CODING FOR DISCRETE MEMORYLESS TWO-WAY CHANNELS

by

Frederick Jelinek

S.B. Massachusetts Institute of Technology (1956)

S.M. Massachusetts Institute of Technology (1958)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January, 1962

Signature of Author --
Department of Electrical Engineering, January 8, 1962

Certified by --
Thesis Supervisor

Accepted by --
Chairman, Departmental Committee on Graduate Students

CODING FOR DISCRETE MEMORYLESS TWO-WAY CHANNELS

by

FREDERICK JELINEK

Submitted to the Department of Electrical Engineering on January 8, 1962 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

Discrete memoryless two-way channels having outputs $y, \bar{y}$ and inputs $x, \bar{x}$ are defined by their set of transmission probabilities $p(y,\bar{y}/x,\bar{x})$. They were first studied by Shannon[(1)] who developed expressions for inner and outer bounds of the capacity region. This thesis investigates strategy codes in which messages are mapped into functions F and $\bar{F}$, where F maps received signal sequences Y into the channel input signal sequences X, and $\bar{F}$ maps received signal sequences $\bar{Y}$ into channel input signal sequences $\bar{X}$. First a coding theorem is developed, based on a generalization of the random argument for independent assignments, extending Shannon's inner bound to the capacity region, and showing that in this extension the probability of decoding error is exponentially bounded with increasing codeword length. The expression for the extended inner bound of the capacity region is then studied, and it is shown that for two-way channels which are not noiseless, all codes involving strategy functions F and $\bar{F}$ will be such that the information passing through the channel in both directions will be strictly less than the information useable for message decoding purposes. Hence the concept of coding information loss is introduced and studied. With its help a tighter outer bound on the capacity region is developed, which, although extremely difficult to evaluate in practice, is used to prove that a certain very broad class of two-way channels has capacity regions strictly interior to Shannon's outer bound. Attention is then focused on classification of different possible cases of binary two-way channels whose transmission probabilities are restricted by the relation $P(y,\bar{y}/x\bar{x}) = p(y/x\bar{x})\ \bar{p}(\bar{y}/x\bar{x})$. It is shown that all such channels can be represented by a cascade combination of two channels one of which has steady, uniform noise, and the other has noise dependent on the identity of transmitted digits only. It is further shown that the latter channels fall into two canonic types, each requiring a different coding strategy for optimal results. Finally, certain convexity theorems are proved, and a way of quasi-optimizing the random ensemble is developed, insuring the use of certain canonical strategy functions F and $\bar{\bar{F}}$ only, thus reducing drastically the number of optimizing variables.

In the process of proving the above coding theorem, use is made of a new theorem proving uniform convergence of certain probabilities associated with Markoff sources (see Appendices III and IV). This theorem may be of mathematical interest for its own sake.

Thesis Supervisor: R. M. Fano
Title: Professor of Electrical Communications.

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to my thesis supervisor, Professor Robert M. Fano for his patient supervision and direction. Without his constant encouragement this work could not have been completed.

The original work of Professor Claude E. Shannon on two-way channels provided the foundation for this work. To him I wish to extend my appreciation for making available to me his unpublished notes on which Appendix II is based and for the helpful discussions we have held on many aspects of the problem.

Finally, I would like to take this opportunity to thank my foremost Teacher, Professor Ernst A. Guillemin who has inspired and guided me throughout my studies at M.I.T.

This work was done with the help of the facilities of and financial support from the Research Laboratory of Electronics. It is a pleasure for me to acknowledge this aid.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

1. Introduction

1.1 Definition of the Discrete Memoryless Two-way Channel

A discrete, memoryless two-way channel, shown schematically in Figure 1-1, can be described as follows:

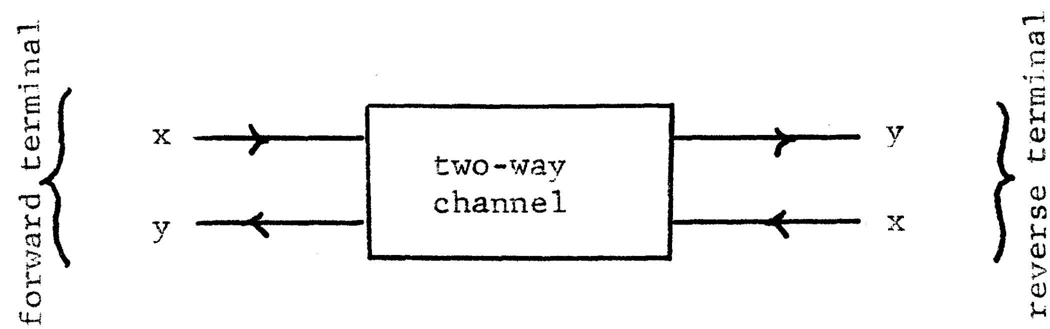


Figure 1-1

The channel consists of two terminals, each of which is provided with an input and an output. The input signal x at the forward terminal is selected from a finite discrete alphabet of $\underline{a}$ symbols. The output y at the same terminal belongs to a discrete alphabet of $\underline{b}$ symbols. The input signal $\bar{x}$ and the output signal $\bar{y}$ at the "reverse" terminal are selected from discrete alphabets of size $\underline{\bar{a}}$ and $\underline{\bar{b}}$, respectively. The channel operates synchronously: at given time intervals inputs x and $\bar{x}$ are chosen and transmitted simultaneously through the channel; outputs y and $\bar{y}$ may then be observed. The definition of the channel is completed by a set of transmission probabilities $\{P(y,\bar{y}/x,\bar{x})\}$. Since the channel is assumed to be memoryless, it will have the property that:

$$Pr(y_1,y_2,\ldots,y_n;\ \bar{y}_1,\bar{y}_2,\ldots,\bar{y}_n/x_1,\ldots,x_n;\ \bar{x}_1,\ldots,\bar{x}_n) = \prod_{i=1}^{n} P(y_i\bar{y}_i/x_i\bar{x}_i) \qquad (1\text{-}1)$$

The most important feature of this channel is the nature of its noise: transmission in one direction interferes with the transmission in the opposite direction. It is thus clear from the start that in an efficient coding scheme the signals to be transmitted will be determined by taking into consideration the identity of past signals received, the latter providing information about the transmission in the opposite direction.

1.2. Examples of Two-way Channels

A trivial example of a two-way channel are two one-way channels transmitting in opposite directions. This instance lacks the essential feature of noise generated by opposite transmission. A more interesting example is pictured in Figure 1-2.

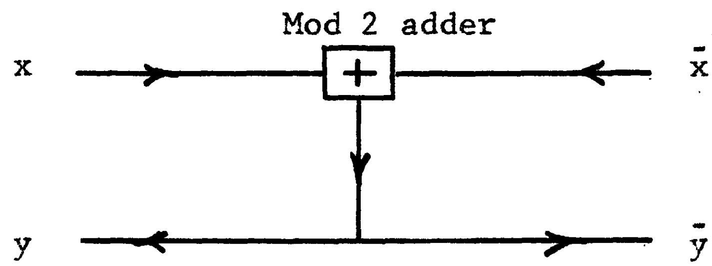


Figure 1-2.

Here inputs and outputs are binary and the channel operation is defined by $y=\bar{y}=x+\bar{x}$(mod 2). Shannon (1) points out that it is possible to transmit one bit per second in each direction simultaneously. "Arbitrarily binary digits may be fed in at x and $\bar{x}$ but, to decode, the observed y's must be corrected to compensate for the influence of the transmitted x. Thus an observed y should be added to the just transmitted x (mod 2) to determine the transmitted $\bar{x}$."

Another example, again a "noiseless" one (in the sense that y and $\bar{y}$ are fully determined by the transmitted pair $x,\bar{x}$), is the binary multiplying channel of Figure 3, suggested by Blackwell

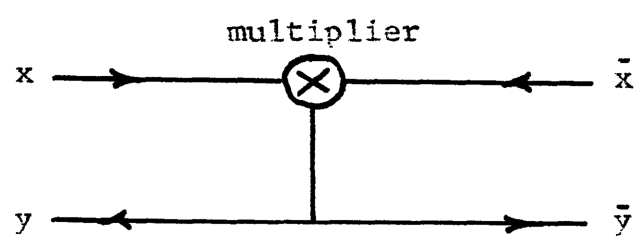


Figure 1-3.

All inputs and outputs are here binary, and the channel operation is $y=\bar{y}=x\bar{x}$. The rate of information transmission in one direction depends on the information rate in the opposite direction: it is clear that one may transmit one bit per signal in the forward direction if $\bar{x}$ is permanently fixed to equal 1, while it is impossible to transmit any information at all if $\bar{x}$ is 0. The region of approachable information rate-pairs is computable by a formula developed later in the thesis.

A final example of a noisy channel is given in Figure 1-4.

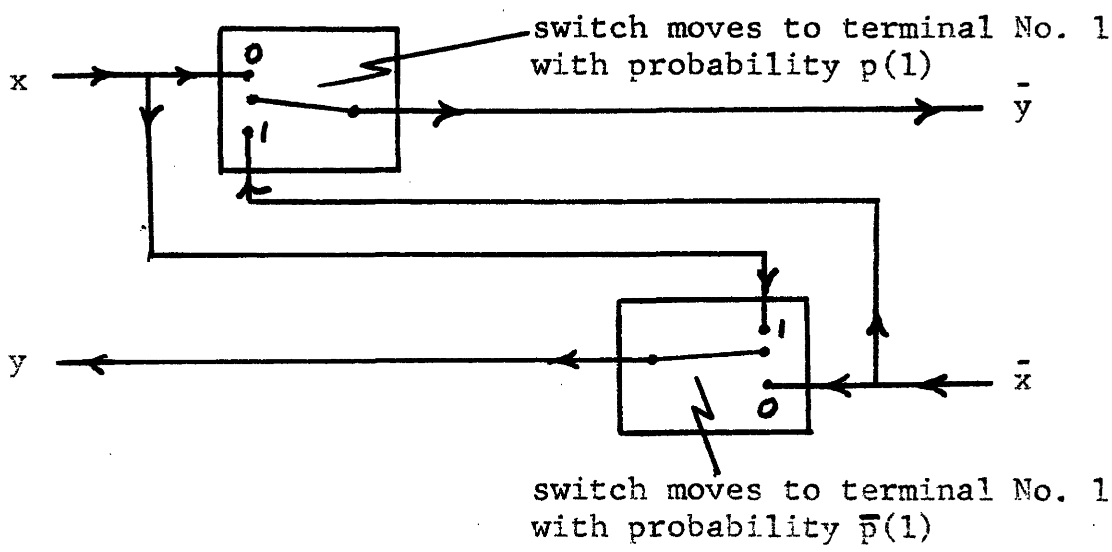


Figure 1-4.

The inputs are again binary. It is clear that with probability $p(1), \bar{y} = \bar{x}$, and with probability $1-p(1)$, $\bar{y} = x$. Hence the channel is a sort of an erasure channel with the important distinction that the receiver does not know when an erasure occurred, although it does occasionally know when an erasure did not occur: whenever $\bar{y} \neq \bar{x}$. In Figure 1-4 one cannot fix the signal x so that the forward transmission would be one bit per signal. The exact region of approachable rate pairs for this very complicated channel is not known, although this thesis develops a formula for computing tight inner and outer bounds for it. It should perhaps be pointed out that the channel of Figure 1-4 is of central importance. It will be shown in Article 8 that it possesses one of two possible cannonical forms.

1.3 A Real Life Example

An example of the concept of the two-way channel is provided by the two-way microwave link with a repeater represented by Figure 1-5.

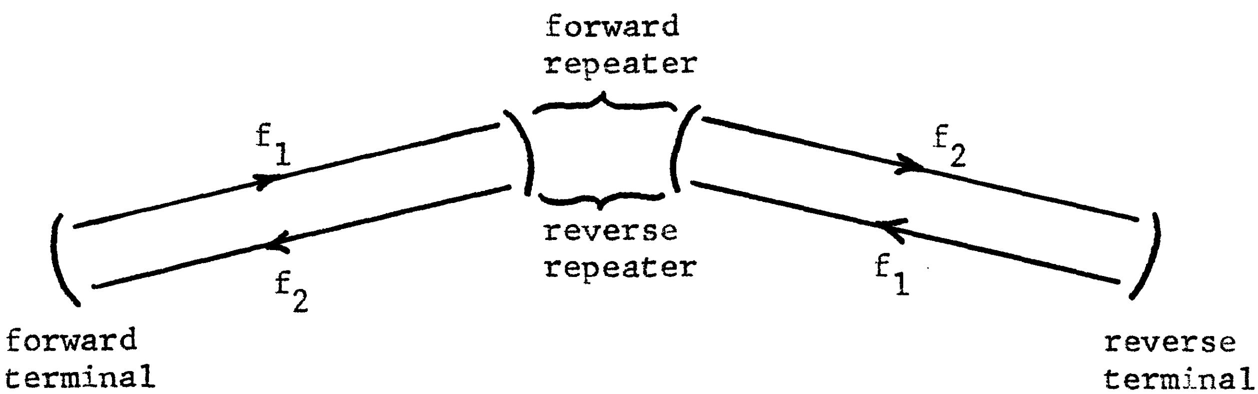


Figure 1-5.

The forward terminal transmits in the frequency band f_1 and its signal is being reshaped by the forward repeater and retransmitted in the frequency band f_2. Conversely, the reverse terminal transmits in the band f_1 and its signal is being reshaped by the reverse repeater and retransmitted in the frequency band f_2. The forward repeater's antenna, although directional, picks up some of the reverse terminal's signal and transmits it at frequency f_2. Hence the reverse terminal's reception depends also on its own transmission. This effect is being compounded by the fact that the reverse repeater's transmitting antenna, although directional, is not perfect and hence that its signal is being picked up, however weakly, by the reverse terminal's receiver. A similar situation exists with respect to the reception at the forward terminal.

2. Shannon's Results

In his pioneering work on two-way channels[1] Shannon derived important results which will be briefly summarized in this article. He has shown that "for a memoryless discrete channel there exists a convex region G of approachable rates. For any point in G, say $(R,\bar{R})$, there exist codes signalling with rates arbitrarily close to the point and with arbitrarily small error probability. The region G is of the form shown in the middle curve of Figure 2-1.

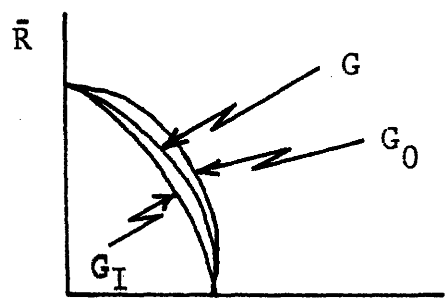


Figure 2-1.

This curve can be described by a limiting expression involving mutual informations for long sequences of inputs and outputs. In addition", Shannon found an "inner and outer bound, G_I and G_0, which are more easily evaluated, involving, as they do, only a maximizing process over single letters of the channel."

2.1 Shannon's Inner Bound G_I

The inner bound is found by use of a random coding argument concerning codes which associate messages strictly to transmitted sequences regardless of the identity of the received sequence. The manner of association is the same as the one used for one-way discrete memoryless channels: in the ensemble a message m is associated to a signal sequence $x_1, x_2, \ldots, x_n$ with probability $p(x_1)p(x_2)\ldots p(x_n)$. Decoding of (e.g.) the reverse message at the forward terminal when sequence $y_1, \ldots, y_n$ was received and sequence $x_1, \ldots, x_n$ was transmitted is done by maximizing the probability $Pr(\bar{x}_1, \ldots \bar{x}_n / y_1, \ldots, y_n, x_1, \ldots, x_n)$ over the signal sequences $\bar{x}_1, \ldots, \bar{x}_n$ associated with the various messages of the reverse terminal.

The inner bound region G_I is then defined by Shannon in his Theorem 3 which we quote here in full:

"<u>Theorem 3</u>

Let G_I be the convex hull of points $(R, \bar{R})$

$$R = E\left\{\log \frac{Pr(x/\bar{x},\bar{y})}{P(x)}\right\}$$

$$\bar{R} = E\left\{\log \frac{Pr(\bar{x}/x,y)}{\bar{P}(\bar{x})}\right\}$$

when $P(x)$ and $\bar{P}(\bar{x})$ are given various probability assignments. All points of G_I are in the capacity region. For any point $(R, \bar{R})$ in G_I and any $\varepsilon > 0$ we can find points whose signalling rates are within ε of R and $\bar{R}$ and whose decoding error probabilities in both directions are less than $e^{-A(\varepsilon)n}$ for all sufficiently large n."

Shannon further remarks: "It may be noted that the convex hull G_I in this theorem is a closed set containing all its limit points". "Furthermore, if G_I contains a point $(R,\bar{R})$, it contains the projections $(R,0)$ and $(0,\bar{R})$."

2.2 Shannon's Outer Bound G_0

Shannon gives a partial converse to Theorem 3 quoted above and an outer bound G_0 to the capacity region. This he obtains not by an argument directly connected to some coding scheme, but rather by examining the change

$$\Delta = H(m/\bar{m},\bar{y}_1,\ldots,\bar{y}_{i-1})-H(m/\bar{m},\bar{y}_1,\ldots,\bar{y}_{i-1},\bar{y}_i)$$

in the uncertainty about the forward message m due to the reception of the signal $\bar{y}_i$ when previously signals $\bar{y}_1,\ldots,\bar{y}_{i-1}$ were received and message m was being transmitted at the reverse terminal. He shows that

$$\Delta \leq E(\log \frac{Pr(x_i/\bar{x}_i\bar{y}_i)}{P(x_i/\bar{x}_i)})$$

and concludes that "the capacity region G is included in the convex hull G_0 of all points $(R,\bar{R})$

$$R = E\ (\log \frac{Pr(x/\bar{x},\bar{y})}{P(x/\bar{x})})$$

$$\bar{R} = E\ (\log \frac{Pr(\bar{x}/x,y)}{P(\bar{x}/x)})$$

when arbitrary joint probability assignments $P(x,\bar{x})$ are made." I.e. whenever a code signals with forward rate R^* and reverse rate $\bar{R}^*$ where the point $(R^*,\bar{R}^*)$ lies outside the region G_0, the probability of decoding error will at least in one direction be bounded away from zero, independent of the code length n.

Shannon further shows that the information rates given above are concave downward functions of the assigned input probabilities $P(x,\bar{x})$, and develops conditions on the transmission probability set $\{P(y,\bar{y}/x,\bar{x})\}$, whose fulfillment would guarantee the coincidence of the regions G_I and G_0. Finally, Shannon proves that bounds of G_I and G_0 intersect the coordinate axes at identical points.

2.3 Shannon's Capacity Region Solution G

Shannon considers the product two-way channel K_n whose inputs and outputs are sequences of n signals of the given two-way channel. An input "strategy letter" to the channel K_n at the forward terminal may be represented by

$$[x_o, f_1, (x_0; y_0), f_2(x_0, x_1; y_0, y_1), \ldots, f_{n-1}(x_0, x_1, \ldots, x_{n-2}; y_0, y_1, \ldots, y_{n-2})] \equiv X^n$$

where the functions f_i map pairs of i-tupples $[x_0, \ldots, x_{i-1}; y_0, \ldots, y_{i-1}]$ into signals x_i. I.e there is a mapping relationship

$$x_i = f_i\ (x_0, \ldots, x_{i-1};\ y_0, \ldots, y_{i-1})$$

for all possible pairs of i-tupples. A reverse strategy letter may similarly be represented by

$$[x_0, f_1(x_0; y_0), f_2(x_0, x_1; y_0, y_1), \ldots, f_{n-1}(x_0, x_1, \ldots, x_{n-2}; y_0, y_1, \ldots, y_{n-2})] \equiv \bar{X}^n.$$

One may then in a straightforward manner define the product channel's transmission probabilities

$$\Pr\ (Y^n, \bar{Y}^n / X^n, \bar{X}^n)$$

where $\quad Y^n \equiv y_0, y_1, \ldots, y_{n-1} \quad \bar{Y}^n \equiv \bar{y}_0, \bar{y}_1, \ldots, \bar{y}_{n-1}$

and evaluate the inner bound G_I^n for the capacity region of the channel K_n by finding the convex hull of points $(R^n, \bar{R}^n)$, where

$$R^n = \frac{1}{n}\ E\left\{\log \frac{\Pr(X^n / \bar{Y}^n\ \bar{X}^n)}{P(X^n)}\right\}$$

$$\bar{R}^n = \frac{1}{n} \; E \left\{ \log \frac{Pr(\bar{X}^n / Y^n \, X^n)}{\bar{P}(\bar{X}^n)} \right\},$$

over the various probability assignments $P(X^n)$ and $\bar{P}(\bar{X}^n)$.

Shannon next shows that $G_I^n \leq G_I^{n+r}$ and that the capacity region

$$G = \lim_{n \to \infty} G_I^n \quad .$$

3. Proposed Attack on the Problem

3.1 Objectives of Two-way Channel Investigation

As illustrated in the introduction, the two-way channel presents a new problem as yet not fully treated in information theory: How to arrange a code so that the opposing terminals cooperate in the selections of signals to be transmitted even in situations in which the messages to be communicated remain independent. It is of course possible to communicate through a two-way channel without any signal cooperation, and Shannon (1) found the region of pairs of signalling rates $(R,\bar{R})$ for the two directions which can be approached with arbitrarily small error probabilities by non-cooperating codes (see Article 2). One of our major objectives would be to see whether or not the above capacity region can be extended by the use of cooperating codes.

Another question to be asked concerns the construction of the cooperating codes: what kind of cooperation is desirable and effective and what are the features of the two-way channel which make it so. In this connection one might ask whether the desirable cooperation is of one kind only, or whether perhaps there are several classes of two-way channels, each class calling for a different sort of cooperation.

The cooperating codes will of course be of an as yet uninvestigated type and their structure, besides being in an obvious way a more complicated one, might present new difficulties of an essential nature not yet encountered by the treatment of the usual codes. One would expect that the increase in the complexity of the structure of the cooperating code will be the price paid for its greater effectiveness. It might then turn out that different grades of cooperation will exist, and their respective effectiveness will be of interest.

Finally, one might compare the cooperating with the non-cooperating codes in their effectiveness of limiting decoding error probability at signalling rate pairs $(R,\bar{R})$ within the capacity region of the non-cooperating codes.

3.2 Comment on Shannon's Results

A comment on Shannon's results in light of the objectives listed in the preceding section is called for at this point. The implications of the inner bound G_I described in section 2.1 are self-evident. The limiting procedure for the capacity region solution described in section 2.3 is certainly suggestive of a way of coding. However, one ought to point out that for a binarychannel K_n the number of different input letters X^n suggested by this scheme is $2^{\frac{4n-1}{4-1}}$. Better results than those of the inner bound non-cooperative coding can be obtained for values of n larger than or equal to 2, and for the latter integer the number of different input letters is 32. Hence the inner bound for K_2 involves optimization over 64 quantities. Moreover, it should be pointed out that such coding is actually quite wasteful in terms of the number of input quantities intreduced compared to the probable effect: thus for n=2 only half of the time any cooperation at all exists, and even this cooperation will most probably not be very effective, since at the next time interval it will again be interrupted. Intuitively it seems that in order for the two terminals to cooperate, the signals must get into some sort of synchronism which the repeated interruptions of cooperation will only hinder.

Now the outer bound results of section 2.2, although not directly suggesting any coding scheme, tell us that cooperation is indeed effective (Actually they only suggest the possibility that the outer bound G_0 differs from the inner bound G_I. However, it is easy to compute in specific instances, such as the multiplying channel mentioned earlier, or the cannonic channel of Figure 1-4, that this is really the case). Before going into more complicated schemes, one might wonder whether a code mapping messages strictly into signal sequences could not be arranged which would result in an average joint probability $Pr(x,\bar{x})$ of simultaneous occurrance of opposing signals x and $\bar{x}$ such that $Pr(x,\bar{x}) \neq Pr(x)\ Pr(\bar{x})$. It turns out (as shown in Appendix I) that any random assignment which maps a forward message into a sequence $x_1, x_2, \ldots, x_n$ with probability $P(x_1)P(x_2)\ldots P(x_n)$ and a reverse message into a sequence $\bar{x}_1$, $\bar{x}_2, \ldots, \bar{x}_n$ with probability $\bar{P}(\bar{x}_1)\bar{P}(\bar{x}_2)\ldots\bar{P}(\bar{x}_n)$ would on the average lead to a code for which the probability $Pr(x,\bar{x}) \rightarrow P(x)\bar{P}(\bar{x})$ as the number of different forward and reverse codewords $k \rightarrow \infty$. On reflection this destroys our hope for extending the region G_I by codes having strict message-to-signal sequence mapping.

Thus we will after all have to resort to coding in which the signal transmitted will depend on the identity of past signals transmitted and received. As will be seen, we will make use of the functions f_i which help define the strategy letters X^n in section 2.3, but we will try to exclude all waste discussed in the paragraph preceding the last one.

3.3 Restriction of Generality of Discussion

In the following work we will restrict our attention to two-way channels characterized by a set of probabilities

$$P(y,\bar{y}/x,\bar{x}) = p(y/x,\bar{x})\ \bar{p}(\bar{y}/x,\bar{x}) \qquad (3\text{-}1)$$

defined over the set of possible input signals x, $\bar{x}$ and output signals y, $\bar{y}$. Above restriction is not necessary for the proof of the key theorems presented in this thesis. In fact, the coding Theorem 6-2 and the Theorem 7-5 introducing the concept of information coding loss are independent of the restriction (3-1). On the other hand, it will be seen that the interpretation of the problems introduced by the two-way channel concept will be much clearer with the restriction (3-1) imposed. By making use of the symmetries created by (3-1) we will be able to separate the effects of transmission in one direction from the effects of transmission in the other direction.

It should be noted that the simplification effected by the restriction (3-1) is due to the fact that the received signal at one terminal will depend only on the two signals sent and not on the signal received at the other terminal.

Throughout the thesis all specific references will be to a binary two-way channel, although the signal alphabets will be unrestricted in all the proofs of the theorems presented. However, the intuitively very important results of Article 8 will apply only to binary channels with restriction (3-1) imposed on them.

Any binary channel whose transmission probabilities are restricted by (3-1) is schematically representable by Figure 3-1.

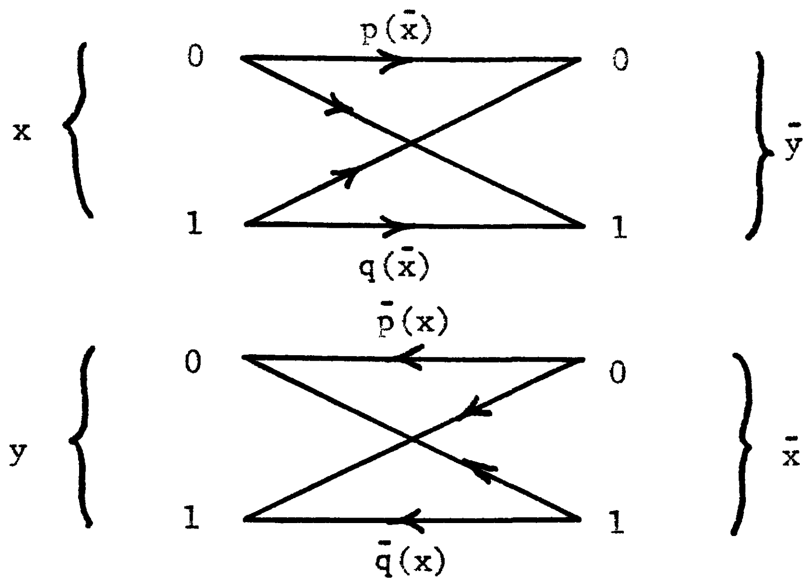


Figure 3-1.

The interpretation of the above diagram is as follows: if the reverse transmitter sends the signal $\bar{x} = 0$ (1) and simultaneously the forward **transmitter** sends the signal $x = 0$ (1), the probability that at the reverse terminal a 0 (1) will be received is $p(\bar{x})$ ($q(\bar{x})$), and the probability that a 1 (0) will be received is $1-p(\bar{x})$ ($1-q(\bar{x})$). A similar interpretation applies to the transmission in the reverse direction. Thus the identification of the symbols introduced in Figure 3-1 with the set of transmission probabilities defining the channel is as given in the table below:

$$p(\bar{x}) \equiv p(0/0,\bar{x}) \qquad \bar{p}(x) \equiv p(0/x,0)$$
$$q(\bar{x}) \equiv p(1/1,\bar{x}) \qquad \bar{q}(x) \equiv p(1/x,1) \qquad (3\text{-}2)$$

3.4 Desirability of Sequential Encoding

In the preceding discussion we have shown that optimal encoding of a message should be such that the signal transmitted in any given time interval would be determined not only by the message but also by all the signals received and transmitted in the previous time intervals.

A signal transmitted when either sequential or block encoding is used, is dependent on n-1 other signals. Hence knowledge at the forward terminal of the signals received up to time t-1 provides a probabilistic indication of the identity of the signal to be transmitted at time t by the reverse terminal. Since the latter signal will in the two-way channel determine part of the noise on the transmission at time t in the forward direction, it is essential that its probable value be considered at the time when the forward terminal's signal is being generated. This being the case, the natural encoding is sequential, since using that scheme the signal to be transmitted at time t is not yet determined at time t-1. We will therefore modify Wozencraft's (2) sequential coding procedure to the needs of two-way channel encoding. The procedure's tree-like structure will be retained, but not its code generator feature. In what follows we will not concern ourselves with the obvious decoding advantages of the sequential procedure.

3.5 State Diagram Representation of Sequential Encoding

It is assumed that the reader is familiar with Wozencraft's work entitled "Sequential Decoding for Reliable Communication" (2). In this section we will make use of its terminology.

Wozencraft determines the binary signal sequence to be transmitted through the channel by means of a generator into which a message, also in binary form, is fed. Because of the way the generator operates (for description see pp 3-4 to 3-7 of (2)), the sequence to be transmitted through the channel has the characteristic that any block of kn_0 signals is statistically constrained, while signals more than kn_0 places apart are statistically independent. The encoding is usually presented in the form of a tree, the path through which is determined by the binary message sequence. The branches of the tree are associated with sequences of n_0 binary signals which are transmitted if the message path leads through the branch. Such tree, however, does not give the complete picture of the generatable signal sequence of potentially infinite length: one would have to extend the tree indefinitely, and indeed the generator and its rules of operation give us the means to do so. On the other hand, a much more compact way to present the entire process presents itself in the form of a state diagram.

Let the state be determined by the sequence $(z_{t-k}, z_{t-k+1}, \ldots, z_{t-1}; t)$ consisting of the time interval t (given modulo k) and the last k binary message digits. For a given interval t there are 2^k states forming a "time string" (the reason for such terminology will become apparent by inspection of the state diagram given in Figure 3-3), and the state diagram consists of a total of $k2^k$ states. Transition is possible only from a state of interval t into a state of interval t+1, and moreover, from any given state $(z_{t-k}, z_{t-k+1}, \ldots, z_{t-1}, t)$ only two transitions exist: into state $(z_{t-k+1}, \ldots, z_{t-1}, 0; t+1)$ and into state $(z_{t-k+1}, \ldots, z_{t-1}, 1; t+1)$. Hence the particular transition taken is determined by the identity of the message digit z_t. Now each transition has associated with it a sequence of n_0 binary signals $x_{t,0}, x_{t,1}, \ldots, x_{t,n_0-1}$. The succession of the signal sequences then forms the encoded transmitted communication.

As an example, consider the particular Wozencraft scheme in which $kn_0=9$; $n_0=3$, k=3, and the generator sequence g=110100. Then the tree of Figure 3-2 is obtained.

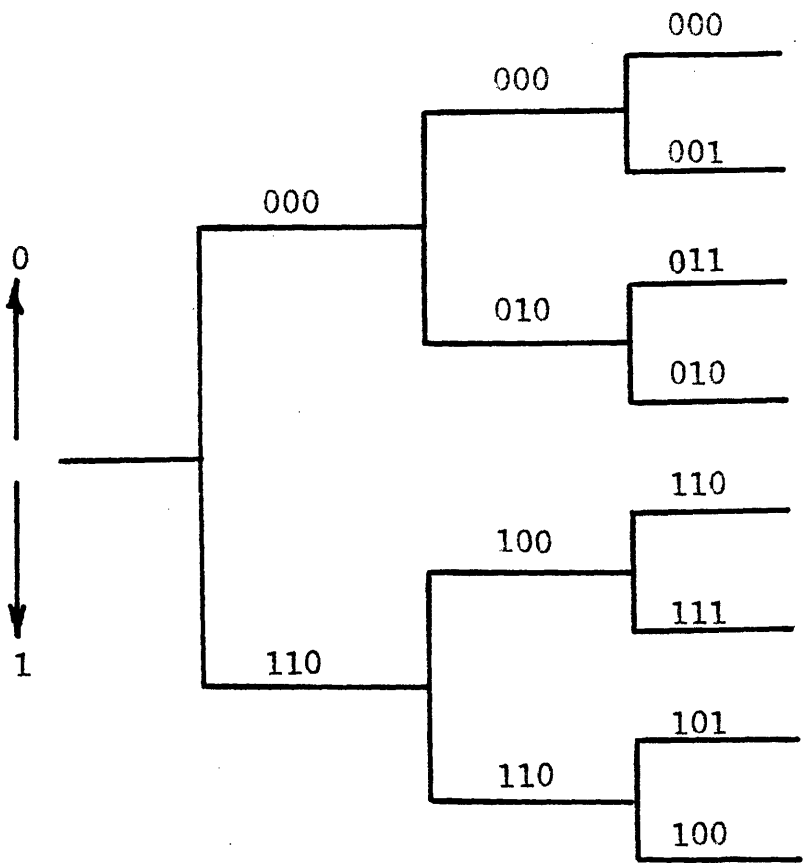


Figure 3-2.

This tree should, as explained above, be extended indefinitely. Its equivalent state diagram is shown in Figure 3-3 in which, with the exceptions of the transitions corresponding to the tree above, the transition sequence values are left out. It is hoped that this will serve to further elucidate the relationship between Figures 3-2 and 3-3.

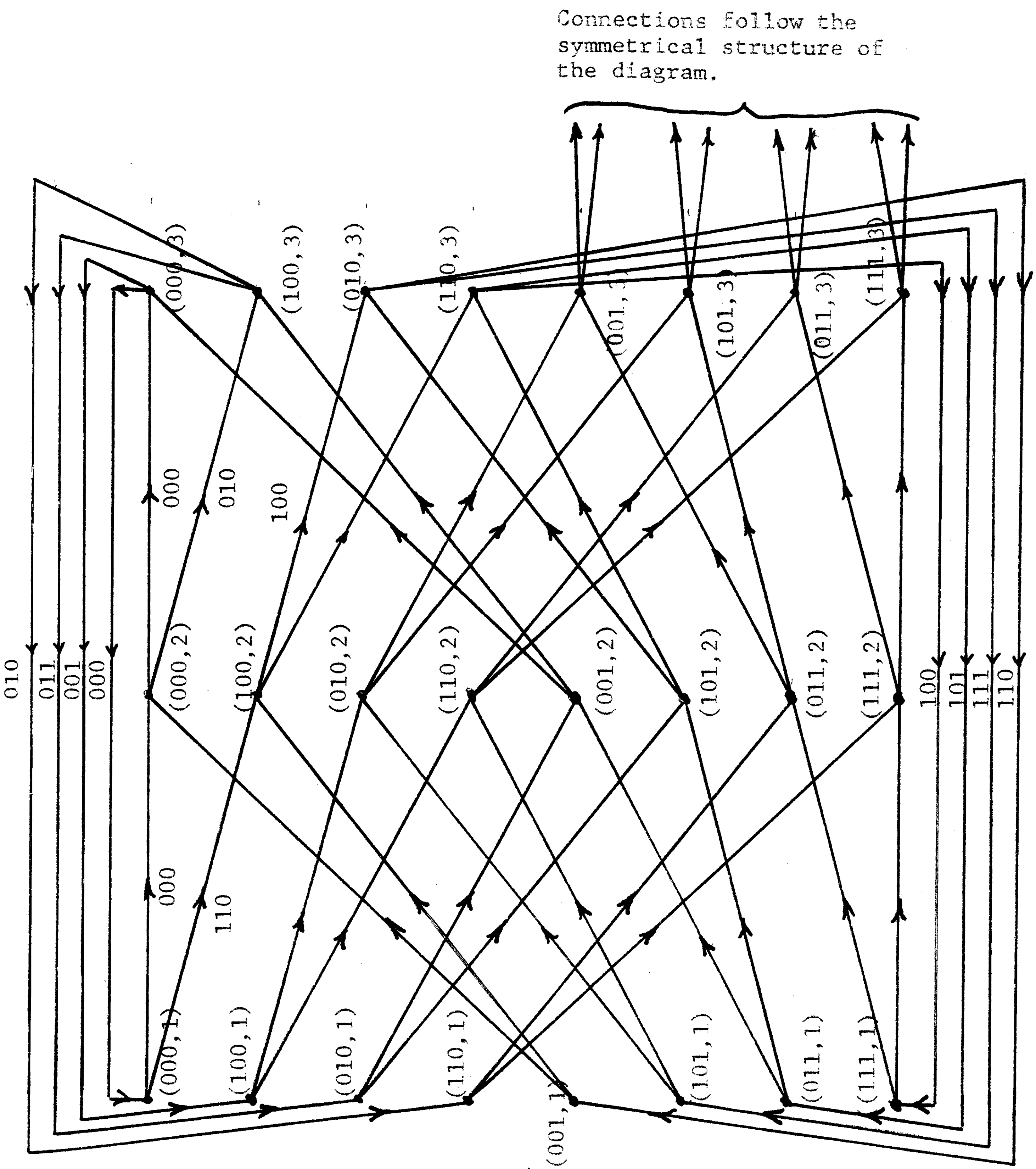

Connections follow the
symmetrical structure of
the diagram.
(000,3)
(100,3)
(010,3)
(101,3)
(011,3)
(000,2)
(100,2)
(010,2)
(110,2)
(001,2)
(101,2)
(011,2)
(111,2)
(000,1)
(100,1)
(010,1)
(110,1)
(101,1)
(011,1)
(111,1)
000
010
100
000
110
010
011
001
000
100
101
111
110

Figure 3-3.

4. Notation Used

As will be seen further on, a treatment of signal cooperating codes for the two-way channel will necessitate the introduction of new, relatively complex notation. In order that a ready reference to this notation be available throughout the reading of this paper, we consider it best to introduce it all at once.

4.1 Indexing Notation

We will be dealing with two kinds of sequences of symbols. One will pertain to sequential coding and the other to block coding. Actually we will be using the sequential approach most of the time but for certain mutual information and entropy results the block approach allows greater notational simplicity and can be used without any loss of generality.

In sequential coding we are concerned with tree branches emanating from a common "departure" coding state whose identity is assumed to be known to the decoder at the opposing terminal. If the departure state was reached at time i, then the two branches emanating from it, leading into states to be reached at time $\underline{i+1}$, will be indexed by the number $\underline{i}$. In a sequential code, to every tree branch there corresponds a sequence of symbols of length n_0. Then, for instance

$x_{i,j}$ $(j = 0,1,\ldots,n_0-1)$ will be the $j+1^{th}$ symbol (4-1)
corresponding to some branch of the i^{th} time interval.

Next, for reasons of notational simplicity that will become apparent, we shall indicate by

$x_{i,j}^{-t}$ the t^{th} preceeding symbol of $x_{i,j}$ in any given (4-2)
sequence. Hence:

$$x_{i,j}^{-t} = x_{r,s} \quad \text{where} \quad (i-r)n_0 + (j-s) = t \tag{4-3}$$

In block coding we will adhere to the usual notation. Thus

x_i will be the $\underline{i}^{th}$ symbol of a block. (4-4)

Similar to the sequential notation,

x_i^{-t} will be the t^{th} preceeding symbol of x_i in any (4-5)
given sequence (it is allowed that $t > i$, so that x_i^{-t} and x_i can belong to different blocks).

4.2 Symbol Notation

We will next display the symbolism used in the thesis. We will do so by using the sequential index notation defined above, any necessary changes for block notation will be straightforward and require no further explanation.

In what follows we will deal with symbols pertinent to the forward terminal. Their counterparts for the reverse terminal are superscribed with a bar " — ". Hence $\bar{x}$ is the symbol at the reverse terminal corresponding to x at the forward terminal.

Message digits:

$\ldots, z_i, z_{i+1}, \ldots, z_{i+j}, \ldots$ is a particular sequence of message digits, where z_i indicates which branch is to be followed out of the i^{th} sequential encoding state. (4-6)

Channel input signals:

$\ldots, x_{i,0}, x_{i,1}, \ldots, x_{i,n_{0-1}}, x_{i+1,0}, \ldots, x_{i+j,k}, \ldots$

is a particular sequence of input signals to the channel. Subscript notation was explained above. (4-7)

Channel output signals:

$$\ldots,y_{i,0},y_{i,1},\ldots,y_{i,n_0-1},y_{i+1,0},\ldots,y_{i+j,k},\ldots$$

is a particular sequence of output signals from the channel. Subscript notation was explained above. (4-8)

In order to simplify notation, we will assign special symbols to sequences of input and output signals of length ℓ.

Thus we will define:

$$\mathcal{X}^{\ell}_{i,j} \equiv (x^{-1}_{i,j},x^{-2}_{i,j},\ldots,x^{-\ell}_{i,j})$$

$$\mathcal{Y}^{\ell}_{i,j} \equiv (y^{-1}_{i,j},\ldots\ldots,y^{-\ell}_{i,j}) \tag{4-9}$$

Finally, we would like to define a single letter notation for sequences of symbols from the beginning of communication to a given time (i,j), and for sequences of symbols from a given time (i,j) to the time (i+K,j), where K is the number of time-strings in an encoding diagram, i.e., the number of branches taken into consideration for a single decoder decision. We will have:

$$X^{-}_{i,j} \equiv (x_{i,j-1},x_{i,j-2},\ldots\ldots,x_{0,0})$$

$$Y^{-}_{i,j} \equiv (y_{i,j-1},\ldots\ldots\ldots\ldots,y_{0,0}) \tag{4-10}$$

and

$$X^{+}_{i,j} \equiv (x_{i,j}, x_{i,j+1}, \ldots\ldots, x_{i+K,j-1})$$
$$Y^{+}_{i,j} \equiv (y_{i,j}, \ldots\ldots\ldots\ldots, y_{i+K,j-1}) \quad (4\text{-}11)$$

A similar capital letter notation will be used for sequentes of any other symbols yet to be defined. If it should become clear from the context that block coding is being considered, then double subscripts will give way to single subscripts. To illustrate, we will write e.g.:

$$X^{-}_{i+1} \equiv (x_{i}, x_{i-1}, \ldots\ldots, x_{0})$$
$$X^{+}_{i+1} \equiv (x_{i+1}, x_{i+2}, \ldots\ldots, x_{i+n}) \quad (4\text{-}12)$$
$$\mathfrak{X}^{\ell}_{i} \equiv (x_{i}^{-1}, x_{i}^{-2}, \ldots\ldots, x_{i}^{-\ell})$$

4.3 Encoding Functions

We will define functions f of "dependency" length ℓ, mapping the space of sequence pairs $(\mathcal{X}_i^\ell, \mathcal{Y}_i^\ell)$ on the space of channel input signals x_i:

$$f(\mathcal{X}_i^\ell, \mathcal{Y}_i^\ell) = x_i \tag{4-13}$$

Similar functions are defined for the reverse terminal.

It is clear from definition (4-13) that the domain of definition of any function f consists of $2^{2\ell}$ points $(\mathcal{X}_i^\ell, \mathcal{Y}_i^\ell)$. Hence the function is fully defined if a table of its values for the $2^{2\ell}$ possible different arguments is given. Thus any function f can be represented by a binary sequence of $2^{2\ell}$ elements, each corresponding to a different point in the domain of definition. We can write:

$$f \equiv (a_0, a_1, \ldots\ldots, a_i, \ldots\ldots, a_{2^{2\ell}-1}) \tag{4-14}$$

where a_i is the value of f when the sequence $(\mathcal{X}^\ell, \mathcal{Y}^\ell)$ constituting its argument is the binary representation of the integer i.

It is then clear from (4-14) that there are altogether $2^{2^{2\ell}}$ different possible functions f.

The script letter convention (4-9) and the capital letter conventions (4-10), (4-11), (4-12) will also apply to encoding functions, as well as the general indexing notation established above.

4.4 Derived Two-Way Channel

Consider the transducer of Figure 4-1.

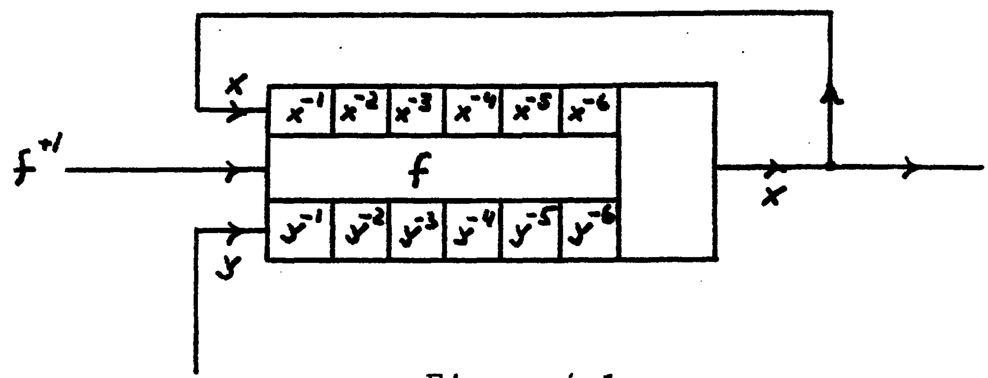


Figure 4-1.

It is designed for encoding functions of dependency length $\ell=6$. It operates as follows: there are memory compartments storing the input signal sequence $(x^{-1},\ldots\ldots,x^{-6})$ and the output sequence $(y^{-1},\ldots.,y^{-6})$ plus the function f. At given time intervals the transducer puts out the signal

$$x = f(x^{-1},\ldots,x^{-6},y^{-1},\ldots,y^{-6})$$

and simultaneously shifts its stored sequences by one place to the right, discarding digits x^{-6}, y^{-6} and the function f. Then its empty x-sequence slot is filled up by the signal x, its empty y-sequence slot by signal y, and its f-slot by the next function f^{+1}. It is then ready to put out the signal

$$x^{+1} = f^{+1}(x,x^{-1},\ldots,x^{-5},y,y^{-1},\ldots,y^{5})$$

and to repeat the same cycle. We will call the device of Figure 4-1 a "function-signal transducer", and it will be understood that the number of x-sequence and y-sequence slots will always be equal to the defined dependency length of the input functions f.

Consider now the channel of Figure 4-2.

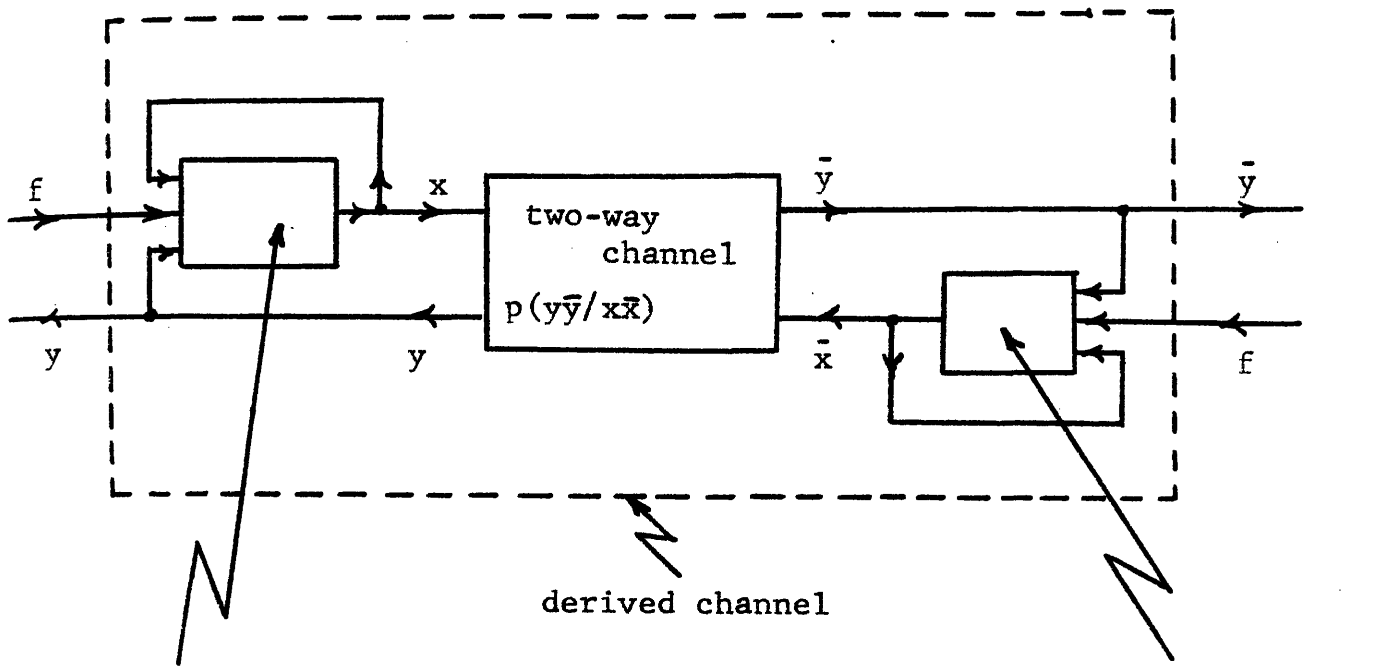


Figure 4-2.

The block arrangement is self-explanatory with the help of Figure 4-1. The entire arrangement within the dotted border can be considered to be a derived two-way channel with $2^{2^{2\ell}}$ different possible inputs f and $\bar{f}$ and binary outputs y and $\bar{y}$.

If it were agreed that communication will always start with prescribed sequences in the memory compartments of both transducers, then the derived channel would have finite state memory. In fact, the derived channel state at time i is a strict function of the state at time i-1, the channel inputs f_{i-1} and $\bar{f}_{i-1}$, and the channel outputs y_{i-1} and $\bar{y}_{i-1}$.

Next, we will develop some more terminology. The derived channel state denoted by $(s_i, \bar{s}_i)$ is fully determined by the sequences in the memory of the two transducers. Hence:

$$(s_i, \bar{s}_i) \equiv (\mathscr{X}_i, \mathscr{Y}_i, \bar{\mathscr{X}}_i, \bar{\mathscr{Y}}_i) \tag{4-15}$$

where $\mathscr{X}_i$ and $\mathscr{Y}_i$ are the sequences in the memory of the forward transducer at time i and $\bar{\mathscr{X}}_i$ and $\bar{\mathscr{Y}}_i$ are the sequences in the memory of the reverse transducer at time i. We have left out the superscripts ℓ to simplify notation. Henceforth, unless otherwise stated, any missing script letter superscripts will be assumed to have the value ℓ.

The derived channel now has the following state transmission probabilities:

$$Pr_{s_i, s_i}(y_i, \bar{y}_i / f_i, \bar{f}_i) = p(y_i, \bar{y}_i / f_i(s_i), \bar{f}_i(\bar{s}_i)) \tag{4-16}$$

It is also clear that there exist functions g and $\bar{g}$ such that:

$$s_{i+m} = g(s_i; f_i, f_{i+1}, \ldots, f_{i+m-1}; y_i, y_{i+1}, \ldots, y_{i+m-1}) \tag{4-17}$$

$$\bar{s}_{i+m} = \bar{g}(\bar{s}_i; \bar{f}_i, \bar{f}_{i+1}, \ldots, \bar{f}_{i+m-1}; \bar{y}_i, \bar{y}_{i+1}, \ldots, \bar{y}_{i+m-1})$$

Above is due to the working of the transducers described on p. 32 and to equation (4-13). There are of course $2^{4\ell}$ possible different derived channel states $(s, \bar{s})$.

We shall assume, without any loss of generality that the state $(s_o, \bar{s}_o)$ at the beginning of communication is always that in which the memories of both transducers are filled with zeros only. Hence, by convention,

$$(s_o, s_o) = (0,0) \tag{4-18}$$

where we have replaced the binary sequences $(\mathcal{x}_o, \mathcal{y}_o)$ and $(\bar{\mathcal{x}}_o, \bar{\mathcal{y}}_o)$ by their respective decimal values, a practice we shall follow whenever convenient.

5. Summary of Results

In view of the fact that the mathematical procedures used in this thesis are quite complicated and that some of the terminology and notation are new and perhaps cumbersome, it seems best to present at this point a summary of the results arrived at in the following articles. It is hoped that in this way motivation will be provided for those steps taken by the author which at first reading might otherwise appear capricious. In all that follows it will be assumed that channel transmission probabilities are restricted by the relation (3-1).

In Article 6 an ensemble of sequential codes is constructed by associating forward messages with sequences $f_1, f_2, \ldots, f_i, \ldots$ of encoding functions (having an arbitrary but fixed dependency length ℓ) independently at random with probability $P(f_1)P(f_2)\ldots P(f_i)\ldots$, and reverse messages with sequences $\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_i, \ldots$ of encoding functions (having the same dependency length ℓ as the forward functions) independently at random with probability $\bar{P}(\bar{f}_1)\bar{P}(\bar{f}_2)\ldots\bar{P}(\bar{f}_i)\ldots$ (for terminology see section 4.3), where $P(\ )$ and $\bar{P}(\ )$ are arbitrary probability distributions. It is then shown that the ensemble probability of forward and reverse decoding error is exponentially bounded with increasing code-word length n (i.e. it is shown that it is smaller than 2^{-nA}, where A is a function of the signalling rates), provided that the forward signalling rate R and the reverse signalling rate $\bar{R}$ are such that

$$R < \log(1-\varepsilon(m)) + E\left\{\log \frac{\Pr(\bar{y}/f,\boldsymbol{f}^m,\bar{f},\bar{\boldsymbol{f}}^m,\bar{\boldsymbol{y}}^m,\bar{\boldsymbol{x}}^{\ell}_{-m},\bar{\boldsymbol{y}}^{\ell}_{-m})}{\Pr(\bar{y}/\bar{f},\bar{\boldsymbol{f}}^m,\bar{\boldsymbol{y}}^m,\bar{\boldsymbol{x}}^{\ell}_{-m},\bar{\boldsymbol{y}}^{\ell}_{-m})}\right\} \qquad (5\text{-}1)$$

$$\bar{R} < \log(1-\bar{\varepsilon}(m)) + E\left\{\log \frac{\Pr(y/f,\boldsymbol{f}^m,\bar{f},\bar{\boldsymbol{f}}^m,\boldsymbol{y}^m,\boldsymbol{x}^{\ell}_{-m},\boldsymbol{y}^{\ell}_{-m})}{\Pr(y/f,\boldsymbol{f}^m,\boldsymbol{y}^m,\boldsymbol{x}^{\ell}_{-m},\boldsymbol{y}^{\ell}_{-m})}\right\}$$

where the expectations are determined from the channel transmission probabilities and from the probabilities P(f) and $\bar{P}(\bar{f})$ used to generate the ensemble, and the functions $\varepsilon(m) \to 0$ and $\bar{\varepsilon}(m) \to 0$ as the integer $m \to \infty$. As the integer m grows larger, the quantities on the right hand sides of the inequalities (5-1) increase in value. It follows from (5-1) that if the point $(R,\bar{R})$ lies in the convex hull of points

$$\lim_{r\to\infty} E\left\{\log \frac{\Pr(\bar{y}_r/f_r,f_{r-1},\ldots,f_1,\bar{f}_r,\bar{f}_{r-1},\ldots,\bar{f}_1,\bar{y}_{r-1},\ldots\bar{y}_1)}{\Pr(\bar{y}_r/\bar{f}_{r-1},\ldots,\bar{f}_1,\bar{y}_{r-1},\ldots,\bar{y}_1)}\right\},$$

$$\lim_{r\to\infty} E\left\{\log \frac{\Pr(y_r/f_r,f_{r-1},\ldots,f_1,\bar{f}_r,\bar{f}_{r-1},\ldots,\bar{f}_1,y_{r-1},\ldots,y_1)}{\Pr(y_r/f_r,f_{r-1},\ldots,f_1,y_{r-1},\ldots,y_1)}\right\} \qquad (5\text{-}2)$$

obtained for different assignments P(f) and $\bar{P}(\bar{f})$, then a code signalling with rate R in the forward direction and with rate $\bar{R}$ in the reverse direction will exist, for which the probability of decoding error will be bounded by $2^{-n\,A(R,\bar{R})}$, where $A(R,\bar{R})$ is a positive real quantity.

It should be pointed out that the convex hull (5-2) is a lower bound to the capacity region, tighter than Shannon's G_I to which it reduces when the assignment distributions $P(f)$ and $\bar{P}(\bar{f})$ are appropriately restricted (by assigning non-zero probabilities only to those $\mathbf{f}$ $(\bar{f})$ which map all the $2^{2\ell}$ different sequences $\boldsymbol{x}, \boldsymbol{y}$ $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ into the same signal x $(\bar{x})$).

The random argument used to reach the conclusion (5-1) and evaluate the character of the function $A(R,\bar{R})$ is complicated by the fact that the random variables with which one deals are dependent. The dependency is of a kind which can only approximately be associated with a Markoff source, and an important step in the proof (see Appendices III and IV) consists of showing that the error induced by the approximation is of an appropriate kind. In fact, the functions $\varepsilon(m)$ and $\bar{\varepsilon}(m)$ of (5-1) are the corrections necessitated by the approximation.

In Article 7 the properties of the expectations (5-2) are examined. It is first pointed out that the expectation pairs (5-2) are identical with the pairs

$$\lim_{n \to \infty} E\left\{\frac{1}{n} I(\bar{Y}^-_n;F^-_n/\bar{F}^-_n)\right\}, \quad \lim_{n\to\infty} E\left\{\frac{1}{n} I(Y^-_n;\bar{F}^-_n/F^-_n)\right\} \qquad (5\text{-}3)$$

of mutual informations, where the sequences $Y^-_n, \bar{Y}^-_n, F^-_n, \bar{F}^-_n$ of length n were defined in (4-12). It is then shown that if sequences F^-_n, $\bar{F}^-_n$ were transmitted and sequences Y^-_n, $\bar{Y}^-_n$ were received, then the signal information which passed through the two-way channel in the forward direction is equal to $I(\bar{Y}_n^-;X_n^-/\bar{X}_n^-)$, in the reverse direction is equal to $I(\bar{Y}_n^-;\bar{X}_n^-/X_n^-)$, and the total information which passed through the channel in both directions is equal to

$$I(\bar{Y}_n^-;X_n^-/\bar{X}_n^-) + I(Y_n^-;\bar{X}_n^-/X_n^-) = \\ = I(X_n^-,Y_n^-;\bar{X}_n^-,\bar{Y}_n^-) = I(F_n^-,Y_n^-;\bar{F}_n^-,\bar{Y}_n^-), \quad (5\text{-}4)$$

where the function F_n^- mapped Y_n^- into the sequence X_n^-, and the function $\bar{F}_n^-$ mapped $\bar{Y}_n^-$ into the sequence $\bar{X}_n^-$. It moreover turns out that if sequences F_n and $\bar{F}_n$ are picked at random with probabilities $Pr(F_n^-) = \prod_{i=0}^{n-1} P(f_i)$ and $Pr(\bar{F}_n^-) = \prod_{i=0}^{n-1} \bar{P}(\bar{f}_i)$ then

$$E\{I(\bar{Y}_n^-;X_n^-/\bar{X}_n^-)\} + E\{I(Y_n^-;\bar{X}_n^-/X_n^-)\} = \\ = E\{I(X_n^-,Y_n^-;\bar{X}_n^-,\bar{Y}_n^-)\} = E\{I(F_n^-,Y_n^-;\bar{F}_n^-,\bar{Y}_n^-)\}, \quad (5\text{-}5)$$

where the average on the right hand side is taken with respect to probabilities $P(f)$, $\bar{P}(\bar{f})$, while the expectations involving as variables the signal sequences $X_n^-,Y_n^-,\bar{X}_n^-,\bar{Y}_n^-$ only may be determined from the communication process represented in Figure 5-1.

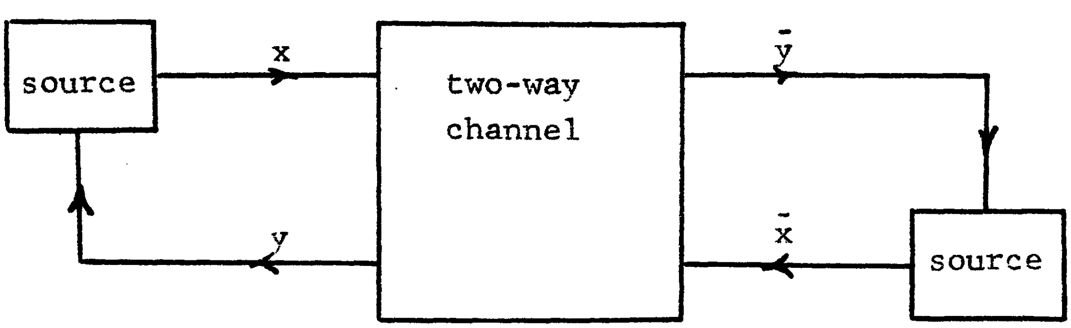


Figure 5-1

where the forward source generates signals x with probability $Pr(x/\mathfrak{x},\mathfrak{y}) = q_{\mathfrak{x}\mathfrak{y}}(x)$ and the reverse source generates signals $\bar{x}$ with probability $Pr(\bar{x}/\bar{\mathfrak{x}},\bar{\mathfrak{y}}) = \bar{q}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(\bar{x})$ (for explanation of notation see (4-9) and (4-12)). The source probabilities $q_{\mathfrak{x}\mathfrak{y}}(x)$ and $\bar{q}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(\bar{x})$ are defined with the help of the probabilities P(f) and $\bar{P}(\bar{f})$ as follows:

$$q_{\mathfrak{x}\mathfrak{y}}(x) = \sum_{f \ni f(\mathfrak{x},\mathfrak{y})=x} P(f) \qquad \bar{q}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(\bar{x}) = \sum_{\bar{f} \ni \bar{f}(\bar{\mathfrak{x}},\bar{\mathfrak{y}})=\bar{x}} \bar{P}(\bar{f}) \tag{5-6}$$

The information (5-4) transmitted through the channel is then examined in relation to the informations (5-3) which in the coding theorem were shown to be useful in message decoding. Theorem 7-6 then states that

$$\begin{aligned} E\left\{I(\bar{Y}_n^-;X_n^-/\bar{X}_n^-)\right\} - E\left\{I(\bar{Y}_n^-;F_n^-/\bar{F}_n^-)\right\} &= E\left\{I(Y_n^-;\bar{Y}_n^-/F_n^-,\bar{X}_n^-)\right\} \\ E\left\{I(Y_n^-;\bar{X}_n^-/X_n^-)\right\} - E\left\{I(Y_n^-;\bar{F}_n^-/F_n^-)\right\} &= E\left\{I(Y_n^-:\bar{Y}_n^-/X_n^-,\bar{F}_n^-)\right\} \end{aligned} \tag{5-7}$$

or, in other words, that not all the information that passes through the channel is useful for message decoding purposes. Further theorems show that the quantities $E\left\{I(Y_n^-;\bar{Y}_n^-/F_n^-,\bar{X}_n^-)\right\}$ and $E\left\{I(Y_n^-;\bar{Y}_n^-/X_n^-,\bar{F}_n^-)\right\}$ can be interpreted as an average loss of information due to forward and reverse coding, respectively, and that, except for the so called noiseless channels, all codes which do not map messages strictly into channel signals are associated with a positive average coding loss. A new formula for the outer bound to the channel capacity region is developed in section 7-6. It is subsequently used, in conjunction with the positiveness of the average coding loss, to prove that for non-degenerate channels the capacity region G is strictly interior to Shannon's outer bound G_0.

In Article 8 binary two-way channels with transmission probabilities restricted by the relation

$$P(y,\bar{y}/x,\bar{x}) = p(y/x\bar{x})\ \bar{p}(\bar{y}/x\bar{x}) \tag{5-8}$$

are discussed. It is recalled that they can be represented by the diagram of Figure 3-1 which can be said to consist of a forward (top) and reverse (bottom) portion. The forward portion, diagrammed in Figure 5-2,

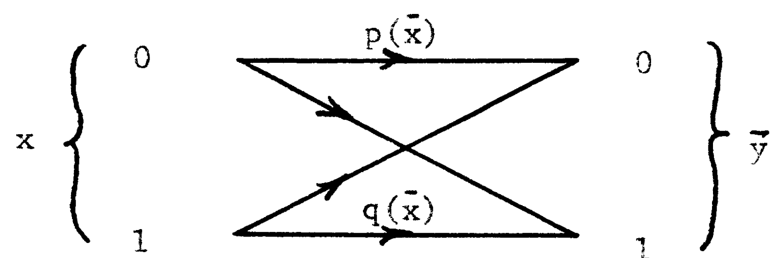


Figure 5-2.

is then examined with the conclusion that it must belong to one of two classes, depending on the probability set $\{p(\bar{x}), q(\bar{x})\}$. Into Class A belong channel portions in which additional noise is caused by simultaneous transmission of signals x and $\bar{x}$ such that $x \neq \bar{x}$. Into Class B belong those portions in which additional noise on the forward transmission is caused by one of the two possible signals $\bar{x}$. It is then shown that all portions of Figure 5-2 can be represented, either directly or after a transformation by the cascade of Figure 5-3,

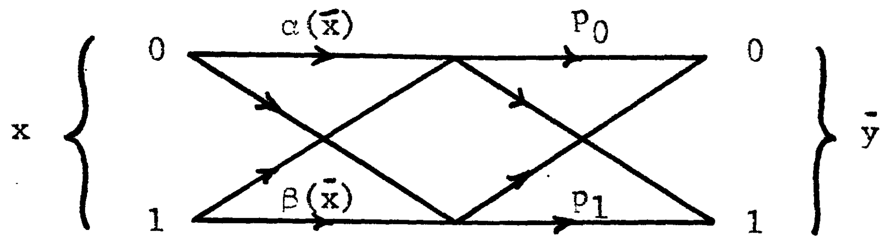


Figure 5-3

where the channel involving the probabilities p_0 and p_1 is a binary one-way channel, and the channel involving probabilities $\alpha(x)$ and $\beta(x)$ can be represented either as in Figure 5-4 or as in Figure 5-5.

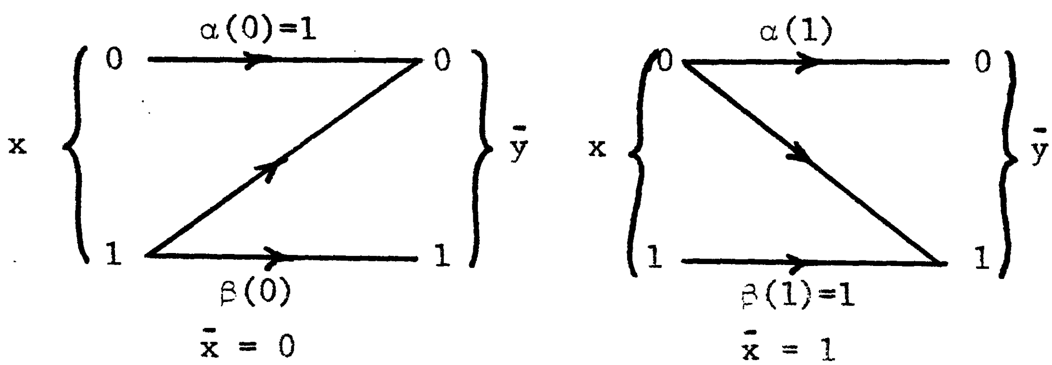


Figure 5-4.

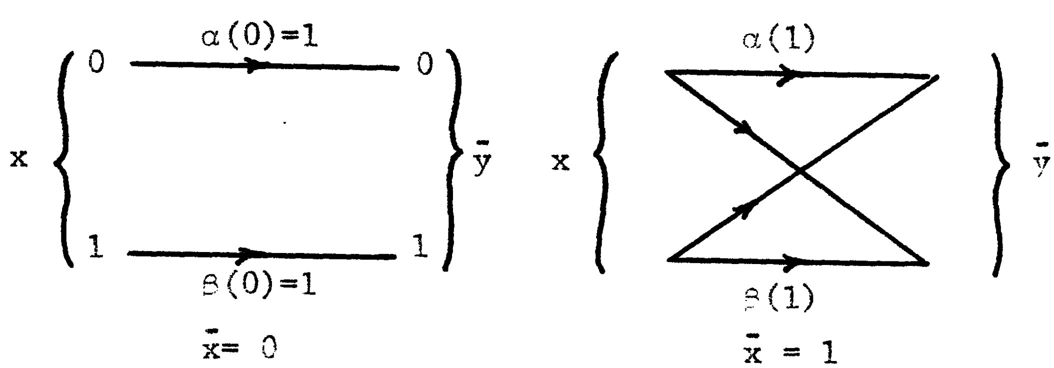


Figure 5-5.

It is clear that if a channel portion is representable by Figure 5-3 directly (i.e. if no transformation is necessary), then if it belongs to Class A, it is representable by Figure 5-4, and if it belongs to Class B, it is representable by Figure 5-5. Finally, interesting examples of binary two-way channels restricted by (5-8) are given, exhibiting significant symmetries.

Article 9 uses the results of Articles 7 and 8 to discuss heuristically what kind of source probabilities $q(x)$ and $\bar{q}(\bar{x})$ (see Figure 5-1) will insure high information transmission through binary two-way channels whose defining transmission probabilities $p(y/x,\bar{x})$ and $\bar{p}(\bar{y}/x,\bar{x})$ are qualitatively described by membership in the classes of Article 8. We then go on to prove that the six mutual information functions appearing in the two equations (5-7) are concave downward functions of the probability assignments $P(F_n^-)$, $P(\bar{F}_n^-)$.

Attention is then focussed on the problem of construction of random ensembles from which to pick, with great likelihood, codes which would signal at close to capacity rates with a small probability of decoding error. If the intended ratio of reverse to forward signalling rate is λ then the probability distributions $P(f)$, $\bar{P}(\bar{f})$ should be determined by maximization of the quantity

$$E\left\{\log \frac{\Pr(\bar{y}/f,\boldsymbol{f}^m,\bar{f},\bar{\boldsymbol{f}}^m,\bar{\boldsymbol{y}}^m,\bar{\boldsymbol{x}}^{\ell}_{-m},\bar{\boldsymbol{y}}^{\ell}_{-m})}{\Pr(\bar{y}/\bar{f},\bar{\boldsymbol{f}}^m,\bar{\boldsymbol{y}}^m,\bar{\boldsymbol{x}}^{\ell}_{-m},\bar{\boldsymbol{y}}^{\ell}_{-m})}\right\} + \\ + \lambda E\left\{\log \frac{\Pr(y/f,\boldsymbol{f}^m,\bar{f},\bar{\boldsymbol{f}}^m,\boldsymbol{y}^m,\boldsymbol{x}^{\ell}_{-m},\boldsymbol{y}^{\ell}_{-m})}{\Pr(y/f,\boldsymbol{f}^m,\boldsymbol{y}^m,\boldsymbol{x}^{\ell}_{-m},\boldsymbol{y}^{\ell}_{-m})}\right\} \tag{5-9}$$

over the set of all possible independent symbol distributions for some sufficiently large integer m. For a given dependency length ℓ there exist, however, $2 \times 2^{2^{2\ell}}$ different functions f and $\bar{f}$, and therefore as many variables over which to maximize (5-9). In view of Theorem 7-6 (see relations (5-7)) it is suggested that one maximize instead of (5-9) the expression

$$E\left\{\log \frac{\bar{p}(\bar{y}/x,\bar{x})}{\Pr(\bar{y}/\bar{x},\bar{\boldsymbol{x}}^m,\bar{\boldsymbol{y}}^m,\bar{\boldsymbol{x}}^{\ell}_{-m},\bar{\boldsymbol{y}}^{\ell}_{-m})}\right\} + \\ + \quad E \log \frac{p(y/x,\bar{x})}{\Pr(y/x,\boldsymbol{x}^m,\boldsymbol{y}^m,\boldsymbol{x}^{\ell}_{-m},\boldsymbol{y}^{\ell}_{-m})} \tag{5-10}$$

over the set of $2 \times 2^{2\ell}$ probabilities $q_{\boldsymbol{xy}}(x)$, $\bar{q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{y}}}(\bar{x})$ and then, using the relations (5-6) as constraints, minimize

$$E\left\{\log \frac{\bar{p}(\bar{y}/x,\bar{x})}{\Pr(\bar{y}/f,\boldsymbol{f}^m,\bar{x},\bar{\boldsymbol{x}}^m,\bar{\boldsymbol{y}}^m,\bar{\boldsymbol{x}}^{\ell}_{-m},\bar{\boldsymbol{y}}^{\ell}_{-m})}\right\} + \\ \lambda E\left\{\log \frac{p(y/x,\bar{x})}{\Pr(y/\bar{f},\bar{\boldsymbol{f}}^m,x,\boldsymbol{x}^m,\boldsymbol{y}^m,\boldsymbol{x}^{\ell}_{-m},\boldsymbol{y}^{\ell}_{-m})}\right\} \tag{5-11}$$

over the possible distributions P(f) and $\bar{P}(\bar{f})$. A theorem is then proven stating that, except for noiseless channels, the quasi-optimization involving (5-10) and (5-11) will never lead to the optimal solution. It is nevertheless concluded that for practical reasons one must be satisfied with carrying out even a weaker quasi-optimization: one involving step (5-10) but replacing step (5-11) by an arbitrary determination of the distribution P(f), $\bar{P}(\bar{f})$ satisfying constraints (5-6). It is shown that for any given set $\{q_{xy}(x)\}$ of size $2^{2\ell}$ there exists a set of $(2^{2\ell}+1)$ non-zero P(f) which will satisfy the constraint (5-6). The $\{P(f)\}$ set obtained is in a definite sense an optimal one, and hence is called canonic. Using the idea of the canonic sets $\{P(f)\}$ and $\{\bar{P}(\bar{f})\}$, another, better quasi-optimization procedure is developed, involving maximization of the expression (5-9) over the sets of different $2(2^{2\ell}+1)$ canonical variables.

We will finally call attention to Appendix III whose results we believe to be of mathematical interest for their own sake. As already mentioned in the second paragraph following (5-2), the theorems derived in Appendix III are necessary for the proof of the coding theorem of Article 6. Suppose a Markoff source is given with an associated matrix $[\mathcal{M}]$ of its transition probabilities. Let the source have k states, $\{s_1, s_2, \ldots, s_k\}$, and let the value of a matrix element m_{rt} be

$$m_{rt} = p(s_t^i / s_r^{i-1}) \qquad (5\text{-}12)$$

(ie. the probability that if the process was at time i-1 in state s_r, that it will be at time i in the state s_t)

Let the states of the source be partitioned into ℓ non-empty classes $A_1, A_2, \ldots, A_\ell$. Appendix III develops the sufficient conditions on the transition matrix $[\mathcal{M}]$ which must be satisfied so that the following statement will hold:

Given any $\varepsilon > 0$, an integer Λ can be found such that for all integers $\lambda > \Lambda$ and $\sigma \geq 1$ and for all sequences of successive states $A^{i-\lambda-\sigma}, A^{i-\lambda-\sigma+1}, \ldots, A^{i-\lambda}, \ldots, A^{i-1}, A^{i}$ for which

$$\Pr(A^i, A^{i-1}, \ldots, A^{i-\lambda}, \ldots, A^{i-\lambda-\sigma}) \neq 0, \tag{5-13}$$

the inequality

$$\left| \Pr(A^i / A^{i-1}, \ldots, A^{i-\lambda}) - \Pr(A^i / A^{i-1}, \ldots, A^{i-\lambda}, \ldots, A^{i-\lambda-\sigma}) \right| < \varepsilon \tag{5-14}$$

is satisfied, where $A^j \in \{A_1, A_2, \ldots, A_\ell\}$,

$$j = i-\lambda-\sigma, \ldots, i.$$

6. Coding Theorem

6.1 Two-way Channel Sequential Function Codes

We will approach the coding problem through the concept of a derived channel described diagrammatically by Figure 4-2. In what follows we will use the terminology and concepts of section 3.5 as well as the notation of Article 4.

Consider a pair of sequential state coding diagrams having the structure of Figure 3-3, one for each input terminal of the derived channel of Figure 4-2. Let the number of time-strings in the forward coding diagram be k. Associate with each of the transitions out of each state of the forward coding diagram a sequence of n_0 encoding functions, say $(f_{i,0}, f_{i,1}, \ldots, f_{i,n_0-1})$ where $f_{i,j}$ is a function mapping the space of possible past signal sequence pairs $(\mathcal{X}^{\ell}_{i,j}, \mathcal{Y}^{\ell}_{i,j}) = (x^{-1}_{i,j}, \ldots, x^{-\ell}_{ij}, y^{-1}_{i,j}, \ldots, y^{-\ell}_{ij})$ on the space of channel inputs signals x_{ij} (for more thorough discussion see section 4.3). It ought to be recalled that the functions f_{ij} constitute the input symbols to the derived channel of Figure 4-2.

If the message is encoded into a string of binary digits, the operation of the sequential coding diagram is identical to the operation described in section 3.5. Hence, for instance, given the time t/mod k and the k most recent message digits, $z_{t-k}, z_{t-k+1}, \ldots, z_{t-1}$, the state of the encoder is determined. The next state and therefore the particular transition to be taken, together with the associated sequence of input symbols to be transmitted through the channel, is then determined by the message digit z_t. The reverse encoding diagram, operates in the same way, except that the function sequences associated with the state transitions will have length $\bar{n}_0$ and will be denoted $\bar{f}_{i,0},\ldots,\bar{f}_{i,n_0-1}$. We can postulate the reverse coding diagram's time-string number to be $\bar{k}$. The appropriate messages are of course $\bar{z}_i$-sequences and in what follows it will be assumed that the forward and reverse messages are independent of each other.

Decoding will be sequential, one message digit at a time. Hence at time i, on the basis of a known forward "departure state" (characterized by time i-k and the message digit sequence $z_{i-2k},\ldots,z_{i-k-1}$) and of a consequent set of all possible function words $F^{+}_{i-k,0}$ allowed by the forward coding diagram, the reverse decoder will determine the next forward encoding state (i.e. the identity of the message digit z_{i-k}) by the use of its knowledge of sequences $\bar{F}^{+}_{i-k,0},\bar{Y}^{+}_{i-k,0},F^{-}_{i-k,0}$, $\bar{F}^{-}_{i-k,0}$ and $\bar{Y}^{-}_{i-k,0}$. Perhaps the reader should be reminded here that in section 4.2, $F^{-}_{i-k,0}$ was defined as the sequence $(f_{i-k,0}$, $f_{i-k,1}\ldots f_{i-1,n_0-1})$ of derived channel input symbols, and $\bar{F}^{-}_{i-k,0}$ was defined as the sequence $(\bar{f}_{0,0},\bar{f}_{0,1},\ldots,\bar{f}_{i-k-1,n_0-1})$. Similar definitions apply for the sequences $\bar{F}^{+}_{i-k,0},\bar{Y}^{+}_{i-k,0}$, $\bar{F}^{-}_{i-k,0},\bar{Y}^{-}_{i-k,0}$, except for a slight modification explained below.

We recall first that the integer n_0 was defined as the number of symbols associated with a transition of the forward encoding diagram, and that the integer $\bar{n}_0$ was defined as the number of symbols associated with a transition of the reverse encoding diagram. But in general $n_0 \neq \bar{n}_0$. From this it follows that when one encoder reaches a state in time-string α and is about to transmit the first symbol associated with the next transition, the other encoder may be just about to transmit the i^{th} symbol of some transition out of time-string β. We also recall that in general the number of time strings k of the forward encoder does not equal the number of time-strings $\bar{k}$ of the reverse encoder. Thus, according to the notation of Article 4, the symbols $f_{i,j}$ and $\bar{f}_{r,s}$ might be transmitted simultaneously, where in general $i \neq r$, $j \neq s$, but $in_0 + j = r\bar{n}_0 + s$, where i and r are considered to be absolute integers (i.e. not integers modulo k and $\bar{k}$, respectively). Such state of affairs confuses very much the notational system. However, in what follows we will always be concerned with either the decoding of the forward or of the reverse messages only. When dealing with the decoding of the forward transmission it will be important to keep the forward indexing notation "straight", while the reverse indexing will matter only in so far as it will indicate which forward and reverse symbol events occurred simultaneously. It will therefore be advantageous to "refer" the reverse indexing to the forward notation. Hence when dealing with the forward message decoding, it will be understood that, for instance, $\bar{f}_{i,j}$ will refer to the reverse symbol transmitted simultaneously with the forward symbol $f_{i,j}$, and that the index (i,j), wherever it should occur, will indicate the position on the forward encoding tree only. It will also be convenient to pick in any given code the integers k and $\bar{k}$ so that $kn_0 = \bar{k}\bar{n}_0$. This can always be done. When, on the other hand, decoding of the reverse message will be considered, all indexes will refer to the position on the reverse encoding tree.

In the paragraph preceding the last one we have discussed the decoding of the forward message, and hence the index notation of the listed quantities $F^+_{i-k,0}, \bar{F}^+_{i-k,0}, \bar{Y}^+_{i-k,0}, F^-_{i-k-0}, \bar{F}^-_{i-k,0}$ and $\bar{Y}^-_{i-k,0}$ is to be referred to the forward side.

Resuming our discussion of decoding, it ought to be noted that the identity of signals transmitted or received at the forward terminal of the two-way channel is never determined by the reverse decoder. It is in fact possible to treat consistently the derived channel as a "black box" into whose insides one never looks.

The decoding procedure will now be considered in detail. An arbitrary, not necessarily optimal, decoding decision rule will be given. Let it then be stated again that decoding is to be performed one message digit at a time and that once a decision is made it is never cancelled or re-considered: all further work is done on the assumption that the decision was a correct one. All derived channel communication follows the convention established at the end of section 4.4; it always starts from the derived channel transducer state (0,0) (see Eq. 4-18). The reverse decoder is at time i in possession of the following knowledge:

(a) primary knowledge:

- (i) initial state (0,0) of the derived channel
- (ii) forward message digit sequence Z^-_{i-k} (6-1)
- (iii) reverse message digit sequences $\bar{Z}^-_{i-k}$, and $\bar{Z}^+_{i-k}$
- (iv) reverse received sequences $\bar{Y}^+_{i-k}$, and $\bar{Y}^-_{i-k}$
- (v) forward and reverse sequential coding diagrams.

(b) secondary knowledge (computable from primary):

- (i) forward input symbol sequence F^-_{i-k}
- (ii) reverse input symbol sequences $\bar{F}^-_{i-k}$, and $\bar{F}^+_{i-k}$.
- (iii) reverse input signal sequences $\bar{X}^-_{i-k}$, and $\bar{X}^+_{i-k}$.
- (iv) reverse portion of the derived channel state sequence $s_{0,0},\ldots,s_{i-1,n_0-1}$.

As mentioned, the determination of the message digit z_{i-k} is desired. The forward coding diagram together with the known departure state specify a set of 2^k symbol words F^+_{i-k} which could have been transmitted through the derived channel in the forward direction. Basing itself on the available information (9-1), let the reverse decoder find the most probable word of the set, say F^{+*}_{i-k}. The word F^{+*}_{i-k} will correspond to some message sequence Z^{+*}_{i-k}, which will include z^{*}_{i-k} as its first digit. In that case let the decoding decision be that z^{*}_{i-k} was actually transmitted. It is clear that the above decision rule is based on finding in the allowable word-set the sequence F^{+}_{i-k} which will maximize the probability

$$\Pr(F^{+}_{i-k}/\bar{F}^{+}_{i-k},\bar{Y}^{+}_{i-k},F^{-}_{i-k},\bar{F}^{-}_{i-k},\bar{Y}^{-}_{i-k}) = \tag{6-2}$$

$$= \frac{\Pr(\bar{Y}^{+}_{i-k}/F^{+}_{i-k}\ \bar{F}^{+}_{i-k}\ F^{-}_{i-k}\ \bar{F}^{-}_{i-k}\ \bar{Y}^{-}_{i-k})\ \Pr(F^{+}_{i-k})}{\Pr(\bar{Y}^{+}_{i-k}/\bar{F}^{+}_{i-k},F^{-}_{i-k},\bar{F}^{-}_{i-k}\ \bar{Y}^{-}_{i-k})} =$$

$$= \frac{1}{2^{k}}\ \frac{\Pr(\bar{Y}^{+}_{i-k}/F^{+}_{i-k}\ \bar{F}^{+}_{i-k}\ F^{-}_{i-k}\ \bar{F}^{-}_{i-k}\ \bar{Y}^{-}_{i-k})}{\Pr(\bar{Y}^{+}_{i-k}/\bar{F}^{+}_{i-k},F^{-}_{i-k},\bar{F}^{-}_{i-k},\bar{Y}^{-}_{i-k})}$$

where we have assumed that the messages transmitted from the opposite terminals are statistically independent and equiprobable. The maximization over the allowable set can be carried out on the probability (6-3) below, since the denominator on the right hand side of (6-2) is not a function of F^{+}_{i-k}.

$$\Pr(\bar{Y}^{+}_{i-k}/F^{+}_{i-k},\bar{F}^{+}_{i-k},F^{-}_{i-k}\ \bar{F}^{-}_{i-k}\ \bar{Y}^{-}_{i-k}) = \tag{6-3}$$

$$\prod_{r=i-k}^{i-1}\prod_{s=0}^{n_0-1}\Pr(\bar{y}_{r,s}/f_{r,s},f_{r,s-1},\ldots,f_{i-k,0},\bar{y}_{r,s-1},\ldots$$

$$\ldots\bar{y}_{i-k,0},\bar{f}_{r,s},\ldots,\bar{f}_{i-k,0},F^{-}_{i-k},\bar{F}^{-}_{i-k},\bar{Y}^{-}_{i-k})$$

6.2 Random Codes

As stated, our aim is to define a region G (which will include Shannon's inner bound region G_I) of points $(R,\bar{R})$ for which the following statement can be made:

> If a point $(R,\bar{R})$ is in G then there exist two-way channel codes, signalling at forward and reverse rates arbitrarily close to R and $\bar{R}$, respectively, for which the probability of erroneous decoding can be made as small as desired.

In what follows we will use the well-known random coding technique. Assume that the desired signalling rates are such that

$$R = \frac{1}{n_0} \quad ; \quad \bar{R} = \frac{1}{\bar{n}_0} \tag{6-5}$$

(assumption (6-5) will assure that in the sequential code used each transition from one state to another will correspond to one message digit only, and hence that there will always be two transitions out of each state. However, all that follows will be readily generalizable for situations in which (6-5) does not hold).

If the time-string numbers k and $\bar{k}$ are also decided upon, the forward and reverse decoding diagrams may be constructed. Let us next assign a probability distribution P(f) and $\bar{P}(\bar{f})$ over the possible forward and reverse input symbols to the derived channel. A random code ensemble can then be generated by assigning the coding sequences $f_{i,0},\ldots,f_{i,n_0-1}$ to the transitions of the forward coding diagram, independently at random with probabilities $P(f_{i,0})P(f_{i,1})\ldots P(f_{i,n_0-1})$. Hence a coding diagram whose coding sequences were chosen in the above way will have an ensemble probability equal to the product of probabilities of all the symbols f appearing in it. The same procedure is to be applied to reverse coding diagrams. The ensemble probability of a forward-reverse code pair will then be equal to the product of probabilities of its forward and reverse components.

In what follows we will be dealing with the decoding of forward messages and our notation will be adjusted accordingly. The ensemble probability that the forward terminal will transmit the symbol sequence $f_{0,0},\ldots,f_{i,j}$ is:

$$Pr(f_{00},f_{01},\ldots,f_{ij})=\left[\prod_{r=0}^{i-1}\prod_{s=0}^{n_0-1}P(f_{rs})\right]\cdot\left[P(f_{i,0})\ldots P(f_{ij})\right] \tag{6-6}$$

The expression for the ensemble probability of transmission of a reverse symbol sequence is of the same form. Hence the ensemble probability of the event $(f_{00},\ldots,f_{i-1,n_0-1}),(\bar{f}_{00},\ldots,\bar{f}_{i-1,n_0-1});(y_{00},\ldots,y_{i-1,n_0-1}),(\bar{y}_{00},\ldots,\bar{y}_{i-1,n_0-1})$ is:

$$\Pr(F^-_{i-k},F^+_{i-k},\bar{F}^-_{i-k},\bar{F}^+_{i-k},Y^-_{i-k},Y^+_{i-k},\bar{Y}^-_{i-k},\bar{Y}^+_{i-k}) = \qquad (6\text{-}7)$$

$$= \prod_{r=0}^{i-1} \prod_{s=0}^{n_0-1} P(f_{r,s})\bar{P}(\bar{f}_{r,s})p(y_{r,s}/x_{rs}\bar{x}_{r,s})\bar{p}(\bar{y}_{r,s}/x_{rs}\bar{x}_{rs})$$

where $x_{rs} = f(\mathcal{X}_{rs},\mathcal{Y}_{rs})$, $\bar{x}_{rs} = \bar{f}(\bar{\mathcal{X}}_{r,s},\bar{\mathcal{Y}}_{r,s})$ and

$$(\mathcal{X}_{00},\mathcal{Y}_{00}) = (\bar{\mathcal{X}}_{00},\bar{\mathcal{Y}}_{00}) = (0,0)$$

where for simplicity we have assumed the restriction $P(y,y/xx) = p(y/x\bar{x})\ \bar{p}(\bar{y}/x\bar{x})$ on the transmission probabilities of the two-way channel, although, as will be seen from the proof, this was not essential. Equation (6-7) shows that our random process can be represented by a Markoff diagram having $2^{4\ell}$ states characterized by different quadruplets of channel signal sequences of length ℓ (i.e. by the states of the derived channel): $(\mathcal{X},\mathcal{Y},\bar{\mathcal{X}},\bar{\mathcal{Y}})$.

Transitions from Markoff state $(\mathcal{X},\mathcal{Y},\bar{\mathcal{X}},\bar{\mathcal{Y}})$ to state $(\mathcal{X}',\mathcal{Y}',\bar{\mathcal{X}}',\bar{\mathcal{Y}}')$ are possible only if

$$\begin{aligned} (x')^{-r} &= x^{-r+1} \\ (y')^{-r} &= y^{-r+1} \\ (\bar{x}')^{-r} &= \bar{x}^{-r+1} \\ (\bar{y}')^{-r} &= \bar{y}^{-r+1} \end{aligned} \qquad (6\text{-}8)$$

for $r\varepsilon(2,3,\ldots,\ell)$

Given a state $(\mathcal{X},\mathcal{Y},\bar{\mathcal{X}},\bar{\mathcal{Y}})$ there are four possible transitions associated with any given pair of symbols $f,\bar{f}$, the next possible state being fixed except for the received signal pair

$((y')^{-1}, (\bar{y}')^{-1})$. This follows from the fact that $(x')^{-1} = f(\mathfrak{x},\mathfrak{y})$ and that $(\bar{x}')^{-1} = \bar{f}(\bar{\mathfrak{x}},\bar{\mathfrak{y}})$. Given the quantities $(\mathfrak{x},\mathfrak{y},x)$ there are $2^{2^{2\ell}-1}$ functions f such that $f(\mathfrak{x}\mathfrak{y}) = x$ and it is therefore clear that there are either 0 or $2^{2(2^{\ell}-1)}$ transitions from any given Markoff state to another.

The probability of a transition from state $(\mathfrak{x},\mathfrak{y},\bar{\mathfrak{x}},\bar{\mathfrak{y}})$ to state $(\mathfrak{x}',\mathfrak{y}',\bar{\mathfrak{x}}',\bar{\mathfrak{y}}')$ associated with the function pair $f,\bar{f}$ is

$$P_{\mathfrak{x}\mathfrak{y}\bar{\mathfrak{x}}\bar{\mathfrak{y}}}((\mathfrak{x}',\mathfrak{y}',\bar{\mathfrak{x}}',\bar{\mathfrak{y}}'), f\bar{f}) = \begin{cases} = p((y')^{-1}/(x')^{-1}(\bar{x}')^{-1})\,\bar{p}((\bar{y}')^{-1}/(x')^{-1}(\bar{x}')^{-1}) \cdot P(f)\bar{P}(\bar{f}) \\ \quad \text{if } f(\mathfrak{x},\mathfrak{y}) = (x')^{-1}, \bar{f}(\bar{\mathfrak{x}},\bar{\mathfrak{y}}) = (\bar{x}')^{-1} \\ \quad \text{and (6-8) is satisfied} \\ = 0 \quad \text{otherwise.} \end{cases} \qquad (6\text{-}9)$$

Thus the appropriate channel ensemble Markoff diagram can at least in principle be constructed.

6.3 Ensemble Average Probability of Error

If in the following argument, we were to try to use the decoding criterion (6-3) we would run into serious difficulty when trying to estimate the average ensemble probability of decoding error. This difficulty, it would turn out, would be due entirely to the fact that, as can be seen from the product on the right-hand side of (6-3), the decoding criterion takes into account the entire past symbol sequences known to the decoder. It will prove necessary to define for our purposes a slightly weaker criterion in which the decoder will base its decision on sequences of length m of past known symbols, where m is an arbitrarily picked integer. It will be seen later that this weaker criterion will not diminish the strength of the resulting coding theorem at all, since after the desired bound expression will be obtained for a given m, it will be possible to carry it to the limit as $m \to \infty$.

Hence we will define in (6-10a) below a probability-like function

$$Q_m\,(\bar{Y}^{+}_{i-k}/F^{+}_{i-k}, \bar{F}^{+}_{i-k}, F^{-}_{i-k}, \bar{F}^{-}_{i-k}, \bar{Y}^{-}_{i-k}) \tag{6-10a}$$

$$\prod_{r=i-k}^{i-1} \prod_{s=0}^{n_\theta-1} \Pr(\bar{y}_{r,s}/f_{r,s}\; \bar{f}_{r,s}, \mathscr{f}^{m}_{r,s}, \bar{\mathscr{f}}^{m}_{r,s}, \bar{\mathscr{y}}^{n}_{r,s}, \bar{s}^{-m}_{r,s}$$

where m is an arbitrary integer and $\bar{s}^{-m}_{r,s}$ are the reverse portions of the derived channel states as listed in (6-1,b,iv).

They are determined from the knowledge of all the symbols $\bar{f}_{n,j}$ and all the signals $\bar{y}_{n,j}$ preceding the state $\bar{s}^{-m}_{r,s}$ in time, by the use of the formula (4-17).

We stress again that the difference between (6-3) and (6-10a) is that in the product on the right-hand side of (6-10a) the conditional probabilities exhibit dependence on sequences of constant length, while the dependency of the probabilities on the right-hand side of (6-3) increases indefinitely with increasing i. This of course accounts for the statement that the criterion (6-10a) is weaker than (6-3).

For any given code the new decoding procedure is then the following one: Using its knowledge of the sequences $F^-_{i-k}, \bar{F}^+_{i-k}$ $\bar{F}^-_{i-k}, \bar{Y}^+_{i-k}, \bar{Y}^-_{i-k}$, the reverse decoder is going to compute the quantity (6-10a) for all the sequences F^+_{i-k} associated in the forward code with the 2^k different messages. Suppose that the highest value of (6-10a) is obtained for the word F^{+*}_{i-k} associated to the message Z^{+*}_{i-k}. Then the reverse decoder will decide that z^*_{i-k}, the first digit of the message word Z^{+*}_{i-k}, was transmitted by the forward terminal.

Suppose therefore that the event $F^-_{i-k}, F^+_{i-k}, \bar{F}^-_{i-k}, \bar{F}^+_{i-k}, \bar{Y}^-_{i-k}$, $\bar{Y}^+_{i-k}$ occurred and that the reverse decoder is trying to determine the message digit z_{i-k}, having decoded correctly all the preceding message digits Z^-_{i-k}. The decoder will make the wrong decision only if there is in the incorrect subset of the sequential encoder (using Wozencraft's terminology) a symbol word F^{+*}_{i-k} such that for it

$$Q_m(\bar{Y}^+_{i-k}/F^{+*}_{i-k},\bar{F}^+_{i-k},F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^-_{i-k}) \geq \geq Q_m(\bar{Y}^+_{i-k}/F^+_{i-k},\bar{F}^+_{i-k},F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^-_{i-k}) \tag{6-10b}$$

We can then define a set $S_{\bar{Y}^+_{i-k},\bar{Y}^-_{i-k}}(F^+_{i-k},F^-_{i-k},\bar{F}^+_{i-k},\bar{F}^-_{i-k})$ of symbol words F^{+*}_{i-k} such that

$$F^{*+}_{i-k} \in S_{\bar{Y}^+_{i-k},\bar{Y}^-_{i-k}}(F^+_{i-k},F^-_{i-k},\bar{F}^+_{i-k},\bar{F}^-_{i-k}) \tag{6-11}$$

if and only if (6-10b) holds.

We then get that

$$\Pr[S_{\bar{Y}^+\bar{Y}^-}(F^+,F^-,\bar{F}^+,\bar{F}^-)] = \sum_{F^{+*} \in S_{\bar{Y}^+\bar{Y}^-}(F^+,F^-,\bar{F}^+,\bar{F}^-)} \Pr(F^{+*}) \tag{6-12}$$

It will be convenient to state the criterion (6-10b) for membership in the set (6-11) in a different way. Define an inverse distance between the signal sequence $\bar{Y}^+_{i-k}$ and the symbol sequence F^+_{i-k}, given that the word $\bar{F}^+_{i-k}$ was transmitted and events $F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^{-1}_{i-k}$ occurred previously:

$$D_m(F^+_{i-k};\bar{Y}^+_{i-k}/\bar{F}^+_{i-k},F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^-_{i-k},) \equiv \equiv \log \frac{Q_m(\bar{Y}^+_{i-k}/F^+_{i-k},\bar{F}^+_{i-k},F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^-_{i-k})}{G\ (\bar{Y}^+_{i-k},\bar{F}^+_{i-k},F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^-_{i-k})} \tag{6-13}$$

where G is an as yet unspecified function of the indicated variables. The structure of G will be determined later so as to facilitate the estimation of the probability of ensemble decoding error. Since G is not a function of the symbol sequence F^{+}_{i-k}, the relation (6-10) can be replaced by (6-14) below as the condition for membership in the set defined by (6-11).

$$D_m(F^{+*}_{i-k};\bar{Y}^{+}_{i-k}/\bar{F}^{+}_{i-k},F^{-}_{i-k},\bar{F}^{-}_{i-k},\bar{Y}^{-}_{i-k}) \geq$$
$$\geq D_m(\bar{F}^{+}_{i-k};Y^{+}_{i-k}/\bar{F}^{+}_{i-k},F^{-}_{i-k},\bar{F}^{-}_{i-k},\bar{Y}^{-}_{i-k}). \quad (6\text{-}14)$$

Now the probability over the code ensemble that no message in the incorrect subset of the sequential encoding graph be associated with a word $F^{+*}_{i-k} \in S_{\bar{Y}^{+}\bar{Y}^{-}}(F^{+}_{i-k},F^{-}_{i-k},\bar{F}^{+}_{i-k},\bar{F}^{-}_{i-k})$ is greater than or equal to

$$[1 - Pr[S_{\bar{Y}^{+}_{i-k}\bar{Y}^{-}_{i-k}}(F^{+}_{i-k}\ F^{-}_{i-k}\ \bar{F}^{+}_{i-k}\ \bar{F}^{-}_{i-k})]]^{2^{k-1}} . \quad (6\text{-}15)$$

Hence the probability of error averaged over the code ensemble characterized by the Markoff source described in section 6.2 is bounded as in (6-16):

$$P(e) \leq \sum_{\substack{F^{+}_{i-k},F^{-}_{i-k},\bar{F}^{+}_{i-k} \\ \bar{F}^{-}_{i-k},\bar{Y}^{-}_{i-k},\bar{Y}^{-}_{i-k}}} Pr(F^{+}_{i-k},F^{-}_{i-k},\bar{F}^{+}_{i-k},\bar{F}^{-}_{i-k},\bar{Y}^{+}_{i-k},\bar{Y}^{-}_{i-k}) \bullet \left\{1- [1-Pr(S_{\bar{Y}^{+}_{i-k}\bar{Y}^{-}_{i-k}}(F^{+}_{i-k}\ F^{-}_{i-k}\ \bar{F}^{+}_{i-k}\ \bar{F}^{-}_{i-k}))]^{2^{l}}\right. \quad (6\text{-}16)$$

In order to bound the right-hand side of (6-16) conveniently we will now divide all events into two complementary sets T and T^c, according to a distance criterion D_0:

Let the event $(F^+_{i-k}, \overline{F}^+_{i-k}, \overline{Y}^+_{i-k}, \overline{F}^-_{i-k}, F^-_{i-k}, \overline{Y}^-_{i-k}) \in T$

$$\text{if and only if } \frac{1}{kn_0} D_m(F^+_{i-k}; \overline{Y}^+_{i-k} / \overline{F}^+_{i-k}, F^-_{i-k}, \overline{F}^-_{i-k}, \overline{Y}^-_{i-k}) \leq D_0 \tag{6-17}$$

otherwise let the event belong to the complementary set T^c.

We will overbound $\overline{P(e)}$ as follows:

$$\overline{P(e)} \leq \overline{P_1(e)} + \overline{P_2(e)} \tag{6-18}$$

where

$$\overline{P_1(e)} = \sum_T \Pr(F^+_{i-k}, F^-_{i-k}, \overline{F}^+_{i-k}, \overline{F}^-_{i-k}, \overline{Y}^+_{i-k}, \overline{Y}^-_{i-k})$$

$$P_2(e) = \sum_{T^c} 2^{k-1} \Pr(F^+_{i-k}, F^-_{i-k}, \overline{F}^+_{i-k}, \overline{F}^-_{i-k}, \overline{Y}^+_{i-k}, \overline{Y}^-_{i-k}) \cdot$$

$$\cdot \Pr(S_{\overline{Y}^+_{i-k} \overline{Y}^-_{i-k}}(F^-_{i-k}, F^+_{i-k}, \overline{F}^-_{i-k}, \overline{F}^+_{i-k}))$$

Our task will then be to evaluate the two terms on the right-hand side of (6-18).

6.4 A Bound on $\overline{P_2(e)}$

In this and in the following sections we will try to find a bound on the decoding error probabilities $\overline{P_1(e)}$ and $\overline{P_2(e)}$. We will succeed in displaying a region of signalling rate points $(R,\bar{R})$ for which the error probabilities will be exponentially bounded with increasing decoding sequence length kn_0. We will adopt here Shannon's (3) approach to error bounding.

In the process of finding the bound on $\overline{P_2(e)}$ we will be forced to choose the structure of the function $G(\bar{Y}^+_{i-k},\bar{F}^+_{i-k}, F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^-_{i-k})$ appearing in the denominator of the logarithm defining the inverse distance D_m (see equation (6-13)). The choice that we will make will turn out to be convenient for the estimation of $\overline{P_2(e)}$ but will introduce a difficulty into the estimation of $\overline{P_1(e)}$. We will be able to overcome this difficulty in the next section by proving in Appendices III and IV a theorem about the convergence of certain probabilities. As will be seen, there is an obvious way of defining the function $G(\bar{Y}^+_{i-k},\bar{F}^+_{i-k},F^-_{i-k},\bar{F}^-_{i-k},\bar{Y}^-_{i-k})$ which would enable the estimation of $\overline{P_1(e)}$ to be carried out in a routine way. It will be shown at the beginning of section 6.5 that such definition would make the estimation of $\overline{P_2(e)}$ impossible. To simplify further notation we will from this point on, without any loss in generality, change the subscripts (i-k) to subscripts i.

From definition (6-17) and (6-13) it follows that whenever $(F_i^+,\bar{F}_i^+,\bar{Y}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-) \in T^c$ then

$$e^{nD_0}\, G(\bar{Y}_i^+\, F_i^+\, F_i^-\, \bar{F}_i^-\, \bar{Y}_i^-) < Q_m(\bar{Y}_i^+/F_i^+\, \bar{F}_i^+\, F_i^-\, Y_i^-\bar{F}_i^-) \tag{6-19}$$

where we write $n \cong kn_0$

Hence if $F_i^{+*} \in S_{\bar{Y}_i^+\bar{Y}_i^-}(F_i^+F_i^-\bar{F}_i^+\bar{F}_i^-)$; then

$$e^{-nD_0}\quad Q_m(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-\bar{Y}_i^-) > G(Y_i^+,\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-) \tag{6-20}$$

so that

$$e^{-nD_0}\quad \frac{Q_m(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\ \Pr(F_i^{+*})}{G(\bar{Y}_i^+,\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)} > \Pr(F_i^{+*}) \tag{6-21}$$

Summing the right hand side of the above over the set $S_{\bar{Y}_i^+\bar{Y}_i^-}(F_i^+\, F_i^-\, \bar{F}_i^+\, \bar{F}_i^-)$, and the left hand side over the larger set of all F_i^{*+} (which includes the set $S_{\bar{Y}_i^+\bar{Y}_i^-}(F_i^+\, F_i^-\, \bar{F}_i^+\, \bar{F}_i^-)$ we get

$$e^{-nD_0}\quad \sum_{F^{*+}} \frac{Q_m(-\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\ \Pr(F_i^{+*})}{G(\bar{Y}_i^+,\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)} >$$

$$> \Pr(S_{\bar{Y}_i^+\bar{Y}_i^-}(F_i^+,F_i^-,\bar{F}_i^+,\bar{F}_i^-)\ ), \tag{6-22}$$

and therefore, according to the definition (6-18) we get a bound on $\overline{P_2(e)}$:

$$\overline{P_2(e)} < 2^{k-1}\ e^{-nD_0} \sum_{T^c} \frac{\Pr(F_i^+,F_i^-,\bar{F}_i^+,\bar{F}_i^-,\bar{Y}_i^+,\bar{Y}_i^-)}{G(\bar{Y}_i^+,\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)}\ \cdot$$

$$\cdot \sum_{F^{+*}} Q_m(\bar{Y}_i/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\ \Pr(F^{+*}). \tag{6-23}$$

We must now find a bound on the coefficient of $2^{k-1} e^{-nD_0}$ on the right hand side of (6-23). The standard technique of handling this problem is to show that we are dealing with a sum of probabilities whose value cannot exceed one. To do this we must define the function $G(\overline{Y}_i^+,\overline{F}_i^+,F_i^-,\overline{F}_i^-,\overline{Y}_i^-)$ in such a way that it will cancel out of the sum expression, so that the latter will have no denominator. Hence the function G should be a factor of the probability

$$\Pr(F_i^+,F_i^-,\overline{F}_i^+,\overline{F}_i^-,\overline{Y}_i^+,\overline{Y}_i^-) = \Pr(F_i^+/F_i^-,\overline{F}_i^+,\overline{F}_i^-,\overline{Y}_i^+,\overline{Y}_i^-) \cdot \Pr(\overline{Y}_i^+,F_i^-,\overline{F}_i^+,\overline{F}_i^-,\overline{Y}_i^-), \quad (6\text{-}24a)$$

or, more precisely, it should be a factor of the probability $\Pr(\overline{Y}_i^+,\overline{F}_i^+,F_i^-,\overline{F}_i^-,\overline{Y}^-)$. We could equate G with $\Pr(\overline{Y}_i^+,\overline{F}_i^+,F_i^-,\overline{F}_i^-,\overline{Y}_i^-)$ and successfully bound the sum on the right hand side of (6-23). However, such identification would bring us into difficulties when estimating $\overline{P_1(e)}$. It will turn out that the most convenient way to define G is (see also discussions at the beginning of this section and in section 6.5):

$$G(\overline{Y}_i^+,\overline{F}_i^+,F_i^-,\overline{F}_i^-,\overline{Y}_i^-) \equiv \Pr(\overline{Y}_i^+/\overline{F}_i^+,F_i^-,\overline{F}_i^-,\overline{Y}_i^-) \quad (6\text{-}24b)$$

Using the above identification the sum on the right hand side of (6-23) can be written as in (6-24c):

$$\sum_{T^c} \frac{P_r(F_i^+,F_i^-,\bar{F}_i^+,\bar{F}_i^-,\bar{Y}_i^+,\bar{Y}_i^-)}{G(\bar{Y}_i^+,\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)} \sum_{F^{+*}} Q_m(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)Pr(F_i^{+*}) =$$

$$= \sum_{T^c} Pr(F_i^+/F_i^-,\bar{F}_i^+,\bar{F}_i^-,\bar{Y}_i^+,\bar{Y}_i^-)\ Pr(\bar{F}_i^+)\ Pr(F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\ \bullet \tag{6-24c}$$

$$\bullet \sum_{F^{+*}} Q(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\ Pr(F_i^{+*})$$

But from the definition (6-10a) it follows that

$$\sum_{F^{+*}} \sum_{F^+} Pr(F_i^{+*})\ Pr(\bar{F}_i^+) \sum_{\bar{Y}^+} Q(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-) = 1 \tag{6-25}$$

Hence we get:

$$\sum_{T^c} Pr(F_i^+/F_i^-,\bar{F}_i^+,\bar{F}_i^-,\bar{Y}_i^+,\bar{Y}_i^-)\ Pr(\bar{F}_i^+)\ Pr(F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\ \bullet$$

$$\bullet \sum_{F_i^{+*}} Q_m(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\ Pr(F_i^{+*}) \leq$$

$$\leq \sum_{\substack{F_i^+,\bar{F}_i^+,\bar{Y}_i^+,F_i^{+*}\\ F_i^-,\bar{F}_i^-,\bar{Y}_i^-}} Pr(F_i^+/F^-,\bar{F}^+,\bar{F}^-,\bar{Y}^+,\bar{Y}^-)Pr(F_i^-,\bar{F}_i^-,\bar{Y}_i^-)Q_m(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\bullet$$

$$\bullet\ Pr(F_i^{+*})Pr(\bar{F}_i^+) =$$

$$= \sum_{F_i^-,\bar{F}_i^-,\bar{Y}_i^-} Pr(F_i^-,\bar{F}_i^-,\bar{Y}_i^-) \sum_{F_i^{+*},\bar{F}_i^+} Pr(F_i^{+*})Pr(\bar{F}_i^+) \sum_{\bar{Y}_i^+} Q_m(\bar{Y}_i^+/F_i^{+*},\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)\bullet$$

$$\bullet \sum_{F_i^+} Pr(F_i^+/F_i^-,\bar{F}_i^+,\bar{F}_i^-,\bar{Y}_i^+,\bar{Y}_i^-) = 1 \tag{6-26}$$

Substituting the right hand side of (6-26) into (6-23) we get finally the desired bound:

$$\overline{P_2(e)} < 2^{k-1}\, e^{-nD_0} < e^{-n(D_0-R)} \tag{6-27}$$

where we have used the relation (6-5).

We may now conclude that if the cut-off level D_0 is such that $D_0 - R > 0$, the quantity $\overline{P_2(e)}$ will be exponentially bounded as $n \to \infty$.

6.5 A Bound on $\overline{P_1(e)}$

Next we wish to find a bound on $\overline{P_1(e)}$, i.e. on the probability

$$\Pr\left[\frac{1}{n}\log\frac{Q_m(\bar{Y}_i^+/F_i^+\,\bar{F}_i^+\,F_i^-\,\bar{F}_i^-\,\bar{Y}_i^-)}{P_r(\bar{Y}_i^+/\bar{F}_i^+\,F_i^-\,\bar{F}_i^-\,\bar{Y}_i^-)} \leq D_0\right] \tag{6-28}$$

It is clear that we may expand

$$\Pr(\bar{Y}_i^+/\bar{F}_i^+\,F_i^-\,\bar{F}_i^-\,\bar{Y}_i^-) =$$

$$= \prod_{r=i}^{i+k-1}\prod_{s=0}^{n_0-1} \Pr(\bar{y}_{r,s}/\bar{f}_{r,s},\bar{f}_{r,s-1},\ldots,\bar{f}_{i,0},\bar{y}_{r,s-1},\ldots,\bar{y}_{i,0},F_i^-\,\bar{F}_i^-\,\bar{Y}_i^-) \tag{6-29}$$

so that using the definition (6-10a) we can re-write (6-28):

$$\Pr\left[\frac{1}{n}\sum_{r=i}^{i+k-1}\sum_{s=0}^{n_0-1}\log\frac{\Pr(\bar{y}_{r,s}/f_{r,s},\bar{f}_{r,s},\mathscr{f}^m_{r,s},\bar{\mathscr{f}}^m_{r,s},\bar{\mathscr{y}}^m_{r,s},\bar{s}^{-m}_{r,s})}{\Pr(\bar{y}_{r,s}/\bar{f}_{r,s},\ldots,\bar{f}_{i,0},\bar{y}_{r,s-1},\ldots,\bar{y}_{i,0},F_i^-,\bar{F}_i^-,\bar{Y}_i^-)} \leq D_0\right] \tag{6-30}$$

Thus we must find a bound on a distribution of a sum of dependent random variables

$$\log\frac{\Pr(\bar{y}_{r,s}/f_{r,s},\bar{f}_{r,s},\mathscr{f}^m_{r,s},\bar{\mathscr{f}}^m_{r,s},\bar{\mathscr{y}}^m_{r,s},\bar{s}^{-m}_{r,s}}{\Pr(\bar{y}_{r,s}/\bar{f}_{r,s},\ldots,\bar{f}_{i,0},\bar{y}_{r,s-1},\ldots,\bar{y}_{i,0},F_i^-,\bar{F}_i^-,\bar{Y}_i^-)}\,. \tag{6-31}$$

The only known approach for finding such a bound is due to Shannon and is presented in Appendix II. Shannon's procedure is however restricted to cases in which the random variables are associated with transitions of a Markoff process. As the process passes from state to state the sum is generated by successive addition of those random variables which are associated with the transitions that have taken place.

Now, it is evident that the random variables (6-31) cannot be associated with transitions of a Markoff process, because the denominator of the logarithm is a probability of an event conditioned on a sequence of past events, whose length increases with increasing integer r and s. The numerator of the logarithm, on the other hand, is a probability of an event conditioned by a sequence of past events of fixed length m, and causes therefore no trouble. The reason for the introduction of the weaker decoding criterion (6-10a) in place of the criterion (6-3) now becomes apparent; had we used the criterion (6-3) the numerator of the logarithm (6-31) would have been a probability exhibiting a dependency on a potentially infinite sequence of past events. We may therefore rightfully ask whether we could have avoided in a similar way the trouble caused by the denominator of the logarithm (6-31): could we have, perhaps, in section (6-4), defined the function G (see the definition (6-13) of the inverse distance D_m) in a manner more convenient than (6-24b) for our present purposes? In particular , was it possible to identify the function $G(\overline{Y}_i^+, \overline{F}_i^+, F_i^-, \overline{F}_i^-, \overline{Y}_i^-)$ with the probability $P_t'(\overline{Y}_i^+/\overline{F}_i^+, F_i^-, \overline{F}_i^-, \overline{Y}_i^-)$ defined below?

$$P_t'(\bar{Y}_i^+/\bar{F}_i^+ F_i^- \bar{F}_i^- \bar{Y}_i) \equiv \prod_{r=i}^{i+k-1} \prod_{s=0}^{n_0-1} \Pr(\bar{y}_{r,s}/\bar{\ell}^t_{r,s}, \bar{f}_{r,s}, \bar{y}^t_{r,s}, \bar{s}^{-t}_{r,s}) \tag{6-32}$$

where $\bar{s}_{i,0} = \bar{g}(\bar{s}_{00}=0,\ \bar{F}_i^-, \bar{Y}_i^-)$,

$\bar{s}^{-j}_{r,s} = g(\bar{s}^{-j-1}_{r,s},\ \bar{f}^{-j-1}_{r,s},\ \bar{y}^{-j-1}_{r,s})$, and $\underline{t}$ is an arbitrary integer.

(for definition of the function g see equation (4-17) and preceding discussion)

To answer the above question we must return to the discussion of the paragraph following equation (6-23). There it was pointed out that it is necessary for successful bounding of the probability $\overline{P_2(e)}$ that the function $G(\bar{Y}_i^+, \bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i)$ be a factor of the probability $\Pr(\bar{Y}_i^+, \bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i)$. Hence we must ask whether the probability $P_t'(\bar{Y}_i^+/\bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i)$ defined above is a factor of $\Pr(\bar{Y}_i^+, \bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i)$. But we may write

$$\Pr(\bar{Y}_i^+, \bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i) =$$

$$= \Pr(F_i^-, \bar{F}_i^-, \bar{Y}^-) \prod_{r=i}^{i+k-1} \prod_{s=0}^{n_0-1} P(\bar{f}_{r,s}) \Pr(\bar{y}_{r,s}/\bar{f}_{r,s}, \ldots, \bar{f}_{i,0}, \bar{y}_{r,s-1}, \ldots$$

$$\ldots \bar{y}_{i,0}, F_i^-, \bar{F}_i^-, \bar{Y}_i^-), \tag{6-33}$$

and since the probability $\Pr(\bar{y}_{r,s}/\bar{\ell}^{-t}_{r,s}, \bar{f}_{r,s}, \bar{y}^t_{r,s}, \bar{s}^{-t}_{r,s})$ is not in general a factor of the probability $\Pr(\bar{y}_{r,s}/\bar{f}_{r,s}, \ldots, \bar{f}_{i,0}, \bar{y}_{r,s-1}, \ldots, \bar{y}_{i,0}, F_i^-, \bar{F}_i^-, \bar{Y}_i)$, we see by comparing (6-32) with (6-33) that the function $G(\bar{Y}_i^+, \bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i^-)$ cannot be identified with the probability $P_t'(\bar{Y}_i^+/\bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i)$ without causing (insurpassable, as will be seen from the discussion at the end of section 6-5) difficulties in the bounding of the probability $\overline{P_2(e)}$.

The above conclusion brings us back to the beginning of the discussion in the last paragraph. We must resolve how to circumvent the difficulty caused by the character of the probability in the denominator of the logarithm (6-31). We will do so by showing that for a sufficiently large integer t the probabilities $Pr(\bar{y}_{r,s}/\bar{f}_{r,s}, \bar{\ell}^{t}_{r,s}, \bar{y}^{t}_{r,s}, \bar{s}^{-t}_{r,s})$ approximate closely the probabilities $Pr(\bar{y}_{r,s}/\bar{f}_{r,s},..,\bar{f}_{i,0},\bar{y}_{r,s-1},..$ $..\bar{y}_{i,0},F^{-}_{i},\bar{F}^{-}_{i},\bar{Y}^{-}_{i})$ and that we may therefore replace the latter probabilities by the former in the expression of the random variable (6-31). It will be seen that we will thus cause an "epsilon change" in the computed capacity region for the two-way channel.

We will now make our approach more precise:
In Appendix III we prove a theorem giving the sufficient conditions for the existence of an integer t such that for all $t > t_o$ the probability of an event A^i conditioned on a sequence $A^{i-1},A^{i-2},...,A^{i-t}$ of preceding events will differ by an arbitrarily small amount from the probability of the event A^i conditioned on any sequence $A^{i-1},A^{i-2},...,A^{i-t},A^{i-t-1},...,$ A^{i-t-v} where v is any positive integer. This result is applied in Appendix IV to certain events generated by a particular class of Markoff sources. It is shown there in particular that the probabilities (6-29) can result from a process of that class, and that therefore these probabilities behave in a way

satisfying the following theorem:

Theorem 6-1

Given any $\varepsilon > 0$ there exists an integer t_0 such that for all $t > t_0$

$$\begin{aligned}&\left|\Pr(\bar{y}/\bar{f},\bar{f}^{-1},\ldots,\bar{f}^{-t},\bar{y}^{-1},\ldots,\bar{y}^{-t},\bar{s}^{-t}) - \right.\\ &- \Pr(\bar{y}/\bar{f},\bar{f}^{-1},\ldots,\bar{f}^{-t},\bar{f}^{-t-1},\ldots,\bar{f}^{-t-r},\bar{y}^{-1},\ldots \\ &\left. \ldots,\bar{y}^{-t},\ldots,\bar{y}^{-t-r},\bar{s}^{-t-r})\right| < \\ &< \varepsilon \Pr(y/f,f^{-1},\ldots,f^{-t},y^{-1},y^{-t},s^{-t})\end{aligned} \tag{6-34}$$

$r = (1,2,\ldots)$

provided that

$$\Pr(\bar{f},\bar{f}^{-1},\ldots,\bar{f}^{-t-r},\bar{y},\ldots,\bar{y}^{-t-r}) \neq 0$$

and $\bar{s}^{-t} = \bar{g}(\bar{s}^{-t-r},\bar{f}^{-t-r},\ldots,\bar{f}^{-t-1},\bar{y}^{-t-r},\ldots,\bar{y}^{-t-1})$

(i.e. the reverse portion of the derived channel state $\bar{s}^{-t}$ is determined from preceding state portion $\bar{s}^{-t-r}$ by use of the knowledge of symbols $\bar{f}$ and signals $\bar{y}$ which occurred between the time intervals $(-t)$ and $(-t-r)$. g is defined in (4-17).)

If use is made of the expansion (6-29) and of the definition (6-32), it follows from Theorem 6-1 that

$$\begin{aligned}&(1-\varepsilon)^n P_t'(\bar{Y}_i^+/\bar{F}_i^+ F_i^- \bar{F}_i^- \bar{Y}_i^-) \leq \\ &\leq \Pr(\bar{Y}_i^+/\bar{F}_i^+ F_i^- \bar{F}_i^- \bar{Y}_i^-) \leq (1+\varepsilon)^n P_t'(\bar{Y}_i^+/\bar{F}_i^+ F_i^- \bar{F}_i^- \bar{Y}_i^-)\end{aligned} \tag{6-35}$$

where $n \equiv n_0 k$.

Thus, using also definition (6-10a), we may write for i sufficiently large

$$\frac{1}{n} \log \frac{Q_m(\bar{Y}_i^+/\bar{F}_i^+ \; F_i^+ \; F_i^- \; \bar{F}_i^- \; \bar{Y}_i^-)}{\Pr(\bar{Y}_i^+/\bar{F}_i^+ \; F_i^- \; \bar{F}_i^- \; \bar{Y}_i^-)} \geq$$

$$\geq \log \frac{1}{1+\varepsilon} + \frac{1}{n} \log \prod_{r=i}^{i+k^{+1}} \prod_{s=0}^{n_\theta - 1} \left[\frac{\Pr(\bar{y}_{r,s}/f_{r,s}, \bar{f}_{r,s}, \ell^m_{r,s}, \bar{\ell}^m_{r,s}, \bar{y}^n_{r,s}, \bar{s}^{-m}_{r,s})}{\Pr(\bar{y}_{r,s}/\bar{f}_{r,s}, \bar{\ell}^t_{r,s}, \bar{y}^{\pm}_{r,s}, \bar{s}^{-t}_{r,s})} \right] \tag{6-36}$$

and consequently, again for i sufficiently large,

$$\Pr\left\{ \frac{1}{n} \log \frac{Q_m(\bar{Y}_i^+/F_i^+, \bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i^-)}{\Pr(\bar{Y}_i^+/\bar{F}_i^+ \; F_i^- \; \bar{F}_i^- \; \bar{Y}_i^-)} \leq D_0 \right\} \leq$$

$$\leq \Pr\left\{ \frac{1}{n} \log \prod_{r=i}^{i+k-1} \prod_{s=0}^{n_\theta - 1} \frac{\Pr(\bar{y}_{r,s}/f_{r,s}, \bar{f}_{r,s}, \ell^m_{r,s}, \bar{\ell}^m_{r,s}, \bar{y}^m_{r,s}, \bar{s}^{-m}_{r,s})}{\Pr(\bar{y}_{r,s}/\bar{f}_{r,s}, \bar{\ell}^t_{r,s}, \bar{y}^{\pm}_{r,s}, \bar{s}^{-t}_{r,s})} \leq D_0 + \delta \right\} \tag{6-37}$$

where $\delta = \log(1+\varepsilon)$

Our task will therefore be accomplished by finding a bound on the right hand side of (6-33). This should be possible by the method of Appendix II, since the newly obtained random variable can be thought of as a cumulative sum of random variables generated by the Markoff process described below.

However, let us first make a slight adjustment which will simplify the description of the mentioned process. Since the probability $\Pr(\bar{Y}_i^+/\bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i^-)$ does not involve the integer m used in the definition (6-10a) of the probability $Q_m(\bar{Y}_i^+/F_i^+,\bar{F}_i^+,F_i^-,\bar{F}_i^-,\bar{Y}_i^-)$, we may, instead of bounding the right hand side of (6-37), bound the probability

$$Pr\left\{\frac{1}{n}\log\prod_{r=i-k}^{i-1}\prod_{s=0}^{n_0-1}\frac{Pr(\bar{y}_{r,s}/f_{r,s},\bar{f}_{r,s},\mathcal{f}^M_{r,s},\bar{\mathcal{f}}^M_{r,s},\bar{\mathcal{y}}^M_{r,s},\bar{s}^{-M}_{r,s})}{Pr(\bar{y}_{r,s}/\bar{f}_{r,s},\bar{\mathcal{f}}^M_{r,s},\bar{\mathcal{y}}^M_{r,s},\bar{s}^{-M}_{r,s})}\leq\right.$$

where $M \equiv \max\ (m,t)$ $$\left.\leq D_0+\delta\right\} \quad (6\text{-}38)$$

If M = m, then the correctness of the relation (6-37) will not be affected if we substitute in it M for t. If, on the other hand, M = t, we may change, without any detriment to decoding, the definition (6-10a) by substituting M for m.

Consider then a Markoff process having states characterized by the different possible sequences.

$$(\mathcal{f}^M,\bar{\mathcal{f}}^M,\mathcal{y}^M,\bar{\mathcal{y}}^M,s^{-M},\bar{s}^{-M}) \equiv W_M \quad (6\text{-}39)$$

We can make the state characterizations clearer by displaying the left hand side of (6-39) as an array whose columns are the sequences $\mathcal{f}^M,\bar{\mathcal{f}}^M,\mathcal{y}^M,\bar{\mathcal{y}}^M$. Hence we can write

$$\left.\begin{array}{l} f^{-1},\ y^{-1},\ \bar{f}^{-1},\ \bar{y}^{-1} \\ f^{-2},\ y^{-2},\ \bar{f}^{-2},\ \bar{y}^{-2} \\ \vdots \\ f^{-M},\ y^{-M},\ \bar{f}^{-M},\ \bar{y}^{-M} \\ s^{-M},\ \bar{s}^{-M} \end{array}\right\} \equiv W_M \quad (6\text{-}40)$$

The states W'_M into which a transition from state W_M exists are representable by

$$\left.\begin{matrix} f, & y, & \bar{f}, & \bar{y}, \\ f^{-1}, & y^{-1}, & \bar{f}^{-1}, & \bar{y}^{-1} \\ \vdots & & & \\ f^{-M+1}, & y^{-M+1}, & \bar{f}^{-M+1}, & \bar{y}^{-M+1} \\ s^{-M+1}, & \bar{s}^{-M+1} & & \end{matrix}\right\} = \left\{\begin{matrix} (f')^{-1}, & (y')^{-1}, & (\bar{f}')^{-1}, & (\bar{y}')^{-1} \\ (f')^{-2}, & (y')^{-2}, & (\bar{f}')^{-2}, & (\bar{y}')^{-2} \\ \vdots & & & \\ (f')^{-M}, & (y')^{-n}, & (\bar{f}')^{-M}, & (\bar{y}')^{-M} \\ (s')^{-M} = f^{-M}(s^{-M}), & (\bar{s}')^{-} = \bar{f}^{-M}(\bar{s}^{-M}) & & \end{matrix}\right\} \equiv W'_M \qquad (6\text{-}41)$$

There are therefore $2^{2(2^{2\ell}+1)}$ possible transitions out of any state W'_M, each leading to a different state W'_M. The transition probabilities will then be defined:

$$P_{W_M}(W'_M) = \begin{cases} P((y')^{-1}/(x')^{-1}(\bar{x}')^{-1})\bar{P}((\bar{y}')^{-1}/(x')^{-1}(\bar{x}')^{-1}) \\ \qquad P((f')^{-1})\bar{P}((\bar{f}')^{-1}) \qquad (6\text{-}42) \\ \text{If } W'_M \text{ and } W_M \text{ are representable as in (6-40) and} \\ \text{(6-41), respectively. In the above product} \\ (x')^{-1} = (f')^{-1}((s')^{-1});\ (\bar{x}')^{-1} = (\bar{f}')^{-1}((\bar{s}')^{-1}) \\ 0 \text{ otherwise.} \end{cases}$$

If we compare the Markoff process defined above with the one of section 6.2, and in particular if we compare equation (6-42) with equation (6-9), we see that the probabilities of the functions and signals attached to the transitions are identical. We may thus compute the bound on $\overline{P_1(e)}$ by using the process described by equation (6-39). Thus to each transition from state W_M to state W'_M, (the states being described as in (6-40) and (6-41), respectively) we shall associate the random variable

$$\log \frac{\Pr((\bar{y}')^{-1}/(f')^{-1},(\bar{f}')^{-1},\boldsymbol{f}^M,\bar{\boldsymbol{f}}^M,\bar{\boldsymbol{y}}^{-M},\bar{s}^{-M})}{\Pr((\bar{y}')^{-1}/(\bar{f}')^{-1},\bar{\boldsymbol{f}}^M,\bar{\boldsymbol{y}}^{M},\bar{s}^{-M})} \tag{6-43}$$

and estimate a bound on the distribution (6-38) by use of the methods of Appendix II. Identifying the distribution (6-38) with the function $G(n,n(D_0+\delta))$, we get from equation (A-II-37):

$$G(n,n\gamma'(t)) \leq g e^{-n(t\gamma'(t)-\gamma(t))} \tag{6-44}$$

where the quantities $\gamma(t)$ and $\gamma'(t)$ are defined in equations (A-II-9) and (A-II-19), respectively. The identification

$$\gamma'(t) = D_0 + \delta \tag{6-45}$$

is then made. We will not bother to make a complete translation here into the notation of Appendix II. It is straightforward but quite complicated. We will only observe that the exponent on the right hand side of (6-44) is in the range $t \leq 0$ equal to zero for $t = 0$ and is negative otherwise, and that in the range $t \leq 0$ the threshold $\gamma'(t)$ increases with increasing t, and is largest when $t = 0$, as can be seen from equations (A-II-28) and (A-II-20). But

$$\gamma'(0) = E\left\{\log \frac{\Pr(\bar{y}/f,\bar{f},\boldsymbol{f}^M,\bar{\boldsymbol{f}}^M,\bar{\boldsymbol{y}}^M,\bar{s}^{-M})}{\Pr(\bar{y}/\bar{f},\bar{\boldsymbol{f}}^M,\bar{\boldsymbol{y}}^M,\bar{s}^{-M})}\right\} = \tag{6-46}$$

$$= E\left\{\log \frac{\Pr(\bar{y}/f,f^{-1},\ldots,f^{-M},\bar{f},\bar{f}^{-1},\ldots,\bar{f}^{-M},\bar{y}^{-1},\ldots\bar{y}^{-M},\bar{s}^{-M})}{\Pr(\bar{y}/\bar{f},\bar{f}^{-1},\ldots,\bar{f}^{-M},\bar{y}^{-1},\ldots,\bar{y}^{-M},\bar{s}^{-M})}\right\}$$

where the expectation is to be computed over the probability distribution of the Markoff process described by (6-41), or, equivalently, over the distribution of the process described in section 6.2, in particular by the transition probabilities (6-9).

The symbol g in (6-44) stands for the number of states in the process and we have

$$g = 2^{M(2^{2\ell+1}+1)} + 4\ell \tag{6-47}$$

Although g is a very large number indeed, it is independent of the code length n, and therefore as long as $(t\gamma'(t)-\gamma(t))\geq 0$, the right hand side of (6-44) can be made arbitrarily small by use of a sufficiently large n. Thus we may conclude that $\overline{P_1(e)}$ is exponentially bounded whenever the quantity (6-46) is larger than $D_0 + \delta$. δ can be made, by a sufficiently large M, as small as desired. Taking into account the results of section 6.4, in particular equation (6-27), it can be stated that as $n \rightarrow \infty$ the average probability of the reverse decoding error decreases exponentially, provided that the forward rate R satisfied the inequality

$$R < -\log\,(1+\varepsilon) + E\left\{\log \frac{\Pr(\bar{y}/f,\bar{f},\varphi^M,\bar{\varphi}^M,\bar{\gamma}^M,\bar{s}^{-M})}{\Pr(\bar{y}/\bar{f},\bar{\varphi}^M,\bar{\gamma}^M,\bar{s}^{-M})}\right\} \tag{6-48}$$

where the positive quantity $\varepsilon > 0$ is a function of M and is determined by the equation (6-34).

It is obvious that since in principle the forward direction differs in no way from the reverse one, all the statements in sections 6.3, 6.4 and 6.5 made with respect to reverse decoding of forward messages apply as well to forward decoding of reverse messages. Hence the probability of forward decoding errors decreases exponentially with increasing n, provided that the reverse rate $\bar{R}$ satisfies the inequality

$$\bar{R} < -\log\ (1+\bar{\varepsilon}) + E\left\{\frac{\Pr(y/f,\bar{f},\mathfrak{f}^M,\bar{\mathfrak{f}}^M,\mathfrak{y}^M,s^{-M})}{\Pr(y/f,\mathfrak{f}^M,\mathfrak{y}^M,s^{-M})}\right\} \qquad (6\text{-}49)$$

where $\bar{\varepsilon} > 0$ is a quantity analogous to ε, pertaining to the estimation of $\Pr(y/f,\mathfrak{f}^M,\mathfrak{y}^M,s^{-M})$ in the manner of equation (6-34).

It might be stressed here than in the foregoing argument we have not attempted to obtain the best possible bound on the probability of the decoding error when the signalling rates R and $\bar{R}$ satisfy equations (6-48) and (6-49), respectively. We have, for instance, made no attempt to determine the proper choice of the cutoff threshold D_0. Thus we are leaving undecided the question of the best obtainable exponent combination in the probability of forward transmission error bound

$$\overline{P(e)} \leq \overline{P_1(e)} + \overline{P_2(e)} \leq e^{-n(\gamma'(t)-\gamma-R)} + g e^{-n(t\gamma'(t)-\gamma(t))} \qquad (6\text{-}50)$$

in cases where the rate R is considerably smaller than $(\gamma'(0)-\gamma)$. Rather, we have limited ourselves to displaying the region of points $(R,\bar{R})$ for which the probabilities of error are exponentially bounded with increasing n.

6.6 The Main Coding Result

We will state here the main result of the present article:

Theorem 6-2

Given any binary memoryless two-way channel having input signals x and $\bar{x}$, and output signals y and $\bar{y}$, characterized by the transmission probabilities

$$Pr(y,\bar{y}/x,\bar{x}) = p(y/x\bar{x})\ \bar{p}(y/x\bar{x}). \qquad (6\text{-}51)$$

For an arbitrary dependence length ℓ a derived two-way channel can be constructed as in Figure 4-2 . For any pair of positive numbers ε, $\bar{\varepsilon}$ and for any probability distributions $P(f)$ and $\bar{P}(\bar{f})$ over the input symbols of the derived channel, an integer M can be found such that for all integers $m > M$

$$|Pr(y/f,\mathscr{f}^m,\mathscr{y}^m,s^{-m})-Pr(y/f,\mathscr{f}^m,\mathscr{y}^m,\mathscr{f}^r_{-m},\mathscr{y}^r_{-m},s^{-m-r})| < \varepsilon\ Pr(y/f,\mathscr{f}^m,\mathscr{y}^m,s^{-m})$$

$$|Pr(\bar{y}/\bar{f},\bar{\mathscr{f}}^m,\bar{\mathscr{y}}^m,\bar{s}^{m})-Pr(\bar{y}/\bar{f},\bar{\mathscr{f}}^m,\bar{\mathscr{y}}^m,\bar{\mathscr{f}}^r_{-m},\bar{\mathscr{y}}^r_{-m},\bar{s}^{-m-r})| < \varepsilon\ Pr(\bar{y}/\bar{f},\bar{\mathscr{f}}^m,\bar{\mathscr{y}}^m,\bar{s}^{-m}) \qquad (6\text{-}52)$$

whenever

$$\begin{aligned} &Pr(f,\mathscr{f}^m,f^{m-1},\ldots,f^{-m-r},y,\mathscr{y}^m,y^{-m-1},\ldots,y^{-m-r},s^{-m-r}) \neq 0 \\ &Pr(\bar{f},\bar{\mathscr{f}}^m,\bar{f}^{-m-1},\ldots,\bar{f}^{-m-r},\bar{y},\bar{\mathscr{y}}^m,\bar{y}^{-m-1},\ldots,\bar{y}^{-m-r},\bar{s}^{-m-r}) \neq 0 \end{aligned} \qquad (6\text{-}53)$$

For the above two-way channel, for all $m > M$, codes signalling simultaneously in the forward direction at a rate R and in the reverse direction at a rate $\bar{R}$ such that

$$R < -\log(1+\varepsilon) + E\left\{\log \frac{Pr(\bar{y}/f,\bar{f},\ell^m,\bar{\ell}^m,\bar{y}^m,\bar{s}^{-m})}{Pr(\bar{y}/\bar{f},\bar{\ell}^m,\bar{y}^m,\bar{s}^{-m})}\right\} \quad (6\text{-}54)$$

$$\bar{R} < -\log(1+\bar{\varepsilon}) + E\left\{\log \frac{Pr(y/f,\bar{f},\ell^m,\bar{\ell}^m,y^m,s^{-m})}{Pr(y/f,\ell^m,y^m,s^{-m})}\right\} \quad (6\text{-}55)$$

can be found, for which the probabilities of reverse and forward decoding errors will decrease exponentially with increasing code length n. Hence for a sufficiently large n there will exist codes signalling at rates R, $\bar{R}$ whose probability of decoding error will be as small as desired.

The consequences and implications of the above theorem will be discussed in the following Articles. The author, however, wishes to make one final comment on the expressions on the right hand sides of inequalities (6-54) and (6-55), and on the technique used to obtain them:

From the discussion preceding equation (6-38) it follows that the right hand side of (6-54) could be written as in (6-56)

$$-\log(1+\varepsilon(t)) + E\left\{\log \frac{Pr(\bar{y}/f,\bar{f},\ell^m,\bar{\ell}^m,\bar{y}^m,\bar{s}^{-m})}{Pr(\bar{y}/\bar{f},\bar{\ell}^t,\bar{y}^t,\bar{s}^{-t})}\right\}, \quad (6\text{-}56)$$

where the integers t and m are controllable independently. As m is increased, the value of the expression (6-56) increases or stays the same, since it is a well known result (see for instance Fano (4), Equation (4.11)) that

$$-E\left\{\log Pr^{-1}(\bar{y}/f,\bar{f},\ell^m,\bar{\ell}^m,\bar{y}^m,\bar{s}^{-m})\right\} \leq$$
$$\leq -E\left\{\log Pr^{-1}(\bar{y}/f,\bar{f},\ell^{m+1},\bar{\ell}^{m+1},\bar{y}^{m+1},\bar{s}^{-m+1})\right\}. \quad (6\text{-}57)$$

This is as we would expect it, since an increase in m corresponds to an improvement of the decoding criterion Q_m defined in (6-10a). Indeed, as $m \rightarrow \infty$ the decoding criterion Q_m becomes optimal, and we therefore conclude that our final result was not weakened at all by the temporary adoption of the less than optimal decoding criterion, while it was this adoption which enabled us to carry the random argument through in the only way known so far: by use of the Shannon method for bounding distributions of sums of random variables generated by a Markoff process (see Appendix II).

The result (6-56) also shows conclusively that in section 6.4 it was not possible to identify the function $G(\bar{Y}_i^+, \bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i^-)$, used in the definition (6-13) of the inverse distance D_m, with the probability $P_t'(\bar{Y}_i^+/\bar{F}_i^+, F_i^-, \bar{F}_i^-, \bar{Y}_i^-)$ defined in (6-32). We recall from the discussion of section 6.5 that such identification would have made the bounding of $\overline{P_1(e)}$ easy. In fact, it would have resulted in the elimination of the term $-\log(1+\varepsilon(t))$ from the expression (6-56). But such elimination would cause an increase of the capacity region of the two-way channel. It should further be noted that

$$E\left\{\log \frac{1}{\Pr(\bar{y}/\bar{f}, \bar{\mathfrak{f}}^{t}, \bar{\mathfrak{y}}^{t}, \bar{s}^{-t})}\right\} \leq$$

$$\leq E\left\{\log \frac{1}{\Pr(\bar{y}/\bar{f}, \bar{\mathfrak{f}}^{t-1}, \bar{\mathfrak{y}}^{t-1}, \bar{s}^{-t})}\right\} \qquad (6\text{-}58)$$

and hence if G were to be identified with P'_t and if then a bounding of $\overline{P_2(e)}$ were possible, the capacity region of the two-way channel could be increased by the simple strategem of choosing t=0. This is absurd and hence it follows that with this choice of G the probability $\overline{P_2(e)}$ cannot be bounded.

Finally we wish to note that it follows from Appendix IV that as $t\to\infty$, $\varepsilon(t)\to 0$. Combining this with the conclusion reached in the paragraph before last, we see that the capacity region includes the convex hull of points

$$E\left\{\log \frac{\Pr(\bar{y}/(f,f^{-1},..)(\bar{f},\bar{f}^{-1},..)(\bar{y}^{-1},..))}{\Pr(\bar{y}/(\bar{f},\bar{f}^{-1},..)(\bar{y}^{-1},..))}\right\}, \tag{6-59}$$

$$E\left\{\log \frac{\Pr(y/(f,f^{-1},..)(\bar{f},\bar{f}^{-1},..)(y^{-1},..))}{\Pr(y/(f,f^{-1},..)(y^{-1},...))}\right\}$$

obtainable by different assignments of probabilities P(f) and $\bar{P}(\bar{f})$. However, the conclusion embodied in (6-59) has no practical value since, because of the infinite dependence implied, the expectation expressions cannot be evaluated.

7. Coding Loss and Bounds on the Capacity Region

In the preceding Article we proved a coding theorem which can be interpreted as giving a lower bound on the two-way channel capacity region. It involves evaluation of the right hand sides of equations (6-54) and (6-55) for some input probability distributions of the derived channel's input symbols, $P(f)$ and $\bar{P}(\bar{f})$. We would like to investigate these distributions.

7.1 Interpretation of Capacity Bound Expression

Consider the derived channel of Figure 4-2 to whose input terminals are attached function sources generating input symbols independently at random, with probabilities $P(f)$ and $\bar{P}(\bar{f})$. In what follows we will switch from sequential to block notation, because of its greater simplicity (see in particular section 4.1). The probability of the symbol sequence F_i^+ will then be:

$$\Pr(F_i^+) = \prod_{j=i}^{i+n-1} P(f_j) \tag{7-1}$$

and similarly, the probability of a sequence $\bar{F}_i^+$ will be:

$$\Pr(\bar{F}_i^+) = \prod_{j=1}^{i+n-1} \bar{P}(\bar{f}_j) \tag{7-2}$$

Consider the information about the forward source transmitted through the channel when a given sequence $\bar{F}_i=\bar{F}_i^+,\bar{F}_i^-$ is transmitted from the reverse source and a given sequence $\bar{Y}_i = \bar{Y}_i^+,\bar{Y}_i^-$ is received. It is:

$$I\ (\bar{Y}_i;F_i/\bar{F}_i) = \log \frac{Pr(\bar{Y}_i/F_i,\bar{F}_i)}{Pr(\bar{Y}_i/\bar{F}_i)} \quad . \tag{7-3}$$

But we have

$$\lim_{n\to\infty} E\left\{\frac{1}{n}\log \frac{Pr(\bar{Y}_0/F_0,\bar{F}_0)}{Pr(\bar{Y}_0/\bar{F}_0)}\right\} = \tag{7-4}$$

$$= \lim_{n\to\infty} \frac{1}{n} E\left\{\sum_{j=1}^{n} \log \frac{Pr(\bar{y}_j/f_j,f_{j-1},..,f_1,\bar{f}_j,\bar{f}_{j-1},..,\bar{f}_1,\bar{y}_{j-1},...,\bar{y}_1)}{Pr(\bar{y}_j/\bar{f}_j,\bar{f}_{j-1},..,\bar{f}_1,\bar{y}_{j-1},...,\bar{y}_1)}\right\} =$$

$$= \lim_{n\to\infty} E\left\{\log \frac{Pr(\bar{y}_n/f_n,f_{n-1},..,f_1,\bar{f}_n,\bar{f}_{n-1},..,\bar{f}_1,\bar{y}_{n-1},..,\bar{y}_1)}{Pr(\bar{y}_n/\bar{f}_n,\bar{f}_{n-1},..,\bar{f}_1,\bar{y}_{n-1},..,\bar{y}_1)}\right\}$$

So that we may conclude that the per symbol expectations of information passing through the channel in the forward and reverse directions, respectively are

$$\lim_{n\to\infty} E\left\{\frac{1}{n} I(\bar{Y}_o;F_o/\bar{F}_o)\right\} = \lim_{n\to\infty} E\left\{\log \frac{Pr(\bar{y}_n/f_n,..,f_1,\bar{f}_n,..,\bar{f}_1,\bar{y}_{n-1},..,\bar{y}_1)}{Pr(\bar{y}_n/\bar{f}_n,..,\bar{f}_1,\bar{y}_{n-1},..,\bar{y}_1)}\right\} \tag{7-5}$$

and

$$\lim_{n\to\infty} E\left\{\frac{1}{n} I(Y_0;\bar{F}_0/F_0)\right\} = \lim_{n\to\infty} E\left\{\log \frac{Pr(y_n/f_n,..,f_1,\bar{f}_n,..,\bar{f}_1,y_{n-1}..y_1)}{Pr(y_n/f_n,..,f_1,y_{n-1},..y_1)}\right\} \tag{7-6}$$

Examining (7-5) and (6-54) it becomes obvious that the two expectation functions appearing in these equations would be identical if the decoding criterion (6-10a) were to be ideal. The same is true for equations (7-6) and (6-55). Hence we see that the coding theorem 6-2 provided us with a result which our experience would have led us to expect: the arbitrarily low error rate bound is associated in the usual manner with the information passing through the channel. This of course suggests very strongly that the capacity region of the channel is associated in the usual manner with the average mutual information for which the successive input symbols to the derived channel are not necessarily independent of each other. Indeed, examination of the argument in Article 6 leading to Theorem 6-2 would show that the random code could have been generated by symbol probability assignments $P(f/f^{-1},f^{-2},\ldots,f^{-\nu})$ and $\bar{P}(\bar{f}/\bar{f}^{-1},\bar{f}^{-2},\ldots,\bar{f}^{-\bar{\nu}})$ where ν and $\bar{\nu}$ is any pair of positive integers. Hence the capacity region includes the convex hull of points $\lim_{n\to\infty} E\left\{\frac{1}{n} I(\bar{Y}_0;F_0/\bar{F}_0)\right\}$, $\lim_{n\to\infty} E\left\{\frac{1}{n} I(Y_0;\bar{F}_0/F_0)\right\}$ obtainable by different assignments $P(f/f^{-1},\ldots,f^{-\nu})$ and $\bar{P}(\bar{f}/\bar{f}^{-1},\ldots,\bar{f}^{-\bar{\nu}})$ as ν and $\bar{\nu}$ vary over all positive integers. We did not prove Theorem 6-2 in all its possible breadth because we believe that to do so would have necessitated consideration of issues extraneous to the main goal of this thesis: the clarification of the essential differences between cooperative and non-cooperative coding.

We will therefore study the expressions (7-5) and (7-6) further, eliminating the restriction of independent assignment of input symbols f and $\bar{f}$. In particular we shall try to learn:

(a) The relationship of Shannon's lower bound to the lower bound of Theorem 6-2.

(b) Whether the information flow through the channel can be improved by removing the independence restriction on the input symbols to the derived channel.

(c) Whether one can interpret Theorem (6-2) and the expressions (7-5) and (7-6) in some way that could prove a guide for appropriate code construction for binary two-way channels.

In this Article we plan to deal with points (a) and (b) above, and we will leave point (c) for Article 9.

7.2 Derived Channel and Signal Source Interpretation

Consider the derived channel of Figure 4-2 to which stationary sources have been attached generating successive symbols independently with probabilities $P(f)$ and $\bar{P}(\bar{f})$. The situation is pictured in Figure 7-1.

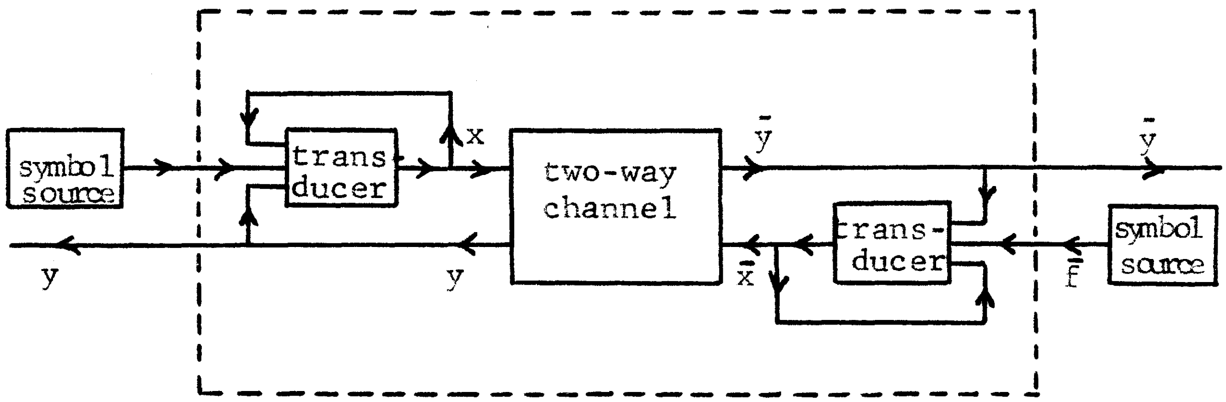


Figure 7-1.

Consider next the binary two-way channel embedded inside the derived channel of Figure 7-1. If one were to consider the input and output signals to the two-way channel only, then the process could be represented by Figure 7-2 in which the signal sources have random outputs depending on the identity of some combinations of previous input and output signals at the terminals of the given two-way channel.

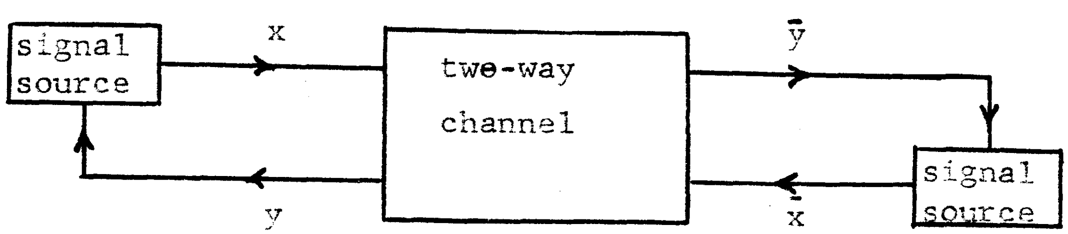


Figure 7-2.

It would be helpful to find out what the signal-source probabilities of Figure 7-2 are for given symbol source probabilities of Figure 7-1. In this respect we can prove the following:

Theorem 7-1

The two-way channel input signal x will depend on a finite number ℓ of immediately preceding channel input and output signals $(x^{-1},\ldots,x^{-\ell},y^{-1},\ldots,y^{-\ell})$ if and only if the successive input symbols f to the derived channel are statistically independent of each other. The integer ℓ is equal to the dependency length of the symbols f used.

Proof:

Comparing Figures 7-1 and 7-2 we see that

$$Pr(x/(x^{-1},x^{-2},\ldots),(y^{-1},y^{-2},\ldots)) = \sum_{f \ni f(x,y)=x} Pr(f/(x^{-1},x^{-2},\ldots),(y^{-1},y^{-2}\ldots)). \tag{7-7}$$

As seen from Figure 7-1, any function f in the above equation can depend on channel signals x^{-j} and y^{-j} only in so far as these are indicative of symbols f^{-i} generated in the past. It is further seen from the summation in (7-7) that even if successive f symbols are independent, the signal x will depend on the immediately preceding input and output signals $(x^{-1},\ldots,x^{-\ell},y^{-1},\ldots,y^{-\ell})$, since on these depends the choice of symbols f to be included in the summation. Similarly, the identity of symbols f^{-1} which could have caused x^{-1} depends on the configuration $(x^{-2},..,x^{-\ell-1},y^{-2},..,y^{-\ell-1})$. Thus if

$$Pr(f^{-1}\ f) \neq P(f^{-1})\ P(f) \tag{7-8}$$

then we can write

$$Pr(f/(x^{-1},x^{-2},...)(y^{-1},y^{-2})) =$$

$$= \sum_{f^{-1} \ni f^{-1}(\mathcal{X}^{-1},\mathcal{Y}^{-1})=x^{-1}} \ \sum_{f^{-2} \ni f^{-2}(\mathcal{X}^{-2},\mathcal{Y}^{-2})=x^{-2}} \cdots\cdots$$

$$\cdots\cdots \sum_{f^{-i} \ni f^{-i}(\mathcal{X}^{-i},\mathcal{Y}^{-i})=x^{-i}} Pr(f/f^{-1},..,f^{-i},(x^{-i-1},..),(y^{-i-1},..)) \cdot$$

$$\cdot\, Pr(f^{-1}/f^{-2},..,f^{-i},(x^{-i-1},..)(y^{-i-1},..))\cdots\cdots$$

$$\cdot\, Pr(f^{-i}/(x^{-i-1},x^{-i-2},..),(y^{-i-1},y^{-i-2})) \tag{7-9}$$

It thus may be seen from (7-9) that as soon as successive symbols are dependent, the signal x will depend on the entire infinite past of received and transmitted signals $(x^{-1},x^{-2},...)$, $(y^{-1},y^{-2},...)$. If, however, the symbol source in Figure 7-1 generates successive symbols independently, then it is certainly true that

$$Pr(x/(x^{-1},x^{-2}..)(y^{-1},y^{-2},..)) = \sum_{f \ni f(\mathcal{X},\mathcal{Y})=x} P(f) \tag{7-10}$$

and thus in such a case x will depend on the finite past $(\mathcal{X},\mathcal{Y})$ only.

Q.E.D.

In the following sections we will deal with sources generating successive symbols independently. In view of equation (7-10) it will be useful to define a set of probabilities $\left\{ q_{\mathbf{xy}}(x) \, , \, \bar{q}_{\bar{\mathbf{x}}\bar{\mathbf{y}}}(\bar{x}) \right\}$ which will determine the operation of the signal sources of Figure 7-2. Let

$$q_{\mathbf{xy}}(x) \equiv \sum_{f \ni f(\mathbf{xy})=x} P(f) \qquad \bar{q}_{\bar{\mathbf{x}}\bar{\mathbf{y}}}(\bar{x}) \equiv \sum_{\bar{f} \ni \bar{f}(\bar{\mathbf{x}}\bar{\mathbf{y}})=\bar{x}} \bar{P}(\bar{f}) \, . \qquad (7\text{-}11)$$

7.3 Information Flow Through the Two-Way Channel

We will here consider the channel of Figure 7-2 for which the signal sources are defined by the set of probabilities (7-11).

We would like to ask what the average information flow is through the channel in the forward and reverse directions, respectively. The information available at the forward terminal about the actual operation of the reverse source is provided by the identity of the signal sequences X and Y, where the implied subscript "0" was left out (i.e. for instance, $X = x_1, x_2, \ldots, x_n$). However, X depends on $\bar{X}$ only because X depends on Y through the signal source, and Y depends on $\bar{X}$. Hence all the information about the output of the reverse source is provided by the received signal sequence Y and depends on X only in so far as the latter sequence modifies the noise in the channel. The manner of generation of X must then be irrelevant to the determination of information flow, which must hence be computable from the situation represented by Figure 7-3.

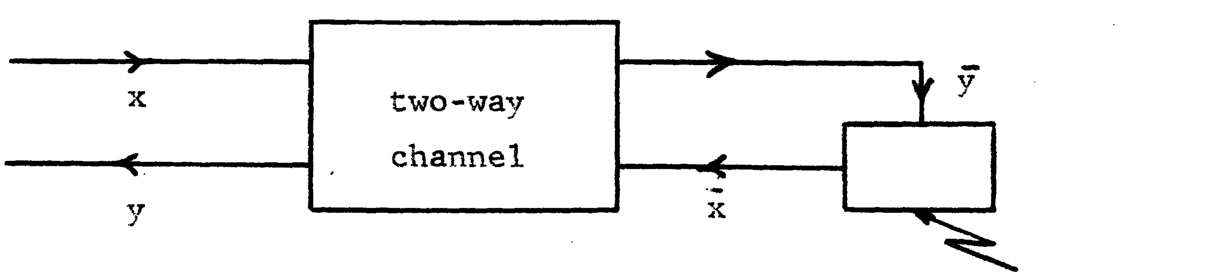


Figure 7-3.

Of course, when averaging the information flow, the weight to be given to the information measure involving quantities X, Y and $\bar{X}$ must be the probability $Pr(X,Y,\bar{X})$ derived from the operation of the channel of Figure 7-2. We therefore conclude that the information flow in the reverse direction is to be measured by the quantity

$$I'(Y;\bar{X}/X) = \log \frac{Pr'(Y/\bar{X},X)}{Pr'(Y/X)} \tag{7-12}$$

where the prime indicates that the conditioning sequence X is fixed independently of the other sequences.

Similarly, the information flow in the forward direction is measured by the quantity

$$I''(\bar{Y};X/\bar{X}) = \log \frac{Pr''(\bar{Y}/\bar{X}X)}{Pr''(\bar{Y}/\bar{X})} \quad . \tag{7-13}$$

The average information flow in the reverse direction through the channel having sources as in Figure 7-2 is

$$E\left\{I'(Y;\bar{X}/X)\right\} = \sum_{XY\bar{X}} Pr(XY\bar{X}) \log \frac{Pr'(Y/X\bar{X})}{Pr'(Y/X)} \tag{7-13a}$$

where

$$Pr(XY\bar{X}) = \sum_{\bar{Y}} Pr(XY\bar{X}\bar{Y}) =$$

$$= \sum_{\bar{Y}} \prod_{i=1}^{n} \underset{\mathcal{X}_i\mathcal{Y}_i}{q(x_i)} \; \underset{\bar{\mathcal{X}}_i\bar{\mathcal{Y}}_i}{\bar{q}(\bar{x}_i)} \; p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i) . \tag{7-14}$$

where $\mathcal{X}_1 = \mathcal{Y}_1 = \bar{\mathcal{X}}_1 = \bar{\mathcal{Y}}_1 = 0$ by convention.

In (7-14) and from here on the fulfillment of the restriction $P(y,\bar{y}/x\bar{x}) = p(y/x\bar{x})\bar{p}(\bar{y}/x\bar{x})$ assumed in (3-1) is essential. Next,

$$P'r(X Y \bar{X}) = \sum_{\bar{Y}} P'r(X Y \bar{X} \bar{Y}) = \sum_{\bar{Y}} \prod_{i=1}^{n} P'(x_i)\, \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)}\;\; p(y_i/x_i \bar{x}_i)\, \bar{p}(\bar{y}_i/x_i \bar{x}_i) \tag{7-15}$$

where P'( ) is an arbitrary distribution.

Hence

$$P'r(Y/X\bar{X}) = \frac{P'r(YX\bar{X})}{\sum_{Y} P'r(YX\bar{X})} = \prod_{i=1}^{n} p(y_i/x_i \bar{x}_i) \tag{7-16}$$

and

$$P'r(Y/X) = \sum_{X} \frac{P'r(YX\bar{X})}{P'r(X)} = \sum_{\bar{X}\bar{Y}} \prod_{i=1}^{n} \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)}\;\; p(y_i/x_i \bar{x}_i)\, \bar{p}(\bar{y}_i/x_i \bar{x}_i). \tag{7-17}$$

Similarly, the average information flow in the forward direction through the channel of Figure 7-2 is

$$E\left\{I''(\bar{Y};X/\bar{X})\right\} = \sum_{X\bar{X}\bar{Y}} Pr(X\bar{X}\bar{Y}) \log \frac{P''r(\bar{Y}/X\bar{X})}{P''r(\bar{Y}/\bar{X})} \tag{7-18}$$

where

$$Pr(X\bar{X}\bar{Y}) = \sum_{Y} \prod_{i=1}^{n} \underset{x_i y_i}{q(x_i)}\;\; \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)}\;\; p(y_i/x_i \bar{x}_i)\, \bar{p}(\bar{y}_i/x_i \bar{x}_i) \tag{7-19}$$

where $x_1 = y_1 = \bar{x}_1 = \bar{y}_1 = 0$ by convention.

$$P''r\;(X\bar{X}\bar{Y}) = \sum_{Y} \prod_{i=1}^{n} \underset{x_i y_i}{q(x_i)}\;\; \bar{p}(\bar{x}_i)\; p(y_i/x_i \bar{x}_i)\; \bar{p}(\bar{y}_i/x_i \bar{x}_i) \tag{7-20}$$

$$P''r(\bar{Y}/X\bar{X}) = \prod_{i=1}^{n} \bar{p}(\bar{y}_i/x_i \bar{x}_i) \tag{7-21}$$

and

$$\Pr(\bar{Y}/\bar{X}) = \sum_{XY} \prod_{i=1}^{n} \underset{x_i y_i}{q(x_i)} \; p(y_i/x_i\bar{x}_i)\; \bar{p}(\bar{y}_i/x_i\bar{x}_i) \; . \qquad (7\text{-}22)$$

Having now fully defined the respective information flows through the two-way channel, we would like to prove

Theorem 7-2

The sum of information flows in the forward and reverse direction through the channel of Figure 7-2 is given by the information measure $I(X,Y;\bar{X},\bar{Y})$. Hence the following relationship holds:

$$I'(Y;\bar{X}/X)+I''(\bar{Y};X/\bar{X}) = I(Y;\bar{X}/X)+I(\bar{Y};X/\bar{X})+I(X;\bar{X}) \qquad (7\text{-}23)$$

Proof:

By elementary information measure properties we get

$$\begin{aligned} I(X,Y;\bar{X}\bar{Y}) &= I(Y;\bar{X}\bar{Y}/X) + I(X;\bar{X}\bar{Y}) \qquad (7\text{-}24) \\ &= I(Y;\bar{Y}/X\bar{X}) + I(Y;\bar{X}/X) + I(\bar{Y};X/\bar{X}) + I(X;\bar{X}) . \end{aligned}$$

But

$$I(Y;\bar{Y}/X\bar{X}) = \log \frac{\Pr(Y,\bar{Y}/X\bar{X})}{\Pr(Y/X\bar{X})\Pr(\bar{Y}/X\bar{X})} \qquad (7\text{-}25)$$

where

$$\Pr(Y;\bar{Y}/X\bar{X}) = \frac{\prod_{i=1}^{n} \underset{x_i y_i}{q(x_i)} \; \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)} \; p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i)}{\sum_{Y,\bar{Y}} \prod_{i=1}^{n} \underset{x_i y_i}{q(x_i)} \; \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)} \; p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i)} \qquad (7\text{-}26)$$

$$\Pr(Y/X\bar{X}) = \sum_{\bar{Y}} \Pr(Y,\bar{Y}/X\bar{X}) \qquad (7\text{-}27)$$

$$\Pr(\bar{Y}/X\bar{X}) = \sum_{Y} \Pr(Y,\bar{Y}/X\bar{X}) . \qquad (7\text{-}28)$$

Therefore we get

$$\frac{\Pr(Y,\bar{Y}/X\bar{X})}{\Pr(Y/X\bar{X})\Pr(\bar{Y}/X\bar{X})} = \frac{\sum\limits_{Y\bar{Y}} \prod\limits_{i=1}^{n} \underset{x_i y_i}{q(x_i)} \; \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)} \; p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i)}{\sum\limits_{\bar{Y}} \prod\limits_{i=1}^{n} \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)} \; \bar{p}(\bar{y}_i/x_i\bar{x}_i) \; \sum\limits_{Y} \prod\limits_{i=1}^{n} q(x_i) \; p(y_i/x_i\bar{x}_i)} = 1, \tag{7-29}$$

and taking (7-25) into consideration conclude that

$$I(X,Y:\bar{X}\bar{Y}) = I(Y;\bar{X}/X) + I(\bar{Y};X/\bar{X}) + I(X;\bar{X}) \; . \tag{7-30}$$

But also

$$I(XY;\bar{X}\bar{Y}) = \log \frac{\Pr(X\bar{X}Y\bar{Y})}{\Pr(X,Y)\Pr(\bar{X}\bar{Y})} \quad , \tag{7-31}$$

where

$$\Pr(XY\bar{X}\bar{Y}) = \prod_{i=1}^{n} \underset{x_i y_i}{q(x_i)} \; \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)} \; p(y_i/x_i\bar{x}_i) \; \bar{p}(\bar{y}_i/x_i\bar{x}_i) \tag{7-32}$$

$$\Pr(XY) = \sum_{\bar{X}\bar{Y}} \Pr(XY\bar{X}\bar{Y}) \tag{7-33}$$

$$\Pr(\bar{X}\bar{Y}) = \sum_{XY} \Pr(XY\bar{X}\bar{Y}) \; . \tag{7-34}$$

Hence we have

$$\frac{\Pr(XY\bar{X}\bar{Y})}{\Pr(XY)\Pr(\bar{X}\bar{Y})} = \frac{\prod\limits_{i=1}^{n} p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i)}{\sum\limits_{\bar{X}\bar{Y}} \prod\limits_{i=1}^{n} \underset{\bar{x}_i \bar{y}_i}{\bar{q}(\bar{x}_i)} \; p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i)} \cdot \frac{1}{\sum\limits_{XY} \prod\limits_{i=1}^{n} \underset{x_i y_i}{q(x_i)} \; p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i)} \; . \tag{7-35}$$

Taking into consideration equations (7-16), (7-17), (7-21), and (7-22) we see that

$$\frac{\Pr(XY\bar{X}\bar{Y})}{\Pr(XY)\Pr(\bar{X}\bar{Y})} = \frac{\Pr'(Y/X\bar{X})\Pr''(\bar{Y}/X\bar{X})}{\Pr'(Y/X)\ \Pr''(\bar{Y}/\bar{X})} \quad . \tag{7-36}$$

Thus from (7-12), (7-13), (7-31), and (7-36) we conclude that

$$I(XY;\bar{X}\bar{Y}) = I'(Y;\bar{X}/X) + I''(\bar{Y};X/\bar{X}) \tag{7-37}$$

Q.E.D.

It can be noted that equation (7-23) confirms that Figure 7-3 represents the reverse information flow situation correctly, since the mutual information $I(X;\bar{X})$ can have a non-zero value only if communication between the two channel terminals is established and causes the two signal sources to be correlated. Hence the quantity $I(Y;\bar{X}/X) + I(\bar{Y};X/\bar{X})$ does not account for the total information flow sum; $I(X;\bar{X})$ must be added to it.

As a consequence of Theorem 7-2 and of the averaging process of (7-13) and (7-18) we can also state

<u>Theorem 7-3</u>

The sum of the average information about signal sources through the two-way channel of Figure 7-2 in forward and reverse directions is

$$\begin{aligned} E\{I(X,Y:\bar{X}\bar{Y})\} &= E\{I'(X;\bar{Y}/\bar{X}) + I''(Y;\bar{X}/X)\} \\ &= E\{I(Y;\bar{X}/X)\} + E\{I(\bar{Y};X/\bar{X})\} + \\ &\quad + E\{I(X;\bar{X})\} . \end{aligned} \tag{7-38}$$

7.4 Information Loss Due to Coding

Consider now the derived channel of Figure 7-1 with both input sources generating successive symbols f and $\bar{f}$ independently at random with probabilities P(f) and $\bar{P}(\bar{f})$. The average mutual information

$$E\ [I(F,Y:\bar{F},\bar{Y})] = E[\log \frac{Pr(F\bar{F}Y\bar{Y})}{Pr(FY)Pr(\bar{F}\bar{Y})}] \qquad (7\text{-}39a)$$

between the symbol sequences of the opposing derived channel terminals can then be computed.

As the sources operate, the transducers of the derived channel generate input symbols x and $\bar{x}$ to the two-way channel which in turn cause the channel output signals y and $\bar{y}$. As pointed out at the beginning of section 7-2, the random signal relationships in the derived channel can be represented by the process of Figure 7-2, and for it the mutual information

$$E[I(XY;\bar{X}\bar{Y})] = E[\ \log\ \frac{Pr(X\bar{X}Y\bar{Y})}{Pr(XY)Pr(\bar{X}\bar{Y})}\ ] \qquad (7\text{-}39b)$$

between signal sequences at the opposing two-way channel terminals can be computed. The signal source porbabilities for Figure 7-2 are fully determined by relations (7-11) from the symbol source probabilities of the derived channel of Figure 7-1. We are interested in the relationship between the corresponding mutual informations $E[I(F,Y;\bar{F},\bar{Y})]$ and $E[I(X,Y;\bar{X},\bar{Y})]$, about which we can prove

Theorem 7-4

Given the derived channel of Figure 7-1 whose input symbol sources generate successive symbols f and $\bar{f}$ independently at random with probabilities P(f) and $\bar{P}(\bar{f})$. Represent the two-way input and output signal generation resulting from the derived channel communication by the process of Figure 7-2. If $E[I(F,Y;\bar{F},\bar{Y})]$ is the average mutual information between the symbol sequences at the opposing derived channel terminals, and $E[I(X,Y;\bar{X}\bar{Y})]$ is the average mutual information between the input and output signal sequences of the corresponding two-way channel of Figure 7-2, then the relationship

$$E[I(F,Y;\bar{F}\bar{Y})] = E[I(X,Y;\bar{X}\bar{Y})] \tag{7-40}$$

holds.

Proof:

First of all let it be understood that if $F = f_1, f_2, \ldots, f_n$ then

$$F(Y) = X \text{ if and only if } f_i(\mathcal{X}_i, \mathcal{Y}_i) = x_i \qquad i = (1,2,\ldots,n)$$

$$\text{where } \mathcal{X}_1 = \mathcal{Y}_1 = 0 \tag{7-41}$$

Then we have

$$Pr(Y,\bar{Y},F,\bar{F}) = \sum_{X,\bar{X}} Pr(Y,\bar{Y},X,\bar{X},F\bar{F}) \tag{7-42}$$

and it can be shown that

$$\Pr(Y,\bar{Y},X,\bar{X},F,\bar{F}) = \begin{cases} \Pr'(Y/X\bar{X})\Pr''(\bar{Y}/X\bar{X})P(F)\bar{P}(\bar{F}) & \\ \quad\quad \text{if } F(Y)=X,\ \bar{F}(\bar{Y})=\bar{X} & \\ 0 \quad\quad \text{otherwise.} \end{cases} \tag{7-43}$$

where

$$P(F) = \prod_{i=1}^{n} P(f_i) \qquad \bar{P}(\bar{F}) = \prod_{i=1}^{n} \bar{P}(\bar{f}_i) \tag{7-44}$$

Thus from (7-42) and (7-43) we can write

$$\Pr(Y,\bar{Y},F,\bar{F}) = \Pr'(Y/F(Y),\bar{F}(\bar{Y}))\Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y}))P(F)\bar{P}(\bar{F}) \tag{7-45}$$

and

$$\Pr(Y,F) = \sum_{\bar{Y},\bar{F}} \Pr(Y,\bar{Y},F,\bar{F}) =$$

$$= \sum_{\bar{X}} \Pr'(Y/F(Y)\bar{X}) \sum_{\bar{Y}} \Pr''(\bar{Y}/F(Y),\bar{X}) \sum_{\bar{F} \ni \bar{F}(\bar{Y})=\bar{X}} \bar{P}(\bar{F})\ P(F). \tag{7-46}$$

Hence we may write from (7-39)

$$I(YF;\bar{F}\bar{Y}) = \log \frac{\Pr'(Y/F(Y),\bar{F}(\bar{Y}))\ \Pr''(\bar{Y}/F(Y),\bar{F}(\bar{Y}))}{[\sum_{\bar{X}} \Pr'(Y/F(Y),\bar{X}) \sum_{\bar{Y}} \Pr''(\bar{Y}/F(Y),\bar{X}) \sum_{\bar{F} \ni \bar{F}(\bar{Y})=\bar{X}} \bar{P}(\bar{F})]} \cdot$$

$$\cdot \frac{1}{[\sum_{X} \Pr''(Y/X,\bar{F}(\bar{Y})) \sum_{Y} \Pr(\bar{Y}/X,\bar{F}(\bar{Y})) \sum_{F \ni F(Y)=X} P(F)]} \tag{7-47}$$

But notice that

$$\sum_{F \ni F(Y)=X} P(F) =$$

$$= \sum_{\substack{(f_n,\ldots,f_1) \ni f_n(\varkappa_n, y_n) = x_n \\ \vdots \\ f_1(\varkappa_1, y_1) = x_1}} \prod_{i=1}^{n} P(f_i) = \tag{7-48}$$

$$= \Big(\sum_{f_n \ni f_n(\mathscr{x}_n, \mathscr{y}_n) = x_n} p(f_n) \Big) \Big(\sum_{f_{n-1} \ni f_{n-1}(\mathscr{x}_{n-1}, \mathscr{y}_{n-1}) = x_{n-1}} P(f_{n-1}) \Big) \cdots$$

$$\cdots \Big(\sum_{f_1 \ni f_1(\mathscr{x}_1, \mathscr{y}_1) = x_1} P(f_1) \Big) .$$

Hence using the definition (7-11) we may write

$$\sum_{F \ni F(Y)=X} P(F) = \prod_{i=1}^{n} q_{\mathscr{x}_i \mathscr{y}_i}(x_i) \tag{7-49}$$

and

$$\sum_{\bar{F} \ni \bar{F}(\bar{Y})=\bar{X}} \bar{P}(\bar{F}) = \prod_{i=1}^{n} \bar{q}_{\bar{\mathscr{x}}_i \bar{\mathscr{y}}_i}(\bar{x}_i) \; . \tag{7-50}$$

It should be noted that for a given X and $\bar{X}$ the quantity (7-47) will have the same value for all F and $\bar{F}$ such that F(Y)=X and $\bar{F}(\bar{Y})=\bar{X}$. Taking into consideration this fact plus equations (7-45), (7-47), (7-49) and (7-50), we get

$$\sum_{Y\bar{Y}F\bar{F}} Pr(Y,\bar{Y},F,\bar{F}) \log \frac{Pr(Y,\bar{Y},F,\bar{F})}{Pr(Y,F)Pr(\bar{Y},\bar{F})} = \tag{7-51}$$

$$= \sum_{X\bar{X}} \sum_{Y,\bar{Y}} Pr'(Y/X\bar{X}) Pr''(\bar{Y}/X\bar{X}) \Big(\sum_{F \ni F(Y)=X} P(F) \Big) \Big(\sum_{\bar{F} \ni \bar{F}(\bar{Y})=\bar{X}} \bar{P}(\bar{F}) \Big) .$$

$$\cdot \log \frac{Pr'(Y/X\bar{X}) Pr''(\bar{Y}/X\bar{X})}{\Big[\sum_{\bar{X}'\bar{Y}'} Pr'(Y/X\bar{X}') Pr''(\bar{Y}'/X\bar{X}') \prod_{i=1}^{n} \bar{q}_{\bar{\mathscr{x}}_i \bar{\mathscr{y}}_i}(\bar{x}_i') \Big] \Big[\sum_{X'Y'} Pr'(Y'/X\bar{X}) Pr''(\bar{Y}/X'\bar{X}') \prod_{i=1}^{n} q_{\mathscr{x}_i \mathscr{y}_i}(x_i') \Big]} =$$

$$= \sum_{X\bar{X}Y\bar{Y}} \prod_{i=1}^{n} p(y_i/x_i\bar{x}_i)\bar{p}(\bar{y}_i/x_i\bar{x}_i)q_{x_i;y_i}(x_i)\ \bar{q}_{x_i;y_i}(\bar{x}_i) \ \cdot$$

$$\cdot \log \frac{\prod_{i=1}^{n} p(y_i/x_i\bar{x}_i)\ \bar{p}(\bar{y}_i/x_i\bar{x}_i)}{\sum_{\bar{X}'\bar{Y}'} \prod_{i=1}^{n} p(y_i/x_i\bar{x}_i')\bar{p}(\bar{y}_i'/x_i\bar{x}_i')\bar{q}_{\bar{x}_i;\bar{y}_i}(\bar{x}_i)\ ],}$$

$$\overline{\cdot\Big[\sum_{X'Y'} \prod_{i=1}^{n} p(y_i'/x_i'\bar{x}_i)\bar{p}(\bar{y}_i/x_i'\bar{x}_i)q_{x_i;y_i}(x_i')\ \Big]}$$

Comparing the right hand side of the previous equation with (7-35) we see from (7-31) that

$$E\{I(YF;\bar{Y}\bar{F})\} = E\{I(XY;\bar{X}\bar{Y})\} \qquad (7\text{-}52)$$

Q.E.D.

The significance of Theorem 7-4 is that it shows that the mutual information between the signals at the two terminals of the two-way channel is equal to the mutual information between the symbols at the two terminals of the derived channel (see Figure 7-1). It should be noted that the signal mutual information in Figure 7-1 is that obtainable from the set-up of Figure 7-3, and that anything obtainable from the latter is obtainable from the former by judicious selection of the probabilities $P(f)$ and $\bar{P}(\bar{f})$. Further importance of Theorem 7-4 will be seen from Theorem 7-5 which follows.

We will see that except for special, rather degenerate two-way channels, the information carried through the channel by the signals cannot be completely used as information about the messages being transmitted. Some of it is lost, and this loss is directly attributable to coding. In fact, any "strategy-like" encoding which does not associate a signal with a message in a one to one way will cause such information loss.

Theorem 7-5

Given transmission through a two-way derived channel in which the sources generate input symbols at opposite terminals independently from one another, the following average mutual information relations between symbols and signals hold:

$$E\{I(X,Y;\bar{X}\bar{Y})\} - (E\{I(Y;\bar{F}/F)\} + E\{I(\bar{Y};F/\bar{F})\}) = E\{I(Y;\bar{Y}/F,\bar{F})\} \geq 0 \quad (7\text{-}53)$$

Equality on the right hand side of (7-53) holds if and only if whenever $P(F) \neq 0$ and $\bar{P}(\bar{F}) \neq 0$ then simultaneously

(a) $Pr'(Y/F(Y),\bar{F}(\bar{Y})) = K_1$ for all $\bar{Y}$ such that $Pr''(\bar{Y}/F(Y),\bar{F}(\bar{Y})) \neq 0$ and

(b) $Pr''(\bar{Y}/F(Y),\bar{F}(\bar{Y})) = K_2$ for all Y such that $Pr'(Y/F(Y),\bar{F}(\bar{Y})) \neq 0$

where K_1 and K_2 are some constants.

Proof:

From Theorem 7-3 we know that (7-52) holds. From elementary information measure algebra we get

$$\begin{aligned} E\{I(X,Y:\bar{X}\bar{Y})\} &= E\{I(FY;\bar{F}\bar{Y})\} = \\ &= E\{I(Y;\bar{F}/F)\} + E\{I(\bar{Y};F/\bar{F})\} + E\{I(Y;\bar{Y}/F\bar{F})\} \\ &+ E\{I(F;\bar{F})\}. \end{aligned} \quad (7\text{-}54)$$

But F and $\bar{F}$ are assumed to be generated independently and therefore

$$E\left\{I(F,\bar{F})\right\} = 0 , \tag{7-55}$$

which proves the validity of the equation (7-53).

Next,

$$E\left\{I(Y;\bar{Y}/F\bar{F})\right\} = \sum_{Y,\bar{Y},F,\bar{F}} Pr(Y,\bar{Y},F,\bar{F}) \log \frac{Pr(Y/\bar{Y},F,\bar{F})}{Pr(Y/F\bar{F})} , \tag{7-56}$$

and therefore

$$-E\left\{I(Y;Y/FF)\right\} \leq \sum_{Y\bar{Y}F\bar{F}} Pr(Y,\bar{Y},F,\bar{F}) \left[\frac{Pr(Y/F\bar{F})}{Pr(Y/\bar{Y}F\bar{F})} - 1\right] \tag{7-57}$$

where equality holds only if

$$Pr(Y/F\bar{F}) = Pr(Y/\bar{Y}F\bar{F}) \quad \text{whenever } Pr(Y,\bar{Y},F,\bar{F}) \neq 0. \tag{7-58}$$

Before we start investigating conditions under which (7-58) can hold, we will break up (7-56). We have:

$$E\left\{I(Y;\bar{Y}/F\bar{F})\right\} = E\left\{\log \frac{Pr'(Y/F(Y),\bar{F}(\bar{Y}))}{\sum_{\bar{Y}'} Pr'(Y/F(Y),\bar{F}(\bar{Y}))Pr''(\bar{Y}'/F(Y)\bar{F}(\bar{Y}'))} + \right. \tag{7-59}$$

$$\left. + \log \frac{Pr''(\bar{Y}/F(Y)\ \bar{F}(\bar{Y}))}{\sum_{Y'} Pr'(Y'/F(Y')\bar{F}(\bar{Y}))Pr''(\bar{Y}/F(Y')\bar{F}(\bar{Y}))}\right\} ,$$

where, using reasoning employed inthe process of proving Theorem 7-4, we can write:

$$E\left\{\log \frac{Pr'(Y/F(Y)\bar{F}(\bar{Y}))}{\sum_{\bar{Y}'} Pr'(Y/F(Y)\bar{F}(\bar{Y}'))Pr''(\bar{Y}'/F(Y)\bar{F}(\bar{Y}'))}\right\} =$$

$$= \sum_{YX\bar{F}\bar{Y}} Pr'(Y/X\ \bar{F}(\bar{Y}))Pr''(\bar{Y}/X\ \bar{F}(\bar{Y}))\ \bar{P}(\bar{F})\Big(\sum_{F \ni F(Y)=X} P(F)\Big) \cdot \tag{7-60}$$

$$\cdot \log \frac{Pr'(Y/X,F(Y))}{\sum_{\bar{X}'} Pr'(Y/X,\bar{X}') \sum_{\bar{Y}' \ni \bar{F}(\bar{Y}')=\bar{X}'} Pr''(\bar{Y}'/X\bar{X}')}$$

and similarly,

$$E\left\{\log \frac{Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y}))}{\sum_{Y'} Pr'(Y'/F(Y')\bar{F}(\bar{Y}))Pr''(\bar{Y}/F(Y')\bar{F}(\bar{Y}))}\right\} =$$

$$= \sum_{\bar{Y}\bar{X}YF} Pr'(Y/F(Y),\bar{X})Pr''(\bar{Y}/F(Y)\bar{X}) \; P(F)\Big(\sum_{\bar{F} \ni \bar{F}(\bar{Y})=\bar{X}} \bar{P}(\bar{F})\Big) \cdot$$

$$\cdot \log \frac{Pr''(\bar{Y}/F(Y)\bar{X})}{\sum_{X'} Pr''(\bar{Y}/X'\bar{X}) \sum_{Y' \ni F(Y')=X'} Pr'(Y'/X'\bar{X})} \quad . \qquad (7\text{-}61)$$

In the next section we will interpret (7-60) and (7-61) "physically". Right now it is clear that if (7-59) is to be zero, both (7-60) and (7-61) must be. But (7-60) is zero if and only if

$$Pr'(Y/X,\bar{F}(\bar{Y})) = \sum_{\bar{X}'} Pr'(Y/X\bar{X}') \sum_{\bar{Y}' \ni \bar{F}(\bar{Y}')=\bar{X}'} Pr''(\bar{Y}'/X\bar{X}') \qquad (7\text{-}62)$$

whenever

$$Pr'(Y/X\bar{F}(\bar{Y}))Pr''(\bar{Y}/X\bar{F}(\bar{Y}))\bar{P}(\bar{F})\Big(\sum_{F \ni F(Y)=X} P(F)\Big) \neq 0 \; .$$

The right hand side of (7-62) is independent of $\bar{Y}$. Hence it is necessary that for all combinations Y,X such that $\sum_{F \ni F(Y)=X} P(F) \neq 0$ whenever $\bar{P}(\bar{F}) \neq 0$ we must have

$$Pr'(Y/X,\bar{F}(\bar{Y})) = \text{const.} \qquad (7\text{-}63)$$

for all $\bar{Y}$ such that $Pr'(Y/X,\bar{F}(\bar{Y}))Pr''(\bar{Y}/X,\bar{F}(\bar{Y})) \neq 0$.

However, as a matter of fact, even more is necessary, namely that

$$\text{For all } \bar{F},Y,X \quad \sum_{F \ni F(Y)=X} P(F) \neq 0 \quad \bar{P}(\bar{F}) \neq 0 \qquad (7\text{-}64)$$

$$Pr'(Y/X,\bar{F}(\bar{Y})) = \text{const.} \quad \text{for all } \bar{Y} \ni Pr''(\bar{Y}/X,\bar{F}(\bar{Y})) \neq 0.$$

Assume that the above is not true, i.e. that there exists a $\bar{Y}'$ such that

$$Pr'(Y/X,\bar{F}(\bar{Y}')) = 0 \quad \text{while } Pr''(\bar{Y}'/X\bar{F}(\bar{Y}')) \neq 0 \qquad (7\text{-}65)$$

and $Pr'(Y/X\bar{F}(\bar{Y})) = K \quad \forall\ \bar{Y} \neq \bar{Y}' \quad Pr''(\bar{Y}/X,\bar{F}(\bar{Y})) \neq 0.$

But then surely

$$\sum_{\bar{Y}} Pr'(Y/X\bar{F}(\bar{Y}))Pr''(\bar{Y}/X\bar{F}(\bar{Y})) \leq K \sum_{\bar{Y}\neq\bar{Y}'} Pr''(\bar{Y}/X\bar{F}(\bar{Y})) < K \qquad (7\text{-}66)$$

since

$$\sum_{\bar{Y}} Pr''(\bar{Y}/X\bar{F}(\bar{Y})) = \sum_{\bar{Y}} \prod_{j=1}^{n} \bar{p}(\bar{y}_j/x_j \bar{f}_j(\mathfrak{X}_j \mathfrak{Y}_j)) = 1 \qquad (7\text{-}67)$$

and it was assumed that $Pr''(\bar{Y}'/X\bar{F}(\bar{Y}')) > 0$,

and applying (7-66) to (7-62) we see that equality cannot hold since for $\bar{Y} \neq \bar{Y}'$, the left hand side is equal to K while the right hand side is smaller than K. This contradiction establishes the necessity of the condition (7-64). The latter is however also seen to be a sufficient condition to establish (7-62) since if (7-64) holds then

$$\sum_{\bar{Y}} Pr'(Y/X\bar{F}(\bar{Y}))Pr''(\bar{Y}/X\bar{F}(\bar{Y}))=K \sum_{\bar{Y}} Pr''(Y/XF(Y))=K \qquad (7\text{-}68)$$

$$\text{for all } \bar{F},Y,X \ni \sum_{F \ni F(Y)} P(F) \neq 0,\ \bar{P}(\bar{F}) \neq 0 .$$

If (7-64) is a necessary and sufficient condition for (7-60) to be zero, then the necessary and sufficient condition for

(7-61) to be zero must be

$$\text{For all } F,\bar{Y},\bar{X} \text{ such that } \sum_{\bar{F} \ni \bar{F}(\bar{Y})=\bar{X}} \bar{P}(\bar{F}) \neq 0 \text{ and } P(F) \neq 0 \tag{7-69}$$

$$Pr''(\bar{Y}/F(Y),\bar{X}) = \text{constant for all } Y \ni Pr'(Y/F(Y),\bar{X}) \neq 0$$

But (7-64) and (7-69) when formulated together give the condition of Theorem 7-5 which we are trying to prove.

Q.E.D.

From the preceding proof it is clear that the quantity $E\{I(Y;\bar{Y}/F,\bar{F})\}$ which is seen from (7-53) to be the difference between the total average information passing through the channel in both directions and the parts of it useful for message transmission, is closely associated with the kind of message encoding for the two-way channel which we are using. We will interpret and study it in the next section, in particular the question under what condition it can be made to equal zero.

7.5 Interpretation of Coding Loss

7.5.1 Forward and Reverse Coding Loss

We have shown in (7-59), (7-60) and (7-61) that the following break-up of the coding-loss function is possible:

$$E\left\{I(Y;\bar{Y}/F,\bar{F})\right\} = E\left\{\log \frac{Pr'(Y/X,\bar{F}(\bar{Y}))}{\sum_{\bar{Y}'} Pr'(Y/X,\bar{F}(\bar{Y}'))Pr''(\bar{Y}'/X,\bar{F}(\bar{Y}'))}\right\} +$$

$$+E\left\{\log \frac{Pr''(Y/F(Y),X)}{\sum_{Y'} Pr'(Y/F(Y')\bar{X})Pr''(\bar{Y}/F(Y')\bar{X})}\right\}, \quad (7\text{-}70)$$

where we can interpret the first term on the right hand side as the average loss of information in the reverse function-signal transducer of the derived channel. We can similarly interpret the second term on the right hand side of (7-70) as the average loss of information in the forward function-signal transducer of the derived channel. Focusing our attention on the first term, i.e. on the expression (7-60), we see that we are averaging the expression

$$\log \frac{Pr'(Y/X\bar{F}(\bar{Y}))}{\sum_{\bar{Y}'} Pr'(Y/X,\bar{F}(\bar{Y}'))P(\bar{Y}'/X,\bar{F}(\bar{Y}'))} \equiv I'(Y;\bar{Y}/\bar{F},X) \qquad (7\text{-}71)$$

which, from our experience gained in section 7.3 with respect to expressions (7-13) and (7-18) we can identify as the information provided by Y about $\bar{Y}$ when X and $\bar{F}$ are known and the signal X is fixed independently of the received signal Y.

Hence we are dealing with a mutual information pertaining to the situation represented in Figure 7-4.

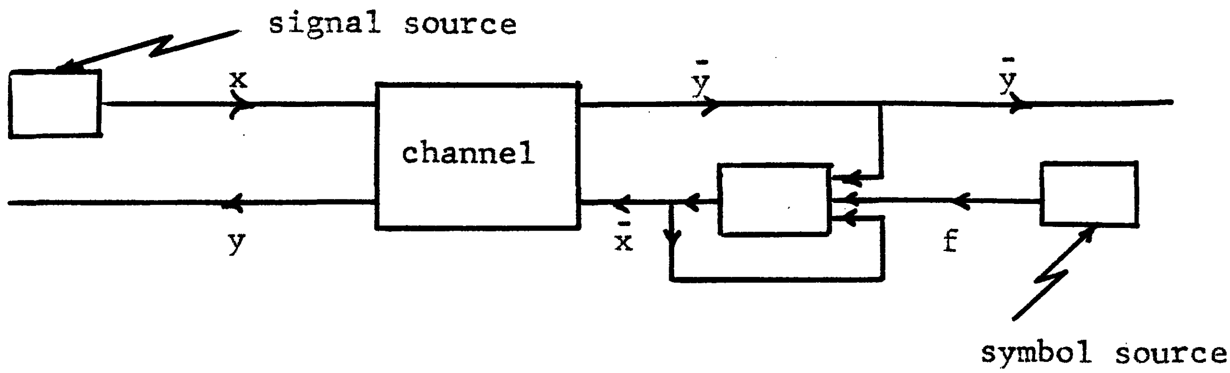


Figure 7-4.

In the expression (7-60) the information (7-71) is being averaged by use of probabilities

$$Pr(X,Y,\bar{F},\bar{Y}) = Pr'(Y/X,\bar{F}(\bar{Y}))Pr''(\bar{Y}/X,\bar{F}(\bar{Y}))\bar{P}(\bar{F}) \sum_{F \ni F(Y)=X} P(F) \quad , \tag{7-72}$$

giving the actual frequency of simultaneous occurrence of the events $X,Y,\bar{F},\bar{Y}$ in the ensemble. If we note that the decoder at the forward terminal is not interested in identification of the received signal $\bar{Y}$ for its own sake, it becomes clear why (7-60) is called the reverse coding information loss: it is the information about $\bar{Y}$ provided by Y after the message $\bar{F}$ was identified.*

In a similar manner, of course, the expression (7-73) can be identified with the information provided by $\bar{Y}$ about Y when F and $\bar{X}$ are known and $\bar{X}$ was fixed independently of the signal $\bar{Y}$:

$$\log \frac{Pr''(\bar{Y}/F(Y),\bar{X})}{\sum_{Y'} Pr'(Y'/F(Y')\bar{X})Pr''(\bar{Y}/F(Y')\bar{X})} \equiv I''(Y;\bar{Y}/F,\bar{X}) \qquad (7\text{-}73)$$

Hence the expectation of (7-73) arrived at by averaging with probabilities

$$Pr(\bar{X},\bar{Y},F,Y)=Pr'(Y/F(Y),\bar{X})Pr''(\bar{Y}/X,\bar{F}(\bar{Y}))P(F)\Big(\sum_{\bar{F}\to\bar{F}(\bar{Y})=\bar{X}} \bar{P}(\bar{F})\Big) \qquad (7\text{-}74)$$

can be called the forward coding information loss, and we can re-write (7-70) as follows

$$E\{I(Y;\bar{Y}/F,\bar{F})\} = E\{I'(Y;\bar{Y}/X,\bar{F})\} + E\{I''(Y;\bar{Y}/F,\bar{X})\} . \qquad (7\text{-}75)$$

The following theorem will prove useful for finding distributions $P(f)$ and $\bar{P}(\bar{f})$ giving good random codes (See Article 9):

Theorem 7-6

If $F(Y) = X$ and if $\bar{F}(\bar{Y}) = \bar{X}$, then

$$I'(Y;\bar{X}/X) - I(\bar{Y};F/\bar{F}) = I'(Y;\bar{Y}/X,\bar{F}) \qquad (7\text{-}76)$$

and

$$I''(\bar{Y};X/\bar{X}) - I(\bar{Y};F/\bar{F}) = I''(Y;\bar{Y}/F,\bar{X}) , \qquad (7\text{-}77)$$

where the quantities in the above equations are defined as in (7-3), (7-12), (7-13), (7-71) and (7-73).

Proof:

It will certainly be sufficient to prove (7-76); (7-77) will then follow.

We have:

$$I'(Y;\bar{X}/X) = \log \frac{Pr'(Y/X\bar{X})}{Pr'(Y/X)} \qquad (7\text{-}78)$$

$$I(Y;\bar{F}/F) = \log \frac{Pr(Y/F,\bar{F})}{Pr(Y/F)} \tag{7-79}$$

$$I'(Y;\bar{Y}/X,\bar{F}) = \log \frac{Pr'(Y/X,\bar{F},\bar{Y})}{Pr'(Y/X,\bar{F})} = \log \frac{Pr'(Y/X,\bar{F}(\bar{Y}))}{Pr(Y/X\bar{F})} \tag{7-80}$$

where $Pr'(Y/X,\bar{X})$, $Pr'(Y/X)$ and $Pr(Y/X,\bar{F})$ are shown in equations (7-16), (7-17) and (7-71), respectively. From there it is clear that if $\bar{F}(\bar{Y})=\bar{X}$, then $Pr'(Y/X\bar{X})=Pr'(Y/X,\bar{F}(\bar{Y}))$.

Also

$$Pr(Y/F\bar{F}) = \frac{\sum_{\bar{Y}} P(F)\bar{P}(\bar{F})Pr'(Y/F(Y)\bar{F}(\bar{Y}))Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y}))}{\sum_{Y,\bar{Y}} P(F)\bar{P}(\bar{F})Pr(Y/F(Y)\bar{F}(\bar{Y}))Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y}))} =$$

$$= \sum_{\bar{Y}} Pr'(Y/F(Y)\bar{F}(\bar{Y}))Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y})) \tag{7-81}$$

where the last equality follows from (7-67). Hence by inspection of (7-71) we see that $Pr(Y/F,\bar{F}) = Pr(Y/X,\bar{F})$, provided $F(Y)=X$. Finally:

$$Pr(Y/F) = \frac{\sum_{\bar{Y}\bar{F}} P(F)\bar{P}(\bar{F})Pr'(Y/F(Y)\bar{F}(\bar{Y}))Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y}))}{P(F)} \tag{7-82}$$

$$= \sum_{\bar{X}'\bar{Y}'} Pr'(Y/X\bar{X}')Pr''(\bar{Y}'/X\bar{X}') \sum_{\bar{F} \ni \bar{F}(\bar{Y}')=\bar{X}'} \bar{P}(\bar{F})$$

$$= \sum_{\bar{X}'\bar{Y}'} \prod_{i=1}^{n} p(y_i/x_i\bar{x}_i')\bar{p}(\bar{y}_i'/x_i\bar{x}_i')\, \bar{q}_{\bar{x}_i'\bar{y}_i}(\bar{x}_i')$$

Hence by inspection of (7-17) we see that if $F(Y)=X$ then $Pr'(Y/X) = Pr(Y/F)$. This completes the proof of (7-76)

Q.E.D.

7.5.2 Conditions for Absence of Coding Loss

It is interesting to inquire in what way the necessary and sufficient condition of Theorem 7-5 for the absence of coding loss can be met, and what coding restrictions this implies. For convenient reference we repeat here the losslessness condition:

The coding loss $E\left\{I(Y;\bar{Y}/F\bar{F})\right\}$ is identically zero if and only if in the given code whenever $P(F) \neq 0$ and $\bar{P}(\bar{F}) \neq 0$ then simultaneously

$$\Pr'(Y/F(Y)\bar{F}(\bar{Y})) = \text{const. for all } \bar{Y} \ni \Pr''(\bar{Y}/F(Y),\bar{F}(\bar{Y})) \neq 0$$

and (7-83)

$$\Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y})) = \text{const. for all } Y \ni \Pr'(Y/F(Y),\bar{F}(\bar{Y})) \neq 0.$$

It is of course immediately obvious that (7-83) will be satisfied if for all allowable Y (i.e. those for which $P(Y/F(Y)\bar{F}(\bar{Y})) \neq 0$ for all $\bar{F},\bar{Y}$), $F(Y) = X(F)$ for all F, and simultaneously for all allowable $\bar{Y}$, $\bar{F}(\bar{Y}) = \bar{X}(\bar{F})$ for all $\bar{F}$. But in such a case, the coding we have is one in which messages are associated with signals in a one to one manner, and we have the situation of "one-way channel coding". There are, however, special channels in which (7-83) is satisfied for any kind of coding. We can state the

Theorem 7-7

The condition (7-83) is satisfied for a derived two-way channel for any probability distribution over possible input symbol words F and $\bar{F}$ if and only if the two-way channel is characterized by transmission probabilities having the following property:

either (a) one cannot communicate through the two-way channel at all. i.e. the probability $p(y/x,\bar{x})$ is a function of x and y only, and the probability $\bar{p}(\bar{y}/x,\bar{x})$ is a function of $\bar{x}$ and $\bar{y}$ only.

or (b) for any signal combination x,x the channel transmission probabilities $p(y/x,\bar{x})$ and $\bar{p}(\bar{y}/x,\bar{x})$ are equal to either 0 or 1.

or (c) there exists one direction, say the forward one, such that $\bar{p}(\bar{y}/x,\bar{x})$ is a function of $\bar{y}$ and $\bar{x}$ only and also equals either 0 or 1.

Proof:

If part (a) is true, it is trivially clear that (7-83) is satisfied. If condition (b) holds then surely for every X and $\bar{F}$ there is only one $\bar{Y}$ such that $Pr''(\bar{Y}/X,\bar{F}(\bar{Y})) \neq 0$. Also for every $\bar{X}$ and F there is only one Y such that $Pr'(Y/F(Y),\bar{X}) \neq 0$. Hence for every F and $\bar{F}$ there exists only one combination $Y,\bar{Y}$ such that $Pr'(Y/F(Y)\bar{F}(\bar{Y})) \neq 0$ $Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y})) \neq 0$, insuring that (7-83) is satisfied. Finally, if condition (c) holds, then for any F, $\bar{F}$, $Pr'(Y/F(Y),\bar{F}(\bar{Y})) = \text{const.}$ for all Y, and there exists only one $\bar{Y}$ such that $Pr(\bar{Y}/F(Y),\bar{F}(\bar{Y})) \neq 0$.

Hence $\bar{F}$ maps into one $\bar{X}$ only and the first part of (7-83) is then satisfied as well.

We next wish to show that unless either of the conditions (a), (b), or (c) is satisfied, there will exist a combination F, $\bar{F}$ for which condition (7-83) will fail to hold. Now in case neither (a), (b) nor (c) hold, communication must be possible at least in one direction, and there will exist a combination $x,\bar{x}$ such that either $p(y/x,\bar{x})$ or $\bar{p}(\bar{y}/x,\bar{x})$ equal neither 0 nor 1.

Case I.: $p(y/x,\bar{x})$ is not a function of y and x only, there is a combination $x,\bar{x}$ such that $0 < p(y/x,\bar{x}) < 1$, and (c) is not satisfied in the forward direction.

(i) for all $x,\bar{x}$, $\bar{p}(\bar{y}/x,\bar{x})$ is either 1 or 0.

In this case, without loss of generality, let

$0 < p(y/0,0) < 1$ and let $\bar{p}(0/0,0) = 1$

$\bar{p}(0/1,0) = 0$

Then we can let $\bar{F}(\bar{Y}) = (0,0,\ldots,0)$ for all $\bar{Y}$

and $F(Y) = (0, f_2(y_1),0,\ldots,0)$ for all Y, where

$f_2(0) = 0 \quad f_2(1) = 1$

Let further $Y' = (0, \; y_2,\ldots,y_n)$

$Y'' = (1, \; y_2,\ldots,y_n)$

In this case clearly $Pr''(\bar{Y}/F(Y')\bar{F}(\bar{Y})) \neq$

$Pr''(\bar{Y}/F(Y'')\bar{F}(\bar{Y}))$

and (7-83) fails.

(ii) there exists $x,\bar{x}$ such that $0 < \bar{p}(y/x,\bar{x}) < 1$.

Without loss of generality let $0 < \bar{p}(\bar{y}/0,0) < 1$

There must also be an x such that $p(y/x,0) \neq p(y/x,1)$

Without loss of generality let this x=1. Then let

$F(Y) = (0,0,1,0,\ldots,0)$ for all Y. Also let

$\bar{F}(\bar{Y}) = (0,0,\bar{f}_3(\bar{y}_2),0,\ldots,0)$ for all $\bar{Y}$ where $\bar{f}_3(0) = 0$
$\bar{f}_3(1) = 1$

Clearly if $\bar{Y}' = (\bar{y}_1,0,\bar{y}_3,\ldots,\bar{y}_n)$
$\bar{Y}'' = (\bar{y}_1,1,\bar{y}_3,\ldots,\bar{y}_n)$

then $Pr'(Y/F(Y)\bar{F}(\bar{Y}')) \neq Pr(Y/F(Y)\bar{F}(\bar{Y}''))$ and (7-83) fails again. (Note both $\bar{Y}'$, $\bar{Y}''$ are allowable!!!)

Case II.: $p(y/x,\bar{x})$ is not a function of y and x only and there is a combination $x,\bar{x}$ such that $0 < \bar{p}(\bar{y}/x,\bar{x}) < 1$.

Without any loss of generality let $0 < p(y/x,0) < 1$ and let $p(y/1,0) \neq p(y/1,1)$.

Then let $F(Y) = (0,0,1,0,\ldots,0)$ for all Y

and let $\bar{F}(\bar{Y}) = (0,0,\bar{f}_3(\bar{y}_2),0,\ldots,0)$ for all $\bar{Y}$

where $\bar{f}_3(0) = 0$
$\bar{f}_3(1) = 1$.

Then surely both of the sequences below are allowable:

$\bar{Y}' = (\bar{y}_1,0,\bar{y}_3,\ldots,\bar{y}_n)$

$\bar{Y}'' = (\bar{y}_1,1,\bar{y}_3,\ldots,\bar{y}_n)$

and $Pr'(Y/F(Y)\bar{F}(\bar{Y}')) \neq Pr(Y/F(Y)\bar{F}(\bar{Y}''))$

Thus condition (7-83) fails even in this case.

Q.E.D.

We will now investigate whether condition (7-83) can be met in channels not fulfilling any of the conditions of Theorem 7-7. Specifically we are interested whether losslessness is possible when a code is used which could potentially get better communication results than a simple one-way channel type code, which maps messages directly into signals. Hence we are not interested in the trivial possibility mentioned in the paragraph preceding Theorem 7-7 where $F(Y) = X(F)$ and $\bar{F}(\bar{Y}) = \bar{X}(\bar{F})$, and neither are we interested in codes whose job could be done as well by a one-way type code.

We use the derived channel type of coding in order to create signal cooperation between the opposing terminals so as to reduce transmission noise in the channel. If the first part of condition (7-83) holds, then it is clear that for all message words F of a code all the possible reverse signals $\bar{F}(\bar{Y})$ for any given $\bar{F}$ are equivalent as far as the reverse transmission is concerned. Hence as far as the reverse noise goes, if the first part of (7-83) is to be fulfilled, there is no advantage gained by the fact that $\bar{F}(\bar{Y})$ may differ for different $\bar{Y}$. However, the forward noise can be improved in the above way, since it is conceivable that for a given $\bar{F}$, $\bar{F}(\bar{Y})$ assumes that value X which will maximize $Pr(Y/F(Y),\bar{F}(\bar{Y}))$. I.E. Given F(Y) let

$$Pr''(\bar{Y}^*/F(Y),\bar{X}^*) = \max_{\bar{Y},\bar{X}} (Pr''(\bar{Y}/F(Y)\bar{X})) \tag{7-84}$$

then it is conceivable that $\bar{F}(\bar{Y}^*) = \bar{X}^*$. Hence decoding discrimination at the reverse terminal can conceivably be made sharper by this device without in any way affecting the probable success of reverse transmission, and it is worthwhile to ask the question proposed on the top of this page.

It turns out that losslessness is possible only in more or less degenerate channel situations, and examples can easily be constructed of such lossless codes which could conceivably accomplish better communication results than one-way type codes,. Whether some of these actually do accomplish more is not known, but this is strongly doubted by the author, In any case, it was possible to prove

<u>Theorem 7-8</u>

Given a two-way channel whose transmission probabilities do not fulfill any of the conditions (a), (b), or (c) of Theorem 7-7, and are such that $0 < p(y/x,\bar{x}) < 1$ and $0 < \bar{p}(\bar{y}/x,\bar{x}) < 1$ for all signals $x,\bar{x},y,\bar{y}$. Then there can exist no code not equivalent to some one-way type code such that for it the condition (7-83) could be satisfied (i.e. there exists no code with zero coding loss).

Proof:

Consider a set of codewords $\{F\}$ and $\{\bar{F}\}$ of any length n. Let at least one of the codewords $\bar{F} = \bar{F}^*$ be such that there exist signal sequences $\bar{Y}^1$ and $\bar{Y}^2$ such that $\bar{F}^* (\bar{Y}^1) \neq \bar{F}^* (\bar{Y}^2)$.

It is clear that above is a necessary condition so that the code be non-equivalent to some one-way type code.

Since $0 < p(\bar{y}/x,\bar{x}) < 1$ for all $x, \bar{x}\bar{y}$ then all sequences $\bar{Y}$ are possible for any given pair F, $\bar{F}$. Consider now any particular of the codewords F and the word $\bar{F}*$. Under the channel transmission probability restrictions, if (7-83) is to be satisfied, then we must have

$$Pr'(Y/F(Y)\bar{F}*(\bar{Y}^1)) = Pr'(Y/F(Y)\bar{F}*(\bar{Y}^2)) \text{ for all } Y. \qquad (7\text{-}85)$$

Pick any sequence Y and let F(Y) = X'. Then, in conformity with our convention $\bar{\mathfrak{x}}_1 = \bar{\mathfrak{y}}_1 = 0$ we can write:

$$Pr'(Y/X'\bar{F}*(\bar{Y})) = p(y_1/x_1' \ \bar{f}_1^*(\bar{\mathfrak{x}}_1, \bar{\mathfrak{y}}_1))p(y_2/x_2' \ \bar{f}_2^*(\bar{\mathfrak{x}}_2, \bar{\mathfrak{y}}_2))\ldots$$

$$\ldots p(y_{n-1}/x_{n-1}'\bar{f}_{n-1}^*(\bar{\mathfrak{x}}_{n-1}, \bar{\mathfrak{y}}_{n-1}) \quad p(y_n/x_n' \ \bar{f}_n^*(\bar{\mathfrak{x}}_n, \bar{\mathfrak{y}}_n)) \qquad (7\text{-}87)$$

$$Pr''(\bar{Y}^1/X'\bar{F}*(\bar{Y}^1)) =$$

$$= \bar{p}(\bar{y}_1^1/x_1' \ \bar{f}_1^*(\mathfrak{x}_1^1, \mathfrak{y}_1^1))\ldots\bar{p}(\bar{y}_{n-1}^1/x_{n-1}' \ \bar{f}_{n-1}^*(\mathfrak{x}_{n-1}^1, \mathfrak{y}_{n-1}^1))$$

$$\bar{p}(\bar{y}_n^1/x_n' \ \bar{f}_n^*(\mathfrak{x}_n^1, \mathfrak{y}_n^1)) \qquad (7\text{-}88)$$

$$Pr''(\bar{Y}^2/X' \ \bar{F}*(\bar{Y}^2)) =$$

$$= \bar{p}(\bar{y}_1^{\,2}/x_1' \ \bar{f}_1^* \ (\bar{\mathfrak{x}}_1^2, \bar{\mathfrak{y}}_1^2))\ldots p(\bar{y}_{n-1}^2/x_{n-1}' \ \bar{f}_{n-1}^*(\bar{\mathfrak{x}}_{n-1}^2, \bar{\mathfrak{y}}_{n-1}^2))$$

$$p(\bar{y}_n^2/x_n' \ f_n^* \ (\bar{\mathfrak{x}}_n^2, \bar{\mathfrak{y}}_n^2)). \qquad (7\text{-}89)$$

Let $\bar{F}*(\bar{Y}^1) = \bar{X}^1$ and $\bar{F}*(\bar{Y}^2) = \bar{X}^2$. Then one possibility of how (7-83) could hold would be that whenever $\bar{x}_i^1 \neq \bar{x}_i^2$ then x_i' would be such that $p(y_i/x_i^1 \bar{x}_i^2) = p(y_i/x_i^1 \bar{x}_i^1)$. On the other hand, by assumption, the channel is such that at least for one X, say x=0, $p(\bar{y}/00) \neq p(\bar{y}/01)$. But then it is only possible that $x_i = 1$ and thus wherever $\bar{x}_i^1 \neq \bar{x}_i^2$ then if (7-83) were to hold for all F, it would do so by virtue of the fact that for all F and Y, $x_i = 1$. But in such a case the codeword $\bar{F}^*$ could just as well be replaced by a constant signalword $\bar{X}$, since any possible reverse signal adjustment in the i^{th} place makes no sense, x_i being in advance known to equal 1.

Therefore if the code is to be non-equivalent to a one-way type code there must be some F such that for at least some Y the signal X be such that when $\bar{x}_i^1 \neq \bar{x}_i^2$ then $x_i = 0$. Consider now that we have picked such a (F, Y)-combination for which in some places i, $x_i = 0$ and $\bar{x}_i^1 \neq \bar{x}_i^2$. Let these places be numbered $i_1, i_2, \ldots, i_k$ where

$$1 \leq i_1 < i_2 < \ldots < i_k \leq n \tag{7-90}$$

Then for the given F and all Y and consequent X=F(Y), if (7-83) is to hold, it must be true that

$$p(y_{i_1}/x_{i_1}\bar{x}^1_{i_1}) \ldots p(y_{i_k}/x_{i_k}\bar{x}^1_{i_k}) = p(y_{i_1}/x_{i_1}\bar{x}^2_{i_1}) \ldots p(y_{i_k}/x_{i_k}\bar{x}^2_{i_k}) \tag{7-91}$$

Consider now whether it be possible that $i_k = n$. In that case, for some Y at least, $p(y_n/x_n\bar{x}_n^1) \neq p(y_n/x_n\bar{x}_n^2)$. But in that case surely,

$$p(y_{i_1}/x_{i_1}\bar{x}^1_{i_1})\ldots p(y_{i_{k-1}}/x_{i_{k-1}}\bar{x}^1_{i_{k-1}}) \neq p(y_{i_1}/x_{i_1}\bar{x}^2_{i_1})\ldots$$

$$\ldots p(y_{i_{k-1}}/x_{i_{k-1}}\bar{x}^2_{i_{k-1}}) \qquad (7\text{-}92)$$

However, since all Y are possible by assumption, then y_n^c could actually occur and it should be noted that y_n cannot have any effect on the identity of the signal x_n. Hence (7-83) could hold only if

$$\begin{aligned} p(y_n/x_n\,\bar{x}^1_n) &= p(y^c_n/x_n\,\bar{x}^1_n) = 1/2 \\ p(y_n/x_n\,\bar{x}^2_n) &= p(y_n/x_n\,\bar{x}^2_n) = 1/2 \end{aligned} \qquad (7\text{-}93)$$

But in that case $p(y_n/x_n\,\bar{x}^1_n) = p(y_n/x_n\,\bar{x}^2_n)$ which is against our assumption. Hence we conclude that $i_k \neq n$.

Next we can ask whether possibly $i_k = n\text{-}1$, and it is clear that by an identical argument we will be led to conclude that since $i_k \neq n$, then also $i_k = n\text{-}1$. Continuing further we will finally end up with the conclusion that in the channel meeting the transmission probability restrictions of the theorem, the condition

$$Pr'(Y/F(Y)\bar{F}^*(\bar{Y})) = \text{const.} \quad \text{for all } \bar{Y} \qquad (7\text{-}94)$$

can be satisfied at least for some Y if and only if either $\bar{F}^*(\bar{Y}) = \bar{X}$ for all $\bar{Y}$, or if $\bar{F}^*(\bar{Y}^1) \neq \bar{F}^*(\bar{Y}^2)$ and whenever $\bar{x}^1_i \neq \bar{x}^2_i$ x_i is such that $p(y_i/x_i\,\bar{x}^1_i) = p(y_i/x_i\,\bar{x}^2_i)$. for all F.

Q.E.D.

7.6 The Outer Bound to the Capacity Region

In view of the results of the preceding section involving the concept of coding loss, it seems desirable to derive a new expression for the outer bound to the capacity region of the two-way channel, involving the expectations of mutual signal information. If we succeeded in showing that this new bound is completely included in the Shannon (1) outer bound G_0 (described in section 2.2), it would follow from Theorems 7-6, 7-7, and 7-8 that, except in degenerate situations, the capacity region G is entirely interior to Shannon's outer bound (this conclusion would be due to the introduction of coding loss by any cooperative code). Above can, in fact, be shown, and we will do so with the help of a slight modification of Shannon's argument proving the outer bound G_0 (see (1), pages 23-26).

A code for the two-way channel will either fall into the block class or into the sequential class (of course it might happen that e.g. the forward terminal encodes into a block code, while the reverse terminal encodes into a sequential code). In a block code messages m are mapped into functions F which map sequences of n received signals Y into sequences of length n of channel input signals, X. The identity of the i^{th} signal in a block is determined by the function F and by the identity of the signals $x_1, x_2, \ldots, x_{i-1}, y_1, y_2, \ldots, y_{i-1}$.

It is uninfluenced by any other signals, in particular by those which did not take place within the block to which x_i belongs. The most general kind of transmission is one consisting of the selection of only one message encoded into some function F mapping into an appropriately long blocklength n. We will therefore consider this kind of coding in what follows.

For this section and this section only we will adopt the following notation:

A function F will be understood to consist of a sequence of functions

$$F = (f_1, f_2, \ldots, f_n)$$

where f_i maps the signal sequence $x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}$ into the signal x_i. Hence, in our previous terminology, f_i will be assumed to have a dependency length equal to i-1. We will also denote by F_i^- and F_i^+ the sequences

$$F_i^- = (f_1, f_2, \ldots, f_i) \quad F_i^+ = (f_{i+1}, f_{i+2}, \ldots, f_n)$$

so that clearly

$$F = (F_i^-, F_i^+).$$

An identical notation will be adopted with respect to the remaining quantities of interest. Hence:

$$\bar{F} = (\bar{F}_i^-, \bar{F}_i^+) = [(\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_i), (\bar{f}_{i+1}, \bar{f}_{i+2}, \ldots, \bar{f}_n)]$$

where $\bar{f}_j$ maps the signal sequence $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{j-1}, \bar{y}_1, \bar{y}_2, \ldots \bar{y}_{j-1}$ into the signal $\bar{x}_j$. Also:

$$X = [X_i^-, X_i^+] = [(x_1, .., x_i)(x_{i+1}, .., x_n)]$$

$$Y = [Y_i^-, Y_i^+] = [(y_1, .., y_i)(y_{i+1}, .., y_n)]$$

$$\bar{X} = [\bar{X}_i^-, \bar{X}_i^+] = [(\bar{x}_1, .., \bar{x}_i)(\bar{x}_{i+1}, .., \bar{x}_n)]$$

$$\bar{Y} = [\bar{Y}_i^-, \bar{Y}_i^+] = [(\bar{y}_1, .., \bar{y}_i)(\bar{y}_{i+1}, .., \bar{y}_n)]$$

Therefore we have, for instance, the functional relationships:

$$F(Y) = X; \ F_i^-(Y_i^-) = X_i^-; \ F_i^+(X_i^-, Y_i^-, Y_i^+) = X_i^+.$$

We are now ready to state our theorem:

<u>Theorem 7-9</u>

Given a discrete memoryless two-way channel.

For any arbitrary positive integer i define the convex hull G_n^i $(n \geq i)$ of points

$$E\left\{\log \frac{Pr(\bar{y}_n/x_n\ \bar{x}_n)}{Pr''(\bar{y}_n/\bar{x}_n, \bar{x}_{n-1}, ..\bar{x}_{n-i+1}, \bar{y}_{n-1}, .., \bar{y}_{n-i+1})}\right\},$$

$$E\left\{\log \frac{Pr(y_n/x_n\ \bar{x}_n)}{Pr'(y_n/x_n, x_{n-1}, .., x_{n-i+1}, y_{n-1}, .., x_{n-i+1})}\right\} \tag{7-95}$$

obtained for different assignments $P(f_1, f_2, ..., f_n)$ $\bar{P}(\bar{f}_1, \bar{f}_2, .., \bar{f}_n)$ of symbol sequence of length n.

Let $\lim_{n \to \infty} G_n^i = G_0^i$. Then for any i, G_0^i is an outer bound on the capacity region of the two-way channel.

The following set-inclusion relationships are satisfied:

$$G_n^n \leq G_n^{n-1} \leq \ldots \leq G_n^1 \leq G_0 \qquad \forall n$$

$$G_i^i \leq G_{i+1}^i \leq \ldots \leq G_0^i \leq G_0 \qquad \forall i \qquad (7\text{-}96)$$

$$G \leq G_0^i \qquad \forall i$$

where G is the capacity region and G_0 is Shannon's outer bound to it.

Proof:

Suppose we have an arbitrary code for the two-way channel which at time zero is to transmit messages m and $\bar{m}$ from the two terminals. The selection of forward and reverse messages for transmission is to be an independent one. The messages will be associated with some codewords $F = F_i^-, F_i^+$ and $\bar{F} = \bar{F}_i^-, \bar{F}_i^+$ of length n, where n and i are arbitrary integers, $i \leq n$. After (i-1) operations of the channel, let Y_{i-1}^- and $\bar{Y}_{i-1}^-$ be the received signal blocks at the two terminals, and let x_i, $\bar{x}_i, y_i, \bar{y}_i$ be the next transmitted and received signals. Consider the change in "equivocation" of message at the two terminals due to the next received signal. At the reverse terminal, for example, this change is

$$\Delta = H(m/\bar{m},\bar{Y}_{i-1}^-) - H(m/\bar{m},\bar{Y}_{i-1}^-,y_i) =$$

$$= E\left\{\log \frac{\Pr(\bar{m},\bar{Y}_{i-1}^-)}{\Pr(m,\bar{m},\bar{Y}_{i-1}^-)}\right\} - E\left\{\log \frac{\Pr(\bar{m}\ \bar{Y}_{i-1}^-,\bar{y}_{i-1})}{\Pr(m,\bar{m},\bar{Y}_{i-1}^-,\bar{y}_{i-1})}\right\} \qquad (7\text{-}97)$$

(cont.)

$$= E\left\{\log \frac{Pr(\bar{y}_i/\bar{Y}^-_{i-1},m,\bar{m})}{Pr(\bar{y}_i/\bar{Y}^-_{i-1},\bar{m})}\right\}$$

$$= H(\bar{y}_i/\bar{Y}^-_{i-1},\bar{m}) - H(\bar{y}_i/\bar{Y}^-_{i-1},m,\bar{m}) .$$

But clearly

$$H(\bar{y}_i/\bar{Y}^-_{i-1},m,\bar{m}) \geq H(\bar{y}_i/\bar{Y}^-_{i-1},Y^-_{i-1},m,\bar{m}) = H(\bar{y}_i/x_i,\bar{x}_i) \tag{7-98}$$

since adding a conditioning variable cannot increase an entropy, and since

$$Pr(\bar{y}_i/\bar{Y}^-_{i-1},\bar{m},Y^-_{i-1},m) =$$

$$= Pr(y_i/F^-_i,Y^-_{i-1},\bar{F}^-_i,\bar{Y}^-_{i-1}) = Pr(\bar{y}_i/x_i,\bar{x}_i) \tag{7-99}$$

Also

$$H(\bar{y}_i/\bar{Y}^-_{i-1},\bar{m}) = H(\bar{y}_i/\bar{Y}^-_{i-1},\bar{F}^-_i) =$$

$$= H(\bar{y}_i/\bar{Y}^-_{i-1},\bar{X}^-_i) \tag{7-100}$$

where $\bar{F}^-_i(\bar{Y}^-_{i-1}) = \bar{X}^-_i$.

Hence from (7-97) we get

$$\Delta \leq H(\bar{y}_i/\bar{Y}^-_{i-1},\bar{X}^-_i) - H(\bar{y}_i/x_i,\bar{x}_i) =$$

$$= E\left\{\log \frac{Pr(\bar{y}_i/x_i\bar{x}_i)}{Pr''(\bar{y}_i/\bar{X}^-_i,\bar{Y}^-_{i-1})}\right\} \tag{7-101}$$

where the necessary averaging is done with respect to the probabilities of the different strategies F^-_i and $\bar{F}^-_i$ for the particular code used, and where the probability $P''(\bar{y}_i/\bar{x}_i,\bar{Y}^-_{i-1})$ is defined by:

$$Pr''(\bar{y}_i/\bar{X}_i^-,\bar{Y}_{i-1}^-) \equiv \frac{\sum_{Y_i,X_i} P'(Y_i^-,\bar{Y}_i^-/X_i^-,\bar{X}_i^-) \sum_{F_i^- \ni F_i^-(Y_i^-)=X_i^-} P(F_i^-)}{\sum_{Y_{i-1}^-,X_{i-1}^-} P'(Y_{i-1}^-,\bar{Y}_{i-1}^-/X_i\bar{X}_i^-) \sum_{F_{i-1}^- \ni F_{i-1}^-(Y_{i-1}^-)=X_{i-1}^-} P(F_i^-)} \quad (7\text{-}101a)$$

where $P(Y_i^-,\bar{Y}_i^-/X_i^-\ \bar{X}_i^-) = \prod_{j=1}^{i} p(y_j\bar{y}_j/x_j\bar{x}_j)$

A similar inequality may be obtained for the change of equivocation at the reverse terminal due to the reception, at time i, of the signal, y_i. It is:

$$\bar{\Delta} \leq E\left\{\log \frac{Pr(y_i/x_i\bar{x}_i)}{Pr'(y_i/X_i^-,Y_{i-1}^-)}\right\} \quad (7\text{-}102)$$

where the probability $Pr'(y_i/X_i^-,Y_{i-1}^-)$ is defined by:

$$Pr'(y_i/X_i^-,Y_{i-1}^-) \equiv \frac{\sum_{\bar{Y}_i^-,\bar{X}_i^-} P'(Y_i^-,\bar{Y}_i^-/X_i^-,\bar{X}_i^-) \sum_{\bar{F}_i^- \ni \bar{F}_i^-(\bar{Y}_i^-)=\bar{X}_i^-} \bar{P}(\bar{F}_i^-)}{\sum_{\bar{Y}_{i-1}^-\bar{X}_{i-1}^-} P(Y_{i-1}^-\bar{Y}_{i-1}^-/X_{i-1}^-,\bar{X}_{i-1}^-) \sum_{\bar{F}_{i-1}^-\ \bar{F}_{i-1}^-(\bar{Y}_{i-1}^-)=\bar{X}_i^-} \bar{P}(\bar{F}_{i-1}^-)} \quad (7\text{-}102a)$$

It therefore follows that in any code, the vector change in equivocation due to the reception of the i^{th} signal must be a vector with components bounded by

$$E\left\{\log \frac{Pr(\bar{y}_i/x_i\bar{x}_i)}{Pr''(\bar{y}_i/\bar{X}_i^-,\bar{Y}_{i-1}^-)}\right\}, \quad E\left\{\log \frac{Pr(y_i/x_i\bar{x}_i)}{Pr'(y_i/X_i^-,Y_{i-1}^-)}\right\} \quad (7\text{-}103)$$

for some distribution $P(F_i)$, $\bar{P}(\bar{F}_i^-)$, where we assumed that the messages m and $\bar{m}$ were selected independently for transmission.

Hence the vector change is included in the convex hull G_i^i of all vectors (7-103) generated by varying the distributions $P(F_i^-)$, $\bar{P}(\bar{F}_i^-)$. Since for any distribution $P(F_i)$, $P(F_i)$ the inequality

$$E\left\{\log \frac{1}{Pr''(\bar{y}_i/\bar{x}_i,\ldots,\bar{x}_2,\bar{y}_{i-1},\ldots\bar{y}_2)}\right\} \geq \tag{7-104}$$

$$\geq E\left\{\log \frac{1}{Pr''(\bar{y}_i/\bar{x}_i,\ldots,\bar{x}_1,\bar{y}_{i-1},\ldots\bar{y}_1)}\right\}$$

holds, it is clear that

$$G_i^j \subseteq G_i^{j-1} \tag{7-105}$$

where G_i^j, $(j \leq i)$, is the convex hull of points

$$E\left\{\log \frac{Pr(\bar{y}_i/x_i\bar{x}_i)}{Pr''(\bar{y}_i/\bar{x}_i,\ldots,\bar{x}_{i-j+1},\bar{y}_i,\ldots,\bar{y}_{i-j+1})}\right\},$$

$$E\left\{\log \frac{Pr(y_i/x_i\bar{x}_i)}{Pr'(y_i/x_i,\ldots,x_{i-j+1},y_{i-1},\ldots,y_{i-j+1})}\right\}. \tag{7-106}$$

obtained for various assignments $\bar{P}(\bar{F}_i^-)$, $P(F_i^-)$.

It is also true that

$$G_i^j\,(F_i^-,\bar{F}_i^-) \subseteq G_{i+1}^j(F_{i+1}^-,\bar{F}_{i+1}^-) \tag{7-107}$$

since we can let only those words F_{i+1}^- and $\bar{F}_{i+1}^-$ have non-zero probabilities for which

$$F^{-}_{i+1}(y_1,y_2,\ldots,y_i) = F^{-}_{i+1}(y_2,\ldots,y_i) = x_0,x_1,\ldots,x_{i+1}$$

$$\bar{F}^{-}_{i+1}(\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_i) = \bar{F}^{-}_{i+1}(\bar{y}_2,\ldots,\bar{y}_i) = \bar{x}_0,\bar{x}_1,\ldots,\bar{x}_{i+1}$$

where x_0 and $\bar{x}_0$ are fixed signals. (7-108)

The convex hull of points

$$E\left\{\log \frac{Pr(\bar{y}_{i+1}/x_{i+1},\bar{x}_{i+1})}{Pr''(\bar{y}_{i+1}/\bar{x}_{i+1},\bar{x}_{i-j+2},\bar{y}_i,\ldots,\bar{y}_{i-j+2})}\right\}, \quad (7\text{-}109)$$

$$E\left\{\log \frac{Pr(y_{i+1}/x_{i+1},\bar{x}_{i+1})}{Pr'(y_{i+1}/x_{i+1},\ldots,x_{i-j+2},y_i,\ldots,y_{i-j+2})}\right\}$$

obtained by varying the probability assignments of the words F^{-}_{i+1}, $\bar{F}^{-}_{i+1}$ allowed by (7-108) will then be identical with the convex hull of points (7-106).

Suppose that a signal block of length n was transmitted and received. Then the total change in equivocation resulting from reception must be a vector lying in the convex hull

$$G^n_n + G^{n-1}_{n-1} + \ldots + G^1_1 \leq$$

$$\leq G^n_n + G^{n-1}_n + \ldots + G^i_n + G^{i-1}_n + \ldots +$$

$$+ G^1_n \leq (n-i+1)\, G^i_n + (i-1)\, G^1_n =$$

$$= n\,[G^i_n + \frac{i-1}{n}\,[G^1_n - G^i_n]] \quad (7\text{-}110)$$

Thus given any positive integer i however large, and any $\epsilon > 0$, there exists an integer n^* such that for all signal blocks of length $n > n^*(i)$ the total change in equivocation resulting from reception must be a vector lying in the convex hull $n[G_n^i + \varepsilon] \subseteq n[G_0^i + \varepsilon]$. Consider now any code signalling at rates $R = \frac{1}{n'} \cdot \log M$ and $\bar{R} = \frac{1}{n'} \cdot \log \bar{M}$ where n' is the length of the signal block used in decoding, and M and $\bar{M}$ is the number of different forward and reverse messages encoded into signal blocks of length n'. If $n' < n^*$ (i) we may, without any detriment to the probability of error, decode blocks of length hn' where h is some positive integer such that $hn' > n^*$. In what follows we will quote Shannon (1), p. 25:

"The initial equivocations of message are nR and $n\bar{R}$ (where n=hn'). Suppose the point $(nR, n\bar{R})$ is outside the convex hull $n[G_0^i + \varepsilon]$ with nearest distance $n\delta$, Figure 7-5.

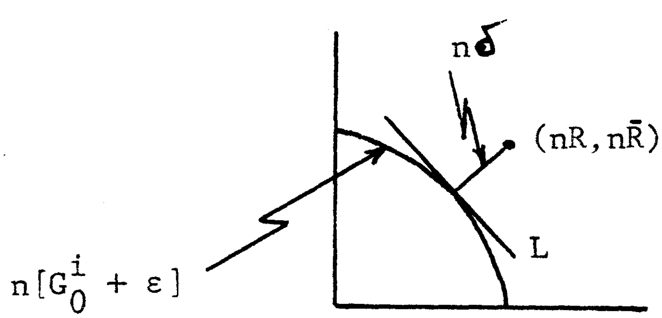


Figure 7-5.

Construct a line L passing through the nearest point of $n[G_0^i + \varepsilon]$ and perpendicular to the nearest approach segment with $n[G_0^i + \varepsilon]$ on one side (using the fact that $n[G_0^i + \varepsilon]$ is a convex region). It is clear that for any point $(nR^*, n\bar{R}^*)$ on the $n[G_0^i + \varepsilon]$ side of L and, particularly, for any point of

$n[G_0^i + \epsilon]$, that we have $|nR - nR^*| + |nR + nR^*| \geq n\delta$ (since the shortest distance is $n\delta$) and furthermore, at least one of $|nR - nR^*|$ and $|nR - nR^*|$ is at least $\frac{n\delta}{\sqrt{2}}$. (In a right triangle at least one leg is as great as the hypotenuse divided by $\sqrt{2}$.)"

"Thus after n uses of the channel, if the signalling rate pair R, $\bar{R}$ is distance δ outside the convex hull $(G_0^i + \varepsilon)$, at least one of the two final equivocations is at least $\frac{\delta}{\sqrt{2}}$. (All equivocations on a per use of the channel basis). Thus for signalling rates δ outside of $(G_0^i + \varepsilon)$ the equivocations per use of the channel are bounded from below independent of the code length n. This implies that the error probability is also bounded from below, that is, at least in one direction the error probability will be $\geq f(\delta) > 0$ independent of n," as shown in Fano (4) p. 186.

Since as $n \to \infty$, $\varepsilon \to 0$, this proves the theorem, except for the assertion in (7-96) that $G_n^1 \subseteq G_0$ (if this is proven, it follows from previous discussion that $G_0^i \subseteq G_0$)

Now G_n^1 is the convex hull of points

$$E\left\{\log \frac{Pr(\bar{y}_n/x_n\bar{x}_n)}{Pr''(\bar{y}_n/\bar{x}_n)}\right\}, \quad E\left\{\log \frac{Pr(y_n/x_n,\bar{x}_n)}{Pr'(y_n/x_n)}\right\} \tag{7-111}$$

obtained by varying the symbol assignments $P(F_n^-)$, $P(\bar{F}_n^-)$, while G_0 is the set of points (7-111) obtained by varying the assignments $P(x_n,\bar{x}_n)$. But, for instance,

$$E\left\{\log \frac{\Pr(\bar{y}_n/x_n\bar{x}_n)}{\Pr'(\bar{y}_n/\bar{x}_n)}\right\} = \sum_{x_n,\bar{x}_n,\bar{y}_n} \Pr(\bar{y}_n/x_n\bar{x}_n)\ \Pr'(x_n,\bar{x}_n)$$

$$\log \frac{\Pr(\bar{y}_n/x_n\bar{x}_n)}{\sum_{x_n} \Pr(\bar{y}_n/x_n\bar{x}_n)\Pr'(x_n\bar{x}_n)} \tag{7-112}$$

and it is obvious that any point (7-112) obtainable by a symbol assignment $\bar{P}(\bar{F}_n^-)$, $P(F_n^-)$ can also be obtained by some signal assignment $P(x_n,\bar{x}_n)$. This completes the proof of the theorem.

Q.E.D.

Whenever Shannon's outer bound G_0 is not equal to his inner bound G_I, it follows from Theorem 7-9, and Theorems 7-5 and 7-8 that the capacity region G will be, under certain conditions, strictly interior to G_0. We will state this in a more precise manner:

Theorem 7-10

Given a discrete, memoryless two-way channel defined by the set of probabilities $P(y,\bar{y}/x\bar{x}) = p(y/x\bar{x})p(\bar{y}/x\bar{x})$. Whenever Shannon's outer bound to the capacity region G_0 does not correspond to his inner bound G_I, the capacity region G will be a proper subset of G_0 provided the channel is not noiseless (i.e. it does not satisfy any of the conditions (a), (b), or (c) of Theorem 7-7) and $0 < p(y/x,\bar{x}) < 1$ and $0 < \bar{p}(\bar{y}/x,\bar{x}) < 1$ for all signals $x,\bar{x},y,\bar{y}$.

7.7 A New Inner Bound on Capacity Region

It is a consequence of Theorem 6-2 that the convex hull of points

$$E\left\{\log \frac{\Pr(\bar{y}/f,\bar{f},\mathscr{f}^{m},\bar{\mathscr{f}}^{m},\bar{\mathscr{y}}^{m},\bar{s}^{-m})}{\Pr(\bar{y}/\bar{f},\bar{\mathscr{f}}^{m},\bar{\mathscr{y}}^{m},\bar{s}^{-m})}\right\},$$

$$E\left\{\log \frac{\Pr(y/f,\bar{f},\mathscr{f}^{m},\bar{\mathscr{f}}^{m},\mathscr{y}^{m},s^{-m})}{\Pr(y/f,\mathscr{f}^{m},\mathscr{y}^{m},s^{-m})}\right\} \qquad (7\text{-}113)$$

where $\Pr(\mathscr{f}^{m}) = \prod_{i=1}^{m} P(f^{-i})$

$$\Pr(\bar{\mathscr{f}}^{m}) = \prod_{i=1}^{m} \bar{P}(\bar{f}^{-i})$$

is certainly included in the totality of points forming the capacity region G, and in turn includes the totality of points in Shannon's inner bound G_I to the capacity region. Hence (7-113) gives a tighter inner bound to the capacity region that G_I did.

From Theorem 7-6 it follows that the convex hull of points

$$E\left\{\frac{1}{n}\, I(\bar{Y};F/\bar{F})\right\},\ E\left\{\frac{1}{n}\, I(Y;\bar{F}/F)\right\} \qquad (7\text{-}114)$$

obtainable from all input symbol probability assignments $P(f)$ and $\bar{P}(\bar{f})$ is included in the convex hull of points

$$E\left\{\frac{1}{n}\, I(\bar{Y};X/\bar{X})\right\},\ E\left\{\frac{1}{n}\, I(Y;\bar{X}/X)\right\} \qquad (7\text{-}115)$$

obtainable from all input symbol probability assignments $P(f)$ and $\bar{P}(\bar{f})$, and thus following the discussion of section 7-2, from all the input signal probability assignments $q_{\boldsymbol{x}\boldsymbol{y}}(x)$, $\bar{q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{y}}}(\bar{x})$ (always dealing with an arbitrary but fixed dependency length ℓ). But as $n \to \infty$, the set (7-113) is included in the set (7-114), while, on the other hand, set (7-115) is, as $n \to \infty$, equal to the convex hull of points

$$\lim_{n\to\infty} E\left\{\log \frac{\bar{p}(\bar{y}_n/x_n,\bar{x}_n)}{Pr''(\bar{y}_n/\bar{x}_n,\bar{x}_{n-1},\ldots,\bar{x}_1,\bar{y}_{n-1},\ldots,\bar{y}_1)}\right\},$$

$$\lim_{n\to\infty} E\left\{\log \frac{p(y_n/x_n\ \bar{x}_n)}{Pr'(y_n/x_n,x_{n-1},\ldots,x_1,y_{n-1},\ldots,y_1)}\right\} \qquad (7\text{-}116)$$

(for proof see for instance Fano (4), p. 87) obtained from all the possible signal probability assignments $q_{\boldsymbol{x}\boldsymbol{y}}(x)$, $\bar{q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{y}}}(\bar{x})$. It follows from the discussion of the previous section that the set of points (7-116) is included in Shannon's outer bound G_0.

Given any set of probabilities $\{q_{\boldsymbol{x}\boldsymbol{y}}(x)\}$, $\{\bar{q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{y}}}(\bar{x})\}$ it can be shown (and will be, in Article 9) that there exists a set of probabilities $\{P(f)\}$, $\{\bar{P}(\bar{f})\}$ such that for them

$$q_{\boldsymbol{x}\boldsymbol{y}}(x) = \sum_{f \ni f(\boldsymbol{x},\boldsymbol{y})=x} P(f) \quad , \quad \bar{q}_{\bar{\boldsymbol{x}}\bar{\boldsymbol{y}}}(\bar{x}) = \sum_{\bar{f} \ni \bar{f}(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})=\bar{x}} \bar{P}(\bar{f}) \qquad (7\text{-}117)$$

Then to any point (7-116) obtained for some probabilities $\{q_{\boldsymbol{x}\boldsymbol{y}}(x)\}$, $\{q_{\bar{\boldsymbol{x}}\bar{\boldsymbol{y}}}(x)\}$ points (7-113) correspond, the latter points being obtained for those probability sets $\{P(f)\}$, $\{\bar{P}(\bar{f})\}$ satisfying (7-117). Since for a given ℓ the number of elements in the set $\{q_{\boldsymbol{x}\boldsymbol{y}}(x)\}$ is $2^{2\ell}$, and the number of elements in the set $\{P(f)\}$ is $2^{2^{2\ell}}$, above correspondence suggests a useful way to bound the boundary of the region of points (7-113) from below: the boundary of the region of points (7-116) can be found, and for each set $\{q_{xy}(x)\}$, $\{\bar{q}_{\bar{x}\bar{y}}(\bar{x})\}$ corresponding to a point on the boundary, a set $\{P(f)\}$, $\{\bar{P}(\bar{f})\}$ is found satisfying (7-117). Using the latter sets, corresponding points (7-113) are found and it is known that these are included inside the boundary of the region containing all the points (7-117). The advantage of this bounding will be shown more fully in Article 9 where related questions of practical coding importance will be discussed.

8. Classification of Two-Way Channels and Development of Canonical Channels.

The results of previous sections have provided us with intuitive insight into the workings of binary two-way channels. In order to further our understanding as to which kinds of codes are effective and why, we wish to analyze and classify the different possible two-way binary channels meeting the restriction

$$p(y,\bar{y}/x\bar{x}) = p(y/x\bar{x})\ p(\bar{y}/x\bar{x}) \qquad (8\text{-}1)$$

In Article 9 we will then carry out a synthesis of the results of Articles 6, 7 and 8 and display some important properties of two-way channels.

The restriction (8-1) insures that the success of the signal transmission in the forward direction will depend on the identity of the signal sent in the reverse direction and not on the consequent signal received at the forward terminal. A similar statement can be made about transmission in the reverse direction. The noise of a two-way channel postulated in this way will then consist of two components:

(a) natural directional channel noise, analogous to the one we are used to in one-way channels

(b) noise induced by the signal transmitted in the opposite direction.

The noise components are superimposed upon one another.

It would be interesting to separate the two components and see more clearly the essential difference between operations of one-way and two-way channels. We hope that such separation would show us a way to reduce the (b) component of the noise by signal cooperation of the opposing terminals. In what follows we will keep ourselves restricted to the binary case.

8.1 Terminal Labeling Rule for Channels Representable by Figure 3-1.

For the general one-way channel representable as in Figure (8-1), the usual practice is to identify the output

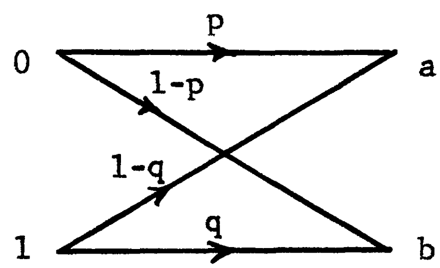


Figure 8-1

alphabet (a,b) with the input alphabet (0,1). A decision must then be made whether a or b is to be made to correspond to 0. Usually

$$\begin{array}{lll} \text{if} & p > 1 - q & \text{we let } a = 0 \text{ and } b = 1 \\ \text{if} & p < 1 - q & \text{we let } a = 1 \text{ and } b = 0 \end{array} \qquad (8\text{-}2)$$

An identification problem similar to the above one exists for the two-way channel, broken up into forward and reverse portions as in Figure 3-1. Since the probability functions p( ), q( ) are defined independently of the functions $\bar{p}$( ), $\bar{q}$( ), we will deal with the two portions separately.

Each portion is characterized by a transition probability set having four elements, e.g. the forward portion is characterized by the set $\{p(0), p(1), q(0), q(1)\}$. Without any loss of generality, the sets of both portions may be required to fulfill the following conditions:

(a) At least one of the following relations must be true:

$$p(1) \neq p(0) \qquad \bar{p}(1) \neq \bar{p}(0)$$

$$q(1) \neq q(0) \qquad \bar{q}(1) \neq \bar{q}(0) \tag{8-3}$$

otherwise no interaction noise is present and the two-way channel breaks up into two separate one-way channels.

(b) At least two members of a probability set have values greater than or equal to 1/2, and in addition, the sum of the values of the four members of a set must be greater than or equal to 2.

(i.e. $p(0) + p(1) + q(0) + q(1) \geq 2$

$p(0) + p(1) + q(0) + q(1) \geq 2$) (8-4)

It should be noted that if necessary the above can always be arranged by switching the "0" and "1" labels at the channel output.

8.2 Classes of Channel Portions

There exist two different classes of channel portions. All channel portions either belong directly to these classes or are equivalent to portions that belong to them. We have:

Class A:

Channels in which additional noise is caused by simultaneous transmission of signals having different digit values.

A channel portion belongs to this class if simultaneously:

$$1 \geq p(0) \geq p(1) \geq 0$$
$$1 \geq q(1) \geq q(0) \geq 0 \qquad (8\text{-}5)$$

Class B:

Channels in which one of the input symbols of the reverse portion causes exceptionally noisy conditions for forward transmission.

A channel portion belongs to this class if simultaneously

$$1 \geq p(0) \geq p(1) \geq 0$$
$$1 \geq q(0) > q(1) \geq 0 \qquad (8\text{-}6)$$

(Under the assumption that $x = 1$ is the noise-causing digit)

Equivalent Class A:

$$1 \geq p(1) > p(0) \geq 0$$
$$1 \geq q(0) > q(1) \geq 0 \qquad (8\text{-}5a)$$

Equivalent Class B:

$$1 \geq p(1) > p(0) \geq 0$$

$$1 \geq q(1) \geq q(0) \geq 0 \qquad (8\text{-}6a)$$

8.3 Equivalent Channel Portion Representations A and B

In order to separate the noise components (a) from (b) we would like to display a channel portion as consisting of a part in which only noise component (a) would be present, and a part having only noise component (b). Such a break-up is evident in Figure 8-2, where the newly introduced probabilities are defined as follows:

$$\alpha(\bar{x}) = \frac{p(\bar{x}) + p_1 - 1}{p_0 + p_1 - 1}$$

$$\beta(\bar{x}) = \frac{q(\bar{x}) + p_0 - 1}{p_1 + p_0 - 1} \qquad (8\text{-}7)$$

Representation by Figure 8-2 is possible only if probabilities p_1 and p_0 can be picked so that simultaneously:

$$0 \leq p_0 \leq 1$$

$$0 \leq p_1 \leq 1$$

$$0 \leq \alpha(\bar{x}) \leq 1 \qquad \bar{x} = 0,1 \qquad (8\text{-}8)$$

$$0 \leq \beta(\bar{x}) \leq 1$$

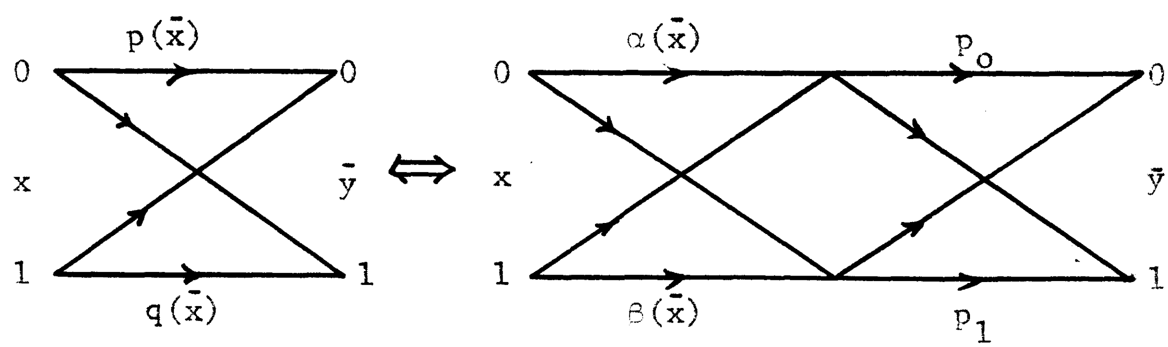


Figure 8-2.

The noise in the second part of Figure 8-2 consists of component (b) only. The noise in the first part will consist only of component (a), provided that $\alpha(\bar{x}) = 1$ for some $\bar{x}$ and $\beta(\bar{\bar{x}}) = 1$ for some $\bar{\bar{x}}$. We shall now see whether this condition together with the requirement (8-8) can be satisfied, and in what case.

For Class A we expect that we will be able to make

$$\alpha(0) = \beta(1) = 1 \tag{8-9}$$

For Class B we expect that we will be able to make

$$\alpha(0) = \beta(0) = 1 \tag{8-10}$$

Equation (8-9) leads to the equivalent channel representation of Figure 8-3. Relations (8-7) then become:

$$\begin{aligned} p_0 &= p(0) \\ p_1 &= q(1) \\ \alpha(1) &= \frac{p(1) + q(1) - 1}{p(0) + q(1) - 1} = \frac{b}{a} \\ \beta(0) &= \frac{p(0) + q(0) - 1}{p(0) + q(1) - 1} = \frac{c}{a} \end{aligned} \tag{8-11}$$

If we limit ourselves to Class A, then from (8-5) it follows that whenever $a < 0$ then also $b < 0$ and $c < 0$. But labeling rule (8-4) requires that $b + c < 0$. Hence we must conclude that in Class A always $a > 0$. Therefore (8-8) can be satisfied if and only if

(a) $p(1) + q(1) \geq 1$

(b) $p(1) \leq p(0)$

(c) $p(0) + q(0) \geq 1$ (8-12)

(d) $q(0) \leq q(1)$

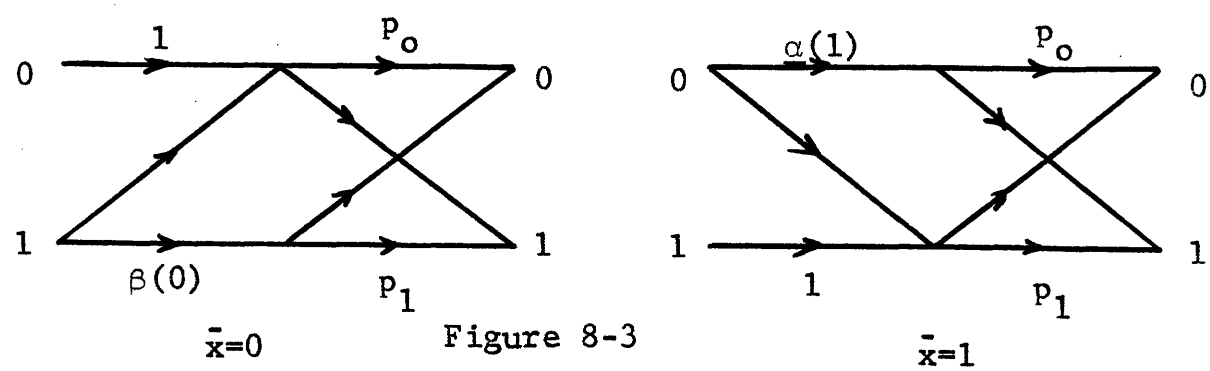


Figure 8-3

Equivalent Channel Portion Representation A

It is furthermore clear that whenever a portion belongs to Class A and labeling rules (8-4) are satisfied, then (8-12b) and (8-12d) will always be fulfilled, and at most one of (8-12a) and (8-12c) will be violated. We will deal with this possibility in the next article.

Consider next relations (8-10) which lead to the Equivalent Channel Portion Representation B of Figure (8-4)

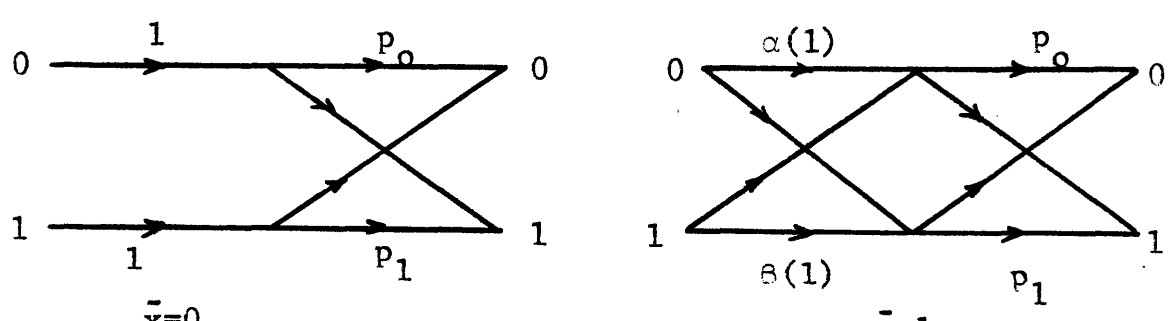


Figure 8-4.

Equivalent Channel Portion Representation B

Relations (8-7) then become:

$$
\begin{aligned}
p_0 &= p(0) \\
p_1 &= q(0) \\
\alpha(1) &= \frac{p(1) + q(0) - 1}{p(0) + q(0) - 1} = \frac{b}{a} \\
\beta(1) &= \frac{p(0) + q(1) - 1}{p(0) + q(0) - 1} = \frac{c}{a}
\end{aligned}
\qquad (8\text{-}13)
$$

Again, if we limit ourselves to Class B, then from (8-6) it will follow that whenever $a < 0$ then also $b < 0$ and $c < 0$. Labeling rule (8-4) requires that $b + c > 0$. Hence we must conclude that in Class B always $a > 0$. Therefore for Class B (8-8) can be satisfied only if

$$
\begin{aligned}
&(a) \quad p(1) + q(0) \geq 1 \\
&(b) \quad p(1) \leq p(0) \\
&(c) \quad q(1) + p(0) \geq 1 \\
&(d) \quad q(1) \leq q(0)
\end{aligned}
\qquad (8\text{-}14)
$$

But whenever labeling rules (8-4) and relations (8-6) are satisfied, then (8-14b) and (8-14d) are always fulfilled, and at most one of (8-14a) and (8-14c) is violated. We will deal also with this possibility in the next article.

8.4 Channel Equivalence Through Transformation

In section 8.1 we have decided upon a labeling rule. The received symbols a and b were identified with the transmitted symbols 0 and 1. If, for the sake of argument, we are dealing with forward transmission, then whenever rules (8-4) are satisfied, a is identified with 0 and b with 1 regardless of the identity of the signal $\bar{x}$ transmitted in the reverse direction. There is, however, no reason why the identification of digits 0, 1 with the constant a should be independent of the digit $\bar{x}$ being transmitted. We could, for instance, identify a with 0 whenever $\bar{x}=0$ and a with 1 whenever $\bar{x}=1$. In such a case we would get instead of the usual channel portion the portion of Figure 8-5, where we assume that rule (8-4) has been kept for the new probabilities $r(\bar{x})$ and $s(\bar{x})$.

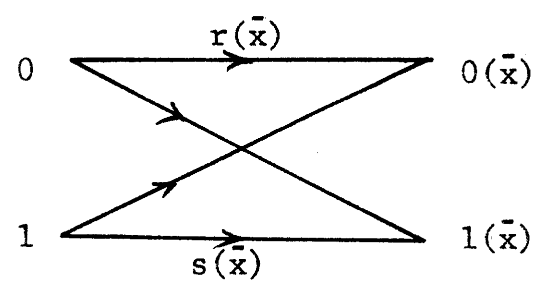


Figure 8-5

From rule (8-4) it follows that if

$$p(0) + q(0) > p(1) + q(1) \tag{8-15}$$

then $0(\bar{x} = 0) = a$ $\quad 0(\bar{x} = 1) = b$

$1(\bar{x} = 0) = b$ $\quad 1(\bar{x} = 0) = a$

and $r(0) = p(0)$ $\quad r(1) = 1 - p(1)$

$s(0) = q(0)$ $\quad s(1) = 1 - q(1)$

and if

$$p(0) + q(0) < p(1) + q(1) \tag{8-16}$$

then

$0(0) = b$, $\quad 0(1) = a$

$1(0) = b$, $\quad 1(1) = a$

and $r(0) = 1-p(0)$; $\quad r(1) = p(1)$

$s(0) = 1-q(0)$ $\quad s(1) = q(1)$

Of course the channel portion of Figure 8-5 may be transformed into equivalent channel portions of Figure 8-3 or Figure 8-4, whichever is applicable. Hence suppose that we are dealing with a channel portion of Class A for which condition (8-12a) fails. We can transform such a portion into the channel of Figure 8-5, where relations (8-15) are applicable. We may now attempt to fit the new channel into the schemes of either Equivalent Channel Portion Representation A or Equivalent Channel Portion Representation B. This, of course, depends on whether the new probabilities $r(\bar{x})$ and $s(\bar{x})$ satisfy conditions (8-5) or (8-6).

(It should be noted that the new channel could also satisfy equivalent conditions (8-5a) or (8-6a) arising from (8-5) and (8-6) by substitution of the argument "1" by the argument "0" and of the argument "0" by the argument "1").

Now if conditions (8-5) are satisfied, then we have

$$1 \geq p(0) > 1 - p(1) \geq 0$$
$$1 \geq 1 - q(1) > q(0) \geq 0 \qquad (8\text{-}17)$$

and the question is whether (8-18) is satisfied:

$$p(1) + q(1) \leq 1$$
$$p(0) + q(0) \leq 1 \qquad (8\text{-}18)$$

But the satisfaction of (8-18) was the reason why we transformed into the equivalent portion of Figure 8-5 in the first place.

Next, if conditions (8-6) are satisfied for the new probabilities, then we have

$$1 \geq p(0) > 1 - p(1) \geq 0$$
$$1 \geq q(0) > 1 - q(1) \geq 0 \qquad (8\text{-}19)$$

and the question is whether (8-20) is satisfied:

$$q(0) \geq p(1)$$
$$p(0) \geq q(1) \qquad (8\text{-}20)$$

But, as pointed out, inequalities (8-18) hold by assumption and (8-20) follow from (8-18) and (8-19).

Finally, the new probabilities might satisfy Equivalent Class A definition. Then we would have:

$$1 \geq 1 - p(1) > p(0) \geq 0$$
$$1 \geq q(0) > 1 - q(1) \geq 0 \qquad (8\text{-}21)$$

and the question is again whether (8-18) is satisfied, and, as has been pointed out, it is.

(Equivalent Class B cannot hold for the new probabilities because, as can be checked, this would mean a violation of the labeling rule for the original probabilities). We have thus successfully finished treating Class A when (8-12a) does not hold.

All together, there are the following possible troublesome cases, which can be treated by the transformation of Figure 8-5, as can be shown by reasoning similar to the one employed above:

(1) Channel portion of Class A where inequality (8-12a) fails to hold (was treated above).

(2) Channel Portion of Class A where inequality (8-12c) is not satisfied. In this case the probabilities $r(\bar{x})$ and $s(\bar{x})$ can fulfill the requirements of: (a) Class A, (b) Equivalent Class A, (c) Equivalent Class B.

(3) Channel portion of Class B where inequality (8-14a) is not satisfied. The transformed channel portion then satisfies inequalities (8-15) and the probabilities $r(\bar{x})$ and $s(\bar{x})$ can fulfill the requirements of: (a) Class A, (b) Class B, (c) Equivalent Class A.

(4) Channel portion of Class B where inequality (8-14c) is not satisfied. The transformed channel then satisfies inequalities (8-16) and the probabilities $r(\bar{x})$ and $s(\bar{x})$ can fulfill the requirements of: (a) Class A, (b) Equivalent Class A, (c) Equivalent Class B.

From the discussion in this and the preceding article we can therefore conclude that every portion of a binary two-way channel whose transition probabilities satisfy equation (8-2) can be treated either directly, or after preliminary transformation into the channel portion of Figure 8-5, as either the Equivalent Channel Portion Representation A, or as the Equivalent Channel Portion Representation B. In other words, it can always be broken up into a cascade of two components, the first of which has noise due to the transmission signal in the opposite direction, and the second of which has steady noise uninfluenced by the opposite signal.

8.5 Symmetrical Equivalent Channels

As pointed out, there are only four different possible channel portions: Class A, Class B, Equivalent Class A, and Equivalent Class B. Every channel consists of a forward and a reverse portion, and consequently there are 10 essentially different two-way channel classes (that is, if we consider a case and its equivalent to be different). However, it is reasonable to assume that in most cases both portions of a two-way channel will belong to the same class.

Further interesting symmetry conditions can, however, be obtained. We will term a channel "horizontally symmetrical" if for both directions the transmission conditions of a $\underline{0}$ and a 1 are the same. We will term a channel "vertically symmetrical" if the transmission conditions for the forward direction are identical with those for the reverse directions. A "completely symmetrical" channel is one which possesses both horizontal and vertical symmetries.

Horizontal symmetry results in very interesting situations for both Class A and B.

Horizontal Class A

Clearly the conditions

$$p(0) = q(1) \qquad \bar{p}(0) = \bar{q}(1)$$
$$p(1) = q(0) \qquad \bar{p}(1) = \bar{q}(0) \tag{8-22}$$

must be satisfied.

Inspecting relations (8-11) we see that

$$p_0 = p_1 = p = p(0)$$
$$\bar{p}_0 = \bar{p}_1 = \bar{p} = \bar{p}(0) \tag{8-23}$$
$$\alpha(1) = \beta(0) = \frac{p(1) + p(0) - 1}{2p(0) - 1}$$
$$\bar{\alpha}(1) = \bar{\beta}(1) = \frac{\bar{p}(1) + \bar{p}(0) - 1}{2\bar{p}(0) - 1}$$

and therefore we obtain the channel of Figure 8-6.

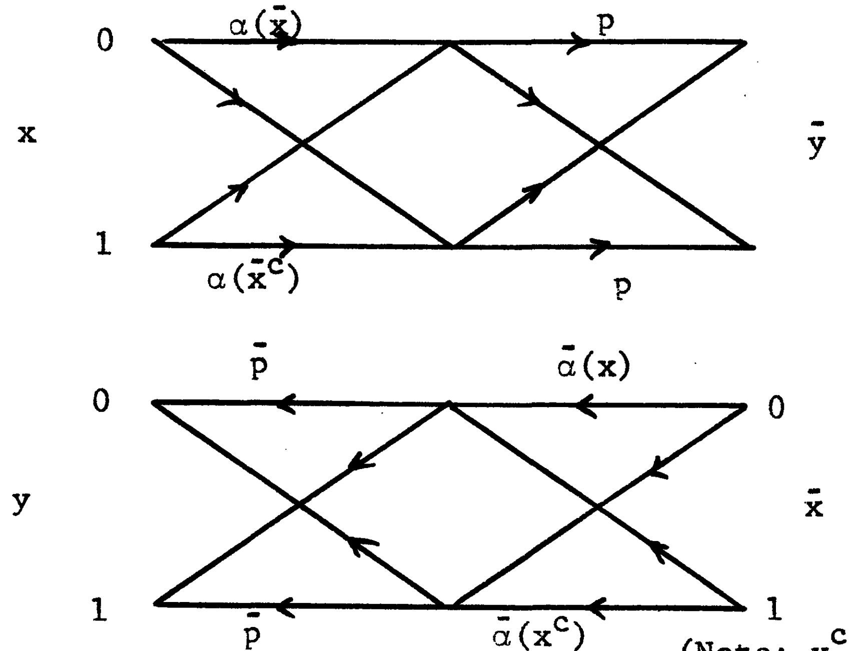


Figure 8-6.

(Note: x^c is the complement of x
$\bar{x}^c$ is the complement of $\bar{x}$)

Thus, we have for both directions a reversible z-channel followed by a constant binary symmetric channel. Further reflection shows that the channel of Figure 8-6 can be represented by the configuration of Figure 8-7, the latter representation being conducive to further interesting interpretations.

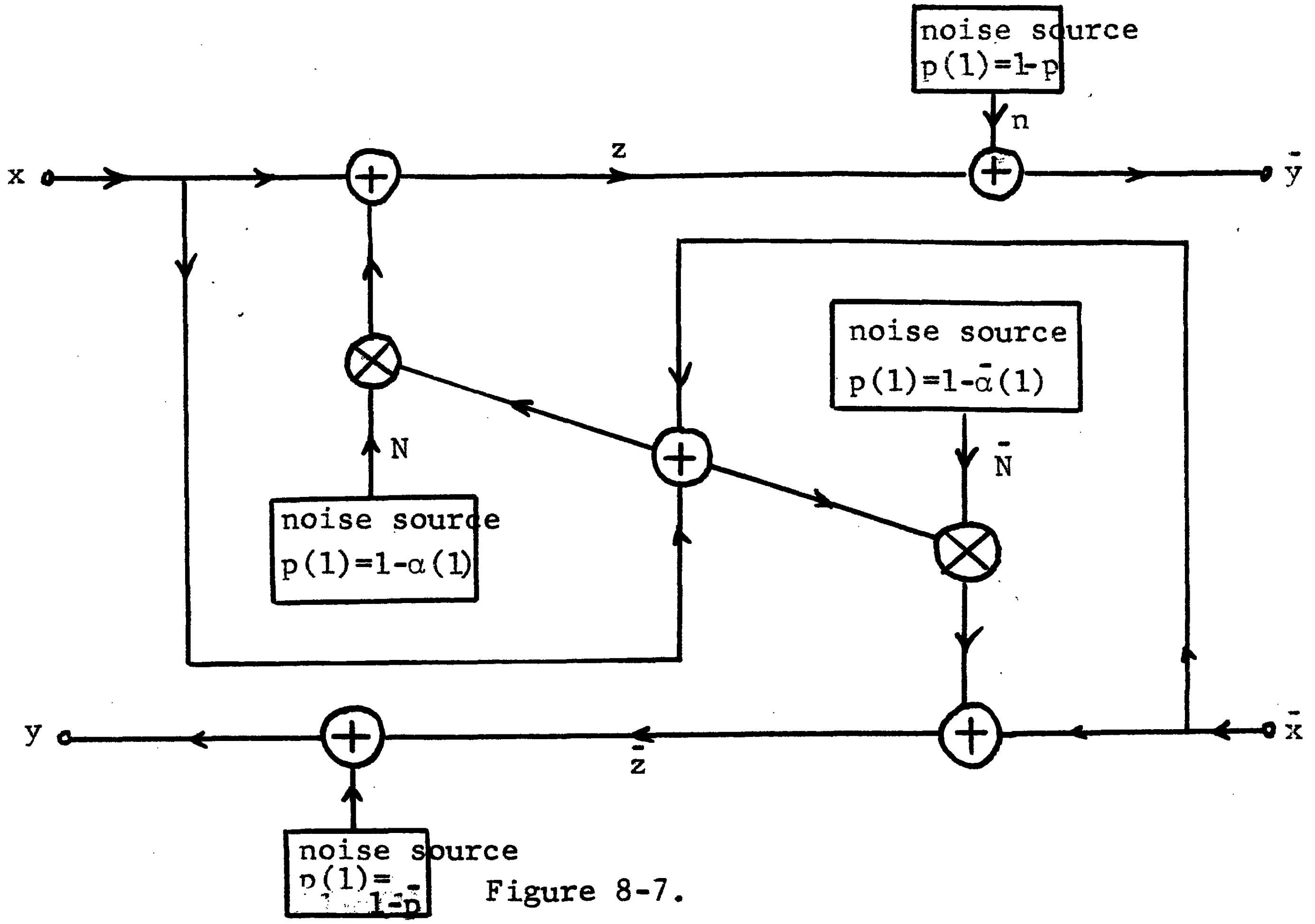


Figure 8-7.

It is seen in Figure 8-7 that the transmission noise is produced by four independent noise sources (outputs 0 and 1 with probabilities indicated), the noises N and $\bar{N}$ being "gated" by multipliers. N and $\bar{N}$ are blocked whenever the signals x and $\bar{x}$ are in agreement. One can derive a boolean expression for the received signals:

$$\begin{aligned} \bar{y} &= z \oplus n = (x \otimes N') \oplus (\bar{x} \otimes N) \oplus n \\ y &= \bar{z} \oplus n = (\bar{x} \otimes \bar{N}') \oplus (x \otimes \bar{N}) \oplus \bar{n} \end{aligned} \qquad (8\text{-}24)$$

and it is understood that

$$P(N = 1) = 1 - \alpha(1)$$

$$P(\overline{N} = 1) = 1 - \alpha(1)$$

$$P(n = 1) = 1 - p \qquad (8\text{-}25)$$

$$P(\overline{n} = 1) = 1 - p$$

From (8-24) we notice that:

a) If $N = 0$ (i.e. no noise) then $\overline{y} = x \oplus n$

If $N = 1$ (i.e. noise) then $\overline{y} = \overline{x} \oplus n$

and therefore the "variable" channel acts like an erasure channel (if there is noise the output tells nothing at all about input) but the receiver does not know where the erasures occurred.

(b) If it were not for the constant channel in cascade, the receiver would know sometimes when an erasure did not occur, i.e. if $n = 0$ and $\overline{y} \neq \overline{x}$ then necessarily $\overline{y} = x$,

c) Noise patterns n and N are independent of each other, and even if they are known, the output $\overline{y}$ will not, in general specify input x.

Because of the results pointed out above, channels belonging to horizontally symmetrical Class A can also be represented as in Figure 8-8.

Horizontal Class B

Here the conditions

$$p(0) = q(0) \qquad \overline{p}(0) = \overline{q}(0)$$

$$p(1) = q(1) \qquad \overline{p}(1) = \overline{q}(1) \qquad (8\text{-}26)$$

must be satisfied.

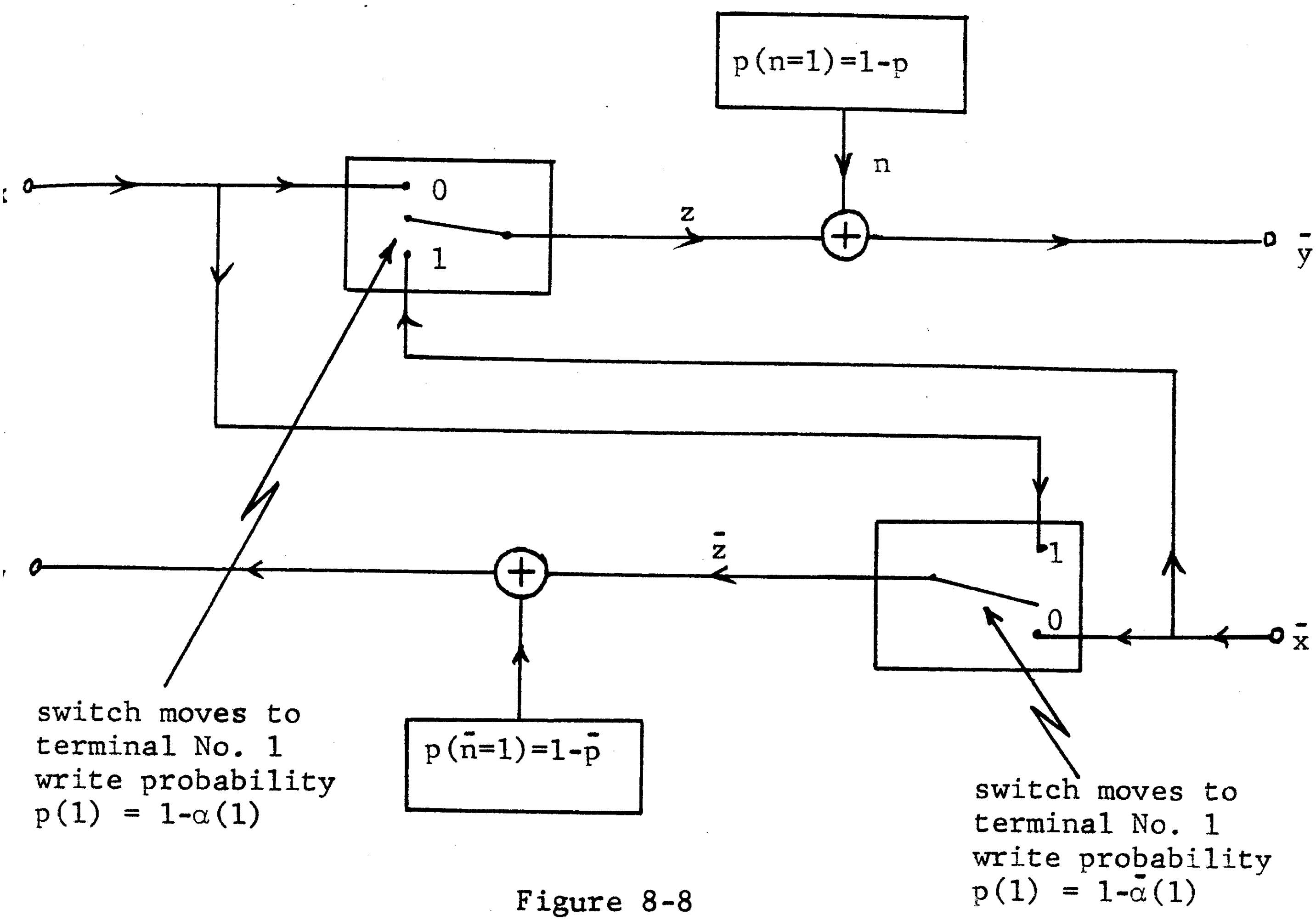


Figure 8-8

Inspecting relations (8-13) we see that we have:

$$p_0 = p_1 = p(0) = p$$

$$\bar{p}_0 = \bar{p}_1 = \bar{p}(0) = \bar{p}$$

$$\alpha(1) = \beta(1) = \frac{p(1) + p(0) - 1}{2p(0) - 1} \qquad (8\text{-}27)$$

$$\alpha(1) = \beta(1) = \frac{p(1) + p(0) - 1}{2p(0) - 1}$$

Therefore we obtain the channel represented by Figure 8-9.

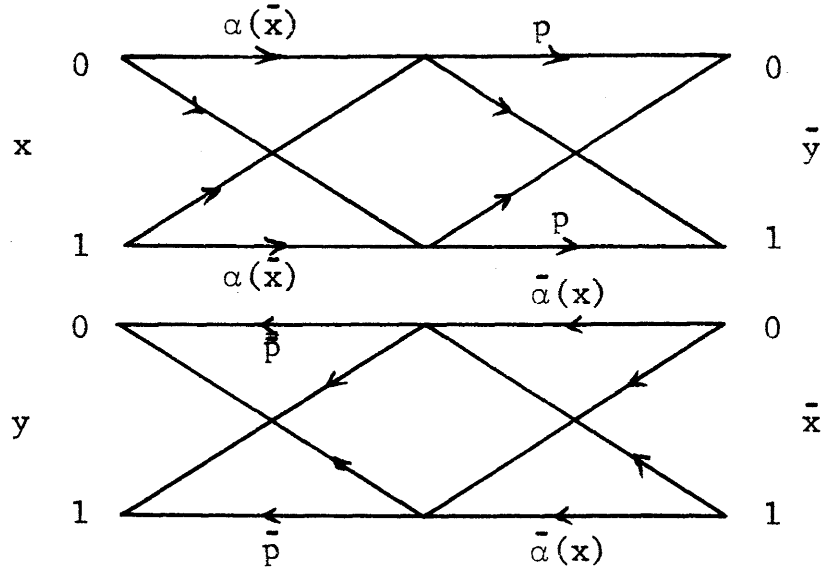


Figure 8-9

We thus have for both directions a variable symmetrical channel (which is noisy if the signal transmitted in the opposite direction was a 1) followed by a binary symmetric channel. Further reflection shows that the channel of Figure 8-9 can be represented by the configuration of Figure 8-10, the latter configuration being conducive to further interesting interpretation.

It is seen that in Figure 8-10 the transmission noise is produced by four independent noise sources, the noises N and $\bar{N}$ being "gated" by multipliers. N and $\bar{N}$ are blocked wherever x and $\bar{x}$ are equal to 0, respectively. The boolean expression for the received signal is:

$$\bar{y} = z \oplus n = x \oplus (N \otimes \bar{x}) \oplus n$$
$$y = \bar{z} \oplus \bar{n} = \bar{x} \oplus (\bar{N} \otimes x) \oplus \bar{n} \qquad (8\text{-}28)$$

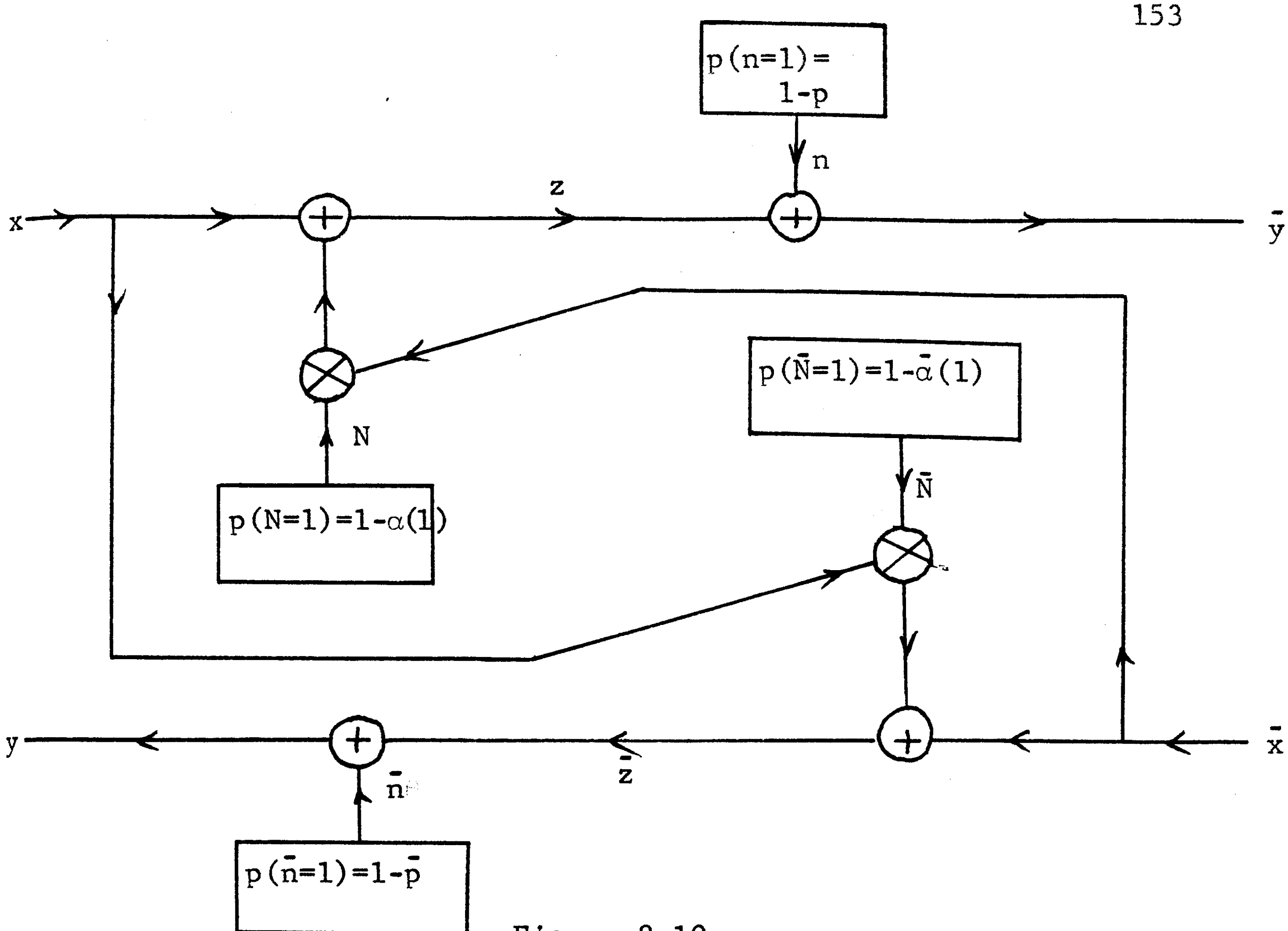


Figure 8-10

From (8-28) we notice that

a) If $N = 0$ (i.e. no noise) then $\bar{y} = x + n$

If $N = 1$ (i.e. noise) then $\bar{y} = x + \bar{x} + n$

In other words knowing the noise patterns N and n the receiver $\bar{y}$ can determine what x was, since it always knows the signal $\bar{x}$.

b) Except for the independent additive noise n, a receiver controls its own noise by its transmission. i.e. if it sends out only 0's, it can guarantee noiselessness in the variable part of the channel. Thus, except for the constant BSC in cascade, by reducing its signalling rate to zero the terminal in question can assure perfect, noiseless reception.

It can easily be seen that channels belonging to the horizontally symmetrical Class B can also be represented as in Figure 8-11.

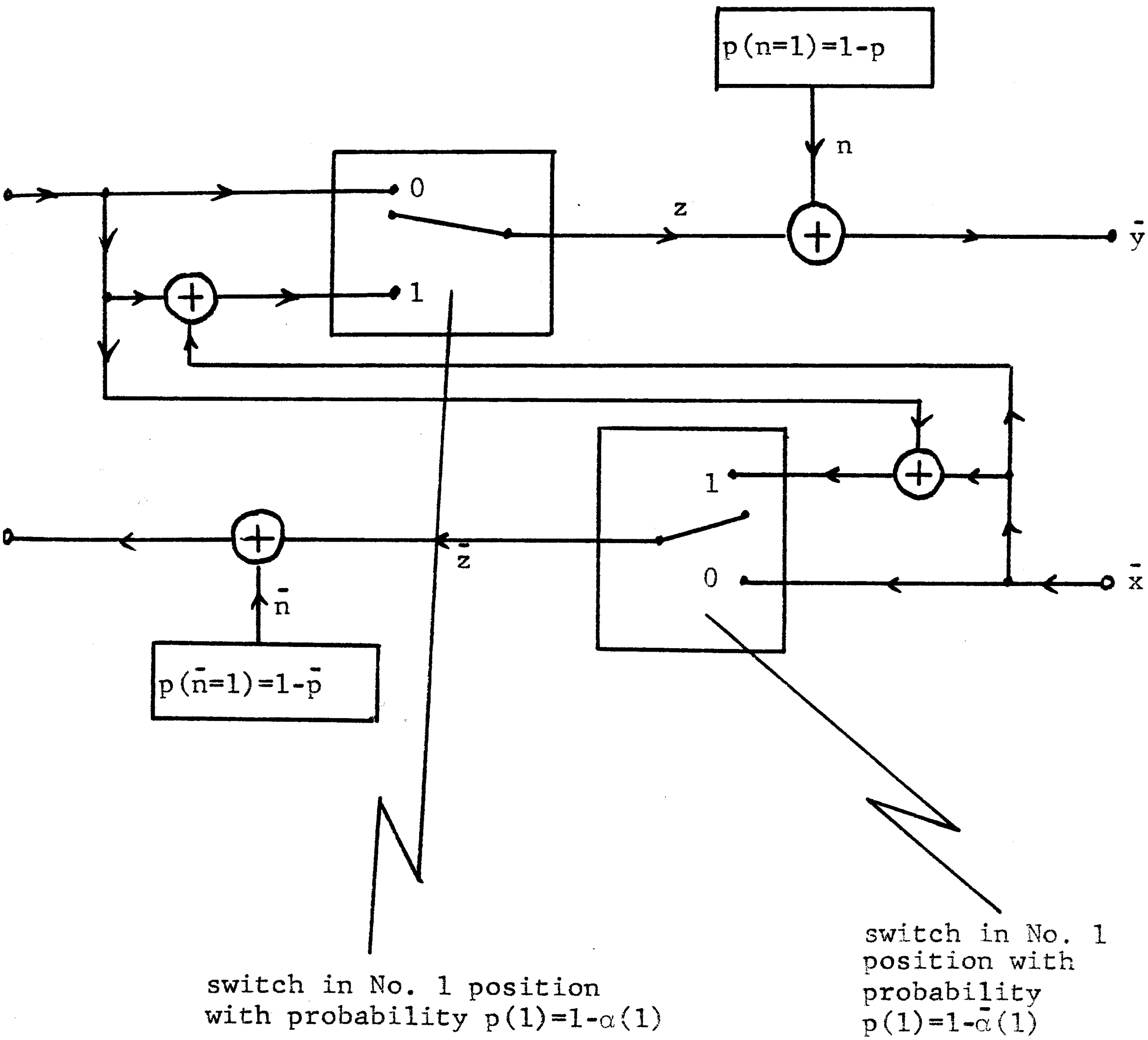


Figure 8-11

From the results of this section we may conclude that a study of channels consisting of portions in which all the noise is due entirely to signal transmission in the opposite direction, would give results very indicative of the situation of two-way binary channels unrestricted otherwise than by the relation (8-1). This may prove crucial in the selection of examples to illustrate the coding results derived in this thesis. In the present article we have also developed a tool for judging possible relevance of particular examples to the general case. Figures 8-8 and 8-11 provide us for given channels with an intuitive guide as to what kind of signal cooperation between opposing terminals could prove useful.

9. Properties of Two-Way Channel Codes

9.1 Philosophy of Strategy Coding

A word F of input symbols to the derived channel of Figure 4-2 can be looked upon as a strategy by which the signal input to the two-way channel is determined. I.e., the signal word F(Y) is decided in advance for any possible received signal word Y. Thus the kind of coding examined in this thesis can rightfully be called "strategy coding". It was demonstrated in Article 8 that a coding strategy for a two-way channel with specified transmission probabilities ought to result in such signal cooperation between the opposing transmitters as would reduce the noise in the variable part of the forward and reverse channel portions (see Figure 8-2). It was further shown that in essence only two different types of channels exist and hence that only two different general types of strategies are necessary.

It is clear from Article 7 that the possible decrease of noise in the variable part of the channel is paid for by a loss of information resulting from the fact that in strategy coding the identification of the signal sent is not, in general, sufficient for the correct decoding of the message associated with a symbol word.

One may then ask whether and how the gains of our method overbalance its drawbacks. It should first be pointed out that in a good code the strategies of the two terminals are correlated in the sense that the strategy of the forward terminal is best for a given strategy of the reverse terminal, and vice versa. (This optimization is accomplished by the maximization of the quantities on the right hand sides of inequalities (6-54) and (6-55) over the different possible input symbol probability distributions P(f) and $\overline{P}(\overline{f})$.). In view of the extreme complexity of the situation, the author is unable to describe in detail the character of those strategies which will insure the optimal correlation of input signals. Instead, he will try to indicate heuristically how such strategies work. The sought for principle is one of variable signalling rate. Namely, when using a good code the forward terminal will increase its rate at times in which by its "judgement" (a) its own decoder can afford the resulting noise increase, and (b) the reverse decoder can cope with the rate increase. Conversely the forward transmitter will lower its rate at times when either the forward or the reverse decoders have trouble operating. All this is done in correlation with the actions of the reverse transmitter, so that, e.g., both do not increase their rates at the same time. We will illustrate the above in a moment, but right now we wish to point out that the "judgement" mentioned is accomplished by virtue of the fact that the selection of the transmitted signal x is based, through the strategy function f, on the identity of ℓ past transmitted and received signals $\mathfrak{x}, \mathfrak{y}$, which in turn are indicative of the ℓ past received and transmitted signals $\bar{\mathfrak{x}}, \bar{\mathfrak{y}}$ at the reverse terminal.

Consider now a two-way channel whose forward and reverse portions are both representable by the Equivalent Channel B of Figure 8-4. A good code will then have the property that when used, the frequency of transmission of the signal 1 through the channel in both forward and reverse directions will be smaller than the frequency of transmisiion of the signal 0. Hence the signal 1 will carry more information than the signal 0. Now the transmission of a 1 from any terminal increases that terminal's reception noise. Hence the transmission frequency of 1's will in a good code be highest at times when the transmitting terminal's decoding is in good shape and the reception noise of the receiving terminal is expected to be low. Thus maximal use will be made of the increased information content of the contemplated signal. It is reasonable to expect that the above strategy can indeed be implemented.

The strategy to be adopted is not as obvious when dealing with a symmetrical channel of class A of Figure 8-6. In the variable part, noise is generated when the signals transmitted simultaneously from the two terminals are different. If the forward transmitter expects a 1 to be transmitted from the reverse side, increasing the probability of x=1 will decrease the channel noise level and also the contemplated signal's information content, and hence the rate. Therefore, information rate can be increased in favorable situations by increasing the frequency of digits apt to cause more noise, and it can be decreased in unfavorable situations by opposite action.. Thus a vehicle of fruitful signal cooperation between opposing terminals is conceivable.

9.2 Dependent Symbol Coding for the Derived Channel

It was shown by Shannon[1] (for discussion see Section 3.3) that the convex hull of the set of points $E(I(F;\bar{Y}/\bar{F}))$; $E(I(\bar{F};Y/F))$ obtained for different probability assignments $P(F)$ and $\bar{P}(\bar{F})$ will constitute the set of points of the capacity region of the two-way channel, as the code word length n and the symbol dependency length ℓ grow appropriately large. (The preceding statement is actually Shannon's theorem given in the notation and terminology of this thesis). According to Theorem 7-5 this capacity region will be included in the convex hull of points $E(I''(X;\bar{Y}/\bar{X}))$; $E(I'(\bar{X};Y/X))$ obtained for different probability assignments $P(F)$, $\bar{P}(\bar{F})$ for different sizes of n and ℓ. We have argued in the preceding section that the aim of the employment of strategic codes is the achievement of signal cooperation between opposing channel terminals, which would result in an increase of the size of the convex hull of average mutual information points $[E\{I''(X;\bar{Y}/\bar{X})\}, E\{I'(\bar{X};Y/X)\}]$ over that obtainable by non-strategic codes. We have discussed the above information measures in sections 7.2, 7.3 and 7.4, and we have concluded (see especially Theorem 7-1), that any value of $E_{P(F)\bar{P}(\bar{F})}(I''(X;\bar{Y}/\bar{X}),I'(\bar{X};Y/X))$ can be approached arbitrarily closely by independent symbol probability assignments $P(f)$ and $\bar{P}(\bar{f})$, if the dependency length ℓ is made long enough. The question therefore arises whether all results achievable by dependent symbol coding (e.g. by word probability assignments $P(F)$, $\bar{P}(\bar{F})$) cannot also be achieved by independent symbol coding associated with the symbol assignment probabilities $P(f)$, $\bar{P}(\bar{f})$, when an appropriately long dependency length ℓ is used. The answer to this question hinges according to the conclusion of Theorem 7-5, on whether or not the use of dependent sources can reduce the coding loss $E(I(Y;\bar{Y}/F,\bar{F}))$.

The author did not succeed in answering the last question. Two possible approaches to the problem exist and the discussion below will give the reader an idea of why neither of them was fruitful.

Firstly, one could try to hold the value of the quantities $E\{I''(X;\bar{Y}/\bar{X})\}$, $E\{I'(\bar{X};Y/X)\}$ constant, and inquire whether it is possible under this constraint to reduce the coding loss $E\{I(Y;\bar{Y}/F,\bar{F})\}$ by switching from independent to dependent symbol sources. Unfortunately it follows from Theorem 7-1 that no such constraint, except in some degenerate cases, can be satisfied, since for independent symbol sources the channel input signals depend on a finite number of preceding input and output signals, while for dependent sources the channel input signals depend on an infinite number of preceding input and output signals.

Secondly, one could attempt to attack the problem in all its breadth, by trying to prove that every point $E(I(F;\bar{\bar{Y}}/\bar{F}))$; $E(I(\bar{F};\bar{Y}/F))$ achievable by dependent sources $P(F)$, $\bar{P}(\bar{F})$ for a given ℓ can be approached arbitrarily closely by independent sources $P(f)$, $\bar{P}(\bar{f})$ when an appropriately long dependency length ℓ^* is used. But the usual approach to similar proofs for the one-way channel (see Fano, (4), p.125), in which the independent distribution $P^*(f)$ is determined from the dependent distribution $P(F)$ by the equation

$$P^*(f_i) = \sum_{f_1,\ldots,f_{i-1},f_{i+1},\ldots,f_n} P(F = f_1, f_2,\ldots,f_i,\ldots,f_n) \qquad (9\text{-}1)$$

will not work. In fact, it is intuitively clear that if the suggested theorem were true, then the dependency lengths ℓ^* and ℓ of the symbols put out by the dependent and independent sources, respectively, would in general not be equal, and hence that the symbols put out by the equivalent dependent and independent sources would be of a different character.

Therefore we are reduced to only asserting that for lossless channels (i.e. those satisfying either of conditions (a), (b) or (c) of Theorem 7-7) the capacity regions of dependent and independent coding are equal.

For reasons similar to those given in the paragraph preceding the last one (made more evident by the proof that follows) the author did not succeed in proving the independent coding version of

<u>Theorem 9-1</u>

Given the transition probabilities $p(y/x,\bar{x})$ and $\bar{p}(\bar{y}/x,\bar{x})$, for a two-way channel, the rates

$$R = E(I(F;\bar{Y}/\bar{F})) \qquad \bar{R} = E(I(\bar{F};Y/F)) \tag{9-2}$$

are concave downward functions of the assigned input probabilities $P(F)$ and $\bar{P}(\bar{F})$.

<u>Proof</u>:

It is only necessary to show that

$$R(\tfrac{1}{2}[P_1(F)+P_2(F)],[\bar{P}_1(\bar{F})+\bar{P}_2(\bar{F})]\tfrac{1}{2}) \geq \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2} R(P_i(F),\bar{P}_j(\bar{F})) \tag{9-3}$$

and the theorem concerning both R and R will follow.

If we could show that

$$R(\tfrac{1}{2}(P_1(F)+P_2(F)),\ \bar{P}(\bar{F})) \geq \tfrac{1}{2}[R(P_1(F),\bar{P}(\bar{F}))+R(P_2(F),\bar{P}(\bar{F})), \tag{9-4}$$

and also

$$R(P(F),\ \tfrac{1}{2}(\bar{P}_1(\bar{F})+\bar{P}_2(\bar{F}))) \geq \tfrac{1}{2}[R(P(F),\bar{P}_1(\bar{F}))+R(P(F),\bar{P}_2(\bar{F}))], \tag{9-5}$$

then (9-3) would follow, since by (9-4)

$$\frac{1}{2}\, R(\tfrac{1}{2}(P_1(F)+P_2(F)),\bar{P}_i(\bar{F})) \geq \frac{1}{4}\sum_{j=1}^{2} R(P_j(F),\bar{P}_i(\bar{F})), \tag{9-6}$$

and by (9-5)

$$R(\tfrac{1}{2}(P_1(F)+P_2(F)),\tfrac{1}{2}(\bar{P}_1(\bar{F})+\tfrac{1}{2}\bar{P}_2(\bar{F}))) \geq \frac{1}{2}\sum_{i} R(\tfrac{1}{2}(P_1(F)+P_2(F));$$

$$\bar{P}_i(\bar{F})) \geq \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2} R(P_j(F),\bar{P}_i(\bar{F})). \tag{9-7}$$

We will first prove (9-4). Let

$$P_0(F) = \frac{1}{2}\,(P_1(F) + P_2(F)). \tag{9-8}$$

then

$$R(P_i(F),\bar{P}(\bar{F}))=E(\log\frac{Pr_i(\bar{Y}/F,\bar{F})}{Pr_i(\bar{Y}(\bar{F})}) = H_i(\bar{Y}/\bar{F})-H_i(\bar{Y}/F\bar{F}), \quad i=0,1,2 \tag{9-9}$$

where by equation (7-81)

$$Pr_i(\bar{Y}/F\bar{F}) = \sum_{Y} Pr'(Y/F(Y)\bar{F}(\bar{Y}))\, Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y})) \quad \text{for all } i, \tag{9-10}$$

and by (7-82)

$$Pr_i(\bar{Y}/\bar{F}) = \sum_{Y,F} Pr'(Y/F(Y)\bar{F}(\bar{Y}))Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y}))P_i(F). \tag{9-11}$$

But

$$H_0(\bar{Y}/F\bar{F})= - \sum_{F,\bar{F}} P_0(F)\bar{P}(\bar{F})\sum_{Y} Pr(\bar{Y}/F\bar{F})\,\log Pr(\bar{Y}/F\bar{F})$$

$$= \frac{1}{2}\,(H_1(\bar{Y}/F\bar{F}) + H_2(\bar{Y}/F\bar{F})). \tag{9-12}$$

Moreover

$$\frac{1}{2}[H_1(\bar{Y}/\bar{F}) + H_2(\bar{Y}/\bar{F})] - H_0(\bar{Y}/\bar{F}) =$$

$$+ \frac{1}{2}[\sum_{\bar{F},\bar{Y}} \bar{P}(\bar{F})Pr_1(\bar{Y}/\bar{F})\,\log\frac{Pr_0(\bar{Y}/\bar{F})}{Pr_1(\bar{Y}/\bar{F})} +$$

$$+ \sum_{\bar{F}\,\bar{Y}} P(F)Pr_2(\bar{Y}/\bar{F})\,\log\frac{Pr_0(\bar{Y}/\bar{F})}{Pr_2(\bar{Y}/\bar{F})}\,] \leq 0. \tag{9-13}$$

Relations (9-12) and (9-13) together with (9-9) prove (9-4).

Our next task is to prove (9-5). Let now

$$R(P(F),\bar{P}_i(\bar{F})) = E(\log \frac{Pr_i^*(\bar{Y}/F\bar{F})}{Pr_i^*(\bar{Y}/\bar{F})}) = H_i^*(\bar{Y}/\bar{F}) - H_i^*(\bar{Y}/F\bar{F}) \qquad (9\text{-}14)$$

where again $\bar{P}_0(\bar{F}) = \frac{1}{2}(\bar{P}_1(\bar{F}) + \bar{P}_2(\bar{F}))$

It follows from (7-81) and (7-82) that probabilities $Pr_i^*(\bar{Y}/F,\bar{F})$ and $Pr_i^*(\bar{Y}/\bar{F})$ are constant over i=0,1,2. Hence

$$H_0^*(\bar{Y}/F\bar{F}) = -\sum_{F,\bar{F}} P(F)\bar{P}_0(\bar{F}) \sum_{\bar{Y}} Pr(\bar{Y}/F\bar{F}) \log Pr(\bar{Y}/F\bar{F}) =$$

$$= \frac{1}{2} H_1^*(\bar{Y}/F\bar{F}) + \frac{1}{2} H_2^*(\bar{Y}/F\bar{F}), \qquad (9\text{-}15)$$

and

$$H_0^*(\bar{Y}/\bar{F}) = -\sum_{\bar{F}\bar{Y}} \bar{P}_0(\bar{F}) Pr(\bar{Y}/\bar{F}) \log Pr(\bar{Y}/\bar{F}) =$$

$$= \frac{1}{2} (H_1^*(\bar{Y}/\bar{F}) + H_2^*(\bar{Y}/\bar{F})). \qquad (9\text{-}16)$$

Probabilities (9-15) and (9-16) together with (9-14) prove (9-5).

Q.E.D.

We next prove a similar theorem for dependent function assignments about the coding loss:

<u>Theorem 9-2</u>

Given the transition probabilities $p(y/x,\bar{x})$ and $\bar{p}(\bar{y}/x,\bar{x})$ for a two-way channel. For a probability assignment of input words to the corresponding derived channel, $P(F)$ and $\bar{P}(\bar{F})$, define the coding loss function

$$L(P(F),\bar{P}(\bar{F})) = E\left\{\log \frac{Pr(Y,\bar{Y}/F\bar{F})}{Pr(Y/F\bar{F})Pr(\bar{Y}/F\bar{F})}\right\}. \qquad (9\text{-}17)$$

Given any two pairs of input probability assignments $(P_1(F),\bar{P}_1(\bar{F}))(P_2(F),\bar{P}_2(\bar{F}))$, the equation

$$L(\frac{1}{2}(P_1(F)+P_2(F)),\ \frac{1}{2}(\bar{P}_1(\bar{F})+\bar{P}_2(\bar{F}))) = \frac{1}{4}\sum_{i,j=1}^{2} L(P_i(F),\bar{P}_i(\bar{F})) \tag{9-18}$$

holds.

Proof:

Because of symmetry, all we must show is that

$$L(\frac{1}{2}(P_1(F)+P_2(F)),\bar{P}(\bar{F})) = \frac{1}{2}[L(P_1(F),\bar{P}(\bar{F}))+L(P_2(F),\bar{P}(\bar{F}))]. \tag{9-19}$$

The equation (9-18) will then follow by the reasoning employed in proving (9-3) from (9-4) and (9-5).

Let equation (9-8) hold. Then

$$L(P_i(F),\bar{P}(\bar{F})) = H_i(Y/F,\bar{F})+H_i(\bar{Y}/F\bar{F})-H_i(Y\bar{Y}/F\bar{F}) \quad i=0,1,2. \tag{9-20}$$

But from (9-12) we get

$$H_0(\bar{Y}/F,\bar{F}) = \frac{1}{2}H_1(\bar{Y}/F,\bar{F}) + \frac{1}{2}H_2(\bar{Y}/F\bar{F}), \tag{9-21}$$

and from (9-16) we get

$$H_0(Y/F\bar{F}) = \frac{1}{2}H_1(Y/F\bar{F}) + \frac{1}{2}H_2(Y/F\bar{F}). \tag{9-22}$$

From (7-45) it follows that

$$Pr_i(Y\bar{Y}/F\bar{F}) = Pr'(Y/F(Y)\bar{F}(\bar{Y}))\ Pr''(\bar{Y}/F(Y)\bar{F}(\bar{Y})) \quad \text{for all } \underline{i}\ , \tag{9-23}$$

so that

$$H_0(Y,\bar{Y}/F\bar{F}) = -\sum_{F\bar{F}} P_0(F)\bar{P}(\bar{F}) \sum_{Y,\bar{Y}} Pr(Y\bar{Y}/F\bar{F})\log Pr(Y\bar{Y}/F\bar{F}) =$$

$$= \frac{1}{2}H_1(Y,\bar{Y}/F\bar{F}) + \frac{1}{2}H_2(Y,\bar{Y}/F\bar{F}) \tag{9-24}$$

Equality (9-19) then follows from (9-20,-21,-22,-24).

Q.E.D.

Finally, from theorems 9-1 and 9-2, by use of theorem 7-5 there follows

<u>Theorem 9-3</u>

Given the transition probabilities $p(y/x,\bar{x})$ and $\bar{p}(y/x,\bar{x})$ for a two-way channel, the average sum of the signal informations received through the channel at the two terminals,

$$E(I(F,Y;\bar{F},\bar{Y})) \quad = \quad E(I(X,Y;\bar{X},\bar{Y})), \qquad (9\text{-}25)$$

is a concave downward function of the assigned input probabilities $P(F)$ and $\bar{P}(\bar{F})$ of the corresponding derived channel.

From the preceding theorems the usual convexity consequences about rate maximizations follow. Amongst these it is worth mentioning that for any given assignment $\bar{P}(\bar{F})$ there is in general only one assignment $P(F)$ which will maximize $E(I(F,Y;\bar{F},\bar{Y}))$ and one assignment $P(F)$ which will maximize $E(I(Y;F/F)+I(Y;F/F))$. (See Fano,[4] p.135).

9.3 Random Construction of Codes

In order to construct a random ensemble from which to pick good codes for signalling at close to capacity rates, one would ordinarily proceed as follows:

(a) Pick a dependency length ℓ (see Section 5.3).

(b) Pick decoding criterion length m (see equation (6-10a) and compare with (6-38)).

(c) Maximize the expression (9-26) over the possible independent probability symbol assignments P(f) and $\bar{P}(\bar{f})$, where λ is a positive real number controlling the intended relation between forward and reverse transmission rates (see equations (6-50) and (6-51))

$$E\left\{\log \frac{Pr(\bar{y}/f,\bar{f},\mathcal{f}^m,\bar{\mathcal{f}}^m,\bar{\mathcal{y}}^m,\bar{s}^{-m})}{Pr(\bar{y}/\bar{f},\bar{\mathcal{f}}^m,\bar{\mathcal{y}}^m,\bar{s}^{-m})}\right\} + \tag{9-26}$$

$$+\lambda\ E\left\{\log \frac{Pr(y/f,\bar{f},\mathcal{f}^m,\bar{\mathcal{f}}^m,\mathcal{y}^m,s^{-m})}{Pr(y/f,\mathcal{f}^m,\mathcal{y}^m,s^{-m})}\right\}.$$

(d) Using the optimizing distributions P(f) and $\bar{P}(\bar{f})$ found in (c), evaluate ε and $\bar{\varepsilon}$ for equations (6-52) and (6-53) by use of the procedure indicated in (IV-47) involving (IV-43) and (IV-45). Compute the expressions $\log(1+\varepsilon)$ and $\log(1+\bar{\varepsilon})$ and decide whether the length m picked in (b) was sufficiently large (see (6-50) and (6-51)). If not, go through the complete procedure starting with step (b) again (note that for a different m the optimizing probabilities P(f) and $\widehat{P}(\bar{f})$ will, in general, have different values).

The procedure listed above is certainly a very complicated one and must in any case be carried out with the help of a computer. Most prohibiting is step (c), not only because it involves evaluation of the complicated expressions (9-26), but especially because optimization over a set of $2^{2^{2\ell}}$ variables is necessary. The enormity of the task is readily realized if it is pointed out that the number $2^{2^{2\ell}}$ equals 32 for ℓ=1, and 131,072 for ℓ=2! The implications of this remark might seem hilarious to a detached observer but not to the author who labored hard to reach this point. Theorem 7-6 suggests a possible way out of this calamity. One might, perhaps maximize the expression

$$E\left\{\log \frac{\bar{p}(\bar{y}/x,\bar{x})}{\Pr(\bar{y}/\bar{x},\mathcal{X}^m,\bar{\mathcal{Y}}^m,\bar{s}^{-m})}\right\} + \\ + \lambda\ E\left\{\log \frac{p(y/x\bar{x})}{\Pr(y/x,\mathcal{X}^m,\mathcal{Y}^m,s^{-m})}\right\} \tag{9-27}$$

over the signal assignment probabilities $q_{\mathcal{X}\mathcal{Y}}(x)$, $\bar{q}_{\bar{\mathcal{X}}\bar{\mathcal{Y}}}(\bar{x})$, and then using the optimizing probabilities $q^*_{\mathcal{X}\mathcal{Y}}(x)$, $\bar{q}^*_{\bar{\mathcal{X}}\bar{\mathcal{Y}}}(\bar{x})$ as constraints in the relations (7-11), minimize the expression

$$E\left\{\log \frac{\Pr(\bar{y}/f,\bar{f},\mathcal{F}^m,\bar{\mathcal{F}}^m,\mathcal{Y}^m,\bar{\mathcal{Y}}^m,s^{-m},\bar{s}^{-m})}{\Pr(\bar{y}/f,\bar{f},\mathcal{F}^m,\bar{\mathcal{F}}^m,\bar{\mathcal{Y}}^m,\bar{s}^{-m})}\right\} + \\ + \lambda\ E\left\{\log \frac{\Pr(y/f,\bar{f},\mathcal{F}^m,\bar{\mathcal{F}}^m,\mathcal{Y}^m,\bar{\mathcal{Y}}^m,s^{-m},\bar{s}^{-m})}{\Pr(y/f,\bar{f},\mathcal{F}^m,\bar{\mathcal{F}}^m,\mathcal{Y}^m,s^{-m})}\right\} \tag{9-28}$$

over the probabilities $P(f)$ and $\bar{P}(\bar{f})$. This can be done anyway for lossless channels satisfying conditions (a), (b), or (c) of Theorem 7-7, and in such cases minimization (9-28) is not even carried out. It is true that in (9-28) the number of variables is the same as in (9-26), but we could perhaps eliminate this last step in practice, if a good way was found to satisfy the constraints (7-11) so that (9-28) would be small compared with (9-27). As will be seen later, such a good way really exists, and so we would like to find out how good or how bad the quasi-optimization involving (9-27) and (9-28) is compared with the totimization (9-26).

Before dealing with the problem stated in the previous sentence, it is worthwhile to point out that because of Shannon's outer and inner bounds and Theorem 7-9, the expression (9-27) is to be optimized only over a restricted range of probabilities $q(x)$, $\bar{q}(\bar{x})$. Namely over those combinations $q(x)$, $\bar{q}(\bar{x})$ for which the expression would give a distribution $Pr(x,\bar{x})$ such that for it the point

$$E\left\{\log \frac{Pr(\bar{y}/x\bar{x})}{Pr(\bar{y}/\bar{x})}\right\}, \quad E\left\{\log \frac{Pr(y/x\bar{x})}{Pr(y/x)}\right\} \qquad (9\text{-}29)$$

would lie outside of Shannon's inner bound to the capacity region of the given two-way channel.

We will now turn our attention to the questions discussed in the paragraph preceding the last one and will prove

Theorem 9-4

Given a two-way channel whose transmission probabilities do not fulfill any of the conditions (a), (b), or (c) of Theorem 7-7, and are such that $0 < p(y/x\bar{x}) < 1$, $0 < \bar{p}(\bar{y}/x\bar{x}) < 1$ for all signals $x, \bar{x}, y, \bar{y}$. If the actual capacity region exceeds Shannon's inner bound, the maximization of the expression (9-26) over the input assignment probabilities $P(f)$ and $\bar{P}(\bar{f})$ will never result through equations (7-11) in signal probabilities $q_{\mathfrak{x}\mathfrak{y}}(x)$, $\bar{q}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(\bar{x})$ which would maximize the expression (9-27). Hence the maximization involving (9-26) and quasi-maximization involving (9-27) and (9-28) will have, in the general case, different results.

Proof:

For notational simplicity we will prove a slightly narrower theorem, but as will be seen, the argument can be generalized to cover the entire extent of Theorem 9-4.

From Theorem 7-5 we have

$$E(I(F;\bar{Y}/\bar{F})) + E(I(\bar{F};Y/F)) = E(I(XY;\bar{X}\bar{Y})) - E(I(Y;\bar{Y}/F,\bar{F})), \quad (9\text{-}30)$$

and we are free to maximize the right hand side of the above.

Let

$$h_{\mathfrak{x}\mathfrak{y}}(q_{\mathfrak{x}\mathfrak{y}}(0), f_0, \ldots, f_{2^{2^{2\ell}}-1}) \equiv q_{\mathfrak{x}\mathfrak{y}}(0) - \sum_{f \ni f(\mathfrak{x},\mathfrak{y})=0} P(f)$$

$$\bar{h}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(\bar{q}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(0), \bar{f}_0, \ldots, \bar{f}_{2^{2^{2\ell}}-1}) \equiv \bar{q}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(0) - \sum_{\bar{f} \ni \bar{f}(\bar{\mathfrak{x}},\bar{\mathfrak{y}})=0} \bar{P}(\bar{f})$$

$$h(f_0,\ldots,f_{2^{2^{2\ell}}-1}) \equiv 1 - \sum_{\forall f} P(f)$$

$$\bar{h}(\bar{f}_0,\ldots,\bar{f}_{2^{2^{2\ell}}-1}) \equiv 1 - \sum_{\forall \bar{f}} \bar{P}(\bar{f}) \ . \qquad (9\text{-}31)$$

Relations (9-31) constitute $2(2^{2\ell}+1)$ constraints between the quantities $q_{xy}(0)$, $\bar{q}_{\bar{x}\bar{y}}(0)$, $P(f)$, $\bar{P}(\bar{f})$. Defining for each constraint a Lagrangian multiplier, we then wish to find the extrema with respect to the variables $q(0)$, $\bar{q}_{\bar{x}\bar{y}}(0)$, $P(f)$, $\bar{P}(\bar{f})$ of the function

$$E(I(X,Y;\bar{X}\bar{Y}))-E(I(Y;\bar{Y}/F,\bar{F})) + \sum_{x,y} \lambda_{xy} h_{xy} +$$

$$+ \sum_{\bar{x},\bar{y}} \bar{\lambda}_{\bar{x}\bar{y}} \bar{h}_{\bar{x}\bar{y}} + \lambda h + \bar{\lambda}\bar{h}, \qquad (9\text{-}32)$$

where we let $q_{xy}(1) = 1-q_{xy}(0)$ and $q_{\bar{x}\bar{y}}(1) = 1-q_{\bar{x}\bar{y}}(0)$. The solution we are looking for will be the maximum of the obtained extrema. The different extrema themselves are found by partially differentiating (9-32) with respect to all the $2^{2\ell}$ variables $q_{xy}(0)$ the $2^{2\ell}$ variables $\bar{q}_{\bar{x}\bar{y}}(0)$, the $2^{2^{2\ell}}$ variables $P(f)$, the $2^{2^{2\ell}}$ variables $\bar{P}(\bar{f})$, the $2^{2\ell}$ variables λ_{xy}, the $2^{2\ell}$ variables $\bar{\lambda}_{\bar{x}\bar{y}}$, and the 2 variables λ and $\bar{\lambda}$. We will thus get $4 2^{2\ell} + 2 2^{2^{2\ell}} + 2$ simultaneous equations which should lead to solutions with $0 \leq q_{xy}(0) \leq 1$; $0 \leq \bar{q}_{\bar{x}\bar{y}}(0) \leq 1$, in order to be acceptable. It is expected that in the general case the solutions will lie on the boundary of the $P(f)$ and $\bar{P}(\bar{f})$ coordinates, i.e. that some $P(f)$ and $\bar{P}(\bar{f})$ will ultimately be made 0. It is further expected that none of the $\bar{q}_{\bar{x}\bar{y}}(0)$, $q_{xy}(0)$ will in the general case come out to be either 0 or 1. It will prove convenient to display the different types of the simultaneous equations mentioned above:

$$
\begin{aligned}
&\text{(i)} \quad \frac{\partial}{\partial q_{xy}(0)} E\{I(XY;\bar{X}\bar{Y})\} + \lambda_{xy} = 0 \\
&\text{(ii)} \quad \frac{\partial}{\partial \bar{q}_{\bar{x}\bar{y}}(0)} E\{I(XY;\bar{X}\bar{Y})\} + \bar{\lambda}_{\bar{x}\bar{y}} = 0 \qquad\qquad (9\text{-}33) \\
&\text{(iii)} \quad -\frac{\partial}{\partial P(f_i)} E\{I(Y;\bar{Y}/F,\bar{F})\} - \sum_{xy \ni f_i(x,y)=0} \lambda_{xy} - \lambda = 0 \\
&\text{(iv)} \quad -\frac{\partial}{\partial \bar{P}(\bar{f}_i)} E\{I(Y;\bar{Y}/F,\bar{F})\} - \sum_{\bar{x}\bar{y} \ni f_i(\bar{x}\bar{y})=0} \bar{\lambda}_{\bar{x}\bar{y}} - \bar{\lambda} = 0 \\
&\text{(v)} \quad h_{xy}(q_{xy}(0),\ f_0,\ldots,f_{2^{2^{2\ell}}-1}) = 0 \\
&\text{(vi)} \quad \bar{h}_{\bar{x}\bar{y}}(\bar{q}_{\bar{x}\bar{y}}(0),\ \bar{f}_0,\ldots,\bar{f}_{2^{2^{2\ell}}-1}) = 0 \\
&\text{(vii)} \quad h(f_0,f_1,\ldots,f_{2^{2^{2\ell}}-1}) = 0 \\
&\text{(viii)} \quad \bar{h}(\bar{f}_0,\bar{f}_1,\ldots,\bar{f}_{2^{2^{2\ell}}-1}) = 0
\end{aligned}
$$

The anticipated setting of $P(f)$ and $\bar{P}(\bar{f})$ to zero for some f and $\bar{f}$ will in no way change the equations (i) and (ii). It will eliminate some equations (iii) and (iv), but will not change the form of the remaining ones. It will leave the number of equations (v), (vi), (vii) and (viii) constant, while changing their form in so far as some of the terms in the right hand side sums of (9-31) will be eliminated.

When maximizing the function $E\{I(X,Y;\bar{X}\bar{Y})\}$, we will get equations of the type

$$\text{(i)} \quad \frac{\partial}{\partial q_{xy}(0)} E\{I(X,Y;\bar{X}\bar{Y})\} = 0$$

$$\text{(ii)} \quad \frac{\partial}{\partial \bar{q}_{\bar{x}\bar{y}}(0)} E\{I(X,Y;\bar{X}\bar{Y})\} = 0 , \tag{9-34}$$

where it is again understood that $q_{xy}(1) = 1-q_{xy}(0)$ and $\bar{q}_{\bar{x}\bar{y}}(1)=1-\bar{q}_{\bar{x}\bar{y}}(1)$. It then follows that if (9-34) is to lead to the same solution as (9-33), then for the absolute maximum at least, the optimizing variables $\lambda_{xy}, \bar{\lambda}_{\bar{x}\bar{y}}$ must come out to be $\lambda_{xy} = \bar{\lambda}_{\bar{x}\bar{y}} = 0$ for all $(xy\bar{x}\bar{y})$. But then the constraints (9-33 -v, -vi) play no role whatever in the minimization of $E\{I(Y;\bar{Y}/F\bar{F})\}$. This in turn implies that the function $E\{I(Y;\bar{Y}/F\bar{F})\}$ will attain its absolute minimum for such $P(f)$ and $\bar{P}(\bar{f})$ as would satisfy the constraints imposed by the distributions $q_{xy}(0)$ and $\bar{q}_{\bar{x}\bar{y}}(0)$ obtained from the optimization of $E\{I(XY;\bar{X}\bar{Y})\}$. However, we know that under the assumptions about the two-way channel listed in this theorem the unique minimum of the function $E\{I(Y;\bar{Y}/F\bar{F})\}$ is obtained for the distribution

$$P(f) = \begin{cases} P(f) & \text{if } f(x,y) = f(x) \quad \forall x,y \\ 0 & \text{otherwise.} \end{cases} \tag{9-35}$$

But the above distribution, if constraints (9-33-v,-vi) are to be satisfied, leads to the conclusion that the $q_{xy}(0)$, $\bar{q}_{\bar{x}\bar{y}}(0)$ optimizing E I(XY;XY) would be

$$q_{xy}(0) = q_{x}(0) \quad \forall x,y \; ; \; \bar{q}_{\bar{x}\bar{y}}(0) = \bar{q}_{\bar{x}}(0) \quad \forall \bar{x},\bar{y}, \tag{9-36}$$

which under this theorem's assumptions about the two-way channel is false. This proves the theorem

Q.E.D.

Actually it is highly probable that the conclusions of the above theorem hold even if the restriction $0 < p(y/xx) < 1;\ 0 < p(y/xx) < 1$ on the transmission probabilities of the two-way channel is removed. We must then conclude that our quasi-optimization scheme will in most cases be just that-it will not be equivalent to a full-fledged optimization. On the other hand, for practical reasons (namely that for $\ell=1$ the number of variables $q_{xy}(0)+\bar{q}_{\bar{x}\bar{y}}(0)$ is 8 and for $\ell=2$ the number is 32) we might still be forced to quasi-optimize rather than optimize, and then by some procedure find the distributions $P(f)$ and $\bar{P}(\bar{f})$ which would satisfy the constraints (9-33-v-vi). Such procedure will be described in the next section.

We wish to conclude this section by observing that a consequence of Theorem 9-4 is that in order to maximize the flow of useful information through the channel, one must not in general maximize the total flow of information through the channel. Rather, one must make a compromise and send less information through the channel, of which however a greater part can be used for message identification. Thus, up to a certain point, an increase in total information flow through the channel due to strategy coding can be made in such a way that an increase in useful information will correspond to it; beyond that point however, any increase in total information will be accompanied by an even greater increase in coding information loss, so that the net amount of useful information will actually decrease.

9.4 Canonical Distributions $P(f), \bar{P}(\bar{f})$ Satisfying the Constraints of $\{q_{xy}(0)\}, \{\bar{q}_{\bar{x}\bar{y}}(0)\}$ Distributions

Given a distribution $\{q_{xy}(0)\}$, we can assign indexes $i \varepsilon (1, 2, \ldots, 2^{2\ell})$ to the various configurations (x,y) according to the following rule:

When $i < j$ then either $q_{(xy)_i}(0) < q_{(xy)_j}(0)$

or $q_{(xy)_i}(0) = q_{(xy)_j}(0)$ and the integer represented by the binary number $(x,y)_i$ is smaller than the integer represented by $(x,y)_j$.

(9-37)

The index rule (9-37)then assigns a unique and different index to each different configuration (x,y).

For the same given distribution $\{q_{xy}(0)\}$ we will define a new notation for the symbols f:

Given a function f such that

$$f(x,y) = \begin{cases} 1 & \text{for } (x,y) \varepsilon \{(x,y)_{i_1}, (x,y)_{i_2}, \ldots, (x,y)_{i_n}\} \\ 0 & \text{for remaining } (x,y) \end{cases} \qquad (9\text{-}38)$$

Let $f \equiv f^*_{i_1, i_2, \ldots, i_n}$.

Also the above rule is unambiguous, but it should be stressed that it applies only for a specified $\{q_{xy}(0)\}$. For a different $\{q_{xy}(0)\}$ the rule operates in general differently. Note that according to (9-38),

$$f \equiv f^* \quad \text{if } f(x,y) = 0 \text{ for all } (x,y)$$

and (9-39)

$$f \equiv f_{1,2,\ldots,2^{2\ell}} \quad \text{if } f(x,y) = 1 \text{ for all } (x,y)$$

We may now construct the distribution P(f) as follows:

Let

$$P(f^*) = q(0)_{(xy)_1}$$

$$P(f_1^*) = q(0)_{(xy)_2} - q(0)_{(xy)_1}$$

$$P(f_{1,2}^*) = q(0)_{(xy)_3} - q(0)_{(xy)_2}$$

$$\vdots$$

$$P(f^*_{1,2,\ldots,2^{2\ell}-1}) = q(0)_{(xy)_{2^{2\ell}}} - q(0)_{(xy)_{2^{2\ell}-1}} \tag{9-40}$$

$$P(f^*_{1,2,\ldots,2^{2\ell}}) = 1 - q(0)_{(xy)_{2^{2\ell}}}$$

$$P(f) = 0 \quad \forall f \notin (f^*, f_1^*, f^*_{1,2}, \ldots, f^*_{1,2,\ldots,2^{2\ell}}) .$$

It is clear from the above and from the indexing convention (9-37) that $P(f) \geq 0, \forall f$ and also $\sum_f P(f) = 1$. Moreover

$$\sum_{f \ni f((xy)_k)=0} P(f) = P(f^*)+P(f_1^*)+\ldots+ P(f^*_{1,2,\ldots,k-1}) =$$

$$= q(0)_{(xy)_1} +[q(0)_{(xy)_2} - q(0)_{(xy)_1}]+\ldots+ [q(0)_{(xy)_k} - q(0)_{(xy)_{k-1}}] = \tag{9-41}$$

$$= q(0)_{(xy)_k} \quad \text{for all } k \in (1,2,\ldots, 2^{2\ell}),$$

so that the assignment (9-40) satisfies the constraints of equations (7-11) for the signal probability distribution $\{q_{xy}(0)\}$. We have thus displayed a procedure for finding distribution P(f) for any constraint (7-11). Such distribution is in a sense a cannonical one, since in general it satisfies (7-11) with a minimal number of symbols which are to have non-zero probabilities. As seen from (9-40) the maximum number of non-zero probabilities necessary for this assignment method is $2^{2\ell}+1$.

Our interest in P(f) and $\bar{P}(\bar{f})$ assignments satisfying the constraints of signal distributions $\{q_{xy}(x)\}, \{\bar{q}_{\bar{x}\bar{y}}(\bar{x})\}$ arose in the preceding section from the consideration of a practical way to optimize expressions (9-26). We were there suggesting that, for distribution $\{q_{xy}(x)\}, \{\bar{q}_{\bar{x}\bar{y}}(\bar{x})\}$ maximizing the expression (9-27), a distribution P(f) and $\bar{P}(\bar{f})$ should be found which would minimize (9-28). The author has not succeeded in proving that the assignment (9-40) is the one which would accomplish that, but an argument can be presented showing that it is a good assignment in some sense:

It follows from Section 7.5.2 that a code giving a minimal coding loss will in general contain such words F whose range F(Y)=X is small over those Y's which occur with overwhelming probability. We would thus wish to include into the code, all other things being equal (and in the next section we will discuss a condition under which they are not), only those symbols f which with overwhelming probability would generate one of the binary digits, and only with a very small probability its complement. Roughly speaking then, the forward codewords ought to consist of those f's for which the number of different configurations $(\mathbf{x},\mathbf{y})$ such that $f(\mathbf{x},\mathbf{y})=x$ would very much exceed the number of different $(\mathbf{x},\mathbf{y})$ such that $f(\mathbf{x},\mathbf{y})=x^c$. We would thus like to give the greatest possible probability weight to those f's exhibiting the above imbalance the most. But inspection of (9-40) convinces us that such is exactly the result of that assignment. Namely, of all possible assignments satisfying the constraints (7-11) for a given distribution $\{q_{\mathbf{x}\mathbf{y}}(x)\}$, $\{\bar{q}_{\mathbf{x}\mathbf{y}}(\bar{x})\}$ the assignment (9-40) gives the greatest possible weight to the symbols f^* and $f^*_{1,2,\ldots,2^{2\ell}}$, and given the previous, it gives the greatest possible weight to f^*_1 and $f^*_{1,2,\ldots,2^{2\ell}-1}$, etc. etc. Hence, in this sense, at least, procedure (9-40) gives the optimal results.

9.5 Further Note on Quasi-optimization of P(f) and $\bar{P}(\bar{f})$ Distributions.

It is reasonable to expect that the optimal distribution P(f) and $\bar{P}(\bar{f})$ maximizing the expression (9-26) will depend on the length m of the decoding criterion defined in (6-10a). This can be seen by consideration of the results of Appendices III and IV. Surely, if the probability $Pr(\bar{y}/f,\bar{f},\ell^m,\bar{\ell}^m,\bar{\zeta}^m,\bar{s}^{-m})$ cannot be said for a sufficiently large m to converge to the probability $Pr(\bar{y}/f,\bar{f},\ell^m,\bar{\ell}^m,\bar{\zeta}^m, f^{-m-1},\bar{f}^{-m-1},\bar{y}^{-m-1},\bar{s}^{-m-1})$ uniformly for all sequences $(\bar{s}^{-m-1},\bar{y}^{-m-1},..,\bar{y},\bar{f}^{-m-1},..\bar{f},f^{-m-1},..,f)$ for which $\bar{s}^{-m} = \bar{g}(\bar{s}^{-m-1},\bar{f}^{-m-1},\bar{y}^{-m-1})$ and $Pr(\bar{y},..,\bar{y}^{-m-1},\bar{f},..,\bar{f}^{-m-1},f,..,f^{-m-1},\bar{s}^{-m-1}) \neq 0$, then it follows that the criterion is not a very accurate one, and that the probabilities of f's and $\bar{f}$'s ought to be weighted so as to make the inaccuracy least damaging. As m increases, such inaccuracy must have a smaller and smaller effect on the resulting probability of a decoding error, and it must hence be taken into consideration less and less when the distributions P(f) and $\bar{P}(\bar{f})$ are being determined.

Now, as shown in Section 6-2 and in particular by equation (6-7) the sequences $(y,..,y^{-m},\bar{y},..,\bar{y}^{-m},f,..,f^{-m},\bar{f},..,\bar{f}^{-m})$ can be considered as having been generated by a Markoff source. Thus the treatment of Appendices III and IV applies to the construction of the probabilities $Pr(y/f,\bar{f},\ell^m,\bar{\ell}^m,\zeta^m,\bar{s}^{-m})$, the Markoff source in question belonging to the class treated in Appendix IV. Set-transition matrices $\mathcal{B}_{\alpha}(\beta)$ can then be constructed, but it can be shown that unless the set of f's and $\bar{f}$'s for which $P(f) \neq 0$ & $\bar{P}(\bar{f}) \neq 0$ is restricted to those for which $f(\mathfrak{x},\mathfrak{z}) = f(\mathfrak{x})$ and $\bar{f}(\bar{\mathfrak{x}}\,\bar{\mathfrak{z}}) = \bar{f}(\bar{\mathfrak{x}})$, the condition of Theorems III-1 or IV-2 on the set-transition matrices guaranteeing the convergence of $Pr(\bar{y}/f,\bar{f},\mathfrak{f}^{m},\bar{\mathfrak{f}}^{m},\bar{\mathfrak{z}}^{m},\bar{s}^{-m})$ cannot be satisfied. Hence in general, the latter probabilities will not converge.

An example of how such non-convergence occurs will be given. It should be pointed out that the convergence required must be of the type of equations (IV-39) of (IV-55), so that reasoning analogous to that involving equation (6-31) can be employed. The latter is necessary if the conclusion is to be made that the magnitude of m does not have a large influence on the optimizing distribution $P(f)$ and $\bar{P}(\bar{f})$. Now note that

$$Pr(\bar{y}/f,\bar{f},\mathfrak{f}^{m},\bar{\mathfrak{f}}^{m},\bar{\mathfrak{z}}^{m},\bar{s}^{-m}) = \bar{p}(\bar{y}/0\bar{x})\ Pr(0/f,\mathfrak{f}^{m},\bar{\mathfrak{x}}^{n},\bar{\mathfrak{z}}^{m},\bar{s}^{-m}) +$$
$$+\ \bar{p}(\bar{y}/1\bar{x})\ Pr(1/f,\mathfrak{f}^{m},\bar{\mathfrak{x}}^{n},\bar{\mathfrak{z}}^{m},\bar{s}^{-m}) \qquad (9\text{-}42)$$

where $\bar{x},\bar{\mathfrak{x}}^{m}$ is determined from $\bar{s}^{-m},\bar{\mathfrak{z}}^{m},\bar{\mathfrak{f}}^{m},\bar{f}$

Hence if it should turn out that for some particular identity of f^{-m-1}, e.g., $Pr(0/f,\mathfrak{f}^{m},\bar{\mathfrak{x}}^{m},\bar{\mathfrak{z}}^{m},\bar{s}^{-m}) = 0$, while for other f^{-m-1} the last probability is non-zero, then the probability (9-42) would not converge as required. This can certainly happen. For suppose that

all f^{i}, $i\varepsilon(0,-1,\ldots,-m)$ are such that

$$f^{i}(\mathfrak{x}\mathfrak{z}) = \begin{cases} 0 & \text{if } (\mathfrak{x},\mathfrak{z}) = 0 \\ 1 & \text{otherwise} \end{cases} \qquad (9\text{-}43)$$

Then if f^{-m-1} were such that $f^{-m-1}(\mathfrak{x}\mathfrak{z}) = 1$ for all $\mathfrak{x},\mathfrak{z}$, the probability $Pr(0/f,\mathfrak{f}^m,\mathfrak{x}^m,\mathfrak{z}^m,s^{-m})$ would indeed be zero. Thus we have exhibited a non-converging example. Of course, much less drastic examples are possible, but the whole problem is best approached from construction of set-transition matrices as mentioned earlier. We conclude therefore, that further safeguards on the quasi-optimization process might prove necessary. We will list two practical possibilities, the second more complicated than the first, but also a more nearly optimal one.

9.5.1 Quasi-optimization: Practical Procedure One

(i) Pick the dependency length ℓ (see section 4.3).

(ii) Pick the decoding criterion length m (see equation (6-10a) and compare it with (6-38)).

(iii) Maximize over all the possible signal probability distributions $\{q(x)\}_{\mathfrak{x}\mathfrak{z}}$, $\{\bar{q}(\bar{x})\}_{\bar{\mathfrak{x}}\bar{\mathfrak{z}}}$ the expression

$$E\left\{\log \frac{\bar{p}(\bar{y}/x\bar{x})}{Pr(\bar{y}/\bar{x},\bar{\mathfrak{x}}^m,\bar{\mathfrak{z}}^m,\bar{s}^{-m})}\right\} + \lambda\, E\left\{\log \frac{p(y/x\bar{x})}{Pr(y/x,\mathfrak{x}^m,\mathfrak{z}^m,s^{-m})}\right\} \tag{9-44}$$

where λ is a positive real number controlling the intended relation between forward and reverse transmission rates (see (6-50) and (6-51).

(iv) Using the distributions $\{q(x)\}_{\mathfrak{x}\mathfrak{z}}$ $\{\bar{q}(\bar{x})\}_{\bar{\mathfrak{x}}\bar{\mathfrak{z}}}$ maximizing (9-44) as a set of constraints, determine by the procedure of Section 9.4 the cannonical sets $\{f\} = S$ and $\{\bar{f}\} = \bar{S}$.

(v) Arbitrarily let

$P(f) = 0$ for all $f \notin S$ $\bar{P}(\bar{f}) = 0$ for all $\bar{f} \notin \bar{S}$ (9-45)

and determine the remainder of the distribution $P(f)$ $\bar{P}(\bar{f})$ by maximizing the expression

$$E\left\{\log \frac{Pr(\bar{y}/f,\bar{f},\rho^m,\bar{\rho}^m,\bar{\zeta}^m,\bar{s}^{-m})}{Pr(\bar{y}/\bar{f},\bar{\rho}^m,\bar{\zeta}^m,\bar{s}^{-m})}\right\} +$$

$$+\lambda\ E\left\{\log \frac{Pr(y/f,\bar{f},\rho^m,\bar{\rho}^m,\zeta^m,s^{-m})}{Pr(y/f,\rho^m,\zeta^m,s^{-m})}\right\} \qquad (9\text{-}46)$$

over the input symbol probabilities left undetermined by (9-45).

(vi) Using the optimized distribution $\{P(f)\}$, $\{\bar{P}(\bar{f})\}$ found in steps (v) add (iv) evaluate the positive constants ε & $\bar{\varepsilon}$ for equation (6-52) by the use of the procedure indicated in (IV-47), involving (IV-43) and (IV-45). Compute the expressions $\log(1+\varepsilon)$ and $\log(1+\bar{\varepsilon})$ and decide whether the criterion length m picked in (ii) was a sufficient one (see (6-54) and (6-55)). If not, increase m and go through the entire procedure again, starting with step (iii).

9.5.2 Quasi-optimization: Practical Procedure Two

Before we describe this method step by step, we wish to return to Section 9.4 and point out that by the rule (9-37) and (9-38) there exist altogether $2^{2\ell}!$ different possible canonical sets $\{f^*\}$, each associated with a different ordering by magnitude of the members of the signal probability set $\{q_{\mathbf{x}\zeta}(x)\}$. Hence, if for any of the $2^{2\ell}!$ canonical sets the probabilities of symbols f not in the set are made equal to zero, and the remaining P(f) left unspecified, any set of signal probabilities $\{q_{\mathbf{x}\zeta}(x)\}$ whose ordering is that which defined the picked canonical set $\{f^*\}$, can be obtained. Hence, when optimizing under the restriction that only one canonical set of f's can end up with non-zero probabilities, we proceed as follows:

(i) Pick a dependency length ℓ (see section 4.3)

(ii) Pick the decoding criterion length m (see equation (6-10a) and compare it with (6-38)).

(iii) There are $(2^{2\ell}!)^2$ possible different combinations of canonical sets $\{f^*\}$ and $\{\bar{f}^*\}$. For all of them in turn maximize the expression

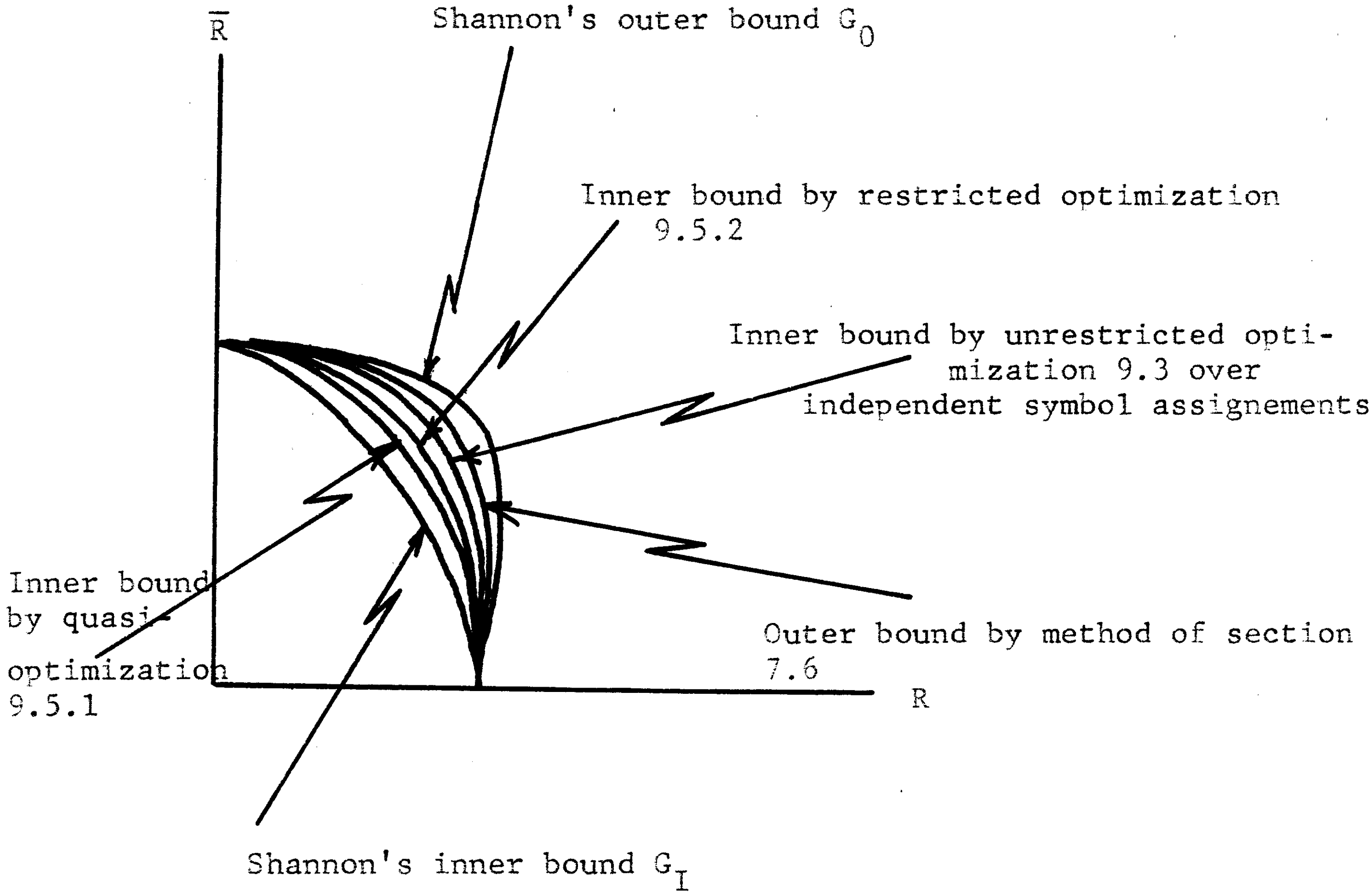
$$E\left\{\log \frac{\Pr(\bar{y}/f,\bar{f},\ell^m,\bar{\ell}^m,\bar{\zeta}^m,\bar{s}^{-m})}{\Pr(\bar{y}/\bar{f},\bar{\ell}^m,\bar{\zeta}^m,\bar{s}^{-m})}\right\} + $$

$$+ \quad E\left\{\log \frac{\Pr(y/f,\bar{f},\ell^m,\bar{\ell}^m,\zeta^m,s^{-m})}{\Pr(y/f,\ell^m,\zeta^m,s^{-m})}\right\} \qquad (9\text{-}47)$$

over the probabilities P(f), $f\epsilon\{f^*\}$, and $\bar{P}(\bar{f})$, $\bar{f}\epsilon\{\bar{f}^*\}$, where arbitrarily P(f)=0 for $f \notin \{f^*\}$ and $\bar{P}(\bar{f})=0$ for $\bar{f} \notin \{\bar{f}^*\}$. In the expression (9-47), λ is a positive real number controlling the intended relation between forward and reverse transmission rates. (See (6-50) and (6-51)). The maxima of (9-47) obtained for each canonical set combination should be recorded and stored.

(iv) Compare the maxima obtained in (iii) and let the greatest of them correspond to the cannonical probability set $\{P^*(f^*)\}$, $\{\bar{P}^*(\bar{f}^*)\}$. For it evaluate the positive constants ε and $\bar{\varepsilon}$ for equation (6-52) by use of the procedure indicated in (IV-47) involving (IV-43) and (IV-45). Compute the expressions $\log(1+\varepsilon)$ and $\log(1+\bar{\varepsilon})$ and decide whether the criterion length m picked in (ii) was a sufficient one (see (6-54) and (6-55)). If not, increase m and go through the entire procedure again, starting with step (iii).

It is clear that all points obtained by the procedure 9.5.1 are included in the convex hull of the points obtained by 9.5.2. The latter in turn are contained in the convex hull of points obtained by the optimization for independent symbol coding of Section 9.3. Thus it might be interesting to draw for a given two-way channel a series of nested capacity regions each corresponding to a different complexity of optimization. One might then obtain a graph having the general shape displayed by Figure 9-1.



$\overline{R}$
Shannon's outer bound G_0
Inner bound by restricted optimization 9.5.2
Inner bound by unrestricted optimization 9.3 over independent symbol assignements
Inner bound by quasi-optimization 9.5.1
Outer bound by method of section 7.6
R
Shannon's inner bound G_I

Figure 9-1

APPENDIX I

It is clear that an effective code will consist of a set of words which are the greatest possible "distance" apart, consistent with the relative signal frequency which would match the channel. A random code which accomplishes this on the average is one whose words $x_1, x_2, \ldots, x_n$ are picked with the probability $P(x_1)P(x_2)\ldots.P(x_n)$, where the probability distribution P( ) is one matching the channel. Now in a two-way channel code the above probability relationship should hold for codewords for both terminals, and hence we would wish that also reverse terminal codewords $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ be picked with probability $\bar{P}(\bar{x}_1)\bar{P}(\bar{x}_2)\ldots\bar{P}(\bar{x}_n)$. One may then ask when picking the word pairs simultaneously, what freedom remains which could be utilized to insure that when the random code is in operation, the signals x and $\bar{x}$ be simultaneously transmitted through the channel with probability $Pr(x,\bar{x})$.

The only way in which one can attempt this is to proceed as outlined below. Suppose that a code consisting of k forward and reverse words is to be picked. If the length of codewords is to be n then one can arrange these in two arrays of k rows and n columns and pick the column pairs as follows: The forward signal x_i $(i=1,2,\ldots,k)$ in the i^{th} row is selected with probability $P(x_i)$, and the reverse signal with probability $Q(\bar{x}_i/x_i)$. If this is done then the probability that in the i^{th} row of the reverse code the signal $\bar{x}_i$ was picked is $\bar{P}(\bar{x}_i) = \sum_{x_i} Q(\bar{x}_i/x_i)P(x_i)$ so that the randomness desired, as described in the first paragraph of this appendix, is achieved. One will then ask what joint probabilities $Pr(x,\bar{x})$ of simultaneous signal transmission can be achieved by judicious choice of the probabilities $Q(\bar{x}_i/x_i)$. Assuming the messages to be equiprobable, we get:

$$Pr(x,\bar{x}) = \sum_{i,j} Pr(\text{messages } m=i,\ \bar{m}=j \text{ were transmitted}).$$

$$\bullet Pr(x_i=x,\bar{x}_j=\bar{x}) = \frac{1}{k^2} \sum_{i,j} Pr(x_i=x,\bar{x}_j=\bar{x}) =$$

$$\frac{1}{k} Q(\bar{x}/x)P(x) + \frac{k-1}{k} \bar{P}(\bar{x})P(x)$$

From the above it follows that regardless of the probabilities $Q(\bar{x}/x)$, as the number of messages $k \to \infty$, the probability $Pr(x,x) \to P(x)\bar{P}(\bar{x})$.

APPENDIX II

Distribution of a cumulative sum of random variables, produced by a Markoff Process.

The method for bounding of random variable distributions discussed here was developed by C. E. Shannon in an unpublished paper portions of which will be paraphrased in what follows. As will be seen it is a generalization of the Chernoff bound to probability dependences derivable from a Markoff process.

Suppose that a Markoff source is given having states $i \epsilon (1,2,\ldots\ldots,g)$, and that there are at most d possible paths from any state i to any state j. Let $\mathcal{P}_i(\alpha,j)$ be the probability of a transition from state i to state j via path $\alpha \epsilon (1,2,\ldots,d)$. With each transition let there be associated a real number $V_{\alpha ij}$. A cumulative sum of these associated numbers is being produced by the process, its value being S_n after n steps. If the process then is in state i and a transition into state j via path α occurs, the cumulative sum becomes:

$$S_{n+1} = S_n + V_{\alpha ij} \tag{II-1}$$

If the system is started in a given state (or in any of the states with certain prescribed probabilities p_j) and operated for n steps, the cumulative sum S_n is a random variable. We wish to investigate the distribution of this random variable, particularly its asymptotic properties for large n.

Let

$$G_{ij}(n,x) = Pr\left\{S_n \leq x;\ \text{final state } j/\text{initial state } i\right\} \qquad \text{(II-2)}$$

i.e. $G_{i,j}(n,x)$ is a distribution function.

Let the corresponding moment generating function be

$$\varphi_{ij}(n,s) = \int_{-\infty}^{\infty} e^{sx}\, d\, G_{ij}(n,x) \qquad \text{(II-3)}$$

We now develop a difference equation for the distribution function and translate it into a relation involving the moment generating functions. By considering the different ways of starting in state i and ending in state j with a sum $G_{ij}(n+1,x)$ after n+1 steps, we have:

$$G_{ij}(n+1,x) = \sum_k \sum_\alpha P_k(\alpha j)\ G_{ik}(n,x-V_{\alpha kj}) \qquad \text{(II-4)}$$

This may also be written in terms of functions $\psi_{kj}(x)$ defined as follows:

$$\psi_{kj}(x) \equiv \sum_{\alpha \ni V_{\alpha,k,j} \leq x} P_k(\alpha,j) \qquad \text{(II-5)}$$

Thus $\psi_{kj}(x)$ is like the distribution function for the change in the cumulative sum when in state k but only for the transition into state j. Then equation (II-4) may be written:

$$G_{ij}(n+1,x) = \sum_k \int G_{ik}(n,x-y)\ d\,\psi_{kj}(y) \qquad \text{(II-6)}$$

The convolution involved above may be translated into a product of the moment generating function for G and the analogue of a moment generating function for ψ_{kj}. The latter is:

$$\beta_{kj}(s) = \int_{-\infty}^{\infty} e^{xs}\, d\Psi_{kj}(x) = \sum_{\alpha} e^{V_{\alpha kj} s}\, P_k(\alpha,j) \quad \text{(II-7)}$$

Taking moment generating functions of each side, equation (II-6) becomes:

$$\varphi_{ij}(n+1,s) = \sum_{k} \varphi_{ik}(n,s)\, \beta_{kj}(s) \quad \text{(II-8)}$$

Thinking of i and j as fixed and the second subscript of φ as ranging over a set of coordinates, this gives a linear difference relationship for generating the n+1 vector from the n^{th} vector. The process for fixed i and s is similar to that familiar in ordinary Markoff analysis, but the $[\beta_{kj}(s)]$ is not, for general s, a stochastic matrix. The $\beta_{kj}(s)$'s are non-negative, but their sum on j is not necessarily 1. However, for s=0 the sum is 1 and, indeed, the equation then becomes the ordinary Markoff equation for the process where $\varphi_{ij}(n,0)$ is identified as the probability of state j after n steps of the process.

Note also that if $\beta_{kj}(0) \neq 0$ (i.e. there is some positive $P_k(\alpha,j)$), then for all s, $\beta_{kj}(s) \neq 0$. This implies that properties of the matrix $[\beta_{kj}(s)]$ which depend only on the vanishing or non-vanishing of elements will be true for every s if they are true for s=0. This includes important properties familiar in the theory of finite state Markoff processes relating to the subdivision of states into transient, recurrent and ergodic subsets. Also, the greatest common divisor of the length of closed paths through a group of states, a quantity of importance in asymptotic behavior, will not depend on s.

Using Frobenius theory, asymptotic behavior of high powers of the matrix $[\beta_{ij}(s)]$ can be evaluated. It is possible to give a reasonably explicit formula for the general term $\beta_{ij}^{(n)}(s)$ of the matrix $[\beta_{ij}(s)]^n$.

Let the characteristic equation

$$\left| \beta_{ij}(s) - \nu(s)\ \delta_{ij} \right| = 0 \qquad \text{(II-9)}$$

have roots $\nu_1(s), \nu_2(s), \ldots, \nu_g(s)$. We assume, for simplicity, that the roots are distinct, but slight alterations of what follows give the general case.

Let the characteristic vector for the root $\nu_t(s)$ be $A_i^{(t)}(s)$, and for the transposed matrix let the characteristic vector be $B_i^{(t)}(s)$. Thus:

$$\sum_i A_i^{(t)} \beta_{ij}(s) = \nu_t(s)\ A_j^{(t)}(s) ; \sum_j B_j^{(t)}(s) \beta_{ij}(s) = \nu_t(s) B_i^{(t)}(s) \qquad \text{(II-10)}$$

The $A_i^{(t)}(s)$ and $B_i^{(t)}(s)$ vectors are orthogonal if $\nu_t(s) \neq \nu_r(s)$ and the inner product of these vectors can be normalized to 1 when r=t. When the normalizing is done, the general element in the n^{th} power of the matrix $[\beta_{ij}(s)]$ is given explicitly in terms of the characteristic values and vectors as follows:

$$\beta_{ij}^{(n)}(s) = \sum_{\ell=1}^{g} A_j^{\ell}(s)\ B_i^{\ell}(s)\ \nu_{\ell}^{n}(s) \qquad \text{(II-11)}$$

In the case where the greatest common divisor of the path lengths is 1 (the case of the most interest to us), one of the eigenvalues is real and dominates all others in absolute value. If we call this $\nu_1(s)$, then $\beta_{ij}^{(n)}(s)$ is asymptotic, and

$$\lim_{n \to \infty} \beta_{ij}^{(n)}(s) = A_j^1(s)\ B_i^1(s)\ \nu_1^{n}(s) \qquad \text{(II-12)}$$

It can be seen from (II-8) and by comparison of (II-7) with (II-3) and of (II-2) with (II-5) that

$$[\varphi_{ij}(n,s)] = [\beta_{ij}(s)]^n \qquad \text{(II-13)}$$

and consequently, taking (II-12) into consideration, that

$$\lim_{n \to \infty} \varphi_{ij}(n,s) = A_j^1(s)\ B_i^1(s)\ \nu_1^{n}(s) \qquad \text{(II-14)}$$

Using the properties of the distribution function for all possible single steps of the Markoff process, we have thus obtained the moment generating function of a cumulative sum of Markoff-dependent random variables. Similarly to the independent case, knowing the moment generating function we can obtain an upper bound on the associated distribution:

$$\varphi_{i,j}(n,s) = \sum_{\forall u \text{ from } i \text{ to } j} e^{sS_{nij}(u)} P_n(i,j,u) \tag{II-15}$$

where u is the n-step transition "route" followed from i to j and on it $S_n = S_{nij}(u)$

Therefore

$$\varphi_{i,j}(n,s) > \sum_{\forall u \ni S_{nij}(u) < xn} e^{sS_{nij}(u)} P_n(i,j,u) > \tag{II-16}$$

$$> e^{snx} \sum_{\forall u \ni S_{nij}(u) < xn} P_n(i,j,u) = e^{snx} G_{ij}(n,nx)$$

if $s \leq 0$

Hence we get the result

$$G_{ij}(n,xn) < e^{-snx} \varphi_{ij}(n,s) \qquad \text{for } s \leq 0 \tag{II-17}$$

which, applying equation (II-14) can, for n sufficiently large, be written as follows:

$$G_{ij}(n;nx) < e^{-n[sx - \ell n\, \nu_1(s)]} \qquad \text{for } s \leq 0 \tag{II-18}$$

As is customary, we may optimize the above bound by defining the function

$$\gamma(s) \equiv \ell n\, \nu_1(s) \tag{II-19}$$

The optimization leads to the result

$$x = \gamma'(s) \tag{II-20}$$

and thus we get

$$G_{ij}(n;n\,\gamma'(s)) < e^{-n(s\gamma'(s) - \gamma(s))} \qquad s \leq 0 \tag{II-21}$$

In order for the bound (II-21) to be of real value to us, we must show that the exponent on the right hand side is always negative. For this purpose, as is usual, we will define a tilted probability distribution $q(i,\alpha,j,s)$ such that for it:

$$\gamma'(s) = \frac{\nu'(s)}{\nu(s)} = \sum_{i,\alpha,j} V_{i\alpha j}\; q\,(i,\alpha,j,s) = \overline{V(s)} \qquad \text{(II-22)}$$

i.e. the cut-off value of the distribution (II-21) becomes the average value of the random variable over the tilted ensemble.

We first need to investigate some properties of the characteristic values of equation (II-9). From (II-10) and the normalizing requirement we get:

$$\nu(s) = \sum_j \nu\, A_j B_j = \sum_j A_j \sum_k \beta_{jk}(s)\; B_k = \sum_{j,k} A_j B_k\; \beta_{jk}(s) \qquad \text{(II-23)}$$

where, to avoid confusion, we have dropped the superscript 1 from the dominant eigenvalue and eigenvectors.

Also from (II-10)

$$\nu'(s)\; A_j B_j = \sum_i A_i' B_j \beta_{ij} + \sum_i A_i B_j \beta'_{ij} - \sum_k A'_j B_k \beta_{jk} \qquad \text{(II-24)}$$

hence

$$\nu'(s) = \nu'(s) \sum_j A_j B_j = \sum_i A_i' \sum_j B_j \beta_{ij} + \sum_i A_i \sum_j B_j \beta'_{ij} - \sum_j A_j' \sum_k B_k \beta_{jk} \qquad \text{(II-25)}$$

and consequently

$$\nu'(s) = \sum_{i,j} A_i(s)\; B_j(s)\; \beta'_{ij}(s) \qquad \text{(II-26)}$$

Therefore we get from (II-22) , (II-26) and (II- 7):

$$\frac{1}{\nu(s)} \sum_{i,j,\alpha} A_i(s)\ B_j(s)\ V_{i\alpha j}\ e^{sV_{i\alpha j}}\ P_i(\alpha j) = \sum_{ij\alpha} V_{i\alpha j}\ q(i\alpha js) \qquad \text{(II-27)}$$

and comparing like terms we get the definition for the tilted distribution:

$$q(i,\alpha,j,s) = \frac{1}{\nu(s)} A_i\ B_j\ e^{sV_{i\alpha j}}\ P_i(\alpha,j) \qquad \text{(II-28)}$$

From (II-23) and (II- 7) it is clear that we are dealing with a well defined probability whose sum is equal to 1 as it should be.

From (II-28) it is clear that

$$\log \frac{q(i,\alpha,j,s)}{P_i(\alpha,j)\ B_j(s)\ A_i(s)} = -\ \gamma(s) + s\ V_{i\alpha j} \qquad \text{(II-29)}$$

Therefore from (II-22) we have the equation

$$s\gamma'(s)-\gamma(s) = \sum_{i\alpha j} q(i,\alpha,j,s)\ \log \frac{q(i,\alpha,j,s)}{A_i(s)B_j(s)\ P_i(\alpha,j)} =$$

$$= \sum_{i\alpha j} q(i,\alpha,j,s)\ \log \frac{q(i,\alpha,j,s)}{A_i(s)B_i(s)P_i(\alpha,j)} + + \log \frac{B_i(s)}{B_j(s)} \qquad \text{(II-30)}$$

However, from (II-28) and (II-7) we get:

$$\sum_{i,j} \ell n \frac{B_i}{B_j} \sum_{\alpha} q(i,\alpha,j,s) = \frac{1}{\nu(s)} \sum_{i,j} A_i B_j \ \beta_{ij} [\ell n \ B_i - \ell n \ B_j]$$

$$= \sum_i A_i B_i \ \ell n \ B_i - \sum_j A_j B_j \ \ell n \ B_j = 0 \qquad \text{(II-31)}$$

The last equality being obtained with the help of Eq. (II-10). But from (II-28) we can also write:

$$q(i,s) = \sum_{j,\alpha} q(i,\alpha,j,s) = \frac{A_i}{\nu(s)} \sum_j B_j \ \beta_{ij}(s) = A_i(s) \ B_i(s) \qquad \text{(II-32)}$$

So that we can finally re-write (II-30):

$$s\gamma'(s) - \gamma(s) = \sum_{i\alpha j} q(i,\alpha,j,s) \ \log \frac{q(i,\alpha,j,s)}{q(i,s)P_i(\alpha,j)} \geq 0 \qquad \forall s \qquad \text{(II-33)}$$

It should be noted that (II-33) equals 0 if and only if s=0, since then $\nu(0) = 1$, $B_j(0) = 1$ for all j, and thus $q(i,\alpha,j,s) = A_i(0)P_i(\alpha,j)$. Note that in such a case

$$\gamma'(0) = V(0) = \overline{V_{i\alpha j}}^{\text{avg}} \qquad \text{(II-34)}$$

Therefore we have proven the usefulness of the bound (II-21).

With reference to equation (II-2) and the discussion of the preceding paragraph, it should finally be noted that the general random variable distribution for n steps starting in state i will be

$$G_i(n,x) = \Pr\left\{S_n \leq x/\text{initial state } i\right\} = \sum_j G_{ij}(n,x) \qquad \text{(II-35)}$$

and if the system is started in state i with probability p_i, then the over-all distribution surely is

$$G(n,x) = \Pr\left\{ S_n \leq x \right\} = \sum_i p_i G_i(n,x) = \sum_{i,j} p_i G_{i,j}(n,x) \tag{II-36}$$

If we wish to find a bound on (II-36) we note that (II-21) is independent of i and j, so that we get the bound

$$G(n,n\gamma'(s)) \leq g\ e^{-n(s\gamma'(s)-\gamma(s))}, s \leq 0. \tag{II-37}$$

This concludes the discussion of this appendix.

APPENDIX III

Suppose a Markoff source is given with an associated matrix [M] of its transition probabilities. Let the source have k states, s_1, s_2, ... s_k, and let the value of a matrix element m_{rt} be:

$$m_{rt} = p(s_t^i / s_r^{i-1}) \quad \text{where the superscript } i \text{ refers to a time interval.} \qquad \text{(III-1)}$$

We may, if we wish, divide the k states into ℓ non-empty classes: A_1, A_2,...A_ℓ. The question we are concerned with is: what is the probability $p(A_i^\nu, A_j^{\nu-1} \ldots, A_q^{'})$ $i,j,\ldots,q\varepsilon(1,2,\ldots,\ell$ that the source will be found, at successive intervals, in states belonging to the sets i,j,...,q ?

Example III-1

To illustrate: let k=3, ℓ=2

$$A_1 = \{s_2\}$$

$$A_2 = \{s_1, s_3\}$$

Then $p(A_1^4, A_2^3, A_2^2, A_1^1) = p(s_2^4, s_1^3, s_1^2, s_2^1) + p(s_2^4, s_1^3, s_3^2, s_2^1)$

$$+ p(s_2^4, s_3^3, s_1^2, s_2^1) + p(s_2^4, s_3^3, s_3^2, s_2^1)$$

A general answer for our problem may be obtained with the help of the matrix[M].

$$p(A_i^\nu, A_j^{\nu-1} \ldots, A_q^1) = \sum_{s^\nu \varepsilon A_i} \sum_{s^{\nu-1} \varepsilon A_j} \cdots \sum_{s^1 \varepsilon A_q} p(s^\nu, s^{\nu-1}, \ldots, s^1) =$$

$$= \sum_{s^\nu \varepsilon A_i} \sum_{s^{\nu-1} \varepsilon A_j} \cdots \sum_{s^1 \varepsilon A_q} p(s^\nu / s^{\nu-1}) \ldots p(s^2/s^1) p(s^1)$$

(III-2)

Define the k x k set-transition matrix $\mathcal{B}_i(j)$, $i,j\ \varepsilon(1,2,\ldots,\ell)$ as follows:

a) $\mathcal{B}_i(j)$ has elements b_{rt}^{ij} $r,t\ \varepsilon(1\ \ldots,\ k)$

$$\text{b)}\quad b_{rt}^{ij} = \begin{cases} m_{rt} & \text{if } s_r \varepsilon A_i \\ & \phantom{\text{if }} s_t \varepsilon A_j \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(III-3)}$$

Define the vector P(i)] in R^k space as follows:

a) P(i)] has elements p_r^i, $r\varepsilon(1,2,\ldots,k)$

$$\text{b)}\quad p_r^i = \begin{cases} p(s_r) & \text{if } s_r \varepsilon A_i \\ 0 & \text{otherwise} \end{cases} \qquad \text{(III-4)}$$

where $p(s_r)$ is the stationary probability of the state s_r of the Markoff source.

Define the vector 1] in R^k space:

all elements of 1] are equal to the integer 1. (III-5)

<u>Lemma 1</u>:

$$p(A_i^{\nu},\ A_j^{\nu-1} \ldots A_w^2,\ A_q^1) = \underline{P(q)} \times \mathcal{B}_q(w) \ldots \mathcal{B}_j(i) \times 1] \qquad \text{(III-6)}$$

Proof:

Let $\mathcal{B}_q(w) \ldots \mathcal{B}_j(i) = \mathcal{D}$

$$\text{then}\quad d_{\nu\sigma} = \sum_{\lambda=1}^{k} \cdots\cdots \sum_{\mu=1}^{k} b_{\nu\lambda}^{qw} \ldots b_{\mu\sigma}^{ji} \qquad \text{(III-7)}$$

But for (III-3) we can write:

$$d_{\nu\sigma} = \sum_{\lambda \ni s_\lambda \varepsilon A_w} \cdots\cdots \sum_{\mu \ni s_\mu \varepsilon A_j} b_{\nu\lambda}^{qw} \ldots b_{\mu\sigma}^{ji} \qquad \text{(III-8)}$$

Therefore we have from (III-3) and (III-4)

$$\underset{\smile}{P(q)}\,[\mathcal{D}]\,1] = \sum_{\nu}\sum_{\sigma} p_{\nu}^{q}\;d_{\nu\sigma}$$

$$= \sum_{\nu \to s_{\nu} \varepsilon A_{q}} \; \sum_{\sigma \to s_{\sigma} \varepsilon A_{i}} p_{\nu}^{q}\, d_{\nu\sigma} \qquad \text{(III-9)}$$

and we see, after inspection of (III-2,-8,-9), that

$$\underset{\smile}{P(q)}\,[\mathcal{D}]\,1] = \sum_{s^{1}\varepsilon A_{q}} \; \sum_{s^{2}\varepsilon A_{w}} \cdots \sum_{s^{\nu-1}\varepsilon A_{j}} \; \sum_{s^{\nu}\varepsilon A_{i}} p(s^{1})p(s^{2}/s^{1})\ldots$$

$$\ldots p(s^{\nu}/s^{\nu-1}) \quad \text{Q.E.D.} \qquad \text{(III-10)}$$

We therefore see that given the r x r matrix [M] and any state partition A_1, A_2, ... A_ℓ, one can, by the use of matrices $\mathcal{B}_i(j)$ defined in (III-3) and vectors $\underset{\smile}{P(i)}$ defined in (III-4) easily derive a notationally compact expression (III-6) for the probability of successive states being members of specified state-sets.

It should be noted that:

$$P(i)\;[\mathcal{B}_j(k)] = 0 \quad \text{if } i \neq j$$

$$[\mathcal{B}_q(r)][\mathcal{B}_j(k)] = [0] \text{ if } r \neq j \qquad \text{(III-11)}$$

Therefore, as a consequence of (III-11), a non-zero answer will be obtained only if the product of matrices $\mathcal{B}_\nu(\lambda)$ premultiplied by a vector $\underset{\smile}{P(i)}$ and postmultiplied by 1] actually makes sense.

We may, however, drop the above mentioned safety precaution and contract further the square k x k matrices $\mathcal{B}_i(j)$ into, in general, rectangular ν x λ matrices $\overline{\mathcal{B}}_i(j)$

where ν is the no. of states in set A_i

and λ is the no. of states in set A_j.

Equally, we may contract the k-dimensional vector $\underline{P(i)}$ into the ν-dimensional vector $\underline{\bar{P}(i)}$, if the set A_i has ν states.

The contraction from $\mathcal{B}_i(j)$, defined in (III-3), to $\bar{\mathcal{B}}_i(j)$ is done by eliminating from $\mathcal{B}_i(j)$ all the rows r such that the state $s_r \notin A_i$ and all the columns t such that the state $s_t \notin A_j$.

Similarly, the contraction from $\underline{P(i)}$, defined in (III-4) to $\underline{\bar{P}(i)}$ is accomplished by eliminating from $\underline{P(i)}$ all the coordinates with indexes r such that the state $s_r \notin A_i$. In both of the above elimination operations the order of rows and columns is to remain unchanged. (III-3b)

On reflection it is obvious that although the precaution (III-11) was dropped, it is still true that

$$p(A_i^{\nu}, A_j^{\nu-1}, \ldots, A_{\omega}^{2}, A_q^{1}) = \underline{P(q)}\,\bar{\mathcal{B}}_q(\omega) \ldots \bar{\mathcal{B}}_j(i)\ 1] \qquad \text{(III-12)}$$

since $\bar{\mathcal{B}}_j(i)\ \bar{\mathcal{B}}_i(q)$ are compatible matrices, being of dimension $\nu \times \lambda$ and $\lambda \times \sigma$, respectively.

<u>Example III-2</u>

An example will be in order:

As in (9), let again: $k = 3$, $\ell = 2$

$$A_1 = \{s_2\}\ ;\quad A_2 = \{s_1, s_3\}$$

Then we have:

$$\underline{P(1)} = \underline{0,\ p(s_2),\ 0} \qquad \underline{1} = \underline{1,\ 1,\ 1}$$

$$\underline{P(2)} = \underline{p(s_1),\ 0,\ p(s_3)}$$

$$\mathcal{B}_1(1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & m_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathcal{B}_1(2) = \begin{bmatrix} 0 & 0 & 0 \\ m_{21} & 0 & m_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

$$p(A_1^4 A_2^3 A_2^2 A_1^1) = \underline{P(1)}\ \mathcal{B}_1(2)\mathcal{B}_2(2)\mathcal{B}_2(1)\ 1]$$

$$= \underline{0, p(s_2), 0} \begin{bmatrix} 0 & 0 & 0 \\ m_{21}m_{11}+m_{23}m_{31} & 0 & m_{21}m_{13}+m_{23}m_{33} \\ 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} m_{12} \\ 0 \\ m_{32} \end{bmatrix}$$

$$= p(s_2)\,[(m_{21}m_{11}+m_{23}m_{31})m_{12} + (m_{21}m_{13}+m_{23}m_{33})m_{32}]$$

$$= p(2,1,1,2) + p(2,\ 1,3,2) + p(2,3,1,2) + p(2,3,3,2)$$

The contracted matrices and vectors are:

$$\underline{\bar{P}(1)} = \underline{p(s_2)}$$

$$\underline{\bar{P}(2)} = \underline{p(s_1), p(s_3)}$$

$$\bar{\mathcal{B}}_1(1) = [m_{22}] \qquad \bar{\mathcal{B}}_1(2) = [m_{21},\ m_{23}]$$

$$\bar{\mathcal{B}}_2(1) = \begin{bmatrix} m_{12} \\ m_{32} \end{bmatrix} \qquad \bar{\mathcal{B}}_2(2) = \begin{bmatrix} m_{11} & m_{13} \\ m_{31} & m_{33} \end{bmatrix}$$

$$p(A_1\ A_2\ A_2\ A_1) = \underline{\overline{P}(1)}\ \overline{\mathcal{B}}_1(2)\ \overline{\mathcal{B}}_2(2)\ \overline{\mathcal{B}}_2(1)\ 1] =$$

$$= \underline{p(s_2)}\ [m_{21}, m_{23}] \begin{bmatrix} m_{11} & m_{13} \\ m_{31} & m_{33} \end{bmatrix} \begin{bmatrix} m_{12} \\ m_{32} \end{bmatrix} \quad 1]$$

$$= p(s_2)\ [(m_{21}m_{11}+m_{23}m_{31})m_{12}+(m_{21}m_{13}+m_{23}m_{33})m_{32}]$$

In what follows we will always be dealing with contracted vectors $\underline{\overline{P}(i)}$ and matrices $[\overline{\mathcal{B}}_i(j)]$. In order to simplify notation, we will henceforth drop the bar " —" from the above quantities - contraction will be assumed as a matter of course. Suppose the Markoff source is now enclosed inside a black box supplied with a printer which, upon the source's arriving in state s_i, prints out the set A_j to which s_i belongs. It is then interesting to ask for the probability of a print-out of A^i given the record of all sets A^{i-1}, A^{i-2},... reached by the source in the past.

It is clear that unless the transition matrix [M] is degenerate in some special way, the best prediction of A^i is obtained by the use of the entire available record of past state sets.

We wish to ask under what conditions the probability of A^i given a past of λ immediately preceeding sets A^{i-1},... $A^{i-\lambda}$ will have a value within a prescribed $\varepsilon > 0$ of the probability of A^i given any longer past $A^{i-1}, \ldots, A^{i-\lambda}$, $A^{i-\lambda-1}$,...

In this regard we will first prove

Lemma 2

Given a Markoff source with states $s_1,\ldots,s_k$ partitioned into non-empty sets $A_1,\ldots,A_\ell$, $\ell < k$. For any sequence of sets $A^i, A^{i-1},\ldots,A^{i-\lambda},\ldots,A^{i-\lambda-\sigma}$ such that for it

$$\Pr(A^i,A^{i-1},\ldots,A^{i-\lambda},\ldots,A^{i-\lambda-\sigma}) > 0 \quad \text{(III-13)}$$

inequality

$$\left|\Pr(A^i/A^{i-1},\ldots,A^{i-\lambda})-\Pr(A^i/A^{i-1},\ldots,A^{i-\lambda},\ldots A^{i-\lambda-\sigma})\right|<2\varepsilon \quad \text{(III-14)}$$

will hold, provided that given any $\varepsilon>0$ an integer Λ can be found such that whenever $\lambda>\Lambda$ then for all possible sequences $A^i, A^{i-1},\ldots,A^{i-\lambda}$, $A^j\varepsilon\{A_1,\ldots,A_\ell\}$ and for all states $s_\mu^{i-\lambda}$, $s_\nu^{i-\lambda}\varepsilon A^{i-\lambda}$ for which

$$\begin{aligned}&\Pr\ (A^i,\ A^{i-1},\ldots,A^{i-\lambda+1}/s_\mu^{i-\lambda}) \neq 0\\ &\Pr\ (A^i,\ A^{i-1},\ldots,A^{i-\lambda+1}/s_\mu^{i-\lambda}) \neq 0\end{aligned} \quad \text{(III-15)}$$

the inequality

$$\left|\Pr(A^i/A^{i-1},\ldots,A^{i-\lambda+1},\ s_\mu^{i-\lambda})-\Pr(A^i/A^{i-1},\ldots,A^{i-\lambda+1},s_\nu^{i-\lambda})\right|<\varepsilon \quad \text{(III-16)}$$

holds.

Proof:

$$P(A^i/A^{i-1},\ldots A^{i-\lambda-\sigma}) = \sum_{s^{i-\lambda}} P(A^i,s^{i-\lambda}/A^{i-1},\ldots,A^{i-\lambda-\sigma}) =$$

$$= \sum_{s^{i-\lambda}} P(A^i/A^{i-1},\ldots,A^{i-\lambda+1},s^{i-\lambda},A^{i-\lambda},\ldots,A^{i-\lambda-\sigma}) \ast$$

$$\ast\ P(s^{i-\lambda}/A^{i-1}\ldots A^{i-\lambda-\sigma}) \quad \text{(III-17)}$$

But

$$P(A^i/A^{i-1},..,s^{i-\lambda},...A^{i-\lambda-\sigma}) = \frac{P(A^i, A^{i-1},..,s^{i-\lambda},..,A^{i-\lambda-\sigma})}{P(A^{i-1},...s^{i-\lambda},..,A^{i-\lambda-\sigma})} =$$

$$= \frac{P(A^i...A^{i-\lambda+1}/s^{i-\lambda})P(s^{i-\lambda},A^{i-\lambda},..,A^{i-\lambda-\sigma})}{P(A^{i-1},..,A^{i-\lambda+1}/s^{i-\lambda})P(s^{i-\lambda},A^{i-\lambda},...A^{i-\lambda-\sigma})} \quad \text{(III-18)}$$

Therefore:

$$P(A^i/A^{i-1},..,s^{i-\lambda}..,A^{i-\lambda-\sigma}) = \begin{cases} P(A^i/A^{i-1},..,A^{i-\lambda+1}s^{i-\lambda}) & \text{if } s^{i-\lambda} \varepsilon A^{i-\lambda} \\ 0 & \text{if } s^{i-\lambda} \notin A^{i-\lambda} \end{cases} \quad \text{(III-19)}$$

We have now from (III-17)

$$P(A^i/A^{i-1}...A^{i-\lambda-\sigma}) = \sum_{s^{i-\lambda} \varepsilon A^{i-\lambda}} P(A^i/A^{i-1},..,A^{i-\lambda+1},s^{i-\lambda}) \times$$

$$\times P(s^{i-\lambda}/A^{i-1}...A^{i-\lambda-\sigma}) \quad \text{(III-20)}$$

Now if (III-16) is satisfied, then

$$P(A^i/A^{i-1},..,A^{i-\lambda+1}s_\mu^{i-\lambda}) - \varepsilon \leq P(A^i/A^{i-1}...A^{i-\lambda-\sigma}) \leq$$

$$\leq P(A^i/A^{i-1},..,A^{i-\lambda+1}, s^{i-\lambda}) + \varepsilon; \; s_\mu^{i-\lambda} \varepsilon A^{i-\lambda} \quad \text{(III-21)}$$

But from (III-20) also

$$P(A^i, A^{i-1},..,A^{i-\lambda+1}, s_\mu^{i-\lambda}) - \varepsilon \leq P(A^i/A^{i-1}...A^{i-\lambda}) \leq$$

$$P(A^i/A^{i-1},...A^{i-\lambda+1}, s_\mu^{i-\lambda}) + \varepsilon \; ; \; s_\mu^{i-\lambda} \varepsilon A^{i-\lambda} \quad \text{(III-22)}$$

and therefore certainly

$$|P(A^i/A^{i-1},..,A^{i-\lambda}) - P(A^i/A^{i-1},..,A^{i-\lambda-\sigma})| < 2\varepsilon$$

Q.E.D.

We would next like to know what circumstances would guarantee the satisfaction of the sufficient condition of the preceeding lemma. For this purpose we will introduce the notion of directional distance $D(\gamma,\delta)$* between two vectors γ & δ in R^m space.

Consider the vectors

$$\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_m]; \quad \delta = [\delta_1, \ldots, \delta_m]$$

having real non-negative elements only.

The ratios γ_j/δ_j of like-numbered coordinates of the two vectors γ, δ can be formed and arranged according to increasing value into an ordered set S from which the ratio γ_i/δ_i will be excluded if and only if $\gamma_i = \delta_i = 0$. The numbers of the set S are to be left in their fractional form, so that if $\delta_i = 0 \neq \gamma_i$ the ratio γ_i/δ_i will be a member of S. If $\delta_i = \delta_j = 0$ and $\gamma_i > \gamma_j \neq 0$ then γ_i/δ_i will be of higher order than γ_j/δ_j in the set S.

$$\text{Let} \quad \frac{\gamma_c}{\delta_c} = \min_i \left[\frac{\gamma_i}{\delta_i} \in S\right] \qquad \text{(III-23)}$$

and

$$\frac{\gamma_d}{\delta_d} = \max_i \left[\frac{\gamma_i}{\delta_i} \in S\right] \qquad \text{(III-24)}$$

Then the directional distance between vector γ and vector δ will be defined as follows:

---

* The measure $D(\gamma,\delta)$ was suggested to the author by Dr. H. Furstenberg of the Mathematics Department of the University of Minnesota.

$$D(\gamma,\delta) \equiv \frac{\gamma_c}{\delta_c} \cdot \frac{\delta_d}{\gamma_d} \qquad \text{(III-25)}$$

We are now ready to state

Lemma 3

Given a Markoff source having a transition probability matrix $[M]$, with states $s_1,\ldots,s_k$ partitioned into non-empty sets $A_1, A_2,\ldots,A_\ell$, $\ell < k$. Let the partitioning lead to the contracted set-transition matrices $\mathcal{B}_i(j)$, $i,j \varepsilon (1,2,\ldots,\ell)$. Choose any possible sequence of sets $A^1,A^2,\ldots,A^\lambda$, $A^i \varepsilon \{A_1,A_2,\ldots,A_\ell\}$. To such a sequence there will correspond a set-transition matrix product

$$\mathcal{B}_g^1(h)\mathcal{B}_h^2(i)\ldots\mathcal{B}_r^{\lambda-1}(t) \qquad \text{(III-26)}$$

where if $A^i = A_\eta$ and $A^{i+1} = A_\xi$ then $\mathcal{B}^i = \mathcal{B}_\eta(\xi)$. Suppose two arbitrary, real, non-negative vectors $\underline{\alpha} = \underline{\alpha_1,\ldots,\alpha_\nu}$; $\underline{\beta} = \underline{\beta_1,\ldots,\beta_\nu}$ are given, where ν is the number of states in the set A^1. Then the above transition matrix product will transform the vectors $\underline{\alpha}$ and $\underline{\beta}$ into vectors $\underline{\gamma}$ and $\underline{\delta}$, respectively:

$$\underline{\gamma} = \underline{\alpha}\, \mathcal{B}^1 \mathcal{B}^2 \ldots \mathcal{B}^{\lambda-1} = \underline{\gamma_1,\ldots,\gamma_\mu}$$
$$\underline{\delta} = \underline{\beta}\, \mathcal{B}^1 \mathcal{B}^2 \ldots \mathcal{B}^{\lambda-1} = \underline{\delta_1,\ldots,\delta_\mu} \qquad \text{(III-27)}$$

where μ is the no. of states in A^λ.

Consider

<u>Proposition 1</u>:

Given any $\varepsilon_1 > 0$ there exists an integer Λ such that whenever $\lambda > \Lambda$ then any set-transition matrix product $\mathcal{B}^1\mathcal{B}^2\ldots\mathcal{B}^{\lambda-1}$ corresponding to a sequence of states $A^1, A^2, \ldots, A^{\lambda-1}, A^\lambda$ for which

$$\Pr\,(A^\lambda, A^{\lambda-1}, \ldots, A^1) \neq 0 \qquad \text{(III-28)}$$

will transform any pair of non-negative vectors $\underline{\alpha}$, $\underline{\beta}$ for which

$$\begin{aligned} \underline{\alpha}\, \mathcal{B}^1\mathcal{B}^2 \ldots \mathcal{B}^{\lambda-1} &\neq 0 \\ \underline{\beta}\, \mathcal{B}^1\mathcal{B}^2 \ldots \mathcal{B}^{\lambda-1} &\neq 0 \end{aligned} \qquad \text{(III-29)}$$

into vectors $\underline{\gamma}$ & $\underline{\delta}$ such that for them the inequality

$$|D(\gamma,\delta) - 1| < \varepsilon_1 \qquad \text{(III-30)}$$

is satisfied.

<u>Proposition 2</u>:

Given any $\varepsilon > 0$ an integer Λ can be found such that whenever $\lambda > \Lambda$ then for all possible sequences $A^1, A^2, \ldots A^\lambda$, $A^j \varepsilon \{A_1, \ldots, A_\ell\}$ and for all states s^1_μ, $s^1_\nu \varepsilon A^1$ for which

$$\begin{aligned} &\Pr\,(A^\lambda, A^{\lambda-1}, \ldots, A^2/s^1_\mu) \neq 0 \\ &\Pr\,(A^\lambda, A^{\lambda-1}, \ldots, A^2/s^1_\nu) \neq 0 \end{aligned} \qquad \text{(III-31)}$$

the inequality

$$|\Pr(A^\lambda/A^{\lambda-1}, \ldots, A^2, s^1_\mu) - \Pr(A^\lambda/A^{\lambda-1}, \ldots, A^2, s^1_\nu)| < \varepsilon$$

holds. (III-32)

Proposition 1 is a sufficient condition for Proposition 2 to hold.

Proof:

First of all we wish to remark that the arbitrary vectors $\lfloor\alpha\rfloor$, $\lfloor\beta\rfloor$ must of course conform dimensionally to the $\mathcal{B}^1$ in question, so that product (III-27) be possible. Thus for different starting sets A^1, in general different vectors $\lfloor\alpha\rfloor$, $\lfloor\beta\rfloor$ must be chosen. Equally the dimension of $\lfloor\sigma\rfloor$ will differ depending on the size of the state set A^λ.

Let state $s_\mu \varepsilon A_\emptyset$. From $\lfloor P(\emptyset)\rfloor$ construct the vector (III-33) $\lfloor C(\mu)\rfloor$ by replacing in $\lfloor P(\emptyset)\rfloor$ all the elements except $p(s_\mu)$ by zeros, and replacing the latter element by one.

Eq. if $\lfloor P(\emptyset)\rfloor = \lfloor p(s_\tau),\ p(s_\mu),\ p(s_\sigma)\rfloor$, then

$$\lfloor C(\emptyset)\rfloor = \lfloor 0 \quad , \quad 1 \quad , \quad 0 \rfloor$$

Using the reasoning employed in Lemma 1, it is clear that we may write

$$Pr(A^\lambda, A^{\lambda-1},\ldots,A^2/s^1_\mu) = \lfloor C(\mu)\rfloor \mathcal{B}^1\mathcal{B}^2 \ldots \mathcal{B}^\lambda\ 1] \qquad \text{(III-34)}$$

and therefore

$$Pr(A^\lambda/A^{\lambda-1},\ldots,A^2\ s^1_\mu) = \frac{\lfloor C(\ )\rfloor\ \mathcal{B}^1\mathcal{B}^2 \ldots \mathcal{B}^\lambda\ 1]}{\lfloor C(\ )\rfloor\ \mathcal{B}^1\mathcal{B}^2 \ldots \mathcal{B}^{\lambda-1}\ 1]} \qquad \text{(III-34)}$$

where the notation introduced in (III-26) was followed.

Let

$$G(2,3,\ldots,\lambda-1) \equiv \mathcal{B}^2\mathcal{B}^3 \ldots \mathcal{B}^{\lambda-1} \qquad \text{(III-35)}$$

Then Proposition 2 states that

$$\left| \frac{\underline{C(\mu)}\, \mathcal{B}^1\, G(2,\ldots,\lambda-1)\, \mathcal{B}^\lambda\, 1]}{\underline{C(\mu)}\, \mathcal{B}^1\, G(2,\ldots,\lambda-1)\, 1]} - \frac{\underline{C(\nu)}\, \mathcal{B}^1\, G(2,\ldots,\lambda-1)\, \mathcal{B}^\lambda\, 1]}{\underline{C(\nu)}\, \mathcal{B}^1\, G(2,\ldots,\lambda-1)\, 1]} \right| < \varepsilon \qquad \text{(III-37)}$$

for all $s_\mu,\ s_\nu\ \varepsilon\ A^1$

Next, let

$$\underline{C(\mu)}\, \mathcal{B}^1 \equiv \underline{\alpha}$$
$$\underline{C(\nu)}\, \mathcal{B}^1 \equiv \underline{\beta} \qquad \text{(III-38)}$$

Further let

$$\mathcal{B}^\lambda\, 1] = B] \qquad \text{(III-39)}$$

and

$$\underline{\alpha}\ [G] = \underline{\gamma}$$
$$\underline{\beta}\ [G] = \underline{\delta} \qquad \text{(III-40)}$$

The requirement (III-37) reduces to:

$$\left| \frac{\underline{\gamma} \cdot B]}{\underline{\gamma} \cdot 1]} - \frac{\underline{\delta} \cdot B]}{\underline{\delta} \cdot 1]} \right| < \varepsilon_1 \qquad \text{(III-41)}$$

But

$$\left| \frac{\underline{\gamma} \cdot B]}{\underline{\gamma} \cdot 1]} - \frac{\underline{\delta} \cdot B]}{\underline{\delta} \cdot 1]} \right| = \left| \frac{\sum_j B_j [\gamma_j (\sum_i \delta_i) - \delta_j (\sum_i \gamma_i)]}{(\sum_i \gamma_i)(\sum_i \delta_i)} \right| \qquad \text{(III-42)}$$

Now

a) if $\gamma_j \neq 0$; $\delta_j \neq 0$

we get

$$\frac{\gamma_j(\sum \delta_i) - \delta_j(\sum \gamma_i)}{(\sum \gamma_i)(\sum \delta_i)} = \frac{\gamma_j}{\gamma_i}\left[1 - \frac{\delta_j}{\gamma_j}\left(\frac{\sum_i \gamma_i}{\sum_i \delta_i}\right)\right] \qquad \text{(III-43)}$$

But following definitions (III-23) and (III-24) it is certainly true that (see Lemma 4)

$$\frac{\gamma_c}{\delta_c} \leq \frac{\sum \gamma_i}{\sum \delta_i} \leq \frac{\gamma_d}{\delta_d} \qquad \text{(III-44)}$$

Then certainly

$$1 - \frac{\delta_c}{\gamma_c}\frac{\gamma_d}{\delta_d} \leq 1 - \frac{\delta_j}{\gamma_j}\left(\frac{\sum \gamma_i}{\sum \delta_i}\right) \leq 1 - \frac{\gamma_c}{\delta_c}\frac{\delta_d}{\gamma_d} \qquad \text{(III-45)}$$

b) If $\gamma_j = \delta_j = 0$

then the quantity (III-43) is identically equal to 0.

Now suppose that for all sequences $A^1\ A^2 \ldots A^\lambda$ a [G] results from (III-36) such that the vectors obtained from (III-40) satisfy (III-30) [it should be noted that in such a case it cannot be that either $\gamma_j = 0 \quad \delta_j \neq 0$ or $\gamma_j \neq 0 \quad \delta_j = 0$]. Then by definition (III-45) becomes:

$$\frac{-\varepsilon_1}{1-\varepsilon_1} \leq 1 - \frac{\delta_j}{\gamma_j}\left(\frac{\sum_i \gamma_i}{\sum_i \delta_i}\right) \leq \varepsilon_1 \qquad \text{(III-46)}$$

and from (III-35, -37, -42) we get:

$$\left| P(A^{\lambda+1}/A^\lambda,..,A^1, s^o_\mu) - P(A^{\lambda+1}/A^\lambda,..,A^1\ s^o_\nu)\right| \leq$$

$$\leq \frac{1}{\sum_i \gamma_i}\ \frac{\varepsilon_1}{1-\varepsilon_1} \sum_j B_j\ \gamma_j \qquad \text{(III-47)}$$

and we can clearly choose ε_1 small enough so that the right-hand side of (III-47) becomes smaller than any given $\varepsilon > 0$ for all vectors B] obtainable from (III-39).

Q.E.D.

In order to be able to determine under what conditions statement (III-30) can be made, we will next prove a lemma about the behavior of ratios (III-23) and (III-24).

Lemma 4

Given any set of finite, non-negative, well defined ratios $x_1/y_1, x_2/y_2, \ldots x_k/y_k$, and any set of non-negative real constants $a_1, a_2, \ldots, a_k$. Let

$$\frac{x_c}{y_c} = \min_i \left(\frac{x_i}{y_i} \right)^* \quad i = 1, 2, \ldots k$$

$$\frac{x_d}{y_d} = \max_i \left(\frac{x_i}{y_i} \right)^* \quad i = 1, 2, \ldots k \qquad \text{(III-48)}$$

then

$$\frac{x_c}{y_c} \leq \frac{a_1x_1 + a_2x_2 + \ldots + a_kx_k}{a_1y_1 + a_2y_2 + \ldots + a_ky_k} \leq \frac{x_d}{y_d} \qquad \text{(III-49)}$$

equality on the left-hand side being possible if and only if

$$a_i = 0 \; ; \; \forall i \ni \frac{x_i}{y_i} \neq \frac{x_c}{y_c} \qquad \text{(III-50)}$$

equality on the right-hand side being possible if and only if

$$a_i = 0 \; ; \; \forall i \ni \frac{x_i}{y_i} \neq \frac{x_d}{y_c} \qquad \text{(III-51)}$$

---

*Note that by assumption of the lemma $y_i \neq 0$ for all $i \varepsilon (1,2,\ldots,k)$ although possibly $x_j = 0$ for some $j \varepsilon (1,2,\ldots,k)$.

Proof:

a) $$\frac{a_1x_1 +\ldots+a_kx_k}{a_1y_1 +\ldots+a_ky_k} - \frac{x_c}{y_c} = \frac{x_cy_c[a_1(\frac{x_1}{x_c} - \frac{y_1}{y_c}) + \ldots + a_c(\frac{x_c}{x_c} - \frac{y_c}{y_c})+\ldots+a_k(\frac{x_k}{x_c} - \frac{y_k}{y_c})}{(a_1y_1 +\ldots+ a_ky_k)\, y_c} \quad \text{(III-52)}$$

but $x_i/x_c - y_i/y_c \geqslant 0$ by assumption (III-48), therefore:

$$\frac{a_1x_1 +\ldots+ a_kx_k}{a_1y_1 +\ldots+ a_ky_k} \geqslant \frac{x_c}{y_c} \quad \text{(III-53)}$$

equality being possible if and only if each of the numerator terms on the right-hand side of (III-52) is equal to <u>0</u>.

b) $$\frac{x_d}{y_d} - \frac{a_1x_1+\ldots+a_kx_k}{a_1y_1+\ldots+a_ky_k} = \frac{x_dy_d[a_1(\frac{y_1}{y_d} - \frac{x_1}{x_d})+\ldots+ a_d(\frac{y_d}{y_d} - \frac{x_d}{x_d})+\ldots+a_k(\frac{y_k}{y_d} - \frac{x_k}{x_d})]}{y_d(a_1y_1 +\ldots+ a_ky_k)} \quad \text{(III-54)}$$

but $\frac{y_i}{y_d} - \frac{x_i}{x_d} \geqslant 0$ by assumption (III-48), therefore:

$$\frac{a_1x_1 +\ldots+ a_kx_k}{a_1y_1 +\ldots+ a_ky_k} \geqslant \frac{x_d}{y_d} \quad \text{(III-55)}$$

equality being possible if and only if each of the numerator terms on the right-hand side of (III-54) is equal to <u>0</u>.

Q.E.D.

It was shown in Lemma 3 that Proposition 1 constitutes a sufficient condition on the product of any λ set-transition matrices $\mathcal{B}_i(j)$ guarenteeing the satisfaction of Proposition 2. In the following Lemma 5 we will state those properties on individual matrices $\mathcal{B}_i(j)$ which insure that Proposition 1 be fulfilled.

Lemma 5

Given any finite set S of rectangular matrices $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ having the following property:

a) If $\mathcal{B} \varepsilon S$ and $\mathcal{B}$ has dimensions r x t then there exists at least one matrix $\mathcal{B}' \varepsilon S$ having dimensions t x v, and at least one matrix $\mathcal{B}'' \varepsilon S$ having dimensions q x r. (III-56)

b) If $\mathcal{B} \varepsilon S$ then all its elements are such that $b_{ij} \geq 0$ and whenever $b_{ij}=0$ then either $b_{\ell j}=0$ for all ℓ, or $b_{ih}=0$ for all h. (III-57)

Given any $\varepsilon > 0$ a positive integer Λ can be found such that for all $\lambda > \Lambda$ all possible products $\mathcal{B}^1\mathcal{B}^2\ldots\mathcal{B}^\lambda$ of λ matrices, each taken from the set S, have the following property:

Given any two non-negative vectors $\underline{\alpha}$, $\underline{\beta}$ in R^t space such that

$$\underline{\alpha}\, \mathcal{B}^1\mathcal{B}^2\ldots\mathcal{B}^\lambda \neq 0$$
$$\underline{\beta}\, \mathcal{B}^1\mathcal{B}^2\ldots\mathcal{B}^\lambda \neq 0 \,, \qquad \text{(III-58)}$$

where t is the number of rows of $\mathcal{B}^1$.

Let

$$\underline{\mathfrak{z}} = \underline{\alpha}\ \mathcal{B}^1\mathcal{B}^2 \ldots \mathcal{B}^\lambda \qquad \text{(III-59)}$$
$$\underline{\eta} = \underline{\beta}\ \mathcal{B}^1\mathcal{B}^2 \ldots \mathcal{B}^\lambda$$

then

$$|1 - D(\mathfrak{z},\eta)| < \varepsilon \qquad \text{(III-60)}$$

where $D(\mathfrak{z},\eta)$ was defined in (III-25).

<u>Proof</u>:

It should be noted that condition (III-56) makes products of matrices in S of any length possible and assumes that whenever the $i^{\underline{th}}$ matrix is $\mathcal{B}$ there exists a matrix $\mathcal{B}'$ in S which can take the $(i+1)^{\underline{th}}$ place in the matrix product.

We will first prove that given any matrix $\mathcal{B}$ of dimension r x t and any two non-zero vectors $\underline{x}$, $\underline{y}$ of dimension r such that

$$\underline{x}\,\mathcal{B} \neq 0 \quad ; \quad \underline{y}\,\mathcal{B} \neq 0 \qquad \text{(III-61)}$$

then whenever

$$0 \leq D(x,y) < 1$$

the inequality

$$0 \leq D(x,y) < D(\underline{x}\,\mathcal{B}, \underline{y}\,\mathcal{B}) \leq 1 \qquad \text{(III-62)}$$

will be satisfied provided $\mathcal{B}$ has the property (III-65). Using definitions (III-25,-26), inequalities (III-62) hold, provided

$$\frac{x_c}{y_c} < \frac{\sum_i x_i b_{ij}}{\sum_i y_i b_{ij}} < \frac{x_d}{y_d} \quad \text{for all j.} \qquad \text{(III-63)}$$

where b_{ij} are the elements of $\mathcal{B}$, because then

$$\frac{x_c}{y_c} \frac{y_d}{x_d} < (\min_j \left[\frac{\sum_i x_i b_{ij}}{\sum_i y_i b_{ij}}\right]) \frac{1}{\max_j \left[\frac{\sum_i x_i b_{ij}}{\sum_i y_i b_{ij}}\right]} \qquad \text{(III-64)}$$

But (III-72) is guaranteed by Lemma 4 whenever $y_d \neq 0$ and $b_{ij} > 0$ for all i, j. (by assumption, $y_c \neq 0$ otherwise $\underline{y}$ would be a zero vector). Furthermore, (III-63) is also satisfied in the case $y_d = 0$, provided $b_{ij} > 0$, $\forall i,j$, since by assumption not all $y_i = 0$.

Next suppose that there exists i,j such that $b_{ij} = 0$. Then, by the assumption of this lemma, either

(a) $b_{ij} = 0 \quad \forall \ell$.

In this case $\sum_\ell x_\ell b_{\ell j} = \sum_\ell y_\ell b_{\ell j} = 0$

and $\dfrac{\sum_\ell x_\ell b_{\ell j}}{\sum_\ell y_\ell b_{\ell j}}$ is not considered in the determination of $D(\underline{x}\mathcal{B}; \underline{y}\mathcal{B})$.

or (b) $b_{ih} = 0$, $\forall$ h.

Case (i) In this case consider first the situation in which $x_c \neq 0$; $y_d \neq 0$.

From Lemma 4 it is clear that either (III-63) is satisfied anyway, or one of the two inequalities in (III-63) becomes an equality.

Assume first that

$$\frac{x_c}{y_c} = \frac{\sum_i x_i b_{ij}}{\sum_i y_i b_{ij}} \qquad \text{(III-65)}$$

But by (III-50) this may be only if

$$b_{ij} = 0 \text{ for all } i \ni \frac{x_i}{y_i} \neq \frac{x_c}{y_c} \qquad \text{(III-66)}$$

But in the above case, by the property (III-57), relation (III-66) must be true for all j and so must (III-65). In such a case $D(\underline{x}\mathcal{B} ; \underline{y}\mathcal{B}) = 1$ and (III-62) is satisfied. Also if

$$\frac{x_d}{y_d} = \frac{\sum_i x_i b_{ij}}{\sum_i y_i b_{ij}} \qquad \text{(III-67)}$$

then the same argument holds and again

$$D(\underline{x}\mathcal{B}; \underline{y}\mathcal{B}) = 1$$

Case (ii) Consider now the situation where $x_c = 0; y_d \neq 0$ In such a case (III-65) cannot be true, because if it were, then by the above argument $\underline{x}\mathcal{B} = 0$, which against our assumption. Thus only (III-67) can hold and such a case is taken care of in (i).

If $y_d = 0$, $x_c \neq 0$, (III-65) and not (III-66) can hold and all is well again.

Case (iii) Finally, if $y_d = y_c = 0$, then neither (III-65) nor (III-67) can hold, and (III-64) must again be satisfied.

We have therefore shown the validity of (III-62) under the assumptions of our Lemma.

Next, let

$$\frac{b_{mq}}{b_{nq}} \frac{b_{np}}{b_{mp}} = \min_{i,j,k,\ell} \left[\frac{b_{i\ell}\ b_{jk}}{b_{j\ell}\ b_{ik}}\right] \tag{III-68}$$

By the assumption (III-57), clearly

$$1 \geq \frac{b_{mq}\ b_{np}}{b_{nq}\ b_{mp}} > 0 \quad , \tag{III-69}$$

equality on the left-hand side being possible only if $\frac{b_{i\ell}}{b_{j\ell}} = \frac{b_{ik}}{b_{jk}}$ for all i,j,k,ℓ, i.e. if the rank of $\mathcal{B}$ is one.

We will show that:

$$D(\underline{x}\mathcal{B};\underline{y}\mathcal{B})-D(x;y) \geq \frac{\dfrac{b_{mq}\ b_{np}}{b_{nq}\ b_{mp}}}{1 + \dfrac{b_{mq}\ b_{np}}{b_{nq}\ b_{mp}} D(x,y)} [1 - D^2(x,y)] \tag{III-70}$$

Now

$$D(\underline{x}\mathcal{B};\underline{y}\mathcal{B})-D(x;y) =$$

$$= \min_{ij} \left[\frac{(\sum_\ell x_\ell\ b_{\ell i})\ \sum_k y_k\ b_{kj}}{(\sum_\ell y_\ell\ b_{\ell i})\ \sum_k x_k\ b_{kj}} - \frac{x_c}{y_c}\ \frac{y_d}{x_d}\right] = \text{(III-71)}$$

$$= \min_{i,j} \left[\frac{\sum_{\ell,k} b_{\ell i}\ b_{kj}[x_\ell y_k y_c x_d - x_k y_\ell x_c y_d]}{y_c\ x_d \sum_{\ell,k} b_{\ell i}\ b_{kj}\ y_\ell\ y_k}\right]$$

From definition (III-23) & (III-24) it is clear that $y_c \neq 0$, $x_d \neq 0$. However, it is certainly possible that in some cases $y_c = 0$ or $x_d = 0$. Furthermore, there might exist some additional i, $i \neq c$, such that $x_i = 0$, or some additional j, $j \neq d$, such that $y_j = 0$. We must take notice of all these possibilities.

Let us first eliminate from the summations in both the numerator and denominator of the right-hand side of (III-71) all terms including coordinate products $x_\ell\, y_k$ such that $x_k\, y_\ell = x_\ell\, y_k = 0$. Also, property (III-57) insures that whenever $b_{\ell i}\, b_{kj} = 0$ then also $b_{ki}\, b_{\ell j} = 0$. We will then further eliminate from the sums of the right-hand side of (III-71) all the terms which will include the matrix element products $b_{\ell i}\, b_{kj} = b_{ki}\, b_{\ell j} = 0$. The remaining terms in the summations then are such that at least one of $(b_{\ell i}\, b_{kj}\, x_\ell\, y_k,\ b_{k\ell}\, b_{\ell j}\, x_k\, y_\ell)$ and one of $(b_{\ell i}\, b_{kj}\, x_k\, y_\ell,\ b_{ki}\, b_{\ell j}\, x_\ell\, y_k)$ are non-zero. Taking this into consideration, it follows from (III-49) of Lemma 4 that

$$\min_{i,j} \left[\frac{\sum_{\ell,k} b_{\ell i}\, b_{kj} [x_\ell y_k - x_k y_\ell \;\; \frac{x_c y_d}{y_c x_d}\;]}{\sum_{\ell,k} b_{\ell i}\, b_{kj}\, y_\ell\, x_k} \right] \geq$$

$$\geq \min_{i,j,k,\ell} \left[\frac{b_{\ell i} b_{kj} [x_\ell y_k - x_k y_\ell \frac{x_c y_d}{y_c x_d}] + b_{ki} b_{\ell j} [x_k y_\ell - x_\ell y_k \frac{x_c y_d}{y_c x_d}]}{b_{\ell i}\, b_{kj}\, x_k\, y_\ell + b_{ki}\, b_{\ell j}\, x_\ell\, y_k} \right]$$

(III-72)

Therefore inequality (III-73) holds,

$$D(\underset{\smile}{x}\mathcal{B}, \underset{\smile}{y}\mathcal{B}) - D(x,y) \geq \min_{i,j,k,\ell} \left[\frac{b_{\ell i} b_{kj} [x_\ell y_k - x_k y_\ell \frac{x_c y_d}{y_c x_d}] + b_{ki} b_{\ell j} [x_k y_\ell - x_\ell y_k \frac{x_c y_d}{y_c x_d}]}{b_{\ell i}\, b_{kj}\, x_k\, y_\ell + b_{ki}\, b_{\ell j}\, x_\ell\, y_k} \right] \tag{III-73}$$

where only those terms which were not eliminated in the above procedure were taken into consideration when the minimum on the right-hand side of (III-73) was being determined.

We will now show that the right-hand side of (III-73) is greater than or equal to the right-hand side of (III-70). This will be accomplished if we succeed in proving that for all i,j,k,ℓ the inequality (III-74) holds:

$$\frac{x_\ell y_k [b_{\ell i} b_{kj} - b_{ki} b_{\ell j} \frac{x_c y_d}{y_c x_d}] + x_k y_\ell [b_{ki} b_{\ell j} - b_{\ell i} b_{kj} \frac{x_c y_d}{y_c x_d}]}{x_\ell y_k\, b_{ki} b_{\ell j} + x_k y_\ell\, b_{\ell i} b_{kj}} \geq$$

$$\geq \frac{b_{mq}\, b_{np}\, y_c x_d\, [1 - (\frac{x_c\, y_d}{y_c\, x_d})^2]}{b_{nq}\, b_{mp}\, y_c x_d + b_{mq}\, b_{np}\, x_c y_d} = \tag{III-74}$$

$$= \frac{y_c x_d\, b_{mq}\, b_{np} + x_c y_d\, [-b_{mq}\, b_{np} \frac{x_c y_d}{y_c x_d}]}{y_c x_d\, b_{nq}\, b_{mp} + x_c y_d\, b_{mq}\, b_{np}}$$

Subtracting the right-hand side of (III-74) from the left-hand side, we get the expression

$$\frac{\begin{array}{l} b_{\ell i}\, b_{kj}\, b_{mq}\, b_{np}[x_\ell y_k x_c y_d - x_k y_\ell x_d y_c] + b_{\ell i} b_{kj}\, b_{mp}\, b_{nq}\,[x_\ell y_k x_d y_c - x_k y_\ell x_c y_d] + b_{ki} b_{\ell j} b_{mq} b_{np}[x_k y_\ell x_c y_d - x_\ell y_k x_d y_c] + \\ + b_{ki}\, b_{\ell j}\, b_{mp}\, b_{nq}[x_k y_\ell x_d y_c - x_\ell y_k x_c y_d] \end{array}}{(x_\ell y_k\, b_{ki}\, b_{\ell j} + x_k y_\ell\, b_{\ell i}\, b_{kj})(y_c x_d\, b_{nq}\, b_{mp} + x_c y_d\, b_{mq}\, b_{np})} = \tag{III-75}$$

$$\frac{\begin{array}{l} [b_{\ell i}\, b_{kj}\, b_{mq}\, b_{np} - b_{ki}\, b_{\ell j}\, b_{mp}\, b_{nq}][x_\ell y_k x_c y_d - x_k y_\ell x_d y_c] + \\ + [b_{\ell i}\, b_{kj}\, b_{mp}\, b_{nq} - b_{ki}\, b_{\ell j}\, b_{mq}\, b_{np}][x_\ell y_k x_d y_c - x_k y_\ell x_c y_d] \end{array}}{(x_\ell y_k\, b_{ki}\, b_{\ell j} + x_k y_\ell\, b_{\ell i}\, b_{kj})(y_c x_d\, b_{nq}\, b_{mp} + x_c y_d\, b_{mq}\, b_{np})}$$

and we should show that (III-75) is greater than or equal to zero.

Now two cases are possible:

Case I:

either $x_c = 0$ or $y_d = 0$

In such a case to prove (III-74) we must only show that for all i,j,k,ℓ

$$b_{\ell i}\, b_{kj}\, b_{mp}\, b_{nq} \left[\frac{b_{ki}\, b_{\ell j}}{b_{\ell i}\, b_{kj}} - \frac{b_{mq}\, b_{np}}{b_{mp}\, b_{nq}} \right] \geq 0$$

$$b_{ki}\, b_{\ell j}\, b_{mp}\, b_{nq} \left[\frac{b_{\ell i}\, b_{kj}}{b_{ki}\, b_{\ell j}} - \frac{b_{mq}\, b_{np}}{b_{mp}\, b_{nq}} \right] \geq 0 \tag{III-75a}$$

But (III-75a) is satisfied by definition (III-68).

Case II.

$x_c \neq 0$ and $y_d \neq 0$

In this case it is clear from (III-75) that in addition to (III-75a) we must also show that for all i,j,k,ℓ:

$$x_\ell y_k x_d y_c \left[\frac{x_k y_\ell}{y_\ell y_k} - \frac{x_c y_d}{y_c x_d} \right] \geqslant 0$$

$$x_k y_\ell x_d y_c \left[\frac{x_\ell y_k}{y_\ell x_k} - \frac{x_c x_d}{y_c y_d} \right] \geqslant 0 \qquad \text{(III-76)}$$

But (III-76) is satisfied by definition (III-25). Hence the assertion (III-70) is proven.

Consider now all the matrices $\mathcal{B}^i$ in the set S, together with their corresponding quantities (III-68). Among the latter there must be one, say $\frac{b^+_{mq}\, b^+_{np}}{b^+_{nq}\, b^+_{mp}}$ having the largest value, and one, say $\frac{b^-_{mq}\, b^-_{np}}{b^-_{nq}\, b^-_{mp}}$ having the smallest value. Then for all matrices $\mathcal{B} \varepsilon S$ the inequality

$$D(\underline{x}\mathcal{B}; \underline{y}\mathcal{B}) - D(x;y) \geqslant$$

$$\geqslant \frac{\dfrac{b^-_{mq}\, b^-_{np}}{b^-_{nq}\, b^-_{mp}}}{1 + D(x;y)\dfrac{b^+_{mq}\, b^+_{np}}{b^+_{nq}\, b^+_{mp}}} [1-D^2(x;y)] \qquad \text{(III-77)}$$

must hold. But (III-77) shows that after each transformation of vectors $\underline{x}$, $\underline{y}$ by the matrix $\mathcal{B}$, the resulting vectors $\underline{x}\mathcal{B}$, $\underline{y}\mathcal{B}$ have a directional distance closer to the value 1, and moreover, that the value 1 is actually approached in the limit, since otherwise the right-hand side of (III-77) would always remain a finite quantity so that the value 1 would actually be exceeded in the end. Thus Lemma 5 is proven.

Q.E.D.

The previously proven Lemmas 1,2,3,4 & 5, taken all together lead to the following, aimed for

<u>Theorem</u>

Given a Markoff source having a transition probability matrix [M], with states $s_1,\ldots,s_k$ partitioned into disjoint non-empty sets $A_1,A_2,\ldots,A_\ell$, $\ell < k$. Construct, by the procedure described in paragraph (III-3b), the contracted set-transition matrices $\mathcal{B}_i(j)$, $i,j\varepsilon(1,2,\ldots,\ell)$. Given any $\varepsilon > 0$ an integer Λ can be found such that for all integers $\lambda > \Lambda$ and $\sigma \geq 1$ and for all sequences $A^i, A^{i-1},\ldots,A^{i-\lambda},\ldots,A^{i-\lambda-\sigma}$ for which

$$Pr(A^i,\ldots,A^{i-\lambda},\ldots,A^{i-\lambda-\sigma}) \neq 0 \qquad \text{(III-78)}$$

the inequality

$$\left|Pr(A^i/A^{i-1},\ldots,A^{i-\lambda})-Pr(A^i/A^{i-1},\ldots,A^{i-\lambda},\ldots,A^{i-\lambda-\sigma})\right|<\varepsilon$$

(III-79)

is satisfied, provided that all set sequences fulfilling (III-78) correspond to set transition matrix products

$$\mathcal{B}^{i-\lambda}\ \mathcal{B}^{i-\lambda+1}\ \ldots.\mathcal{B}^{i-1} \qquad \text{(III-80)}$$

(where notation (III-26) is used) having the following property:

There exists a fixed integer t such that if any of the products (III-80) is expressed as a product of subproducts of t matrices:

$$[\mathcal{B}^{i-\lambda} \ldots \mathcal{B}^{i-\lambda+t-1}][\mathcal{B}^{i-\lambda+t} \ldots \mathcal{B}^{i-\lambda+2t-1}] \ldots [\mathcal{B}^{i-\lambda+st} \ldots \mathcal{B}^{i-1}]$$

where $st \leq \lambda \leq (s+1)t$, (III-81)

then each of the subproducts results in a matrix

$$G(i-\lambda+jt, \ldots, i-\lambda+(j+1)t-1) = [\mathcal{B}^{i-\lambda+jt} \ldots \mathcal{B}^{i-\lambda+(j+1)t-1}]$$

(III-82)

having elements $g_{i,j}$ such that whenever $g_{ij} = 0$ then either $g_{ip} = 0$ for all p, or $g_{qj} = 0$ for all q.

Appendix IV

Uniform Convergence of Probabilities of Disjoint Sets of Markoff States conditioned by a State Set Past of Fixed Length, when the Source States are Determined by a Finite Past of Discrete Outputs.

In this appendix we will be concerned with the question of when the set-transition matrices of a special class of Markoff sources satisfy the sufficient condition for set probability convergence given in Theorem III-1.

As the heading implies, the states of the Markoff sources of interest are to be determined by discrete outputs in a manner to be described below. Without any loss in generality we will in the following exposition limit the alphabet of the outputs to a binary one. It will be clear that a straightforward generalization to n-ary outputs exists.

Consider now the following class of Markoff sources. At given discrete time intervals the source changes state and puts out a binary sequence of length n whose identity is a function of the newly assumed state. Hence at time i the source switches from state s^{i-1} to state s^{i}, putting out simultaneously the sequence

$$(x_1^{i-1}, x_2^{i-2}, \dots , x_n^{i-1}) = \phi(s^i) \qquad x_j^{i-1} \in (0,1) \qquad \text{(IV-1)}$$

There are $2^{n\ell}$ states and at any time the present state is determined by ℓ immediately preceding outputs. Consequently, $2^{n(\ell-1)}$ states have identical outputs and there is a one-to-one mapping of each point in the state space onto the output space of ordered sequences of ℓ binary n-tupples.

Therefore we can write that

$$s^i = \psi \begin{pmatrix} x_1^{i-1}, x_2^{i-1}, \ldots, x_n^{i-1} \\ x_1^{i-2}, x_2^{i-2}, \ldots, x_n^{i-2} \\ \vdots \\ x_1^{i-\ell}, x_2^{i-\ell}, \ldots, x_n^{i-\ell} \end{pmatrix} \qquad \text{(IV - 2)}$$

and an inverse function ψ^{-1} from arrays of ℓ n-tupples to states exists. If s^i occurred then a transition is possible to only those states s^{i+1} which are characterized in terms of the array on the right hand side of (IV-2) as follows:

(a) they can have any 1^{st} row

(b) their k^{th} row $(k=2,3,\ldots,\ell)$ is identical with the $(k-1)^{st}$ row of s^i. (IV-3)

Therefore, given state s^i, there are 2^m possible next states, and 2^m possible immediately preceding states. By this we mean that transitions into any other state but those characterized as in (IV-3) must have the probability 0.

Each row in (IV-2) can be considered as a binary number and as such has a decimal equivalent. Let us denote it by

$$\xi_k^i \equiv x_1^{i-k} 2^{n-1} + x_2^{i-k} 2^{n-2} + \ldots + x_n^{i-k} \qquad \text{(IV-4)}$$

Similarly we can characterize the state s^i of (IV-2) by the decimal number σ^i,

$$s^i \leftrightarrow \sigma^i \equiv \xi_\ell^i 2^{n(\ell-1)} + \xi_{\ell-1}^i 2^{n(\ell-2)} + \cdots\cdots + \xi_1^i \qquad \text{(IV-6)}$$

where ξ^i_k were the decimal equivalents to the rows of the array on the right hand side of (IV-2) as described in (IV-4). Identifying now states and outputs by their decimal correspondents, we can describe the transition matrix $[\mathcal{M}]$ of the Markoff source. Its elements are

$$m_{\sigma\tau} = p\,(s^{i+1} \leftrightarrow \tau / s^i \leftrightarrow \sigma^i) \qquad \text{(IV-7)}$$

and we know that

$$m_{\sigma\tau} = 0$$

$$\text{unless} \quad (\sigma 2^n - \xi_\ell(\sigma) 2^{n\ell}) \leq \tau < (\sigma 2^n - \xi_\ell(\sigma) 2^{n\ell}) + 2^n \qquad \text{(IV-8)}$$

Thus the matrix will have at most 2^n non-zero entries in each row or column.

Suppose that in the expression (IV-1) we suppress w of the n digits ($w < n$). Without loss in generality, we may suppose that we suppressed the first w digits. Thus the new output will be, instead of the word (IV-1), the word (IV-9)

$$(x^{i-1}_{w+1}\,, x^{i-1}_{w+2}\,, \ldots , x^{i-1}_n) \qquad \text{(IV-9)}$$

If we were to observe the new output now, we would never be able to determine what state the Markoff source was in, no matter how much of the output past we took into account. In fact, unless some unusual degeneracies occurred, if we wished to make the best possible prediction of the next output, we would have to use our knowledge of the entire past output since the beginning of the source's operation.

If should be observed that now $2^{w+n(\ell-1)}$ different states will give the same observable output (IV-9). In fact, any state is, as far as its output past and present goes, indistinguishable from 2^w other states. Thus we may partition the $2^{n\ell}$ states of the Markoff source into $2^{(n-w)\ell}$ disjoint sets, each containing all the states characterized by the same array of non-suppressed outputs. After the manner of Appendix III we will denote the sets by the letters A_i, there being a one-to-one mapping of each point in the set-space onto the space of ordered sequences of binary (n-w)-tupple outputs. We can therefore write that

$$A^i \equiv \Omega \begin{pmatrix} x_{w+1}^{i-1}, & x_{w+2}^{i-1}, & \dots, & x_n^{i-1} \\ x_{w+1}^{i-2}, & x_{w+2}^{i-2}, & \dots, & x_n^{i-2} \\ \vdots & & & \\ x_{w+1}^{i-\ell}, & x_{w+2}^{i-\ell}, & \dots, & x_n^{i-\ell} \end{pmatrix} \tag{IV-10}$$

and an inverse function Ω^{-1} from arrays of ℓ(n-w)-tupples to state-sets exists.

Again, each row in (IV-10) can be considered a binary number with a decimal equivalent denoted by

$$\eta_k^i \equiv x_{w+1}^{i-k} 2^{n-(w+1)} + x_{w+2}^{i-k} 2^{n-(w+2)} + \dots + x_n^{i-k} \tag{IV-11}$$

Then also sets A^i can be characterized by a decimal number α^i defined below

$$A^i \leftrightarrow \alpha^i \equiv \eta_\ell^i 2^{(n-w)(\ell-1)} + \eta_{\ell-1}^i 2^{(n-w)(\ell-2)} + \dots\dots + \eta_1^i \tag{IV-12}$$

where η_k^i were the decimal equivalents of the rows of the array on the right hand side of (IV-10), as described in (IV-11). A transition from a state $s^i \epsilon A^i_\alpha$ to a state $s^{i+1} \epsilon A^{i+1}_\beta$ is possible if and only if

$$(\alpha 2^{n-w} - \eta_\ell(\alpha) 2^{(n-w)\ell}) \le \beta < (\alpha 2^{n-w} - \eta_\ell(\alpha) 2^{(n-w)\ell}) + 2^{n-w} \qquad \text{(IV-13)}$$

The operation of the source with some of its outputs suppressed can be illustrated by Figure IV-1

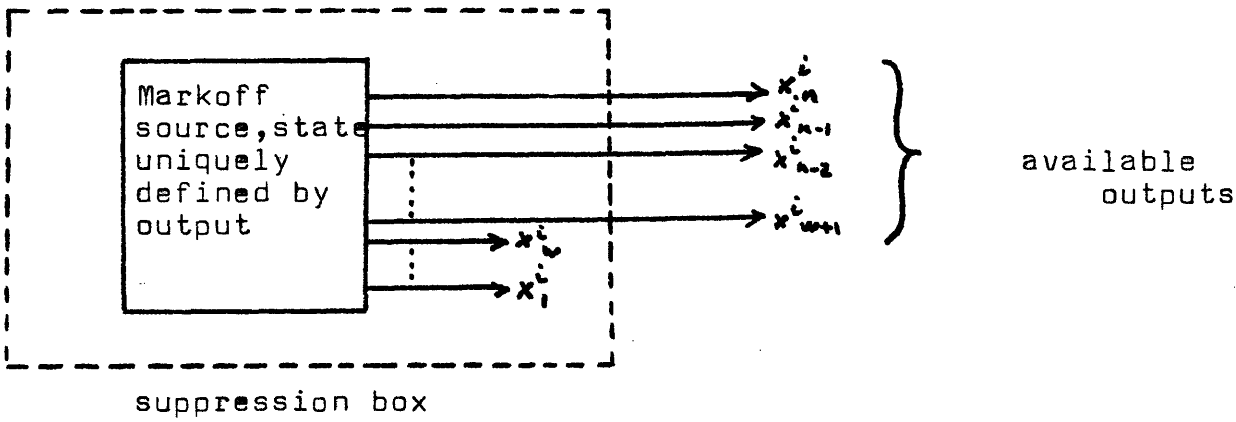


Figure IV-1

It's obvious that a sequence of state sets $A^i, A^{i+1}, \ldots A^{i+\nu}$ identifies outputs (n-w)-tupples for times $i-\ell, i-\ell+1, \ldots, i+\nu-1$, and vice versa. Thus the best possible prediction of the output (n-w)-tupple at time i is obtained by the use of our knowledge of the sequence of sets $A^\ell, A^{\ell+1}, \ldots$ $\ldots, A^i$. The proper conditional probability is

$$p(A^{i+1}/A^i, A^{i-1}, \ldots, A^\ell) \qquad \forall\, i > \ell \qquad \text{(IV-14)}$$

where the first row of the array analogous to (IV-10) characterizing A^{i+1} is the (n-w)-tupple whose probability we want to determine.

Clearly:

$$p(A^{i+1}/A^i, \ldots, A^\ell) = \frac{P(A^{i+1}, A^i, \ldots, A^\ell)}{P(A^i, \ldots, A^\ell)} \qquad i > \ell \qquad \text{(IV-15)}$$

To find the numerator and denominator probabilities in (IV-15) we proceed as in Appendix III.

Suppose we know the original Markoff matrix $[\mathcal{M}]$ with elements defined in (IV-7) and (IV-8). By the procedure (III-3b) we will construct the contracted set-transition matrices $\mathcal{B}_\alpha(\beta)$ from set A_α to set A_β (Note that by our notation set A_α has the decimal equivalent representation α). They will have dimensions $2^{(n-w)\ell} \times 2^{(n-w)\ell}$. Similarly, by the same procedure, we will construct the contracted vectors $\lfloor P(\alpha) \rfloor$ having $2^{(n-w)\ell}$ coordinates. Then by Lemma III-1 we can write:

$$p(A^i/A^{i-1},\dots,A^{i-k}) = \frac{\lfloor P(\alpha) \rfloor \mathcal{B}^{i-k}\mathcal{B}^{i-k+1}\dots\mathcal{B}^{i-1}\, 1]}{\lfloor P(\alpha) \rfloor \mathcal{B}^{i-k}\mathcal{B}^{i-k+1}\dots\mathcal{B}^{i-2}\, 1]} \qquad \text{(IV-16)}$$

where it was assumed that $A^{i-k} = A_\alpha$ and if $A^j = A_\beta$ and $A^{j+1} = A_\gamma$ then $\mathcal{B}^j = \mathcal{B}_\beta(\gamma)$.

The preceding notation and definitions will now be illustrated by an example.

Example IV-1:

Let $\ell = 2$, n=1, w=1 .

The source will then have the output (x_1, x_2) and 16 states whose decimal equivalence representation is illustrated below:

$$0 \leftrightarrow \begin{pmatrix} x_1^{-1} & x_2^{-1} \\ x_1^{-2} & x_2^{-2} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad 5 \leftrightarrow \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$$1 \leftrightarrow \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad 9 \leftrightarrow \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$\vdots$$

$$4 \leftrightarrow \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \qquad 15 \leftrightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

As seen in Figure IV-2, the output x_1 is to be suppressed.

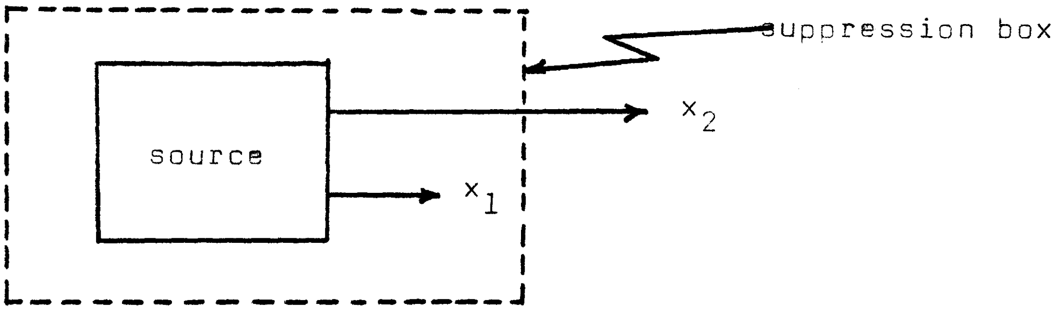


Figure IV-2

This creates four state subsets as follows:

$$A_o = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \right\}$$

$$A_1 = \left\{ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right\}$$

$$A_2 = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right\}$$

$$A_3 = \left\{ \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right\}$$

The pertinent contracted state transition matrices are:

$$\mathcal{B}_0(0) = \begin{bmatrix} m_{0,0} & m_{0,2} & 0 & 0 \\ 0 & 0 & m_{2,8} & m_{2,10} \\ m_{8,0} & m_{8,2} & 0 & 0 \\ 0 & 0 & m_{10,8} & m_{10,10} \end{bmatrix} \qquad \mathcal{B}_2(0) = \begin{bmatrix} m_{4,0} & m_{4,2} & 0 & 0 \\ 0 & 0 & m_{6,8} & m_{6,10} \\ m_{12,0} & m_{12,2} & 0 & 0 \\ 0 & 0 & m_{14,8} & m_{14,10} \end{bmatrix}$$

$$\mathcal{B}_0(1) = \begin{bmatrix} m_{0,1} & m_{0,3} & 0 & 0 \\ 0 & 0 & m_{2,9} & m_{2,11} \\ m_{8,1} & m_{8,3} & 0 & 0 \\ 0 & 0 & m_{10,9} & m_{10,11} \end{bmatrix} \qquad \mathcal{B}_2(1) = \begin{bmatrix} m_{4,1} & m_{4,3} & 0 & 0 \\ 0 & 0 & m_{6,9} & m_{6,11} \\ m_{12,1} & m_{12,3} & 0 & 0 \\ 0 & 0 & m_{14,9} & m_{14,11} \end{bmatrix}$$

$$\mathcal{B}_0(2) = \mathcal{B}_0(3) = [0] \qquad \mathcal{B}_2(2) = \mathcal{B}_2(3) = [0]$$

$$\mathcal{B}_1(0) = \mathcal{B}_1(1) = [0] \qquad \mathcal{B}_3(0) = \mathcal{B}_3(1) = [0]$$

$$\mathcal{B}_1(2) = \begin{bmatrix} m_{1,4} & m_{1,6} & 0 & 0 \\ 0 & 0 & m_{3,12} & m_{3,14} \\ m_{9,4} & m_{9,6} & 0 & 0 \\ 0 & 0 & m_{11,12} & m_{11,14} \end{bmatrix} \qquad \mathcal{B}_3(2) = \begin{bmatrix} m_{5,4} & m_{5,6} & 0 & 0 \\ 0 & 0 & m_{7,12} & m_{7,14} \\ m_{13,4} & m_{13,6} & 0 & 0 \\ 0 & 0 & m_{15,12} & m_{15,14} \end{bmatrix}$$

$$\mathcal{B}_1(3) = \begin{bmatrix} m_{1,5} & m_{1,7} & 0 & 0 \\ 0 & 0 & m_{3,13} & m_{3,15} \\ m_{9,5} & m_{9,7} & 0 & 0 \\ 0 & 0 & m_{11,13} & m_{11,15} \end{bmatrix} \qquad \mathcal{B}_3(3) = \begin{bmatrix} m_{5,5} & m_{5,7} & 0 & 0 \\ 0 & 0 & m_{7,13} & m_{7,15} \\ m_{13,5} & m_{13,7} & 0 & 0 \\ 0 & 0 & m_{15,13} & m_{15,15} \end{bmatrix}$$

while the pertinent contracted vectors are:

$\underline{P(0)} = \underline{p(0),\ p(2),\ p(8),\ p(10)}$

$\underline{P(1)} = \underline{p(1),\ p(3),\ p(9),\ p(11)}$

$\underline{P(2)} = \underline{p(4),\ p(6),\ p(12),\ p(14)}$

$\underline{P(3)} = \underline{p(5),\ p(11),\ p(13),\ p(15)}$

As an example of a conditional probability we have:

$$p(A_0^5 / A_2^4 \ A_1^3 \ A_0^2) = \frac{\underline{P(0)}\ \mathcal{B}_0(1)\ \mathcal{B}_1(2)\ \mathcal{B}_2(0)\ \ 1]}{\underline{P(0)}\ \mathcal{B}_0(1)\ \mathcal{B}_1(2)\ \ 1]} =$$

$$= p(x_2^5 = 0 / x_2^4 = 0,\ x_2^3 = 1,\ x_2^2 = 0,\ x_1^2 = 0)$$

Viewing the matrices of this example, we notice that if all the <u>allowed</u> transitions have non-zero probability, then any possible product of two matrices, $\mathcal{B}_i(j)\ \mathcal{B}_j(k)$, results in a matrix having non-zero elements only. If that were the case for all the set-transition matrix products of the class of Markoff sources introduced in this appendix, then for all sources of the class the sufficient condition of Theorem III-1 for convergence of conditional probabilities of sets would be fulfilled. As a matter of fact we will prove just that in

Theorem IV-1

Given the special class of Markoff sources described in this appendix with the n binary outputs occurring at time i being functions of the state s^i reached:

$$(x_1^{i-1}, \ldots, x_n^{i-1}) = \phi(s^i) \qquad \text{(IV-17)}$$

Let the first w output digits be suppressed so that only outputs

$$(x_{w+1}^{i-1}, \ldots, x_n^{i-1}) \qquad \text{(IV-18)}$$

are observable. Let the states be determined as in (IV-2), represented by decimal number equivalents as in (IV-6), and classified into sets according the ℓ immediately preceding output (n-w)-tupples as in (IV-10). Let the sets in turn be represented by their decimal number equivalents as in (IV-12), and let the state transition rules described by (IV-8) apply. The corresponding set transition matrices $\mathcal{B}_\alpha(s)$ constructed by the rule (III-3b) from the Markoff transition matrix $[m]$ will then have the following property:

If all the transitions allowed by rule (IV-8) have non-zero probability, then any set-transition matrix product of length ℓ,

$$\mathcal{B}^{i-\ell}\,\mathcal{B}^{i-\ell+1}\cdots\,\mathcal{B}^{i-1} = \mathcal{C} \qquad \text{(IV-19)}$$

pertaining to a set sequence $\{A^{i-\ell}, A^{i-\ell+1}, \ldots, A^{i}\}$ such that for it

$$Pr(A^i, A^{i-1}, \ldots, A^{i-\ell}) \neq 0 \qquad \text{(IV-20)}$$

results in a matrix whose every element is positive, i.e.:

$$c_{i,j} > 0 \qquad \text{for all } i,j\,. \qquad \text{(IV-21)}$$

As a consequence it follows from Theorem III-1 that for all $\varepsilon > 0$ there exists an integer Λ such that for all $\lambda > \Lambda$ and $\sigma \geq 1$ the inequality

$$\left| Pr(A^i/A^{i-1}, \ldots, A^{i-\lambda}) - Pr(A^i/A^{i-1}, \ldots, A^{i-\lambda}, \ldots, A^{i-\lambda-\sigma}) \right| < \varepsilon \qquad \text{(IV-22)}$$

holds for any set sequence $\{A^{i-\lambda-\sigma}, \ldots, A^i\}$ for which

$$Pr(A^i, \ldots, A^{i-\lambda-\sigma}) \neq 0 \qquad \text{(IV-23)}$$

Proof:

If (IV-20) is to hold then the succeeding sets in the sequence $\{A^{i-\ell}, A^{i-\ell+1}, \ldots, A^i\}$ must be such that if $A^j = A_\alpha$ and $A^{j+1} = A_\beta$ then relation (IV-13) is satisfied. Any given matrix $\mathcal{C}$ in such a situation is then completely determined by the sets $A^{i-\ell}$ and A^i since the sets in the sequence $\{A^{i-\ell}, \ldots, A^i\}$ for which (IV-20) holds must be determined by the sets $A^{i-\ell}$ & A^i. In that case

$$\mathcal{C} = \mathcal{C}(A^i/A^{i-\ell}) \qquad \text{(IV-24)}$$

is a matrix whose elements $c_{i,j}$ have the value:

$c_{i,j}$ = probability of a transition in ℓ steps from the k^{th} state of set $A^{i-\ell}$ to the j^{th} state of set A^i.

If $A^{i-\ell} = A_\alpha$ then $A^{i-\ell}$ can be represented by

$$\alpha^{i-\ell} = \eta_\ell^{i-\ell} 2^{(n-w)(\ell-1)} + \eta_{\ell-1}^{i-\ell} 2^{(n-w)(\ell-2)} + \cdots + \eta_1^{i-\ell} \qquad \text{(IV-26)}$$

as pointed out in (IV-12). It is then clear from (IV-2) and (IV-10) that $s^{i-\ell} \varepsilon A^{i-\ell}$ if and only if $s^{i-\ell} = s_\sigma$ where

$$\begin{aligned} \sigma^{i-\ell} &= \xi_\ell^{i-\ell} 2^{n(\ell-1)} + \xi_{\ell-1}^{i-\ell} 2^{n(\ell-2)} + \cdots + \xi_1^{i-\ell} = \\ &= \eta_1^{i-\ell} + \zeta_1^{i-\ell} 2^{n-w} + \eta_2^{i-\ell} 2^n + \zeta_2^{i-\ell} 2^{n-w} + \eta_3^{i-\ell} 2^{2n} + \cdots \\ &\quad + \cdots + \eta_\ell^{i-\ell} 2^{n(\ell-1)} + \zeta_\ell 2^{\ell n-w} \end{aligned} \qquad \text{(IV-27)}$$

$$\xi_k^{i-\ell} = \zeta_k^{i-\ell} 2^{n-w} + \eta_k^{i-\ell} \quad ; \quad \zeta_k^{i-\ell} \varepsilon (0,1,\ldots,2^w-1)$$

It then follows from (IV-27) that the k^{th} state of A^i is one represented by the decimal number (IV-27) where

$$k = \zeta_1^{i-\ell} + \zeta_2^{i-\ell} 2^w + \cdots + \zeta_\ell^{i-\ell} 2^{w(\ell-1)}, \quad k \varepsilon (0,1,\ldots,2^{w\ell}-1) \qquad \text{(IV-28)}$$

If $A^i = A_\beta$ where

$$\beta^i = \eta_\ell^i 2^{(n-w)(\ell-1)} + \eta_{\ell-1}^i 2^{(n-w)(\ell-2)} + \cdots + \eta_1^i \qquad \text{(IV-29)}$$

then we must have $A^{i-\ell+h} = A_\gamma$, $h \varepsilon (0,1,\dots,\ell)$ where

$$\gamma^{i-\ell+h} = \eta^{i-\ell}_{\ell-h} 2^{(n-w)(\ell-1)} + \eta^{i-\ell}_{\ell-h-1} 2^{(n-w)(\ell-2)} + \cdots\cdots + \eta^{i-\ell}_1 2^{(n-w)h} + \eta^i_\ell 2^{(n-w)(h-1)} + \cdots + \cdots\cdots + \eta^i_{\ell-h+1} \quad \text{(IV-30)}$$

Similarly, if $s^{i-\ell} = s_\sigma \varepsilon A^{i-\ell}$ is represented as in (IV-27) then a transition to state $s^i = s_\tau \varepsilon A^i$ represented by

$$\tau^i = \eta^i_1 + \xi^i_1 2^{(n-w)} + \eta^i_2 2^n + \xi^i_2 2^{2n-w} + \cdots\cdots + \eta^i_\ell 2^{(\ell-1)n} + \xi^i_\ell 2^{\ell n-w} \quad \text{(IV-31)}$$

is possible only through state $s^{i-\ell+h} = s_\rho \varepsilon A^{i-\ell+h}$ which is represented by

$$\rho^{i-\ell+h} = \eta^i_{\ell-(h-1)} + \xi^i_{\ell-(h-1)} 2^{n-w} + \cdots\cdots + \eta^i_\ell 2^{n(h-1)} + \xi^i_\ell 2^{hn-w} + \eta^{i-\ell}_1 2^{nh} + \xi^{i-\ell}_1 2^{n(h+1)-w} + \cdots\cdots + \eta^{i-\ell}_{\ell-h} 2^{n(\ell-1)} + \xi^{i-\ell}_{\ell-h} 2^{n\ell-w} \quad \text{(IV-32)}$$

It therefore follows that for every state $s^{i-\ell} \varepsilon A^{i-\ell}$ there is a unique transition path to any state $s^i \varepsilon A^i$ and that therefore every element $c_{i,j}$ in the matrix (IV-24) is non-zero if all the state transitions allowed by rule (IV-8) have a non-zero probability.

Q.E.D.

It is interesting to display diagrammatically the appearance of a matrix $\mathcal{B}_\alpha(\Lambda)$. By equation (IV-28) a certain order of states in any set A^i was implied. From rule (IV-8) it follows that given any state $s^i \varepsilon A^i$ there are 2^w states $s^{i+1} \varepsilon A^{i+1}$ into which a transition from s^i is possible. By convention (IV-28) these will occupy in the ordered set A^{i+1} neighboring places, because they are determined by the value of ξ^{i+1}_1 only.

Consequently each row of the matrix $\mathcal{B}_{\alpha}(\beta)$ where α & β satisfy the relation (IV-13), will have 2^w non-zero elements appearing in a block of 2^w neighboring columns. By inspection of (IV-26), (IV-27), (IV-30), and (IV-32) we see that the first state $s^i \epsilon A^i$ has transitions to the first block of 2^w states $s^{i+1} \epsilon A^{i+1}$, the second state to the second block of 2^w, etc. Finally the $2^{w(\ell-1)}$ th state $s^i \epsilon A^i$ has transitions to the last block of 2^w states $s^{i+1} \epsilon A^{i+1}$. (Note that the $2^{w(\ell-1)}$ th state is represented by numbers $\xi_j = 2^w - 1$ for $j = 1, 2, \ldots, \ell-1$ and by $\xi_\ell = 0$).

Now the $(2^{w(\ell-1)}+1)^{\text{th}}$ state in A^i is represented by the number:

$$\sigma^i = \eta_1^i + \eta_2^i 2^n + \cdots + \eta_3^i 2^{2n} + \cdots + \eta_\ell^i 2^{(\ell-1)n} + 2^{\ell n - w} \qquad \text{(IV-33)}$$

and consequently has transitions into the same states in A^{i+1} as the first state in A^i had, i.e., it has transitions into the first block of 2^w states of set A^{i+1}. The cycle now repeats, so that the $(2^{w(\ell-1)}+2)^{\text{th}}$ state ϵA^i has transitions into the second block of 2^w states in A^{i+1}, etc. We can therefore represent $\mathcal{B}_{A^i}(A^{i+1})$ in the following block form, regardless of the identity of the sets A^i and A^{i+1}, as long as a transition from A^i to A^{i+1} is possible by rule (IV-13).

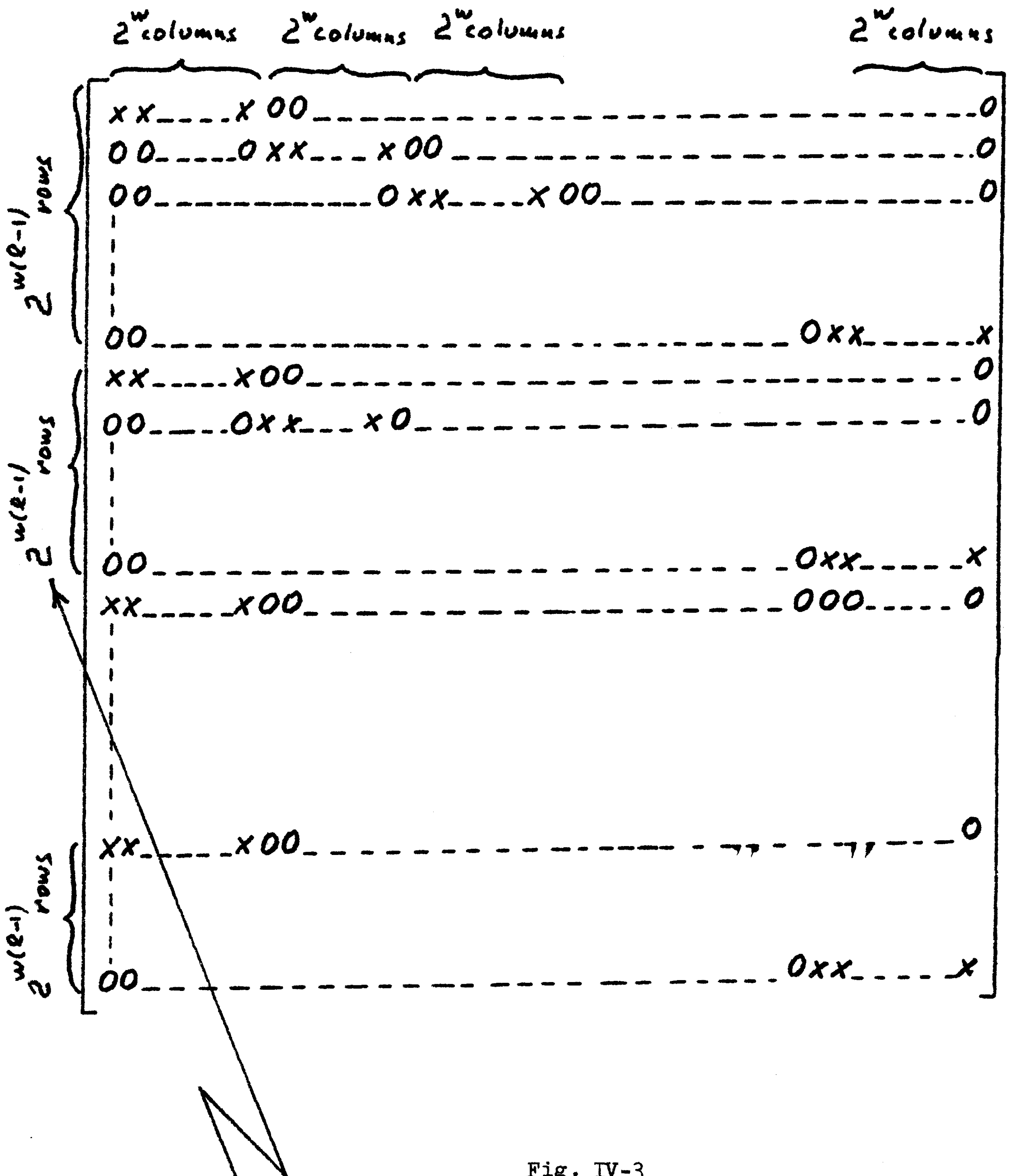


Fig. IV-3

Row cycle repeats. There are altogether 2^W such cycles. The elements marked "x" are potentially positive.

Considering multiplications of matrices of the type of Figure IV-3 it is not difficult to see that the product of r of them, where $r \leq \ell$ will have the diagrammatical appearance of Figure IV-4. Such will then be the appearance of the product

$$\mathcal{B}^{i}\mathcal{B}^{i+1}\ldots\mathcal{B}^{i+r-1} \qquad \text{(IV-34)}$$

which corresponds to a sequence of sets $\{A^{i}, A^{i+1}, \ldots, A^{i+r}\}$ for which

$$\Pr(A^{i+r}, A^{i+r-1}, \ldots, A^{i}) \neq 0 \quad . \qquad \text{(IV-35)}$$

We will next prove a stronger statement of Theorem III-1.

<u>Theorem IV-2</u>

Given a Markoff process and a state partitioning with corresponding set-transition matrices $\mathcal{B}_{i}(j)$ as described in Theorem III-1. Then whenever there exists an integer r such that there corresponds to all set sequences $\{A^{i-r}, \ldots, A^{i}\}$ for which

$$\Pr(A^{i}, A^{i-1}, \ldots, A^{i-r}) \neq 0 \qquad \text{(IV-36)}$$

a set-transition matrix product

$$\mathcal{B}^{i-r}\mathcal{B}^{i-r+1}\ldots\ldots\mathcal{B}^{i-1} = \mathcal{C} \qquad \text{(IV-37)}$$

for which the element $c_{kj} = 0$ if and only if either $c_{kt} = 0$ for all t, or $c_{sj} = 0$ for all j, there will also exist for any $\varepsilon > 0$ an integer Λ such that for all set sequences $\{A^{i-\lambda-\sigma}, \ldots, A^{i-\lambda}, \ldots, A^{i}\}, \lambda > \Lambda$ for which

$$\Pr(A^{i}, A^{i-1}, \ldots, A^{i-\lambda}, \ldots, A^{i-\lambda-\sigma}) \neq 0 \qquad \sigma \geq 1 \qquad \text{(IV-38)}$$

the inequality

$$\left|\Pr(A^{i}/A^{i-1},\ldots,A^{i-\lambda}) - \Pr(A^{i}/A^{i-1},\ldots,A^{i-\lambda},\ldots,A^{i-\lambda-\sigma})\right| < \varepsilon \Pr(A^{i}/A^{i-1},\ldots,A^{i-\lambda}) \qquad \text{(IV-39)}$$

will be satisfied.

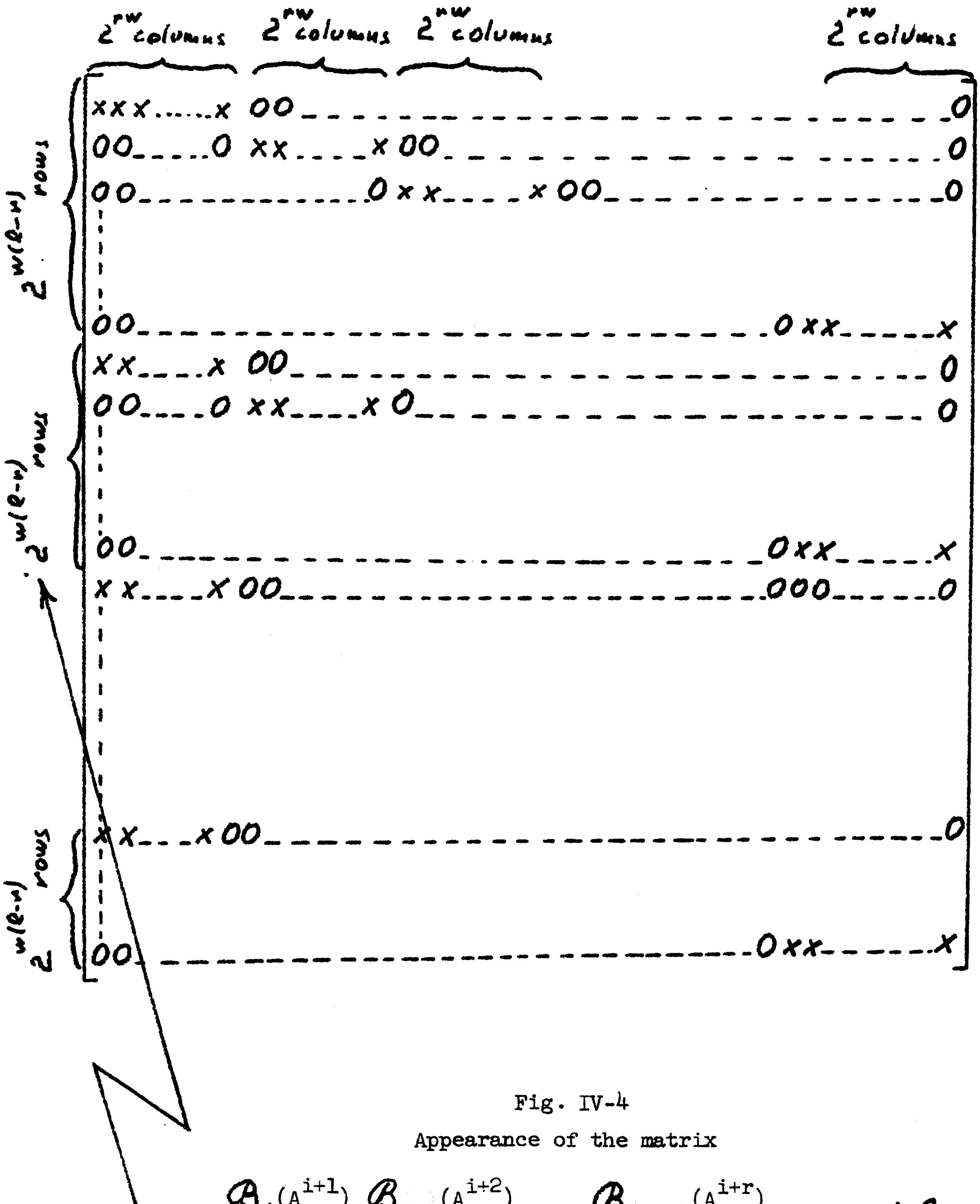


Fig. IV-4

Appearance of the matrix

$$\mathcal{B}_{A^i}(A^{i+1})\ \mathcal{B}_{A^{i+1}}(A^{i+2})\ \ldots\ \mathcal{B}_{A^{i+r-1}}(A^{i+r}) \quad , \quad r \leq \ell$$

Row cycle repeats. There are altogether 2^{rw} such cycles. Elements marked "x" are potentially positive.

Proof:

It should first be noted that except for (IV-39), the above theorem is stated identically with Theorem 1 of Appendix III. In fact, it is a stronger statement of it.

Now, for any fixed λ we are dealing here with a finite number of possible sequences $\{A^i, A^{i-1}, \ldots, A^{i-\lambda}\}$ such that

$$Pr(A^i, A^{i-1}, \ldots, A^{i-\lambda}) \neq 0 \tag{IV-40}$$

Hence, if (IV-40) holds, it is possible to define a set of states $S(A^i)$ where

$$s^{i-1} \in S(A^i) \text{ if and only if } Pr(A^i, s^{i-1}) \neq 0. \tag{IV-41}$$

(Note that $Pr(A^i, s^{i-1})$ is the probability that at a certain time interval $i-1$ the system will be in state s^{i-1} and that at the next interval it will be in one of the states of the set A^i).

It is clear that for all sequences satisfying (IV-38) the following inequality will hold:

$$Pr(A^i/A^{i-1}, \ldots, A^{i-\lambda-}\ \ ) \geq \min_{s \in S(A^i)} \left[P(A^i/s^{i-1}=s)\right] > 0 \tag{IV-42}$$

Thus we can define a quantity

$$\delta = \min_{A^i} \left(\min_{s \in S(A^i)} \left[P(A^i/s^{i-1}=s)\right]\right) \tag{IV-43}$$

which will satisfy inequality

$$Pr(A^i/A^{i-1}, \ldots, A^{i-\lambda}) \geq \delta > 0 \tag{IV-44}$$

for all $\{A^i, A^{i-1}, \ldots, A^{i-\lambda}\}$

such that (IV-40) holds.

It should be noted that (IV-44) holds independently of the value of the integer λ . Thus, given any $\varepsilon > 0$ we can define the quantity

$$\varepsilon_1 = \varepsilon \delta \quad \text{(IV-45)}$$

and since under the conditions of the present theorem the Theorem III-1 holds, we can find the integer Λ for which

$$\left| p(A^i/A^{i-1},\ldots,A^{i-\lambda}) - p(A^i/A^{i-1},\ldots,A^{i-\lambda+\sigma}) \right| < \varepsilon_1 \quad \text{(IV-46)}$$

for all $\lambda > \Lambda$ and $\sigma \geq 1$ whenever (IV-38) is satisfied. But the integer Λ which satisfies (IV-46) will surely satisfy (IV-39) as well, and therefore the theorem is proven.

Q.E.D.

It should be noted here, that because of Lemma 3 of Appendix III, the integer Λ is found for a given $\varepsilon_1 > o$ by examination of the differences

$$\left| p(A^i/A^{i-1},\ldots,A^{i-\Lambda},s_j^{\,i-k-1}) - p(A^i/A^{i-1},\ldots,A^{i-\Lambda},s_n^{\,i-k-1}) \right| \quad \text{(IV-47)}$$

for all s ε $S(A^{i-\Lambda})$. The difference (IV-47) then suggests a very definite procedure for finding an appropriate Λ .

Q.E.D.

Finally, we would like to deal with the actual aim of this Appendix. We would like to prove the theorem needed in Section 6.5 and listed there under the heading 6-1.

Theorem IV-3

Given the special class of Markoff sources described in this appendix with states characterized by configurations $(\mathfrak{x}, \mathfrak{y}, \bar{\mathfrak{x}}, \bar{\mathfrak{y}})_{2\ell}$ (c.f. definition (4-9) with (IV-2)), having $2^{2(2^{2\ell}-1)}$ transitions from a state $(\mathfrak{x}, \mathfrak{y}, \bar{\mathfrak{x}}, \bar{\mathfrak{y}})$ to state $(\mathfrak{x}', \mathfrak{y}', \bar{\mathfrak{x}}', \bar{\mathfrak{y}}')$ provided that

$$x'^{-r} = x^{-r+1} \qquad \bar{x}'^{-r} = x^{-r+1}$$
$$y'^{-r} = y^{-r+1} \qquad \bar{y}'^{-r} = y^{-r+1} \qquad r \; (2,3,\ldots,\ell)\,, \tag{IV-48}$$

each transition associated with a different pair $f, \bar{f}$ such that

$$f(\mathfrak{x}, \mathfrak{y}) = x'^{-1} \quad ; \quad \bar{f}(\bar{\mathfrak{x}}, \bar{\mathfrak{y}}) = \bar{x}'^{-1}\,, \tag{IV-49}$$

the transition probabilities of the allowed transitions being defined by

$$P_{\mathfrak{x}\mathfrak{y}\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(\mathfrak{x}', \mathfrak{y}', \bar{\mathfrak{x}}', \bar{\mathfrak{y}}', f, \bar{f}) = \begin{cases} p(y'^{-1}/x'^{-1}, \bar{x}'^{-1})\,\bar{p}(\bar{y}'^{-1}/x'^{-1}\bar{x}'^{-1})\,P(f)\,\bar{P}(\bar{f}) & \text{if (IV-49) holds} \\ 0 & \text{otherwise} \end{cases} \tag{IV-50}$$

Whenever for all $\mathfrak{x}, \mathfrak{y}, \bar{\mathfrak{x}}, \bar{\mathfrak{y}}, x, \bar{x}$

$$q_{\mathfrak{x}\mathfrak{y}}(x) = \sum_{f \ni f(\mathfrak{x}, \mathfrak{y}) = x} P(f) \neq 0 \quad ; \quad \bar{q}_{\bar{\mathfrak{x}}\bar{\mathfrak{y}}}(\bar{x}) = \sum_{\bar{f} \ni \bar{f}(\bar{\mathfrak{x}}, \bar{\mathfrak{y}}) = \bar{x}} \bar{P}(\bar{f}) \neq 0 \tag{IV-51}$$

then for any $\varepsilon > 0$ there exists an integer t_0 such that whenever

$$\Pr(f, f^{-1}, \ldots, f^{-t-r}, y, y^{-1}, \ldots, y^{-t-r}, s^{t-r}) \neq 0 \quad t > t_0 \quad r \geq 1 \tag{IV-52}$$

the inequality

$$\left| \Pr(y / f, f^{-1}, \ldots, f^{-t}, y^{-1}, \ldots, y^{-t}, s^{-t}) - \Pr(y/f, f^{-1}, \ldots, f^{-t-r}, y^{-1}, \ldots, y^{-t-r}, s^{-t-r}) \right| < \varepsilon \Pr(y / f, f^{-1}, \ldots, f^{-t}, y^{-1}, \ldots, y^{-t}, s^{-t})$$

$$s^{-t} = g(s^{-t-r}, f^{-t-r}, \ldots, f^{-t-1}, y^{-t-r}, \ldots, y^{-t-1}) \tag{IV-53}$$

holds, and for any $\bar{\varepsilon} > 0$ there exists an integer $\bar{t}_0$ such that whenever

$$\Pr(\bar{f}, \bar{f}^{-1}, \ldots, \bar{f}^{-t-r}, \bar{y}, \bar{y}^{-1}, \ldots, \bar{y}^{-t-r}, \bar{s}^{-t-r}) \neq 0 \qquad t > \bar{t}_0, \quad r \geq 1 \tag{IV-54}$$

the inequality

$$\left| \Pr(\bar{y}/\bar{f}, \bar{f}^{-1}, \ldots, \bar{f}^{-t}, \bar{y}^{-1}, \ldots, \bar{y}^{-t}, \bar{s}^{-t}) - \Pr(\bar{y}/\bar{f}, \bar{f}^{-1}, \ldots, \bar{f}^{-t-r}, \bar{y}^{-1}, \ldots, \bar{y}^{-t-r}, \bar{s}^{-t-r}) \right| < \bar{\varepsilon} \Pr(\bar{y}/\bar{f}, \bar{f}^{-1}, \ldots, \bar{f}^{-t}, \bar{y}^{-1}, \ldots, \bar{y}^{-t}, \bar{s}^{-t})$$

$$\bar{s}^{-t} = \bar{g}(\bar{s}^{-t-r}, \bar{f}^{-t-r}, \ldots, \bar{f}^{-t-1}, \bar{y}^{-t-r}, \ldots, \bar{y}^{-t-1}) \tag{IV-55}$$

holds.

Proof:

We are dealing here with a special case of the general situation described in (IV-9) where n=4 and w=2. We will prove the part involving (IV-54) and (IV-55) only, the rest will follow. We will first define notation for identification and ordering of states along the lines developed earlier in this appendix. A state $(\mathfrak{x},\mathfrak{y},\bar{\mathfrak{x}},\bar{\mathfrak{y}})$ can be displayed as

$$(\mathfrak{x},\mathfrak{y},\bar{\mathfrak{x}},\bar{\mathfrak{y}}) \longleftrightarrow \begin{pmatrix} x^{-1} & y^{-1} & \bar{x}^{-1} & \bar{y}^{-1} \\ x^{-2} & y^{-2} & \bar{x}^{-2} & \bar{y}^{-2} \\ \vdots & & & \\ x^{-\ell} & y^{-\ell} & \bar{x}^{-\ell} & \bar{y}^{-\ell} \end{pmatrix} \tag{IV-56}$$

and the state set to which it belongs as

$$(\bar{\mathfrak{x}},\bar{\mathfrak{y}}) \longleftrightarrow \begin{pmatrix} \bar{x}^{-1} & \bar{y}^{-1} \\ \bar{x}^{-2} & \bar{y}^{-2} \\ \vdots & \\ \bar{x}^{-\ell} & \bar{y}^{-\ell} \end{pmatrix} \tag{IV-57}$$

The latter can then be represented by the number

$$\bar{\alpha} = \bar{\eta}_{\ell}\, 2^{2(\ell-1)} + \bar{\eta}_{\ell-1}\, 2^{2(\ell-2)} + \cdots\cdots + \bar{\eta}_1 \tag{IV-58}$$

where

$$\bar{\eta}_k = \bar{x}_{\ell}^{-k}\, 2 + \bar{y}^{-k} \qquad\qquad k \in (1,2,\ldots,\ell) \tag{IV-59}$$

The state $(x, y, \bar{x}, \bar{y})$ can then be represented by the decimal number

$$\sigma = \eta_1 + \xi_1 2^2 + \eta_2 2^4 + \xi_2 2^6 + \cdots + \eta_\ell 2^{2(\ell-1)} + \xi_\ell 2^{2\ell-2} \tag{IV-60}$$

where

$$\xi_k = x^{-k} 2 + y^{-k} \qquad k \varepsilon (1, 2, \ldots, \ell) \tag{IV-61}$$

The order of the state $(x, y, \bar{x}, \bar{y})$ in the set $(\bar{x}\bar{y})$ will depend on the magnitude of the number

$$h = \xi_1 + \xi_2 2^2 + \cdots + \xi_\ell 2^{2(\ell-1)} \tag{IV-62}$$

A transition from state $(x, y, \bar{x}, \bar{y})$ represented by σ to state $(x', y', \bar{x}', \bar{y}')$ represented by τ is possible only if (IV-8) is satisfied for $n=4$. Similarly, a transition from set $(\bar{x}, \bar{y})$ represented by α to set $(\bar{x}', \bar{y}')$ represented by β is possible only if (IV-13) is satisfied for $n-w=2$.

Any state set $(\bar{x}, \bar{y})$ plus "output" $(\bar{f}, \bar{y})$ specify the next state set $(\bar{x}', \bar{y}')$. We are dealing here with a little more general case than the one defined by (IV-7) since whenever a transition between two states is not precluded by rule (IV-48), there actually exist $2^{2(2^{2\ell}-1)}$ possible different transitions. Hence a somewhat different set-transition matrix

$$\mathcal{B}_{\bar{x},\bar{y}}(\xi, \bar{y}; \bar{x}', \bar{y}') \tag{IV-63}$$

of dimension $2^{\ell} \times 2^{\ell}$ for each transition determined by $\bar{f}$ from a state set $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ to state set $(\bar{\mathbf{x}}',\bar{\mathbf{y}}')$ is defined by elements

$$b_{hj} = \begin{cases} p(y'^{-1})/x'^{-1}\,\bar{x}'^{-1})\bar{p}(\bar{y}'^{-1}/x'^{-1},\bar{x}'^{-1})\,\bar{P}(f) \sum_{f \ni f(\mathbf{x},\mathbf{y}) = x'^{-1}} P(f) & \text{if (IV-48) and (IV-49) hold} \\ 0 & \text{otherwise} \end{cases} \quad \text{(IV-64)}$$

where the ordering h for state $(\mathbf{x},\mathbf{y},\bar{\mathbf{x}},\bar{\mathbf{y}})$ and j for state $(\mathbf{x}',\mathbf{y}',\bar{\mathbf{x}}',\bar{\mathbf{y}}')$ is defined by (IV-62).

From Lemma III-1 and the above discussion it is then clear that we may write

$$Pr(\bar{f},\bar{f}^{-1},\ldots,\bar{f}^{-t},\bar{y},\bar{y}^{-1},\ldots,\bar{y}^{-t},\bar{\mathbf{x}}^{-t},\bar{\mathbf{y}}^{-t}) = \quad \text{(IV-65)}$$

$$= \underline{P(\bar{\mathbf{x}}^{-t}\bar{\mathbf{y}}^{-t})}\; \mathcal{B}_{\bar{\mathbf{x}}^{-t}\bar{\mathbf{y}}^{-t}}(\bar{f}^{-t}\bar{y}^{-t};\bar{\mathbf{x}}^{-t+1}\bar{\mathbf{y}}^{-t+1})\; \mathcal{B}_{\bar{\mathbf{x}}^{-t+1}\bar{\mathbf{y}}^{-t+1}}(\bar{f}^{-t+1}\bar{y}^{-t+1};\bar{\mathbf{x}}^{-t+2}\bar{\mathbf{y}}^{-t+2})\ldots$$

$$\ldots\; \mathcal{B}_{\bar{\mathbf{x}}\bar{\mathbf{y}}}(\bar{f},\bar{y};\bar{\mathbf{x}}'\bar{\mathbf{y}}')\; \underline{1}]$$

where $\bar{x}_j^{-1} = \bar{f}^{j-1}(\bar{\mathbf{x}}^{j-1},\bar{\mathbf{y}}^{j-1})$; $\bar{y}_j^{-1} = \bar{y}^{j}$

and vector $\underline{P(\bar{\mathbf{x}}^{-t},\bar{\mathbf{y}}^{-t})}$ was defined by (III-36).

Thus, by the discussion preceding Theorem III-1, it follows from Theorem (IV-2) that (IV-55) will hold if it can be shown that the matrices defined by (IV-63) can be diagrammatically represented by Figure (IV-3) for w=2. But from (IV-64) we see that this is precisely the case whenever (IV-51) and (IV-54) hold.

Q.E.D.

BIOGRAPHICAL NOTE

The author was born in Prague, Czechoslovákia on November 18, 1932. He emigrated into the United States in October 1949, and in June 1950 graduated from the High School of Commerce in New York. He then went to work for the Universal Mfg. Co. in Paterson, N.J. and enrolled as a student of Electrical Engineering in the Evening Session of the City College of New York. In June 1953 he was awarded a scholarship from the Mid European Studies Center in New York and as a consequence transferred to Day Session of the City College. In June 1954 the scholarship was extended to cover tuition expenses at the Massachusetts Institute of Technology which the author started attending in September. He graduated in June 1956, having written a thesis entitled "Realization of Minimum Phase Functions in the Form of Ladder Networks" under the supervision of Professor E. A. Guillemin.

The author was appointed a teaching assistant at M.I.T. in February 1956, and continued his studies receiving his Master's degree in June 1958. This thesis was supervised by Professor R. M. Fano and was entitled, "Coding and Decoding of Binary Group Codes". In June 1959 the author was promoted to the rank of instructor. While at M.I.T. the author has been teaching various graduate and undergraduate courses in Network Theory and Mathematics. In the Summer of 1959 he was a consultant on coding theory to R.C.A. in Camden, N. J., and is currently consulting on Network Theory for Burnell & Co.

The author is a member of Tau Beta Pi, an associate member of the Sigma Xi, and a student member of I.R.E.

The author is married and has a one month old daughter.

BIBLIOGRAPHY

1. C. E. Shannon, Two-Way Communication Channels", Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (to be published).

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

3. C. E. Shannon, "Certain Results in Coding Theory for Noisy Channels", Information and Control, Vol. 1, pp. 6-25, September, 1957.

4. R. M. Fano, Transmission of Information, The M.I.T. Press and John Wiley and Sons, New York, 1961.