, section C, we shall be able to meaningfully bound the average number of computations required for processing $\{ X_w^i\}$ with respect to criterion K. However, we shall not be able to similarly bound the average number of computations required for processing $\{ X_w^c\}$. This is so because by condition 3, the elements of $\{ U_w^i\}$ are independent of V_w, but clearly, the elements of $\{ U_w^c\}$ are statistically related to V_w.

With the decoding scheme as modified in Chapter III, section D, we shall be able to meaningfully bound the average number of computations required for processing both $\{ X_w^i\}$ and $\{X_w^c\}$. This is so because of the following considerations. Suppose $X_n = (x_1, \ldots x_n)$ and $X_n' = (x_1', \ldots x_n')$ are elements of $\{ X_n^c\}$. Suppose the first place in which they differ is the r^{th}, i.e., $x_w = x_w'$ for $1 \leqq w < r$, $x_r \neq x_r'$. If X_n maps into $U_n = (u_1, \ldots u_n)$ and

* We use the following notation here and elsewhere in this research: If $\{ A\}$ is a set of elements, $\{ A|B\}$ is the subset of $\{A\}$ the elements of which satisfy condition B.

and X'_n maps into $(u'_1, \ldots u'_n)$, then by condition 3, over the ensemble of $\{F\}$ and $\{G\}$, the sequence $(u_r, \ldots u_n)$ is independent of the sequence $(u'_r, \ldots u'_n)$. This characteristic is used in section F of this chapter to bound the average number of decoding computations for channels symmetric at their output.

D. Number of Computations to Process $\{X_n^i\}$ for Fixed K - Decoding Procedure of Chapter III, Section C.

With D_w defined as in section B of this chapter, we proceed to bound the average number of computations required to process $\{X_n^i\}$. We define "a computation" as the generation of the $w^{\underline{th}}$ symbol of an X sequence being tested, formation of d_w from d_{w-1} and I_w, comparison of d_w against D_w and deciding whether to continue the test. This definition is applicable for $1 \leqq w \leqq n$.

Suppose there are $M_K(w)$ sequences of length w in the incorrect subset that are probable according to criterion K. If $N_K(w+1)$ is the number of computations necessary to check all previously admissible sequences at the $(w+1)^{\underline{st}}$ level.

$$N_K(w+1) \leqq \Delta(w)\, M_K(w), \quad w \geqq 1$$

$$N_K(1) = m - 1$$

where $\Delta(w) = 1$ or m depending on whether $\{X_w\}$ generates 1 or m elements of $\{X_{w+1}\}$ respectively. (See Figure 5 .)

The inequality holds rather than equality since some of the $M_K(w)$ admissible sequences may have been previously rejected at shorter lengths.

The number of computations required to search the entire incorrect subset is N_K.

$$N_K = \sum_{w=0}^{n-1} N_K(w+1) \leqq \sum_{w=1}^{n} \Delta(w) M_K(w)$$

We now desire to average N_K over the ensemble implied by the equally likely use of the elements of $\{F\}$ and $\{G\}$. Let this average be $\overline{N_K}$. Then,

$$\overline{N_K} \leqq \sum_{w=1}^{n} \Delta(w) \overline{M_K(w)} \tag{4.5}$$

where $\overline{M_K(w)}$ is the mean of $M_K(w)$ averaged over the ensemble. We obtain $\overline{M_K(w)}$ by the following argument: Suppose the elements of $\{U_w^i\}$ are indexed by their position in the tree of Figure 5. Consider a fixed position, the k^{th}. From condition 2 we know that over the ensemble of encoding schemes (equally likely use of the elements of $\{F\}$ and $\{G\}$) the probability that the k^{th} position becomes U_w is P_{U_w}. Further, from condition 3 we know that over the ensemble, the element of $\{U_w^i\}$ in the k^{th} position is independent of the actual U_w sequence transmitted. Since the V_w sequence is statistically related only to the transmitted U_w sequence, the element of $\{U_w^i\}$ under consideration is independent of the V_w received. Condition 2 implies that over the ensemble, the probability of receiving a particular V_w sequence is R_{V_w}. Thus the probability of the pair (U_w, V_w), where U_w is in the k^{th} position of the set $\{U_w^i\}$, is $Pr_2[d_w \geqq D_w]$. The same argument applies to all portions of $\{U_w^i\}$ of which there are $|X_w^i|$. Thus

$$M_K(w) = |X_w^i| \; Pr_2 \; [d_w \geqq D_w] \tag{4.6}$$

If $R = \log m$, $|X_w^i| = (m-1)\, e^{(w-1)R} = \frac{m-1}{m}\, e^{wR}$. If $R < \log m$, we have described in Chapter II how the encoding is to proceed. We set

$\frac{R}{\log m} = \frac{a}{\beta}$, a rational fraction less than one, where β divides n. Since we are in the incorrect subset, the first node is a branch point. Thus if β divides w, $|X_w^i| = \frac{m-1}{m} e^{wR}$. If β does not divide w, let the remainder be ρ_w, $1 \leqq \rho_w < \beta$. Then $|X_w^i| \leqq \frac{m-1}{m} e^{aR} e^{(w-\rho_w)R} = \frac{m-1}{m} e^{(a-\rho_w)R} e^{wR}$.

We have already reduced all situations of $R \neq \log m$ to the case $R < \log m$. Thus we have succeeded in evaluating $|X_w^i|$:

$$|X_w^i| = \delta(w) e^{wR} \tag{4.7}$$

where

$$\delta(w) = \begin{cases} \frac{m-1}{m} & \text{if } \beta \text{ divides } w \\ \frac{m-1}{m} e^{(a-\rho_w)R} & \text{if } \beta \text{ does not divide } w. \end{cases}$$

We will use the Chernoff bound* to overestimate $Pr_2 [d_w \geqq D_w]$

$$Pr_2 [d_w \geqq w\mu_2'(\lambda_w)] \leqq e^{w[\mu_2(\lambda_w) - \lambda_w \mu_2'(\lambda_w)]} \tag{4.8}$$

$$\text{for } \lambda_w > 0.$$

However, $\mu_2(\sigma + 1) = \mu_1(\sigma)$.* Since $D_w = w\mu_1'(\sigma_w)$ for $\sigma_w \geqq 0$, over the range $-1 < \sigma_w \leqq 0$, we may set $\mu_2(\lambda_w) = \mu_1(\sigma_w)$ by setting $\lambda_w = \sigma_w + 1$.

Over this range we obtain from Equation (4.8):

$$Pr_2 [d_w \geqq D_w] \leqq e^{w[\mu_1(\sigma_w) - (\sigma_w + 1)\mu_1'(\sigma_w)]}$$

$$= e^{-K_e - w\mu_1'(\sigma_w)} \tag{4.8a}$$

---

* See Appendix B.

If the channel is not degenerate*, a sufficiently accurate bound to $Pr_2[d_w \geq D_w]$, for $D_w \leq w\mu_1'(-1)$, is unity.

For degenerate channels

$$Pr_2[d_w > -\infty] = \left[\sum_{\{u, v | q_u(v) \neq 0\}} p_u r_v\right]^w \cdot = \exp[w \log \Sigma p_u r_v]$$

But, we have shown in Appendix B that

$$\log \sum_{\{u, v | q_u(v) \neq 0\}} p_u r_v = \mu_1(-1)$$

Thus we have the general result that for $D_w \leq w\mu_1'(-1)$

$$Pr_2[d_w \geq D_w] \leq \exp[w\mu_1(-1)]$$

From Figure 6 it is clear that we are never concerned with values of σ_w greater than or equal to zero. Thus we may write

$$Pr_2[d_w \geq D_w] \leq \begin{cases} e^{-K} e^{-w\mu_1'(\sigma_w)}, & -1 < \sigma_w < 0 \\ e^{w\mu_1(-1)}, & \sigma_w \leq -1 \end{cases} \qquad (4.9)$$

Substituting Eqs. (4.6), (4.7), (4.8), (4.9) into Eq. (4.5) we obtain

$$\overline{N_K} < \sum_{\{w | \sigma_w \leq -1\}} \Delta(w)\, \delta(w)\, e^{w[R+\mu_1(-1)} + e^{-K} \sum_{\{w | \sigma_w > -1\}} \Delta(w)\delta(w) e^{wR} e^{-w\mu_1'(\sigma_w)} \qquad (4.10)$$

* A degenerate channel is defined by the condition $q_i(j) = 0$ for at least one pair (i, j). A non-degenerate channel has $q_i(j) \neq 0$ for all (i, j).

We can overbound the second summation by underbounding $\mu_1'(\sigma_w)$. We note that the plot of $E_1(\sigma)$ vs. $\mu_1'(\sigma)$ is convex; its slope is σ and its second derivative is $\frac{1}{\mu_1''(\sigma)}$ which is non-negative since $\mu_1''(\sigma)$ is a variance.* Thus the plot of $E_1(\sigma)$ vs. $\mu_1'(\sigma)$ may be underestimated in a piece-wise linear manner by a set of tangents. We call this approximation $\hat{E}$. (See Figure 7). We construct tangents at the points corresponding to $\sigma = -1$ and $\sigma = \sigma_1$ with $-1 < \sigma_1 < 0$. The value of σ_1 remains to be specified. In general the straight line tangent to the E vs. μ' curve at the point $\sigma = \sigma_0$ has the equation

$$\frac{\hat{E} - E(\sigma_0)}{\mu' - \mu'(\sigma_0)} = \sigma_0, \text{ or } \hat{E} = \sigma_0\mu' - \mu(\sigma_0)$$

The tangents of slope σ_0 and σ_1 intersect when $\sigma_0\mu' - \mu(\sigma_0) = \sigma_1\mu' - \mu(\sigma_1)$ or $\mu' = \frac{\mu(\sigma_0) - \mu(\sigma_1)}{\sigma_0 - \sigma_1}$

The corresponding ordinate is

$$\hat{E} = \frac{\sigma_1\mu(\sigma_0) - \sigma_0\mu(\sigma_1)}{\sigma_0 - \sigma_1}$$

Letting $\sigma_0 = -1$, we have the intersection of the two tangents at the point:

$$\mu_1' = \frac{\mu_1(\sigma_1) - \mu_1(-1)}{1 + \sigma_1} \qquad (4.11)$$

$$\hat{E}_1 = -\frac{\mu_1(\sigma_1) + \sigma_1\mu_1(-1)}{1 + \sigma_1}$$

---

* See Appendix B.

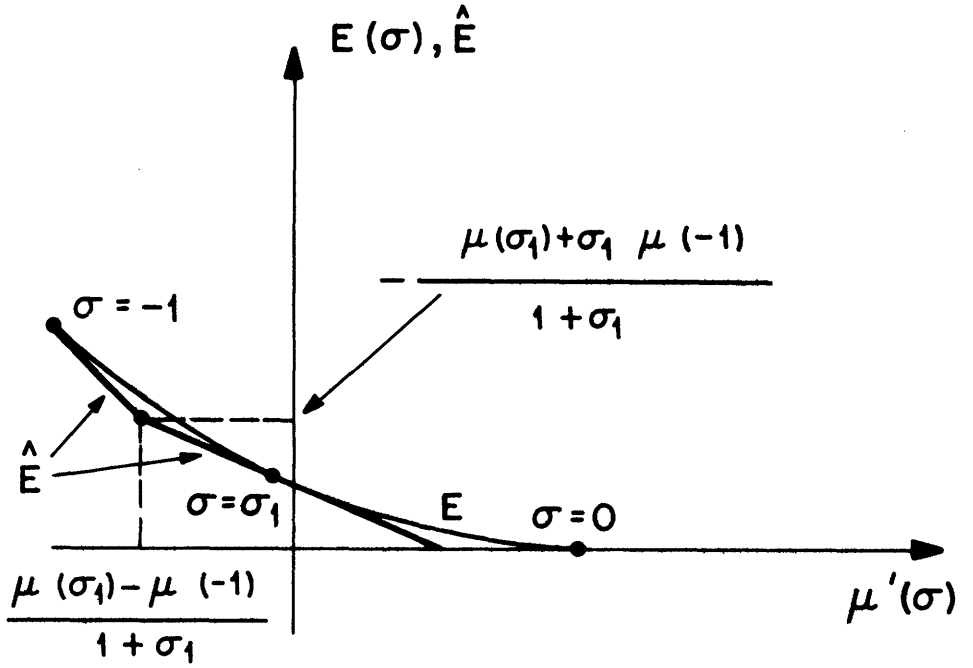


FIG. 7 PIECEWISE LINEAR APPROXIMATION TO $E(\sigma)$ vs. $\mu'(\sigma)$ CHARACTERISTIC

Note in the special case of a non-degenerate channel, $\mu_2(0) = \mu_1(-1) = 0$, in which case we have the intersection of the tangents at the point

$$\mu_1' = \frac{\mu_1(\sigma_1)}{1+\sigma_1}\ ,\ \hat{E}_1 = -\frac{\mu_1(\sigma_1)}{1+\sigma_1} \tag{4.11a}$$

If we let $\hat{E}_1 = \frac{K}{w}$, it is clear that the corresponding abscissa is less than the true value of $\mu_1'(\sigma_w)$ defined by $E_1 = \frac{K}{w}$. Define n_1 such that

$$\frac{K}{n_1+1} \leqq -\frac{\mu_1(\sigma_1) + \sigma_1\mu_1(-1)}{1+\sigma_1} < \frac{K}{n_1} \tag{4.12}$$

Thus

$$\mu_1'(\sigma_w) \geqq \begin{cases} -\left[\frac{K}{w} + \mu_1(-1)\right], & n_0 < w \leqq n_1 \\ \frac{1}{\sigma_1}\left[\frac{K}{w} + \mu_1(\sigma_1)\right], & n_1 < w \leqq n \end{cases} \tag{4.13}$$

Substituting these results in Eq. (4.10) we obtain:

$$\overline{N_K} \leqq \sum_{w=1}^{n_1} \Delta(w)\,\delta(w)\,e^{w[R+\mu_1(-1)]}$$

$$+ e^{-K\left(1+\frac{1}{\sigma_1}\right)} \sum_{w=n_1+1}^{n} \Delta(w)\,\delta(w)\,e^{w\left[R-\frac{\mu_1(\sigma_1)}{\sigma_1}\right]} \tag{4.14}$$

In a particular situation, we may be able to obtain a tight bound to $\Delta(w)\delta(w)$ applicable for all w. However, $\Delta(w) \leqq m$; $\delta(w) \leqq \frac{m-1}{m} e^{(\alpha-1)R}$. Thus we may write:

$$\overline{N_K} \leqq (m-1)\, e^{(a-1)R} \left\{ \sum_{w=1}^{n_1} e^{w[R+\mu_1(-1)]} + e^{-K\left(1+\frac{1}{\sigma_1}\right)} \sum_{w=n_1+1}^{n} e^{w\left[R-\frac{\mu_1(\sigma_1)}{\sigma_1}\right]} \right\} \quad (4.15)$$

We proceed to upper bound the summations in the brackets:

For

$$R + \mu_1(-1) < 0$$

$$\sum_{w=1}^{n_1} e^{w[R+\mu_1(-1)]} < e^{R+\mu_1(-1)} \sum_{w=0}^{\infty} e^{w[R+\mu_1(-1)]}$$

$$= \frac{e^{R+\mu_1(-1)}}{1-e^{R+\mu_1(-1)}} < \frac{1}{1-e^{R+\mu_1(-1)}} \quad (4.16a)$$

For

$$R + \mu_1(-1) > 0$$

$$\sum_{w=1}^{n_1} e^{w[R+\mu_1(-1)]} < e^{n_1[R+\mu_1(-1)]} \sum_{w=0}^{\infty} e^{-w[R+\mu_1(-1)]}$$

$$= \frac{e^{n_1[R+\mu_1(-1)]}}{1-e^{-[R+\mu_1(-1)]}}$$

But, from Eq. (4.12)

$$n_1 < -\frac{K(1+\sigma_1)}{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)}$$

Thus

$$\sum_{w=1}^{n_1} e^{w[R+\mu_1(-1)} < \frac{\exp - \frac{K(1+\sigma_1)[R+\mu_1(-1)]}{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)}}{1 - \exp(-[R+\mu_1(-1)])} \tag{4.16b}$$

For

$$R < \frac{\mu_1(\sigma_1)}{\sigma_1}$$

$$\sum_{w=n_1+1}^{n} e^{w[R - \frac{\mu_1(\sigma_1)}{\sigma_1}]} < e^{(n_1+1)[R - \frac{\mu(\sigma_1)}{\sigma_1}]} \sum_{w=0}^{\infty} e^{w[R - \frac{\mu_1(\sigma_1)}{\sigma_1}]}$$

$$= \frac{e^{(n_1+1)[R - \frac{n_1(\sigma_1)}{\sigma_1}]}}{1 - e^{[R - \frac{\mu_1(\sigma_1)}{\sigma_1}]}}$$

But, from Eq. (4.12)

$$n_1 + 1 \geqq - \frac{K(1+\sigma_1)}{\mu_1(\sigma_1) + \sigma_1\mu_1(-1)}$$

Thus

$$\sum_{w=n_1+1}^{n} e^{w[R - \frac{\mu_1(\sigma_1)}{\sigma_1}]} < \frac{\exp\left[- \frac{K(1+\sigma_1)[R - \frac{\mu_1(\sigma_1)}{\sigma_1}]}{\mu_1(\sigma_1) + \sigma_1\mu_1(-1)} \right]}{1 - \exp[R - \frac{\mu_1(\sigma_1)}{\sigma_1}]} \tag{4.17a}$$

Further,

$$\exp\left[-K\left(1+\frac{1}{\sigma_1}\right)\right]\exp\left[-\frac{K(1+\sigma_1)\left[R-\frac{\mu_1(\sigma_1)}{\sigma_1}\right]}{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)}\right]$$

$$= \exp\left[-\frac{K(1+\sigma_1)\left[R+\mu_1(-1)\right]}{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)}\right] \qquad (4.17b)$$

It is to be observed that the expression

$$-\mu_1(\sigma_1)-\sigma_1\mu_1(-1) = -\sigma_1\left[\frac{\mu_1(\sigma_1)}{\sigma_1}-\frac{\mu_1(-1)}{-1}\right] > 0.$$

This is so because the function $\frac{\mu_1(\sigma)}{\sigma}$ is non-negative and monotonically increasing with σ for $-1 \leqq \sigma < 0$.

Substituting Eqs. (4.16) and (4.17) into Eq. (4.15) we obtain:

$$\overline{N_K} < \begin{cases} C_3, & R < \mu_1(-1) \\ C_3 \exp\left[-\frac{K(1+\sigma_1)\left[R+\mu_1(-1)\right]}{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)}\right], & -\mu_1(-1) < R < \frac{\mu_1(\sigma_1)}{\sigma_1} \end{cases} \qquad (4.18)$$

where

$$C_3 = (m-1)\, e^{(\alpha-1)R}\left\{\frac{1}{1-\exp-\left|R+\mu_1(-1)\right|}+\frac{1}{1-\exp\left[R-\frac{\mu_1(\sigma_1)}{\sigma_1}\right]}\right\} \qquad (4.19)$$

In the special case of the non-degenerate channel we have

$$\overline{N_K} < C_3 \exp\left[- \frac{K (1 + \sigma_1) R}{\mu_1 (\sigma_1)} \right] \tag{4.18a}$$

for $R < \frac{\mu_1 (\sigma_1)}{\sigma_1}$, where

$$C_3 = (m-1)\, e^{(\alpha - 1)R} \left\{ \frac{1}{1-\exp[-R]} + \frac{1}{1-\exp[R - \frac{\mu_1(\sigma_1)}{\sigma_1}]} \right\} \tag{4.19a}$$

E. The Computation Cutoff Rate for Symmetric Channels

In section C, we bounded $\overline{N_{K_j}}$, the average number of computations necessary to process the incorrect subset with respect to a fixed criterion K_j. It is important to note that this bound is not conditioned on whether or not any element of the correct subset satisfies criterion K_j.

We now approach the problem of bounding $\overline{N}$, the average number of computations that must be made on the incorrect subset before the decoding scheme as described in Chapter III, section C makes a decision on x_1. We have pointed out that there exists a sequence of criteria: K_1, K_2, ... It will be convenient to let

$$K_{j+1} = K_j + \Delta K \tag{4.20}$$

There are two cases to consider. We shall dispose of the less interesting one first.

1. $R \leq -\mu_1(-1)$

In this case we modify the decoding procedure as stated by agreeing to terminate and accept an error if no sequence is found which satisfies some maximum criterion $K_{j_{max}}$. In this case the average number of computations can never exceed

$$\overline{N} \leqq j_{max} C_3 \tag{4.21}$$

The selection of j_{max} and its relationship to the probability of error is discussed in Chapter V, section B. For the choice described there, Eq. (4.21) becomes

$$N < \left[2 + \frac{E_1(\sigma)}{\Delta K}\right] C_3, \; R = \mu_1'(\sigma); \; \sigma < 0 \tag{4.21a}$$

2. $-\mu_1(-1) < R < -2\mu_1(-\frac{1}{2})$

We need never eliminate $\{X_n^i\}$ with respect to K_j unless the correct sequence fails to meet K_{j-1}. Let A_j be the event that the correct sequence fails to meet K_{j-1}. Over the ensemble, event A_j has probability $\Pr[A_j]$ where

$$\Pr[A_j] \leqq \begin{cases} 1, & j = 1 \\ n e^{-K_{j-1}} = n e^{\Delta K} e^{-K_j}, & j > 1 \end{cases} \tag{4.22}$$

We have made use of Eq. (4.20) in writing Eq. (4.22) for $j > 1$.

We note that the decoding of x_1 will terminate no later than after $\{X_n^i\}$ is eliminated with respect to K_j if <u>any</u> element of the <u>correct</u> subset satisfies K_j. Thus

$$\overline{N} \leqq \sum_{j=1}^{\infty} \Pr[A_j] \, \overline{N_{K_j} | A_j} \tag{4.23}$$

where $\overline{N_{K_j}|A_j}$ is the average number of computations required to eliminate $\{X_n^i\}$ with respect to K_j conditional on the event A_j.

We have not been able to find a useful upper bound to $\overline{N_{K_j}|A_j}$ for the general channel. However, we will be able to bound $\overline{N}$ if N_{K_j} is independent of the event A_j. This will be the case in channels which are symmetric at their output* if their input probabilities are equally likely. This may be appreciated from the following considerations. We have been able to bound $\overline{N_K}$ because we know the probability measure $\Pr[U_w, V_w | U_w \epsilon \{U_w^i\}] = P_{U_w} R_{V_w}$. This measure has the properties that U_w and V_w are independent and appear with their ensemble probabilities. However, if we impose a condition A_j on d_n, i.e., on the transmitted U_n and received V_n, then V_n will no longer in general appear with the ensemble probability R_{V_n}. Thus,

$$\Pr[U_w, V_w | U_w \epsilon \{U_w^i\}, A_j] = P_{U_w} \Pr[V_w | A_j]$$

We have a tractible solution whenever $\Pr[V_w|A_j] = R_{V_w}$.

The event A_j is a condition on $d_n = \sum_{k=1}^{n} I_k$. This is certainly weaker than a condition on the individual $I_k = \log \frac{q_{u_k}(v_k)}{r_{v_k}}$. It is clear that with a channel symmetric at the output, the output symbols will be equally likely if the input symbols are equally likely, i.e., r_{v_k} is a constant. Further, with such a channel, the specification of I_k (i.e., $q_{u_k}(v_k)$) gives no information as to which v_k was received, i.e., $\Pr[v_k|I_k] = r_{v_k}$. Thus for a channel symmetric at the output with equally likely inputs, $\Pr[V_w|A_j] = R_{V_w}$, and $\overline{N_{K_j}|A_j} = \overline{N_{K_j}}$. Thus, for these channels,

$$\overline{N} \leqq \sum_{j=1}^{\infty} \Pr[A_j] \overline{N_{K_j}}$$

$$\leqq \overline{N_{K_1}} + C_3 \, n e^{\Delta K} \sum_{j=2}^{\infty} e^{-K_j - \frac{K_j(1+\sigma_1)[R+\mu_1(-1)]}{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)}} \quad (4.24)$$

* A channel with transition probability matrix $q_1(j)$ is symmetric at its output if the set of probabilities $\{q_1(j), q_2(j), \ldots q_a(j)\}$ is the same for all output symbols j.

The exponent in the summation is

$$-K_j \left\{ 1 + \frac{(1+\sigma_1)[R+\mu_1(-1)]}{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)} \right\}$$

The summation clearly converges for

$$R + \mu_1(-1) < - \frac{\mu_1(\sigma_1) + \sigma_1\mu_1(-1)}{1+\sigma_1}$$

or

$$R < -\mu_1(-1) - \frac{\mu_1(\sigma_1)+\sigma_1\mu_1(-1)}{1+\sigma_1} = - \frac{(1+2\sigma_1)\mu_1(-1)+\mu_1(\sigma_1)}{1+\sigma_1}$$

We have the further condition $R < \dfrac{\mu_1(\sigma_1)}{\sigma_1}$

The two conditions are clearly identical for $\sigma_1 = -\frac{1}{2}$

Define

$$\left.\frac{\mu_1(\sigma_1)}{\sigma_1}\right|_{\sigma_1 = -\frac{1}{2}} = R_{cutoff} \tag{4.25}$$

Substituting $\sigma_1 = -\frac{1}{2}$, the exponent reduces to

$$-K_j\left[\frac{R_{cutoff} - R}{R_{cutoff}+\mu_1(-1)}\right]$$

Substituting Eq. (4.18) into Eq. (4.24) we obtain

$$\frac{\overline{N}}{C_3} < \exp(BK_1) + n\exp(\Delta K)\sum_{j=2}^{\infty}\exp[(B-1)K_j] \tag{4.26}$$

where

$$B = \frac{R+\mu_1(-1)}{R_{cutoff}+\mu_1(-1)} \tag{4.27}$$

Since $K_j = K_1 + (j-1)\,\Delta K$,

$$\frac{\overline{N}}{C_3} < \exp(BK_1) + n \exp(\Delta K) \sum_{j=2}^{\infty} \exp\{(B-1)[K_1 + (j-1)\Delta K]\}$$

$$= \exp(BK_1) + n \exp[(B-1)K_1] \exp(\Delta K) \sum_{j=1}^{\infty} \exp[j(B-1)\Delta K]$$

$$= \exp(BK_1) + n \exp[(B-1)K_1] \frac{\exp[B\Delta K]}{1 - \exp[(B-1)\Delta K]} \qquad (4.26a)$$

We desire to select K_1 and ΔK to minimize $\overline{N}$. Differentiating with respect to these variables and setting the resulting equations equal to zero we obtain:

$$\Delta K = \frac{\log B}{B - 1} \qquad (4.28a)$$

$$K_1 = \Delta K + \log n \qquad (4.28b)$$

Substituting Eq. (4.28) into Eq. (4.26a) we obtain

$$\overline{N} < \frac{C_3 B}{1 - B} \left(\frac{B}{B-1}\right)^{B} n \qquad (4.29)$$

Since $B < 1$ for $R < R_{cutoff}$, $\overline{N}$ increases with n less quickly than linearly.

The result $R_{cutoff} = \left.\frac{\mu_1(\sigma_1)}{\sigma_1}\right|_{\sigma_1 = -\frac{1}{2}}$ may be obtained from another point of view. From Eq. (4.10) we see that $\overline{N_{K_j}}$ contains a term of the form

$$e^{-K_j} \sum_w e^{w\,[R-\mu_1'(\sigma_w^{(j)})]}$$

However, we weight $\overline{N_{K_j}}$ with the exponential e^{-K_j} and sum on j:

$$\sum_j \sum_w e^{w[R-\mu_1'(\sigma_w^{(j)}) - \frac{2K_j}{w}]}$$

However, $\frac{K_j}{w} = E_1(\sigma_w^{(j)})$. Thus at least one term can be found in the double summation corresponding to the minimum value of $2E_1 + \mu_1'$. This expression has a minimum value at $\sigma = -\frac{1}{2}$.

But

$$2E_1(-\tfrac{1}{2}) + (-\tfrac{1}{2}) = -2\mu_1(-\tfrac{1}{2}) = \left.\frac{\mu_1(\sigma)}{\sigma}\right|_{\sigma = -\frac{1}{2}}$$

$$= R_{cutoff}$$

This shows that the piecewise linear approximation to the $E_1(\sigma)$ vs. $\mu_1'(\sigma)$ curve made in bounding $\overline{N_K}$ did not deteriorate the determination of the cutoff rate. This cutoff rate is inherent in the bounding of $\overline{N_K}$ by Eq. (4.10) which precedes any approximation to the E vs. μ' curve.

F. Number of Decoding Computations - Procedure of Chapter III, Section D

For the modified decoding procedure of Chapter III, section D, we will count the comparison of $d_w - d_r$ with D_{w-r} as one "computation". Thus for each of the computations for the procedure of Chapter III, section C, there will be no more than n computations for the procedure of Chapter III, section D.

Referring to Figure 5, we may define the level of a node as the length of the sequence that terminates on it, i.e., X_w terminates on a node of level w. Consider the correct X_n sequence and the set of nodes it impinges on.

Defining a branching node as one from which $m > 1$ branches emerge, we observe that there are $\frac{nR}{\log m}$ branching nodes in this set. For that branching node of level w in this set there diverges m-1 branches other than the correct branch. Each of these other branches induces U sequences which, over the ensemble of $\{ F \}$ and $\{ G \}$, are independent of the transmitted U sequence. Thus, the processing of incorrect branches out of each of these branching nodes can require no more average computations than $n\overline{N}_K$. Thus the number of computations required for processing all of $\{ X_n \}$ with respect to criterion K, which we call $\overline{N}_K{}'$ is bounded by

$$\overline{N}_K{}' \leq n^2 \frac{R}{\log m} \overline{N}_K \tag{4.30}$$

For symmetric channels, we may proceed to bound the average number of decoding computations. We do this only for the interesting case of $R > -\mu_1(-1)$.

Before the transmitted sequence of length n is accepted with respect to criterion K, it is subjected to $\frac{n(n+1)}{2}$ comparisons. The probability that it fails one or more of these comparisons is less than or equal to $\frac{n(n+1)}{2} e^{-K}$. We conservatively assume that processing of $\{ X_n \}$ with respect to K_j is required whenever the correct sequence fails to satisfy criterion K_{j-1}, we have, by analogy with Eq. (4.25):

$$\frac{\overline{N}'}{C_3 \frac{R}{\log m}} < n^2 \exp(BK_1) + n^4 \exp(\Delta K) \sum_{j=2}^{\infty} \exp[(B-1)K_j] \tag{4.31}$$

In Eq. (4.31) we have upper bounded $\frac{n(n+1)}{2}$ with n^2.

Since $K_j = K_1 + (j-1)\Delta K$, Eq. (4.31) may be rewritten as:

$$\frac{\overline{N}'}{C_3 \frac{R}{\log m}} < n^2 \exp(BK_1) + n^4 \exp[(B-1)K_1] \frac{\exp[B\Delta K]}{1 - \exp[(B-1)\Delta K]} \tag{4.31a}$$

We desire to select K_1 and ΔK to minimize $\overline{N}'$. Differentiating with respect to these variables and setting the resulting equations equal to zero, we obtain:

$$\Delta K = \frac{\log B}{B - 1} \tag{4.28a}$$

$$K_1 = \Delta K + 2 \log n \tag{4.32}$$

Substitution of Eq. (4.28a) and (4.32) into Eq. (4.31a) yields

$$\overline{N}' < C_3 \frac{R}{\log m} \frac{B^{\frac{B}{B-1}}}{1 - B} n^2 (1 + B) \tag{4.33}$$

To recapitulate, the bound of Eq. (4.33) refers to the average number of decoding computation for the procedure of Chapter III, section D. This contrasts with the bound of Eq. (4.29) which refers to the average number of computations required to process $\{X_n^i\}$ before a decision is made, for the procedure of Chapter 3, section C. For the latter procedure, a meaningful bound to the number of computations required to process $\{X_n^c\}$ is not known.

CHAPTER V

PROBABILITY OF ERROR

A. General Bound - Decoding Procedure of Chapter III, Section C

Suppose we conservatively count as a decoding error the occurrence of either or both of the following events:

1. The transmitted sequence U_n and the received sequence V_n are such that they fail to meet some fixed criterion, say K_j. The probability of this event, over the ensemble, is less than $n e^{-K_j}$.

2. Any element of $\{ U_n^i \}$ together with the received V_n satisfies K_j. An element of $\{ U_n^i \}$ picked at random, together with the received V_n has a probability of satisfying K_j equal to $Pr_2 [d_n \geqq D_n^{(j)}]$. However, there are $\frac{m-1}{m} e^{nR}$ elements of $\{ U_n^i \}$. Since the probability of the union of events is upper bounded by the sum of the probabilities of the individual events, the probability that any element of $\{U_n^i\}$ together with the received V_n satisfies K_j is less than $\frac{m-1}{m} e^{nR} Pr_2 [d_n \geqq D_n^{(j)}]$.

The two events alluded to above are not in general independent. However, the probability of their union is upper bounded by the sum of their probabilities. Thus the probability of error, P_e, may be bounded:

$$P_e \leqq n e^{-K_j} + \frac{m-1}{m} e^{nR} Pr_2 [d_n \geqq D_n^{(j)}] \tag{5.1}$$

In Chapter IV we defined $D_n^{(j)}$ by setting it equal to $n\mu_1'(\sigma_n^{(j)})$, where $E_1(\sigma_n^{(j)}) = \frac{K_j}{n}$, $\sigma_n^{(j)} < 0$. The bound will take different forms depending on whether $R > \mu_1'(-1)$ or $R \leqq \mu_1'(-1)$.* Thus we select j conveniently for these two cases. Let

* In Appendix B it is shown that $\mu_1'(-1) < 0$ for non-degenerate channels. Thus $R > \mu_1'(-1)$ is the case of most practical interest.

$$\mu_1'(\sigma_n^{(j+1)}) < R \leqq \mu_1'(\sigma_n^{(j)}),\ R > \mu_1'(-1) \tag{5.2a}$$*

$$\mu_1'(\sigma_n^{(j)}) \leqq R < \mu_1'(\sigma_n^{(j-1)}),\ R \leqq \mu_1'(-1) \tag{5.2b}$$*

From the fact that μ_1' is monotomic with σ, $\sigma_n^{(j)} > -1$ for $R > \mu_1'(-1)$ and $\sigma_n^{(j)} \leqq -1$ for $R \leqq \mu_1'(-1)$. For $\sigma_n^{(j)} > -1$, there exists a $\lambda_n^{(j)} > 0$ such that $D_n^{(j)} = n\mu_2'(\lambda_n^{(j)})$. In fact $\lambda_n^{(j)} = \sigma_n^{(j)} + 1$. In this case

$$Pr_2\ [d_w \geqq D_n^{(j)}] \leqq e^{-nE_2(\lambda_n^{(j)})}$$

where $E_2(\lambda) = \lambda\mu_2'(\lambda) - \mu_2(\lambda)$. Since $E_2(\lambda_n^{(j)}) = E_1(\sigma_n^{(j)}) + \mu_1'(\sigma_n^{(j)})$ for $\lambda_n^{(j)} = \sigma_n^{(j)} + 1$, we may rewrite Eq. (5.1) for $R > \mu_1(-1)$ as:

$$P_e \leqq e^{-nE_1(\sigma_n^{(j)})} \left[n + \frac{m-1}{m}\, e^{n\,[R-\mu_1'(\sigma_n^{(j)})]} \right] \tag{5.1a}$$

Since $R - \mu_1'(\sigma_n^{(j)}) \leqq 0$, $\exp n\ [R-\mu_1'(\sigma_n^{(j)})] \leqq 1$ and Eq. (5.1a) further simplifies to

$$P_e \leqq (n+1)\, e^{-nE_1(\sigma_n^{(j)})} \tag{5.1b}$$

On the other hand, for $R \leqq \mu_1'(-1)$, $D_n^{(j)} \leqq n\mu_1'(-1)$. We have observed in Chapter IV that if $D_n^{(j)} \leqq n\mu_1'(-1)$,

$$Pr_2\ [d_n \geqq D_n^{(j)}] \leqq e^{n\mu_1(-1)}$$

---

* We have observed that the $E_1(\sigma)$ vs. $\mu_1'(\sigma)$ curve has negative slope for $\sigma < 0$. Thus, since $E_1(\sigma_n^{(j)}) > E_1(\sigma_n^{(j-1)})$, it follows that $\mu_1'(\sigma_n^{(j)}) < \mu_1'(\sigma_n^{(j-1)})$.

In this case Eq. (5.1) becomes

$$P_e \leqq n e^{-nE_1(\sigma_n^{(j)})} + \frac{m-1}{m} e^{n[R+\mu_1(-1)]} \tag{5.1c}$$

Since the $E_1(\sigma)$ vs. $\mu_1'(\sigma)$ is convex, it is above the tangent at $\sigma = -1$ which has the equation $-\mu_1'(\sigma) - \mu_1(-1)$. Thus

$$E_1(\sigma_n^{(j)}) \geqq -\mu_1'(\sigma_n^{(j)}) - \mu_1(-1) \tag{5.3}$$

Further we have selected j such that $R \geqq \mu_1'(\sigma_n^{(j)})$. Substituting this inequality in Eq. (5.2) we obtain

$$E_1(\sigma_n^{(j)}) \geqq -[R + \mu_1(-1)] \tag{5.4}$$

Equation (5.4) shows that the second exponential in Equation (5.1c) will dominate the first. Thus we may write for $R \leqq \mu_1'(-1)$:

$$P_e \leqq (n+1) e^{n[R+\mu_1(-1)]} \tag{5.1d}$$

Recapitulating, with j related to R as given in Equation (5.2),

$$P_e \leqq \begin{cases} (n+1) e^{-nE_1(\sigma_n^{(j)})}, & R > \mu_1'(-1) \quad (5.1b) \\ (n+1) e^{n[R+\mu_1(-1)]}, & R \leqq \mu_1'(-1) \quad (5.1d) \end{cases}$$

We have observed that the $E_1(\sigma)$ vs. $\mu_1'(\sigma)$ curve is convex with slope σ. Then for any j,

$$\left| \frac{E_1(\sigma_n^{(j+1)}) - E_1(\sigma_n^{(j)})}{\mu_1'(\sigma_n^{(j+1)} - \mu_1'(\sigma_n^{(j)})} \right| \geq |\sigma_n^{(j)}|$$

Since $E_1(\sigma_n^{(j+1)}) - E_1(\sigma_n^{(j)}) = \frac{K_{j+1}}{n} - \frac{K_j}{n} = \frac{\Delta K}{n}$,

we have

$$|\mu_1'(\sigma_n^{(j+1)}) - \mu_1'(\sigma_n^{(j)})| \leq \frac{\Delta K}{\sigma_n^{(j)}} \cdot \frac{1}{n} \tag{5.5}$$

Equation (5.5) shows that the upper and lower bounds to R in Eq. (5.2) approach each other for large n. Thus for large n we can write Eq. (5.1b) in the approximate form:

$$P_e \lesssim (n+1)\, e^{-nE_1(\sigma)}, \quad R = \mu_1'(\sigma) > \mu_1'(-1) \tag{5.1e}$$

It is to be noted that the above discussion is applicable to all $R < C$. In particular it applies to $R_{cutoff} \leq R < C$.

The exponents in Equations (5.1d) and (5.1e) are identical to the exponents derived by Shannon[2] for block coding. Shannon's results are:

$$P_e \leq \begin{cases} 2\, e^{-E(\sigma)}, & R = \mu_1'(\sigma) > \mu_1'(-1) \quad (5.6a) \\ 2\, e^{n[R+\mu_1(-1)]}, & R \leq \mu_1'(-1) \quad (5.6b) \end{cases}$$

Equations (5.6a) and (5.6b) differ from Equations (5.1e) and (5.1d) respectively only insofar as the coefficient (n+1) in the latter pair of equations is replaced by 2 in the former pair.

B. Bound for $R < -\mu_1(-1)$

Since $0 < E_1(-1) = -\mu_1'(-1) - \mu_1(-1)$ we may conclude that $\mu_1'(-1) < -\mu_1(-1)$. For the general case of $R < -\mu_1(-1)$, we have been able to bound the number of decoding computations by slightly modifying the decoding procedure: we accept an error whenever no sequence satisfies some maximum criterion $K_{j_{max}}$. It is clear that if we select j_{max} to satisfy Equation (5.2), the probability of error results of Equations (5.1b) and (5.1d) follow.

Define $\sigma < 0$ such that $R = \mu_1'(\sigma)$. From Equation (5.2) and our method of selecting j_{max}, we may write

$$\frac{K_{j_{max}} - 1}{n} = \frac{K_1 + (j_{max} - 2)\,\Delta K}{n} < E_1(\sigma) \tag{5.7}$$

Solving Equation (5.7) for j_{max} we obtain

$$j_{max} < 2 + \frac{E_1(\sigma)}{\Delta K}\, n \tag{5.8}$$

In Equation (4.21) we have bounded $\overline{N}$ as a function of j_{max}. Substituting Equation (5.8) into the expression for $\overline{N}$ we have

$$\overline{N} < \left[2 + \frac{E_1(\sigma)}{\Delta K}\, n\right] C_3 \text{ for } R = \mu_1'(\sigma) < -\mu_1(-1) \tag{4.21a}$$

C. Bound for Channels Symmetric At Their Output - Decoding Procedure of Chapter III, Section C

We have observed in Chapter IV that for a channel symmetric at the output with inputs that are equally likely, the outputs will be equally likely. Further, with such a channel the specification of I_k (i.e., $q_{u_k}(v_k)$) gives no information as to which v_k was received, i.e., $\Pr[v_k | I_k] = r_{v_k}$. Thus over the ensemble of encoding schemes for such a channel, the specification that the smallest criterion satisfied by the correct sequence is the $j^{\underline{th}}$ does not bias the probability measure on the pair (U_n, V_n) for U_n an element of $\{U_n^i\}$ and V_n received.

Let B_j be the event that the $j^{\underline{th}}$ is the smallest criterion satisfied by the transmitted sequence. For $j > 1$, B_j implies that the correct sequence fails to meet the $(j-1)^{\underline{st}}$ criterion. This has probability upper bounded by ne^{-Kj-1}. Thus

$$\Pr[B_j] \leqq \left\{ \begin{array}{ll} 1 & , \ j = 1 \\ ne^{-Kj-1} = ne^{\Delta K} e^{-K_j} & , \ j > 1 \end{array} \right\} \tag{5.9}$$

Conditional on the event B_j, we will always accomplish a correct decoding if no element of $\{U_n^i\}$ satisfies K_j. Due to the symmetry of the channel, the probability measure on the pair (U_n, V_n) for U_n an element of $\{U_n^i\}$ and V_n received is $\Pr_2[U_n, V_n]$ and the probability that this (U_n, V_n) pair satisfies the j^{th} criterion is $\Pr_2[d_n \geqq D_n^{(j)}]$. Let C_j be the event that any elements of $\{U_n^i\}$ together with the received V_n satisfies K_j. C_j is the union of $\frac{m-1}{m} e^{nR}$ events: the satisfying of K_j by the individual elements of $\{U_n^i\}$ of which there are $\frac{m-1}{m} e^{nR}$ in number. Since the probability of a union of events is upper bounded by the sum of the probabilities of the individual events, we may write

$$\Pr[C_j] \leq \frac{m-1}{m} e^{nR} \Pr_2[d_n \geq D_n^{(j)}] \tag{5.10}$$

We will use the bound of Equation (5.10) for $1 \leq j \leq \hat{j}$, where $\hat{j}$ remains to be defined. If we select $\hat{j}$ such that $D_n^{(\hat{j})} = n\mu_1'(\sigma_n^{(\hat{j})}) > n\mu_1'(-1)$, then we have shown that $\mu_2'(\lambda_n^{(j)}) = \mu_1'(\sigma_n^{(j)})$ for $\lambda_n^{(j)} = \sigma_n^{(j)} + 1$ and $-1 < \sigma_n^{(j)} < 0$. This holds for all $j \leq \hat{j}$. In this range,

$$\Pr_2[d_n \geq D_n^{(j)}] \leq e^{-nE_2(\lambda_n^{(j)})} = e^{-n[E_1(\sigma_n^{(j)}) + \mu_1'(\sigma_n^{(j)})]} \tag{5.11}$$

Thus the right hand side of Equation (5.10) is less than unity for $R \leq E_1(\sigma_n^{(j)}) + \mu_1'(\sigma_n^{(j)})$ with the restriction that $\sigma_n^{(j)} > -1$.

It will be convenient to define the quantity $E_1(-\frac{1}{2}) + \mu_1'(-\frac{1}{2})$.

$$R_{crit} = E_1(-\tfrac{1}{2}) + \mu_1'(-\tfrac{1}{2}) \tag{5.12}$$

We now select $\hat{j}$ such that

$$E_1(\sigma_n^{(\hat{j}+1)}) + \mu_1'(\sigma_n^{(\hat{j}+1)}) < R \leq E_1(\sigma_n^{(\hat{j})}) + \mu_1'(\sigma_n^{(\hat{j})}), \text{ for } R \geq R_{crit} \tag{5.13a}$$

$$E_1(\sigma_n^{(\hat{j})}) + \mu_1'(\sigma_n^{(\hat{j})}) \leq R < E_1(\sigma_n^{(\hat{j}-1)}) + \mu_1'(\sigma_n^{(\hat{j}-1)}), \text{ for } -\mu_1(-1) < R < R_{crit} \tag{5.13b}$$

$\hat{j}$ = maximum value of j such that $-\mu_1(-1) < E_1(\sigma_n^{(j)}) + \mu_1'(\sigma_n^{(j)})$ for $R \leq -\mu_1(-1)$ (5.13c)

* The function $E(\sigma) + \mu'(\sigma)$ vs. σ has the slope $(1+\sigma)\mu''(\sigma)$ which is always positive for $-1 < \sigma \leq 0$. Thus $E(\sigma_1)+\mu'(\sigma_1) < E(\sigma_2)+\mu'(\sigma_2)$ if $-1 \leq \sigma_1 < \sigma_2 \leq 0$.

We have already defined the event A_j as being the failure of the correct sequence to satisfy K_{j-1}. Let us count as an error whenever the event $A_{\hat{j}+1}$ occurs. From Equation (4.22),

$$\Pr [A_{\hat{j}+1}] \leq n e^{-K_{\hat{j}}} \tag{5.14}$$

If $A_{\hat{j}+1}$ does not occur, we will bound the probability of error by the summation

$$P'_e \leq \sum_{j=1}^{\hat{j}} \Pr [B_j] \Pr [C_j] \tag{5.15}$$

Substituting Equations (5.9), (5.10) and (5.11) in Equation (5.15), we obtain

$$P'_e \leq \frac{m-1}{m} e^{-n [E_1(\sigma_n^{(1)}) + \mu'_1(\sigma_n^{(1)}) - R]} + \frac{n(m-1)}{m} e^{\Delta K + nR} \sum_{j=2}^{\hat{j}} e^{-n[E_1(\sigma_n^{(j)}) + \mu'_1(\sigma_n^{(j)})] - K_j} \tag{5.16}$$

For each of the elements in the summation of Equation (5.16), $j \leq \hat{j}$. Thus $\sigma_n^{(j)} \geq \sigma_n^{(\hat{j})} > -1$ and $K_j = n E_1(\sigma_n^{(j)})$ and we may rewrite Equation (5.16) as:

$$P'_e \leq e^{-n [E_1(\sigma_n^{(j)}) + \mu'_1(\sigma_n^{(1)}) - R]} + n e^{\Delta K + nR} \sum_{j=2}^{\hat{j}} e^{-n [2E_1(\sigma_n^{(j)}) + \mu'_1(\sigma_n^{(j)})]} \tag{5.16a}$$

The function $2E_1(\sigma) + \mu_1'(\sigma)$ has a minimum at $\sigma = -\frac{1}{2}$ at which point the function equals $-2\mu_1(-\frac{1}{2}) = R_{cutoff}$. For $\sigma_n^{(\hat{j})} \geqq -\frac{1}{2}$, the maximum term in the summation of Equation (5.15a) is $\exp(-n[2E_1(\sigma_n^{(\hat{j})}) + \mu_1'(\sigma_n^{(\hat{j})})])$. For $\sigma_n^{(\hat{j})} < -\frac{1}{2}$, the maximum term in the summation is less than or equal to $e^{-nR_{cutoff}}$. From Equation (5.13a) we see that for $R \geqq R_{crit}$, $\sigma_n^{(\hat{j})} \geqq -\frac{1}{2}$, and the maximum term in the summation is $\exp(-n[2E_1(\sigma_n^{(\hat{j})}) + \mu_1'(\sigma_n^{(\hat{j})})])$. For $R < R_{crit}$, $\sigma_n^{(\hat{j})} < -\frac{1}{2}$ and the maximum term is less than or equal to $e^{-nR_{cutoff}}$.

In Equation (5.13) we defined $\sigma_n^{(\hat{j})} > -1$. Thus

$$E_1(-1) > E_1(\sigma_n^{(\hat{j})}) = \frac{1}{n}[K_1 + (\hat{j}-1)\Delta K] \tag{5.17}$$

Solving Equation (5.17) for j in terms of known constants, we have:

$$\hat{j} < \frac{nE_1(-1)}{\Delta K} + 1 \tag{5.18}$$

The summation of Equation (5.15b) is upper bounded by the maximum term multiplied by the number of terms. Thus for $R \geqq R_{crit}$,

$$P_e' < e^{-n[E_1(\sigma_n^{(1)}) + \mu_1'(\sigma_n^{(1)}) - R]} + \frac{n^2 e^{\Delta K} E_1(-1)}{\Delta K} e^{-nE_1(\sigma_n^{(\hat{j})})} \tag{5.19}$$

In Equation (5.19) we have made use of Equation (5.13a) in upper bounding $\exp(-n[E_1(\sigma_n^{(1)}) + \mu_1'(\sigma_n^{(1)}) - R])$ by unity. For $R < R_{crit}$

$$P_e' < e^{-n[E_1(\sigma_n^{(1)}) + \mu_1'(\sigma_n^{(1)}) - R]} \frac{n^2 e^{\Delta K} E_1(-1)}{\Delta K} e^{-n[R_{cutoff} - R]} \tag{5.20}$$

We have bounded the probability of error conditional on the occurrence of event A_{j+1} (Equation (5.14)) and conditional on the non-occurrence of event A_{j+1} (Equations (5.19) and (5.20)). The over-all probability of error, P_e, is upper bounded by the sum of these conditional probabilities. In Equation (5.14) we may set $K_{\hat{j}} = nE_1(\sigma_n^{(\hat{j})})$ since $-1 < \sigma_n^{(\hat{j})} < 0$.

Thus for $R \geqq R_{crit}$

$$P_e < e^{-n[E_1(\sigma_n^{(1)}) + \mu_1'(\sigma_n^{(1)}) - R]} + ne^{\Delta K}\left[\frac{nE_1(-1)}{\Delta K} + 1\right]e^{-nE_1(\sigma_n^{(\hat{j})})} \tag{5.21a}$$

For $R < R_{crit}$

$$P_e < e^{-n[E_1(\sigma_n^{(1)}) + \mu_1'(\sigma_n^{(1)}) - R]} + ne^{\Delta K}\, e^{-nE_1(\sigma_n^{(\hat{j})})} + \left[\frac{n^2 e^{\Delta K} E_1(-1)}{\Delta K}\right] e^{-n[R_{cutoff} - R]} \tag{5.21b}$$

The plot of $E_1(\sigma)$ vs. $E_1(\sigma) + \mu_1'(\sigma)$ has slope $\frac{\sigma}{\sigma + 1}$ and second derivative $\frac{1}{(1+\sigma)^3 \mu_1''(\sigma)}$ which is positive for $-1 < \sigma < 0$. In this range, the curve is above the tangent at $\sigma = -\frac{1}{2}$ which has the equation $R_{cutoff} - [E_1(\sigma) + \mu_1'(\sigma)]$.

From Equation (5.13b), $R \geq E_1(\sigma_n^{(\hat{j})}) + \mu_1'(\sigma_n^{(\hat{j})})$ for $R < R_{crit}$. Thus,

$$R_{cutoff} - R \leq R_{cutoff} - [E_1(\sigma_n^{(\hat{j})}) + \mu_1'(\sigma_n^{(\hat{j})})] \tag{5.22}$$

The right hand side of Equation (5.22) is the expression for the tangent to the $E_1(\sigma)$ vs. $E_1(\sigma) + \mu_1'(\sigma)$ curve at $\sigma = -\frac{1}{2}$. From the fact that the curve is above the tangent, we conclude $R_{cutoff} - R < E_1(\sigma_n^{(\hat{j})})$ for $R < R_{crit}$. Furthermore, $R_{cutoff} < E_1(\sigma_n^{(1)}) + \mu_1'(\sigma_n^{(1)})$ for reasonable values of n. Thus the third exponential dominates the first and second exponentials in Equation (5.21b) and the equation may be simplified to:

$$P_e < \left[n^2 \frac{e^{\Delta K} E_1(-1)}{\Delta K} + ne^{\Delta K} + 1\right] e^{-n[R_{cutoff}-R]} \tag{5.21c}$$

We are concerned with the asymptotic forms of Equations (5.21a) and (5.21c) for large n. By an argument similar to the one in section A of this chapter, it is easy to show that the upper and lower bounds to R in Equation (5.13a) approach each other. Thus for large n, we may define σ by the equation

$$R = E_1(\sigma) + \mu_1'(\sigma).$$

Furthermore, for large n, $\sigma_n^{(1)} \to 0$ since $E_1(\sigma_n^{(1)}) = \frac{K_1}{n}$ and K_1 varies as log n. Thus for sufficiently large n, the first exponential of Equation (5.21a) will be dominated by the second. Dropping all coefficients but those of highest degree in n, we have for large n:

$$P_e \lesssim \begin{cases} n^2 \dfrac{e^{\Delta K E_1(-1)}}{\Delta K} e^{-nE_1(\sigma)}, & R = E_1(\sigma)+\mu_1'(\sigma) \geqq R_{crit} \quad (5.22a) \\ n^2 \dfrac{e^{\Delta K E_1(-1)}}{\Delta K} e^{-n[R_{cutoff} - R]}, & R < R_{crit} \quad (5.22b) \end{cases}$$

Channels which are symmetric at the output and with equally likely inputs have been studied by Fano and Shannon. In unpublished work, Shannon has shown that for block codes of length n on these channels, and $R = E_1(\sigma) + \mu_1'(\sigma) \geqq R_{crit}$, the exponent of Equation (5.22a) is indeed optimum, i.e., the best block code of length n will have a probability of error with an exponential term of the form $e^{-nE_1(\sigma)}$. Fano[6] and Shannon in unpublished work have also shown that for $R < R_{crit}$, an upper bound to the probability of error has an exponential term of the form $e^{-n[R_{cutoff} - R]}$.

The error exponent E(R) vs. rate R for a typical symmetric channel is shown in Figure 8.

D. Probability of Error-Decoding Procedure of Chapter III, Section D.

The analysis of sections A, B, and C of this chapter may be repeated for the modified decoding procedure of Chapter III, section D. It is straightforward and will be omitted. It differs from the previous analyses only in that the probability that the correct sequence of length n fails to satisfy K_j is upper bounded by $\frac{n(n+1)}{2} e^{-K_{j-1}}$ rather than $ne^{-K_{j-1}}$. We present here the limiting forms for n very large.

General bound:

$$P_e \lesssim \begin{cases} \dfrac{n^2}{2} e^{-nE_1(\sigma)}, & R = \mu_1'(\sigma) > \mu_1'(-1) \quad (5.23a) \\ \dfrac{n^2}{2} e^{n[R+\mu_1(-1)]}, & R \leqq \mu_1'(-1) \quad (5.23b) \end{cases}$$

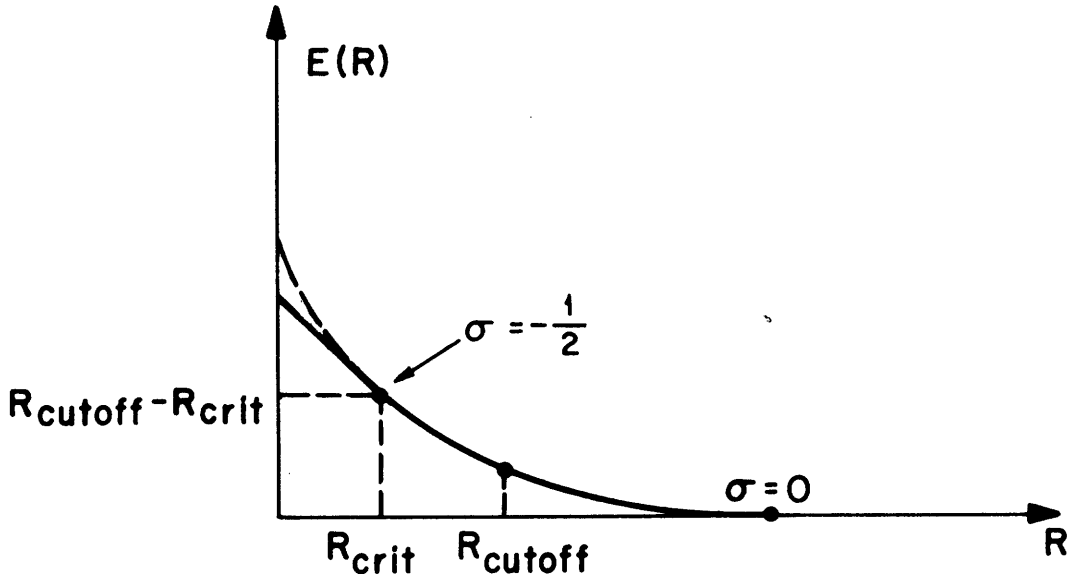


FIG. 8 ERROR EXPONENT vs. RATE FOR A CHANNEL SYMMETRIC AT OUTPUT WITH EQUALLY LIKELY INPUTS.

Bound for channel symmetric at output with equally likely inputs:

$$P_e \lesssim \begin{cases} n^3 \dfrac{e^{\Delta K} E_1(-1)}{2\Delta K} e^{-nE_1(\sigma)}, & R = E_1(\sigma)+\mu_1'(\sigma) \geqq R_{crit} \quad (5.24a) \\ n^3 \dfrac{e^{\Delta K} E_1(-1)}{2\Delta K} e^{-n[R_{cutoff}-R]}, & R < R_{crit} \quad (5.24b) \end{cases}$$

CHAPTER VI

MISCELLANEOUS TOPICS

A. The Generation of Symmetric from Asymmetric Channels

Equation (4.29) bounds the average number of computations required to process the incorrect subset for channels which are symmetric at their output with equally likely inputs. No such non-exponential bound has been obtained for the general asymmetric channel except in the almost trivial case of $R < -\mu_1(-1)$. We conjecture that the number of decoding computations for the general channel is an algebraic function of n. However, this is only a conjecture, and we are motivated to try to make a stronger statement. To this end we investigate means of converting a general asymmetric channel into a symmetric channel for which Equation (4.29) is applicable.

Our starting point is a paper by Shannon[7] where he defines channel inclusion. We quote the definition from Shannon, with slight changes in notation:

Definition: Let $q_i(j)$ (i=1, ..., a; j = 1, ..., b) be the transition probabilities for a discrete memoryless channel C_1 and $p_k(\ell)$ (k = 1, ..., c; ℓ = 1, ..., d) be those for C_2. We shall say that C_1 includes C_2, $C_1 \geqq C_2$, if and only if there exist two sets of transition probabilities, $r_{ak}(i)$ and $t_{aj}(\ell)$, with

$$r_{ak}(i) \geqq 0, \quad \sum_i r_{ak}(i) = 1 \tag{6.1a}$$

and

$$t_{aj}(\ell) \geqq 0, \quad \sum_\ell t_{aj}(\ell) = 1 \tag{6.1b}$$

and there exists

$$g_a \geqq 0, \quad \sum_a g_a = 1$$

with

$$\sum_{a,\,i,\,j} g_a\, r_{ak}(i)\, q_i(j)\, t_{aj}(\ell) = p_k(\ell) \tag{6.2}$$

The expression $\sum_{i,\,j} r_{ak}(i)\, q_i(j)\, t_{aj}(\ell)$ may be visualized as the entry in the k^{th} row and ℓ^{th} column of the matrix $[C_{2a}]$ defined by

$$[C_{2a}] = [R_a]\,[C_1]\,[T_a] \tag{6.3}$$

$[C_1]$ is the matrix of channel C_1 whose entry in its i^{th} row and j^{th} column is $q_i(j)$. $[C_1]$ is, of course, a stochastic matrix, i.e., one whose entries are all positive and whose rows add to unity. $[R_a]$ is a matrix whose entry in the k^{th} row and i^{th} column is $r_{ak}(i)$; $[T_a]$ is a matrix whose entry in the j^{th} row and ℓ^{th} column is $t_{aj}(\ell)$. From conditions (6.1a) and (6.1b), $[R_a]$ and $[T_a]$ are also stochastic matrices and thus they may be thought of as channels. $[C_{2a}]$ is a stochastic matrix, since the product of stochastic matrices is stochastic. In other words, C_{2a} is a channel obtained by the tandem connection of channels T_a, C_1, and R_a in the order named. That is, the i^{th} output of T_a feeds into the i^{th} input of C_1 and the j^{th} output of C_1 feeds into the j^{th} input of R_a.

Using Equation (6.3) we may rewrite Equation (6.2) as:

$$[C_2] = \Sigma\, g_a\, [C_{2a}] \tag{6.2a}$$

where $[C_2]$ is the transition probability matrix of C_2. The pre-channel T_a and the post-channel R_a are used as a pair with probability g_a. When this is the case we have made channel C_1 "look like" C_2.

Shannon points out that no loss of generality is implied if the channels T_a and R_a are all pure channels, i.e., all the entries in the $[T_a]$ and $[R_a]$

matrices are zero or unity. This implies that each of the channels T_a and R_a has the property that a fixed input always produces the same output (although different inputs may produce the same output).

If we are given a channel C_1 and it includes a channel C_2 symmetric at its output, and if the corresponding g_a is identically simulated at transmitter and receiver, then by means of the transmitter switching operations implied by pure channel T_a and the simultaneous receiver switching operations implied by pure channel R_a, C_1 is made to appear like C_2. We then may encode and decode as if C_2 were actually the channel available and be certain that the analyses of Chapter IV, Sections E and F, are applicable. The fact that C_1 includes at least one channel C_2 which is symmetric at its output is proven in Theorem 4.

Theorem 4: Let C_1 be a discrete memoryless channel with probability $q_i(j)$ of receiving the j^{th} output when the i^{th} input is used, $(i=0, \ldots, a-1; j=0, \ldots, b-1)$. Then, C_1 includes a channel C_2 with the following properties:

1. C_2 is symmetric at both input* and output
2. C_2 has υ inputs and outpus where $\upsilon = \min(a, b)$.
3. The corresponding pairs of pure channels (T_a, R_a) are υ in number and are equally likely, i.e., $q_a = \frac{1}{\upsilon}$ for all a.
4. If C_1 has non-zero capacity, then C_2 has non-zero capacity.

The proof of Theorem 4 is in Appendix A.

---

* A channel with transition probabilities $q_i(j)$ $(i=1, \ldots, a; j=1, \ldots, b)$ is symmetric at its input if the set $\{ q_i(1), \ldots, q_i(b)\}$ is the same for all i. A channel symmetric at both input and output must have an equal number of inputs and outputs.

Property 3 of Theorem 4 makes the simulation of g_a at the transmitter and receiver relatively simple. For the derived channel C_2 used with sequential encoding and decoding of constraint length n, we will want to generate sequences of a choices which are independent over length n. One method that comes to mind is to make use of the properties of maximal-length shift register sequences.* A discussion of this method will not be attempted here, since it would deviate from the central theme of the research and is only of incidental interest.

B. Sequential Decoding and Incorrect Decisions

Our discussions of the number of decoding computations and probability of error have been predicated on the assumption that no prior errors have been made. If an error occurs, say x_0 is decoding incorrectly, then the decoding of x_1** will result in a prohibatively large number of computations. This is so because the encoding and decoding procedures will make the incorrect x_0 enter into the determination of u_i for $1 \leqq i \leqq n - 1$. Thus, over the ensemble, the probability measure on the pair (U_w, V_w) for <u>any</u> U_w tested and V_w received will be $Pr_2 [U_w, V_w]$. This holds for $1 \leqq w \leqq n - 1$. Thus, from a practical point of view, communication is impossible after the receiver makes an error for the system as described. However, this situation may be mitigated by the following procedure:

Suppose the X sequence, which is the ϕ map of the source sequence S is expanded by placing after every ℓ symbols of X a sequence of n zeros. This

---

* These are discussed by W. Peterson[8] in his work on error correcting codes soon to be published as a book. Peterson gives a bibliography of the mathematical theory of these sequences.

** We assume no redundancy places in the X sequence.

defines the new sequence X' which has the property that after every sequence of length ℓ, the encoder is primed, i.e., returned to a state or condition known with certainty by the receiver. Stated in an equivalent way, X' permits the encoder and decoder to periodically "forget" the effects of all previous outputs of the message sequence S.

We may ask: given an X' sequence of length $\ell + n$ starting with the n priming symbols, what is the probability that one or more of the ℓ information symbols will be incorrectly decoded? Call this probability P_e'. Clearly

$$P_e' \leqq \ell P_e \tag{6.4}$$

where P_e is bounded for various cases in Chapter V.

Equations (5.23) and (6.4) show that for symmetrical channels we can accomplish block transmission of data in blocks of length ℓ, decode sequentially and attain a probability of error upper bounded by a quantity proportional to $\ell n^2 e^{-nE(R)}$. Here n is the decoding delay. Optimum block coding with delay n would have a probability of error proportional to $e^{-nE(R')}$ for $R' \geqq R_{crit}$, where R' is the reduced rate implied by the expansion of X into X':

$$R' = \frac{\ell}{\ell + n} R \tag{6.5}$$

The scheme described deteriorates the exponent from E(R') to E(R)*, but exhibits the same type of exponential decrease with delay n. If $\ell >> n$, the deterioration is negligible.

---

* Since $R' < R$, $E(R') > E(R)$.

C. The Computation Cutoff Rate and Critical Rate for the BSC

If the crossover probability of the BSC is $p = 1-q$, then the semi-invariant moment generating function, $\mu(\sigma)$, is

$$\mu_1(\sigma) = \log \left[p(2p)^{\sigma} + q(2q)^{\sigma} \right] \tag{6.6}$$

$$R_{cutoff} = -2\mu_1\left(-\frac{1}{2}\right) = \log 2 - \log \left[1 + 2\sqrt{pq} \right] \tag{6.7}$$

Making the substitution $p = \frac{1 - \Delta}{2}$, $q = \frac{1 + \Delta}{2}$, Equation (6.7) becomes:

$$R_{cutoff} = \log 2 - \log \left[1 + \sqrt{1 - \Delta^2} \right] \tag{6.7a}$$

The capacity of the BSC is:

$$C = \log 2 + p \log p + q \log q \tag{6.8}$$

In terms of Δ, Equation (6.8) becomes:

$$C = \frac{1}{2} \left[(1+\Delta) \log (1+\Delta) + (1-\Delta) \log (1-\Delta) \right] \tag{6.8a}$$

The critical rate, R_{crit}, is most easily expressed as the capacity of a BSC with crossover probability of p_{crit} where

$$p_{crit} = \frac{\sqrt{p}}{\sqrt{p} + \sqrt{q}} = 1 - q_{crit} \tag{6.9}$$

Thus

$$R_{crit} = \log 2 + p_{crit} \log p_{crit} + q_{crit} \log q_{crit} \qquad (6.10)$$

Defining Δ_{crit} by the relation $p_{crit} = \frac{1 - \Delta_{crit}}{2}$, we have as in Equation (6.8a)

$$R_{crit} = \frac{1}{2}\left[(1+\Delta_{crit}) \log (1+\Delta_{crit}) + (1-\Delta_{crit}) \log (1-\Delta_{crit})\right] \qquad (6.10a)$$

For a very good BSC, $p \to 0$ which implies that C and R_{cutoff} both approach log 2. Furthermore, $p_{crit} \to 0$ which implies that R_{crit} approaches log 2. Thus for very good channels, $R_{crit} \approx R_{cutoff} \approx C \approx \log 2$.

For a very poor BSC, $p \to \frac{1}{2}$ or $\Delta \to 0$. Expanding Equations (6.7a) and (6.8a) in a power series in Δ and preserving only the dominant term in each equation we have

$$R_{cutoff} \approx \frac{\Delta^2}{4} \qquad (6.7b)$$

$$C \approx \frac{\Delta^2}{2} \qquad (6.8b)$$

From Equation (6.9), as $p \to \frac{1}{2}$, $p_{crit} \to \frac{1}{2}$ which implies that $\Delta_{crit} \to 0$. From Equations (6.10a) and (6.8b) we then have

$$R_{crit} \approx \frac{\Delta_{crit}^2}{2} \qquad (6.10b)$$

But from Equation (6.9) we have for small Δ,

$$p_{crit} \approx \frac{1 - \frac{\Delta}{2}}{2}$$

This implies that $\Delta_{crit} \approx \frac{\Delta}{2}$ which when substituted into Equation (6.10b) yields:

$$R_{crit} \approx \frac{\Delta^2}{8} \tag{6.10c}$$

Comparison of Equations (6.7b), (6.8b) and (6.10c) show that for a very poor BSC, $4\,R_{crit} \approx 2\,R_{cutoff} \approx C \approx \frac{\Delta^2}{2}$.

Figure 9 is a plot of the ratios $\frac{R_{crit}}{C}$ and $\frac{R_{cutoff}}{C}$ for the range of C values. This plot shows that both ratios are monotonically increasing with C. Furthermore, except for the trivial channel corresponding to $p = 0$, $R_{crit} < R_{cutoff}$. We already had this result when we observed that $R_{crit} = E_1(-\frac{1}{2}) + \mu_1'(-\frac{1}{2})$ and $R_{cutoff} = 2E_1(-\frac{1}{2}) + \mu_1'(-\frac{1}{2})$. Thus $R_{cutoff} - R_{crit} = E_1(-\frac{1}{2})$ which is positive for non-trivial channels.

D. Deviation from Capacity due to Improperly Chosen Input Probabilities

Consider the discrete memoryless channel with transition probability $q_i(j)$ of receiving the j^{th} output when the i^{th} output is sent. Suppose its capacity, C, is attained when the i^{th} input is used with probability p_i'. Due to diaphantine constraints imposed by an encoding scheme for this channel, the i^{th} input is used with probability $p_i = p_i' + \delta_i$, where $\sum_i \delta_i = 0$. We desire an expression for the reduction in average mutual information between input and output for $\delta_i \neq 0$.

Observe that

$$C = \max_{p_i} \overline{I(i;j)} \tag{6.11}$$

where

$$\overline{I(i;j)} = \sum_{i,j} p_i\, q_i(j) \log \frac{q_i(j)}{r_j} \tag{6.12}$$

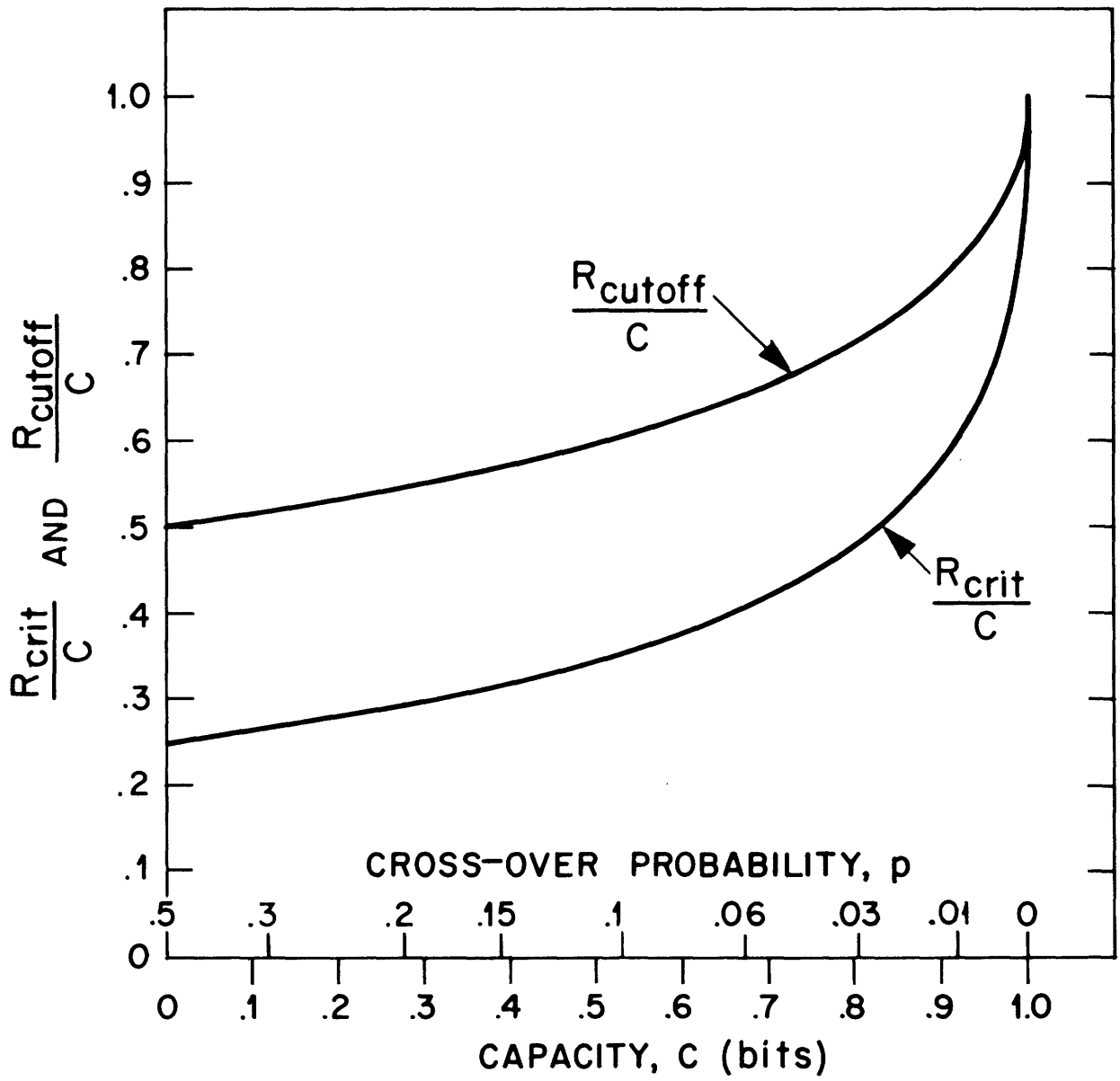


FIG. 9: $\frac{R_{crit}}{C}$ AND $\frac{R_{cutoff}}{C}$ VS. C AND p FOR THE BINARY SYMMETRIC CHANNEL

$$r_j = \sum_i p_i q_i(j) \tag{6.13}$$

Suppose

$$r_j = r'_j + \epsilon_j \tag{6.14}$$

where

$$r'_j = \sum_i p'_i q_i(j) \tag{6.15}$$

From Equations (6.13), (6.14) and (6.15) we see

$$\epsilon_j = \sum_i \delta_i q_i(j) \tag{6.16}$$

Suppose we expand Equation (6.12) in a power series about the set of points $\{p'_i\}$, and preserve only the lowest order deviation terms, since we assume the δ_i are small.

$$\bar{I} \approx C + \sum_k \left.\frac{\partial \bar{I}}{\partial p_k}\right|_{\{p_i = p'_i\}} \delta_k + \frac{1}{2} \sum_{k,\ell} \left.\frac{\partial^2 \bar{I}}{\partial p_k \partial p_\ell}\right|_{\{p_i = p'_i\}} \delta_k \delta_\ell \tag{6.17}$$

The summation $\sum_k \left.\frac{\partial \bar{I}}{\partial p_k}\right|_{\{p_i = p'_i\}} \delta_k$ must equal zero since p'_i corresponds

to capacity. Taking the second partial derivative of Equation (6.12), we have:

$$\frac{\delta^2 \bar{I}}{\partial p_k \, \partial p_\ell} = - \sum_j \frac{q_k(j)\, q_\ell(j)}{r_j} \tag{6.18}$$

Substituting Equation (6.18) in Equation (6.17) we have:

$$\bar{I} \approx C - \frac{1}{2} \sum_{j,k,\ell} \frac{\delta_k \delta_\ell \, q_k(j)\, q_\ell(j)}{r'_j} \tag{6.19}$$

Using Equation (6.16), Equation (6.19) simplifies to:

$$\bar{I} \approx C - \frac{1}{2} \sum_j \frac{(\epsilon_j)^2}{r'_j} \approx C - \frac{1}{2} \sum_j \frac{(\epsilon_j)^2}{r_j} \tag{6.19a}$$

Thus the fractional <u>loss</u> in rate, $\frac{\Delta \bar{I}}{C}$ is:

$$\frac{\Delta \bar{I}}{C} \approx \frac{\sum_j \frac{(\epsilon_j)^2}{r_j}}{2C} \tag{6.20}$$

E. The Zero Error Capacity of Symmetric Channels

The concept of zero error capacity of a noisy channel was introduced by Shannon[9]. The zero error capacity, C_0, is the least upper bound of rates at which it is possible to transmit information at zero probability of error. Only degenerate channels can have $C_0 > 0$; non-degenerate channels have $C_0 = 0$.

We show in Theorem 5, that for a discrete memoryless channel symmetric at input and output,

$$C_0 \geqq -\mu_1(-1)$$

Thus for a fully symmetric channel, the result of Equation (4.21a) is applicable for $R < C_0$.

CHAPTER VII

DISCUSSION OF RESULTS AND RECOMMENDATIONS FOR FUTURE WORK

The most striking feature of this research is that it presents a method of sending data over a discrete memoryless channel with a probability of error no more than a quantity proportional to $n^3 \exp[-nE(R)]$ and an average number of decoding computations no more than a quantity proportional to n^4, where n is the decoding delay and E(R) is an exponent independent of n, but dependent on rate, R. There is a suggestion, but not proof, that the average number of computations can in fact be upper bounded by a quantity proportional to n.* These results actually hold for a channel symmetric at its output. However, we show in Chapter VI that all channels may be reduced to symmetric channels for which these results hold.

The fact that the average number of computations varies as n raised to some power rather than exponentially with n, while the probability of error varies exponentially with n, permits us to consider the practical realization of sequential decoders for interestingly large values of n, assured that the complexity of these decoders will not be prohibitive.

The philosophy of this research clearly extends to the case of a semi-continuous channel, i.e., a channel with discrete inputs and a continuum of outputs. Here, the mutual information function, I(u; v), is a continuous rather than a discrete random variable. The thresholds $D_w^{(j)}$ may be defined

---

* We have shown in Chapter IV that the average number of computations required for the processing of the incorrect subset is upper bounded by a quantity proportional to n for the decoding scheme of Chapter III, Section C. The experimental work of Horstein[10] on sequential decoding for the BSC suggests that this quantity is, in fact, an upper bound to the average total number of decoding computations. Further experimental work is clearly required for verification of this point.

as in Chapter IV and the decoding can proceed as described in Chapter III. There is one practical difficulty here: one cannot expect the decoding computer, whether it be analog or digital, to form the function I(u; v) without error. The effect of these errors can only deteriorate the performance. The practical treatment of the situation, at least for digital computers, is to quantize the range of I or equivalently quantize the range of the output v. However, this reduces the channel to a new channel with discrete outputs: the case we have considered.

There is a major unsatisfying aspect of the foregoing summary: this is the fact that we have not made any meaningful statement about $\overline{N}$ for asymmetric channels. True, we have shown that any asymmetric channel may be reduced to a symmetric channel for which our results are valid, but this reduction is bound to cause a degradation in P_e if not in $\overline{N}$. We conjecture that an asymmetric channel also has an R_{cutoff} below which $\overline{N}$ is proportional to some power of n. The treatment of this problem is considerably more difficult than the treatment in Chapter IV, due to the correlation between d(X) and d(X') where X is the correct sequence and X' is an element of the incorrect subset. The problem may be solvable, however, by means of a generalization of the Chernoff bound to the case of two dependent variables. Let I and I' be two random variables, with probability measure Pr[I, I'] and consider n independent trails of the pair (I, I'). Let d_n be the sum of the n values of I and d'_w be the sum of the first w values of I' where $w \leqq n$. One can bound expressions of the form $Pr[d_n \leqq D_n, d'_w \geqq D'_w]$.* The realization that this technique might be applicable occurred to the writer during the final write-up of this research and as such, any results following from this technique are not part of this report.

---

* This generalization was first suggested by Shannon in 1956. Fano was the first to apply it for $w = n$[6]. Recently, Gallager considered the case $w \neq n$.[11]

Our bounds on $\bar{N}$ have been conservative in several respects which are worthy of special mention. First of all, we estimate the average number of elements of $\{ X_w^i \}$ which satisfy the criterion, K, without regard to whether or not these elements might have been rejected at some length less than w. As a consequence of this, the computation cutoff rate, R_{cutoff}, results with $R_{cutoff} < C$. It is an interesting theoretical and practical question whether R_{cutoff} is inherent in the sequential decoding procedure or whether it arises due to imperfect analytic techniques. Secondly, no attempt has been made to make use of the information available at the receiver about the correct x_2 at the time the decision on x_1 is made. Clearly, if a sequence, $X_n = (x_1, \ldots, x_n)$ satisfies criterion K_j at all lengths, and no other element of $\{ X_n \}$ satisfies K_{j-1}, then we decide on x_1. However, x_2 is very likely to be correct. How best to exploit this characteristic to reduce decoding complexity without deteriorating the error exponent is a question worthy of further study.

Although we have taken the effective constraint length of the encoding operation to be equal to the decoding delay, our derivations require only that the constraint length be greater than or equal to the decoding delay. An increase in the effective constraint length requires only a modest increase in encoding and decoding complexity if the decoding delay is kept constant. These considerations suggest that it might be profitable to use variable decoding delay as a parameter. When the channel is behaving normally, we may expect the correct sequence to satisfy a low criterion, perhaps K_1 or K_2. However, when the channel is especially noisy, the smallest criterion satisfied by the correct sequence, K_j, will be such that $j >> 1$. This suggests that it might be profitable to use smaller decoding delays with lower criteria. This point of view should be expecially applicable to time varying channels. These considerations are worthy of further study.

Considerable experimental work is in order to properly evaluate and understand the full potentialities of sequential decoding. This work naturally divides into two phases: first, a general purpose digital computer should be programmed to perform the decoding described in Chapter III. We may expect from this work an estimate of the mean and variance of the number of decoding computations. Further, this work is bound to suggest modifications to the decoding procedure which tend to minimize the decoding complexity. Horstein[10] has made a start in this direction. The second phase of the work should concern itself with the logical design of a real-time digital computer especially adapted to sequential decoding so that the physical hardware requirements may be more fully appreciated. It is only after several special-purpose computers are constructed for various channels, delays and rates that sequential decoding can take a natural place in the set of techniques available to the communications systems designer.

APPENDIX A

PROOF OF THEOREMS

We have stated in Chapter II that Theorems 1-3 hold for $D = p^k$ with p prime and k a positive integer. However, we have described the encoding scheme assuming that D was prime (k=1), so that it may be understood by those readers not familiar with the theory of finite fields. In keeping with this intent, the proofs below will assume that D is prime. However, the proofs are easily generalized for $k > 1$ by making use of the characteristics of the finite field $GF(p^k)$ of order $D = p^k$. This field is isomorphic to the set of polynomials in one variable of degree less than k with coefficients in Z_p. Addition of field elements is accomplished by naturally adding their corresponding polynomials, with coefficients adding mod p. Multiplication of field elements is accomplished by expressing the natural product of the corresponding polynomials mod the irreducible polynomial of order k, a root of which generates $GF(p^k)$.

To modify the encoding procedure and the following proofs for $k > 1$, x^ℓ, g_i, and f_j are chosen as elements of $GF(p^k)$ for all i, j, ℓ. The convolution operation $X * G$ and the addition operation $Y \oplus F$ is performed mod the irreducible polynomial. The proof of the theorems below hold for $k > 1$ if the following phrase changes are made:

Notation for k = 1 (D prime)	Notation for k > 1
Z_D	GF (D)
mod D	mod irreducible polynomial
integer	element of GF (D)
D is prime	GF (D) is a field

Theorem 1: Condition 1 is satisfied if the elements of $\{F\}$ are equally likely.

Proof: Regardless of which Θ map is used, $\Pr[u_t^i] = p_i$ if the D possible z_t integers are equally likely. Since $z_t = y_t + f_t \pmod D$,

$$\Pr[z_t^k] = \sum_i \Pr[f_t^i] \Pr[y_t = z_t^k - f_t^i \pmod D]$$

If the elements of $\{F\}$ are equally likely, $\Pr[f_t^i] = \frac{1}{D}$ for all i and t. Furthermore,

$$\sum_i \Pr[y_t = z_t^k - f_t^i \pmod D] = 1$$

Thus $\Pr[z_t^k] = \frac{1}{D}$ for all t and k.

Theorem 2: Condition 2 is satisfied if the elements of $\{F\}$ are equally likely.

Proof: Regardless of which Θ map is used, condition 2 is satisfied if all Z sequences of length w are equally likely. Let a Z sequence under consideration be $Z = (z_1, \ldots, z_w)$. Then $Z = Y \oplus F$, $Y = (y_1, \ldots, y_w)$ and $F = (f_1, \ldots, f_w)$. We have already defined the notation $\oplus$ to mean term by term addition mod D. Now,

$$\Pr[Z] = \sum_{\{F\}} \Pr[F] \Pr[Y = Z \ominus F]$$

where the symbol $\ominus$ denotes term by term subtraction mod D. Since $w \leqq n$ by condition 2, all the elements of $\{F\}$ are equally likely: $\Pr[F] = \frac{1}{D^w}$. Furthermore,

$$\sum_{\{F\}} \Pr[Y = Z \ominus F] = 1.$$

Thus $\Pr[Z] = \frac{1}{D^w}$, i.e., all Z sequences of length $w \leqq n$ are equally likely.

Theorem 3: Condition 3 is satisfied for any pair (X, X') where $x_t = x'_t$ for $t \leqq 0$ and $x_1 \neq x'_1$ if

(a) the elements of $\{F\}$ are equally likely, and

(b) the elements of $\{G\}$ are equally likely.

Proof: Clearly U and U' will be independent if the sequences $Z = (z_1, \ldots, z_n)$ and $Z' = (z'_1, \ldots, z'_n)$ are independent. Suppose the pair (X, X') is fixed. The space of pairs $\{Y, Y'\}$ has a probability measure induced on it by the measure on $\{G\}$ and possibly dependent on the pair (X, X').

Now

$$\Pr[Z, Z'] = \sum_{i=1}^{D^n} \Pr[Z, Z', F_i]$$

$$= \sum_{i=1}^{D^n} \Pr[Z, Z' | F_i] \Pr[F_i] \qquad \text{(A.1)}$$

If all elements of $\{F\}$ are equally likely, $\Pr[F_i] = D^{-n}$. Further, a fixed pair (Z, Z') and F_i uniquely determine the pair (Y_i, Y'_i) where $Y_i = Z \ominus F_i$, $Y'_i = Z' \ominus F_i$. Thus,

$\Pr[Z, Z' | F_i] = \Pr[Y_i, Y'_i]$ and Equation (A.1) may be rewritten as:

$$\Pr[Z, Z'] = D^{-n} \sum_{i=1}^{D^n} \Pr[Y_i, Y'_i] \qquad \text{(A.2)}$$

Consider the sum $\sum_{i=1}^{D^n} \Pr[Y_i, Y'_i]$. This is a sum of probabilities of D^n points in the space $\{Y, Y'_i\}$ which contains D^{2n} points. The pairs $\{Y_i, Y'_i\}$ have one thing in common: they have the same difference, i.e., $Y_i \ominus Y'_i = Z \ominus Z' = \hat{\eta}$.

That is, $\hat{\eta}$ is uniquely defined by the pair (Z, Z') under consideration. Since in the space { Y, Y'} there are exactly D^n points satisfying the condition $Y \ominus Y' = \eta$, the sum $\sum_{i=1}^{D^n} \Pr [Y_i, Y_i']$ may be replaced by the single term $\Pr [\hat{\eta}]$. In other words we have defined a new space { η} containing D^n points. We propose to show that for a fixed (X, X'), the elements of { η} will be equally likely if the elements of { G} are equally likely.

Now if $Y = X * G$ and $Y' = X' * G$, $\eta = Y - Y' = X * G - X' * G = (X-X') * G$. The last equation follows from the way the convolution operation $*$ was defined and the rules of modular arithmetic. Define $\xi = X \ominus X'$ where $\xi_t = x_t - x_t'$. Now $\xi_t = 0$ for $t \leqq 0$ by hypothesis. Also $\xi_1 \neq 0$ since $x_1 \neq x_1'$ by hypothesis. We may write the equation $\xi * G = \eta$ in matrix form with η and G taken as column vectors:

$$\begin{bmatrix} \xi_1 & & & & & \\ \xi_2 & \xi_1 & & & & \\ \vdots & \ddots & \ddots & & & \\ & & & & & \\ \xi_n & & \xi_2 & \xi_1 & & \end{bmatrix} \begin{bmatrix} g_0 \\ \vdots \\ \\ \\ \\ g_{n-1} \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \vdots \\ \\ \\ \\ \eta_n \end{bmatrix}$$

The ξ matrix of Equation (A. 3) is triangular. Thus its determinant is the product of the major diagonal terms and equals ξ_1^n. Since D is prime and $\xi_1 \neq 0$, $\xi_1^n \neq 0$ and the ξ matrix is non-singular. Thus η uniquely specifies G and G uniquely specifies η. For the ξ matrix defined by (X, X'), suppose

$\eta = \hat{\eta}$ when $G = \hat{G}$. Thus $Pr[\eta] = Pr[\hat{G}]$. Since all elements of $\{G\}$ are equally likely be hypothesis, $Pr[\hat{\eta}] = D^{-n}$.

Combining our results we have

$$Pr[Z, Z'] = D^{-n} \sum_{i=1}^{D^n} Pr[Y_i, Y_1']$$

$$= D^{-n} Pr[\hat{\eta}] = D^{-n} Pr[\hat{G}] = D^{-2n} \qquad (A.4)$$

From the hypothesis of Theorem 3, we know from Theorem 2 that $Pr[Z] = Pr[Z'] = D^{-n}$. Thus:

$$Pr[Z] Pr[Z'] = D^{-2n} \qquad (A.5)$$

Combining Equations (A.4) and (A.5) we obtain the desired result:

$$Pr[Z, Z'] = Pr[Z] Pr[Z'],$$

the condition for independence of Z and Z'.

Theorem 4: Let C_1 be a discrete memoryless channel with probability $q_i(j)$ of receiving the j^{th} output when the i^{th} input is used, $(i=0,\ldots, a-1; j=0,\ldots, b-1)$. Then, C_1 includes a channel C_2 with the following properties:

1. C_2 is symmetric at both input and output
2. C_2 has υ inputs, where $\upsilon = \min(a, b)$
3. The corresponding pairs of pure channels (T_α, R_α) are υ in number and are equally likely, i.e., $q_\alpha = \frac{1}{\upsilon}$ for all α
4. If C_1 has non-zero capacity, then C_2 has non-zero capacity.

Proof: If $a > b$, we can reach the capacity of C_1 by using only b of the inputs. If $a < b$, we may place C_1 in tandem with a pure post-channel which connects

a-1 of the outputs of C_1 to a-1 outputs of the post-channel. The remaining b-a+1 outputs of C_1 are connected to the a^{th} output of the post-channel. Thus we have defined a channel included in C_1 with a inputs and a outputs. The post-channel can always be chosen so that if C_1 has non-zero capacity, the included channel has non-zero capacity. Thus without loss of generality, we may assume for the purposes of the following that C_1 has υ inputs and υ outputs, where $\upsilon = \min(a, b)$. Consider the υ by υ matrix $[T]$ with entries t_{ij}, $(i = 0, \ldots, \upsilon-1; j=0, \ldots, \upsilon-1)$ such that $t_{0,\,\upsilon-1} = 1$, $t_{i+1,\,i} = 1$ for $0 \leqq i \leqq \upsilon - 2$, and $t_{ij} = 0$ otherwise. In other words, all the t_{ij} are zero except those below the major diagonal and $t_{0,\,\upsilon-1}$, all of which are unity. It is easy to show that $[T]$ is an orthogonal matrix whose transpose is its inverse, i.e., $[T]_t = [T]^{-1}$.

Pre-multiplication of $[C_1]$ by $[T]$ performs a cyclic permutation of the rows of $[C_1]$; post-multiplication of $[C_1]$ by $[T]$ performs a cyclic permutation of the columns of $[C_1]$. In both cases, the i^{th} row/column becomes the $(i+1)^{\text{st}}$ row/column, where i+1 is expressed mod υ.

Let $[C_1]$ be partitioned as shown in Equation (A. 6)

$$[C_1] = \left[\begin{array}{c|c} C_{11} & C_{12} \\ \hline C_{21} & C_{22} \end{array}\right] \tag{A.6}$$

where C_{12} is the column vector $(q_0(\upsilon-1), \ldots, q_{\upsilon-2}(\upsilon-1))$

C_{21} is the row vector $(q_{\upsilon-1}(0), \ldots, q_{\upsilon-1}(\upsilon-2))$

$C_{22} = q_{\upsilon-1}(\upsilon-1)$

In view of the effects of pre-multiplication by $[T]$ and post-multiplication by $[T]_t$, we may write

$$[T]\,[C_1]\,[T]_t = \left[\begin{array}{c|c} C_{22} & C_{21} \\ \hline C_{12} & C_{11} \end{array}\right] \tag{A.7}$$

Let $p_{1i}(j)$ be the entry in the i^{th} row and j^{th} column of the matrix $[T]\,[C_1]\,[T]_t$. Then Equation (A.7) states that

$$p_{1i}(j) = q_{i+1}\,(j+1) \tag{A.8}$$

where $i + 1$ and $j + 1$ are taken mod υ.

For α a non-negative integer, we may define the matrix operation $[T]^{\alpha}\,[C_1]\,[T]^{\alpha}_t$ whose entry in the i^{th} row and j^{th} column we call $p_{\alpha i}(j)$. $[T]^0$ is taken as the υ by υ identity matrix. By an argument similar to the preceding, we may show

$$P_{\alpha i}(j) = q_{i+\alpha}\,(j+\alpha) \tag{A.9}$$

where $i + \alpha$ and $j + \alpha$ are taken mod υ.

Define the channel C_2 with transition probability matrix $[C_2]$ with entries $p_i(j)$ as:

$$[C_2] = \frac{1}{\upsilon} \sum_{\alpha=0}^{\upsilon-1} [T]^{\alpha}[C_1][T]^{\alpha}_t \tag{A.10}$$

Using the notation of Chapter V, section A, we have defined the channel C_2 included in C_1 by means of pure pre-channels T_α with matrix $[T]^{\alpha}$ and pure post-channels R_α with matrix $[T]^{\alpha}_t$. The index α runs from zero to $\upsilon - 1$ and $g_\alpha = \frac{1}{\upsilon}$ for all α.

From Equations (A.9) and (A.10)

$$p_i(j) = \frac{1}{\upsilon} \sum_{\alpha=0}^{\upsilon-1} p_{\alpha i}(j) = \frac{1}{\upsilon} \sum_{\alpha=0}^{\upsilon-1} q_{i+\alpha}\,(j+\alpha) \tag{A.11}$$

For any integer k, $0 \leqq k \leqq \upsilon - 1$,

$$p_{i+k}(j+k) = \frac{1}{\upsilon} \sum_{\alpha=0}^{\upsilon-1} q_{i+k+\alpha}(j+k+\alpha)$$

$$= \frac{1}{\upsilon} \sum_{\alpha=k}^{k+\upsilon-1} q_{i+\alpha}(j+\alpha) = p_i(j) \qquad \text{(A. 12)}$$

Equation (A. 12) indicates that channel C_2 is symmetric at input and output.

The capacity of the symmetric channel C_2 will be obtained when the inputs are equally likely. This will induce the outputs to be equally likely. Thus the capacity is given by Equation (A. 13):

$$\text{Capacity of } C_2 = \log \upsilon + \sum_{j=0}^{\upsilon-1} p_i(j) \log p_i(j) \qquad \text{(A. 13)}$$

Note that due to the symmetry of C_2, the summation of Equation (A. 13) is independent of the input i. However,

$$-\sum_{j=0}^{\upsilon-1} p_i(j) \log p_i(j) \leqq \log \upsilon \qquad \text{(A. 14)}$$

with equality if and only if $p_i(j) = \frac{1}{\upsilon}$ for all j. But if $p_i(j) = \frac{1}{\upsilon}$ for all j, C_1 has zero capacity. Thus if C_1 has non-zero capacity, the right hand side of Equation (A. 13) is positive, and C_2 has non-zero capacity.

It is significant to note that we have completed the proof of Theorem 4 with an arbitrary ordering of the inputs and outputs of C_1. If we were going to construct C_2 from C_1, as described above, we would chose an ordering which would maximize the capacity of C_2.

Theorem 5: If a discrete memoryless channel is symmetric at input and output, then $C_0 \geqq - \mu_1(-1)$.

Proof: Our starting point is a result of Shannon( ):

$$C_0 \geqq - \log \min_{p_i} \sum_{i,j} A_{ij} p_i p_j \tag{A.15}$$

where p_i is the probability of using the i^{th} input letter and A_{ij} is the connectivity matrix of the channel, i.e., if $q_i(k)$ is the transition probability of sending the i^{th} input and receiving the k^{th} output, $A_{ij} = 1$ if there exists an output k such that $q_i(k) \neq 0$, $q_j(k) \neq 0$. Otherwise $A_{ij} = 0$.

From Equation (A.15) we observe

$$C_0 \geqq - \log \sum_{i,j} A_{ij} p_i p_j \tag{A.16}$$

where p_i is any input probability measure.

For a channel symmetric at input and output, there are an equal number, say a, of inputs and outputs. We say that input i is connected to output j if $q_i(j) \neq 0$. If input i is connected to k outputs, then it must be connected to k inputs, since the channel is symmetric. Further, all inputs must be connected to k inputs.

If we take $p_i = \frac{1}{a}$ for all i,

$$\sum p_i p_j A_{ij} = \frac{1}{a^2} ka = \frac{k}{a} \tag{A.17}$$

Defining $r_j = \sum_i p_i q_i(j)$, we see that $r_j = \frac{1}{a}$ for all j if $p_i = \frac{1}{a}$ for all i.

Consider the expression $\sum_{\{ i,j \mid q_i(j) \neq 0 \}} p_i r_j$. We have already observed that for each i, $q_i(j) \neq 0$ for k values of the index j. Thus,

$$\sum_{\{ i,j \mid q_i(j) \neq 0 \}} p_i r_j = \frac{1}{a^2} \cdot ka = \frac{k}{a} \tag{A.18}$$

Comparison of Equations (A.17) and (A.18) shows

$$\sum_{i,j} p_i p_j A_{ij} = \sum_{\{ i,j \mid q_i(j) \neq 0 \}} p_i r_j \tag{A.19}$$

$$\text{Thus } C_0 \geqq -\log \sum_{\{ i,j \mid q_i(j) \neq 0 \}} p_i r_j \tag{A.20}$$

But we show in Appendix B that the right hand side of Equation (A.20) is $-\mu_1(-1)$. This is the desired result.

APPENDIX B

CHERNOFF BOUNDS

Consider the problem of estimating the probability that the sum of independent and identically distributed random variables exceeds (or is less than) some value. A technique for accomplishing this has been described by Chernoff[12] and has been extensively used by workers in the field of information theory starting with Shannon[1]. The technique will be described here in the context of this research.

Suppose the discrete random variable I, $-\infty < I < \infty$, has probability measure Pr [I]. The moment generating function (m.g.f.) of I, g(σ), is:

$$g(\sigma) = \sum_{\{I\}} e^{\sigma I} \Pr[I] \tag{B.1}$$

Clearly g(σ) is a continuous function of σ with all derivatives defined for $-\infty < \sigma < \infty$.

Let ($I_1, \ldots, I_n$) be n independent occurrences of the random variable I. Form the sum

$$d_n = \sum_{k=1}^{n} I_k \tag{B.2}$$

The m.g.f. of the random variable d_n, g_n, is:

$$g_n(\sigma) = \sum_{\{I_1, \ldots, I_n\}} e^{\sigma \sum_k I_k} \Pr[I_1, \ldots, I_n] \tag{B.3}$$

From the independence condition, $\Pr[I_1, \ldots, I_n] = \prod_{k=1}^{n} \Pr[I_k]$. This causes Equation (B. 3) to reduce to

$$g_n(\sigma) = \prod_{k=1}^{n} \sum_{\{I_k\}} e^{\sigma I_k} \Pr[I_k] = [g(\sigma)]^n \tag{B.3a}$$

We are interested in the probability that d_n is greater than or equal to some value, say A. For the set $\{I_1, \ldots, I_n | d_n \geq A\}$, $e^{\sigma d_n} \geq e^{\sigma A}$ for $\sigma \geq 0$. Using this fact, we may rewrite Equation (B. 3) as:

$$g_n(\sigma) \geq e^{\sigma A} \sum_{\{I_1, \ldots, I_n | d_n \geq A\}} \Pr[I_1, \ldots, I_n] \tag{B.4}$$

Using Equation (B. 3a), Equation (B. 4) may be rewritten as:

$$\Pr[d_n \geq A] \leq e^{n\mu(\sigma) - \sigma A} \tag{B.5}$$

where we have defined $\mu(\sigma) = \log g(\sigma)$.

Equation (B. 5) is valid for all $\sigma \geq 0$. We may choose σ such that the exponent is minimized. Differentiating with respect to σ and setting the result equal to zero, we obtain:

$$\Pr[d_n \geq n\mu'(\sigma)] \leq e^{n[\mu(\sigma) - \sigma\mu'(\sigma)]}, \quad \sigma \geq 0 \tag{B.6}$$

where $\mu'(\sigma) = \frac{d\mu(\sigma)}{d\sigma}$.

A similar derivation will show:

$$\Pr[d_n \leq n\mu'(\sigma)] \leq e^{n[\mu(\sigma) - \sigma\mu'(\sigma)]}, \quad \sigma \leq 0 \tag{B.7}$$

It is easy to show that

$$\mu'(\sigma) = \sum_{\{I\}} I\, Q_\sigma[I] \tag{B.8}$$

where

$$Q_\sigma[I] = \frac{e^{\sigma I} \Pr[I]}{\sum_{\{I\}} e^{\sigma I} \Pr[I]} \tag{B.9}$$

The function $Q_\sigma[I]$ may be viewed as a "tilted" probability measure on the random variable I, a formulation introduced by Cramér[13]. The mean of I with respect to this tilted measure is $\mu'(\sigma)$.

Similarly one can show

$$\mu''(\sigma) = \frac{d^2\mu(\sigma)}{d\sigma^2} = \sum_{\{I\}} [I-\mu'(\sigma)]^2\, Q_\sigma[I] \tag{B.10}$$

That is, $\mu''(\sigma)$ is the variance of I with respect to the tilted measure. Except in the trivial case of I = constant, $\mu''(\sigma)$ is positive for $-\infty < \sigma < +\infty$. Thus $\mu'(\sigma)$ varies monotonically with σ.

The exponent in Equations (B.6) and (B.7) is $E(\sigma) = \sigma\mu'(\sigma)-\mu(\sigma)$. Considered as a function of the parameter σ, $E(\sigma)$ has a slope equal to $\sigma\mu''(\sigma)$. Thus, except for the trivial case of I = constant, $E(\sigma) \geqq 0$, $-\infty<\sigma<\infty$, with equality only at $\sigma = 0$.

The mean of I with respect to $\Pr[I]$ is given by $\mu'(0)$. Thus Equation (B.6) is applicable for $A = n\mu'(\sigma)$ greater than or equal to the mean of d_n, while Equation (B.7) is applicable for $A = n\mu'(\sigma)$ less than or equal to the mean of d_n. Further, from Equations (B.8) and (B.9) we observe:

$$\lim_{\sigma \to -\infty} \mu'(\sigma) = I_{min} \tag{B.11a}$$

$$\lim_{\sigma \to \infty} \mu'(\sigma) = I_{max} \tag{B.11b}$$

where I_{min} and I_{max} are the minimum and maximum values of the random variable I. Thus σ is uniquely defined by $\mu'(\sigma)$. The range of $E(\sigma)$ and $\mu'(\sigma)$ and their mutual relationship is shown graphically in Figure 6 .

Suppose we desire to permit I to become $-\infty$ with a positive probability but never become $+\infty$. Then the function $g(\sigma)$ is well defined only for $\sigma > 0$. As such, $g(\sigma)$ is no longer a moment generating function. However, the derivation of Equation (B. 6) will follow nevertheless and be applicable for $\sigma > 0$. Similarly if we permit I to become $+\infty$ with a positive probability, but never become $-\infty$, then the function $g(\sigma)$ is well defined only for $\sigma < 0$.

In this research, the random variable I is the mutual information between input letter i and output letter j for the channel with transition probability matrix $q_i(j)$ with the input i used with probability p_i. That is,

$$I(i;j) = \log \frac{q_i(j)}{r_j} \tag{B.12}$$

where $$r_j = \sum_i p_i q_i(j)$$

We are concerned with two probability measures on the pair (i, j):

$$Pr_1[i, j] = p_i q_i(j) \tag{B.13}$$

$$Pr_2[i, j] = p_i r_j \tag{B.14}$$

Associated with these measures are the functions $g_1(\sigma)$ and $g_2(\sigma)$ defined as in Equation (B. 1).

For non-degenerate channels ($q_i(j) \neq 0$ for all (i, j)), the random variable I(i; j) can assume only finite values and $g_1(\sigma)$ and $g_2(\sigma)$ are well defined for $-\infty < \sigma < \infty$.

In the case of degenerate channels, for any pair (i, j) such that $q_i(j) = 0$, $I(i; j) = -\infty$. With respect to Pr_1, the event will occur with zero probability. Thus $g_1(\sigma)$ is well defined for $-\infty < \sigma < \infty$. However, with respect to Pr_2, the event $I(i; j) = -\infty$ may occur with positive probability. Thus $g_2(\sigma)$ is well defined only for $\sigma > 0$.

Using Equations (B. 1), (B. 13) and (B. 14), we may write for the general case:

$$g_1(\sigma) = \sum_{\{i,j \mid q_i(j) \neq 0\}} p_i \frac{[q_i(j)]^{\sigma+1}}{[r_j]^{\sigma}} \tag{B.15}$$

$$g_2(\sigma) = \sum_{i,j} p_i \frac{[q_i(j)]^{\sigma}}{[r_j]^{\sigma-1}} \tag{B.16}$$

From Equation (B. 15) we have the useful relationship

$$\mu_1(-1) = \log g_1(-1) = \log \sum_{\{i,j \mid q_i(j) \neq 0\}} p_i r_j \tag{B.17}$$

From Equations (B. 15) and (B. 16) we have

$$g_1(\sigma) = g_2(\sigma+1), \quad \sigma > -1 \tag{B.18}$$

For non-degenerate channels, Equation (B.18) is valid for $-\infty < \sigma < \infty$. Thus $\mu_1'(-1)=\mu_2'(0)$= the mean of I with respect to Pr_2.

It is the relationship expressed in Equation (B.18) which makes the choice of the random variable $I(i;j) = \log \frac{q_i(j)}{r_j}$ so useful in this problem.

Let $\bar{I}^1$ and $\bar{I}^2$ be the means of the random variable I with respect to Pr_1 and Pr_2 respectively. We proceed to show $\bar{I}^2 \leqq 0$, with equality if and only if the channel has zero capacity.

$$\bar{I}^2 = \sum_{i,j} p_i r_j \log \frac{q_i(j)}{r_j} = -\sum_j r_j \log r_j + \sum_{i,j} p_i r_j \log q_i(j) \tag{B.19}$$

By the inequality of the geometric and arithmetic mean,[14]

$$\sum_i p_i \log q_i(j) = \log \prod_i [q_i(j)]^{p_i} \leqq \log \sum_i p_i q_i(j) = \log r_j \tag{B.20}$$

with equality if and only if the transition probabilities from every input letter to the j^{th} output letter are equal. Substituting Equation (B.20) in Equation (B.19) we have $\bar{I}^2 \leqq 0$. Equality can hold if and only if for <u>every</u> j, the transition probabilities from all inputs are equal. But this situation describes a channel with zero capacity. This yields the desired result.

In any practical coding problem, the capacity $C > \bar{I}^1 > 0$. Thus we have the situation $\bar{I}^2 < 0 < \bar{I}^1$.

In this research we have occasion to select a quantity D_n such that

$$n\bar{I}^2 < D_n < n\bar{I}^1 \tag{B.21}$$

and desire to bound the probabilities $Pr_1 [d_n \leqq D_n]$ and $Pr_2 [d_n \geqq D_n]$. For D_n selected as in Equation (B.21), there exist unique quantities $\sigma < 0$ and $\lambda > 0$ such that $D_n = n\mu_1'(\sigma) = n\mu_2'(\lambda)$. For $D_n > n\mu_1'(-1)$, as will be the case for all non-degenerate channels, we may use Equation (B.18) to observe that $\lambda = \sigma + 1$. Substituting this result in Equation (B.6), we have

$$Pr_2 [d_n \geqq D_n] \leqq e^{-n\mu_1'(\sigma)} e^{n[\mu_1(\sigma) - \sigma\mu_1'(\sigma)]} \tag{B.22}$$

Direct use of Equation (B.7) yields:

$$Pr_1 [d_n \leqq D_n] \leqq e^{n[\mu_1(\sigma) - \sigma\mu_1'(\sigma)]} \tag{B.23}$$

REFERENCES

1. Shannon, C. E., "Certain Results in Coding Theory for Noisy Channels," Information and Control 1, 6-25 (1957).

2. Shannon, C. E., "Error Probability Bounds for Noisy Channels," Unpublished Report (1959).

3. Elias, P., "Coding for Noisy Channels," I.R.E. Convention Record, Part IV (1955).

4. Wozencraft, J. M., "Sequential Decoding for Reliable Communications," Technical Report No. 325, Research Laboratory of Electronics, M.I.T. (9 August 1957).

5. Epstein, M. A., "Algebraic Decoding for the Binary Erasure Channel," Technical Report No. 340, Research Laboratory of Electronics, M.I.T. (14 March 1958).

6. Fano, R. M., "An Upper Bound on the Probability of Error for Noisy Channels," Unpublished Report (1960).

7. Shannon, C. E., "A Note on Partial Ordering for Communications Channels," Information and Control 1, 390-397 (1958).

8. Peterson, W., Class Notes for Course No. 6.575, Spring Term (1960).

9. Shannon, C. E., "The Zero Error Capacity of a Noisy Channel," I.R.E. Transactions on Information Theory, IT-2, No. 3 (1956).

10. Horstein, M., "An Experimental Study of Sequential Decoding for the Binary Symmetric Channel," Group Report 34-74, Lincoln Laboratory, M.I.T. (20 November 1958).

11. Gallager, R. G., "Low Density Parity Check Codes," Sc.D. Thesis submitted to E. E. Department, M.I.T. (August, 1960).

12. Chernoff, H., "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on a Sum of Observations," Ann. Math. Stat., 23 (1952).

13. Cramer, H., "Sur un Nouveau Theoreme-Limite de la Theorie des Probabilites," Colloque d'Octobre, 1937 sur la Theorie des Probabilites, Hermann et Cie, Paris, France (1938).

14. Hardy, Littlewood and Polya, Inequalities (Cambridge University Press, London, 1952).