ESD-TR-68-54

ESD TR-68-54
File Copy

ESD RECORD COPY
RETURN TO
SCIENTIFIC & TECHNICAL INFORMATION DIVISION
(ESTI), BUILDING 1211

ESD ACCESSION LIST
ESTI Call No. 61808
Copy No. 1 of 1 cys.

ESLL

Technical Report 447

Joel Max

Parallel Channels Without Crosstalk

9 April 1968

Prepared under Electronic Systems Division Contract AF 19(628)-5167 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Lexington, Massachusetts

DIC

AD 673475

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the U.S. Air Force under Contract AF 19(628)-5167.

This report may be reproduced to satisfy needs of U.S. Government agencies.

This document has been approved for public release and sale; its distribution is unlimited.

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

PARALLEL CHANNELS WITHOUT CROSSTALK

JOEL MAX

Group 63

TECHNICAL REPORT 447

9 APRIL 1968

LEXINGTON MASSACHUSETTS

PARALLEL CHANNELS WITHOUT CROSSTALK*

ABSTRACT

In this report, a study is made of information theoretic channels which are decomposable into a number of parallel subchannels which will, in general, be dependent. For this situation, two models are constructed in which each subchannel input affects only the corresponding subchannel output (no crosstalk). In the first model (MC channel), the lack of crosstalk is ensured by constraints on the channel conditional probability distribution. The second model (MS channel) is a channel with an underlying state structure with states independent of the input. Both models are memoryless. All MS channels are MC, but the reverse does not hold.

The effect of subchannel dependencies on capacity and random coding exponent (RCE) is investigated. It is proved that these dependencies cannot decrease the capacity of our channels. However, subchannel dependencies may either increase or decrease the RCE. It is also proved that the capacity of the channel is not less than the sum of the capacities of the individual subchannels. When the state model is used, the above two quantities are equal if the receiver has knowledge of the channel state.

A definition of partial state knowledge is given. It is proved that, when the receiver has partial state knowledge, the resulting capacity and RCE are not decreased. For complete state knowledge at the receiver, the capacity and RCE are not less than those obtained for partial state knowledge.

A restricted class of MS channels is defined wherein all the subchannels are in the same state during each use of the channel; these channels are called MSCC channels. For these channels, a number of results are given, most of which concern the limiting behavior of the capacity per subchannel and the RCE as the number of subchannels becomes large. The principal results are: (1) the capacity per subchannel has a finite limit; and (2) the RCE has a finite limit if the rate per subchannel is kept constant and the constant is sufficiently large. These results hold whether or not the state is known at the receiver.

Systematic coding and decoding, using BCH codes and minimum distance decoding rules, are considered for MSCC channels. Various coding alternatives are discussed, and formulas are given for computing or bounding performance.

Accepted for the Air Force
Franklin C. Hudson
Chief, Lincoln Laboratory Office

---

* This report is based on a thesis of the same title submitted to the Department of Electrical Engineering at the Massachusetts Institute of Technology on 5 May 1967 in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The report differs from the thesis principally in that Appendix F has been added and that the references to M.I.T. course notes have been updated to references to a book by R.G. Gallager which evolved from those notes.

CONTENTS

Abstract iii

CHAPTER 1 – INTRODUCTION 1

CHAPTER 2 – MODELS, DUALITY, AND SOME BASIC THEOREMS 3

A. Parallel Channel Models 3

B. Duality Between Time and Parallel Directions 7

C. Mutual Information and Capacity 9

CHAPTER 3 – STATE REPRESENTATIONS AND BOUNDS FOR MUTUAL INFORMATION AND PROBABILITY OF ERROR 17

A. State Representations 17

B. Entropy 20

C. Natural State Representations 20

D. Bounds on Mutual Information 21

E. Random Coding Bound 22

F. State Knowledge – Some General Considerations 24

G. State Knowledge, Mutual Information and Capacity 26

H. State Knowledge and Random Coding Exponent (RCE) 27

I. Subchannel Dependencies and RCE 31

CHAPTER 4 – THE COMPLETELY CONSTRAINED CHANNEL 33

A. Definition of Channel 33

B. Examples of MSCC Channels and Their Properties 33

C. Capacity Theorems for MSCC Channels 35

D. Further Properties of Examples 1 and 2 36

E. Random Coding Exponent (RCE) for MSCC Channels 37

CHAPTER 5 – SYSTEMATIC CODING FOR COMPLETELY CONSTRAINED CHANNELS 67

A. Introduction 67

B. Coding Alternatives 67

C. Dimensionless Rate 68

D. BCH Codes and Simple Coding Schemes 69

E. State Information and Reliability 70

F. Minimum Distance Decoding 71

G. Single-Letter Erasure and Error Probabilities 72

H. Probability of Correct Decoding 73

I. Chernoff Bound 73

J. Chernoff Bounds for Erasures and/or Errors Decoding 74

K. Bounds on Total Probability of Decoding Failure 76

L. Error Exponents for Some Examples of MSCC Channels 77

M. Compound Coding 81

CHAPTER 6 – MORE GENERAL CHANNEL MODELS 83

A. Markov Parallel Channel 83

B. Capacity of MPC 84

C. Random Coding Exponent for MPC 84

D. Systematie Coding for MPC 84

E. Other MS Channels 86

F. Channels with Both Time and Parallel Dependeneies 86

G. Bloek Model 86

H. Constrained-Markov Model 87

I. Other Models 87

APPENDIX A – Channels Whieh are MC but not MS, and Related Topies 89

APPENDIX B – Proofs of Theorems 2.3 and 2.4 95

APPENDIX C – Some Useful Inequalities 98

APPENDIX D – Canonieal Representations 99

APPENDIX E – A Completely Constrained Channel with a Continuous Parameter 102

APPENDIX F – An MSCC Channel for Whieh $E'_M(R_s) \neq \tilde{E}_M(R_s)$ 104

PARALLEL CHANNELS WITHOUT CROSSTALK

CHAPTER 1
INTRODUCTION

In a typical point-to-point discrete communication situation (see Fig. 1), we have as input to a transmitter a random message m which may take on one of K values. Corresponding to each message value m_i, there is a distinct waveform $s_i(t)$ which is transmitted in response to the message input. The transmitted waveform s(t) is corrupted by the waveform channel (fading, additive noise, attenuation, etc.), and a resultant signal r(t) is the input to the receiver. The

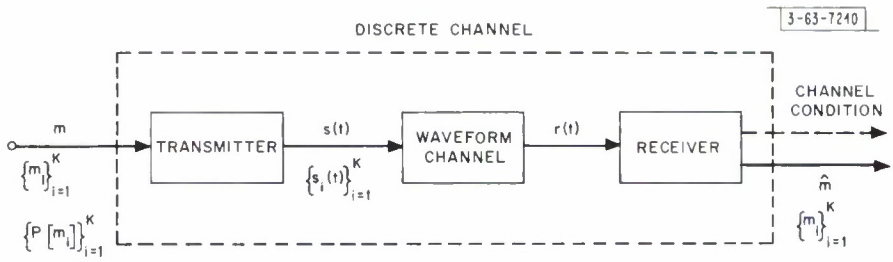


Fig. 1. Discrete communication system model.

receiver then must decide which message was the input to the transmitter; its decision is denoted in Fig. 1 as $\hat{m}$. Discrete information theory generally deals with situations where the modulation (transmitter), waveform channel, and receiver are considered as fixed, and the problems addressed concern the properties and proper utilization of the resulting combination. This combination is called the discrete channel. For the purpose of properly utilizing the discrete channel, we shall be willing to add both pre-transmission and post-reception processing devices. These are usually called coders and decoders, respectively (see Fig. 2). Sometimes, the receiver will have knowledge of the condition (state) of the channel, in which case it is assumed that this information is passed on to the decoder.

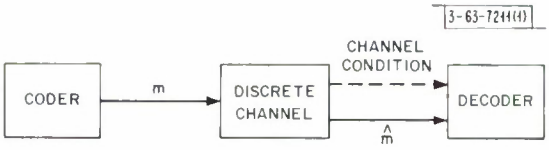


Fig. 2. Discrete channel with coder and decoder.

Often, in practice, the transmitter, receiver, and waveform channel are such that the single discrete channel may be profitably viewed as an aggregate of parallel subchannels. This situation is usually associated with modulation schemes where each subchannel corresponds to a transmitter frequency interval which does not significantly overlap any of the others. We will, in

general, assume that our parallel subchannels are dependent but without crosstalk. The absence of crosstalk implies that each subchannel input affects only the corresponding subchannel output. Such a set of subchannels may, however, be dependent (i.e., the subchannel input-output pairs are dependent in the usual statistical sense) if the natural disturbance (e.g., fading) affecting them is itself not independent from subchannel to subchannel.

Some examples of the dependent parallel channel situation we have in mind are scatter channels (e.g., tropospheric and ionospheric scatter), channels with additive colored Gaussian noise of unknown spectrum, and channels subject to jamming.

Multiple subchannels taken together are usually a less general type of single channel than that which the physical constraints on the communication problem alone would suggest we consider. However, the study of parallel channels is important for three principal reasons. In the first place, many existing communication systems are built in a multiple-channel form. These include HF systems, tropospheric scatter systems, satellite systems, and telephone company equipment of various types. In such situations, the multiple-channel structure is forced upon the user. A second situation is one which is thought to obtain in optical communication systems. Here, the bandwidths are so great that no method is presently available or immediately foreseeable which would allow one to modulate across the entire channel bandwidth at once. A division of the channel bandwidth into subchannels is a technological necessity. Finally, given certain physical constraints on a communication problem, a multiple-channel communication system may always be a candidate for consideration as a solution. It may, in fact, be the most general type of realization that one is able to analyze, but this will depend on the behavior of the physical channel.

The communication systems we have been considering are point-to-point systems, where all information originates at a single point and is to he ultimately received at a single point physically removed from the first. In what follows, we shall take the discrete information theoretic point of view and always assume the discrete channel to be given.

CHAPTER 2
MODELS, DUALITY, AND SOME BASIC THEOREMS

A. PARALLEL CHANNEL MODELS

To model a time discrete channel in the information theoretic sense, we need to define several elements. First, corresponding to the basic unit of signal duration implied by the time discreteness, we define an input space X and an output space Y.† We refer to a member x of X as an input, and to a member y of Y as an output. Let X^N denote the space of sequences of inputs of length N, and Y^N denote the space of sequences of outputs of length N. We denote a member of X^N by $\vec{x}^N = (x_1, \ldots, x_N)$, and a member of Y^N by $\vec{y}^N = (y_1, \ldots, y_N)$. Then, we define a set of conditional probability distributions or densities‡ $p_N(\vec{y}^N/\vec{x}^N)$, N = 1, 2, . . . , on sequences of inputs and outputs of arbitrary length. Sometimes, a channel state variable is introduced into the description to account for the memory of the channel, if any. If the channel is memoryless,

$$p_N(\vec{y}^N/\vec{x}^N) = \prod_{i=1}^{N} p_1(y_i/x_i) \qquad N = 1, 2, \ldots$$

and X, Y, and $p_1(y/x)$ suffice to specify the channel.

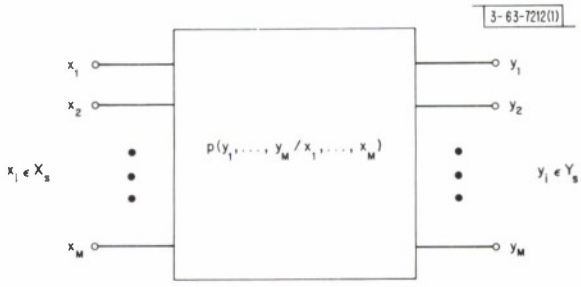


Fig. 3. Conceptual diagram of parallel channels.

The parallel channel models we shall discuss assume that each input x_i is decomposable into M subelements $x_{1i}, \ldots, x_{Mi}$ which we shall call subchannel inputs, and each output y_i is decomposable into M subelements $y_{1i}, \ldots, y_{Mi}$ which we shall call subchannel outputs. We shall assume that the space of the kth subchannel input (output) at "time" i is independent of both k and i and denote it by $X_s(Y_s)$. Now, in general, we can simply substitute $(x_{1i}, \ldots, x_{Mi})$ for x_i, and $(y_{1i}, \ldots, y_{Mi})$ for y_i in the probability distributions which describe the channel and we shall have a description in terms of subchannel inputs and outputs. It will be convenient to make a simplifying assumption which will be in effect throughout most of this report: the channel will be assumed memoryless. Hence, we shall be interested in a channel description given by subchannel input and output spaces X_s and Y_s and probability distributions of the form $p(y_1, \ldots, y_M/x_1, \ldots, x_M)$,§ $x_i \in X_s$, $y_i \in Y_s$ (see Fig. 3).

We have not yet finished imposing structure on our channel. Further structure is desirable in order to model the physical channels we have mentioned in Chapter 1 and to restrict the situations we wish to consider in order to obtain meaningful results. Moreover, we shall provide two structural descriptions (models) of quite different sorts and shall have

† In Chapter 1, we (tacitly) assumed Y = X. Here, we consider a more general situation.

‡ To avoid the tedium of repeating the words "or densities" when the random variables referred to may be either continuous or discrete, this may be assumed unless otherwise stated.

§ If, contrary to what is implied but not required by the notation, $p(y_1, \ldots, y_M/x_1, \ldots, x_M)$ depends on fewer than M subchannel inputs, we have a highly degenerate situation. We do not wish to consider such situations.

something to say about their relation to each other. The first model will be in the form of a set of constraints on the probability distribution $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$; the second will utilize a channel state structure to describe $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$.

1. Model 1 – The MC Channel (M Subchannel, Crosstalkless Channel)

Suppose a set of $M - k$ subchannel inputs and their corresponding outputs are not to be used in communicating. These $M - k$ inputs are set to fixed values. For purposes of communication, we are interested in the conditional probability which relates the k inputs which are used to their corresponding outputs. Denote the unused inputs and outputs by x_{j_ℓ}, y_{j_ℓ} $(\ell = 1, \ldots, M - k)$, and the used inputs and outputs by x_{i_ℓ}, y_{i_ℓ} $(\ell = 1, \ldots, k)$. (Note that $\{j_\ell\}_{\ell=1}^{M-k}$ and $\{i_\ell\}_{\ell=1}^{k}$ are disjoint sets and that their union is the set of integers from 1 to M.) What we wish is $p\left(y_{i_1}, \ldots, y_{i_k} / x_{i_1}, \ldots, x_{i_k}\right)$.

Now,

$$p\left(y_{i_1}, \ldots, y_{i_k} / x_1, \ldots, x_M\right) = \sum_{y_{j_1} \in Y_s} \cdots \sum_{y_{j_{M-k}} \in Y_s} p(y_1, \ldots, y_M / x_1, \ldots, x_M) \quad . \tag{2-1}$$

As the notation implies, the LHS of Eq. (2-1) is, in general, dependent upon all M subchannel inputs $x_1, \ldots, x_M$. However, if for a particular $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$, k and $\{j_\ell\}_{\ell=1}^{M-k}$ the LHS of Eq. (2-1) does not depend on $x_{j_1}, \ldots, x_{j_{M-k}}$, then the values to which these latter are set do not affect the used inputs and outputs in the least. If this is the case,

$$p\left(y_{i_1}, \ldots, y_{i_k} / x_1, \ldots, x_M\right) = p\left(y_{i_1}, \ldots, y_{i_k} / x_{i_1}, \ldots, x_{i_k}\right) \quad . \tag{2-2}$$

We then say that there is no crosstalk between the used and unused subchannels. If this is the case for all $k = 1, \ldots, M - 1$ and all $\{i_\ell\}_{\ell=1}^{k}$, then we refer to our channel as an MC channel. This name is chosen for brevity rather than explicit descriptiveness, because we will need to repeat it often. The condition we have derived can be stated very simply: A parallel channel with M subchannels is an MC channel if a summation† of $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$ over all values of the members of any subset of $\{y_i\}_{i=1}^{M}$ destroys the dependence on the corresponding subset of $\{x_i\}_{i=1}^{M}$.‡

Here, a terminological note is appropriate. We have assumed that the channel input is decomposable into the same number of elements as the channel output. Hence, we may refer to a pair consisting of an input subelement and its corresponding output subelement as a subchannel. The correspondence we speak of is only clear if the channel is an MC channel. In fact, we may

---

† An integration is required if $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$ is a density. We shall not bother to state this explicitly again.

‡ In the discussion above, the values of $x_{i_1}, \ldots, x_{i_{M-k}}$ need not be considered fixed. We could have assumed at the beginning of the description of the MC channel that k subchannels were used by user A and the remaining $M - k$ by user B. If it is desired that user B's input not affect user A's output, we get our MC channel model.

say that a channel is an MC channel if and only if there is some way to pair the input and output subelements so that Eq. (2-2) is true for all k, $1 \leqslant k \leqslant M-1$, and all $\{i_\ell\}_{\ell=1}^{k}$, $1 \leqslant i_\ell \leqslant M$, $1 \leqslant \ell \leqslant k$. The pairing need not be unique.

We summarize with the following definition: A memoryless channel consisting of M subchannels each with input space X_s and output space Y_s and characterized by the conditional probability $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$ is an <u>MC channel</u> if for each k, $1 \leqslant k \leqslant M-1$, and for each $\{i_\ell\}_{\ell=1}^{k}$, $1 \leqslant i_\ell \leqslant M$, $1 \leqslant \ell \leqslant k$,

$$p\left(y_{i_1}, \ldots, y_{i_k} / x_1, \ldots, x_M\right) = p\left(y_{i_1}, \ldots, y_{i_k} / x_{i_1}, \ldots, x_{i_k}\right) \quad . \qquad \text{[Eq. (2.2)]}$$

Although the definition of MC channel we have used assures us that disjoint sets of subchannels are mutually noninterfering however they are composed, the verification of the MC property is rather tedious if the number of subchannels is large. In fact, the number z of different equations of the form of Eq. (2-2) which must be satisfied is given by

$$z = \sum_{k=1}^{M-1} \binom{M}{k} = 2^M - 2 \quad . \tag{2-3}$$

Fortunately, this number can be reduced to M by making use of the following theorem.

Theorem 2.1.

A memoryless channel consisting of M subchannels each with input space X_s and output space Y_s and characterized by the conditional probability $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$ is an MC channel if and only if for each i, $1 \leqslant i \leqslant M$.

$$\begin{aligned} &p(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_M / x_1, \ldots x_M) \\ &\quad = p(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_M / x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_M) \quad . \end{aligned} \tag{2-4}$$

Proof.

(1) Necessity is proved simply by observing that Eq. (2-4) is equivalent to Eq. (2-2) for $k = M-1$ and $i_1, \ldots, i_k$ distinct.

(2) To prove sufficiency, pick k, $1 \leqslant k \leqslant M$. Let $\{i_\ell\}_{\ell=1}^{k}$ be a set of k distinct integers each satisfying $1 \leqslant i_\ell \leqslant M$. Let $\{j_\ell\}_{\ell=1}^{M-k}$ consist of the remaining $(M-k)$ integers satisfying $1 \leqslant j_\ell \leqslant M$. Recall

$$\begin{aligned} &p\left(y_{i_1}, \ldots, y_{i_k} / x_1, \ldots, x_M\right) \\ &\quad = \sum_{y_{j_1} \in Y_s} \cdots \sum_{y_{j_{M-k}} \in Y_s} p(y_1, \ldots, y_M / x_1, \ldots, x_M) \quad . \qquad \text{[Eq. (2-1)]} \end{aligned}$$

If we assume the theorem is false, then for some integer q, $1 \leqslant q \leqslant M-k$, the LHS of Eq. (2-1) depends on x_{j_q}. Now the sums on the RHS of Eq. (2-1) may be formed in any order.† Hence,

† This is true even if Y_s is not finite; see, for example, W. Rudin, <u>Principles of Mathematical Analysis</u>, 2nd edition (McGraw-Hill, New York, 1964), Theorem 8.3. If $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$ is a conditional density and integrals replace sums, then the Fubini Theorem allows us to integrate in any order; see, for example, H. L. Royden, <u>Real Analysis</u> (Macmillan, New York, 1963), p. 233.

$$p\left(y_{i_1}, \ldots, y_{i_k} / x_1, \ldots, x_M\right)$$

$$= \sum_{y_{j_1}} \cdots \sum_{y_{j_{q-1}}} \sum_{y_{j_{q+1}}} \cdots \sum_{y_{j_{M-k}}} \sum_{y_{j_q}} p(y_1, \ldots, y_M / x_1, \ldots, x_M) \quad . \tag{2-5}$$

From Eq. (2-4),

$$p\left(y_{i_1}, \ldots, y_{i_k} / x_1, \ldots, x_M\right) = \sum_{y_{j_1}} \cdots \sum_{y_{j_{q-1}}} \sum_{y_{j_{q+1}}} \cdots \sum_{y_{j_{M-k}}}$$

$$p\left(y_1, \ldots, y_{j_{q-1}}, y_{j_{q+1}}, \ldots, y_M / x_1, \ldots, x_{j_{q-1}}, x_{j_{q+1}}, \ldots, x_M\right) \quad . \tag{2-6}$$

Since each summand on the RHS of Eq. (2-6) is independent of x_{j_q}, the sum, and hence the LHS of Eq. (2-6), is independent of x_{j_q}. Thus, we have a proof by contradiction.

We note that in the MC channel model, although we have defined subchannel inputs and outputs, the subchannels themselves are not identifiable. Our second model will have identifiable subchannels.

2. Model 2 – The MS Channel (M Subchannel, State Description Channel)

Suppose we have a set of M subchannels each of which may be in one of a number of states. We call the set of subchannel states Λ. Associated with each $\alpha \in \Lambda$, there is a subchannel conditional probability distribution† $p_\alpha(\xi/\eta)$, $\xi \in Y_s$, $\eta \in X_s$. We let α_i denote the state of the ith subchannel, and $\vec{\alpha} = (\alpha_1, \ldots, \alpha_M)$ denote the state of the (whole) channel consisting of M subchannels. We call $\vec{\alpha}$ the channel state vector. We assume that a probability distribution‡ $p(\alpha_1, \ldots, \alpha_M)$ on the subchannel states is given. This is equivalent to a distribution $p(\vec{\alpha})$ on the (whole) channel state. Let§

$$p(y_1, \ldots, y_M / x_1, \ldots, x_M)$$

$$= \sum_{\alpha_1 \in \Lambda} \cdots \sum_{\alpha_M \in \Lambda} p(\alpha_1, \ldots, \alpha_M)\, p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M) \quad . \tag{2-7}$$

If we write $p(y/x)$ for $p(y_1, \ldots, y_M/x_1, \ldots, x_M)$ and $p_{\vec{\alpha}}(y/x)$ for $p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M)$, then Eq. (2-7) can be written in the more condensed form

$$p(y/x) = \sum_{\vec{\alpha} \in \Lambda^M} p(\vec{\alpha})\, p_{\vec{\alpha}}(y/x) \quad . \tag{2-8}$$

---

† These may be densities.

‡ This may be a density.

§ If $p(a_1, \ldots, a_M)$ is a density, the sums over $a_1, \ldots, a_M$ become integrals.

A memoryless channel consisting of M subchannels each with input space X_s and output space Y_s and characterized by the conditional probability $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$ defined by Eq. (2-7) is called an MS channel (see Fig. 4).

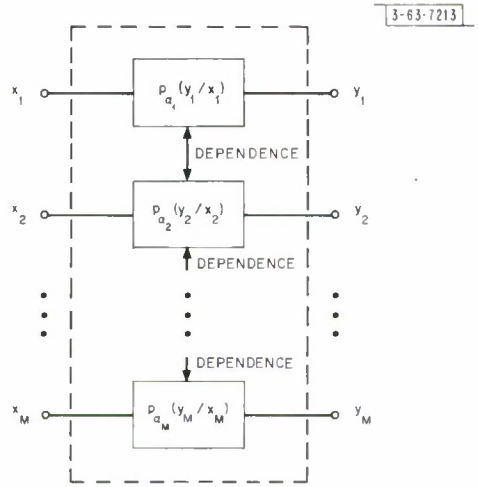


Fig. 4. The MS channel.

$$p(y_1, \ldots, y_M / x_1, \ldots, x_M) = \sum_{\alpha_1 \in \Lambda} \cdots \sum_{\alpha_M \in \Lambda} p(\alpha_1, \ldots, \alpha_M)\, p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M)$$

Suppose that we chose instead the apparently more general model where the sets Λ_i, $i = 1, \ldots, M$ of subchannel states were allowed to be different. In fact, letting $\Lambda = \bigcup_{i=1}^{M} \Lambda_i$, we can represent this situation as an MS channel. Hence, these two models are equivalent and we have chosen the one which is notationally a trifle more simple.

Some relations between the MC and MS channels will now be given. Referring to Eq. (2-7), one may see that a summation of any y_i over Y_s destroys the dependence of the summed expression on x_i $\left[\text{because } \sum_{y_i \in Y_s} p_{\alpha_i}(y_i/x_i) = 1, \text{ for all } \alpha_i \in \Lambda, \text{ and } x_i \in X_s\right]$. Hence, by Theorem 2.1, we see immediately that every MS channel is an MC channel.

Although the fact is somewhat surprising, it is not true that every MC channel is an MS channel. A counterexample is discussed in Appendix A.

The emphasis in this work will be on MS rather than on MC channels, which latter are of doubtful engineering interest when they cannot be modeled as MS channels. We shall, however, assume the more general MC channel model when a result follows naturally from this assumption.

B. DUALITY BETWEEN TIME AND PARALLEL DIRECTIONS

The mathematical descriptions of a memoryless parallel channel bear a strong resemblance to those of a single channel with memory. In both cases, we start with base spaces X_s and Y_s for the basic indecomposable inputs and outputs. An input to or output of an MS or MC channel is a member of the product space X_s^M or Y_s^M. Similarly, if we have but a single channel, a sequence of its inputs or outputs of length N is a member of X_s^N or Y_s^N. The MS and MC channels are characterized by a conditional probability distribution defined over $Y_s^M \times X_s^M$. Insofar as one is interested only in transmitting a sequence of inputs of length N,† a single channel with

† N may be the block length of a block code, or large enough so that N times the basic unit of signal duration is equal to the lifetime of the equipment.

memory is sufficiently characterized by a conditional probability distribution defined over $Y_s^N \times X_s^N$. This is not to say that there may not exist simpler characterizations in particular cases.

The purpose of mentioning the duality between time and parallel directions is twofold: first, it enables us to borrow results obtained for single channels, possibly with memory, to use for our parallel channel model; second, some of the results obtained here will apply to single channels with memory.

We may define channels with memory to correspond to our MC channels: a channel with no intersymbol interference (N II channel) is defined as one for which given any integer i, $1 \leqslant i \leqslant N$,

$$\sum_{y_i \in Y_s} p(y_1, \ldots, y_N / x_1, \ldots, x_N)$$

$$= p(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N / x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N) \quad . \tag{2-9}$$

The correspondence between N II and MC channels is made clear by Theorem 2.1. We may also define the time analog of the MS channel: let Λ be a set of channel states; $p_\alpha(\xi/\eta)$, $\xi \in Y_s$, $\eta \in X_s$, $\alpha \in \Lambda$, a conditional probability; α_i denote the channel state at the ith time instant and $p(\alpha_1, \ldots, \alpha_N)$ the probability distribution over channel state sequences of length N. Let

$$p(y_1, \ldots, y_N / x_1, \ldots, x_N)$$

$$= \sum_{\alpha_1 \in \Lambda} \cdots \sum_{\alpha_N \in \Lambda} p(\alpha_1, \ldots, \alpha_N)\, p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_N}(y_N/x_N) \quad . \tag{2-10}$$

A channel with conditional probability distribution given by Eq. (2-10) is an MST channel.

Now, all MST channels are N II for the same reason that MS channels are MC. All N II channels are <u>not</u> MST. Basically, the same counterexample which was used to show that all MC channels are not MS can be used here. (See Appendix A.)

For N II channels, if k, ℓ are chosen so that $1 \leqslant k < \ell \leqslant N$, then $p(y_k, \ldots, y_\ell / x_k, \ldots, x_\ell)$ is independent of the input distribution $p(x_1, \ldots, x_N)$ and is characteristic of the channel alone.† Thus, for purposes of block coding, we may take N as the block length and obtain a sufficient and unique characterization of the channel.‡

For N II channels, we may also define stationarity. An N II channel is stationary if for any integers j, k, and ℓ satisfying $0 \leqslant j \leqslant j + \ell \leqslant N$, $0 \leqslant k \leqslant k + \ell \leqslant N$ we have

$$p(y_{j+1}, \ldots, y_{j+\ell} / x_{j+1}, \ldots, x_{j+\ell}) = p(y'_{k+1}, \ldots, y'_{k+\ell} / x'_{k+1}, \ldots, x'_{k+\ell})$$

whenever

$$y'_{k+i} = y_{j+i} \qquad 1 \leqslant i \leqslant \ell$$

and

$$x'_{k+i} = x_{j+i} \qquad 1 \leqslant i \leqslant \ell \quad . \tag{2-11}$$

† This statement will appear subsequently in its "parallel" form. It may be proved by applying Eq. (2-9) and Theorem 2.1 to the calculation of $p(y_k, \ldots, y_\ell / x_k, \ldots, x_\ell)$.

‡ Since no stationarity condition has been imposed, there is no guarantee that the channel will remain the same from block to block.

In other words, only the length and values of an input-output sequence matter, not the starting point.

For MST channels, if the distribution of α_i's is stationary, then the channel will be stationary as well.

C. MUTUAL INFORMATION AND CAPACITY

We now wish to examine the effect of subchannel dependencies on the mutual information between input and output of a channel with independent subchannel inputs, and to compare this mutual information with the mutual information between input and output of the individual subchannels. A comparison is implied between an original channel with dependencies and a derived channel without them. Suppose we are given a channel with subchannel inputs $x_1, \ldots, x_M$, subchannel outputs $y_1, \ldots, y_M$, and conditional probability distribution $p(y_1, \ldots, y_M / x_1, \ldots, x_M)$. Suppose, too, that we are given an input probability distribution

$$p(x_1, \ldots, x_M) = \prod_{i=1}^{M} p_i(x_i)$$

where $p_i(\xi)$ is a single-subchannel input distribution, $i = 1, \ldots, M$. Furthermore, let $\sum_{\overline{X}_i}$ and $\sum_{\overline{Y}_i}$ denote summation over all but the ith input and output, respectively. Then, the single-subchannel conditional probability distributions $p_i(y_i/x_i)$, $1 \leqslant i \leqslant M$, are given by

$$p_i(y_i/x_i) = \frac{1}{p_i(x_i)} \sum_{\overline{X}_i} \sum_{\overline{Y}_i} p(y_1, \ldots, y_M / x_1, \ldots, x_M)\, p(x_1, \ldots, x_M) \quad . \tag{2-12}$$

We define a dependence-removed (DR) channel with conditional probability distribution given by

$$p_{DR}(y_1, \ldots, y_M / x_1, \ldots, x_M) = \prod_{i=1}^{M} p_i(y_i/x_i) \quad . \tag{2-13}$$

Theorem 2.2.

Taking the usual definition of mutual information†

$$I(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M) \geqslant \sum_{i=1}^{M} I(X_i;\ Y_i) \quad . \tag{2-14}$$

If we denote the mutual information between input and output of the dependence-removed channel by $I_{DR}(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M)$, we have

$$\sum_{i=1}^{M} I(X_i;\ Y_i) = I_{DR}(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M) \quad . \tag{2-15}$$

---

† The use of X_i or Y_i, $i = 1, \ldots, M$, as an argument of the informational expressions implies that on expression which is a function of x_i or y_i is averaged over X_s or Y_s.

Proof.

First we note that[1†]

$$I(X_i;\ Y_i) = H(X_i) - H(X_i/Y_i) \qquad i = 1, \ldots, M \tag{2-16}$$

and

$$I(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M) = H(X_1, \ldots, X_M) - H(X_1, \ldots, X_M/Y_1, \ldots, Y_M) \quad . \tag{2-17}$$

These expressions hold for both discrete and continuously distributed variables. Since the sub-channel inputs are independently distributed,[2]

$$H(X_1, \ldots, X_M) = \sum_{i=1}^{M} H(X_i) \quad . \tag{2-18}$$

We have also that

$$H(X_1, \ldots, X_M/Y_1, \ldots, Y_M) = H(X_1/Y_1, \ldots, Y_M) + H(X_2/Y_1, \ldots, Y_M, X_1)$$
$$+ \ldots + H(X_M/Y_1, \ldots, Y_M, X_1, \ldots, X_{M-1}) \quad .$$

Since for any random variables[3] U, V, W,

$$H(U/VW) \leqslant H(U/V)$$

we obtain the inequality

$$H(X_1, \ldots, X_M/Y_1, \ldots, Y_M) \leqslant \sum_{i=1}^{M} H(X_i/Y_i) \quad . \tag{2-19}$$

Hence, combining Eqs. (2-16) through (2-19), we get

$$I(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M) \geqslant \sum_{i=1}^{M} I(X_i;\ Y_i)$$

as required. The proof of Eq. (2-15) is an immediate consequence of the definition of I_{DR}.

Note that for an MC channel the values of $p_i(y_i/x_i)$ computed from Eq. (2-12) are independent of the input distribution $p(x_1, \ldots, x_M)$. Hence, corresponding to each MC channel there is a unique dependence-removed channel. This is not true, in general.

In the remainder of this report, we will have frequent need to compare constants defined as maxima of functions of several variables and to compare functions of one variable defined as maxima over the remaining variables of functions of several variables. This comparison, which will usually take the form of an inequality between two non-negative quantities, will be facilitated by the two theorems which will be stated below. First, it will be necessary to explain some notation and give a definition.

A probability measure over an input space consisting of a finite number K of points can be represented as a vector $\vec{p}$ in K-dimensional Euclidean space $\Re$.

---

† Numbered references appear at the end of each chapter.

Suppose we have a parallel channel consisting of M subchannels. If $p_i(x_i)$, $i = 1, \ldots, M$ are probability distributions over X_s, then $\prod_{i=1}^{M} p_i(x_i)$ is a <u>product distribution</u> over $X = X_s^M$.

Theorem 2.3.

Let P be the set of probability distributions over a finite† product space X_s^M, and D the set of product distributions over X_s^M. Let ρ and R be real variables with $0 \leqslant \rho \leqslant 1$, and $\vec{p} \in P$. Let f and g each be continuous real valued functions of ρ, $\vec{p}$, and R. Define

$$F_D(R) = \max_{\substack{0 \leqslant \rho \leqslant 1 \\ \vec{p} \in D}} f(\rho, \vec{p}, R)$$

$$G_D(R) = \max_{\substack{0 \leqslant \rho \leqslant 1 \\ \vec{p} \in D}} g(\rho, \vec{p}, R)$$

$$F_P(R) = \max_{\substack{0 \leqslant \rho \leqslant 1 \\ \vec{p} \in P}} f(\rho, \vec{p}, R)$$

$$G_P(R) = \max_{\substack{0 \leqslant \rho \leqslant 1 \\ \vec{p} \in P}} g(\rho, \vec{p}, R) \quad .$$

Then,

(1) $F_D(R)$, $G_D(R)$, $F_P(R)$, and $G_P(R)$ are all finite.

(2) Suppose $f \leqslant g$ for all $0 \leqslant \rho \leqslant 1$ and $\vec{p} \in D$; then

$$F_D(R) \leqslant G_D(R) \leqslant G_P(R) \tag{2-20}$$

and $F_D(R) < G_D(R)$ if $f < g$ for all $0 \leqslant \rho \leqslant 1$ and $\vec{p} \in D$.

(3) Suppose $f \leqslant g$ for all $0 \leqslant \rho \leqslant 1$ and $\vec{p} \in P$; then Eq. (2-20) holds and, in addition,

$$F_D(R) \leqslant F_P(R) \leqslant G_P(R) \quad . \tag{2-21}$$

Furthermore, $F_P(R) < G_P(R)$ if $f < g$ for all $0 \leqslant \rho \leqslant 1$ and $\vec{p} \in P$.

The proof is given in Appendix B.

Theorem 2.4.

Let P be the set of probability distributions over a product space X_s^M, and D the set of product distributions over X_s^M. Let ρ and R be real variables with $0 \leqslant \rho \leqslant 1$ and $p \in P$. Let f and g each be real valued functions (functionals) of ρ, p, and R. Define

† This implies that X_s is a finite set, and M a finite positive integer.

$$F_D(R) = \underset{\substack{0 \leqslant \rho \leqslant 1 \\ p \in D}}{\text{l. u. b.}} \; f(\rho, p, R)$$

$$G_D(R) = \underset{\substack{0 \leqslant \rho \leqslant 1 \\ p \in D}}{\text{l. u. b.}} \; g(\rho, p, R)$$

$$F_P(R) = \underset{\substack{0 \leqslant \rho \leqslant 1 \\ p \in P}}{\text{l. u. b.}} \; f(\rho, p, R)$$

$$G_P(R) = \underset{\substack{0 \leqslant \rho \leqslant 1 \\ p \in P}}{\text{l. u. b.}} \; g(\rho, p, R) \quad .$$

Then,

(1) If $f \leqslant g$ for all $0 \leqslant \rho \leqslant 1$ and $p \in D$, we have

$$F_D(R) \leqslant G_D(R) \leqslant G_P(R) \quad . \tag{2-22}$$

(2) If $f \leqslant g$ for all $0 \leqslant \rho \leqslant 1$ and $p \in P$, Eq. (2-22) holds and, in addition,

$$F_D(R) \leqslant F_P(R) \leqslant G_P(R) \quad . \tag{2-23}$$

The proof is given in Appendix B.

Theorem 2.5.

Suppose we are given an MC channel, with X_s finite. Denote the capacity of the dependence-removed channel by C_{DR}, the capacity of the ith subchannel [with conditional probability distribution given by Eq. (2-12)] by C_i, and the capacity of the original channel (consisting of M subchannels) by C. Then,

$$C \geqslant \sum_{i=1}^{M} C_i = C_{DR} \quad . \tag{2-24}$$

Proof.

Since the $p_i(y_i/x_i)$ are unique, the verification of Eq. (2-24) is straightforward. First, the capacity of the dependence-removed channel is achieved with independent[4] subchannel inputs (i.e., a product distribution maximizes the mutual information). Hence, the equality part of Eq. (2-24) comes from Eq. (2-15). Let us use a superscript $\vec{p}$ to make explicit the input distribution which is involved in the calculation of mutual information between input and output of our channel. Then, since

$$C = \max_{\vec{p} \in P} I^{\vec{p}}(X_1, \ldots, X_M; \; Y_1, \ldots, Y_M)$$

and

$$C_{DR} = \max_{\vec{p} \in D} I^{\vec{p}}_{DR}(X_1, \ldots, X_M; \; Y_1, \ldots, Y_M)$$

Eqs. (2-14), (2-15), and Theorem 2.3 give us the inequality part of Eq. (2-24).

Theorem 2.6.

Suppose we are given an MC channel. Let C, C_{DR}, and C_i be as above. Then, Eq. (2-24) holds.

Proof.

The proof is analogous to that of Theorem 2.5. Theorem 7.2.2 replaces Theorem 4.2.1 in Ref. 4, and our Theorem 2.4 replaces Theorem 2.3.

Hence, whether or not X_s is finite, the capacity of an MC channel cannot be increased by removal of its dependencies.

Obviously, this result applies to N II channels as well. However, if it is the capacity per use c of an N II channel that we are concerned with, conditional probability distributions are defined on $Y_s^N \times X_s^N$ for all positive integers N, and we define

$$c = \lim_{N\to\infty} \frac{C}{N}$$

and

$$c_{DR} = \lim_{N\to\infty} \frac{C_{DR}}{N} \quad .$$

Then,

$$c \geqslant \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} C_i = c_{DR} \tag{2-25}$$

so that the capacity per use of an N II channel is not decreased by removal of its dependencies. If the channel is stationary, the middle expression in Eq. (2-25) is simply the capacity for one-shot use of the channel.

To abbreviate the description of channel examples in the remainder of this report, we will agree to call an MC channel with M subchannels, $X_s = \{1, \ldots, L\}$ and $Y_s = \{1, \ldots, Q\}$, an <u>$M \times L \times Q$ channel</u>.

Both the inequality of Eq. (2-14) and the inequality part of Eq. (2-24) may be strict. This can be shown by the following example of a $2 \times 2 \times 2$ channel:

$$\begin{array}{cc|c|c|c|c|}
 & & \multicolumn{4}{c}{x_1x_2} \\
 & & 00 & 01 & 10 & 11 \\
\hline
 & 00 & 5/8 & 1/8 & 1/8 & 1/8 \\
\hline
y_1y_2 & 01 & 1/8 & 5/8 & 1/8 & 1/8 \\
\hline
 & 10 & 1/8 & 1/8 & 5/8 & 1/8 \\
\hline
 & 11 & 1/8 & 1/8 & 1/8 & 5/8 \\
\hline
\end{array} \quad p(y_1y_2/x_1x_2)$$

This example is MS† as well as MC. Capacity (0.452 bit) is achieved by the following input distribution‡ $p(x_1, x_2)$:

$$p(00) = p(01) = p(10) = p(11) = 1/4 \quad . \tag{2-26}$$

The dependence-removed channel is:

$p_{DR}(y_1y_2/x_1x_2)$

y_1y_2 \ x_1x_2	00	01	10	11
00	9/16	3/16	3/16	1/16
01	3/16	9/16	1/16	3/16
10	3/16	1/16	9/16	3/16
11	1/16	3/16	3/16	9/16

C_{DR}(0.378 bit) is also achieved by the input distribution Eq. (2-26).‡ Hence, the inequality part of Eq. (2-24) may be strict. Since Eq. (2-26) is a product distribution, both C and C_{DR} are achieved with independent inputs. Thus, the hypothesis for Theorem 2.2 is obeyed and the inequality in Eq. (2-14) may be strict as well. A continuity argument shows that Eq. (2-14) may still hold for dependent subchannel inputs.

The example we have been discussing may also be used to show that for dependent subchannel inputs neither Eqs. (2-14) nor (2-15) need hold. For the input distribution $p(x_1, x_2)$ given by

$$p(00) = p(11) = 1/2$$
$$p(01) = p(10) = 0$$

we have

$$I(X_1, X_2;\ Y_1, Y_2) = 0.262 \text{ bit}$$

$$\sum_{i=1}^{2} I(X_i;\ Y_i) = 0.379 \text{ bit}$$

$$I_{DR}(X_1, X_2;\ Y_1, Y_2) = 0.329 \text{ bit} \quad .$$

One might be tempted to conjecture that an MC channel always achieves capacity for independent subchannel inputs. The following example of a $2 \times 2 \times 2$ channel disproves this conjecture:

$p(y_1y_2/x_1x_2)$

y_1y_2 \ x_1x_2	00	01	10	11
00	0.5	0	0	0
01	0	0.5	0	0
10	0	0	0.5	0
11	0.5	0.5	0.5	1

† See Example 1 in Chapter 3, p. 17.

‡ Theorem 4.5.1 of Ref. 4 provides the means of proof.

One may verify that this is an MC channel (it is MS as well). Capacity (0.806 bit) is achieved only by the following distribution† $p(x_1, x_2)$:

$$p(00) = p(01) = p(10) = 2/7 \quad ; \quad p(11) = 1/7 \quad .$$

Since $p_1(x_1)$ and $p_2(x_2)$ are both given by $p(0) = 4/7$ and $p(1) = 3/7$, capacity is not achieved by independent subchannel inputs.

Since Theorems 2.2 and 2.5 deal with channels without a prescribed state structure, they naturally have nothing to say about the situation where the channel state is known to the receiver. Suppose, however, we are given an MS channel. If we consider the "output" in the state known case to be a doublet $(y, \vec{\alpha}) \in (Y, \Lambda^M)$, then we see that this is just a special case of the general channel with state unknown to the receiver. Furthermore, since the channel state is independent of the input and the conditional probability distribution corresponding to a single channel state (a product distribution) satisfies the MC constraints, the channel with doublet output is MC. Hence, Theorems 2.2 and 2.5 hold for MS channels whether or not the channel state is known to the receiver.‡

We conclude Chapter 2 with a theorem which applies only to the situation where the channel state is known to the receiver.

Theorem 2.7.

For an MS channel whose state is known to the receiver during each transmission, the channel capacity is equal to the sum of the individual subchannel capacities.

Proof.

If the receiver knows the channel state, the applicable conditional probability for the channel, corresponding to the state $\vec{\alpha}$, is

$$p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M) \quad .$$

Thus, the MS channel with state known at the receiver is already a dependence-removed channel. We have from Eq. (2-15)

$$\sum_{i=1}^{M} I_{p_i^*}^{\alpha_i}(X_i;\ Y_i) = I_{p^*}^{\vec{\alpha}}(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M) \tag{2-27}$$

where the channel state $\vec{\alpha}$, and input (product) distribution

$$p^*(x_1, \ldots, x_M) = \prod_{i=1}^{M} p_i^*(x_i)$$

are now both made explicit parameters of the informational expressions. Averaging over the channel states, we have

† Theorem 4.5.1 of Ref. 4 shows that the distribution given yields capacity. Corollary 2 to this theorem states that there is a unique output distribution corresponding to capacity. Since the transition matrix for the channel is nonsingular, the input distribution must be unique as well.

‡ There is no difficulty posed by the fact that we use an augmented output in our definition of mutual information. Since channel state is independent of input, $I(X;\ Y\Lambda^M) = I(X;\ Y/\Lambda^M)$.

$$\sum_{\alpha_1 \in \Lambda} \cdots \sum_{\alpha_M \in \Lambda} p(\alpha_1, \ldots, \alpha_M) \sum_{i=1}^{M} I_{p_i^*}^{\alpha_i}(X_i;\ Y_i)$$

$$= \sum_{\vec{\alpha} \in \Lambda^M} p(\vec{\alpha})\, I_{p^*}^{\vec{\alpha}}(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M) \tag{2-28}$$

which becomes

$$\sum_{i=1}^{M} \sum_{\alpha_i \in \Lambda} p_i(\alpha_i)\, I_{p_i^*}^{\alpha_i}(X_i;\ Y_i) = \sum_{\vec{\alpha} \in \Lambda^M} p(\vec{\alpha})\, I_{p^*}^{\vec{\alpha}}(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M) \quad . \tag{2-29}$$

Given a dependence-removed channel, the mutual information between input and output for any joint distribution on the subchannel inputs is never greater than that corresponding to the product distribution with the same single-subchannel marginal distributions as the original joint distribution.† Hence, for the purpose of discussing capacity, we need only consider independent subchannel inputs. The capacity C' of the MS channel with state known at the receiver is then obtained by maximizing the RHS of Eq. (2-29) over all product input distributions p^*. Hence,

$$C' = \max_{p^*} \sum_{\vec{\alpha} \in \Lambda^M} p(\vec{\alpha})\, I_{p^*}^{\vec{\alpha}}(X_1, \ldots, X_M;\ Y_1, \ldots, Y_M)$$

$$= \max_{p^*} \sum_{i=1}^{M} \sum_{\alpha_i \in \Lambda} p_i(\alpha_i)\, I_{p_i^*}^{\alpha_i}(X_i;\ Y_i)$$

$$= \sum_{i=1}^{M} \max_{p_i^*} \sum_{\alpha_i \in \Lambda} p_i(\alpha_i)\, I_{p_i^*}^{\alpha_i}(X_i;\ Y_i) \quad . \tag{2-30}$$

But, the last expression in Eq. (2-30) is clearly just a sum of individual subchannel capacities C_i', with the state known at the receiver. Hence,

$$C' = \sum_{i=1}^{M} C_i' \tag{2-31}$$

as required.

REFERENCES

1. R.M. Fano, Transmission of Information (M.I.T. Press/Wiley, New York, 1961), Eqs. (2.119) and (2.150).
2. Ibid., Eqs. (2.98), (2.100), and (2.101).
3. Ibid., Eq. (2.105).
4. R.G. Gallager, Information Theory and Reliable Communication (Wiley, New York, 1968), Theorem 4.2.1.

† See proof of Theorem 4.2.1 in Ref. 4.

CHAPTER 3
STATE REPRESENTATIONS AND BOUNDS FOR MUTUAL INFORMATION AND PROBABILITY OF ERROR

A. STATE REPRESENTATIONS

The formula for the conditional probability of the output $y_1, \ldots, y_M$ of an MS channel given the input $x_1, \ldots, x_M$ is given by Eq. (2-7) which is rewritten below:

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M) = \sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M) \times p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M) \quad . \tag{3-1}$$

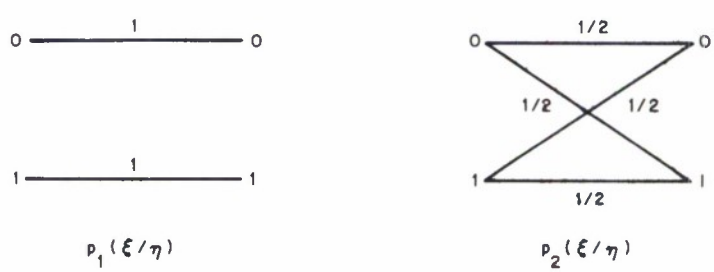
The definitions of Chapter 2 apply to all the expressions in this formula (see p. 6). Suppose we are given a conditional probability distribution $p(y_1, \ldots, y_M/x_1, \ldots, x_M)$ which can be expressed in the form of the RHS of Eq. (3-1) and is therefore the conditional probability distribution associated with an MS channel. Is the representation unique or are there other sets of subchannel conditional probabilities $r_\gamma(\xi/\eta)$, $\gamma \epsilon \Gamma$, and probability distributions $p(\gamma_1, \ldots, \gamma_M)$ such that

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M) = \sum_{\gamma_1 \epsilon \Gamma} \cdots \sum_{\gamma_M \epsilon \Gamma} p(\gamma_1, \ldots, \gamma_M) \times r_{\gamma_1}(y_1/x_1) \cdots r_{\gamma_M}(y_M/x_M) \quad ?$$

The answer to this question is that the representation Eq. (3-1) is <u>not</u>, in general, unique. This may best be shown by an example.

<u>Example 1</u>

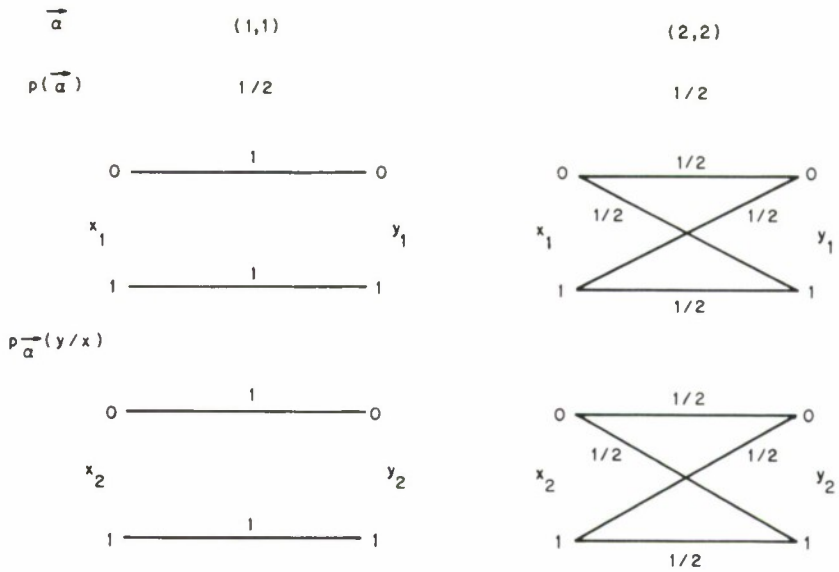
Let $p_1(\xi/\eta)$ and $p_2(\xi/\eta)$ be given below:



Let $p(\alpha_1, \alpha_2)$ be given by

$$p(1,1) = p(2,2) = 1/2 \qquad p(1,2) = p(2,1) = 0 \quad .$$

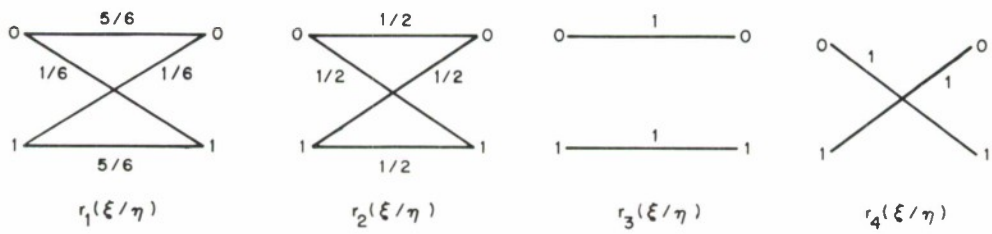
Then the representation for $p(y_1 y_2/x_1 x_2)$ can itself be represented by the diagram:



$p(y_1y_2/x_1x_2)$ can be represented by the matrix

y_1y_2 \ x_1x_2	00	01	10	11
00	5/8	1/8	1/8	1/8
01	1/8	5/8	1/8	1/8
10	1/8	1/8	5/8	1/8
11	1/8	1/8	1/8	5/8

Now, let $r_1(\xi/\eta), \ldots, r_4(\xi/\eta)$ be given below:



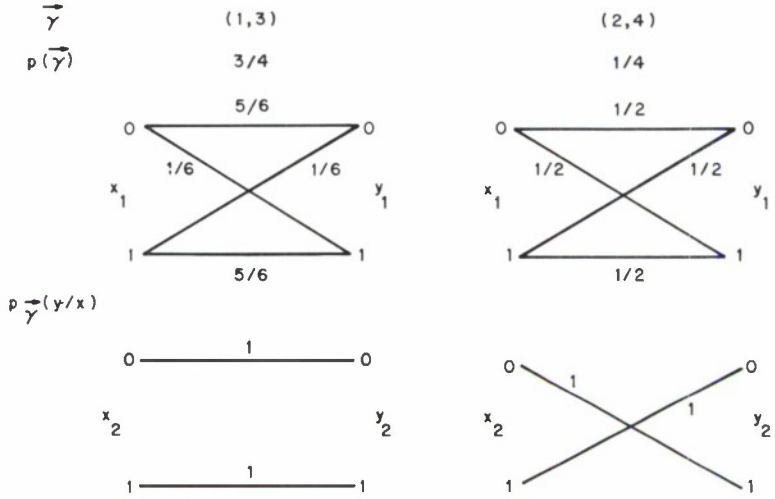
Let $p(\gamma_1, \gamma_2)$ be given by

$$p(1, 3) = 3/4$$

$$p(2, 4) = 1/4$$

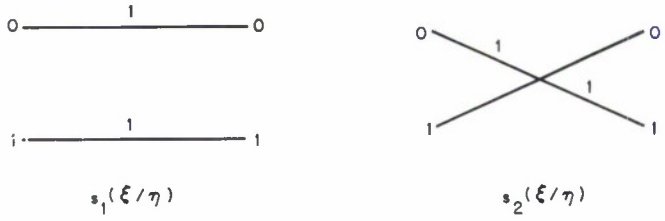
$$p(\gamma_1, \gamma_2) = 0 \qquad \text{unless } (\gamma_1, \gamma_2) = (1, 3) \text{ or } (2, 4) \quad .$$

The diagram for the representation is:



One may check directly that this new representation yields the same probability distribution $p(y_1 y_2/x_1 x_2)$, and hence the same matrix, as is given above.

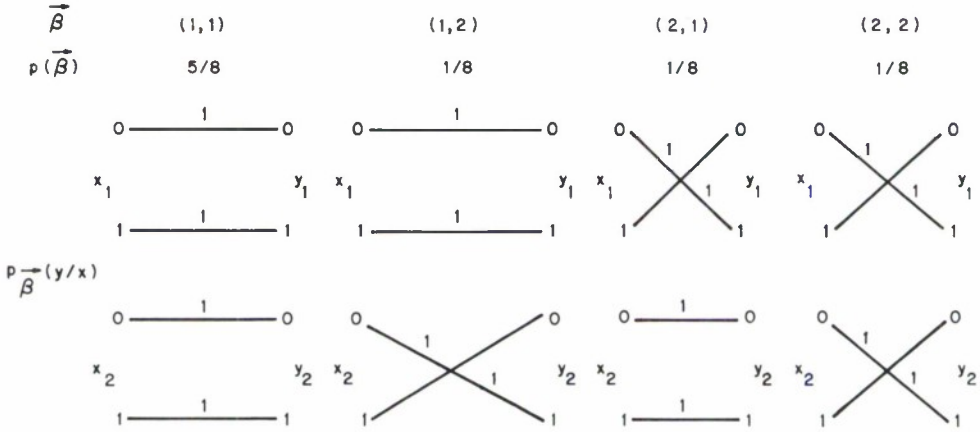
To conclude this example, we give yet another representation for the $p(y_1 y_2/x_1 x_2)$ given by the matrix above. Let $s_1(\xi/\eta)$ and $s_2(\xi/\eta)$ be given below:



Let $p(\beta_1, \beta_2)$ be given by

$$p(1,1) = 5/8 \qquad p(1,2) = p(2,1) = p(2,2) = 1/8 \quad .$$

The diagram for the representation is given below:



We note that the channel states used here are "pure" channels, i.e., given the input and the channel state, the output is completely determined. The "burden of randomness" is placed entirely on the channel state probability distribution. Thus, in some sense, this last representation is a "canonic" representation. The existence of canonic representations is not peculiar to the channel example given above. For any MS channel with finite input and output alphabets, there always exists a representation for which in each channel state the subchannel conditional probabilities are all either 0 or 1, and hence for which the channel state probability distribution supplies all the randomness. This fact is proven in Appendix D. There is often more than one canonic representation for a given MS channel.

B. ENTROPY

The notation of channel representation leads naturally to the idea of channel entropy. If the input-output statistics of a channel are given by an expression of the form in Eq. (3-1), one might wish to define the entropy H of the channel representation by the formula†

$$H = -\sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M) \log p(\alpha_1, \ldots, \alpha_M) \quad . \tag{3-2}$$

If we compute the entropies of the three representations in Example 1, we obtain, in order,

$$H_1 = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \text{ bit}$$

$$H_2 = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.811 \text{ bit}$$

$$H_3 = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{1}{8} = 1.549 \text{ bits} \quad .$$

Thus, the entropy of a channel representation is not determined by the channel's input-output conditional probability distribution alone. Therefore, we may not simply associate the quantity given by Eq. (3-2) with the entropy of the channel. We note, however, that the entropy is both non-negative and continuous in $p(\alpha_1, \ldots, \alpha_M)$. Hence, among all representations of the channel, there must be at least one which gives a smallest value for the entropy of the representation.

Despite the problem with uniqueness, the entropy of a channel representation is, in some instances, a simple and natural quantity to use in bounding the mutual information between its input and output. This fact will be demonstrated in the sequel.

C. NATURAL STATE REPRESENTATIONS

It should be clear at this point that it is impossible to decide which channel representation is a "natural" one from the input-output probabilities alone. The naturalness of a channel representation will depend on the relationship between the states $\vec{\alpha}$ of the mathematical model and the processes which take place in the physical channel. The choice of a natural state representation is important because we will often talk about the situation where the receiver has knowledge of the channel state. If the representation is natural, this knowledge may usually be obtained

† We will limit our discussion of channel entropy and its properties to cases where the state distribution is discrete.

through measurement of some physical quantity. Generally, the models we shall use (e.g., the first representation of Example 1) are natural ones for a set of fading subchannels with equal energy orthogonal signaling on each subchannel. The observable which the receiver may use to obtain state (corresponding to depth of fade) knowledge is received signal energy.

D. BOUNDS ON MUTUAL INFORMATION

We now proceed to derive some relations involving the mutual information between the input and output of a channel with discrete states. The channel is not necessarily MS.

Theorem 3.1.

Suppose we have an input random variable x which may take on values in a space X, an output random variable y which may take on values in a space Y, and a discrete collection G of channels g, each with input alphabet X and output alphabet Y. If probability distributions are given over G and X, and x and g are independent, then

$$I(X;\ Y/G) - H(G) \leqslant I(X;\ Y) \leqslant I(X;\ Y/G) \tag{3-3}$$

where H(G) is the entropy of the probability distribution over G. If G is not discrete, the right-hand inequality in Eq. (3-3) still holds. The situation is shown schematically in Fig. 5.

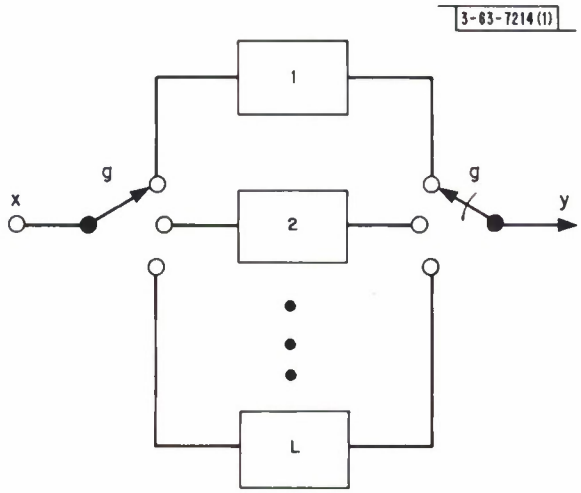


Fig. 5. General channel with state structure.

Proof.

$$I(X;\ YG) = I(X;\ Y) + I(X;\ G/Y)$$

$$I(X;\ YG) = I(X;\ G) + I(X;\ Y/G) \quad .$$

But $I(X;\ G) = 0$, since x and g are independent. Thus,

$$I(X;\ Y/G) = I(X;\ Y) + I(X;\ G/Y) \quad . \tag{3-4}$$

Also,

$$0 \leqslant I(X;\ G/Y) \leqslant H(G) \quad . \tag{3-5}$$

The right-hand inequality in Eq. (3-5) holds if G is discrete. Combining Eqs. (3-4) and (3-5), we have Eq. (3-3).

Now we may interpret Eq. (3-3). In the first place, g represents a "channel state" in just the sense we have been using this term. Hence, the rightmost inequality of Eq. (3-3) implies that knowing the channel state increases the mutual information between input and output. It is a somewhat disguised form of the statement that mutual information is a convex downward function of the channel transition probabilities.

From our assumptions, it is clear that

$$p(y/x) = \sum_{g \epsilon G} p(g)\, p_g(y/x) = \sum_{g \epsilon G} p(g)\, p(y/xg)^{\dagger} \quad .$$

Hence, H(G) is the entropy of a channel representation, and the leftmost inequality of Eq. (3-3) states that we need subtract only this entropy from the upper bound to I(X; Y) to obtain a lower bound. H(G) is thus a measure of the tightness of the bounds.

E. RANDOM CODING BOUND

Coding is a subject we have not discussed, as yet. For a discrete channel without parallel structure, a random coding bound is derived by choosing a probability distribution over all input letter sequences of length N, picking the requisite number of code words independently at random according to this distribution, computing an upper bound to the probability of error given that a particular message is transmitted, and averaging over the ensemble of possible codes. Since the bound thus obtained is independent of the particular message chosen, it is a bound to the average probability of error for the code.

In the discussion to follow, the bounding technique of Gallager[1‡] will be used; we shall state some of his results below. First, we must give the notation and assumptions. Let X^N be the set of all sequences of length N that can be transmitted on a given channel, and let Y^N be the set of all sequences of length N that can be received. We assume that both X^N and Y^N are finite sets. Let $p(\vec{y}/\vec{x})$, for $\vec{y} \epsilon Y^N$ and $\vec{x} \epsilon X^N$, be the conditional probability of receiving sequence $\vec{y}$ given that $\vec{x}$ was transmitted. We assume that we have a code consisting of W code words, that is, a mapping of the integers from 1 to W into a set of code words $\vec{x}_1, \ldots, \vec{x}_W$, where $\vec{x}_m \epsilon X^N$; $1 \leqslant m \leqslant W$. We also assume that maximum-likelihood decoding is performed at the receiver. Finally, we define a probability measure $p(\vec{x})$ on X^N and use $\overline{P}_{em}$ to denote the average over the ensemble of codes of the probability of error, given that the mth code word was transmitted.

Now we state the following result of Gallager,[2]

$$\overline{P}_{em} \leqslant (W-1)^{\rho} \sum_{\vec{y} \epsilon Y^N} \left[\sum_{\vec{x} \epsilon X^N} p(\vec{x})\, p(\vec{y}/\vec{x})^{1/1+\rho} \right]^{1+\rho} \tag{3-6}$$

for any ρ, $0 \leqslant \rho \leqslant 1$.

If we make some further assumptions, we can simplify the bound of Eq. (3-6). Let $x_1, \ldots, x_N$ be the individual letters in an input sequence $\vec{x}$, and let $y_1, \ldots, y_N$ be the letters in an output

† Here and in the remainder of this report, we will freely use the nototional convention that $p_t(u/v) = p(u/vt)$.

‡ Numbered references appear at the end of each chapter.

sequence $\vec{y}$. We now assume that the channel is memoryless and time invariant so that

$$p(\vec{y}/\vec{x}) = \prod_{n=1}^{N} p(y_n/x_n) \tag{3-7}$$

and that the probability distribution $p(\vec{x})$ on input sequences factors into a product of individual letter probability distributions as follows:

$$p(\vec{x}) = \prod_{n=1}^{N} p(x_n) \quad . \tag{3-8}$$

Then, a bound on the ensemble probability of decoding error P_e, which is independent of the probabilities with which the code words are used, is obtained in the form

$$P_e \leqslant \exp[-NE(R)] \tag{3-9}$$

where E(R) is called the random coding exponent and is defined by the equations

$$E_o(\rho, \vec{p}) = -\ln \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p(j/k)^{1/1+\rho} \right]^{1+\rho} \tag{3-10}$$

and

$$E(R) = \max_{\rho, \vec{p}} [-\rho R + E_o(\rho, \vec{p})] \quad . \tag{3-11}$$

We have assumed that the channel input alphabet consists of the integers from 1 to K, and that the channel output alphabet consists of the integers from 1 to J. The maximization in Eq. (3-11) is over all ρ, $0 \leqslant \rho \leqslant 1$, and all (input letter) probability vectors $\vec{p}$. R is the rate in natural units (i.e., $W = \exp[NR]$).

Now, the question arises of what to do about the parallel structure of a channel if such structure exists. In the first place, Gallager's bounds in Eqs. (3-6) and (3-9) apply without change to a channel with parallel structure if we understand that a "letter" (x_n or y_n) in the sense used above is an M-tuple which is composed of the inputs to or outputs of the M subchannels of an arbitrary M-input, M-output channel. Then, if each subchannel input has an L letter alphabet and each subchannel output has a Q letter alphabet, $K = L^M$ and $J = Q^M$. We note that these statements do not depend on the assumptions of Eqs. (3-7) and (3-8). Although Gallager's bounds are fully applicable to the situation we wish to study, some further structure will have to be imposed so that these bounds will be productive of insight in spite of the additional complexity of our channel model. Some of this structure is already implicit in our MS channel model. In addition, we will make notational changes which will facilitate the explanation of some of our results. Let the rate per subchannel R_s be defined by

$$R_s = \frac{R}{M} \quad . \tag{3-12}$$

Define

$$E_M(R_s) = E(R) \tag{3-13}$$

where there are M subchannels and E(R) is given by Eq. (3-11). Then, we have from Eqs. (3-13) and (3-9) that

$$P_e \leq \exp[-NE_M(R_s)] \quad . \tag{3-14}$$

Now, let us consider a special case. Suppose that the MS channel consists of M identical and independent subchannels. Then, Theorem 5 of Gallager[3] implies that $E_M(R_s)$ may be further decomposed so that

$$E_M(R_s) = ME_1(R_s) \tag{3-15}$$

where

$$E_1(U) = \max_{\rho, \vec{p}_s} \left\{ -\rho U - \ln \sum_{q=1}^{Q} \left[\sum_{\ell=1}^{L} p(\ell)\, p(q/\ell)^{1/1+\rho} \right]^{1+\rho} \right\} \tag{3-16}$$

and the maximization is performed over all ρ, $0 \leq \rho \leq 1$, and all probability vectors $\vec{p}_s$ defined on the subchannel input alphabet. All the quantities in Eq. (3-16) refer to a single subchannel. From Eqs. (3-14) and (3-15), we have

$$P_e \leq \exp[-NME_1(R_s)] \quad . \tag{3-17}$$

It should be strongly emphasized here that this result depends on our coding simultaneously in the "parallel" and "time" directions. According to Gallager's Theorem 5, one of the conditions for the maximum required in the definition of $E_M(R_s)$ is that each subchannel letter be chosen independently of the other subchannel letters at that instant of time, and independently of all subchannel letters at other instants of time. A code word may be thought of as a matrix with M rows and N columns. Each element of the matrix is a letter in the subchannel input alphabet. Each column of the matrix is a letter in the (whole) channel input alphabet. The bound of Eq. (3-17) assumes each matrix element is chosen independently of the others. Clearly, each output word may also be thought of as an $M \times N$ matrix with elements equal to letters of the subchannel output alphabet.

In what follows, we shall generally be studying MS channels whose subchannels are not independent, although the MS channel itself is memoryless. Thus, we shall obtain bounds of the form of Eq. (3-14).

F. STATE KNOWLEDGE – SOME GENERAL CONSIDERATIONS†

When dealing with a channel that has a state structure, one naturally expects that knowledge of the state at the receiver will be advantageous, both in terms of increasing the capacity and decreasing the probability of error. It would also be expected that partial knowledge of the state at the receiver is better than no knowledge, but not as good as complete knowledge.

In dealing with capacity, we may work directly on the mathematical expressions involved. The situation with regard to probability of error is somewhat different. Here, we know that receiver knowledge cannot increase the probability of error because the receiver uses this

† The remarks and results in the remainder of this chapter are not limited to channels with parallel structure.

knowledge optimally (i.e., it computes the likelihoods of code words based on what state knowledge it has). Since one of the options available to the receiver is to ignore any state knowledge it may have, the optimal receiver must do at least as well as this. This same inequality must apply to the ensemble probability of decoding error because it applies to each member of the ensemble. However, we do not generally compute this probability of error; we compute the random coding exponent (RCE). We would hope that an inequality between ensemble probabilities of decoding error for two categories of receiver state knowledge would be reflected in the opposite inequality between the corresponding RCE's. This is indeed the case, but it is necessary to pursue the mathematical properties of the RCE in order to prove it.

Suppose we have an input random variable x which may take on values in a space X, an output random variable y which may take on values in a space Y, and a collection G of channels g, each with input alphabet X and output alphabet Y. By complete receiver knowledge of the channel state, we mean that the receiver knows g. By partial receiver knowledge of the channel state, we mean that the receiver knows some observable $t \in T$† which is related to g.‡ We shall assume that a distribution $p(xygt)$ on $X \times Y \times G \times T$ is given, and that

$$p(y/xgt) = p(y/xg) \tag{3-18}$$

and

$$p(gt/x) = p(gt)^{\S} \quad . \tag{3-19}$$

The first assumption, Eq. (3-18), is consistent with our terminology of "partial" and "complete" knowledge, i.e., once g is known, t becomes irrelevant for the computation of $p(y/x)$. The second assumption, Eq. (3-19), is equivalent to the statement that the pair (g, t) conveys no information about x. Equations (3-18) and (3-19) taken together imply

$$p(t/gyx) = p(t/g) \quad . \tag{3-20}$$

This assures us that the receiver need not consider t if it knows g (see Ref. 4).

In the remarks following the proof of Theorem 3.1, it was noted that the effect of state knowledge in increasing mutual information was related to the convexity (downward) of the mutual information as a function of the input-output conditional probabilities. The following theorem on convex functions will be useful in the sequel.

Theorem 3.2.

Let f be a convex downward function of transition probabilities $p(y/x)$, and assume that a probability distribution $p(xygt)$ on $X \times Y \times G \times T$ is given such that Eqs. (3-18) and (3-19) are satisfied. Then,

---

† None of the spaces X, Y, G, and T need be finite or even discrete. We will proceed as though all the spaces were discrete, and remark that an appropriate replacement of sums by integrals covers the other cases, until we reach Theorem 3.6 which requires G to be finite.

‡ For example, if we are dealing with a single fading channel with binary input and output alphabets and equal transmitted energy allotted to 0 and 1, g would be the bit crossover probability, and t would be the energy of the received waveform.

§ Assumptions in Eqs. (3-18) and (3-19) and all the statements in the paragraph preceding them shall be in effect for the remainder of this chapter.

$$f\,[p(y/x)] \leqslant \sum_{t \epsilon T} p(t)\, f\,[p(y/xt)] \leqslant \sum_{g \epsilon G} p(g)\, f\,[p(y/xg)] \quad . \tag{3-21}$$

[If f is convex upward, the inequalities in Eq. (3-21) are reversed.]

Proof.

$$p(y/xt) = \sum_{g \epsilon G} p(g/tx)\, p(y/gtx)$$

$$= \sum_{g \epsilon G} p(g/t)\, p(y/xg) \tag{3-22}$$

from Eqs. (3-18) and (3-19). Also,

$$p(y/x) = \sum_{t \epsilon T} p(y/xt)\, p(t/x) = \sum_{t \epsilon T} p(y/xt)\, p(t) \tag{3-23}$$

from Eq. (3-19). Hence,

$$f\,[p(y/x)] \leqslant \sum_{t \epsilon T} p(t)\, f\,[p(y/xt)] \tag{3-24}$$

from convexity of f and Eq. (3-23), and

$$f\,[p(y/xt)] \leqslant \sum_{g \epsilon G} p(g/t)\, f\,[p(y/xg)] \tag{3-25}$$

from convexity of f and Eqs. (3-22). Inequality (3-25) implies

$$\sum_{t \epsilon T} p(t)\, f\,[p(y/xt)] \leqslant \sum_{t \epsilon T} \sum_{g \epsilon G} p(t)\, p(g/t)\, f\,[p(y/xg)]$$

$$\leqslant \sum_{g \epsilon G} p(g)\, f\,[p(y/xg)] \quad . \tag{3-26}$$

Equations (3-24) and (3-26) are equivalent to Eq. (3-21).

G. STATE KNOWLEDGE, MUTUAL INFORMATION AND CAPACITY

Theorem 3.3.

$$I(X;\ Y) \leqslant I(X;\ Y/T) \leqslant I(X;\ Y/G) \quad . \tag{3-27}$$

Proof.

Since the mutual information is a convex downward function of the transition probabilities, this follows directly from Theorem 3.2.

Theorem 3.4.

Denote the capacity of the channel when the receiver knows neither t nor g as C, the capacity when the receiver knows t as C_t, and the capacity when the receiver knows g as C_g. Then,

$$C \leqslant C_t \leqslant C_g \quad . \tag{3-28}$$

Proof.

Since Eq. (3-27) holds for all input distributions p(x), Theorems 2.3 or 2.4 give us Eq. (3-28).

This concludes our discussion of the effect of channel state information on the mutual information between input and output and on capacity.

H. STATE KNOWLEDGE AND RANDOM CODING EXPONENT (RCE)

We shall begin this section by deriving some mathematical expressions involved in the definition of the RCE when an auxiliary variable $v \in V$, independent of the input, is known to the receiver.

We note that the RHS of Eq. (3-6) is independent of m. Hence, it is a bound on the ensemble probability of decoding error and is independent of the probabilities with which the code words are used. If, during an input sequence of length N, the variable v assumes the values $v_1, \ldots, v_N$, then we assume the conditional probability relating input and output sequences to be given by

$$p_{\vec{v}}(\vec{y}/\vec{x}) = \prod_{n=1}^{N} p_{v_n}(y_n/x_n) \quad . \tag{3-29}$$

To obtain the RCE $[E^V(R)]$ corresponding to receiver knowledge of v, we must substitute Eqs. (3-8) and (3-29) in Eq. (3-6), replace $(W-1)$ by $W = e^{RN}$, average over the distribution† of $\vec{v}$, divide the negative of the natural logarithm of the result by N, and perform a maximization. Thus, if we define

$$E^V(\rho, \vec{p}, R) = -\frac{1}{N} \ln \left[e^{\rho RN} \sum_{\vec{v} \in V^N} p(\vec{v}) \sum_{\vec{y} \in Y^N} \left\{ \sum_{\vec{x} \in X^N} \left[\prod_{n=1}^{N} p(x_n) \right] \right.\right.$$

$$\left.\left. \times \left[\prod_{n=1}^{N} p_{v_n}(y_n/x_n) \right]^{1/1+\rho} \right\}^{1+\rho} \right] \tag{3-30}$$

where V^N is the set of all sequences of v's of length N, $\vec{v} \in V^N$, we have

$$E^V(R) = \max_{\substack{0 \leq \rho \leq 1 \\ \vec{p} \in P}} E^V(\rho, \vec{p}, R) \quad . \tag{3-31}$$

We shall assume

$$p(\vec{v}) = \prod_{n=1}^{N} p(v_n) \quad . \tag{3-32}$$

This corresponds to the assumption of time invariance and memorylessness if v is the state variable. Substituting Eq. (3-32) in Eq. (3-30) and reducing the result, we get

† We will now assume v discrete, $X = \{1, \ldots, K\}$, and $Y = \{1, \ldots, J\}$ for purposes of notation. Similar results may be obtained if any or all of these assumptions are dropped.

$$E^V(\rho, \vec{p}, R) = -\frac{1}{N} \ln \left\{ e^{\rho R} \sum_{v \epsilon V} p(v) \sum_{y \epsilon Y} \left[\sum_{x \epsilon X} p(x)\, p_v(y/x)^{1/1+\rho} \right]^{1+\rho} \right\}^N$$

$$= -\rho R - \ln \left\{ \sum_{v \epsilon V} p(v) \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p_v(j/k)^{1/1+\rho} \right]^{1+\rho} \right\} \quad . \tag{3-33}$$

Define

$$F^V(\rho, \vec{p}) = \sum_{v \epsilon V} p(v) \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p_v(j/k)^{1/1+\rho} \right]^{1+\rho} \quad . \tag{3-34}$$

Thus,

$$E^V(\rho, \vec{p}, R) = -\rho R - \ln F^V(\rho, \vec{p}) \quad . \tag{3-35}$$

We may also define

$$E(\rho, \vec{p}, R) = -\rho R - \ln \left\{ \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p(j/k)^{1/1+\rho} \right]^{1+\rho} \right\} \tag{3-36}$$

and

$$F(\rho, \vec{p}) = \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p(j/k)^{1/1+\rho} \right]^{1+\rho} \quad . \tag{3-37}$$

Thus, we have

$$E(\rho, \vec{p}, R) = -\rho R - \ln F(\rho, \vec{p}) \tag{3-38}$$

and

$$E(R) = \max_{\substack{0 \leq \rho \leq 1 \\ \vec{p} \epsilon P}} E(\rho, \vec{p}, R) \quad . \tag{3-39}$$

Now, we may easily show that F is a convex upward function of the conditional probabilities included in its definition. Suppose

$$p(j/k) = \sum_{v \epsilon V} p(v)\, p_v(j/k) \quad . \tag{3-40}$$

Hence,

$$F(\rho, \vec{p}) = \sum_{j=1}^{K} \left\{ \sum_{k=1}^{K} p(k) \left[\sum_{v \epsilon V} p(v)\, p_v(j/k) \right]^{1/1+\rho} \right\}^{1+\rho} \quad . \tag{3-41}$$

By applying Eq.(C-3) (Minkowski's inequality) to the inner two sums of the RHS of Eq.(3-41), we get

$$\sum_{j=1}^{J} \left\{ \sum_{k=1}^{K} p(k) \left[\sum_{v \in V} p(v)\, p_v(j/k) \right]^{1/1+\rho} \right\}^{1+\rho}$$

$$\geqslant \sum_{j=1}^{J} \sum_{v \in V} \left[\sum_{k=1}^{K} p(k)\, p(v)^{1/1+\rho}\, p_v(j/k)^{1/1+\rho} \right]^{1+\rho}$$

$$= \sum_{v \in V} p(v) \left\{ \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p_v(j/k)^{1/1+\rho} \right]^{1+\rho} \right\} \tag{3-42}$$

as required.

Theorem 3.5.

If $E(R)$ is the RCE corresponding to no state knowledge at the receiver, $E^t(R)$ and $E^g(R)$ are the RCE's corresponding to receiver knowledge of t and g, respectively, and Eqs. (3-29) and (3-32) hold for t, g, or a blank replacing v, then

$$E(R) \leqslant E^t(R) \leqslant E^g(R) \quad . \tag{3-43}$$

Proof.

By the convexity upward of F and Theorem 3.2, we have

$$F^g(\rho, \vec{p}) \leqslant F^t(\rho, \vec{p}) \leqslant F(\rho, \vec{p}) \quad .$$

Hence,

$$-\rho R - \ln F(\rho, \vec{p}) \leqslant -\rho R - \ln F^t(\rho, \vec{p}) \leqslant -\rho R - \ln F^g(\rho, \vec{p}) \quad .$$

By Eqs. (3-35) and (3-38), we have

$$E(\rho, \vec{p}, R) \leqslant E^t(\rho, \vec{p}, R) \leqslant E^g(\rho, \vec{p}, R) \quad .$$

Our result follows from Eqs. (3-31), (3-39), and Theorem 2.3 (or Theorem 2.4).

It should be emphasized that not only must the state variable g and partial knowledge variable t satisfy Eqs. (3-18) and (3-19), but successive values of g and t must be independent and identically distributed. In addition, the single-letter conditional probability of the channel must depend only on the value g or t assumes during the transmission of a single letter. If all these assumptions hold, we shall say that the channel with complete or partial state knowledge is still memoryless and time invariant.

We avoided making these additional assumptions in Theorems 3.3 and 3.4, but the results there are "one-shot" results. If the additional assumptions are made, the results become "per-transmitted-letter" results as well.

We have devoted a fair amount of space to showing that state knowledge increases the RCE, a result which is analogous to the result that state knowledge increases mutual information. Included in the mathematical statement [Eq. (3-3)] of this last fact is a bound on the magnitude of the increase. We shall now derive an analogous result for the case of the RCE. Our notation and assumptions remain the same.

Theorem 3.6.

Let G, the set of channel states, contain S elements, and let the channel with or without state information be memoryless and time invariant. Let ρ' be the value of ρ which achieves the maximum required in the definition of $E^g(R)$. Then,

$$E^g(R) - \rho' \ln S \leqslant E(R) \leqslant E^g(R) \quad . \tag{3-44}$$

Proof.

From Eq. (C-1),

$$p(j/k)^{1/1+\rho} = \left[\sum_{g \epsilon G} p(g)\, p_g(j/k) \right]^{1/1+\rho} \leqslant \sum_{g \epsilon G} p(g)^{1/1+\rho}\, p_g(j/k)^{1/1+\rho} \quad . \tag{3-45}$$

Hence,

$$\sum_{j=1}^{J} \left\{ \sum_{k=1}^{K} p(k) \left[\sum_{g \epsilon G} p(g)\, p_g(j/k) \right]^{1/1+\rho} \right\}^{1+\rho}$$

$$\leqslant \sum_{j=1}^{J} \left[\sum_{g \epsilon G} p(g)^{1/1+\rho} \sum_{k=1}^{K} p(k)\, p_g(j/k)^{1/1+\rho} \right]^{1+\rho}$$

$$= S^{1+\rho} \sum_{j=1}^{J} \left[\sum_{g \epsilon G} \frac{1}{S}\, p(g)^{1/1+\rho} \sum_{k=1}^{K} p(k)\, p_g(j/k)^{1/1+\rho} \right]^{1+\rho} \quad . \tag{3-46}$$

Using Eq. (C-2) on the sum over G, we get

$$S^{1+\rho} \sum_{j=1}^{J} \left[\sum_{g \epsilon G} \frac{1}{S}\, p(g)^{1/1+\rho} \sum_{k=1}^{K} p(k)\, p_g(j/k)^{1/1+\rho} \right]^{1+\rho}$$

$$\leqslant S^{1+\rho} \sum_{j=1}^{J} \left\{ \sum_{g \epsilon G} \frac{1}{S}\, p(g) \left[\sum_{k=1}^{K} p(k)\, p_g(j/k)^{1/1+\rho} \right]^{1+\rho} \right\}$$

$$= S^{\rho} \sum_{g \epsilon G} p(g) \left\{ \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p_g(j/k)^{1/1+\rho} \right]^{1+\rho} \right\}$$

$$= S^{\rho} F^g(\rho, \vec{p}) \tag{3-47}$$

where we use Eq. (3-34) with g replacing v in the last step. Hence,

$$F(\rho, \vec{p}) \leqslant S^{\rho} F^g(\rho, \vec{p}) \tag{3-48}$$

and

$$E^g(\rho, \vec{p}, R) - \rho \ln S \leqslant E(\rho, \vec{p}, R) \quad . \tag{3-49}$$

Suppose

$$E^g(R) = E^g(\rho', \vec{p}', R) \quad . \tag{3-50}$$

Then,

$$E^g(R) - \rho' \ln S \leqslant E(\rho', \vec{p}', R) \leqslant E(R) \quad . \tag{3-51}$$

Since $0 \leqslant \rho' \leqslant 1$, we have immediately

$$E^g(R) - \ln S \leqslant E(R) \tag{3-52}$$

which bears a strong resemblance to the left inequality of Eq. (3-3). Equation (3-51) contains the left inequality in Eq. (3-44). The right inequality comes directly from Theorem 3.5.

It is important to note that ρ' is implicitly a function of R and is the value of ρ which achieves the maximum in Eq. (3-31), with g replacing v.

I. SUBCHANNEL DEPENDENCIES AND RCE

It now seems appropriate to remark that there is no RCE counterpart to Theorem 2.5. Subchannel dependencies may either increase or decrease the RCE. An example will illustrate this fact.

Example 2

Let $p_1(\xi/\eta)$, $p_2(\xi/\eta)$, and three state distributions q, r, and s be given below:

$$q(1,1) = q(2,2) = 1/2 \qquad q(1,2) = q(2,1) = 0$$

$$r(1,1) = r(1,2) = r(2,1) = r(2,2) = 1/4$$

$$s(1,2) = s(2,1) = 1/2 \qquad s(1,1) = s(2,2) = 0 \quad .$$

Each state distribution leads to a different 2S channel. We note that the channel corresponding to r has independent subchannels. It may be verified that the "r channel" is the dependence-removed channel derived from either the q or s channels. The input distribution which achieves the maximum required by the definition of the RCE is the same for all three cases:

$$p(00) = p(01) = p(10) = p(11) = 1/4 \quad .$$

This may be verified by using Gallager's Theorem 4 (Ref. 5). The curves of $E_2(R_s)$ vs R_s for the three cases are given in Fig. 6. Since the curve for the independent subchannels case lies between the other two, we see that subchannel dependencies may either increase or decrease the RCE.†

† In fact, the s channel has a zero-error capacity equal to its capacity of 1 bit.

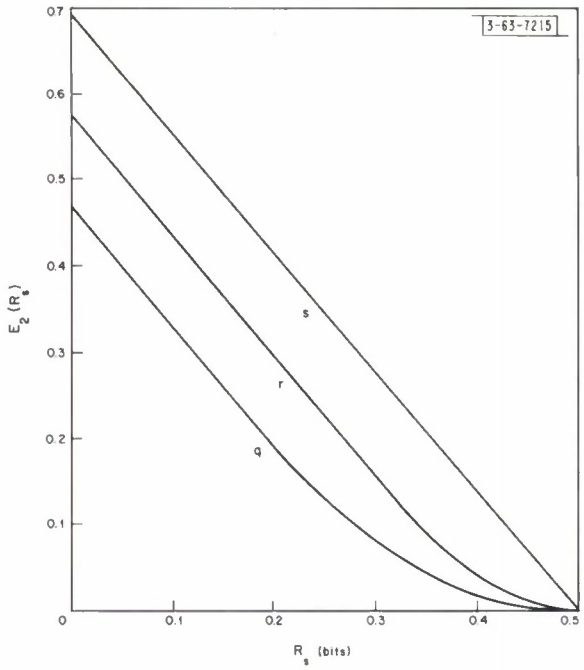


Fig. 6. $E_2(R_S)$ vs R_S for three different channels.

REFERENCES

1. R. G. Gallager, "A Simple Derivation of the Coding Theorem and Some Applications," IEEE Trans. Inform. Theory IT-11, No. 1, 3 – 18 (1965).
2. Ibid., p. 4, Eq. (11).
3. Ibid., p. 10.
4. J. M. Wozencraft and I. M. Jacobs, Principles of Communication Engineering (Wiley, New York, 1965), p. 220.
5. R. G. Gallager, op. cit., p. 8.

CHAPTER 4
THE COMPLETELY CONSTRAINED CHANNEL

A. DEFINITION OF CHANNEL

An important limiting case of the MS channel is the class of MS channels with the property that during the transmission of a single input letter all the subchannel states are the same. We will call such channels completely constrained (MSCC) channels. For MSCC channels, Eq. (2-7) becomes

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M) = \sum_{\alpha \in \Lambda} p(\alpha)\, p_\alpha(y_1/x_1) \cdots p_\alpha(y_M/x_M) \tag{4-1}$$

where we will always assume $p(\alpha) > 0$, $\alpha \in \Lambda$. Because of the complete dependence of the subchannel states, the MSCC channel has some very striking properties; in fact, these are properties of sequences of MSCC channels $\{\mathcal{C}_M\}_{M=1}^{\infty}$, defined as follows:

(1) Λ, the set of subchannel states, is the same for all M.

(2) $p(\alpha)$, $\alpha \in \Lambda$ is the same for all M.

(3) For the Mth channel $\mathcal{C}_M$ in the sequence, $p(y_1, \ldots, y_M/x_1, \ldots, x_M)$ is given by Eq. (4-1).

B. EXAMPLES OF MSCC CHANNELS AND THEIR PROPERTIES

To illuminate the definition of sequences of MSCC channels and provide specific examples of their general properties, we will discuss two examples.

Example 1

Let $\Lambda = \{1, 2\}$, $p(1) = p(2) = 1/2$, and $Y_s = X_s = \{0, 1\}$. Let $p_1(y_i/x_i)$ and $p_2(y_i/x_i)$ be the binary symmetric distributions with crossover probabilities equal to zero and one-half, respectively. Then, the Mth channel in our sequence may be represented as in Fig. 7.

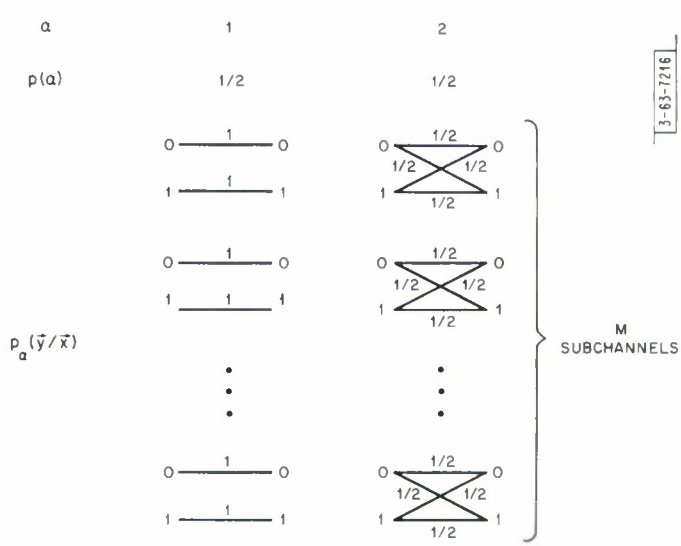


Fig. 7. An MSCC channel – Example 1.

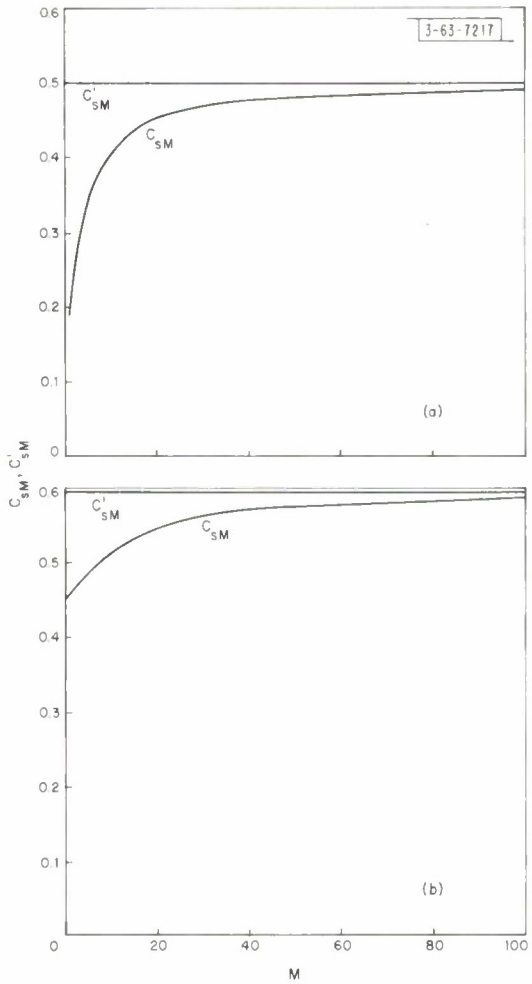


Fig. 8. Capacity per subchannel vs number of subchannels for (a) Example 1, and (b) Example 2.

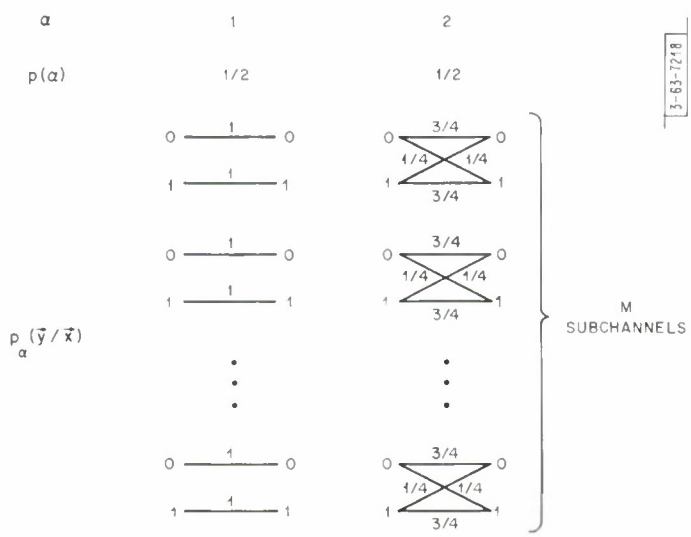


Fig. 9. An MSCC channel — Example 2.

Now, we will introduce some further notation which will be in effect for the remainder of this chapter.

Associated with the Mth member of a sequence of MSCC channels, we have a capacity C_M and a random coding exponent $E_M(R_s)$. We define the capacity per subchannel C_{sM} by

$$C_{sM} = \frac{C_M}{M} \quad . \tag{4-2}$$

We introduce the convention that when the channel state is known at the receiver, the corresponding capacity and RCE will be represented by primed quantities [e.g., C'_M, C'_{sM}, $E'_M(R_s)$], and when it is unknown at the receiver, by unprimed quantities.

Plots of C_{sM} and C'_{sM} vs M for Example 1 are found in Fig. 8(a).

Example 2

Let $\Lambda = \{1, 2\}$, $p(1) = p(2) = 1/2$, and $Y_s = X_s = \{0, 1\}$. Let $p_1(y_i/x_i)$ and $p_2(y_i/x_i)$ be the binary symmetric distributions with crossover probabilities equal to zero and one-quarter, respectively. Then, the Mth channel in our sequence may be represented as in Fig. 9.

Plots of C_{sM} and C'_{sM} vs M for Example 2 are found in Fig. 8(b).

C. CAPACITY THEOREMS FOR MSCC CHANNELS

Theorem 4.1.

When the channel state is known at the receiver, the capacity per subchannel [defined by Eq. (4-2)] is the same as the capacity of a single subchannel standing alone, i.e.,

$$C'_{sM} = C'_1 \quad . \tag{4-3}$$

Proof.

The result follows directly from Theorem 2.7.

Theorem 4.1 is illustrated by the two horizontal lines in Fig. 8(a-b). However, an MS channel need not be MSCC for the theorem to hold; it holds whenever the individual subchannel capacities (with state known at the receiver) are all equal. This last is certainly true if for each $\alpha \in \Lambda$ the probability that the ith subchannel is in state α is independent of i.

Theorem 4.2.

If

$$H = -\sum_{\alpha \in \Lambda} p(\alpha) \log p(\alpha)$$

is finite, then

$$\lim_{M \to \infty} C_{sM} = C'_1 \quad . \tag{4-4}$$

Proof.

Applying Theorems 2.3 or 2.4 to Eq. (3-3), we get

$$C'_M - H \leqslant C_M \leqslant C'_M \quad .$$

Hence,

$$\frac{C'_M}{M} - \frac{H}{M} \leqslant \frac{C_M}{M} \leqslant \frac{C'_M}{M} \quad .$$

Applying Eqs. (4-2) and (4-3), we get

$$C'_1 - \frac{H}{M} \leqslant C_{sM} \leqslant C'_1 \quad .$$

Hence, passing to the limit, we obtain Eq. (4-4) directly.

The curves of Fig. 8 illustrate the limiting property of C_{sM} which was just proved. We note again that the channel need not be MSCC for the theorem to hold; in fact, Eq. (4-4) may obtain even if a discrete entropy H does not exist. (See Appendix E.)

D. FURTHER PROPERTIES OF EXAMPLES 1 AND 2

Just as the capacity theorems were illustrated by previously given curves, so we shall provide curves relating to the RCE's of our examples to illustrate the theorems which are to come. The data on which the curves are based are as follows.

For M = 1, 2, 5, 10, 20, 50, and 100, we computed $E_M(R_s)$ and $E'_M(R_s)$ for equally spaced values of R_s from zero to C_{sM} or C'_{sM}. The spacing was 0.025 bit. Furthermore, for each such computation, the value of ρ which achieves the maximum required by the definition of $E_M(R_s)$ or $E'_M(R_s)$ is provided as output. The input probability vector which, for the case of M subchannels, achieves the requisite maximum is the probability vector with each of its 2^M components equal to $(1/2)^M$ (see Ref. 1†).

In Figs. 10 through 23, the curves plotted from the data are:

Example 1

Fig. 10	$E_M(R_s)$ vs R_s for M = 1, 2, 5, 10, 20, 50, 100
Fig. 11	$E'_M(R_s)$ vs R_s for M = 1, 2, 5, 10, 20, 50, 100
Fig. 12	$E'_M(R_s) - E_M(R_s)$ vs R_s for M = 1, 2, 5, 10, 20, 50, 100
Fig. 13	$E_M(R_s)$ vs M for R_s = 0 to 0.3 in steps of 0.025 bit
	R_s = 0.3 to 0.45 in steps of 0.05 bit
Fig. 14	$E'_M(R_s)$ vs M for R_s = 0 to 0.3 in steps of 0.025 bit
	R_s = 0.3 to 0.45 in steps of 0.05 bit
Fig. 15 (state unknown)	Maximizing ρ vs M for R_s = 0 to 0.45 in steps of 0.05 bit
Fig. 16 (state known)	Maximizing ρ vs M for R_s = 0 to 0.45 in steps of 0.05 bit

†Numbered references appear at the end of each chapter.

Example 2

Fig. 17	$E_M(R_s)$ vs R_s for M = 1, 2, 5, 10, 20, 50, 100
Fig. 18	$E'_M(R_s)$ vs R_s for M = 1, 2, 5, 10, 20, 50, 100
Fig. 19	$E'_M(R_s) - E_M(R_s)$ vs R_s for M = 1, 2, 5, 10, 20, 50, 100
Fig. 20	$E_M(R_s)$ vs M for R_s = 0 to 0.2 in steps of 0.025 bit
	R_s = 0.2 to 0.3 in steps of 0.05 bit
	R_s = 0.3 to 0.5 in steps of 0.1 bit
Fig. 21	$E'_M(R_s)$ vs M for R_s = 0 to 0.2 in steps of 0.025 bit
	R_s = 0.2 to 0.3 in steps of 0.05 bit
	R_s = 0.3 to 0.5 in steps of 0.1 bit
Fig. 22 (state unknown)	Maximizing ρ vs M for R_s = 0 to 0.55 in steps of 0.05 bit
Fig. 23 (state known)	Maximizing ρ vs M for R_s = 0 to 0.55 in steps of 0.05 bit

We will refer to these figures in subsequent sections of this chapter.

E. RANDOM CODING EXPONENT (RCE) FOR MSCC CHANNELS

We shall now undertake to prove a number of general properties of the RCE's of a sequence of MSCC channels. We begin with a definition.

Corresponding to each subchannel state α, there is a unique subchannel conditional probability distribution $p_\alpha(\xi/\eta)$, $\xi \in Y_s$, $\eta \in X_s$. This defines a channel with a capacity C_α. If there exists $a \in \Lambda$ with $p(a) > 0$ and

$$C_a \leqslant C_\alpha \qquad \text{all } \alpha \in \Lambda \tag{4-5}$$

then we say that <u>there exists a worst subchannel state a</u>. (We have, in fact, <u>not</u> even assumed that Λ is purely discrete, but only that a has a positive probability, as opposed to a positive probability density.)

Theorem 4.3.

Suppose there exists a worst subchannel state a, with probability of occurrence p(a). Then, for all $R_s > C_a$, we have

$$E_M(R_s) \leqslant E'_M(R_s) \leqslant -\ln p(a) \tag{4-6}$$

for all M.

Proof.

Recall that, when the receiver knows the channel state,

$$P_e \leqslant \exp\left[-NE'_M(R_s)\right] \quad . \tag{4-7}$$

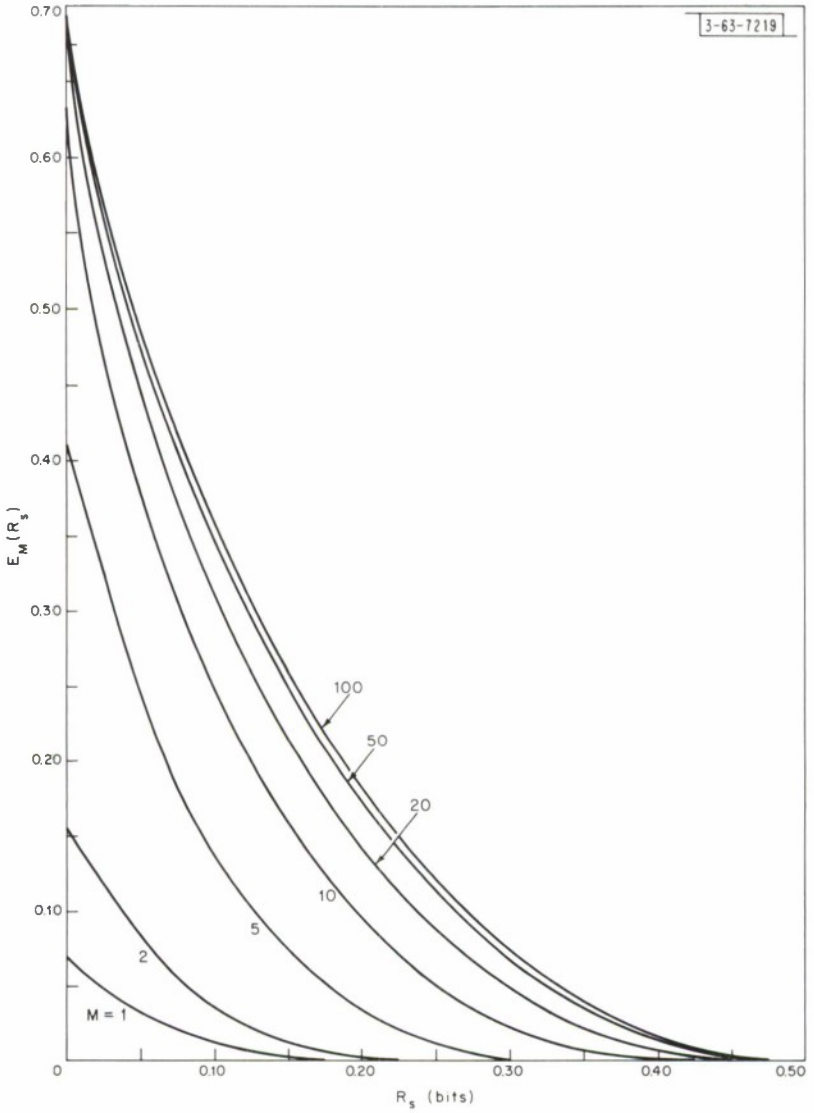


Fig. 10. Random coding exponent vs rate per subchannel (state unknown) — Example 1.

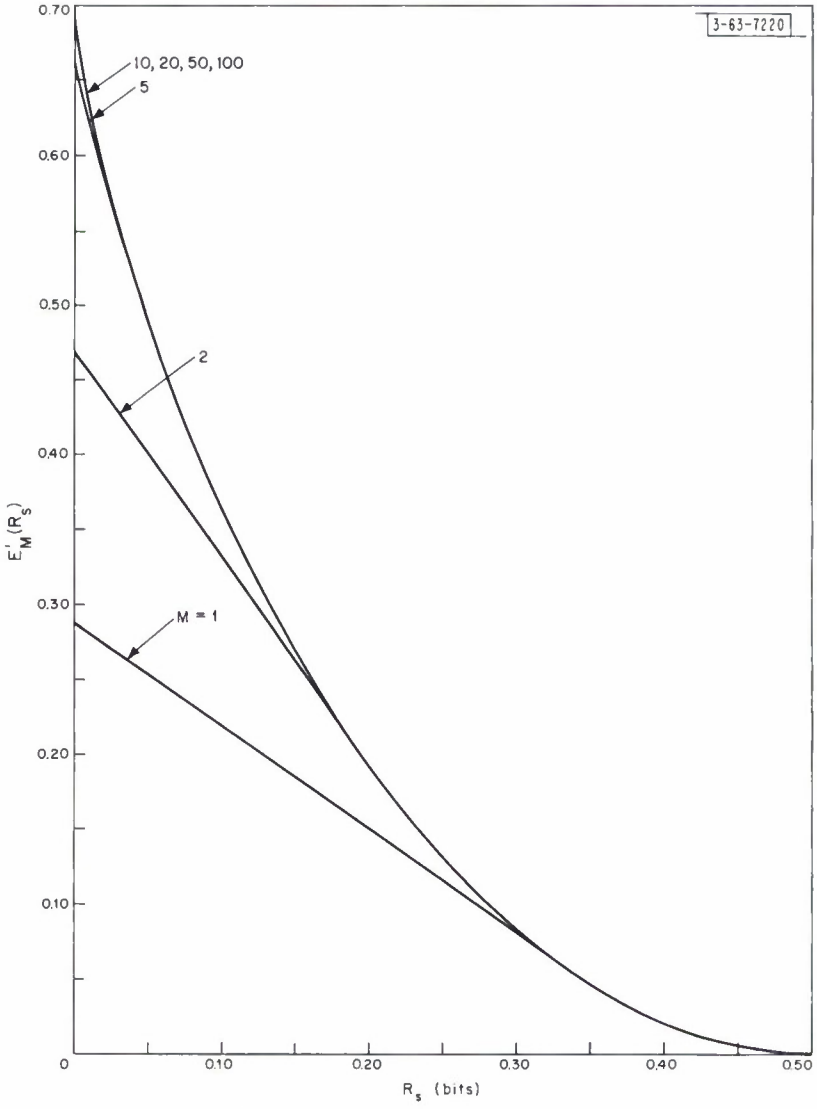


Fig. 11. Random coding exponent vs rate per subchannel (state known) – Example 1.

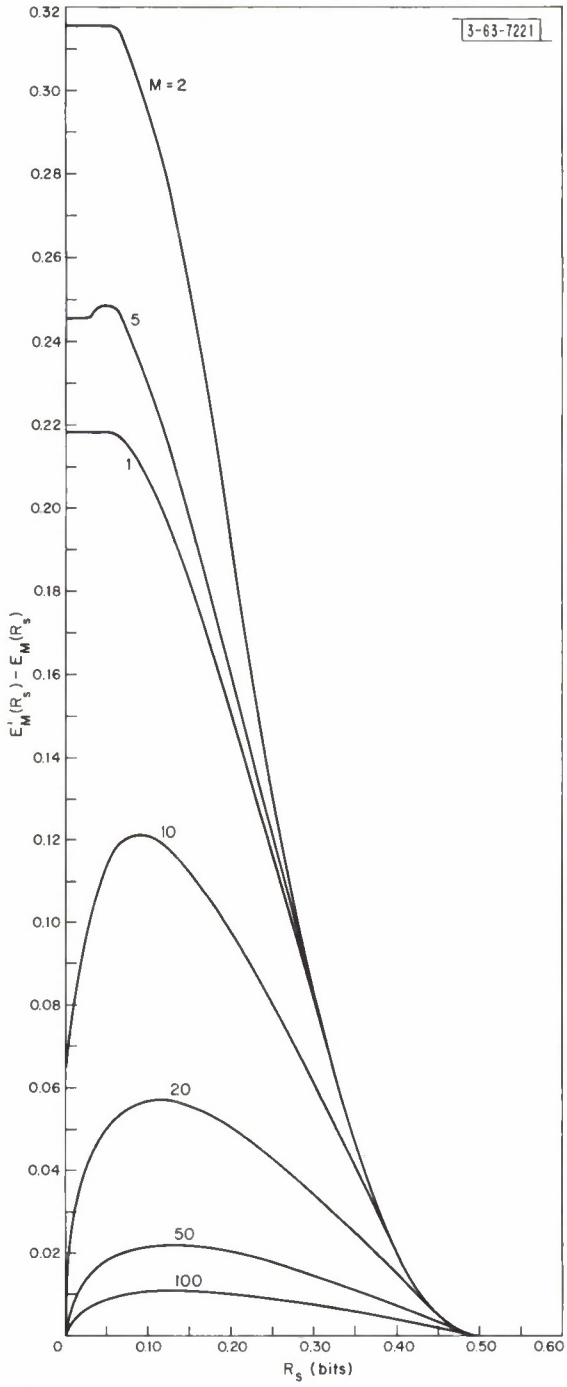


Fig. 12. Difference between state knawn and unknown randam coding expanents vs rate per subchannel – Example 1.

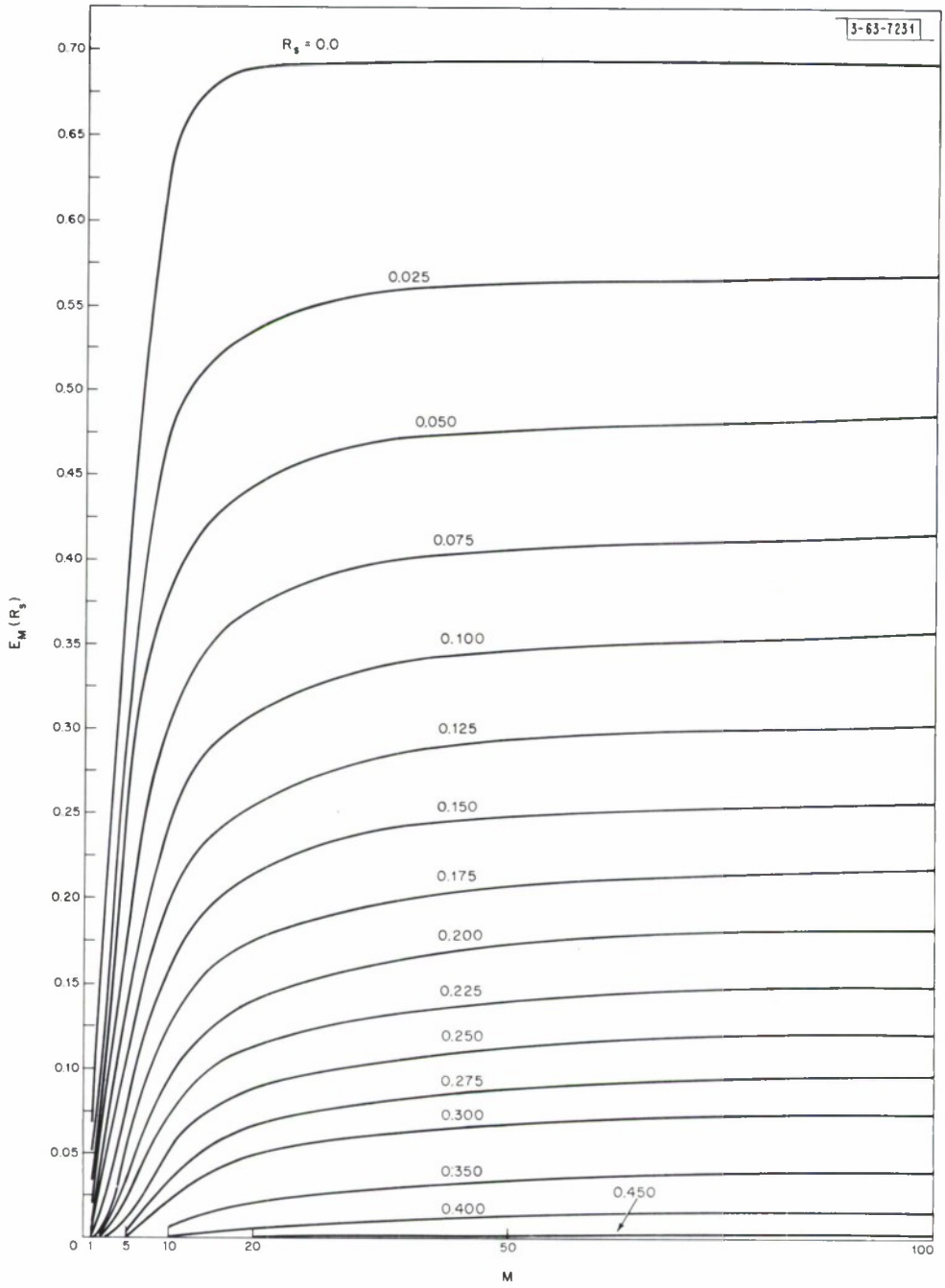


Fig. 13. Rondam coding exponent vs number of subchonnels (state unknawn) – Example 1.

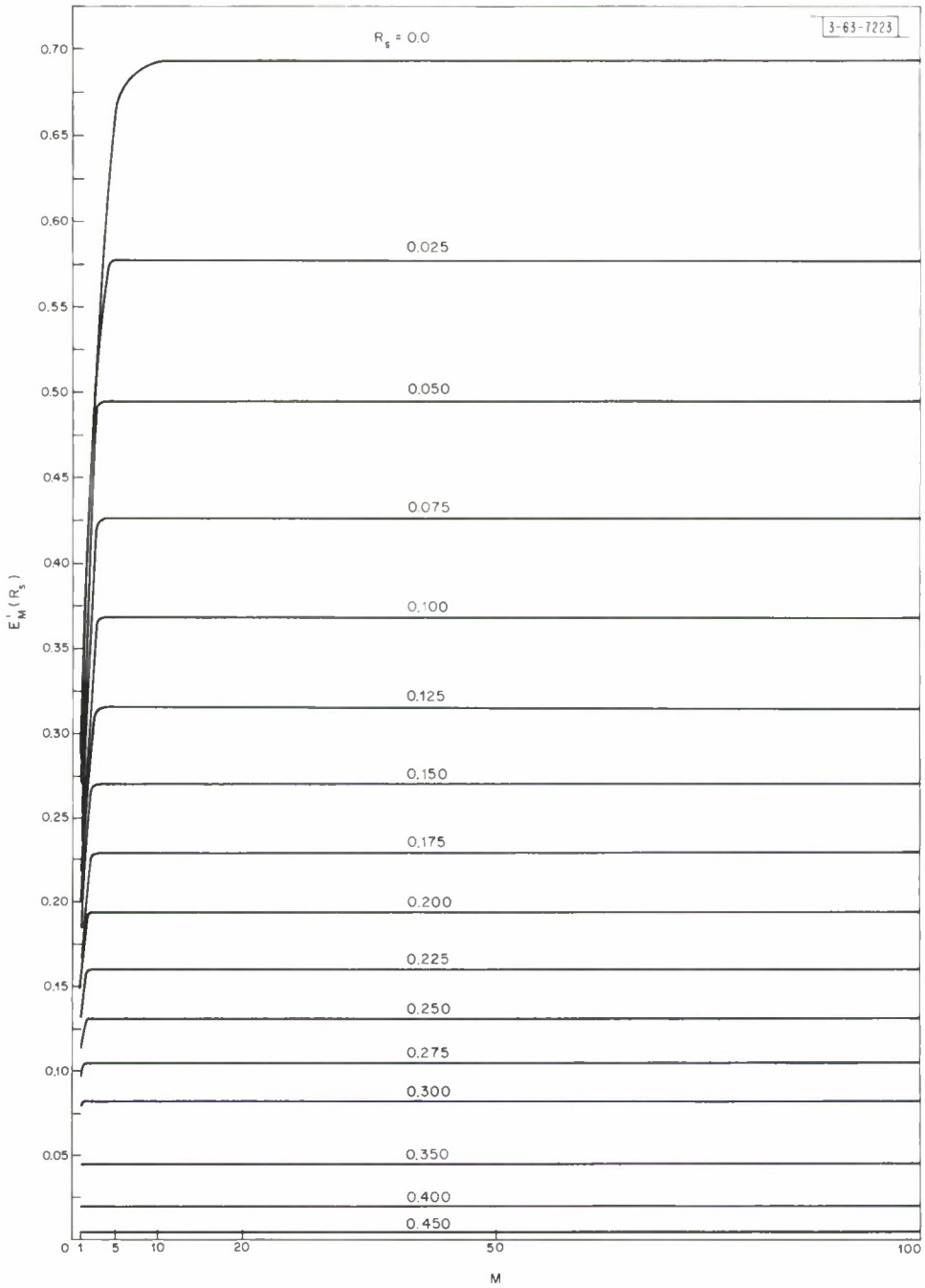


Fig. 14. Random coding exponent vs number of subchannels (state known) – Example 1.

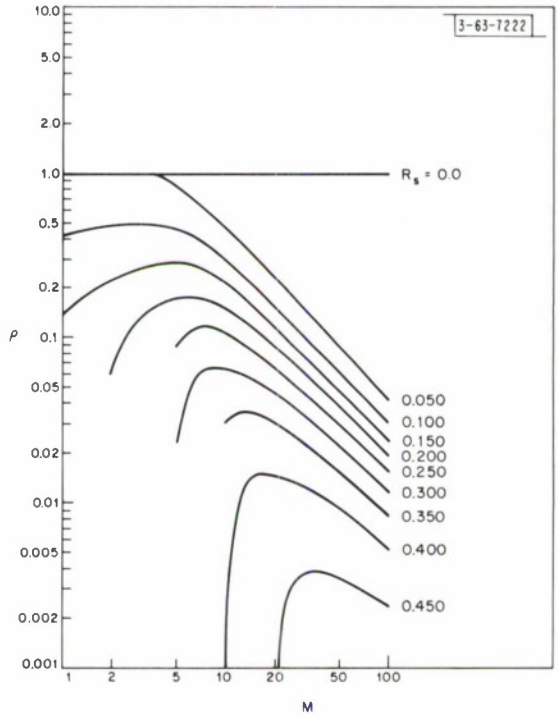


Fig. 15. Maximizing ρ vs number af subchannels (state unknawn) – Example 1.

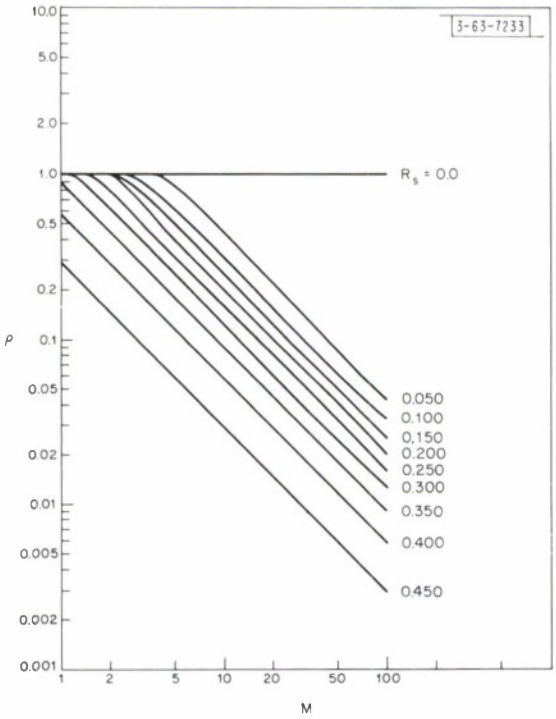


Fig. 16. Maximizing ρ vs number af subchannels (state knawn) – Example 1.

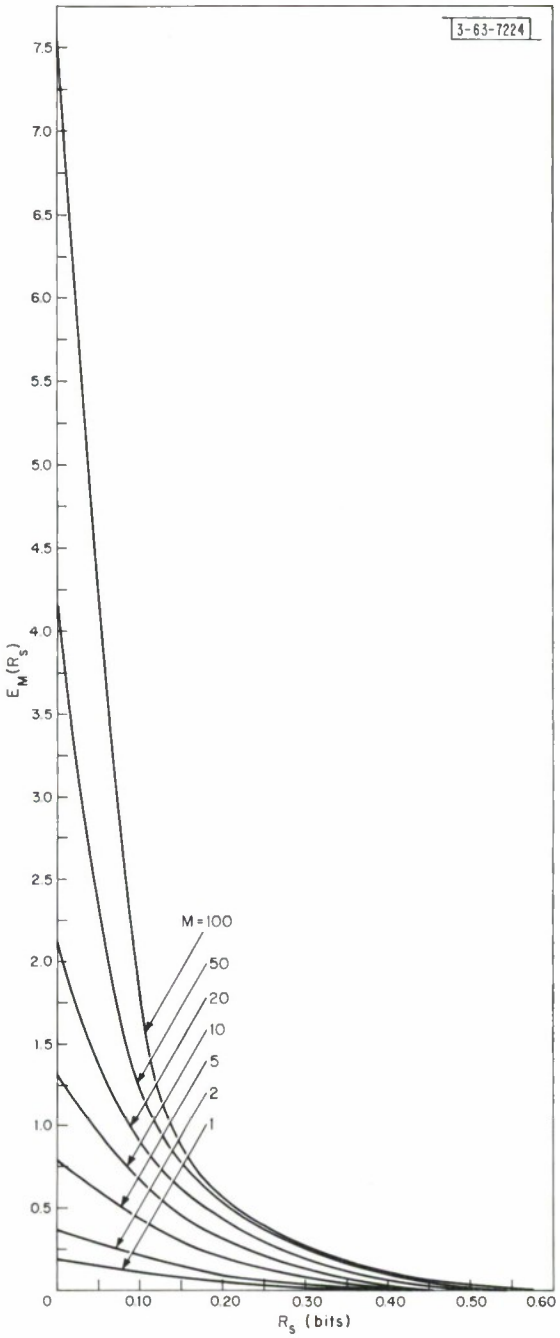


Fig. 17. Random coding exponent vs rate per subchannel (state unknown) – Example 2.

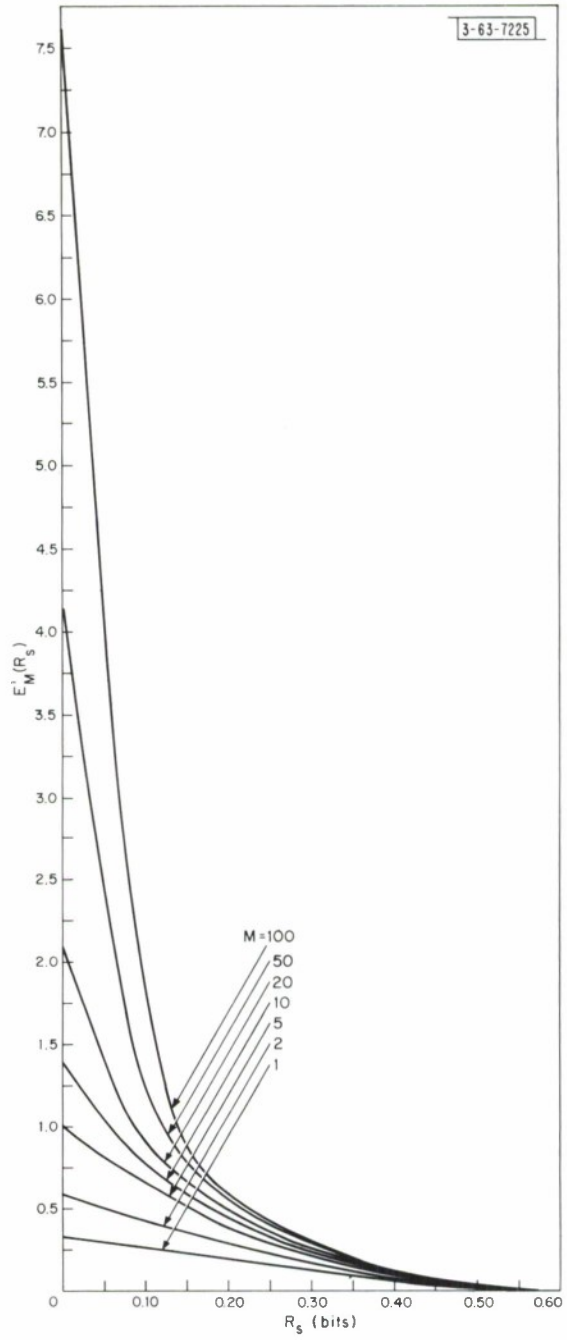


Fig. 18. Random coding exponent vs rate per subchannel (state known) – Example 2.

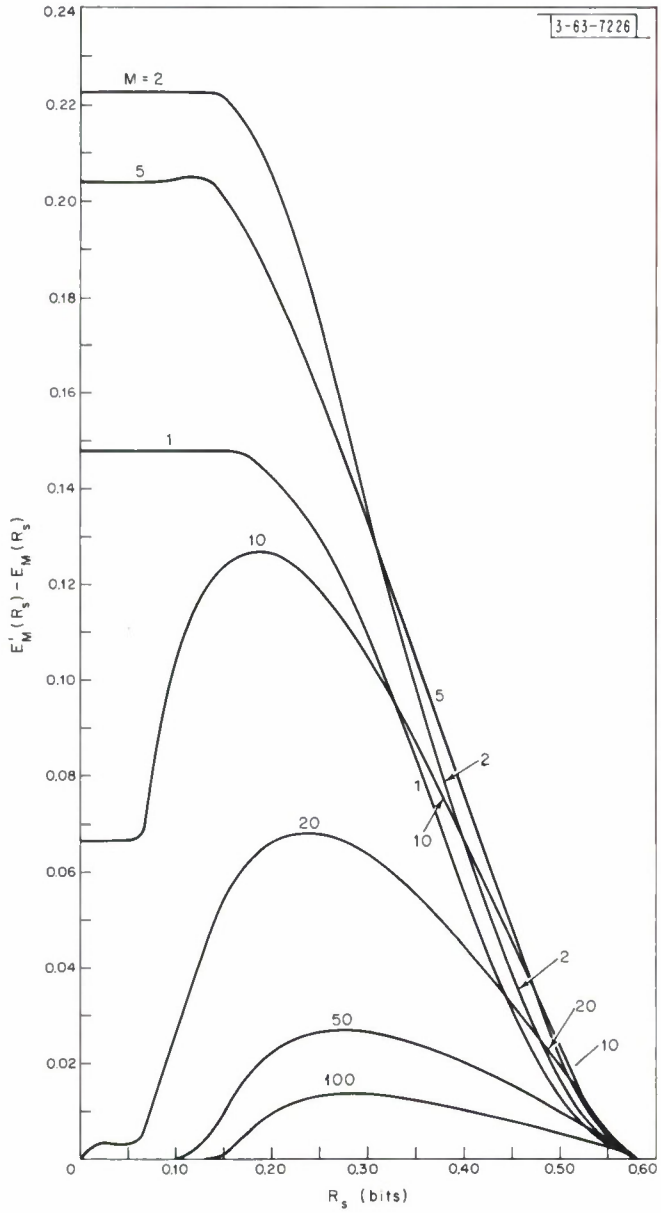


Fig. 19. Difference between state known and unknown random coding exponents vs rate per subchannel – Example 2.

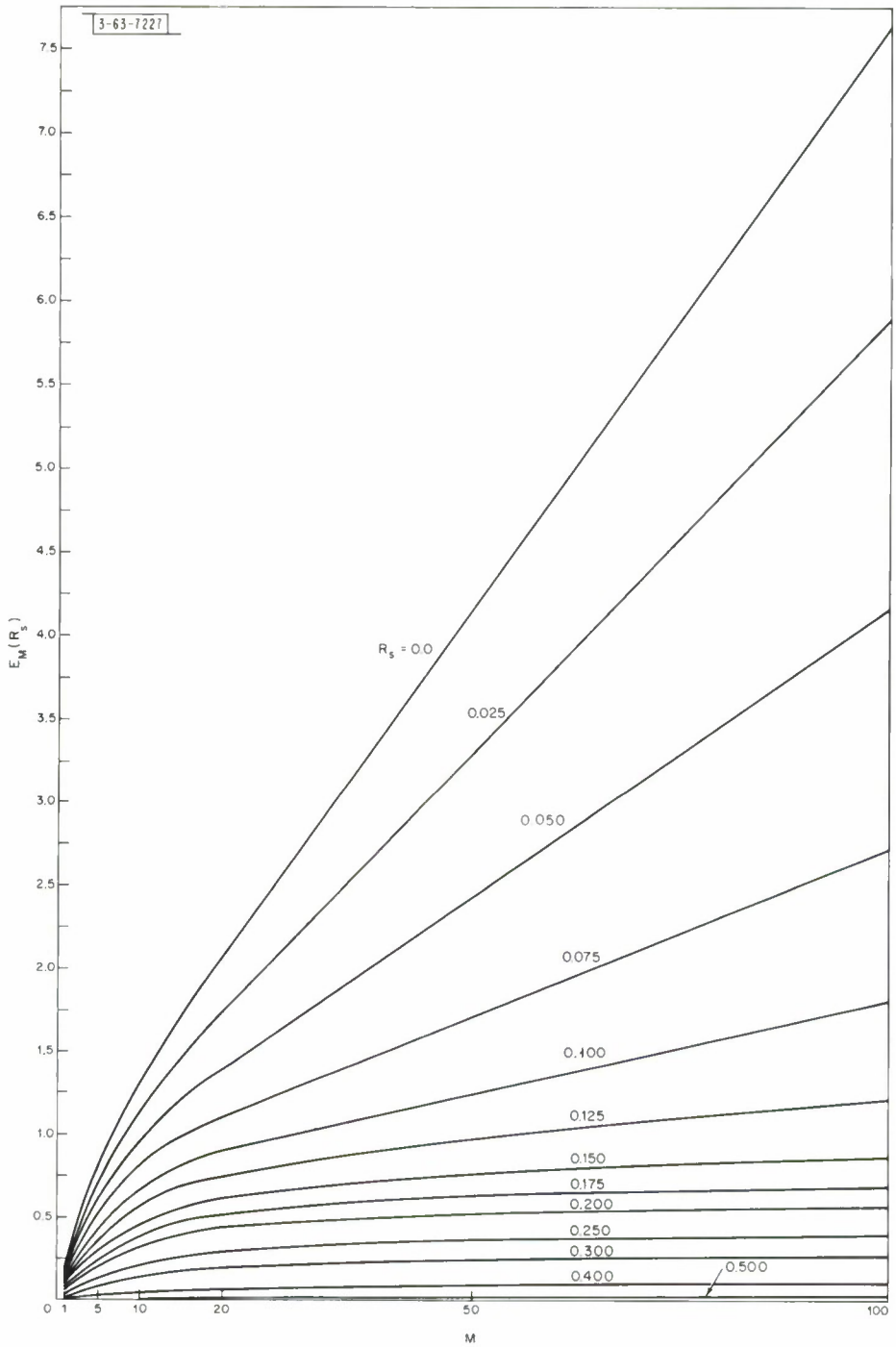


Fig. 20. Random coding exponent vs number of subchannels (state unknown) – Example 2.

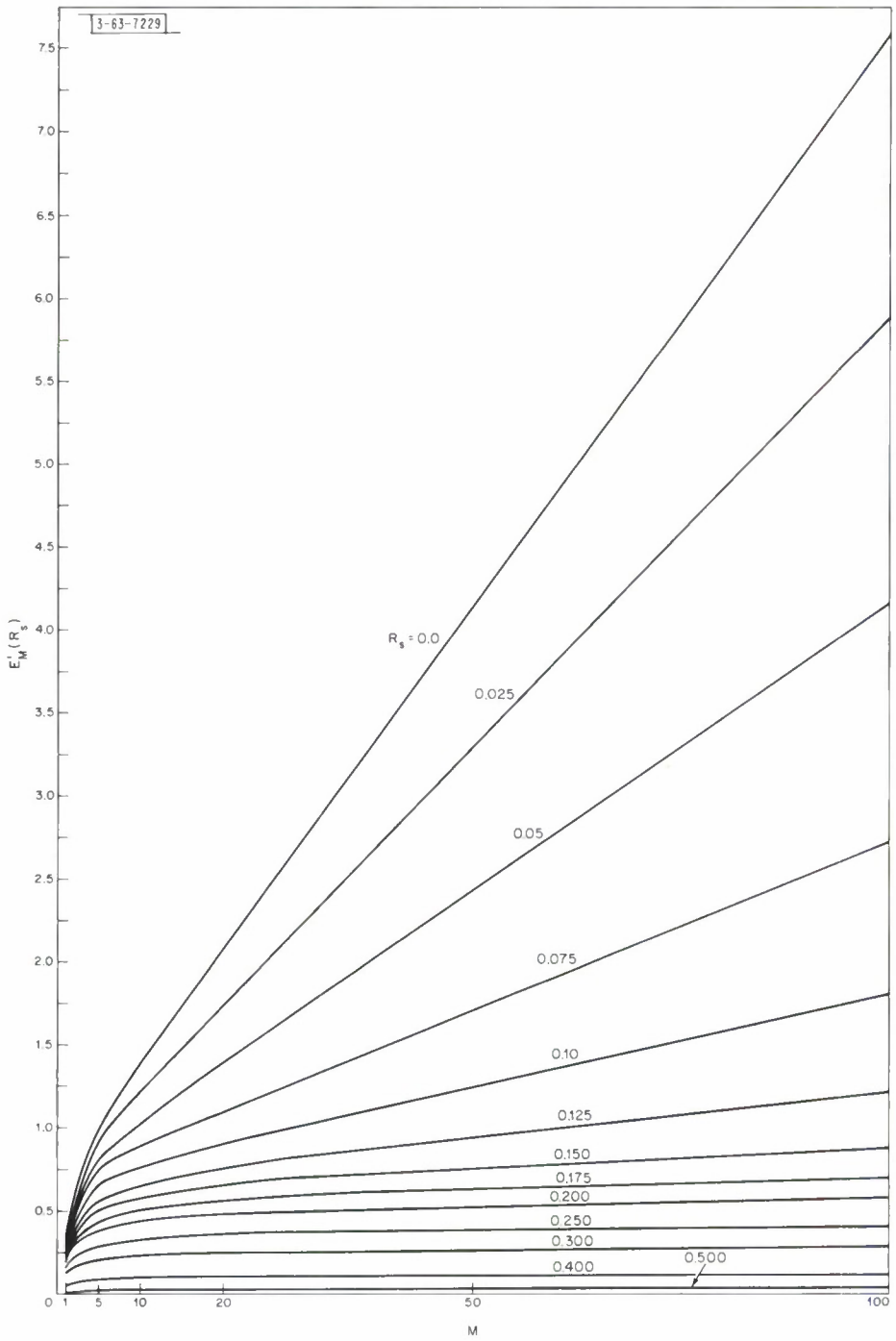


Fig. 21. Random coding exponent vs number of subchannels (state known) — Example 2.

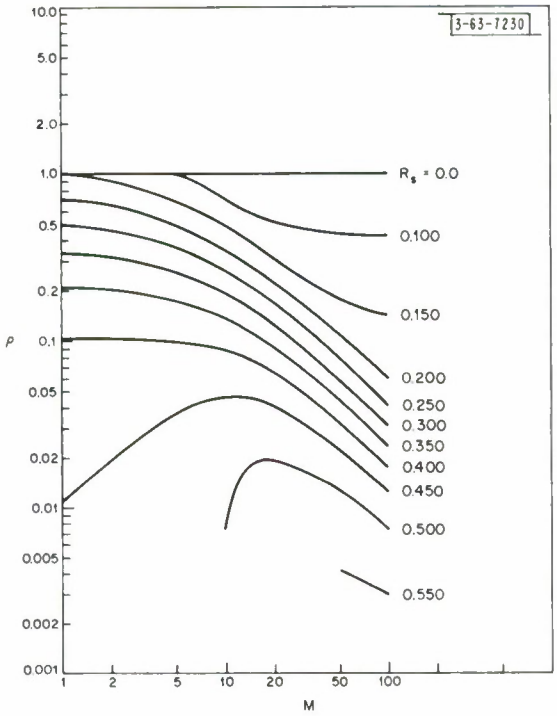


Fig. 22. Maximizing ρ vs number of subchannels (state unknown) – Example 2.

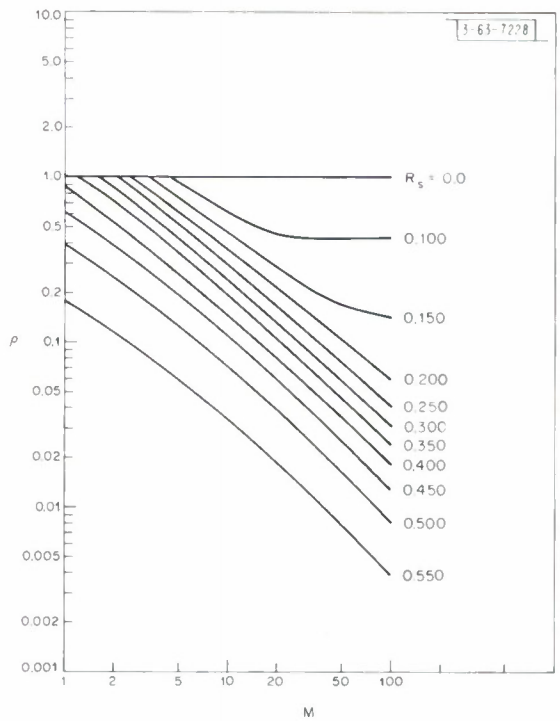


Fig. 23. Maximizing ρ vs number of subchannels (state known) – Example 2.

This is just Eq. (3-14) for the state known case. Since Eq. (4-7) is derived using maximum-likelihood decoding at the receiver, it is also true for maximum a posteriori probability (MAP) decoding when the inputs are equiprobable. Recall, too, that P_e is an average probability of error over an ensemble of codes.

For a particular code, with block length N, the probability of error p(e) satisfies

$$p(e) \geqslant p(e, a^N) = p(a^N)\, p(e/a^N) \tag{4-8}$$

where a^N refers to N consecutive occurrences of the worst state a. Since our channel is memoryless,

$$p(a^N) = [p(a)]^N \quad . \tag{4-9}$$

Define

$$H(e/a^N) = -p(e/a^N) \ln p(e/a^N) - [1 - p(e/a^N)] \ln [1 - p(e/a^N)] \quad . \tag{4-10}$$

Then, letting $W = \exp[NMR_s]$ be the number of (equiprobable) code words, we have[2]

$$H(e/a^N) + p(e/a^N) \ln (W - 1) \geqslant NM(R_s - C_a) \quad . \tag{4-11}$$

Thus,

$$p(e/a^N) \geqslant \frac{NM(R_s - C_a) - H(e/a^N)}{\ln (W - 1)} > \frac{NM(R_s - C_a)}{NMR_s} - \frac{\ln 2}{NMR_s}$$

and

$$\frac{R_s - C_a}{R_s} - \frac{\ln 2}{NMR_s} < p(e/a^N) \leqslant 1 \quad . \tag{4-12}$$

Since inequalities in Eqs. (4-8) and (4-12) hold for each code in an ensemble, they must hold after being averaged over the ensemble of codes. Thus,† Eqs. (4-7), (4-8), and (4-9) become

$$[p(a)]^N P_{e/a^N} \leqslant P_e \leqslant \exp[-NE'_M(R_s)] \tag{4-13}$$

and Eq. (4-12) becomes

$$\frac{R_s - C_a}{R_s} - \frac{\ln 2}{NMR_s} \leqslant P_{e/a^N} \leqslant 1 \quad . \tag{4-14}$$

From Eq. (4-13), we get

$$E'_M(R_s) \leqslant -\ln p(a) - \frac{1}{N} \ln P_{e/a^N} \tag{4-15}$$

for all N. Passing to the limit $N \to \infty$, we get

$$E'_M(R_s) \leqslant -\ln p(a) - \lim_{N \to \infty} \left(\frac{1}{N} \ln P_{e/a^N}\right) \quad . \tag{4-16}$$

From Eq. (4-14), we obtain

$$\lim_{N \to \infty} \left(\frac{1}{N} \ln P_{e/a^N}\right) = 0 \quad . \tag{4-17}$$

† We denote the ensemble averages of p(e) and $p(e/a^N)$ by P_e and P_{e/a^N}.

Thus, Eqs. (4-16) and (4-17) combine to give us

$$E'_M(R_s) \leqslant -\ln p(a)$$

as required. By Theorem 3.5 [$E'_M(R_s)$ is denoted $E^g(R)$ in the statement of the theorem], we have also

$$E_M(R_s) \leqslant E'_M(R_s) \tag{4-18}$$

for all M.

We note that $C_a = 0$ in Example 1, and $C_a = 0.1887$ bit in Example 2. Hence, the bounded behavior of the RCE's is seen in all the curves of Figs. 13 and 14, and in those curves in Figs. 20 and 21 for which $R_s > 0.1887$ bit.

It will now be convenient to restate some of the results of Chapter 3 in MSCC channel notation, and to provide further definitions which will be useful in the sequel. Following Eq. (3-34), we define

$$F'_M(\rho, \vec{p}) = \sum_{\alpha \in \Lambda} p(\alpha) \left\{ \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p_\alpha(j/k)^{1/1+\rho} \right]^{1+\rho} \right\}$$

$$= \sum_{\alpha \in \Lambda} p(\alpha) \left\{ \sum_{y_1, \ldots, y_M} \left[\sum_{x_1, \ldots, x_M} p(x_1, \ldots, x_M) \prod_{i=1}^{M} p_\alpha(y_i/x_i)^{1/1+\rho} \right]^{1+\rho} \right\} . \tag{4-19}$$

Then, we define

$$E'_{oM}(\rho, \vec{p}) = -\ln F'_M(\rho, \vec{p}) \tag{4-20}$$

$$E'_M(\rho, \vec{p}, R_s) = -\rho M R_s + E'_{oM}(\rho, \vec{p}) \tag{4-21}$$

and

$$E'_M(R_s) = \max_{\substack{0 \leqslant \rho \leqslant 1 \\ \vec{p} \in P}} E'_M(\rho, \vec{p}, R_s) \tag{4-22}$$

where, again, P is the space of all input probability vectors. If we define

$$F_{\alpha M}(\rho, \vec{p}) = \sum_{y_1, \ldots, y_M} \left[\sum_{x_1, \ldots, x_M} p(x_1, \ldots, x_M) \prod_{i=1}^{M} p_\alpha(y_i/x_i)^{1/1+\rho} \right]^{1+\rho} \tag{4-23}$$

then

$$F'_M(\rho, \vec{p}) = \sum_{\alpha \in \Lambda} p(\alpha)\, F_{\alpha M}(\rho, \vec{p}) \quad . \tag{4-24}$$

Finally, we define

$$E_{o\alpha M}(\rho, \vec{p}) = -\ln F_{\alpha M}(\rho, \vec{p}) \quad . \tag{4-25}$$

If $M = 1$, $\vec{p}$ becomes $\vec{p}_s$ (a subchannel input probability vector), and we shall generally drop the M subscript on $E_{o\alpha M}$ and $F_{\alpha M}$. Thus, $E_{o\alpha}(\rho, \vec{p}_s) = E_{o\alpha 1}(\rho, \vec{p})$, and $F_{\alpha}(\rho, \vec{p}_s) = F_{\alpha 1}(\rho, \vec{p})$ are quantities relating to a single subchannel.

Theorem 4.4.

Suppose there is a worst subchannel state a, and $R_s > C_a$. Let ρ'_M be the value of ρ which achieves the maximum required by the definition of $E'_M(R_s)$ [Eq. (4-22)]. Then,

$$\lim_{M\to\infty} \rho'_M = 0 \tag{4-26}$$

and

$$\lim_{M\to\infty} (\rho'_M)^2 M = 0 \quad . \tag{4-27}$$

Proof.

From Eqs. (4-6), (4-21), and (4-22),

$$-\rho M R_s + E'_{oM}(\rho, \vec{p}) \leqslant -\ln p(a) \tag{4-28}$$

for all $R_s > C_a$, $0 \leqslant \rho \leqslant 1$, $\vec{p} \in P$. Thus,

$$E'_{oM}(\rho, \vec{p}) \leqslant -\ln p(a) + \rho M C_a \tag{4-29}$$

for all $0 \leqslant \rho \leqslant 1$, $\vec{p} \in P$. Equations (4-28) and (4-29) combine to give

$$-\rho M R_s + E'_{oM}(\rho, \vec{p}) \leqslant -\ln p(a) + \rho M(C_a - R_s) \tag{4-30}$$

for all $0 \leqslant \rho \leqslant 1$, $\vec{p} \in P$, and $R_s > C_a$. Since $E'_M(R_s)$ is non-negative, Eqs. (4-30), (4-21), and (4-22) combine to give

$$0 \leqslant E'_M(R_s) \leqslant -\ln p(a) + \rho'_M M(C_a - R_s)$$

and thus we obtain

$$0 \leqslant \rho'_M \leqslant \frac{-\ln p(a)}{M(R_s - C_a)} \quad . \tag{4-31}$$

From Eq. (4-31), we get Eqs. (4-26) and (4-27) directly.

We note that the proof could just as well be carried through if ρ'_M were the value of ρ which achieves the maximum required by the definition of $E_M(R_s)$.

The behavior of ρ'_M just proved is illustrated in the curves of Figs. 15 and 16, and in those curves of Figs. 22 and 23 for which $R_s > 0.1887$ bit. Since the slopes of these curves are all minus one for large M, they suggest that indeed $\rho'_M M$ is equal to a constant independent of M for M sufficiently large. However, the constant is smaller than that suggested by the rightmost expression in Eq. (4-31).

Although the assertions of the theorem just proved are technical in the sense that they are not subject to immediate physical interpretation, their consequences are quite striking. One such consequence is given by the following theorem.

Theorem 4.5.

If there exists a worst subchannel state a, $R_s > C_a$, and Λ is finite, then,

$$\lim_{M\to\infty} [E'_M(R_s) - E_M(R_s)] = 0 \quad . \tag{4-32}$$

Proof.

The result follows directly from Eq. (4-26) and Theorem 3.6.

The result of Theorem 4.5 is illustrated in Figs. 12 and 19. For the examples computed, Eq. (4-32) appears to hold at all rates. Note that the difference in RCE's is not monotone in M.

Since the difference $E'_M(R_s) - E_M(R_s)$ approaches zero with increasing M, under the conditions stated, it is natural to ask whether either term (and hence both terms) approaches a limit under similar conditions. This question will be answered in the affirmative after some labor.

First, it will be necessary to study the properties of the input probability vector $\vec{p}'_M$, which achieves the maximum in Eq. (4-22). Now, from Eqs. (4-20), (4-21), and (4-22), we have

$$\begin{aligned} E'_M(R_s) &= \max_{\substack{0\leqslant\rho\leqslant 1\\ \vec{p}\epsilon P}} [-\rho M R_s - \ln F'_M(\rho, \vec{p}] \\ &= \max_{0\leqslant\rho\leqslant 1} [-\rho M R_s - \ln \min_{\vec{p}\epsilon P} F'_M(\rho, \vec{p})] \quad . \end{aligned} \tag{4-33}$$

Thus, we shall be concerned with the properties of probability vectors (distributions) which minimize $F'_M(\rho, \vec{p})$.

Definition.

A distribution p over a product space V^M is said to have <u>permutational symmetry</u> if

$$p(v_1, \ldots, v_M) = p(v_{j_1}, \ldots, v_{j_M}) \tag{4-34}$$

for all permutations $\{j_\ell\}_{\ell=1}^{M}$ of the integers from 1 to M.

Theorem 4.6.

For an MSCC channel, the $\min_{\vec{p}} F'_M(\rho, \vec{p})$ may, for any ρ, $0 \leqslant \rho \leqslant 1$, be achieved with an input distribution having permutational symmetry.

Proof.

Let

$$F'_M(\rho, \vec{p}*) = \min_{\vec{p}} F'_M(\rho, \vec{p}) \tag{4-35}$$

where $\vec{p}*$ is implicitly a function of ρ and M, and we may write

$$\vec{p}* = p*(x_1, \ldots, x_M) \quad .$$

Define

$$\vec{p}^{\,r} = p^r(x_1, \ldots, x_M) = \frac{1}{M!} \sum_{PT} p*\left(x_{j_1}, \ldots, x_{j_M}\right) \tag{4-36}$$

where $\sum_{PT}$ denotes the sum over all M! possible permutations of the integers from 1 to M. For any permutation $\{j_\ell\}_{\ell=1}^{M}$ and any input distribution $p(x_1, \ldots, x_M)$, we have

$$\sum_{y_1, \ldots, y_M} \left[\sum_{x_1, \ldots, x_M} p\left(x_{j_1}, \ldots, x_{j_M}\right) \prod_{i=1}^{M} p_\alpha(y_i/x_i)^{1/1+\rho} \right]^{1+\rho}$$

$$= \sum_{y_1, \ldots, y_M} \left[\sum_{x_1, \ldots, x_M} p\left(x_{j_1}, \ldots, x_{j_M}\right) \prod_{i=1}^{M} p_\alpha\left(y_{j_i}/x_{j_i}\right)^{1/1+\rho} \right]^{1+\rho}$$

$$= \sum_{y_{j_1}, \ldots, y_{j_M}} \left[\sum_{x_{j_1}, \ldots, x_{j_M}} p\left(x_{j_1}, \ldots, x_{j_M}\right) \prod_{i=1}^{M} p_\alpha\left(y_{j_i}/x_{j_i}\right)^{1/1+\rho} \right]^{1+\rho}$$

$$= \sum_{y_1, \ldots, y_M} \left[\sum_{x_1, \ldots, x_M} p(x_1, \ldots, x_M) \prod_{i=1}^{M} p_\alpha(y_i/x_i)^{1/1+\rho} \right]^{1+\rho} \tag{4-37}$$

Since $F_{\alpha M}(\rho, \vec{p})$ is a convex downward function† of $\vec{p}$,

$$F_{\alpha M}(\rho, \vec{p}^{\,r}) \leqslant \frac{1}{M!} \sum_{PT} \sum_{y_1, \ldots, y_M} \left[\sum_{x_1, \ldots, x_M} p^*\left(x_{j_1}, \ldots, x_{j_M}\right) \prod_{i=1}^{M} p_\alpha(y_i/x_i)^{1/1+\rho} \right]^{1+\rho}$$

$$= \sum_{y_1, \ldots, y_M} \left[\sum_{x_1, \ldots, x_M} p^*(x_1, \ldots, x_M) \prod_{i=1}^{M} p_\alpha(y_i/x_i)^{1/1+\rho} \right]^{1+\rho}$$

$$= F_{\alpha M}(\rho, \vec{p}^{\,*}) \qquad \text{all } \alpha \in \Lambda \tag{4-38}$$

where we use Eq. (4-37) with $p = p^*$. From Eqs. (4-38) and (4-24),

$$F'_M(\rho, \vec{p}^{\,r}) \leqslant F'_M(\rho, \vec{p}^{\,*}) \quad . \tag{4-39}$$

But, by Eq. (4-35),

$$F'_M(\rho, \vec{p}^{\,*}) \leqslant F'_M(\rho, \vec{p}^{\,r}) \quad . \tag{4-40}$$

Hence,

$$F'_M(\rho, \vec{p}^{\,r}) = F'_M(\rho, \vec{p}^{\,*}) = \min_{\vec{p} \in P} F'_M(\rho, \vec{p}) \tag{4-41}$$

as required.

We note that the essential property of the MSCC channel which allows us to prove the result is that the subchannel state distribution $p(\alpha_1, \ldots, \alpha_M)$ has permutational symmetry. This may be demonstrated by a minor modification of the proof of Theorem 4.6.

† See Ref. 1, Theorem 4.

Let P_r be the space of all input probability vectors with permutational symmetry. We have shown that for an MSCC channel,

$$\min_{\vec{p} \epsilon P} F'_M(\rho, \vec{p}) = \min_{\vec{p} \epsilon P_r} F'_M(\rho, \vec{p}) \tag{4-42}$$

for all ρ, $0 \leqslant \rho \leqslant 1$.

Definition.

If $p(\eta)$ is a probability distribution on X_s, and

$$p(x_1, \ldots, x_M) = \prod_{i=1}^{M} p(x_i) \tag{4-43}$$

then we say $p(x_1, \ldots, x_M)$ is a product distribution. Let D be the set of all product distributions.† Clearly, $D \subset P_r$. If $\vec{p}_s$ is the probability vector corresponding to $p(\eta)$, $\vec{p}$ is the probability vector corresponding to $p(x_1, \ldots, x_M)$, and Eq. (4-43) holds, then we shall write

$$\vec{p} = (\vec{p}_s)^M \quad . \tag{4-44}$$

We shall also write $\vec{p} \epsilon D$. Finally, we shall denote the set of subchannel input probability vectors by U.

We shall now examine the properties of functions from which $E'_M(R)$ is derived if $\vec{p} \epsilon D$. From Eqs. (4-23) and (4-43),

$$F_{\alpha M}(\rho, \vec{p}) = \left\{ \sum_{y_1} \left[\sum_{x_1} p(x_1)\, p_\alpha(y_1/x_1)^{1/1+\rho} \right]^{1+\rho} \right\}^M$$

$$= [F_\alpha(\rho, \vec{p}_s)]^M \quad . \tag{4-45}$$

Henceforth, we shall assume that $X_s = \{1, \ldots, L\}$, and $Y_s = \{1, \ldots, Q\}$. Thus, we have

$$F_\alpha(\rho, \vec{p}_s) = \sum_{q=1}^{Q} \left[\sum_{\ell=1}^{L} p(\ell)\, p_\alpha(q/\ell)^{1/1+\rho} \right]^{1+\rho} \quad . \tag{4-46}$$

Using Eqs. (4-24), (4-45), and (4-25), we have

$$F'_M(\rho, \vec{p}) = \sum_{\alpha \epsilon \Lambda} p(\alpha)\, [F_\alpha(\rho, \vec{p}_s)]^M$$

$$= \sum_{\alpha \epsilon \Lambda} p(\alpha) \exp[-ME_{o\alpha}(\rho, \vec{p}_s)] \tag{4-47}$$

for all $\vec{p} \epsilon D$ and all ρ, $0 \leqslant \rho \leqslant 1$. More generally, Eqs. (4-24) and (4-25) yield

$$F'_M(\rho, \vec{p}) = \sum_{\alpha \epsilon \Lambda} p(\alpha) \exp[-E_{o\alpha M}(\rho, \vec{p})] \tag{4-48}$$

for all $\vec{p} \epsilon P$ and all ρ, $0 \leqslant \rho \leqslant 1$.

---

† Note that this is a more restrictive definition than in Chapter 2, because here we ask that all the individual subchannel marginal distributions be the same.

Because of the functional form of $E_{o\alpha}(\rho, \vec{p}_s)$ and the fact that the sums used in its definition are finite, all derivatives of $E_{o\alpha}(\rho, \vec{p}_s)$ with respect to ρ are continuous with respect to ρ and the L + 1 probability vectors involved in the definition of $E_{o\alpha}$ (one subchannel input probability vector and L subchannel output probability vectors, each conditioned on one input). Hence, by an argument identical to that used in the proof of part (1) of Theorem 2.3, there exists a positive constant B(L, Q) for which

$$\left| \frac{\partial^2 E_{o\alpha}}{\partial \rho^2} \right| \leqslant B(L, Q) \tag{4-49}$$

for all ρ, $0 \leqslant \rho \leqslant 1$, all subchannel input distributions $p(\ell)$, and all conditional probability distributions $p_\alpha(q/\ell)$.

Theorem 4.7. (Gallager)

Consider a channel with $X = \{1, \ldots, K\}$, $Y = \{1, \ldots, J\}$, and transition probabilities $p_\alpha(j/k)$, $1 \leqslant j \leqslant J$, $1 \leqslant k \leqslant K$. Let $\vec{p} = [p(1), \ldots, p(K)]$ be an input probability vector, and assume that the average mutual information

$$I_{\alpha M}(\vec{p}) = \sum_{k=1}^{K} \sum_{j=1}^{J} p(k)\, p_\alpha(j/k) \ln \frac{p(j/k)}{\sum_{i=1}^{K} p(i)\, p_\alpha(j/i)} \tag{4-50}$$

is nonzero. Define

$$E_{o\alpha M}(\rho, \vec{p}) = -\ln \left\{ \sum_{j=1}^{J} \left[\sum_{k=1}^{K} p(k)\, p_\alpha(j/k)^{1/1+\rho} \right]^{1+\rho} \right\} . \tag{4-51}$$

Then, for $\rho \geqslant 0$,

$$E_{o\alpha M}(o, \vec{p}) = 0 \tag{4-52}$$

$$E_{o\alpha M}(\rho, \vec{p}) > 0 \qquad \text{for } \rho > 0 \tag{4-53}$$

$$\frac{\partial E_{o\alpha M}(\rho, \vec{p})}{\partial \rho} > 0 \qquad \text{for } \rho > 0 \tag{4-54}$$

$$\left. \frac{\partial E_{o\alpha M}(\rho, \vec{p})}{\partial \rho} \right|_{\rho=0} = I_{\alpha M}(\vec{p}) \tag{4-55}$$

$$\frac{\partial^2 E_{o\alpha M}(\rho, \vec{p})}{\partial \rho^2} \leqslant 0 \tag{4-56}$$

with equality in Eq. (4-56) if and only if both of the following conditions are satisfied:

(1) $p_\alpha(j/k)$ is independent of k for j, k such that $p(k)\, p_\alpha(j/k) \neq 0$.

(2) $\sum_{k:p_\alpha(j/k)\neq 0} p(k)$ is independent of j.

This is Theorem 2 of Gallager,† so the proof will not be given here. If $I_{\alpha M}(\vec{p}) = 0$, input and output are independent.[3] Thus, $p_\alpha(j/k) = p_\alpha(j)$ if $p(k) \neq 0$ $(1 \leqslant j \leqslant J)$. Then, using Eq. (4-51), we have $E_{o\alpha M}(\rho, \vec{p}) = 0$ for all $\rho \geqslant 0$.

Note that if the channel referred to in Theorem 4.7 is MSCC, the definition of $E_{o\alpha M}$ by Eq. (4-51) is consistent with the definition of $E_{o\alpha M}$ by Eqs. (4-23) and (4-25). Note, too, that all the results of Theorem 4.7 apply to a single subchannel as well as to the whole channel. In our notation, this means that the results hold if all M's are deleted and we make the following changes:

$j \to q$

$k \to \ell$

$J \to Q$

$K \to L$

$\vec{p} \to \vec{p}_s$

For $\vec{p}_s \in U$, define‡

$$EN(t, \vec{p}_s, R_s) = -tR_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-tI_\alpha(\vec{p}_s)] \tag{4-57}$$

and

$$EN(R_s) = \underset{0 \leqslant t < \infty}{\text{l.u.b.}} \max_{\vec{p}_s \in U} EN(t, \vec{p}_s, R_s) \quad . \tag{4-58}$$

Theorem 4.8.

(a) Suppose there exists a worst subchannel state a, and $C_a < R_s < C_1'$. Then, there exists a positive number t_o with

$$EN(R_s) = \max_{\vec{p}_s \in U} EN(t_o, \vec{p}_s, R_s) \quad . \tag{4-59}$$

(b) If, in addition, there exists a single-subchannel probability vector $\vec{p}_s$ with

$$C_\alpha = I_\alpha(\vec{p}_s) \qquad \text{all } \alpha \in \Lambda \tag{4-60}$$

then,

$$EN(R_s) = -t_o R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-t_o C_\alpha] \tag{4-61}$$

---

† See Ref. 1, p. 6.

‡Note that for each $\vec{p}_s$ and R_s, $\sum_{a \in \Lambda} p(a) \exp\{t[R_s - I_a(\vec{p}_s)]\}$ is the moment generating function $g(t, \vec{p}_s, R_s)$ associated with the random variable $R_s - I_a(\vec{p}_s)$. Since by Eq. (4-57) $EN(t, \vec{p}_s, R_s) = -\ln g(t, \vec{p}_s, R_s)$, some of the properties of $EN(t, \vec{p}_s, R_s)$ which we shall derive may be obtained from the theory of moment generating functions. See, for example, Chapter 8 of Ref. 2.

where t_o is given implieitly and uniquely by

$$R_s = \frac{\sum_{\alpha \in \Lambda} p(\alpha)\, C_\alpha \exp[-t_o C_\alpha]}{\sum_{\alpha \in \Lambda} p(\alpha) \exp[-t_o C_\alpha]} \quad . \tag{4-62}$$

Proof.

(a) All that really needs to be proved is that

$$EN(R_s) \neq \limsup_{t \to \infty} \max_{\vec{p}_s \in U} EN(t, \vec{p}_s, R_s) \quad . \tag{4-63}$$

From Eq. (4-57),

$$\begin{aligned} EN(t, \vec{p}_s, R_s) &\leqslant -t\,[R_s - I_a(\vec{p}_s)] - \ln p(a) \\ &\leqslant -t(R_s - C_a) - \ln p(a) \end{aligned} \tag{4-64}$$

for all $\vec{p}_s \in U$. If we define

$$t^* = -\frac{\ln p(a)}{R_s - C_a} \tag{4-65}$$

$t > t^*$ implies

$$EN(t, \vec{p}_s, R_s) < 0 = EN(o, \vec{p}_s, R_s) \tag{4-66}$$

for all $\vec{p}_s \in U$. Thus, we have proven Eq. (4-63).

(b) Using the definition of C_α, we obtain

$$\max_{\vec{p}_s \in U} \left\{ -tR_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-tI_\alpha(\vec{p}_s)] \right\} \leqslant -tR_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-tC_\alpha] \tag{4-67}$$

for all $t \geqslant 0$, with equality if

$$I_\alpha(\vec{p}_s) = C_\alpha \qquad \text{all } \alpha \in \Lambda \quad . \qquad \text{[Eq. (4-60)]}$$

Thus, if Eq. (4-60) holds, we have from Eqs. (4-57) and (4-58) that

$$\begin{aligned} EN(R_s) &= \underset{0 \leqslant t < \infty}{\text{l. u. b.}} \left\{ -tR_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-tC_\alpha] \right\} \\ &= -\ln\left(\underset{0 \leqslant t < \infty}{\text{g. l. b.}} \left\{ \sum_{\alpha \in \Lambda} p(\alpha) \exp[t(R_s - C_\alpha)] \right\} \right) \quad . \end{aligned} \tag{4-68}$$

Let

$$\varphi(t) = \sum_{\alpha \in \Lambda} p(\alpha) \exp[t(R_s - C_\alpha)] \quad . \tag{4-69}$$

Then,

$$\varphi'(t) = \frac{d\varphi}{dt} = \sum_{\alpha \in \Lambda} p(\alpha)\,(R_s - C_\alpha) \exp[t(R_s - C_\alpha)] \tag{4-70}$$

and

$$\varphi''(t) = \frac{d^2\varphi}{dt^2} = \sum_{\alpha \in \Lambda} p(\alpha)\,(R_s - C_\alpha)^2 \exp[t(R_s - C_\alpha)] \quad . \tag{4-71}$$

From Eq. (4-70),

$$\varphi'(0) = R_s - \sum_{\alpha \in \Lambda} p(\alpha)\, C_\alpha$$

$$= R_s - C_1' < 0 \quad . \tag{4-72}$$

From Eq. (4-71),

$$\varphi''(t) > 0 \tag{4-73}$$

for all $t \geqslant 0$. Thus, $\varphi(t)$ is strictly convex downward and must, by Eqs. (4-72), (4-68), and part (a) of this theorem, have its g. l. b. at its stationary point. Thus, setting $\varphi'(t_o)$ to zero, we obtain Eq. (4-62).

For $\vec{p} \in D$, define

$$\overline{E}_M(\rho, \vec{p}, R_s) = -\rho M R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-\rho M I_\alpha(\vec{p}_s)] \tag{4-74}$$

where $\vec{p} = (\vec{p}_s)^M$, and

$$\overline{E}_M(R_s) = \max_{\substack{0 \leqslant \rho \leqslant 1 \\ \vec{p} \in D}} \overline{E}_M(\rho, \vec{p}, R_s) \quad . \tag{4-75}$$

Theorem 4.9.

Suppose there exists a worst subchannel state a, and $C_a < R_s < C_1'$. Then, if t_o is defined by Eq. (4-59), and $\overline{\rho}(M)$ is the value of ρ which achieves the maximum in Eq. (4-75), we have for $M \geqslant t_o$

$$\overline{E}_M(R_s) = EN(R_s) \tag{4-76}$$

and

$$\overline{\rho}(M) = t_o/M \quad . \tag{4-77}$$

Furthermore,

$$\lim_{M \to \infty} \overline{E}_M(R_s) = EN(R_s) \quad . \tag{4-78}$$

Proof.

A comparison of Eqs. (4-74) and (4-75) with Eqs. (4-57) and (4-58) makes Eqs. (4-76) and (4-77) obvious consequences of Theorem 4.8. Equation (4-78) is a consequence of Eq. (4-76).

Define

$$\tilde{E}_M(R_s) = \max_{\substack{0\leq\rho\leq 1 \\ \vec{p}\in D}} E'_M(\rho, \vec{p}, R_s)$$

$$= \max_{\substack{0\leq\rho\leq 1 \\ \vec{p}\in D}} [-\rho M R_s - \ln F'_M(\rho, \vec{p})]$$

$$= \max_{0\leq\rho\leq 1} [-\rho M R_s - \ln \min_{\vec{p}\in D} F'_M(\rho, \vec{p})] \quad . \tag{4-79}$$

This definition differs from that of $E'_M(R_s)$ only in that the maximization over input probability vectors is over D rather than P.

Note that Eq. (4-33) and Theorem 4.6 imply

$$E'_M(R_s) = \max_{\substack{0\leq\rho\leq 1 \\ \vec{p}\in P_r}} [-\rho M R_s - \ln F'_M(\rho, \vec{p})]$$

$$= \max_{0\leq\rho\leq 1} [-\rho M R_s - \ln \min_{\vec{p}\in P_r} F'_M(\rho, \vec{p})] \quad . \tag{4-80}$$

Equations (4-74) and (4-75) imply

$$\bar{E}_M(R_s) = \max_{\substack{0\leq\rho\leq 1 \\ \vec{p}_s\in U}} \left\{-\rho M R_s - \ln \sum_{\alpha\in\Lambda} p(\alpha) \exp[-\rho M I_\alpha(\vec{p}_s)]\right\} \quad . \tag{4-81}$$

Theorem 4.10.

(a) If there exists a worst subchannel state a, and $C_a < R_s < C'_1$, then

$$\lim_{M\to\infty} E'_M(R_s) = \lim_{M\to\infty} \tilde{E}_M(R_s) = EN(R_s) \quad . \tag{4-82}$$

(b) Suppose there exists a single subchannel input probability vector $\vec{p}_s$ satisfying Eq. (4-60). Associate $p(\ell)$ with $\vec{p}_s$ and assume that the subchannel conditional probability distributions $p_\alpha(q/\ell)$ satisfy

$$p_\alpha(q/\ell) \text{ is independent of } \ell \tag{4-83}$$

for q, ℓ with $p(\ell)\, p_\alpha(q/\ell) \neq 0$ and all $\alpha \in \Lambda$. Assume, too, that

$$\sum_{\ell : p_\alpha(q/\ell)\neq 0} p(\ell) \text{ is independent of } q \tag{4-84}$$

for all $\alpha \in \Lambda$. Then, there exists a positive number t_o defined by Eq. (4-62) such that $M \geq t_o$ implies

$$E'_M(R_s) = \lim_{T\to\infty} E'_T(R_s) \quad . \tag{4-85}$$

Proof.

(a) From Taylor's theorem and Theorem 4.7, we have for any $0 < \rho \leqslant 1$, $\vec{p} \in P$, and $\vec{p}_s \in U$,

$$E_{o\alpha M}(\rho, \vec{p}) = \rho I_{\alpha M}(\vec{p}) + \frac{\rho^2}{2} \left. \frac{\partial^2 E_{o\alpha M}(\rho, \vec{p})}{\partial \rho^2} \right|_{\xi(\rho)} \leqslant \rho I_{\alpha M}(\vec{p}) \tag{4-86}$$

where $0 < \xi(\rho) < \rho \leqslant 1$, and

$$E_{o\alpha}(\rho, \vec{p}_s) = \rho I_{\alpha}(\vec{p}_s) + \frac{\rho^2}{2} \left. \frac{\partial^2 E_{o\alpha}(\rho, \vec{p}_s)}{\partial \rho^2} \right|_{\xi(\rho)} \tag{4-87}$$

where $0 < \xi(\rho) < \rho \leqslant 1$. From Eq. (4-49) and Theorem 4.7, we get

$$B(L, Q) \leqslant \frac{\partial^2 E_{o\alpha}(\rho, \vec{p}_s)}{\partial \rho^2} \leqslant 0 \quad . \tag{4-88}$$

Thus, from Eqs. (4-48) and (4-86),

$$F'_M(\rho, \vec{p}) \geqslant \sum_{\alpha \in \Lambda} p(\alpha) \exp[-\rho I_{\alpha M}(\vec{p})] \quad . \tag{4-89}$$

From Eq. (4-80), for purposes of minimizing $F'_M(\rho, \vec{p})$, we may assume $\vec{p} \in P_r$. Making this assumption, we define

$$p_i(x_i) = \sum_{\overline{X}_i} p(x_1, \ldots, x_M) \quad . \tag{4-90}$$

Since $\vec{p} \in P_r$, we have

$$p_i(x_i) = p(x_i) \qquad \text{all } i, \ 1 \leqslant i \leqslant M \quad . \tag{4-91}$$

Define

$$p_{DR}(x_1, \ldots, x_M) = \prod_{i=1}^{M} p(x_i) \tag{4-92}$$

and associate $\vec{p}_{DR}$ with $p_{DR}(x_1, \ldots, x_M)$, and $\vec{p}_s$ with $p(x_i)$. Then, by the remark following Eq. (2-29),

$$M I_{\alpha}(\vec{p}_s) \geqslant I_{\alpha M}(\vec{p}) \quad . \tag{4-93}$$

By Eqs. (4-89) and (4-93),

$$F'_M(\rho, \vec{p}) \geqslant \sum_{\alpha \in \Lambda} p(\alpha) \exp[-\rho M I_{\alpha}(\vec{p}_s)] \tag{4-94}$$

for all ρ, $0 \leqslant \rho \leqslant 1$, and all $\vec{p} \in P_r$. From Eqs. (4-79), (4-80), and Theorem 2.3, we get the left inequality below:

$$\tilde{E}_M(R_s) \leqslant E'_M(R_s) \leqslant \overline{E}_M(R_s) \quad . \tag{4-95}$$

The right inequality is obtained from Eqs. (4-80), (4-81), (4-94), and Theorem 2.3.

Let $\bar{\rho}$, $\vec{p}_s'$ be such as to achieve the maximum in Eq. (4-75). From Eq. (4-95),

$$0 \leqslant \overline{E}_M(R_s) - E_M'(R_s) \leqslant \overline{E}_M(R_s) - \tilde{E}_M(R_s) \quad . \tag{4-96}$$

From Eq. (4-79),

$$\overline{E}_M(R_s) - \tilde{E}_M(R_s) \leqslant \overline{E}_M(R_s) - E_M' [\bar{\rho}, (\vec{p}_s')^M, R_s] \quad . \tag{4-97}$$

From Eqs. (4-20), (4-21), (4-47), and (4-87),

$$E_M' [\rho, (\vec{p}_s)^M, R_s] = -\rho M R_s - \ln \sum_{\alpha \epsilon \Lambda} p(\alpha) \exp \left[-\rho M I_\alpha(\vec{p}_s) - \frac{\rho^2 M}{2} \left. \frac{\partial^2 E_{o\alpha}(\rho, \vec{p}_s)}{\partial \rho^2} \right|_{\xi(\rho)} \right] \tag{4-98}$$

for all ρ, $0 \leqslant \rho \leqslant 1$, and $\vec{p}_s \epsilon U$. Using Eq. (4-88), we have

$$-E_M' [\bar{\rho}, (\vec{p}_s')^M, R_s] \leqslant \bar{\rho} M R_s + \ln \left\{ \sum_{\alpha \epsilon \Lambda} p(\alpha) \exp[-\bar{\rho} M I_\alpha(\vec{p}_s')] \right\} + \frac{\bar{\rho}^2 M}{2} B(L, Q) \quad . \tag{4-99}$$

Thus, using Eqs. (4-81) and (4-77), we get

$$\overline{E}_M(R_s) - E_M' [\bar{\rho}, (\vec{p}_s')^M, R_s] \leqslant \frac{\bar{\rho}^2 M}{2} B(L, Q) = \frac{t_o^2}{2M} B(L, Q) \quad . \tag{4-100}$$

Combining Eqs. (4-96), (4-97), and (4-100), we get

$$0 \leqslant \overline{E}_M(R_s) - E_M'(R_s) \leqslant \overline{E}_M(R_s) - \tilde{E}_M(R_s) \leqslant \frac{t_o^2}{2M} B(L, Q) \quad . \tag{4-101}$$

Thus,

$$\lim_{M \to \infty} [\overline{E}_M(R_s) - E_M'(R_s)] = \lim_{M \to \infty} [\overline{E}_M(R_s) - \tilde{E}_M(R_s)] = 0 \quad . \tag{4-102}$$

From Eqs. (4-78) and (4-102), we get Eq. (4-82), as required.

(b) Equations (4-67), (4-74), and (4-75) imply that for $\vec{p}_s$ satisfying Eq. (4-60),

$$\overline{E}_M(R_s) = \max_{0 \leqslant \rho \leqslant 1} \overline{E}_M [\rho, (\vec{p}_s)^M, R_s] \quad . \tag{4-103}$$

Equations (4-83), (4-84), and Theorem 4.7 imply

$$\frac{\partial^2 E_{o\alpha}(\rho, \vec{p}_s)}{\partial \rho^2} = 0 \tag{4-104}$$

for all ρ, $0 \leqslant \rho \leqslant 1$. From Eqs. (4-74), (4-98), and (4-104), for $\vec{p}_s$ satisfying Eq. (4-60) and all ρ, $0 \leqslant \rho \leqslant 1$, we have

$$\overline{E}_M [\rho, (\vec{p}_s)^M, R_s] = E_M' [\rho, (\vec{p}_s)^M, R_s] \quad . \tag{4-105}$$

Hence, by Eqs. (4-95) and (4-103),

$$\overline{E}_M(R_s) - E_M'(R_s) = 0 \qquad \text{all } M \quad . \tag{4-106}$$

Thus, for $M \geqslant t_o$, Eqs. (4-76), (4-78), and (4-106) give the result.

The curves of Fig. 14 illustrate both parts of Theorem 4.10. In this case, the $\vec{p}_s$ of part (b) is given by $\vec{p}_s = (\frac{1}{2}, \frac{1}{2})$. The curves of Fig. 21, corresponding to $R_s > 0.1887$ bit, illustrate part (a) of the theorem. Part (b) of Theorem 4.8 applies to both examples, with $\vec{p}_s = (\frac{1}{2}, \frac{1}{2})$.

Corollary 1.

Suppose Λ is finite. Under the assumptions of part (a) of Theorem 4.10,

$$\lim_{M\to\infty} E_M(R_s) = EN(R_s) \quad . \tag{4-107}$$

Proof.

Combine Theorems 4.10 and 4.5.

The curves of Figs. 13 and 20, corresponding to $R_s > 0.1887$ bit, illustrate the corollary. Again, part (b) of Theorem 4.8 applies to both examples with $\vec{p}_s = (\frac{1}{2}, \frac{1}{2})$.

Theorem 4.11.

If for each ρ, $0 \leqslant \rho \leqslant 1$, there exists a single $\vec{p}'_s \in U$ with

$$\min_{\vec{p}_s \in U} F_\alpha(\rho, \vec{p}_s) = F_\alpha(\rho, \vec{p}'_s) \tag{4-108}$$

for all $\alpha \in \Lambda$, then

$$E'_M(R_s) = \tilde{E}_M(R_s) \tag{4-109}$$

for all M.

Proof.

Theorem 5 of Gallager (see Ref. 1, p.10) and our Eqs. (4-20) and (4-45) imply

$$\min_{\vec{p}\in P} F_{\alpha M}(\rho, \vec{p}) = \min_{\vec{p}\in D} F_{\alpha M}(\rho, \vec{p}) = \left[\min_{\vec{p}_s \in U} F_\alpha(\rho, \vec{p}_s)\right]^M \quad . \tag{4-110}$$

Equations (4-108), (4-44), (4-45), and (4-110) imply

$$\min_{\vec{p}\in P} F_{\alpha M}(\rho, \vec{p}) = F_{\alpha M}[\rho, (\vec{p}'_s)^M] \tag{4-111}$$

for all $\alpha \in \Lambda$. Thus, from Eqs. (4-24) and (4-111),

$$\begin{aligned} \min_{\vec{p}\in P} F'_M(\rho, \vec{p}) &= F'_M[\rho, (\vec{p}'_s)^M] \\ &= \min_{\vec{p}\in D} F'_M(\rho, \vec{p}) \end{aligned} \tag{4-112}$$

and from Eqs. (4-33), (4-79), and (4-112) we have our result.

For both Examples 1 and 2, Eq. (4-108) is satisfied for all ρ, $0 \leqslant \rho \leqslant 1$, if $\vec{p}_s = (\frac{1}{2}, \frac{1}{2})$. Thus, Eq. (4-109) holds for our examples.

One might wonder if Eq. (4-109) holds for all MSCC channels. The answer, although far from obvious, is that it does not. An example which demonstrates this fact is discussed in Appendix F.

Thus far, we have considered only the case where there exists a worst subchannel state a, and $C_a < R_s < C_1'$. Now, if $R_s > C_1'$,

$$E_M(R_s) = E_M'(R_s) = 0$$

for all M, by the converse to the coding theorem. It remains to investigate the behavior of $E_M'(R_s)$ when $R_s < C_a$ and $R_s < C_1'$.

Theorem 4.12.

If there exists a subchannel input distribution $\vec{p}_s$ and a positive number I, for which

$$I_\alpha(\vec{p}_s) \geqslant I > R_s \qquad \text{all } \alpha \in \Lambda \tag{4-113}$$

then,

$$E_M'(R_s) \to \infty \qquad \text{as } M \to \infty \quad . \tag{4-114}$$

If, in addition, Λ is finite, then

$$E_M(R_s) \to \infty \qquad \text{as } M \to \infty \quad . \tag{4-115}$$

Proof.

By Eqs. (4-87) and (4-88) for $\vec{p}_s$ satisfying Eq. (4-113), all $\rho \geqslant 0$ and all $\alpha \in \Lambda$, we have

$$E_{o\alpha}(\rho, \vec{p}_s, R_s) = \rho I_\alpha(\vec{p}_s) + \frac{\rho^2}{2} \left. \frac{\partial^2 E_{o\alpha}(\rho, \vec{p}_s)}{\partial \rho^2} \right|_{\xi(\rho)}$$

$$\geqslant \rho I - \frac{\rho^2}{2} B(L, Q) \quad . \tag{4-116}$$

Hence, by Eqs. (4-20), (4-21), (4-47), and (4-116),

$$E_M'[\rho, (\vec{p}_s)^M, R_s] \geqslant -\rho M R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp\left[-\rho M I + \frac{\rho^2 M}{2} B(L, Q)\right]$$

$$\geqslant \rho M(I - R_s) - \frac{\rho^2 M}{2} B(L, Q) \quad .$$

Let

$$\rho^* = \min\left[1, \frac{I - R_s}{B(L, Q)}\right] \quad .$$

Then,

$$E_M'[\rho^*, (\vec{p}_s)^M, R_s] \geqslant \frac{M}{2} \min\left[(I - R_s), \frac{(I - R_s)^2}{B(L, Q)}\right] \quad . \tag{4-117}$$

Clearly,

$$\frac{M}{2} \min\left[(I - R_s), \frac{(I - R_s)^2}{B(L, Q)}\right] \to \infty \qquad \text{as } M \to \infty \quad . \tag{4-118}$$

Thus, Eqs. (4-22), (4-117), and (4-118) imply that Eq. (4-114) holds. If Λ is finite, Eqs. (3-52) and (4-114) combine to yield Eq. (4-115).

The curves of Figs. 20 and 21, corresponding to $R_s < 0.1887 = I$, illustrate the theorem $[\vec{p}_s = (\frac{1}{2}, \frac{1}{2})]$.

Theorem 4.13.

Let Λ be finite. If for each subchannel input distribution $\vec{p}_s$ there exists $\beta \in \Lambda$ with

$$I_\beta(\vec{p}_s) < R_s \tag{4-119}$$

then,

$$E_M(R_s) \leqslant E'_M(R_s) \leqslant d \tag{4-120}$$

for all M, where†

$$d = \max_{\alpha \in \Lambda} \{-\ln p(\alpha)\} < \infty \quad . \tag{4-121}$$

Proof.

Let

$$E'_M(R_s) = E'_M(\rho', \vec{p}', R_s)$$

where $\vec{p}'$ is chosen to have permutation symmetry. Using Eqs. (4-33) and (4-89), we have

$$E'_M(R_s) \leqslant -\rho' M R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-\rho' I_{\alpha M}(\vec{p}')] \quad . \tag{4-122}$$

If $\vec{p}_s$ is the single-subchannel marginal distribution corresponding to $\vec{p}'$, Eqs. (4-122) and (4-93) imply

$$E'_M(R_s) \leqslant -\rho' M R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-\rho' M I_\alpha(\vec{p}_s)] \quad .$$

Then, for β satisfying Eq. (4-119) for this particular $\vec{p}_s$,

$$E'_M(R_s) \leqslant -\rho' M\,[R_s - I_\beta(\vec{p}_s)] - \ln p(\beta) \leqslant -\ln p(\beta) \quad . \tag{4-123}$$

Thus, using Eq. (4-121), we have

$$E'_M(R_s) \leqslant d$$

independently of $\vec{p}_s$, and hence independently of M. The remainder of Eq. (4-120) is provided by Theorem 3.5.

Theorem 4.13 extends the conditions under which the conclusion of Theorem 4.3 holds [with substitution of d for $-\ln p(a)$]. One would expect that a similar extension is possible for Theorems 4.4, 4.5, 4.8, 4.9, and 4.10. This is indeed the case. Of course, some modification of the proofs of these theorems is required.

We shall close this chapter with a result on monotonicity.

---

† Recall that $p(a) > 0$ for all $a \in \Lambda$.

Theorem 4.14.

$\tilde{E}_M(R_s)$ is a monotone nondecreasing function of M.

Proof.

First, note that Eq. (4-56) implies that $E_{o\alpha}(\rho, \vec{p}_s)$ is a convex upward function of ρ for each $\vec{p}_s \in U$ and $\alpha \in \Lambda$. By definition of convexity,

$$\frac{M}{M+1} E_{o\alpha}(\rho, \vec{p}_s) + \frac{1}{M+1} E_{o\alpha}(o, \vec{p}_s) \leqslant E_{o\alpha}(\frac{M}{M+1}\rho, \vec{p}_s) \quad . \tag{4-124}$$

By Eq. (4-52), this becomes

$$ME_{o\alpha}(\rho, \vec{p}_s) \leqslant (M+1) E_{o\alpha}(\frac{\rho M}{M+1}, \vec{p}_s) \tag{4-125}$$

for all ρ, $0 \leqslant \rho \leqslant 1$, $\alpha \in \Lambda$, and $\vec{p}_s \in U$. From Eqs. (4-20), (4-21), and (4-47), we have

$$E'_M [\rho, (\vec{p}_s)^M, R_s] = -\rho M R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-ME_{o\alpha}(\rho, \vec{p}_s)] \quad . \tag{4-126}$$

Define $\rho', \vec{p}'_s$ by

$$\tilde{E}_M(R_s) = E'_M [\rho', (\vec{p}'_s)^M, R_s] \quad . \tag{4-127}$$

Then,

$$\tilde{E}_M(R_s) = -\rho' M R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-ME_{o\alpha}(\rho', \vec{p}'_s)] \tag{4-128}$$

and

$$E'_{M+1} [\frac{\rho' M}{M+1}, (\vec{p}'_s)^{M+1}, R_s] = -\rho' M R_s - \ln \sum_{\alpha \in \Lambda} p(\alpha) \exp[-(M+1) E_{o\alpha}(\frac{\rho' M}{M+1}, \vec{p}'_s)] \quad . \tag{4-129}$$

Since $0 \leqslant \rho' \leqslant 1$, $0 \leqslant \rho' M/(M+1) \leqslant 1$ also, and by Eqs. (4-125), (4-128), and (4-129),

$$\tilde{E}_M(R_s) \leqslant E'_{M+1} [\frac{\rho' M}{M+1}, (\vec{p}'_s)^M, R_s] \leqslant \max_{\substack{0 \leqslant \rho \leqslant 1 \\ \vec{p} \in D}} E'_{M+1}(\rho, \vec{p}, R_s) = \tilde{E}_{M+1}(R_s) \quad .$$

Note that if $E'_M(R_s) = \tilde{E}_M(R_s)$, the monotonicity above carries over to $E'_M(R_s)$. It is not known whether $E'_M(R_s)$ is always monotone for MSCC channels. I would conjecture that the answer is in the negative. However, since

$$\frac{\partial^2 E_{o\alpha k}(\rho, \vec{p})}{\partial \rho^2} \leqslant 0$$

for $0 \leqslant \rho \leqslant 1$ and $\vec{p}$ defined on X_s^k, we may derive in a manner analogous to the derivation of Eq. (4-125)

$$E_{o\alpha k}(\rho, \vec{p}) \leqslant \ell E_{o\alpha k}(\rho/\ell, \vec{p})$$

for ℓ an integer. Thus, again proceeding as in Theorem 4.14, we get

$$E'_M(R_s) \geqslant E'_\ell(R_s)$$

and

$$E'_M(R_s) \geqslant E'_k(R_s)$$

whenever $M = k\ell$.

For our examples, $\tilde{E}_M(R_s) = E'_M(R_s)$. Thus, the monotone behavior of their RCE's may be observed in Figs. 11 and 18.

REFERENCES

1. R. G. Gallager, "A Simple Derivation of the Coding Theorem and Some Applications," IEEE Trans. Inform. Theory IT-11, No. 1, 3 – 18 (1965). Theorem 4 of this paper provides the means of proof.
2. R. M. Fano, Transmission of Information (M. I. T. Press/Wiley, New York, 1961), Eq. (6.26).
3. A. Feinstein, Foundations of Information Theory (McGraw-Hill, New York, 1958), p. 27.

CHAPTER 5
SYSTEMATIC CODING FOR COMPLETELY CONSTRAINED CHANNELS

A. INTRODUCTION

Our study of coding for parallel channels has thus far been confined to an exploration of the properties of the applicable random coding exponent (RCE). This exponent presupposes maximum-likelihood decoding, as has been previously stated. For any given code, with code words used equiprobably, maximum-likelihood decoding yields the minimum probability of error. Unfortunately, the amount of computational effort required to perform maximum-likelihood decoding is a positive exponential function of block length. Thus, for long codes, this effort becomes prohibitive.

Fortunately, for a class of block codes known as BCH codes [which class includes the Reed-Solomon (RS) codes], the computational effort involved in decoding can be reduced to a practical level through the use of minimum distance decoding techniques. The use of these techniques will, however, involve some sacrifice in performance relative to maximum-likelihood decoding.

In this chapter, we shall examine a class of procedures for BCH coding on a channel with parallel structure. For the case of an MSCC channel, we shall develop a set of formulas which, in combination, will enable us to calculate or bound the probability of error associated with each procedure. Although general results concerning performance will not be given, some examples are computed out at the end of the chapter.

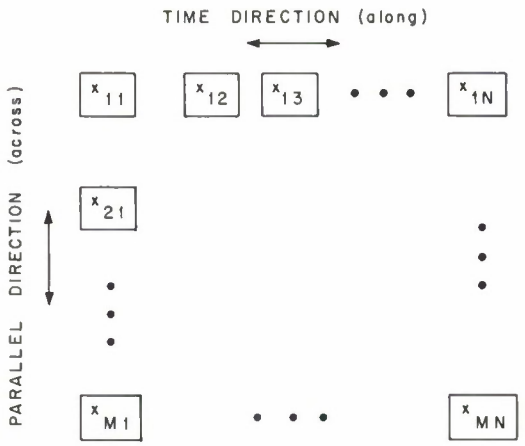
B. CODING ALTERNATIVES

The presence of a number of parallel subchannels presents us with a number of coding alternatives. One of the decisions which must be made in choosing among them is to decide on the number $m^{\dagger}$ of subchannels to be coded on at once. Such a decision implies that the M subchannels will be divided into M/m sets of m subchannels each. For each such set, a code letter will be defined as the m-tuple consisting of the m subchannel inputs in the set at some one instant of time, i.e., a code letter is a member of X_s^m. If we choose the code alphabet to be X_s^m, we shall classify our coding technique as <u>simple</u>. However, we may wish to increase the reliability of individual code letters by choosing the code alphabet to be a proper subset of X_s^m, in which case we classify our coding technique as <u>compound</u>.

The code letters corresponding to each set of m subchannels are then encoded to form code words of length N (this is done, separately, for each set). Each set of m subchannel outputs is then separately decoded (although there may be state knowledge used in common by all sets). An error is considered to have occurred if a decoding error is made in any of the sets.

The distinction, defined above, between simple and compound coding may be made more graphic by referring to the following diagram which shows a code word of length N on M subchannels:

† We assume m divides M.



In compound coding, we code in the parallel direction (with dimensionless rate less than unity) before coding in the time direction. In simple coding, we code in the time direction only.

Simple Coding Example

Let $M = 10$, $m = 2$, $N = 3$, and $X_s = \{0, 1\}$. Then,

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{are the four possible code letters.}$$

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \text{is one of the 64 possible code words.}$$

Compound Coding Example

Let $M = 10$, $m = 2$, $N = 3$, and $X_s = \{0, 1\}$. Let

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{be the only two permitted code letters.}$$

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{is one of the eight possible code words.}$$

C. DIMENSIONLESS RATE

Obviously, we shall be interested in comparing the performance of various coding techniques on particular MSCC channels. To make the comparisons meaningful, we must define our input and output parameters with some care. We have already done this for the probability of error, i.e., an "error" means the same thing regardless of the value of m. Suppose X_s is a set consisting of L members, and that on each set of m subchannels we define W_m code words of length N, with

$$1 \leqslant W_m \leqslant L^{mN} \quad . \tag{5-1}$$

Then, for some real number r satisfying

$$0 \leqslant r \leqslant 1 \tag{5-2}$$

we have

$$W_m = L^{rmN} \quad . \tag{5-3}$$

We shall call r the dimensionless rate. If we consider each M/m-tuple of the above code words to be a code word on the whole channel, then

$$W_M = (W_m)^{M/m} = L^{rMN} \quad . \tag{5-4}$$

Since the RHS of Eq. (5-4) is independent of m, we may define

$$r = \frac{\log_L W_m}{mN} \tag{5-5}$$

as the appropriate input parameter. The <u>dimensionless rate</u> is related to the rate per subchannel R_s by

$$R_s = r \ln L \tag{5-6}$$

where R_s is in natural units.

D. BCH CODES AND SIMPLE CODING SCHEMES

The properties of BCH codes and various decoding schemes for them are developed and described in a fairly extensive literature.[1-3†] In a BCH code, the code letters are equal to (isomorphic with) the elements of a finite (Galois) field with q elements GF(q). Such fields exist whenever $q = p^n$, where p is a prime and n is a positive integer. Hence, if a BCH code is to be used for simple coding over m subchannels of a channel with parallel structure and subchannel input alphabet $\{1, \ldots, L\}$, we require $L^m = p^n$ for some prime p and positive integer n. The requirement can be met if and only if for some prime p and positive integer k,

$$L = p^k \quad . \tag{5-7}$$

Then, n = km. Thus, we must restrict our discussion to situations where the subchannel input alphabet size is an integer power of a prime.

If there is a Galois field GF(q) with q elements, then for any positive integer ℓ there exists an extension field $GF(q^\ell)$ with q^ℓ elements. Let

$$N = q^\ell - 1 \quad .$$

A sequence $(u_{N-1}, \ldots, u_o)$ of code letters[‡] may be represented by a polynomial u(t) of degree at most N − 1

$$u(t) = u_{N-1} t^{N-1} + \ldots + u_o \qquad u_i \in GF(q) \quad .$$

Pick $\gamma \in GF(q^\ell)$ so that γ is primitive, and pick d a positive integer less than N. Let the code words of a code of block length N be given by the set of polynomials of degree N − 1 or less with coefficients in GF(q) which have $\gamma, \gamma^2, \ldots, \gamma^{d-1}$ as roots. A code generated in this way is defined as a BCH code. If $\ell = 1$, the code is an RS code. If r is the dimensionless rate of the code

† Numbered references appear at the end of each chapter.

‡ Note that the usual subscript order for the letters is reversed here.

$$(d-1) \geqslant \frac{(1-r)\,N}{\ell} \quad . \tag{5-8}$$

For RS codes,

$$(d-1) = (1-r)\,N \quad . \tag{5-9}$$

The actual value of the ratio of $(1-r)\,N$ to $(d-1)$ for BCH codes with $\ell \neq 1$ may be obtained by making use of the fact that $(1-r)\,N$ is equal to the degree of the polynomial which is the least-common multiple of the minimal polynomials for the $(d-1)$ field elements $\gamma, \ldots, \gamma^{d-1}$. This can be a tedious calculation. Note, however, that for $q = 2$, we have

$$(d-1) \geqslant \frac{2(1-r)\,N}{\ell} \tag{5-10}$$

and the parameters of a number of such binary BCH codes are tabulated by Peterson.†

We note that the parameter d of a BCH code is not necessarily the same as its minimum distance, although it serves as a lower bound to the minimum distance.

E. STATE INFORMATION AND RELIABILITY

We shall restrict our consideration to those channels for which, for each $\alpha \in \Lambda$,

$$\sum_{q:q\neq\ell} p_\alpha(q/\ell) = f(\alpha) \tag{5-11}$$

independent of ℓ. This will have the effect of making the probability of correct decoding by a minimum distance algorithm independent of the code word sent, and thus greatly simplify the calculation or bounding of code performance. The symmetry requirement, Eq. (5-11), is usually met in practice.

The probability of correct decoding will, in general, be affected by the choice of m in the simple coding and decoding schemes described above. It will also be affected by what state information is available at the receiver and how it is used. As was pointed out in Chapter 2, where the physical channels we are modeling are fading channels, it is usually possible to obtain partial state information at the receiver by making an energy measurement. (We may also obtain this information by using some of the channels as test channels.) This information will often enable us to assign a number representing reliability to each received letter. Suppose m subchannels are coded at once, and the reliability b_m of an m-tuple received at a particular instant of time is defined as the probability of its being correctly received conditioned on whatever state information the receiver possesses. If the receiver has complete state knowledge, we have from Eq. (5-11)

$$b_m(\alpha) = [1 - f(\alpha)]^m \quad . \tag{5-12}$$

Suppose the receiver has partial channel state information represented by knowledge of a random variable β, for which

$$p(q/\ell\alpha\beta) = p(q/\ell\alpha) = p_\alpha(q/\ell) \tag{5-13}$$

and

$$p(\ell\alpha\beta) = p(\ell)\,p(\alpha\beta) \quad . \tag{5-14}$$

† See p. 166 of Ref. 1.

One can readily show that Eqs. (5-11), (5-13), and (5-14) imply

$$\sum_{q:q\neq\ell} p(q/\ell\beta) = g(\beta) \tag{5-15}$$

independently of ℓ, with

$$g(\beta) = \sum_{\alpha\in\Lambda} f(\alpha)\, p(\alpha/\beta) \quad . \tag{5-16}$$

We then have

$$b_m(\beta) = [1 - g(\beta)]^m \quad . \tag{5-17}$$

F. MINIMUM DISTANCE DECODING

Let $X_s = Y_s$, and $A = X_s^{mN} = Y_s^{mN}$ be the set of possible received sequences. Let g be a function defined on $A \times A$ with

$$g(\vec{x},\vec{y}) = g(\vec{y},\vec{x}) \geqslant 0 \qquad \text{all } \vec{x},\vec{y} \in A$$

$$g(\vec{x},\vec{x}) = 0 \qquad \text{all } \vec{x} \in A$$

and

$$g(\vec{x},\vec{z}) \leqslant g(\vec{x},\vec{y}) + g(\vec{y},\vec{z}) \qquad \text{all } \vec{x},\vec{y},\vec{z} \in A \quad .$$

Then, g is a distance function. A minimum distance decoding scheme is one which decodes a received word $\vec{y}$ into the <u>code word</u> $\vec{x}$ for which $g(\vec{x},\vec{y})$ is minimum.

The simplest choice for $g(\vec{x},\vec{y})$ is the Hamming distance, which is simply the number of code letters in which $\vec{x}$ and $\vec{y}$ differ. The Hamming distance treats each code letter equally and makes no use of reliability information. Decoding with a Hamming distance is also referred to as errors-only decoding. Efficient algorithms exist for errors-only decoding of BCH codes. These algorithms succeed whenever twice the number of code letters received in error is less than d, where d is the code parameter.

If the receiver has partial state information, it is no longer logical to use a distance function which treats all received letters equally. One may, for example, establish a reliability threshold r_t, $0 < r_t < 1$, and erase the ith received letter if its reliability $b_m^i(\beta)$ satisfies

$$b_m^i(\beta) \leqslant r_t \quad . \tag{5-18}$$

One may then define $g(\vec{x},\vec{y})$ as the number of non-erased positions in which $\vec{x}$ and $\vec{y}$ differ. This distance is called the Elias distance. Decoding with this distance is called erasures and errors decoding. Efficient algorithms exist which succeed whenever the number of errors e and number of erasures k satisfy

$$2e + k < d \quad . \tag{5-19}$$

Finally, define $v(x_i, y_i)$, $x_i \in X_s^m$, and $y_i \in X_s^m$ by

$$v(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \quad . \tag{5-20}$$

Then, if b_m^i is the reliability of the ith received code letter, and $h(t)$, $0 \leq h(t) \leq 1$, is a monotone nondecreasing function of t, $0 \leq t \leq 1$, we may define $g(\vec{x}, \vec{y})$ by

$$g(\vec{x}, \vec{y}) = \sum_{i=1}^{N} h(b_m^i)\, v(x_i, y_i) \quad . \tag{5-21}$$

Decoding with this distance function is called generalized minimum distance decoding.† Efficient algorithms exist which are successful when the transmitted word $\vec{x}$ and received word $\vec{y}$ obey

$$2g(\vec{x}, \vec{y}) < d - N + \sum_{i=1}^{N} h(b_m^i) \quad .$$

Since explicit analytical error bounds do not exist for generalized minimum distance decoding, we shall confine our analyses to decoding with erasures, errors, or both. It should be pointed out, however, that when $b_m(\beta)$ may take on many widely separated values (e.g., 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0), generalized minimum distance decoding should promise sufficient advantage over erasures and errors decoding to justify the numerical calculation of a bound in a practical situation.

Note that the Hamming distance is obtained from Eq. (5-21) by setting

$$h(t) = 1 \qquad \text{all } t,\ 0 \leq t \leq 1 \quad .$$

The Elias distance is obtained by setting

$$h(t) = \begin{cases} 0 & \text{if } t \leq r_t \\ 1 & \text{if } t > r_t \end{cases} \quad .$$

G. SINGLE-LETTER ERASURE AND ERROR PROBABILITIES

We are now in a position to compute the single-letter error, erasure, and correct reception probabilities (p_e, p_s, and p_c, respectively) for an MSCC channel when the receiver has knowledge of β, and m subchannels are coded at once. Let Γ be the set of all possible β. For convenience, we shall assume that β is a discrete variable. Define $\Gamma_s \subseteq \Gamma$ as the set of β, for which

$$b_m(\beta) \leq r_t \quad . \tag{5-22}$$

Then, let

$$p_s = \sum_{\beta \in \Gamma_s} p(\beta) \tag{5-23}$$

$$p_e = \sum_{\beta \in \overline{\Gamma}_s} p(\beta)\, \{1 - [1 - g(\beta)]^m\} \tag{5-24}$$

† See Ref. 2, pp. 12-24.

$$p_c = 1 - p_s - p_e \tag{5-25}$$

where $\overline{\Gamma}_s$ is the complement of Γ_s.

For the special case of complete channel state knowledge, we may define Λ_s as the set of α for which

$$b_m(\alpha) \leqslant r_t \tag{5-26}$$

and $\overline{\Lambda}_s$ as the complement of Λ_s. Then, Eqs. (5-23), (5-24), and (5-25) hold if α replaces β, $f(\alpha)$ replaces $g(\beta)$, and Λ replaces Γ.

For the special case of no channel state knowledge,

$$p_s = 0 \tag{5-27}$$

$$\begin{aligned} p_e &= \sum_{\alpha \in \Lambda} p(\alpha) \{1 - [1 - f(\alpha)]^m\} \\ &= 1 - p_c \quad . \end{aligned} \tag{5-28}$$

H. PROBABILITY OF CORRECT DECODING

Using the probabilities derived in Sec. G above and Eq. (5-19), it is possible to calculate the probability $P_c(m)$ of correct decoding of a single set of m subchannels:

$$P_c(m) = \sum_{\substack{\ell,k: \\ 2\ell+k<d}} \frac{N!}{\ell!\ k!\ (N-\ell-k)!} p_e^{\ell} p_s^{k} (1 - p_e - p_s)^{N-\ell-k} \quad . \tag{5-29}$$

Equation (5-29) assumes erasures and errors decoding, but reduces to errors only or erasures only if $p_s = 0$ or $p_e = 0$, respectively.

I. CHERNOFF BOUND

It is often difficult to evaluate Eq. (5-29). To evaluate conveniently the performance of BCH codes, we shall need to use the Chernoff bound.[4] Let u_i, $1 \leqslant i \leqslant N$ be a set of independent identically distributed random variables with mean $\bar{u}$; let $\epsilon > 0$ and $\lambda = \bar{u} + \epsilon$; let $a_{-1}(t)$ be defined by

$$a_{-1}(t) = \begin{cases} 1 & \text{for } t \geqslant 0 \\ 0 & \text{for } t < 0 \end{cases} \quad . \tag{5-30}$$

Then,

$$\begin{aligned} P\left(\frac{1}{N}\sum_{i=1}^{N} u_i \geqslant \lambda\right) &= E\left[a_{-1}\left(\frac{1}{N}\sum_{i=1}^{N} u_i - \lambda\right)\right] \\ &= E\left[a_{-1}\left(\sum_{i=1}^{N} u_i - N\lambda\right)\right] \quad . \end{aligned} \tag{5-31}$$

For any $s \geqslant 0$,

$$a_{-1}(t) \leqslant \exp[st] \quad . \tag{5-32}$$

Thus,

$$E\left[a_{-1}\left(\sum_{i=1}^{N} u_i - N\lambda\right)\right] \leqslant E\left\{\exp\left[s\left(\sum_{i=1}^{N} u_i - N\lambda\right)\right]\right\} \tag{5-33}$$

and

$$E\left\{\exp\left[s\left(\sum_{i=1}^{N} u_i - N\lambda\right)\right]\right\} = E\left\{\prod_{i=1}^{N} \exp[s(u_i - \lambda)]\right\} \tag{5-34}$$

$$= \left(E\{\exp[s(u-\lambda)]\}\right)^N \tag{5-35}$$

where we use the fact that the u_i are independent and identically distributed. Thus, if

$$D(s,\lambda) = -\ln\{E(\exp[s-\lambda])\} \tag{5-36}$$

we have for $s \geqslant 0$,

$$P\left(\frac{1}{N}\sum_{i=1}^{N} u_i \geqslant \lambda\right) \leqslant \exp[-ND(s,\lambda)] \quad . \tag{5-37}$$

The bound is tightest for s satisfying

$$\frac{d}{ds} E\{\exp[s(u-\lambda)]\} = E\{(u-\lambda)\exp[s(u-\lambda)]\} = 0$$

or

$$\lambda = \frac{E(u \exp[su])}{E(\exp[su])} \quad . \tag{5-38}$$

Since $\lambda > \bar{u}$, a unique positive solution s_o to Eq. (5-38) is guaranteed to exist† if the variance of u is positive.

J. CHERNOFF BOUNDS FOR ERASURES AND/OR ERRORS DECODING

Theorem 5.1. (Chernoff Bound for Errors and Erasures)

Suppose we use a BCH code of block length N and parameter d over a channel with erasure probability p_s, error probability p_e, and probability of correct reception p_c. Let

$$p_c + p_s + p_e = 1 \tag{5-39}$$

$$p_e > 0 \tag{5-40}$$

and

$$2p_e + p_s < \frac{d}{N} = t < 1 \quad . \tag{5-41}$$

---

† This is proved in essentially the same way as the similar result in part (b) of Theorem 4.8 in Chapter 4. $E(\exp[su])$ is the moment generating function associated with the random variable u.

Then, the best Chernoff bound to the probability that decoding will fail, P_e, is given as follows:

$$P_e \leqslant \exp[-ND(d)] \tag{5-42}$$

where

$$D(d) = s_o t - \ln\left(p_c + p_s e^{s_o} + p_e e^{2s_o}\right) \tag{5-43}$$

and

$$\dot{s}_o = \ln\left\{\frac{-p_s(t-1)}{2p_e(t-2)} + \sqrt{\left[\frac{p_s(t-1)}{2p_e(t-2)}\right]^2 + \frac{p_c t}{p_e(2-t)}}\right\} \quad . \tag{5-44}$$

Proof.

By Eq. (5-19), the probability of error in decoding is at most the probability that $2e + k \geqslant d$. Now, define random variables u_i, $1 \leqslant i \leqslant N$ as follows:

$u_i = 0$ with probability p_c

$u_i = 1$ with probability p_s

$u_i = 2$ with probability p_e.

These N random variables will be assumed to be independently distributed. Clearly,

$$P(2e + k \geqslant d) = P\left(\frac{1}{N}\sum_{i=1}^{N} u_i \geqslant t\right) \quad . \tag{5-45}$$

The RHS of Eq. (5-45) may be Chernoff bounded as per Eqs. (5-36), (5-37), and (5-38), where $\lambda \to t$. Now,

$$E\left(e^{s_o u}\right) = p_c + p_s e^{s_o} + p_e e^{2s_o} \tag{5-46}$$

$$E\left(ue^{s_o u}\right) = p_s e^{s_o} + 2p_e e^{2s_o} \quad . \tag{5-47}$$

Thus, by Eq. (5-38),

$$t = \frac{p_s e^{s_o} + 2p_e e^{2s_o}}{p_c + p_s e^{s_o} + 2p_e e^{2s_o}} \quad . \tag{5-48}$$

Equation (5-48) is a quadratic in e^{s_o}. When it is solved and the natural logarithm is taken, the RHS of Eq. (5-44) results. From Eq. (5-36),

$$D = -\ln E\left[e^{s_o(u-t)}\right]$$

$$= +s_o t - \ln E\left(e^{s_o u}\right)$$

$$= s_o t - \ln\left(p_c + p_s e^{s_o} + p_e e^{2s_o}\right) \quad .$$

Theorem 5.2. (Chernoff Bound for Erasures Only)

Suppose we use a BCH code of block length N and parameter d over a channel with erasure probability p_s and probability of correct reception $(1 - p_s)$. Suppose

$$0 < p_s < \frac{d}{N} = t < 1 \quad . \tag{5-49}$$

Then, the probability that decoding will fail, P_e, is bounded as follows:

$$P_e \leq \exp\{-N[-t \ln p_s - (1-t)\ln(1-p_s) - \mathcal{H}(t)]\} \tag{5-50}$$

where

$$\mathcal{H}(t) = -t \ln t - (1-t)\ln(1-t) \quad . \tag{5-51}$$

Proof.

The proof proceeds as that of Theorem 5.1 up to Eq. (5-48), which is now linear in e^{s_o}. Substituting the value of s_o obtained in Eq. (5-36) gives our result.

Theorem 5.3. (Chernoff Bound for Errors Only)

Suppose we use a BCH code of block length N and parameter d over a channel with error probability p_e and probability of correct reception $(1 - p_e)$. Suppose

$$0 < 2p_e < \frac{d}{N} = t < 1 \quad . \tag{5-52}$$

Then, the probability that decoding will fail, P_e, is bounded as follows:

$$P_e \leq \exp\{-N[-\tfrac{t}{2}\ln p_e - (1-\tfrac{t}{2})\ln(1-p_e) - \mathcal{H}(t/2)]\} \tag{5-53}$$

where $\mathcal{H}(t)$ is given by Eq. (5-51).

Proof.

This is really a corollary of Theorem 5.1. We set $p_s = 0$, $p_c = 1 - p_e$ in Eq. (5-44), and substitute the result in Eqs. (5-43) and (5-42) to obtain our result.

K. BOUNDS ON TOTAL PROBABILITY OF DECODING FAILURE

We have one further topic we must explore before proceeding with the analysis of simple coding schemes on some specific MSCC channels. Earlier, we agreed to count the decoding as being in error if a decoding failure occurs in any one of the $M/m = S$ sets of m subchannels each, where m subchannels are encoded at once. If the probability of error $P_e(m)$ is computed or bounded above for each such set, separately, then the total probability of error P_e may be estimated by the use of the union bound. Thus, if for each set of m subchannels,

$$P_e(m) \leqslant c^{-ND} \tag{5-54}$$

we have

$$P_e \leqslant SP_e(m) \leqslant Se^{-ND} \quad . \tag{5-55}$$

Since, at each instant of time, all the subchannels of an MSCC channel are required to be in the same state, we would normally expect that errors in the various subchannel sets would tend to occur together, and thus that a better estimate than the union bound exists. Suppose $\vec{\alpha}$ is a sequence of N subchannel states, and Λ^N is the set of such sequences. Then, we have

$$P_e = 1 - \sum_{\vec{\alpha} \in \Lambda^N} p(\alpha)\,[1 - P_e(m/\vec{\alpha})]^S \tag{5-56}$$

where $P_e(m/\vec{\alpha})$ is the probability of decoding failure on a single set of m subchannels when the sequence of subchannel states was $\vec{\alpha}$. Equation (5-56) is obtained from the elementary rules of probability after noting that when the state vector $\vec{\alpha}$ is given, the subchannels become independent of each other. The principal limitation on the use of Eq. (5-56) is the fact that even though $P_e(m/\vec{\alpha})$ depends only on the number of each subchannel state in the sequence $\vec{\alpha}$, it is not generally easy to compute or bound. Equation (5-56) is most easily used when $P_e(m/\vec{\alpha})$ is equal either to zero or to unity for all $\vec{\alpha}$. Then,

$$P_e = P_e(m) \qquad \text{all } m \quad . \tag{5-57}$$

This situation occurs in erasures-only decoding. Equation (5-57) reflects the fact that decoding will fail if there are d or more erasures, and that the same number of erasures occur for each set of m subchannels because our channel is MSCC.

L. ERROR EXPONENTS FOR SOME EXAMPLES OF MSCC CHANNELS

We shall now compute error bounds for simple BCH coding on some particular MSCC channels. The subchannel input and output alphabets will be binary, and the number of subchannels M will be seven. Since seven is prime, we shall have only two simple coding alternatives to consider: first, to code and decode on all seven subchannels simultaneously, using an RS code for which the alphabet size is $2^7 = 128$, and block length is $2^7 - 1 = 127$; second, to code and decode each subchannel separately, using a binary BCH code of the same block length. The relationship among d, r, and N for the RS code is given by Eq. (5-9); for the binary BCH code, the relationship is obtained from Table 9.1 of Ref. 1. The results we need are summarized in Table I.

We define a new exponent, B(r), by

$$B(r) = D[d(r)] - \frac{\ln S}{127} \tag{5-58}$$

and an upper bound to the total probability of decoding failure $\overline{P}_e$ by

$$\overline{P}_e = \exp[-127B(r)] \quad . \tag{5-59}$$

Hence, we have

$$P_c \leqslant \exp[-127B(r)] = \overline{P}_e \quad . \tag{5-60}$$

TABLE I RATE AND MINIMUM DISTANCE FOR SOME BINARY BCH CODES (N = 127)			
r × N	d	r × N	d
120	3	57	23
113	5	50	27
106	7	43	29
99	9	36	31
92	11	29	43
85	13	22	47
78	15	15	55
71	19	8	63
64	21		

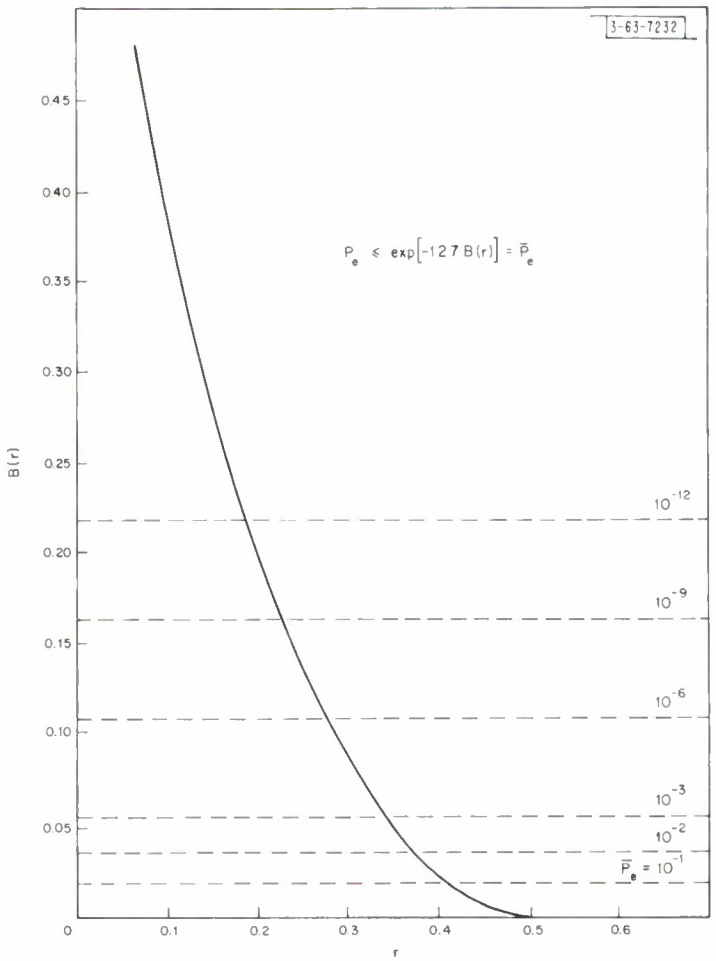
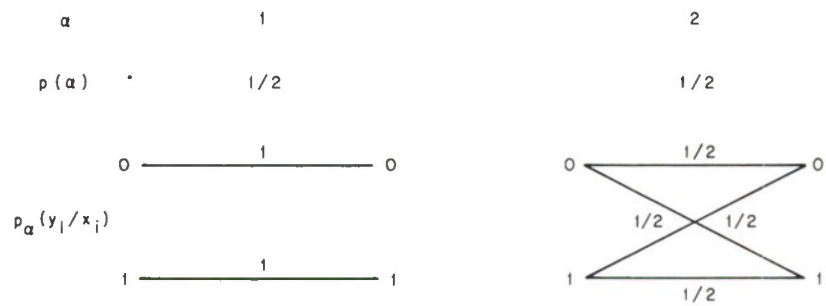


Fig. 24. Exponent vs dimensionless rate for Reed-Solomon code – Example 1.

Example 1

M = 7 State Known at Receiver

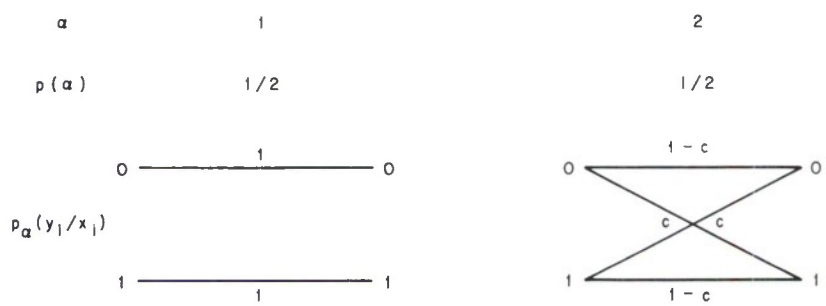


We note that the erasures-only bounds are applicable here, and $p_s = 1/2$. For the RS code, $r \leqslant 1/2$ implies that Eq. (5-49) holds and we may compute a positive error exponent using Eqs. (5-50) and (5-51). The result is plotted in Fig. 24.

Now, we consider a binary BCH code on one subchannel of this channel, and note that the expected number of erasures is 63.5. Even the lowest rate binary BCH code given in Table I corrects at most 62 erasures. Hence, the probability of decoding failure exceeds 0.5 for all positive rates. Thus, coding over all subchannels at once is clearly a superior procedure here.

Example 2

M = 7 State Unknown at Receiver



c = 0.05, 0.10, 0.15, 0.20

For the RS codes,

$$p_e = \frac{1}{2}\left[1 - (1-c)^7\right] \quad . \tag{5-61}$$

For each value of c, Eqs. (5-52), (5-61), and (5-9) tell us at which rates a positive exponent may be expected. If Eq. (5-52) is satisfied, the exponent is the expression in square brackets in Eq. (5-53). These exponents are plotted in Fig. 25 for c = 0.05, 0.10, 0.15, and 0.20 (curves labeled S).

For the BCH codes,

$$p_e = c/2 \quad . \tag{5-62}$$

For each value of c, Eqs. (5-52), (5-62), and Table I tell us at which rates a positive exponent

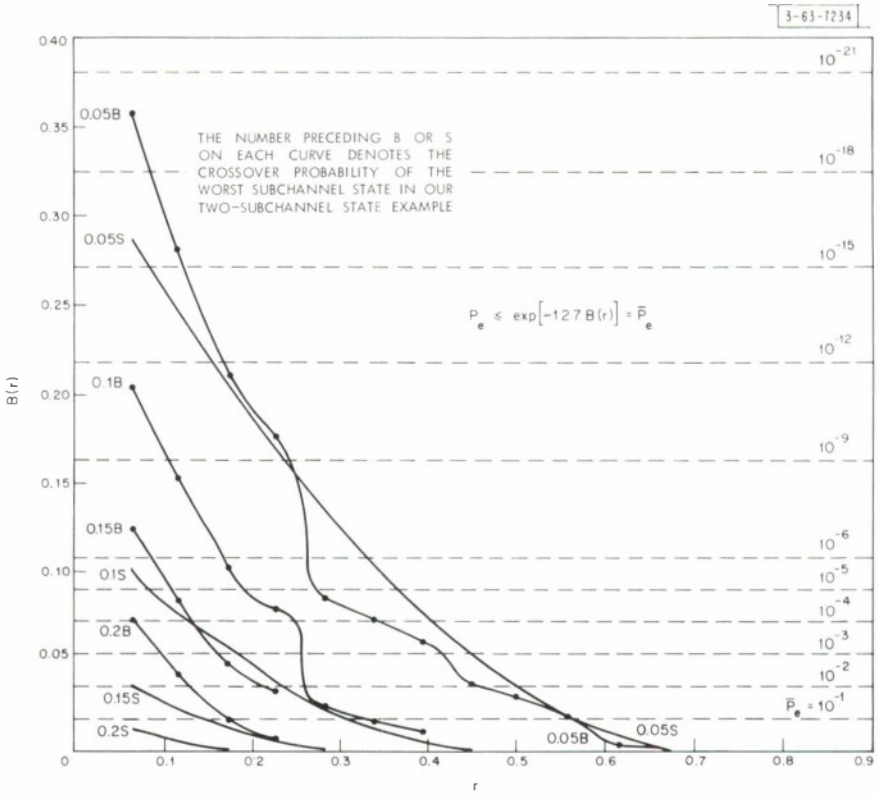


Fig. 25. Exponent vs dimensionless rate for Reed-Solomon (S) and binary BCH (B) codes on several MSCC channels – Example 2.

may be expected in the bound for $P_e(1)$. If Eq. (5-52) is satisfied, the exponent D(r) is the expression in square brackets in Eq. (5-53). We plot the exponent B(r) given by Eq. (5-58) whenever it is positive; plots are shown in Fig. 25 for c = 0.05, 0.10, 0.15, and 0.20 (curves labeled B).

We see from the curves that, for c = 0.10, 0.15, and 0.20, the binary BCH exponent is greater than the RS exponent at all rates. Hence, a binary BCH code would be our choice for these examples. In addition, the binary BCH code is easier to implement. If c = 0.05, it appears as though the favored code depends on the rate. (Note that the curves for the binary BCH codes serve only to connect the data points and have no meaning between them. Thus, the RS and binary BCH codes may only be compared at the data points for the latter.)

Now, we can examine the reasons why we have obtained the above results. When we increase the number of subchannels coded over at once, two things happen. First, the alphabet size increases, with a resultant increase in the code parameter d for a fixed dimensionless rate and block length N. The largest value of d for a fixed dimensionless rate and block length is that given by Eq. (5-9) and is achieved for an alphabet size one greater than the block length. Second, the probability of a code letter being received in error (or erased if the receiver has state knowledge) increases [see Eqs. (5-17), (5-22), (5-23), and (5-24)], with a resultant increase in the expected number of errors and erasures. Example 1 is a special case in that the erasure probability remains the same regardless of how many subchannels are coded over. Hence, the increase in alphabet size is entirely beneficial, and the RS code is superior. In Example 2, the increase in error probability with the number of subchannels coded over is the dominant effect for c = 0.10,

0.15, and 0.20. Unlike the situation in Example 1, we can only determine this after making a calculation.

In general, we would expect the optimal number of subchannels coded over at once to lie somewhere between 1 and M. The determination of this optimum number (which may depend upon the rate), in more general situations than we have considered, involves considerable labor. This labor is primarily due to the difficulty of finding the precise relationship among code parameter, dimensionless rate, and block length for non-binary, non-RS BCH codes. Inequality (5-8) is of some help here. If assuming (5-8) were satisfied with an equal sign, we compute our exponent and find it to be greater (for the same block length) than that for a RS code, then we are secure in concluding that the RS code is not best.

M. COMPOUND CODING

The problem mentioned above, of an increase in code letter probability of error with increasing number of subchannels, has an obvious solution – to code across the subchannels before coding along them. This is what we have called compound coding.† The number of channel code letters A (i.e., the number of possible input M-tuples, at a given instant of time) is given by

$$A = L^{r_1 M}$$

for some $0 \leq r_1 < 1$. The number of channel code words W is given by

$$W = A^{r_2 N} = L^{r_1 r_2 M}$$

for some $0 \leq r_2 \leq 1$. Thus, $r_1 r_2$ for compound coding is comparable to r for simple coding.

Compound coding does not seem to be an attractive technique for MSCC channels. In the first place, we must make $r_1 \ln L$ smaller than the capacity (in natural units) of the worst subchannel state whose reliability we wish to improve. Furthermore, for a number of subchannels of order 100, the improvement in reliability is generally not too marked unless $r_1 \ln L$ is one-half or less the capacity of the worst subchannel state. Thus, compound coding is generally applicable only to low rates. That the exponents obtained even at these low rates are not generally as large as those obtained for simple coding is not so obvious. Indeed, we cannot be sure that compound coding is not advantageous in some instances, although intuition suggests that coding in a direction in which we have no "diversity" will not be advantageous.

REFERENCES

1. W.W. Peterson, Error Correcting Codes (M.I.T. Press/Wiley, New York, 1961).
2. G.D. Forney, "Concatenated Codes," Technical Report 440, Research Laboratory of Electronics, M.I.T. (1 December 1965).
3. R.G. Gallager, Information Theory and Reliable Communication (Wiley, New York, 1968), Chapter 6.
4. H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on a Sum of Observations," Ann. Math. Stat. 23, 493 (1952).

† This resembles the approach taken in concatenated coding (see Ref. 2).

CHAPTER 6
MORE GENERAL CHANNEL MODELS

In Chapters 4 and 5, we restricted our discussion to channels which are MSCC. Roughly speaking, such channels have three main properties: state structure, memorylessness, and identical subchannel states at each instant of time. We shall here consider channels in which the second and third properties may not obtain.

A. MARKOV PARALLEL CHANNEL

Let p_{ij} be such that

$$0 \leqslant p_{ij} \leqslant 1 \qquad 1 \leqslant i,j \leqslant S$$

and

$$\sum_{j=1}^{S} p_{ij} = 1 \qquad 1 \leqslant i \leqslant S \quad . \tag{6-1}$$

Let $\{u_j\}_{j=1}^{S}$ be a solution of

$$u_j = \sum_{i=1}^{S} u_i p_{ij} \qquad 1 \leqslant j \leqslant S \tag{6-2}$$

$\{u_j\}$ is called a stationary distribution. Let

$$u_j p_{ji} = u_i p_{ij} \qquad 1 \leqslant i,j \leqslant S \quad . \tag{6-3}$$

Definition

A Markov Parallel Channel (MPC) is an MS channel with $\Lambda = \{1, \ldots, S\}$ in which the probability $p(e_1, \ldots, e_k)$ that <u>any</u> k successive subchannels are in states $e_1, \ldots, e_k$, respectively, is given by

$$p(e_1, \ldots, e_k) = u_{e_1} p_{e_1 e_2} \cdots p_{e_{k-1} e_k} \tag{6-4}$$

irrespective†‡ of the direction of progression across the subchannels.

The channel we have just defined has relatively simple subchannel dependencies while not requiring the states of all the subchannels to be the same during a single-letter transmission. The MSCC channel is the special case of the MPC with $p_{ij} = \delta_{ij}$. By our remarks on time-parallel duality in Chapter 2-B, we should be able to make use of known results on single channels with a Markov state dependence in time§ to analyze the MPC.

---

† Numbered references appear at the end of each chapter.

‡ Equation (6-3) is the consistency condition which allows this (see Ref. 1).

§ The channel with Markov state dependence in time is called a "discrete finite state channel" in Ref. 2.

B. CAPACITY OF MPC

The first problem we might wish to consider is that of the computation of the capacity of the MPC. Unfortunately, there is no simple formula for capacity in terms of the channel parameters given. Gilbert[3] computes the capacity of a single channel with Markov state dependence in time for the special case where both input and output are binary and there are only two channel states. The first state guarantees that the output and input be the same. In the second state, all transition probabilities are one-half. (The states correspond to $p_1(\xi/\eta)$ and $p_2(\xi/\eta)$ of Example 1, Chapter 3.) Since Gilbert's result is a capacity per use of the channel defined by a limiting process, Theorem 2.5 suggests that his results will serve as upper bounds to the capacity per subchannel C_{sM}. In general, Theorems 2.3, 2.4, 2.7, and 3.1 may be combined to obtain upper and lower bounds to C_{sM} which are relatively easily calculable. These are given by

$$\max_{p \in U} \sum_{j=1}^{S} u_j I_p^j(X_s;\ Y_s) + \frac{1}{M} \sum_{j=1}^{S} u_j \log u_j + \left(1 - \frac{1}{M}\right) \sum_{j=1}^{S} \sum_{i=1}^{S} u_i p_{ij} \log p_{ij}$$

$$\leqslant C_{sM} \leqslant \max_{p \in U} \sum_{j=1}^{S} u_j I_p^j(X_s;\ Y_s) \tag{6-5}$$

where, if

$$X_s = \{1, \ldots, L\} \quad , \quad Y_s = \{1, \ldots, Q\}$$

then,

$$I_p^j(X_s;\ Y_s) = \sum_{q=1}^{Q} \sum_{\ell=1}^{L} p(\ell)\ p_j(q/\ell) \log \frac{p_j(q/\ell)}{\sum_{k=1}^{L} p_j(q/k)\ p(k)} \quad . \tag{6-6}$$

[Recall that j is the subchannel state, and $p(\ell)$ is the subchannel input distribution.]

C. RANDOM CODING EXPONENT FOR MPC

Unfortunately, no results are available concerning maximum-likelihood RCE's for Markov channels. Yudkin† derives RCE's for Markov channels with a type of nonmaximum-likelihood decoding. By duality, his results carry over to the MPC without essential change.

D. SYSTEMATIC CODING FOR MPC

In contrast to the situation which exists for the RCE, the performance of BCH codes with simple coding schemes and minimum distance decoding can be evaluated almost as readily for the MPC as for the MSCC. We continue to assume that Eq. (5-11) holds. What is changed is our computation of the reliability b_m (probability of correct reception) of an m-tuple of subchannel inputs. Suppose, for convenience, we choose the m-tuple to consist of the first m subchannel inputs. The state vector we are concerned with is $\vec{\alpha} = (\alpha_1, \ldots, \alpha_m)$. Thus, for the case of complete state knowledge,

† See Ref. 2, Chapter IV.

$$b_m(\vec{\alpha}) = \prod_{i=1}^{m} [1 - f(\alpha_i)] \quad . \tag{6-7}$$

If we have no state knowledge,

$$b_m = \sum_{\alpha_1 \in \Lambda} \cdots \sum_{\alpha_m \in \Lambda} u_{\alpha_1} p_{\alpha_1 \alpha_2} \cdots p_{\alpha_{m-1} \alpha_m} \prod_{i=1}^{m} [1 - f(\alpha_i)] \quad . \tag{6-8}$$

Suppose we have a random variable $\vec{\beta} = (\beta_1, \ldots, \beta_m)$, representing partial knowledge, with

$$p(y/x\vec{\alpha}\vec{\beta}) = p(y/x\vec{\alpha}) = p_{\vec{\alpha}}(y/x) \tag{6-9}$$

$$p(x\vec{\alpha}\vec{\beta}) = p(x)\, p(\vec{\alpha}\vec{\beta}) \tag{6-10}$$

and

$$p_i(y_i/x_i\vec{\beta}) = p_i(y_i/x_i\beta_i) \tag{6-11}$$

for all i, $1 \leqslant i \leqslant m$. Then, if we define $g_i(\beta_i)$ by

$$g_i(\beta_i) = \sum_{y_i : y_i \neq x_i} p_i(y_i/x_i\beta_i) \tag{6-12}$$

$g_i(\beta_i)$ is independent of x_i, and

$$g_i(\beta_i) = \sum_{\alpha_i \in \Lambda} f(\alpha_i)\, p_i(\alpha_i/\beta_i) \tag{6-13}$$

$g_i(\beta_i)$ is the probability that the ith subchannel symbol will be incorrectly received given that the ith component of the partial knowledge vector is β_i. Thus, if the receiver knows $\vec{\beta}$,

$$b_m(\vec{\beta}) = \prod_{i=1}^{m} [1 - g_i(\beta_i)] \quad . \tag{6-14}$$

The erasure criterion is of the form

$$b_m(\vec{\beta}) \leqslant r_t \quad . \tag{6-15}$$

If we define Γ_s^m as the set of $\vec{\beta}$ for which Eq. (6-15) holds, Eqs. (5-23), (5-24), and (5-25) become

$$p_s = \sum_{\vec{\beta} \in \Gamma_s^m} p(\vec{\beta}) \tag{6-16}$$

$$p_e = \sum_{\beta \in \overline{\Gamma_s^m}} p(\vec{\beta}) \left\{ 1 - \prod_{i=1}^{m} [1 - g_i(\beta_i)] \right\} \tag{6-17}$$

$$p_c = 1 - p_s - p_e \tag{6-18}$$

where $\overline{\Gamma_s^m}$ is the complement of Γ_s^m. If the receiver has no channel state knowledge, Eqs. (5-27) and (5-28) become

$$p_s = 0 \tag{6-19}$$

$$p_e = \sum_{\vec{\alpha} \in \Lambda^m} p(\vec{\alpha}) \left\{ 1 - \prod_{i=1}^{m} [1 - f(\alpha_i)] \right\}$$

$$= 1 - p_c \quad . \tag{6-20}$$

E. OTHER MS CHANNELS

Needless to say, the MPC and MSCC channel are not the only possible MS channels. However, together with the independent subchannel case, they are the only MS channels for which random coding results of any generality are known. If, for some reason, it seems desirable to use some other MS channel model, numerical computation may provide the only guide to the behavior of the RCE. The performance of minimum distance decoding of BCH codes may nonetheless be evaluated with an effort comparable to that involved in a similar evaluation for the MPC. The details of this evaluation involve an obvious extension of the material in the preceding section.

F. CHANNELS WITH BOTH TIME AND PARALLEL DEPENDENCIES

Thus far, we have considered only memoryless channels or, equivalently, channels with memory but no parallel dependencies. An obvious generalization is to channels with dependencies in both the time and parallel directions. We shall assume a state structure where Λ is the set of subchannel states, and a conditional probability distribution $p_\alpha(\xi/\eta)$, $\xi \in Y_s$, $\eta \in X_s$ is associated with each $\alpha \in \Lambda$.

G. BLOCK MODEL

Suppose we have a channel consisting of M_1 subchannels which, at each transmission instant, are all in the same state. Suppose, too, that there is an integer M_2 such that for any integer k, the state which is in effect at time $kM_2 + 1$ must persist until time $(k + 1) M_2$, and that the state corresponding to each value of k is independent of all the others.† The channel is cyclostationary rather than stationary, because a change in state may occur only at specified times.

The significant facts about the channel are that a block of

$$M = M_1 M_2 \tag{6-21}$$

subchannel letters is transmitted while the corresponding subchannel states are all the same, and that the state for each block is independent of the states for the others. Hence, we may make use of our MSCC results for the block model. Let R_b be the rate per block of length M_2. Then, the rate per subchannel per channel use R_s is given by

† Note that if $M_1 = 1$, we have the dual of the MSCC channel. This serves as a simple model for a single channel with memory.

$$R_s = \frac{R}{M_1} = \frac{R_b}{M} \quad . \tag{6-22}$$

Results concerning C_{sM} in Chapter 4 remain true, where M is now the number of subchannel letters in the block. We have also, for block codes of length NM_2,† that

$$P_e \leqslant \exp[-NE_M(R_s)] \tag{6-23}$$

where $E_M(R_s)$ is precisely the same as in Chapter 4, with M given by Eq. (6-21).

H. CONSTRAINED-MARKOV MODEL

Suppose we have a channel consisting of M subchannels where, at each transmission instant, all the subchannels must be in the same state. Suppose, further, that the state sequence in time is a Markov chain. Then, using the results of Yudkin,[2] we might hope to pursue a line of reasoning similar to that in Chapter 4 to prove theorems such as those in Chapter 4 for nonmaximum-likelihood decoding of block codes on this channel. This seems like a promising area for future research.

Obviously, the constrained-Markov model has a dual. This dual has a Markov state dependence in the parallel direction. In the time direction, each subchannel state persists for a "block length" of M_2 uses of the channel. At the start of a new block, the set of subchannel states is chosen independently of prior states according to the Markovian rule given. This dual seems less attractive as a model for physical channels than the constrained Markov model itself.

I. OTHER MODELS

Clearly, any one-dimensional discrete-time random process which is not independent from shot to shot may be combined with complete constraint in the parallel direction to yield a state process for a channel with both time and parallel dependencies. Since general results concerning the single-subchannel versions of such channels are not available, one would anticipate difficulty in analyzing the multiple-subchannel case.

When we consider the case of channels which are neither completely constrained nor independent in either the time or the parallel direction, it becomes difficult even to find simple models for the underlying state process. Here, the prospect for other than numerical results is slim indeed, and even numerical results can only be obtained with great difficulty.

REFERENCES

1. W. Feller, An Introduction to Probability Theory and Its Applications, Vol. 1, 2nd edition (Wiley, New York, 1957), p. 373.
2. H. L. Yudkin, "Channel State Testing in Information Decoding," Ph.D. Thesis, Department of Electrical Engineering, M.I.T. (September 1964).
3. E.N. Gilbert, "Capacity of a Burst Noise Channel," Bell System Tech. J. 39, 1253 – 1266 (1960).

† We assume that the first letter of each code word occurs at time $kM_2 + 1$, for some integer k.

APPENDIX A
CHANNELS WHICH ARE MC BUT NOT MS, AND RELATED TOPICS

Theorem A.1.

There exists a two-subchannel MC channel with $X_s = Y_s = \{0, 1\}$ which is not MS.

Proof.

Let the conditional probability distribution $p(y_1 y_2 / x_1 x_2)$ of a $2 \times 2 \times 2$ MC channel be given by the entries in the following matrix:

$y_1 y_2 \backslash x_1 x_2$	00	01	10	11
00	0.5	0	0	0
01	0	0.5	0.5	0.5
10	0	0.5	0.5	0.5
11	0.5	0	0	0

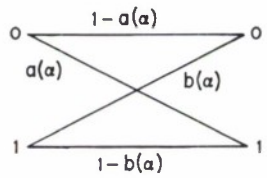
$p(y_1 y_2 / x_1 x_2)$.

It may be verified [using Eq. (2-2)] that this channel is MC. Suppose this channel is MS; then, for some set of subchannel conditional probability distributions $\{p_\alpha(\xi/\eta)\}_{\alpha \in \Lambda}$, $\xi, \eta \in \{0, 1\}$, and some joint distribution $p(\alpha_1, \alpha_2)$, we have

$$p(y_1 y_2 / x_1 x_2) = \sum_{\alpha_1 \in \Lambda} \sum_{\alpha_2 \in \Lambda} p(\alpha_1, \alpha_2)\, p_{\alpha_1}(y_1/x_1)\, p_{\alpha_2}(y_2/x_2) \quad . \tag{A-1}$$

For each value of α, we may depict $p_\alpha(\xi/\eta)$ as follows:

$p_\alpha(\xi/\eta)$

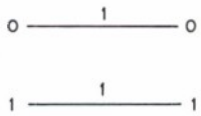


$0 \leq a(\alpha) \leq 1$

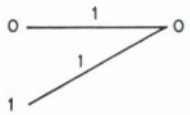
$0 \leq b(\alpha) \leq 1$

Let us consider the four "pure" channels which are diagrammed below:

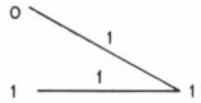
$q_1(\xi/\eta)$



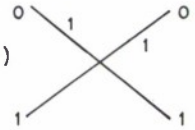
$q_2(\xi/\eta)$



$q_3(\xi/\eta)$



$q_4(\xi/\eta)$



One can easily see that $p_\alpha(\xi/\eta)$ may be represented as follows:

$$p_\alpha(\xi/\eta) = (1-a)(1-b)\, q_1(\xi/\eta) + b(1-a)\, q_2(\xi/\eta) + a(1-b)\, q_3(\xi/\eta) + abq_4(\xi/\eta) \quad . \qquad \text{(A-2)}$$

Note that all the coefficients of the q_k's are non-negative, and that a and b are functions of α. Since $p(\alpha_1, \alpha_2) \geq 0$ for all $\alpha_1, \alpha_2 \in \Lambda$, Eqs. (A-1) and (A-2) imply that there exist variables β_1 and β_2, each of which may take on values 1, 2, 3, and 4, and a joint probability distribution $p(\beta_1, \beta_2)$, such that

$$p(y_1 y_2/x_1 x_2) = \sum_{\beta_1=1}^{4} \sum_{\beta_2=2}^{4} p(\beta_1, \beta_2)\, q_{\beta_1}(y_1/x_1)\, q_{\beta_2}(y_2/x_2) \quad . \qquad \text{(A-3)}$$

[We know that $p(\beta_1, \beta_2)$ is a probability distribution because the non-negativity of the coefficients in Eqs. (A-1) and (A-2) implies $p(\beta_1, \beta_2) \geq 0$ and $\sum_{y_i \in Y_s} q_{\beta_i}(y_i/x_i) = 1$, $i = 1, 2$ implies $\sum_{\beta_1=1}^{4} \sum_{\beta_2=1}^{4} p(\beta_1, \beta_2) = 1$ from Eq. (A-3).] It is clear that Eq. (A-3) holds even if $p(\alpha_1, \alpha_2)$ is a joint density and the sums in Eq. (A-1) are replaced by integrals.

Suppose Eq. (A-3) holds for the MC channel given. Setting $x_1 = y_1 = 0$, we have

$$p(0y_2/0x_2) = \sum_{\beta_1=1}^{4} \sum_{\beta_2=1}^{4} p(\beta_1, \beta_2)\, q_{\beta_1}(0/0)\, q_{\beta_2}(y_2/x_2)$$

$$= \sum_{\beta_2=1}^{4} p(1, \beta_2)\, q_{\beta_2}(y_2/x_2) + \sum_{\beta_2=1}^{4} p(2, \beta_2)\, q_{\beta_2}(y_2/x_2) \quad . \qquad \text{(A-4)}$$

Since $p(\beta_1, \beta_2) \geq 0$ and $p(0y_2/0x_2) = 0.5 q_1(y_2/x_2)$, only the terms of Eq. (A-4) with $\beta_2 = 1$ may be nonzero. [Otherwise $p(01/00) > 0$ or $p(00/01) > 0$ or both, contrary to assumption.] Hence,

$$0.5 = p(1, 1) + p(2, 1) \quad .$$

From Eq. (A-3),

$$p(11/11) = p(1, 1) + p(1, 3) + p(3, 1) + p(3, 3) \geq p(1, 1) \geq 0 \quad .$$

Since $p(11/11) = 0$, $p(1, 1) = 0$ as well. Thus, $p(2, 1) = 0.5$. We have

$$p(00/10) = p(2, 1) + p(2, 2) + p(4, 1) + p(4, 2) \geq p(2, 1) = 0.5 \quad .$$

But, $p(00/10)$ was given to be zero; hence, we have a contradiction which proves that the channel whose transition probability matrix is given above is not MS.

Theorem A.2.

There exists an M subchannel MC channel with $X_s = Y_s = \{0, 1\}$ which is not MS.

Proof.

Pick a single-subchannel conditional probability distribution $p(\xi/\eta)$, $\xi \in Y_s$, $\eta \in X_s$ and let

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M) = p(y_1 y_2/x_1 x_2) \prod_{i=3}^{M} p(y_i/x_i) \qquad \text{(A-5)}$$

where $p(y_1 y_2/x_1 x_2)$ is given by the matrix shown in the proof of Theorem A.1. It may easily be verified that the RHS of Eq. (A-5) is the conditional probability distribution of an MC channel. Suppose that for some $\{p_\alpha(\xi/\eta)\}_{\alpha\in\Lambda}$ and joint distribution $p(\alpha_1, \ldots, \alpha_M)$, we have

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M) = \sum_{\alpha_1\in\Lambda} \cdots \sum_{\alpha_M\in\Lambda} p(\alpha_1, \ldots, \alpha_M) \times p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M) \quad . \tag{A-6}$$

Then,

$$\sum_{y_3\in Y_s} \cdots \sum_{y_M\in Y_s} p(y_1, \ldots, y_M/x_1, \ldots, x_M)$$

$$= \sum_{\alpha_1\in\Lambda} \sum_{\alpha_2\in\Lambda} p(\alpha_1, \alpha_2)\, p_{\alpha_1}(y_1/x_1)\, p_{\alpha_2}(y_2/x_2)$$

$$= p(y_1 y_2/x_1 x_2) \quad . \tag{A-7}$$

But this implies that the channel whose transition probability matrix is given in the proof of Theorem A.1 is MS, contrary to Theorem A.1.

Theorem A.3.

For any integer M, subchannel input space X_s, and subchannel output space Y_s, there exist MC channels which are not MS.

Proof.

Let $X'_s = Y'_s = \{0, 1\}$ and $p'(y'_1, \ldots, y'_M/x'_1, \ldots, x'_M)$ be given by the RHS of Eq. (A-5) with $x_i \to x'_i$, and $y_i \to y'_i$. Define a function f(x) from X_s onto X'_s, and two probability distributions (or densities) $p_0(\xi)$ and $p_1(\xi)$ over Y_s such that $p_0(\xi)\, p_1(\xi) = 0$ for all $\xi \in Y_s$. Define†

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M) = \sum_{y'_1\in Y'_s} \cdots \sum_{y'_M\in Y'_s} \left[\prod_{i=1}^{M} p_{y'_i}(y_i) \right] \times p'[y'_1, \ldots, y'_M/f(x_1), \ldots, f(x_M)] \quad . \tag{A-8}$$

The relationship between the primed (original) and unprimed (derived) channels is shown in Fig. A-1. Since the primed channel is MC, clearly the unprimed channel is MC, too. Now, suppose for some $\{p_\alpha(\xi/\eta)\}_{\alpha\in\Lambda}$, $\xi \in Y_s$, $\eta \in X_s$ and some joint distribution $p(\alpha_1, \ldots, \alpha_M)$, we have

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M) = \sum_{\alpha_1\in\Lambda} \cdots \sum_{\alpha_M\in\Lambda} p(\alpha_1, \ldots, \alpha_M) \times p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M) \quad . \tag{A-9}$$

† As usual, $p_{y'_i}(y_i) = p(y_i/y'_i)$.

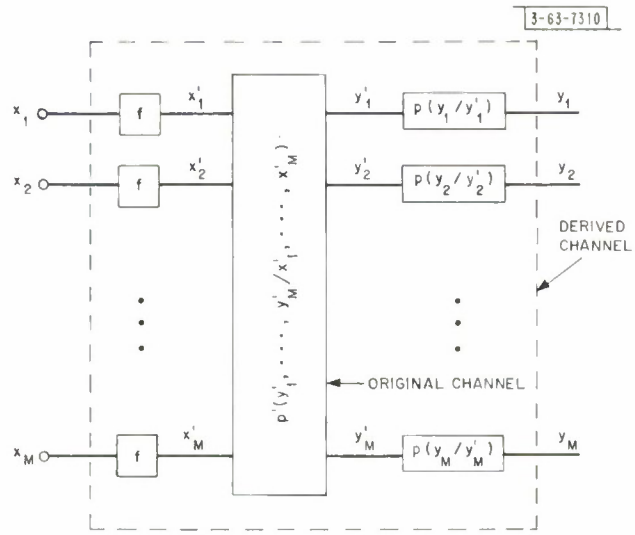


Fig. A-1. Relotionship between original and derived channels.

Let

$$Y_{so} \text{ be the set of } \xi \in Y_s \text{ for which } p_o(\xi) > 0$$

and

$$Y_{s1} \text{ be the set of } \xi \in Y_s \text{ for which } p_1(\xi) > 0 \quad .$$

Since $p_o(\xi)\, p_1(\xi) = 0$ for all $\xi \in Y_s$, Y_{so} and Y_{s1} are disjoint. If $r_i \in Y'_s$, define

$$P_{\alpha_i}(r_i/x_i) = \sum_{y_i \in Y_{sr_i}} p_{\alpha_i}(y_i/x_i) \quad . \tag{A-10}$$

Note that

$$\sum_{y_i \in Y_{sr_i}} p_{y'_i}(y_i) = \delta_{y'_i}^{r_i} \quad . \tag{A-11}$$

Pick

$$(r_1, \ldots, r_M) \quad , \quad r_i \in Y'_s \quad , \quad i = 1, \ldots, M$$

and

$$(u_1, \ldots, u_M) \quad , \quad u_i \in X'_s \quad , \quad i = 1, \ldots, M$$

Choose

$$(x_1, \ldots, x_M) \quad , \quad x_i \in X_s \quad , \quad i = 1, \ldots, M$$

so that

$$f(x_i) = u_i \qquad i = 1, \ldots, M \quad . \tag{A-12}$$

From Eqs. (A-8) and (A-9),

$$\sum_{y'_1 \epsilon Y'_s} \cdots \sum_{y'_M \epsilon Y'_s} \left[\prod_{i=1}^{M} p_{y'_i}(y_i) \right] p'[y'_1, \ldots, y'_M / f(x_1), \ldots, f(x_M)]$$

$$= \sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M)\, p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M) \quad . \quad \text{(A-13)}$$

Summing both sides of Eq. (A-13) over all $y_i \epsilon Y_{sr_i}$, $i \leqslant i \leqslant M$, and using Eqs. (A-10), (A-11), and (A-12), we get

$$\sum_{y'_1 \epsilon Y'_s} \cdots \sum_{y'_M \epsilon Y'_s} \left[\prod_{i=1}^{M} \delta_{y'_i}^{r_i} \right] p'(y'_1, \ldots, y'_M / u_1, \ldots, u_M)$$

$$= \sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M)\, P_{\alpha_1}(r_1/x_1) \cdots P_{\alpha_M}(r_M/x_M) \quad . \quad \text{(A-14)}$$

This reduces further to

$$p'(r_1, \ldots, r_M / u_1, \ldots, u_M) = \sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M)$$

$$\times\, w_{\alpha_1}(r_1/u_1) \cdots w_{\alpha_M}(r_M/u_M) \qquad \text{(A-15)}$$

where $w_{\alpha_i}(r_i/u_i)$ is defined by

$$w_{\alpha_i}[r_i/f(x_i)] = P_{\alpha_i}(r_i/x_i) \qquad \text{(A-16)}$$

for all $\alpha_i \epsilon \Lambda$, $r_i \epsilon Y'_s$, and $x_i \epsilon X_s$. Clearly, w_α is a conditional distribution on $Y'_s \times X'_s$. Since $(r_1, \ldots, r_M)$ and $(u_1, \ldots, u_M)$ were arbitrary, Eq. (A-15) implies that the primed channel is MS, contrary to Theorem A.2.

Theorem A.4.

There exists an NII channel with $X_s = Y_s = \{0, 1\}$, which is not MST.

Proof.

For ℓ an even integer, and N an even positive integer, define

$$p(y_{\ell+1}, \ldots, y_{\ell+N} / x_{\ell+1}, \ldots, x_{\ell+N}) = \prod_{k=1}^{N/2} p(y_{\ell+2k-1}, y_{\ell+2k} / x_{\ell+2k-1}, x_{\ell+2k}) \qquad \text{(A-17)}$$

where the bivariate conditional probability is given by the matrix shown in the proof of Theorem A.1. The conditional probability for other values of ℓ and N can be obtained by

summing conditional probabilities of the form given over appropriate outputs. Since each odd input-output pair and the succeeding even one "stand alone," it is clear (an argument similar to that used to prove Theorem A.2 may be used) that the channel given by Eq. (A-17) is NII but not MST.

The restriction to binary subchannel alphabets may be removed as in Theorem A.3.

We note that the channel given by Eq. (A-17) is cyclostationary rather than stationary. In fact, there is a unique stationary channel with bivariate conditional probabilities given by the entries in the following matrix:

$y_i y_{i+1}$ \ $x_i x_{i+1}$	00	01	10	11
00	0.5	0	0	0
01	0	0.5	0.5	0.5
10	0	0.5	0.5	0.5
11	0.5	0	0	0

$p(y_i y_{i+1}/x_i x_{i+1})$.

This channel has the property that given a single input sequence, there are only two possible output sequences, each having probability one-half. This is true regardless of the length of the input sequence. Unfortunately, this channel is not NII, and this fact is proven as follows:

(1) The input $(x_1, x_2, x_3) = (0, 0, 1)$ may result in the outputs $(y_1, y_2, y_3) = (0, 0, 1)$ or $(1, 1, 0)$, each with probability one-half.

(2) The input $(x_1, x_2, x_3) = (0, 1, 1)$ may result in the outputs $(y_1, y_2, y_3) = (0, 1, 0)$ or $(1, 0, 1)$, each with probability one-half.

Hence,

$$p(001/001) + p(011/001) = \frac{1}{2} + 0 = \frac{1}{2}$$

and

$$p(001/011) + p(011/011) = 0 + 0 = 0 \quad .$$

Thus, the channel referred to is not NII. It is not known whether there exist strictly stationary NII channels which are not MST.

APPENDIX B
PROOFS OF THEOREMS 2.3 AND 2.4

Let P be the set of probability distributions over a finite product space X_s^M, and D be the set of product distributions over X_s^M. Let X_s^M consist of K points. Clearly,

$$D \subseteq P \subset \Re^K \tag{B-1}$$

and P is compact.†

Lemma.

D is compact in $\Re^K$.

Proof.

D is bounded because $D \subseteq P$, and P is bounded. Let $p(x_1, \ldots, x_M)$ be a limit point of D. Then, there exists a sequence $\left\{ \prod_{i=1}^{M} q_{\ell_i}(x_i) \right\}_{\ell=1}^{\infty}$ of product distributions with

$$\lim_{\ell \to \infty} \sum_{x_1 \epsilon X_s} \cdots \sum_{x_M \epsilon X_s} \left| p(x_1, \ldots, x_M) - \prod_{i=1}^{M} q_{\ell_i}(x_i) \right|^2 = 0 \quad . \tag{B-2}$$

Since the sums are finite, this implies

$$\lim_{\ell \to \infty} \sum_{x_1 \epsilon X_s} \cdots \sum_{x_M \epsilon X_s} \left| p(x_1, \ldots, x_M) - \prod_{i=1}^{M} q_{\ell_i}(x_i) \right| = 0 \quad . \tag{B-3}$$

Now,

$$\sum_{x_1 \epsilon X_s} \cdots \sum_{x_M \epsilon X_s} \left| p(x_1, \ldots, x_M) - \prod_{i=1}^{M} q_{\ell_i}(x_i) \right|$$

$$\geqslant \sum_{x_i \epsilon X_s} \left| p_i(x_i) - q_{\ell_i}(x_i) \right| \geqslant 0 \qquad i = 1, \ldots, M \tag{B-4}$$

where $p_i(x_i)$ is the marginal distribution of the ith subchannel input associated with the joint distribution $p(x_1, \ldots, x_M)$. From Eqs. (B-3) and (B-4),

$$\lim_{\ell \to \infty} q_{\ell_i}(x_i) = p_i(x_i) \quad . \tag{B-5}$$

From Eqs. (B-2) and (B-5),

$$p(x_1, \ldots, x_M) = \lim_{\ell \to \infty} \prod_{i=1}^{M} q_{\ell_i}(x_i) = \prod_{i=1}^{M} p_i(x_i) \quad . \tag{B-6}$$

---

† That is, closed and bounded. See W. Rudin, Principles of Mathematical Analysis, 2nd edition (McGraw-Hill, New York, 1964), Theorem 2.41.

Hence, $p(x_1, \ldots, x_M)$ is a product distribution, and D is closed. Since D is closed and bounded, it is compact.

Proof of Theorem 2.3.

We note that since D and P are compact subspaces of $\mathcal{R}^K$, $D \times \{[0, 1]\}$ and $P \times \{[0, 1]\}$ are compact subspaces of $\mathcal{R}^{K+1}$.

(1) Since a continuous function defined on a compact set is bounded,† $F_D(R)$, $G_D(R)$, $F_P(R)$, and $G_P(R)$ are finite.

(2) Since a continuous function defined on a compact set achieves its maximum,‡ for each value of R there exist $0 \leqslant \rho^* \leqslant 1$ and $\vec{p}^* \in D$ with

$$F_D(R) = f(\rho^*, \vec{p}^*, R) \quad . \tag{B-7}$$

Thus,

$$F_D(R) = f(\rho^*, \vec{p}^*, R) \leqslant g(\rho^*, \vec{p}^*, R) \leqslant G_D(R) \leqslant G_P(R) \tag{B-8}$$

as required. The last inequality is an obvious consequence of Eq. (B-1). Clearly, if $f < g$, the inequality between F_D and G_D is strict.

(3) For the same reason as in (2) above, for each value of R there exist $0 \leqslant \rho' \leqslant 1$ and $\vec{p}' \in P$ with

$$F_P(R) = f(\rho', \vec{p}', R) \quad . \tag{B-9}$$

Hence,

$$F_D(R) \leqslant F_P(R) = f(\rho', \vec{p}', R) \leqslant g(\rho', \vec{p}', R) \leqslant G_P(R) \quad . \tag{B-10}$$

The first inequality is an obvious consequence of Eq. (B-1). It is clear that if $f < g$, the inequality between F_P and G_P is strict.

Proof of Theorem 2.4.

(1) For each R, pick $\epsilon > 0$. There exist $0 \leqslant \rho' \leqslant 1$ and $p' \in D$ with

$$0 \leqslant F_D(R) - f(\rho', p', R) \leqslant \epsilon \quad . \tag{B-11}$$

Thus,

$$G_D(R) - F_D(R) = [G_D(R) - g(\rho', p', R)] + [g(\rho', p', R) - f(\rho', p', R)]$$
$$+ [f(\rho', p', R) - F_D(R)] \geqslant 0 + 0 - \epsilon = -\epsilon \quad .$$

† W. Rudin, op. cit., Theorem 4.15.

‡ Ibid., Theorem 4.16. If this were not so, we would have used l.u.b. instead of max in the definitions of F_D, G_D, F_P, and G_P.

Since ϵ was arbitrary, we have

$$G_D(R) - F_D(R) \geqslant 0$$

as required. The rest of Eq. (2-22) follows from Eq. (B-1).

(2) The proof here proceeds as in (1) with P replacing D.

We note that if f and/or g fail to depend on any or all of ρ, $\vec{p}$, and R, the conclusions of Theorems 2.3 and 2.4 remain valid.

APPENDIX C
SOME USEFUL INEQUALITIES

This appendix contains statements of the important nontrivial inequalities used in the text. A reference is given for each inequality stated. In any inequality in which λ appears, we assume $0 < \lambda \leqslant 1$.

(1) Let $\{a_i\}_{i=1}^{N}$ be a sequence of non-negative numbers. Then†

$$\left(\sum_{i=1}^{N} a_i\right)^{\lambda} \leqslant \sum_{i=1}^{N} a_i^{\lambda} \quad . \tag{C-1}$$

(2) Let $\{a_i\}_{i=1}^{N}$ and $\{b_i\}_{i=1}^{N}$ each be sequences of non-negative numbers with $\sum_{i=1}^{N} b_i = 1$. Then,‡

$$\sum_{i=1}^{N} b_i a_i^{\lambda} \leqslant \left(\sum_{i=1}^{N} b_i a_i\right)^{\lambda} \quad . \tag{C-2}$$

(3) Minkowski's inequality: let $\{a_{ij}\}_{i=1}^{N}{}_{j=1}^{M}$, and $\{p_j\}_{j=1}^{M}$ be sets of non-negative numbers with $\sum_{j=1}^{M} p_j = 1$. Then,§

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{M} p_j a_{ij}^{\lambda}\right)^{1/\lambda} \leqslant \left[\sum_{j=1}^{M} p_j\left(\sum_{i=1}^{N} a_{ij}\right)^{\lambda}\right]^{1/\lambda} \quad . \tag{C-3}$$

---

† G.H. Hardy, J.E. Littlewood, and G. Polya, Inequalities (Cambridge University Press, Cambridge, England, 1959). See Theorem 19, p. 28.

‡ Ibid., Theorem 16, p. 22.

§ Ibid., Theorem 24, p. 30.

APPENDIX D
CANONICAL REPRESENTATIONS

Theorem D.1.

For any MS channel with finite input and output alphabets, there always exists a representation for which in each channel state the subchannel conditional probabilities are all either zero or unity.

Proof.

Let Λ, $X_s = \{1, \ldots, L\}$, $Y_s = \{1, \ldots, Q\}$, $p_\alpha(q/\ell)$ $\alpha \in \Lambda$, $q \in Y_s$, $\ell \in X_s$, $p(\alpha_1, \ldots, \alpha_M)$ be given and

$$p(y_1, \ldots, y_M/x_1, \ldots, x_M)$$

$$= \sum_{\alpha_1 \in \Lambda} \cdots \sum_{\alpha_M \in \Lambda} p(\alpha_1, \ldots, \alpha_M)\, p_{\alpha_1}(y_1/x_1) \cdots p_{\alpha_M}(y_M/x_M) \quad . \tag{D-1}$$

We define a pure (sub) channel as one for which the input completely determines the output. There are Q^L possible different pure subchannels. We denote the conditional probability distribution associated with a pure subchannel by $s_\beta(q/\ell)$, $\ell \in X_s$, $q \in Y_s$, where $1 \leqslant \beta \leqslant Q^L$. We note that for each β, ℓ there is a unique value of q with $s_\beta(q/\ell) = 1$.

If $1 \leqslant b_k \leqslant Q^L$ for $1 \leqslant k \leqslant M$, then

$$p(y/x) = \prod_{k=1}^{M} s_{b_k}(y_k/x_k) \tag{D-2}$$

is the conditional probability distribution of a pure channel with L^M inputs and Q^M outputs.

For each $\alpha \in \Lambda$, we shall show that it is possible to expand $p_\alpha(q/\ell)$ in a series of $s_\beta(q/\ell)$ with non-negative coefficients, as shown in Eq. (D-3).

$$p_\alpha(q/\ell) = \sum_{n=1}^{Q^L} C_n^\alpha s_n(q/\ell) \tag{D-3}$$

where $C_n^\alpha \geqslant 0$, $\alpha \in \Lambda$, and $1 \leqslant n \leqslant Q^L$. The expansion Eq. (D-3) is not generally unique, but one way of finding the C_n^α is to proceed as follows: Let

$$p_\alpha^o(q/\ell) = p_\alpha(q/\ell) \quad . \tag{D-4}$$

By definition of a probability distribution, $p_\alpha^o(q/\ell) \geqslant 0$, $1 \leqslant \ell \leqslant L$, $1 \leqslant q \leqslant Q$. Find the smallest nonzero transition probability in the set $\{p_\alpha^o(q/\ell)\}_{\ell=1}^{L}{}_{q=1}^{Q}$. Call it r_α^1. Now,

$$\sum_{q=1}^{Q} p_\alpha^o(q/\ell) = 1 > 0 \tag{D-5}$$

for all ℓ, $1 \leqslant \ell \leqslant L$. Thus, there must be a (not necessarily unique) function $q^1(\ell)$, $1 \leqslant q^1(\ell) \leqslant Q$, with $p_\alpha^o\,[q^1(\ell)/\ell] > 0$, $1 \leqslant \ell \leqslant L$. By definition of r_α^1,

$$p_\alpha^o\,[q^1(\ell)/\ell] \geqslant r_\alpha^1 > 0 \tag{D-6}$$

for all ℓ, $1 \leqslant \ell \leqslant L$. Choose $q^1(\ell)$ so that Eq. (D-6) is satisfied with equality for at least one value of ℓ. Let s_{b_1} be the pure subchannel with

$$s_{b_1}[q^1(\ell)/\ell] = 1 \quad . \tag{D-7}$$

Let

$$p_\alpha^1(q/\ell) = p_\alpha^o(q/\ell) - r_\alpha^1 s_{b_1}(q/\ell) \quad . \tag{D-8}$$

Clearly,

$$p_\alpha^1(q/\ell) \geqslant 0 \quad , \quad 1 \leqslant \ell \leqslant L \quad , \quad 1 \leqslant q \leqslant Q \quad .$$

If $p_\alpha^1(q/\ell)$ is not identically zero, the process may be continued. Note that at each step of the process, a positive value of $p_\alpha^m(q/\ell)$ is converted to a zero value of $p_\alpha^{m+1}(q/\ell)$. Since there are, at most, LQ positive values of $p_\alpha^o(q/\ell)$, the process must terminate after, at most, LQ steps, and we may write

$$p_\alpha(q/\ell) = \sum_{m=1}^{N} r_\alpha^m s_{b_m}(q/\ell) \tag{D-9}$$

where

$$N \leqslant LQ \quad . \tag{D-10}$$

Equation (D-9) may be converted to Eq. (D-3) if $r_\alpha^m \to C_m^\alpha$ and the $s_\beta(q/\ell)$ are renumbered so that $b_m \to m$. Now, from Eq. (D-3),

$$\prod_{i=1}^{M} p_{\alpha_i}(y_i/x_i) = \prod_{i=1}^{M} \left[\sum_{n_i=1}^{Q^L} C_{n_i}^{\alpha_i} s_{n_i}(y_i/x_i) \right] \quad . \tag{D-11}$$

Thus, from Eqs. (D-1) and (D-11),

$$\begin{aligned}
&p(y_1, \ldots, y_M/x_1, \ldots, x_M) \\
&= \sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M) \prod_{i=1}^{M} \left[\sum_{n_i=1}^{Q^L} C_{n_i}^{\alpha_i} s_{n_i}(y_i/x_i) \right] \\
&= \sum_{n_1=1}^{Q^L} \cdots \sum_{n_M=1}^{Q^L} \left[\sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M) C_{n_1}^{\alpha_1} \cdots C_{n_M}^{\alpha_M} \right] \\
&\quad \times s_{n_1}(y_1/x_1) \cdots s_{n_M}(y_M/x_M) \\
&= \sum_{n_1=1}^{Q^L} \cdots \sum_{n_M=1}^{Q^L} d(n_1, \ldots, n_M)\, s_{n_1}(y_1/x_1) \cdots s_{n_M}(y_M/x_M)
\end{aligned} \tag{D-12}$$

where

$$d(n_1, \ldots, n_M) = \sum_{\alpha_1 \epsilon \Lambda} \cdots \sum_{\alpha_M \epsilon \Lambda} p(\alpha_1, \ldots, \alpha_M)\, C_{n_1}^{\alpha_1} \cdots C_{n_M}^{\alpha_M} \quad . \tag{D-13}$$

$d(n_1, \ldots, n_M)$ is clearly a probability distribution because it is non-negative valued by Eq. (D-13), and a summation of both sides of Eq. (D-12) over $y_1, \ldots, y_M$ shows that it is properly normalized. Thus, we have our desired representation.

APPENDIX E
A COMPLETELY CONSTRAINED CHANNEL WITH A CONTINUOUS PARAMETER

In this appendix, we shall give an example of an MSCC channel whose capacity per subchannel with state unknown approaches the capacity of a single subchannel with state known as the number of subchannels increases, although the state representation for the channel is not discrete. Suppose we have an MS channel consisting of M parallel binary symmetric subchannels. For each use of the channel, the crossover probability g is the same for all the subchannels. Thus, this channel is MSCC. A probability density p(g) is given with $p(g) = 0$, unless $0 \leqslant g \leqslant 1$. We will suppose that each possible input is used with probability $(1/2)^M$, independently of the channel state. It is easy to see that this is the input distribution which achieves capacity, whether or not the state is known at the receiver. Now,

$$I(XY;\ G) = I(Y;\ G) + I(X;\ G/Y) \quad . \tag{E-1}$$

By symmetry, $p(y/g) = (1/2)^M = p(y)$. Thus, y and g are independent, and $I(Y;\ G) = 0$. Hence,

$$I(XY;\ G) = I(X;\ G/Y) \quad . \tag{E-2}$$

Combining Eqs. (3-4) and (E-2), we have

$$I(X;\ Y) = I(X;\ Y/G) - I(XY;\ G) \quad . \tag{E-3}$$

For each value of g, p(y/xg) depends only on the Hamming distance $d_H(x, y)$ between x and y. Let D be the ensemble of such distances.

$$\begin{aligned} I(XYD;\ G) &= I(XY;\ G) + I(D;\ G/XY) \\ &= I(D;\ G) + I(XY;\ G/D) \quad . \end{aligned} \tag{E-4}$$

Now,

$$p(d/gxy) = p(d/xy)$$

and

$$p(xy/gd) = p(xy/d) \quad . \tag{E-5}$$

Hence,

$$I(D;\ G/XY) = 0 = I(XY;\ G/D) \tag{E-6}$$

and, combining Eqs. (E-6) and (E-4), we get

$$I(XY;\ G) = I(D;\ G) \quad . \tag{E-7}$$

Combining Eqs. (E-7) and (E-3), we get

$$I(X;\ Y) = I(X;\ Y/G) - I(D;\ G) \quad . \tag{E-8}$$

Now, $D = \{0, 1, \ldots, M\}$; hence,

$$I(D;\ G) \leqslant H(D) \leqslant \log(M + 1) \quad . \tag{E-9}$$

Thus, using Theorem 3.1, Eqs. (E-8) and (E-9), we have

$$I(X;\ Y/G) - \log(M + 1) \leqslant I(X;\ Y) \leqslant I(X;\ Y/G) \quad . \tag{E-10}$$

We may readily compute

$$I(X;\ Y/G) = M \int_0^1 [1 + g \log g + (1 - g) \log(1 - g)]\ p(g)\ dg$$

$$= M\ I(X_1;\ Y_1/G) \quad . \tag{E-11}$$

Thus, from Eqs. (E-10) and (E-11),

$$I(X_1;\ Y_1/G) - \frac{\log(M + 1)}{M} \leqslant \frac{I(X;\ Y)}{M} \leqslant I(X_1;\ Y_1/G) \quad . \tag{E-12}$$

Now,

$$\lim_{M \to \infty} \frac{\log(M + 1)}{M} = 0 \quad . \tag{E-13}$$

Hence,

$$\lim_{M \to \infty} \frac{I(X;\ Y)}{M} = I(X_1;\ Y_1/G) \quad . \tag{E-14}$$

From the remarks preceding Eq. (E-1), Eq. (E-14) implies

$$\lim_{M \to \infty} C_{sM} = C'_1 \quad . \tag{E-15}$$

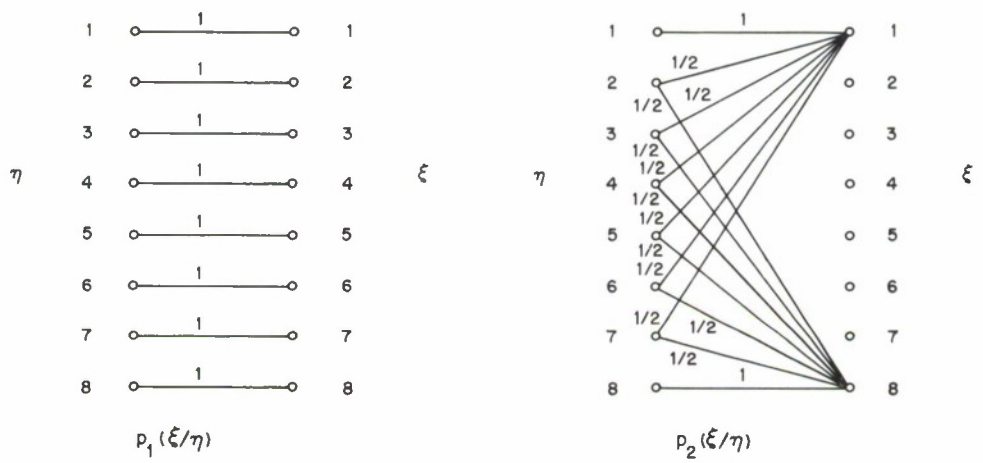
APPENDIX F
AN MSCC CHANNEL FOR WHICH $E'_M(R_s) \neq \tilde{E}_M(R_s)$

Consider the 2SCC channel defined [see Eq. (4-1)] by $X_s = Y_s = \{1, \ldots, 8\}$, $p(\alpha) = 1/2$, $\alpha = 1, 2$, and

$$p_1(\xi/\eta) = \begin{cases} 1 & \eta = \xi \\ 0 & \text{otherwise} \end{cases}$$

$$p_2(\xi/\eta) = \begin{cases} 0 & \text{if } \xi = 2, \ldots, 7 \\ \frac{1}{2} & \text{if } \xi = 1, 8 \text{ and } \eta = 2, \ldots, 7 \\ 1 & \text{if } (\xi, \eta) = (1, 1) \text{ or } (8, 8) \\ 0 & \text{if } (\xi, \eta) = (1, 8) \text{ or } (8, 1) \end{cases} .$$

The subchannel states are depicted as follows:



In the computation of $E'_2(R_s)$, we are first concerned with the minimization for each value of ρ, $0 \leqslant \rho \leqslant 1$, of the function $F'_2(\rho, \vec{p})$ over all $\vec{p} \in P$ [see Eq. (4-33)]. Using Eq. (4-19) and taking advantage of the available symmetries, we find that for purposes of minimization we may consider F'_2 as a function of a reduced probability vector $\vec{p}_r$:

$$F'_2(\rho, \vec{p}_r) = \frac{1}{2} \left[4p_r(1, 1)^{1+\rho} + 24p_r(1, 2)^{1+\rho} + 36p_r(2, 2)^{1+\rho}\right]$$

$$+ 2 \left[p_r(1, 1) + 12p_r(1, 2) \left(\frac{1}{2}\right)^{1/1+\rho} + 36p_r(2, 2) \left(\frac{1}{4}\right)^{1/1+\rho}\right]^{1+\rho}$$

where

$$\vec{p}_r = [p_r(1, 1),\ p_r(1, 2),\ p_r(2, 2)]$$

$$p_r(1, 1) \geqslant 0 \cdot , \quad p_r(1, 2) \geqslant 0 \quad , \quad p_r(2, 2) \geqslant 0$$

and

$$4p_r(1, 1) + 24p_r(1, 2) + 36p_r(2, 2) = 1 \quad .$$

The components $p(\eta_1, \eta_2)$, $\eta_1, \eta_2 = 1, \ldots, 8$, of the original probability vector $\vec{p}$ may be obtained from those of the reduced probability vector $\vec{p}_r$ as follows.

Define

$$S_1 = \{1, 8\}$$

$$S_2 = \{2, \ldots, 7\}$$

$$A_{11} = \{(\eta_1, \eta_2)/\eta_1 \in S_1,\ \eta_2 \in S_1\}$$

$$A_{12} = \{(\eta_1, \eta_2)/\eta_1 \in S_1,\ \eta_2 \in S_2 \text{ or } \eta_1 \in S_2,\ \eta_2 \in S_1\}$$

$$A_{22} = \{(\eta_1, \eta_2)/\eta_1 \in S_2,\ \eta_2 \in S_2\} \quad .$$

Then, $(\eta_1, \eta_2) \in A_{ij}$ implies

$$p(\eta_1, \eta_2) = p_r(i, j) \quad , \quad i, j = 1, 2 \quad .$$

In the computation of $\tilde{E}_2(R_s)$, we are first concerned with the minimization for each value of ρ, $0 \leqslant \rho \leqslant 1$, of the function $F_2'(\rho, \vec{p})$ over all $\vec{p} \in D$ [see Eq. (4-79)]. Using Eq. (4-47) [replacing the LHS by $G_2'(\rho, \vec{q})$, and $\vec{p}_s$ by $\vec{q}_s$ to avoid confusion] and taking advantage of the available symmetries, we find that for purposes of minimization we may consider G_2' as a function of a reduced probability vector $\vec{q}_r$:

$$G_2'(\rho, \vec{q}_r) = \frac{1}{2}\,[2q_{sr}(1)^{1+\rho} + 6q_{sr}(2)^{1+\rho}]^2 + 2\,[q_{sr}(1) + 6q_{sr}(2)\,(\tfrac{1}{2})^{1/1+\rho}]^{2(1+\rho)}$$

where

$$\vec{q}_r = (\vec{q}_{sr})^2$$

$$\vec{q}_{sr} = [q_{sr}(1),\ q_{sr}(2)]$$

$$q_{sr}(1) \geqslant 0 \quad , \quad q_{sr}(2) \geqslant 0$$

and

$$2q_{sr}(1) + 6q_{sr}(2) = 1 \quad .$$

The components $q_s(\eta)$, $\eta = 1, \ldots, 8$, of the original subchannel probability vector $\vec{q}_s$ may be obtained from those of the reduced probability vector $\vec{q}_{sr}$ as follows:

$$\eta \in S_i \quad \text{implies} \quad q_s(\eta) = q_{sr}(i) \quad , \quad i = 1, 2 \quad .$$

Thus, we may compute (for R_s in bits)

$$E_2'(R_s) = \max_{0 \leqslant \rho \leqslant 1} \left[-2\rho R_s \ln 2 - \ln \min_{\vec{p}_r} F_2'(\rho, \vec{p}_r) \right]$$

and

$$\tilde{E}_2(R_s) = \max_{0 \leqslant \rho \leqslant 1} \left[-2\rho R_s \ln 2 - \ln \min_{\vec{q}_r} G_2'(\rho, \vec{q}_r) \right] \quad .$$

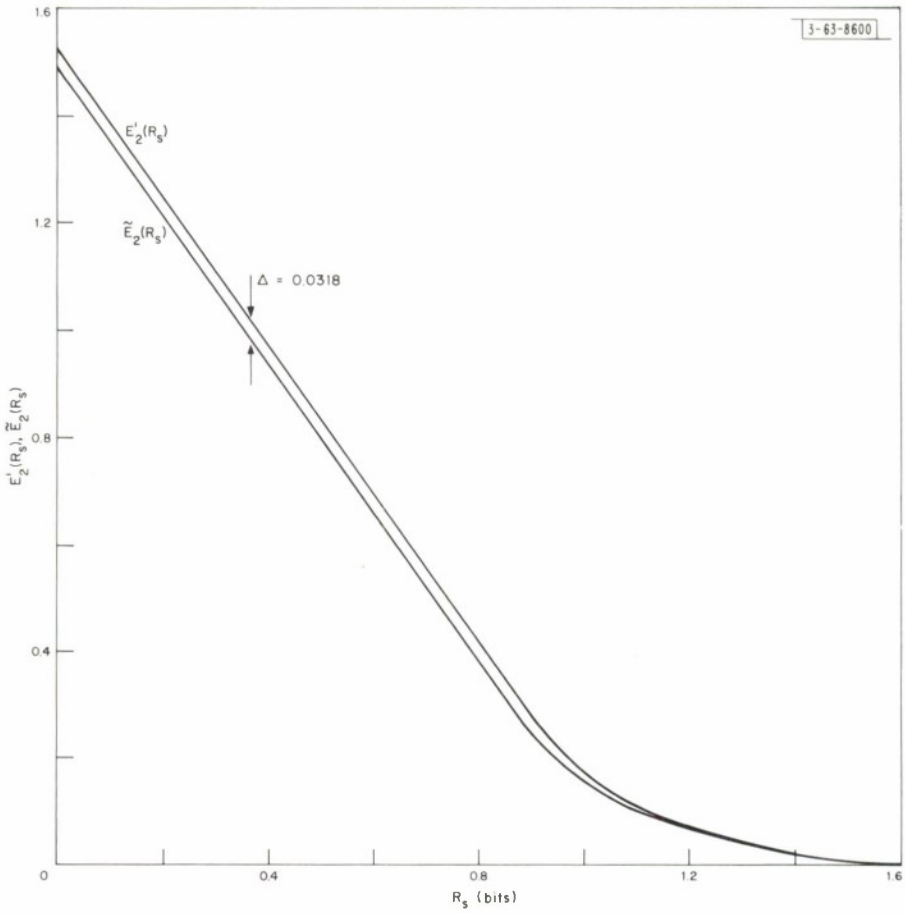


Fig. F-1. $E_2'(R_s)$ and $\tilde{E}_2(R_s)$ vs R_s for a particular MSCC channel.

The minimizations over $\vec{p}_r$ and $\vec{q}_r$ were performed using the method of Zoutendijk.† The entire computation was programmed on the IBM 360. Results are given in Fig. F-1, where the vertical distance Δ between the straight-line portions of the two curves is 0.0318. The guaranteed accuracy of Δ is given by the bound $0.0301 < \Delta < 0.0319$. This bound is believed to be conservative (i.e., Δ is believed to be given by 0.0318 to three significant figures). The results clearly demonstrate that $E_2'(R_s) \neq \tilde{E}_2(R_s)$ for this channel.

Table F-1 gives some of the values of $E_2'(R_s)$ and $\tilde{E}_2(R_s)$, together with the values of ρ, $\vec{p}_r$, and $\vec{q}_r$ which achieve the maxima required by the definitions of $E_2'(R_s)$ and $\tilde{E}_2(R_s)$.

The marginal distribution of subchannel inputs $p(\eta)$ (the same for both subchannels) corresponding to $\vec{p}_r$ may be computed as follows:

$$p_r(1) = 6p_r(1, 2) + 2p_r(1, 1)$$

$$p_r(2) = 6p_r(2, 2) + 2p_r(1, 2)$$

$$\eta \in S_i \quad \text{implies} \quad p(\eta) = p_r(i) \quad , \quad i = 1, 2 \quad .$$

The capacity C_2' for this channel is 1.660964 bits, achieved by a product distribution (see proof of Theorem 2.7) with $\vec{q}_r = (0.19992979, 0.10002340)$.

† G. Zoutendijk, Methods of Feosible Directions (Elsevier, Amsterdam, 1960).

TABLE F-1

VARIOUS PARAMETERS INVOLVED IN THE CALCULATION OF $E_2'(R_s)$ AND $\tilde{E}_2(R_s)$ FOR A PARTICULAR MSCC CHANNEL

R_s	$E_2'(R_s)$			$\tilde{E}_2(R_s)$		
	Value	Maximizing ρ	Maximizing $\vec{p}_r$	Value	Maximizing ρ	Maximizing $\vec{q}_r$
0	1.5235	1	0.14054, 0.01824, 0.0	1.4916	1	0.40918, 0.03027
0.9	0.2758	1	0.14054, 0.01824, 0.0	0.2440	1	0.40918, 0.03027
1.0	0.1737	0.58	0.09850, 0.02525, 0.0	0.1566	0.44	0.31483, 0.06172
1.1	0.1085	0.35	0.07477, 0.02538, 0.00255	0.1057	0.31	0.27989, 0.07337
1.2	0.0690	0.24	0.06147, 0.02343, 0.00533	0.0683	0.23	0.25683, 0.08106
1.3	0.0408	0.17	0.05396, 0.02226, 0.00694	0.0406	0.17	0.24011, 0.08663
1.4	0.0207	0.12	0.04903, 0.02155, 0.00796	0.0207	0.12	0.22701, 0.09100
1.5	0.0077	0.07	0.04478, 0.02088, 0.00889	0.0077	0.07	0.21493, 0.09502
1.6	0.0011	0.03	0.04162, 0.02036, 0.00958	0.0011	0.03	0.20601, 0.09800

ACKNOWLEDGMENTS

It is with great pleasure that I acknowledge the contributions made by Professor R. G. Gallager, upon whose formulation of the random coding exponent a substantial portion of this report rests. His continued strong interest in this work was of the greatest help in the development of my ideas.

Also, I wish to acknowledge the contributions of Professors P. Elias and R. S. Kennedy, whose stimulating discussions aided this effort in no small measure.

I would like to express my gratitude to Miss M. M. Pennell and Mrs. H. S. Davis, who were principally responsible for the machine computation of some of the examples in this report.

I want to express my appreciation to M. I. T.'s Lincoln Laboratory for the generous financial support given to me under the Staff Associate Program. Finally, I wish to thank the Research Laboratory of Electronics at M. I. T. for the use of their facilities.

DOCUMENT CONTROL DATA – R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Lincoln Laboratory, M.I.T.	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP None
3. REPORT TITLE Parallel Channels Without Crosstalk	
4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)* Technical Report	
5. AUTHOR(S) *(Last name, first name, initial)* Max, Joel	
6. REPORT DATE 9 April 1968	7a. TOTAL NO. OF PAGES 116 7b. NO. OF REFS 15
8a. CONTRACT OR GRANT NO. AF 19 (628)-5167 b. PROJECT NO. 649L c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Report 447 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* ESD-TR-68-54
10. AVAILABILITY/LIMITATION NOTICES This document has been approved for public release and sale; its distribution is unlimited.	
11. SUPPLEMENTARY NOTES None	12. SPONSORING MILITARY ACTIVITY Air Force Systems Command, USAF

13. ABSTRACT

In this report, a study is made of information theoretic channels which are decomposable into a number of parallel subchannels which will, in general, be dependent. For this situation, two models are constructed in which each subchannel input affects only the corresponding subchannel output (no crosstalk). In the first model (MC channel), the lack of crosstalk is ensured by constraints on the channel conditional probability distribution. The second model (MS channel) is a channel with an underlying state structure with states independent of the input. Both models are memoryless. All MS channels are MC, but the reverse does not hold.

The effect of subchannel dependencies on capacity and random coding exponent (RCE) is investigated. It is proved that these dependencies cannot decrease the capacity of our channels. However, subchannel dependencies may either increase or decrease the RCE. It is also proved that the capacity of the channel is not less than the sum of the capacities of the individual subchannels. When the state model is used, the above two quantities are equal if the receiver has knowledge of the channel state.

A definition of partial state knowledge is given. It is proved that, when the receiver has partial state knowledge, the resulting capacity and RCE are not decreased. For complete state knowledge at the receiver, the capacity and RCE are not less than those obtained for partial state knowledge.

A restricted class of MS channels is defined wherein all the subchannels are in the same state during each use of the channel; these channels are called MSCC channels. For these channels, a number of results are given, most of which concern the limiting behavior of the capacity per subchannel and the RCE as the number of subchannels becomes large. The principal results are: (1) the capacity per subchannel has a finite limit; and (2) the RCE has a finite limit if the rate per subchannel is kept constant and the constant is sufficiently large. These results hold whether or not the state is known at the receiver.

Systematic coding and decoding, using BCH codes and minimum distance decoding rules, are considered for MSCC channels. Various coding alternatives are discussed, and formulas are given for computing or bounding performance.

14. KEY WORDS

information theory
crosstalk
space communications
coding
dependent channels
multiple channels