UNCLASSIFIED

AD NUMBER

AD241216

NEW LIMITATION CHANGE

TO
Approved for public release, distribution unlimited

FROM
Distribution authorized to U.S. Gov’t. agencies and their contractors; Administrative/Operational Use; APR 1960. Other requests shall be referred to Air Force Cambridge Research Center, Cambridge, MA.

AUTHORITY

AFCRL ltr, 3 Nov 1971

THIS PAGE IS UNCLASSIFIED
NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.
Topics in the Theory of Discrete Information Channels

RICHARD A. SILVERMAN and SZE-HOU CHANG

Contract No. AF 19(604)5238
APRIL, 1960
NEW YORK UNIVERSITY
Institute of Mathematical Sciences
Division of Electromagnetic Research

Research Report No. EM-152

TOPICS IN THE THEORY
OF DISCRETE INFORMATION CHANNELS

Richard A. Silverman and Sze-Hou Chang

The research reported in this document has been sponsored by the Electronics Research Directorate of the Air Force Cambridge Research Center, Air Research and Development Command, under Contracts Nos. AF 19(604)5238 and AF 19(604)3053.
Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to the:

ARMED SERVICES TECHNICAL INFORMATION AGENCY
DOCUMENTS SERVICE CENTER
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

Department of Defense contractors must be established for ASTIA services or have their 'need-to-know' certified by the cognizant military agency of their project or contract.

All other persons and organizations should apply to the:

U.S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D.C.
Abstract

The present article represents the authors' contribution to the URSI Information Theory monograph edited by J. Loeb, Vice Chairman of Commission 6. It discusses various topics in the theory of discrete information channels, including the general binary channel, channels with fading, cascaded channels, and channels with memory.

Table of Contents

1. Introduction 1
2. Information sources and information rate 1
3. Channels and mutual information rate 3
4. Channel capacity 9
5. The general binary channel 17
6. Channels with fading 22
7. Cascaded channels 31
8. Channels with memory 35

References 42
1. Introduction

Many interesting aspects of information theory can be illustrated by studying discrete channels with small input and output alphabets. In fact, much important work has been done on the simplest of all channels, the (memoryless) binary symmetric channel (BSC). We refer to Elias' study of the way in which rate and probability of error in the BSC depend on the length of the blocks in which the coding is done [1]. In the present article, we shall be concerned exclusively with asymptotic properties of channels, i.e. properties which are based on the assumption that coding is done in arbitrarily long blocks. In this study, we shall sometimes consider ternary and higher order channels as well as binary channels. In fact, the binary channel is too simple to be representative of the general discrete channel, as can be seen from the fact that neither input letter can be suppressed without destroying its information rate*.

2. Information sources and information rate

The first concept of interest in information theory is that of an information source, i.e. a device which generates a random sequence of letters from some alphabet; the random sequence then serves as the input to a channel (see Section 3). Let the input alphabet consist of the m letters $x_1, ..., x_m$. We shall confine our attention to independent sources, for which the probability $P_i$, $1 \leq i \leq m$, that the input letter $x_i$ is emitted at time t is statistically independent of which letters were emitted at times prior to t and of which

---

*Even in the most asymmetric binary channel one does not have to use an input symbol more often than 63 per cent (or less often than 37 per cent) of the time to achieve capacity (see Section 5).
letters will be emitted at times subsequent to \( t \). Thus, the probability of a sequence in which \( x_i \) appears \( N_i \) times, \( 1 \leq i \leq m \), is just

\[
(P_1)^{N_1}(P_2)^{N_2} \cdots (P_m)^{N_m}
\]

According to Shannon [2], it is particularly meaningful and fruitful in studying information sources to introduce the concept of the entropy or information associated with a source; for a source \( S \) of the type under consideration, this quantity is defined by

\[
H(S) = - \sum_{i=1}^{m} P_i \log P_i
\]

The base to which the logarithm is taken is conventionally chosen to be 2, in which case (1) is said to give the number of bits of information per source letter (see remarks at the end of this section); whenever we write \( \log \) we shall mean \( \log_2 \). Henceforth, it will be assumed that the source emits one letter per second (and that the channel accepts one letter per second). With this convention, (1) gives either the entropy of the source in bits/symbol or the information rate of the source in bits/second. This convention allows us to use the terms entropy (or information) and rate interchangeably; the adjustment needed in case the source emits one letter every \( T \) seconds is obvious.

We shall not linger on the derivation of (1) from a set of properties which it seems reasonable to expect information to have. Such derivations are given in detail by Shannon [2], Khinchin [3] and Faddeyev [4]. An important property of \( H(S) \) is that it vanishes if the source can emit only one of the
letters \( x_1, \ldots, x_m \) and takes its maximum value \( \log m \) when the source emits all of the letters \( x_1, \ldots, x_m \) with equal probability. (This is in accord with intuitive ideas of information.) Another important property of \( H(S) \) is the fact that although \( S \) can emit \( m^N \) possible sequences of length \( N \), one of a much smaller set of \( 2^{NH(S)} \) sequences is very likely to occur if \( N \) is very large. (An exceptional case occurs if all \( m \) source letters are equally likely; then \( H(S) = \log m \) and \( 2^{NH(S)} = m^N \).) This so-called asymptotic equipartition property is fundamental in information theory and can be demonstrated for much more general sources than those studied here (in fact for any stationary ergodic source; see McMillan [5], Khinchin [3]). Finally, we remark that the first of Shannon's two coding theorems (the so-called noiseless coding theorem) consists in showing that precisely \( H(S) \) binary digits (bits) per symbol are needed to noiselessly encode the output of \( S \) into binary digits; in general it takes code blocks of infinite length to effect this encoding. The noiseless coding theorem gives theoretical justification for measuring \( H(S) \) in bits/symbol (or bits/second).

3. Channels and mutual information rate

A discrete \( m \times n \) memoryless (information) channel is a probabilistic device which accepts any of \( m \) possible input (or "transmitted") letters \( x_1, \ldots, x_m \) and emits any of \( n \) possible output (or "received") letters \( y_1, \ldots, y_n \) in accordance with the following rules:

1) For every input letter \( x_i \) and output letter \( y_j \) there is a definite number \( p_{ij} \): \( 0 \leq p_{ij} \leq 1 \), which represents the probability that if \( x_i \) is transmitted, \( y_j \) is received.
2) Every input letter gives rise to at least one output letter, i.e.
\[ \sum_{j=1}^{n} p_{ij} = 1. \]

3) The response of the device to any input letter \( x_i \) is statistically independent of its response to any past or future letter. In Section 8 we shall abolish this requirement and consider a special kind of discrete channel with memory.

A convenient representation of such a discrete channel is as an \( m \times n \) channel matrix \( C \), i.e. as the \( m \times n \) rectangular array of numbers

\[
C = \begin{bmatrix}
P_{11} & P_{12} & \cdots & P_{1n} \\
P_{21} & P_{22} & \cdots & P_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
P_{m1} & P_{m2} & \cdots & P_{mn}
\end{bmatrix}
\]

(We shall use \( C \) to denote either a channel or the associated matrix.) By 1) and 2) each \( p_{ij} \) lies between 0 and 1, and the sum of every row of the matrix (2) is unity; these properties are summarized by calling \( C \) a stochastic matrix. The study of stochastic matrices has received a great deal of attention in the mathematical literature, especially in connection with the theory of Markov chains (see Feller [6]). We note that the case of a noiseless channel corresponds to the case where each row contains one 1 and \((m - 1)\) 0's, while each column contains one 1 and \((n - 1)\) 0's.

Suppose now that the input letter \( x_1, 1 \leq i \leq m \), is used with probability
$P_i$ and define $p(i,j)$, the joint probability that $x_i$ is emitted and $y_j$ is received, by the relation

$$p(i,j) = P_i P_{ij}$$

Then the probability that $y_j$, $1 \leq j \leq n$, is received, regardless of which $x_i$ is transmitted, is given by

$$P_j = \sum_{j=1}^{m} p(i,j)$$

It is convenient to represent both the numbers $P_i$, $1 \leq i \leq m$, and the numbers $P'_j$, $1 \leq j \leq n$, as the components of corresponding column vectors $P$ and $P'$, i.e.

$$P = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_m \end{bmatrix} \quad \text{and} \quad P' = \begin{bmatrix} P'_1 \\ P'_2 \\ \vdots \\ P'_n \end{bmatrix}$$

We now give a fundamental definition introduced by Shannon [2]: The average mutual information rate (or simply the rate) of the channel $C$ when it is driven by a random sequence of input letters chosen independently with probabilities given by the vector $P$ (equivalently, when it is driven by

*It can be shown that for a memoryless channel any dependence between input letters diminishes the rate of the channel (see Feinstein [7]).
an independent source characterized by $P$ is

$$R(P) = \sum_{i=1}^{m} \sum_{j=1}^{n} p(i,j) \log \frac{p(i,j)}{P_i P_j}$$

bits/second. (Recall that by our convention one input letter per second is emitted by the source and accepted by the channel.)

To justify (3), we specialize to the case where the input and output alphabets are the same (so that in particular $m = n$) and use Shannon's correction channel argument [2], which asserts that

$$\text{correction rate}^* = \text{mutual information rate} = \text{source rate}.$$  

The source rate is of course just

$$-\sum_{i=1}^{m} P_i \log P_i$$

The correction rate is derived as follows: Whenever the letter $y_j$ is received, it may be correct or incorrect. Since the probability that $y_j$ originated from $x_i$ is $p_j(i) = p(i,j)/P_j$, $1 \leq i, j \leq m$, whenever $y_j$ is received we must supply an amount of entropy

$$-\sum_{i=1}^{m} p_j(i) \log p_j(i)$$

to correct it or to leave it stand uncorrected with the assurance that it is

*More precisely, the average rate of correction entropy.
correct. Since the letter $y_j$ is received with probability $P'_j$, the average rate at which correction entropy must be supplied is

$$- \sum_{j=1}^{m} P'_j \sum_{i=1}^{m} p_j(i) \log p_j(i) = - \sum_{j=1}^{m} \sum_{i=1}^{n} p(i,j) \log \frac{p(i,j)}{P'_j}.$$

Since

$$\sum_{j=1}^{m} p_{i,j} = 1,$$

we can also write the source rate as

$$- \sum_{i=1}^{m} \sum_{j=1}^{m} P_i P_{i,j} \log P_i = - \sum_{i=1}^{m} \sum_{j=1}^{n} p(i,j) \log P_i.$$

Finally, using (4), we obtain

$$\text{mutual information rate} = - \sum_{i=1}^{m} \sum_{j=1}^{m} p(i,j) \log \frac{p(i,j)}{P_i} + \sum_{i=1}^{m} \sum_{j=1}^{n} p(i,j) \log \frac{P(i,j)}{P'_j}.$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} p(i,j) \log \frac{P(i,j)}{P_i P'_j},$$

which agrees with (3) for the case $m = n$.

In what follows we shall find it convenient to introduce the symbol $\langle \rangle$ (angular brackets) to denote averaging with respect to the joint probability distribution $p(i,j)$, i.e., if $f(i,j)$ is a function of the two
integral arguments $i$ and $j$, where $1 \leq i \leq m$, $1 \leq j \leq n$, then

$$\langle f(i,j) \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} p(i,j) f(i,j).$$

With this notation (3) becomes

$$\overrightarrow{R}(P) = - \langle \log P(i) \rangle + \langle \log p_j(i) \rangle = \langle \log \frac{p(i,j)}{P_i P_j} \rangle.$$

As an example of (3), consider the general binary channel

$$C(\alpha, \beta) = \begin{cases} \alpha & \text{if } i = 0, \\ (1 - \alpha) & \text{if } i = 1, \end{cases} \quad 0 \leq \alpha, \beta \leq 1,$$

fed by a source producing 0's and 1's independently, with probabilities $P_0$ and $P_1 = 1 - P_0$, respectively. (Eq. (5) means that transmitted 0's are received as 0's with probability $\alpha$, while transmitted 1's are received as 0's with probability $\beta$.) Then it is easily verified that the rate associated with (5) and the input vector

$$\overrightarrow{P} = \begin{bmatrix} P_0 \\ P_1 \end{bmatrix}$$

is

$$(6) \overrightarrow{R}(\alpha, \beta; \overrightarrow{P}) = \alpha P_0 \log \frac{\alpha}{P_0'} + (1-\alpha) P_0 \log \frac{(1-\alpha)}{P_1'} + \beta P_1 \log \frac{\beta}{P_1'} + (1-\beta) P_1 \log \frac{(1-\beta)}{P_1'},$$

where $P'_0 = \alpha P_0 + \beta (1 - P_0)$ is the probability of a received zero and $P'_1 = 1 - P'_0$ is the probability of a received 1.
4. **Channel capacity**

Following Shannon \[2\], we define the capacity \( c(C) \) of the (discrete memoryless) channel \( C \) as the largest value of the rate which is achieved when the input probabilities are varied over all possible values, i.e.

\[
\text{Max}_{P} \rightarrow R(P)
\]

(7)

As an example consider the BSC where the channel matrix

\[
C_{\text{sym}} = \begin{pmatrix}
\alpha & \beta \\
\beta & \alpha
\end{pmatrix}
\]

is obtained by setting \( \beta = 1 - \alpha \) in (5). In this case it is clear from symmetry (and it can be verified by direct calculation) that capacity is achieved for the choice \( P_0 = P_1 = 1/2 \), which implies \( P'_0 = P'_1 = 1/2 \) as well. Substituting these values in (6) and using \( \beta = 1 - \alpha \), we find

\[
c(C_{\text{sym}}) = 1 + \alpha \log \alpha + \beta \log \beta
\]

for the capacity of the BSC.

The fundamental importance of the channel capacity as an information-theoretic quantity stems from the role that it plays in Shannon's second coding theorem (the so-called noisy coding theorem), which it is safe to say contains most of the substantive content and technical promise of information theory (taken together with corresponding studies of finite-length block coding like \([1]\)). This theorem asserts that with proper encoding it is possible to transmit information at any rate less than capacity with arbitrarily small
probability of error, provided that the block length of the code is long enough,
and furthermore that regardless of the encoding scheme, errors will always be made
if one attempts to transmit information at a rate greater than capacity. Much
space has been devoted in the literature to a rigorous demonstration of this
theorem for an appropriately large class of sources and channels, and the whole
subject is a difficult one which we shall not go into here. The interested
reader is referred to the papers of Khinchin [3] and the book by Feinstein [7];
the latter author played an important role in developing a rigorous proof of
the noisy coding theorem.

We turn now to the question of how the mathematical operation symbolized
by (7) is to be carried out in general. This operation involves more than a
simple maximization problem, since the vector $P$ in question is subject to the
constraint that it be a vector with non-negative components which add up to
unity. We begin by writing (3) in a form which explicitly exhibits its
dependence on the input probabilities $P_i$; this form is

$$ R(P) = -\langle \log P_j \rangle + \langle \log P_{ij} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} P_{ij} \log \frac{m}{P_{ij}} + \sum_{i=1}^{m} \sum_{j=1}^{n} P_{ij} \log P_{ij}. $$

To incorporate the constraint*

$$ \sum_{i=1}^{m} P_i = 1, $$

we add

$$ \lambda \sum_{i=1}^{m} P_i $$

*For the time being we neglect the additional constraint that $P_i \geq 0$, $1 \leq i \leq m$, which will be discussed below in connection with Muroga's work.
to (8), where \( \lambda \) is an undetermined Lagrange multiplier. Differentiating the sum of these two terms with respect to \( P_i, 1 \leq i \leq m \), and equating the result to zero, we obtain

\[
\sum_{j=1}^{n} P_{ij} \log \frac{P_{ij}}{\sum_{i=1}^{m} P_{ipij}} = \mu, \quad 1 \leq i \leq m,
\]

where \( \mu = \lambda - (1/\log e) \) is a new constant. Multiplying (9) by \( P_i \) and summing over \( i \), we find that

\[
\mu = c,
\]

where \( c \) is the capacity of the channel \( C \).

Suppose now that the channel \( C \) under consideration, with channel matrix (2), is square \( (m = n) \) and that \( \det(C) \), the determinant of (2), is non-vanishing. Then the inverse matrix \( C^{-1} \), satisfying the matrix equation

\[
CC^{-1} = C^{-1}C = I
\]

(where \( I \) denotes the unit matrix) exists. If we denote the elements of \( C^{-1} \) by \( P_{ij}^{-1}, 1 \leq i, j \leq m \), then (10) reads

\[
\sum_{j=1}^{m} P_{ij} P_{jk} = \sum_{j=1}^{m} P_{ij} P_{jk} - \delta_{ik}, \quad 1 \leq i, k \leq m,
\]

where \( \delta_{ik} \) is the Kronecker delta symbol, equal to 1 when \( i = k \) and 0 when \( i \neq k \). Multiplying (9) by \( P_{ki}^{-1} \) and summing over \( i \), we obtain

\[
\sum_{i=1}^{m} \sum_{j=1}^{m} P_{ki}^{-1} P_{ij} \log P_{ij} - \log \sum_{i=1}^{m} P_{ipik} = \mu \sum_{i=1}^{m} P_{ki}^{-1}, \quad 1 \leq k \leq m.
\]
Hence
\[
\sum_{k=1}^{m} P_{jk} = \exp \left\{ -c \sum_{k=1}^{m} P_{kk} + \sum_{i=1}^{m} \sum_{j=1}^{m} P_{kj} \log p_{ij} \right\}, \quad 1 \leq k \leq m,
\]
where by \( \exp(x) \) we mean \( 2^x \), the base appropriate to our choice of units.

Multiplying by \( P_{kj}^{-1} \) and summing over \( k \), we obtain
\[
(11) \quad P_i = \sum_{k=1}^{m} P_{ki} \exp \left\{ -c \sum_{k=1}^{m} P_{kk} + \sum_{i=1}^{m} \sum_{j=1}^{m} P_{kj} \log p_{ij} \right\}, \quad 1 \leq i \leq m.
\]

Finally, following Muroga [8], we note that
\[
(12) \quad \sum_{j=1}^{m} P_{ij} = \frac{1}{2} \sum_{j=1}^{m} P_{ij} \sum_{k=1}^{m} P_{kj} = \sum_{k=1}^{m} \delta_{ik} = 1,
\]
so that (11) can be simplified to
\[
(13) \quad P_i = \sum_{k=1}^{m} P_{ki} \exp \left\{ -c + \sum_{i=1}^{m} \sum_{j=1}^{m} P_{kj} \log p_{ij} \right\}, \quad 1 \leq i \leq m.
\]

Thus, the channel capacity is the number \( c \) which when substituted in (13) gives \( P_i > 0 \) and
\[
\sum_{i=1}^{m} P_i = 1.
\]

Then with this value of \( c \), (13) gives the corresponding rate-maximizing values of the input probabilities \( P_i, 1 \leq i \leq m \). Explicitly, we form the sum
\[
\sum_{\ell=1}^{m} P_{\ell} = 1
\]
and use the relation (12) again, obtaining
\[
1 = \sum_{k=1}^{m} \sum_{i=1}^{m} p_{ki} \exp \left\{ -c + \sum_{i=1}^{m} \sum_{j=1}^{m} p_{ki} p_{ij} \log p_{ij} \right\} = \\
= \sum_{k=1}^{m} \exp \left\{ -c + \sum_{i=1}^{m} \sum_{j=1}^{m} p_{ki} p_{ij} \log p_{ij} \right\}.
\]

Solving for \( c \), we get

\[
(14) \quad c = \log \left[ \sum_{k=1}^{m} \exp \left\{ \sum_{i=1}^{m} \sum_{j=1}^{m} p_{ki} p_{ij} \log p_{ij} \right\} \right],
\]

where it will be recalled that both \( \log \) and \( \exp \) are taken to the base 2.

Computations are simplified by defining (after Muroga [6]) the auxiliary vector

\[
\begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_m
\end{bmatrix} = \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_m
\end{bmatrix}
\]

which satisfies the matrix equation

\[
(15) \quad C \ X = -H,
\]

where \( H \) is the row-entropy vector of the channel \( C \), i.e., the vector

\[
(16) \quad H = \begin{bmatrix}
-\sum_{i=1}^{m} p_{1i} \log p_{1i} \\
-\sum_{i=1}^{m} p_{2i} \log p_{2i} \\
\vdots \\
-\sum_{i=1}^{m} p_{mi} \log p_{mi}
\end{bmatrix}.
\]
It follows from (15) and (16) that $X = \mathbf{C}^{-1} \mathbf{H}$, i.e. that

$$X_i = \sum_{j=1}^{m} p_{ij}^{-1} \sum_{k=1}^{m} p_{jk} \log p_{jk}, \quad 1 \leq i \leq m,$$

so that in terms of the components of $X$, (14) becomes simply

$$c = \log \sum_{i=1}^{m} \exp X_i.$$

Moreover (13) becomes

$$P_i = \exp(-c) \sum_{k=1}^{m} p_{ki}^{-1} \exp X_k, \quad 1 \leq i \leq m.$$

Since

$$p_{ki}^{-1} = \text{cof}(p_{ik}) / \det(C),$$

where $\text{cof}(p_{ik})$ is the cofactor of the element $p_{ik}$, we finally have

$$P_i = \frac{\exp(-c)}{\det(C)} \text{det} \begin{vmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ P_{11} & \cdots & \cdots & P_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ P_{m1} & \cdots & \cdots & P_{mm} \end{vmatrix},$$

where the matrix in (18) differs from the channel matrix $C$ by having the entries $\exp(X_1), \ldots, \exp(X_m)$ instead of $p_{11}, \ldots, p_{1m}$ in the $i$'th row.

The method just described requires modification if it leads to negative input probabilities, and more generally if $\det(C) = 0$ or the sizes of the input
and output alphabets are different. Details of how to deal with these various cases are given in Muroga's paper [8], to which the interested reader is referred. There is also available a different and more easily visualized approach to the general problem of capacity due to Shannon [9].

We conclude our present discussion by discussing a one-parameter family of ternary channels which for suitable values of the parameter leads to a negative probability for one of the input symbols, which must therefore then be suppressed according to Muroga. Consider the channel

\[
C = \begin{bmatrix}
\alpha & 1-\alpha & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} \\
0 & \frac{1}{2} & \alpha
\end{bmatrix}
\]

where \(0 < \alpha < 1\). The inverse matrix is

\[
C^{-1} = \begin{bmatrix}
\frac{1}{2\alpha} & 1 & -\frac{1}{2\alpha} \\
\frac{1}{2(1-\alpha)} & \frac{\alpha}{\alpha-1} & \frac{1}{2(1-\alpha)} \\
-\frac{1}{2\alpha} & 1 & \frac{1}{2\alpha}
\end{bmatrix}
\]

The row-entropy vector is

\[
\rightarrow H = \begin{bmatrix}
H \\
1 \\
H
\end{bmatrix}
\]
and the vector $\mathbf{x}$ is

$$
\mathbf{x} = \begin{bmatrix}
-l \\
\frac{-H-a}{1-a} \\
-l 
\end{bmatrix},
$$

where

$$
H = -a \log a - (1-a) \log (1-a).
$$

Using (17) and (18) we find that the symbol corresponding to the middle row of $C$ should be used with the probability

$$
p = \frac{1 + \frac{a}{a-1} \exp \left( -\frac{H-a}{1-a} \right)}{1 + \exp \left( -\frac{H-a}{1-a} \right)}
$$

if capacity is to be achieved. This symbol should be suppressed when the numerator goes negative. This happens when $a > a_o$, where $a_o$ is the solution of

$$
\frac{H(a_o) - a_o}{1 - a_o} = \log \frac{a_o}{1-a_o}.
$$

or

$$
(19) \quad \log a_o = -a_o.
$$

The solution of the transcendental equation (19) is $a_o \approx 0.641$. When $p > 0$, the capacity of the channel is

$$
\log \left( 1 + \exp \left( -\frac{H-a}{1-a} \right) \right) \text{ bits/second},
$$
whereas when \( p \leq 0 \), the channel matrix reduces to

\[
\begin{bmatrix}
\alpha & 1-\alpha & 0 \\
0 & 1-\alpha & \alpha
\end{bmatrix}
\]

This is the matrix of the binary erasure channel (BEC), i.e. the channel in which \( 0 \)'s and \( 1 \)'s are transmitted and \( 0 \)'s, \( 1 \)'s and \( x \)'s are received; \( \alpha \) is the probability that a \( 0 \) or \( 1 \) is received correctly and \( 1-\alpha \) the probability that a \( 0 \) or \( 1 \) is received as an \( x \). Since received \( 0 \)'s and \( 1 \)'s are always correct, whereas a received \( x \) is always wrong and equally likely to have come from a \( 0 \) or \( 1 \), the capacity of the BEC is obviously \( \alpha \) bits/second. It is interesting to note that the capacity of \( C \) is 1 bit/second both when \( \alpha = 0 \) and \( \alpha = 1 \), but when \( \alpha = 0 \) the symbol corresponding to the middle row should be sent, whereas when \( \alpha = 1 \) it should be suppressed.

5. The general binary channel

Using Muroga's method, Silverman [10] has made a detailed study of the general binary channel \( C(\alpha, \beta) \) defined by (5)*. We sketch without proof some of the results obtained in his paper:

1) The capacity \( c(\alpha, \beta) \) of the general binary channel is given by the formula

\[
c(\alpha, \beta) = \frac{-\beta H(\alpha) + \alpha H(\beta)}{\beta - \alpha} + \log \left[ 1 + \exp \left( \frac{H(\alpha) - H(\beta)}{\beta - \alpha} \right) \right].
\]

The function \( c(\alpha, \beta) \) has the symmetries

\[
c(\alpha, \beta) = c(\beta, \alpha) = c(1 - \alpha, 1 - \beta) = c(1 - \beta, 1 - \alpha),
\]

*Loeb [11] has also studied some aspects of the general binary channel.
and defines a surface over the unit square \(0 \leq \alpha, \beta \leq 1\). Lines of constant \(c(\alpha, \beta)\) are shown in Fig. 1.

2) Capacity is achieved if 0's are transmitted with probability

\[
P_0(\alpha, \beta) = \beta(\beta - \alpha)^{-1} - (\beta - \alpha)^{-1} \left[ 1 + \exp \left( \frac{H(\beta) - H(\alpha)}{\beta - \alpha} \right) \right]^{-1}.
\]

\(P_0(\alpha, \beta)\) satisfies the relation

\[0.37 \leq \frac{1}{6} \leq P_0(\alpha, \beta) \leq 1 - \frac{1}{6} \approx 0.63,\]

which explains the footnote in Section 1. The function \(P_0(\alpha, \beta)\) is discontinuous at the points \(\alpha = \beta = 0\) and \(\alpha = \beta = 1\); it has the symmetries

\[P_0(\alpha, \beta) = P_0(1-\alpha, 1-\beta) = 1 - P_0(\beta, \alpha) = 1 - P_0(1-\beta, 1-\alpha).
\]

Lines of constant \(P_0(\alpha, \beta)\) are shown in Fig. 2.

3) At capacity, 0's are received with probability

\[
P'_0(\alpha, \beta) = \left[ 1 + \exp \left( \frac{H(\beta) - H(\alpha)}{\beta - \alpha} \right) \right]^{-1}.
\]

The function \(P'_0(\alpha, \beta)\) has the symmetries

\[P'_0(\alpha, \beta) = P'_0(\beta, \alpha) = 1 - P'_0(1-\alpha, 1-\beta) = 1 - P'_0(1-\beta, 1-\alpha).
\]

Lines of constant \(P'_0(\alpha, \beta)\) are shown in Fig. 3.

The reader interested in other properties of the general binary channel, e.g., the form of the channels giving the maximally asymmetric input probability distributions \((P_0 \sim 1/e\) or \(P_0 \sim 1 - 1/e\)), the probability of error for the
Fig. 1 - Lines of constant $c(\alpha, \beta)$. 
Fig. 2 - Lines of constant $P_0(\alpha, \beta)$. 
Fig. 3 - Lines of constant $P_0'(\alpha, \beta)$.
general binary channel, etc., is referred to Silverman's paper [10].

6. **Channels with fading**

The communication situation dealt with so far can be indicated schematically by the diagram

\[ T \rightarrow C \rightarrow R \]

i.e., information is sent from a transmitter T to a receiver R through a channel C, characterized by a stochastic matrix (see Section 3). (We make no distinction between the primary information source S of Section 2 and the transmitter T, although in general there is the problem of (noiselessly) encoding the output of S so as to match the input of T.) We now consider the following generalization of this situation: Instead of one (memoryless) channel C, let there be a family of (memoryless) channels \( C_a \) where the index \( a \) ranges from 1 to \( s \), and let the channel which is actually present at a given time of transmission depend on the state of a random device N (N for "nature"). Schematically, we have

\[ N \]
\[ T \rightarrow C \rightarrow R \]

i.e., information is sent from the transmitter T to the receiver R through the channel C, the state of which depends on the state of nature N. We assume that the state \( a, 1 \leq a \leq s \), chosen by nature is statistically independent of past and future states of N and of the transmitted sequences as well; let \( p_a \) be the probability that nature chooses the state \( a \). The channel C is now represented by the family of stochastic matrices
Since the transition probabilities $p_{ij}(a)$ are now random variables, a new element of randomness has entered into the problem. Such a model might be used to give an abstract representation of communication in the presence of "fading", and with this in mind we refer to the totality of channels $C_a$, $1 \leq a \leq s$, as a channel with fading.

A natural problem in the theory of channels with fading is that of finding the channel capacity for various conditions of knowledge of $N$ at the transmitter and receiver. There are four possible cases, which can be schematically represented as follows:

Case 1. $T \rightarrow C \rightarrow R$

Case 2. $T \rightarrow C \rightarrow R$

Case 3. $T \rightarrow C \rightarrow R$

Case 4. $T \rightarrow C \rightarrow R$
Case 1, which we have already encountered, represents the situation in which neither the transmitter nor the receiver knows nature's state, so that $N$ is effectively just more noise in addition to that already included in $C$. Case 2 is the situation in which the receiver but not the transmitter knows nature's state. Case 3 is the situation in which both the transmitter and the receiver know nature's state, and Case 4 is the situation in which the transmitter but not the receiver knows nature's state. We now give expressions for the capacity in all four cases.

Let $p(a,i,j)$ be the joint probability that nature chooses the state $a$, that $x_i$ is transmitted and that $y_j$ is received. In cases 1 and 2, where the transmitter is ignorant of nature's state, we have

$$p(a,i,j) = p_a P_i P_j(a)$$

In case 3, where the transmitter's action can depend on nature's state, we have

$$p(a,i,j) = p_a P_{i,a} P_j(a)$$

where $P_{i,a}$ is the probability of choosing $x_i$, given that nature's state is $a$. (Case 4 requires special treatment; see below.) Generalizing the angular bracket notation of Section 3, we write

$$\langle f(a,i,j) \rangle = \sum_{a=1}^A \sum_{i=1}^m \sum_{j=1}^n p(a,i,j) f(a,i,j)$$

We now derive expressions for the channel capacity in the four different cases.

**Case 1.** Since neither the transmitter nor the receiver knows nature's,
state, the channel has the same capacity as the average channel
\[
\bar{C} = \sum_{a=1}^{s} P_a C_a
\]
with transition probabilities
\[
\bar{P}_{ij} = \sum_{a=1}^{s} P_a P_{ij}(a)
\]
Explicitly we have, from (8)
\[
c_1 = \max_{p} \left\langle \log \sum_{i=1}^{m} P_i \bar{P}_{ij} \right\rangle
\]
for the capacity in this case.

Case 2. Since now the receiver can use its knowledge of the pair \((a,y_j)\) to infer \(x_i\), we have, from (3)
\[
c_2 = \max_{p} \left\langle \log \frac{p(a,i,j)}{P_i p(a,j)} \right\rangle
\]
for the capacity in this case, where \(p(a,j)\) is the probability of the pair \((a,y_j)\), i.e.
\[
p(a,j) = \sum_{i=1}^{m} p(a,i,j)
\]
Using (20) we can rewrite (22) as
Let $R_\alpha(P)$ be the rate in the channel $C_\alpha$ when using the input vector $P$. Then

\[ c_2 = \max_P \left\langle \log \frac{p_{ij}(\alpha)}{\sum_{\alpha=1}^m P_\alpha p_{ij}(\alpha)} \right\rangle. \]

(23)

(23) is just

\[ c_2 = \max_P \sum_{\alpha=1}^s p_\alpha R_\alpha(P). \]

(24)

(24)

In other words, the capacity in case 2 is the maximum average rate when driving all the channels $C_\alpha$ with the same input probability vector $P$. Clearly, $c_2 \geq c_1$, since we are now using more detailed knowledge of the channel.

Case 3. Now the transmitter can base its choice of input probabilities on nature's state, choosing a suitable input vector

\[
P_\alpha = \begin{bmatrix}
p_{1,\alpha} \\
p_{2,\alpha} \\
\vdots \\
p_{m,\alpha}
\end{bmatrix}
\]

for each state, $1 \leq \alpha \leq s$. Moreover, as in case 2, the receiver has the pair $(x_i, y_j)$ available to infer $x_i$. Therefore we have

\[ c_3 = \max_{P_\alpha} \left\langle \log \frac{p(\alpha, i, j)}{\sum_j p_{\alpha}(\alpha, j) p(\alpha, i, j)} \right\rangle. \]

(25)
for the capacity in this case, where the maximization is over all $P_\alpha$, $1 \leq \alpha \leq s$. Using (21), we can rewrite (25) as

$$c_3 = \max_{P_\alpha} \sum_{a=1}^{s} p_\alpha r(P_\alpha).$$

In other words, the capacity in case 3 is the maximum average rate when driving each channel $C_\alpha$ with its own input probability vector $P_\alpha$. It follows at once that

$$c_3 = \sum_{a=1}^{s} p_\alpha c(C_\alpha),$$

i.e., the capacity is just the average capacity of the channels $C_\alpha$, $1 \leq \alpha \leq s$. Clearly $c_3 > c_2$, since now the transmitter as well as the receiver is using the extra information about the channel.

Case 4 is more complicated. In this case Shannon [12] has shown that the capacity of the channel with fading is the same as the capacity of a new $m^3 \times n$ (non-fading) channel $C'$, with input letters consisting of the $m^3$ $s$-tuples $(x_1, x_2, \ldots, x_s)$, $1 \leq i_1, i_2, \ldots, i_s \leq m$ and output letters consisting of the $n$ letters $y_j$, $1 \leq j \leq n$, where the transition probabilities for $C'$ are defined by

$$p_{i_1, i_2, \ldots, i_s} \rightarrow y_j = \sum_{a=1}^{s} p_\alpha p_{i_\alpha j}(s).$$

Thus the capacity in this case is

$$c_4 = c(C').$$
and clearly \( c_1 \leq c_4 \leq c_3 \). We now give two examples which illustrate the various cases.

**Example 1.** Suppose nature has two states 1 and 2 with probabilities \( \alpha \) and \( \beta \), respectively, and suppose

\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\quad \text{and}
\quad
\begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}
\]

Then the average channel \( \overline{C} \) is the BSC, i.e.

\[
\overline{C} = \alpha C_1 + \beta C_2 = \begin{pmatrix}
\alpha & \beta \\
\beta & \alpha
\end{pmatrix}
\]

so that \( c_1 = 1 + \alpha \log \alpha + \beta \log \beta \) bits/second (see Section 4). If nature's state is known by the receiver or by both the transmitter and the receiver, then the capacity of the fading channel is obviously 1 bit/second, achieved by transmitting 0's and 1's with equal probability and interchanging 0's and 1's at the receiver when the channel is in state 2. Eqs. (24) and (26) confirm that \( c_2 = c_3 = 1 \). When only the transmitter knows nature's state, the capacity is again 1 bit/second, achieved by transmitting 0's and 1's with equal probability and interchanging 0's and 1's at the transmitter when the channel is in state 2. In this instance (case 4) Shannon's construction asserts that \( c_4 = c(C') \), where

\[
C' = \begin{pmatrix}
\alpha & \beta \\
1 & 0 \\
0 & 1 \\
\beta & \alpha
\end{pmatrix}
\]
Denoting the rows of this matrix by the vectors $A_1, \ldots, A_4$ (e.g. $A_1 = (a, \beta)$), we have

$$
\rightarrow \rightarrow \rightarrow \\
A_1 = \alpha A_2 + \beta A_3 \\
\rightarrow \rightarrow \rightarrow \\
A_4 = \beta A_2 + \alpha A_3
$$

It follows by an argument due to Shannon [9] that the symbols corresponding to the first and fourth rows of $C'$ have to be suppressed if capacity is to be achieved. Dropping the first and fourth rows of $C'$, we get the matrix $C_1$, so that $c_4 = c(C_1) = 1$ bit/second, as required.

**Example 2.** Let there be three states with equal probability, and let

$$
C_1 = \begin{bmatrix}
1 & 0 \\
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}
\end{bmatrix}, \quad C_2 = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} \\
0 & 1
\end{bmatrix}, \quad C_3 = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}
\end{bmatrix}
$$

Then if neither the transmitter nor the receiver knows nature's state, the capacity of the fading channel is that of the average channel

$$
\overline{C} = \begin{bmatrix}
\frac{2}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{2}{3} \\
\frac{1}{3} & \frac{2}{3}
\end{bmatrix}
$$

i.e. $c_1 = c(\overline{C}) = 1 + (1/3) \log (1/3) + (2/3) \log (2/3) = (5/3) - \log 3 \sim .082$ bits/second. If only the receiver knows nature's state, then by (24)

$$
c_2 = \frac{\text{Max}}{p} \frac{1}{3} \left[ R_1(P) + R_2(P) + R_3(P) \right] 
$$
By symmetry, the maximum is achieved for

$$P = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix},$$

and an elementary calculation shows that $c_2 = 1 - (1/2) \log 3 \approx 0.208$ bits/second $> c_1$. If both the transmitter and the receiver know nature's state, then by (26)

$$c_3 = \frac{2}{3} c(C_1),$$

since $c(C_1) = c(C_2)$ and $c(C_2) = 0$. Applying Muroga's method (see Section 4), we easily find that $c(C_1) = \log 5 - 2$, so that $c_3 = 2/3 \log 5 - 4/3 \approx 0.215$ bits/second $> c_2$.

If the transmitter but not the receiver knows nature's state, then according to Shannon's construction, $c_4 = c(C')$, where

$$C' = \begin{bmatrix} 0 \\ 1 \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}.$$
By a simple extension of the argument given in connection with the preceding example, we have

\[ c_4 = c(C') = c(C) = c_1 \sim 0.82 \text{ bits/second} \]

so that in this case, unlike example 1, the transmitter cannot use its knowledge of nature's state to increase the channel capacity.

7. Cascaded channels

Let C be an \( l \times m \) channel matrix with output alphabet \( y_1, \ldots, y_m \), and let \( C' \) be an \( m \times n \) channel matrix with input alphabet \( y_1, \ldots, y_m \) (i.e. identical to the output alphabet of C). Then we agree to apply \( y_i \) to the input of \( C' \) whenever \( y_i \) is received at the output of C; this mode of channel combination is called cascading. Denote by \( p_{ij} \) and \( p'_{ij} \) the elements of the matrices \( C \) and \( C' \), and denote by \( p''_{ij} \) the elements of the matrix \( C'' \) obtained by cascading \( C \) and \( C' \), as just described. Then, since obviously

\[ p''_{ij} = \sum_{k=1}^{m} p_{ik} p'_{kj} \]

\( C'' \) is obtained from \( C \) and \( C' \) by matrix multiplication, i.e. \( C'' = CC' \). Since matrix multiplication is in general non-commutative, the same is true of channel cascading.

The problem of cascading identical binary channels \( C(\alpha, \beta) = \begin{vmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{vmatrix} \) is especially simple, in view of the identity

\[
C^n(\alpha, \beta) = \begin{vmatrix} \alpha(n) & 1-\alpha(n) \\ \beta(n) & 1-\beta(n) \end{vmatrix} = \frac{1}{1-\alpha+\beta} \left\{ \begin{vmatrix} \beta & 1-\alpha \\ \beta & 1-\alpha \end{vmatrix} - (\alpha-\beta)^n \right\} \begin{vmatrix} \alpha-1 & 1-\alpha \\ \beta & -\beta \end{vmatrix},
\]

\[
C^n(\alpha, \beta) = \begin{vmatrix} \alpha(n) & 1-\alpha(n) \\ \beta(n) & 1-\beta(n) \end{vmatrix} = \frac{1}{1-\alpha+\beta} \left\{ \begin{vmatrix} \beta & 1-\alpha \\ \beta & 1-\alpha \end{vmatrix} - (\alpha-\beta)^n \right\} \begin{vmatrix} \alpha-1 & 1-\alpha \\ \beta & -\beta \end{vmatrix},
\]
which is easily derived from a general formula of matrix algebra (see e.g. [13]).

(We exclude the trivial cases $\alpha = 1$, $\beta = 0$ and $\alpha = 0$, $\beta = 1$, corresponding to the noiseless channels $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.) A number of simple consequences can be derived from (27):

1) Since $|\alpha - \beta| < 1$, we have

$$\lim_{n \to \infty} C_n(\alpha, \beta) = C(\infty)(\alpha, \beta) = \frac{1}{1 - \alpha + \beta} \begin{bmatrix} \beta & 1 - \alpha \\ 1 - \alpha & \beta \end{bmatrix} = \begin{bmatrix} \alpha(\infty) & 1 - \alpha(\infty) \\ 1 - \alpha(\infty) & \beta(\infty) \end{bmatrix}.$$

The limiting channel $C(\infty)(\alpha, \beta)$ obviously has zero capacity, in accord with the intuitive idea that an infinite cascade of noisy channels must destroy any information fed into it.

2) Since

$$\frac{-\beta(\infty)}{1 - \alpha(\infty)} = \frac{\beta(n) - \beta(\infty)}{\alpha(n) - \alpha(\infty)} = \frac{\beta}{\alpha - 1},$$

all the channels $C_n(\alpha, \beta)$ lie on a straight line in the $(\alpha, \beta)$ square passing through the point corresponding to the limiting channel $C(\infty)(\alpha, \beta)$ and the point $(0,1)$. Thus, referring to Fig. 1, we have the following simple interpretation of the operation of cascading a binary channel with itself: Using (28), draw the straight line just described; then as $n$ increases, $C_n(\alpha, \beta)$ approaches $C(\infty)(\alpha, \beta)$ along this straight line, moving alternately from one side of the zero-capacity line $\alpha = \beta$ to the other if $\alpha - \beta < 0$.

The operation of channel cascading, or equivalently of multiplying channel matrices, is a partial ordering in the following sense: Given any two channels $C_1$ and $C_2$, then $C_1$ is said to include $C_2$, written $C_1 \supseteq C_2$, if there
exists a channel $C$ such that $C_1 C = C_2$, i.e. if $C_2$ can be obtained from $C_1$ by cascading. Channel inclusion has the defining properties of a partial ordering, namely

1) $C \succeq C$, for any $C$;
2) $C_1 \succeq C_2$, $C_2 \succeq C_3$ implies $C_1 \succeq C_3$.

On the other hand, given arbitrary channels $C_1$ and $C_2$, neither of the relations $C_1 \succeq C_2$, $C_2 \succeq C_1$ may hold, i.e. $C_1$ and $C_2$ may not be comparable.

In the case of binary channels, the structure of the partial ordering is very simply displayed. Suppose we have two binary channels

$$C = \begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix}, \quad C' = \begin{bmatrix} \alpha' & 1-\alpha' \\ \beta' & 1-\beta' \end{bmatrix},$$

and assume, as we can without loss of generality, that $\beta \leq \alpha$, $\beta' \leq \alpha'$. Then, following a procedure suggested by Birnbaum (private communication), we first prove the following lemma:

A necessary and sufficient condition for $C \succeq C'$ is that the interval $(L_1, L_2)$ contain the interval $(L_1', L_2')$, where

$$L_1 = \beta/\alpha, \quad L_1' = \beta'/\alpha', \quad L_2 = (1-\beta)/(1-\alpha), \quad L_2' = (1-\beta')/(1-\alpha').$$

(Note that in our case $0 \leq L_1 \leq L_2 \leq \infty$, $0 \leq L_1' \leq L_2' \leq \infty$.)

To show the necessity, we suppose that

$$C'' = \begin{bmatrix} x & 1-x \\ y & 1-y \end{bmatrix}$$

If $c(C)$ denotes the capacity of $C$, then $c(C) \geq c(C')$ is a necessary but not sufficient condition for $C \succeq C'$. 
is the channel such that $C'' = C'$. Then, doing the matrix multiplication explicitly, we find that

$$L_1' = \frac{\beta' / a'}{\alpha x + (1 - \alpha) y} \geq \frac{\beta / a}{L_1},$$

since $y \geq 0$, and

$$L_2' = \frac{(1 - \beta') / (1 - a')}{1 - \alpha x - (1 - \alpha) y} \leq \frac{(1 - \beta) / (1 - a)}{L_2},$$

since $x \leq 1$.

To show the sufficiency, we solve formally for the parameters $x, y$ of $C''$, finding

$$y = \frac{\alpha \beta' - a' \beta}{\alpha - \beta},$$

$$x = \frac{a' - \beta'}{\alpha - \beta} + y.$$

Since $\alpha' - \beta' \geq 0$, $\alpha - \beta \geq 0$, and $\alpha \beta' - a' \beta \geq 0$ by the hypothesis that $L_1 \leq L_1'$, we see that $0 \leq y \leq x$. Moreover $y \leq x \leq 1$, since $\alpha \beta' - a' \beta + a' - \beta \leq \alpha - \beta$ is an easy consequence of the hypothesis $L_2' \leq L_2$. This completes the proof that $C \supseteq C'$ and $(L_1', L_2') \supseteq (L_1, L_2)$ are equivalent statements.

Using the lemma, we can give a simple geometrical model of the partial ordering of binary channels under cascading. Plot the point $(a, \beta)$ corresponding to $C$ in the unit square, and draw the straight lines from the point $(a, \beta)$ to the points $(0,0)$ and $(1,1)$, as shown in Fig. 1. Then all the points in the region shaded with vertical lines represent channels contained in $C$, i.e. channels which can be obtained from $C$ by cascading. This follows by applying the lemma in the part of the region lying below the line $\alpha = \beta$, and then noting that multiplication by $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ changes a channel into its reflection in the line $\alpha = \beta$. Moreover, it follows from this construction that all points
in the region shaded by horizontal lines represent channels containing C. Similarly, points lying outside of the shaded regions represent channels which are not comparable with C, i.e. which cannot be reached from C either by premultiplication or postmultiplication by any binary channel.

Fig. 4 - Illustrating the partial ordering of binary channels.

For further discussion of topics related to channel cascading, see Shannon [14], Birnbaum [15] and Silverman [10].

8. Channels with memory

We now consider a simple idealized communication system suggested by Chang [16], which leads naturally to the study of channels with memory. Let the channel input be a source emitting binary digits, and let the channel be such that 0's and 1's emitted at the time \( t = 0 \) are represented by the waveforms

\[
f_0(t) = A \exp\left(-\frac{t}{\tau}\right),
\]

\[
f_1(t) = -A \exp\left(-\frac{t}{\tau}\right),
\]
respectively, in the absence of noise*. The effect of noise will be represented by the addition of white Gaussian noise** of r.m.s. voltage $\sigma'$. 

Suppose, to consider the simplest non-trivial case, that the source produces the digits in isolated doublets, i.e., let a pair of binary digits separated by an amount $\gamma \leq \tau$ be transmitted, and then let there be a pause in transmission long enough to allow substantially complete decay of the exponential $\exp(-t/\tau)$. Finally, suppose that we use synchronous, threshold detection at the receiver, i.e., at the times $t_0$ and $t_0 + \gamma$ corresponding to transmission of a doublet, we interpret a positive voltage as a 0 and a negative voltage as a 1.

We now proceed to find the capacity of this simple channel. At a time $t_0$ such that the channel has recovered from the effects of previous transmitted signals the channel is described by the matrix

$$
C = \begin{pmatrix}
\alpha & 1-\alpha \\
\beta & 1-\beta
\end{pmatrix},
$$

where

$$
\alpha = \text{Prob} \left[ f_0(0) + \xi > 0 \right],
$$

$$
\beta = \text{Prob} \left[ f_1(0) + \xi > 0 \right],
$$

and $\xi$ is a Gaussian random variable with mean 0 and variance $\sigma^2$. (Specifically,

---

*We have in mind, for example, a situation where 0's and 1's are encoded into sharp positive and negative pulses of amplitude $A$ and then sent through a channel whose transmission characteristics resemble those of an RC filter with time constant $\tau$.

**By "white" noise, we mean noise with a constant power spectral density, or equivalently with a correlation function which is a delta function. Gaussian white noise has no memory. This is, of course, a limiting case, approached only when the noise bandwidth is much greater than that of the signals.
α is the probability that a 0 transmitted at time \( t_0 \) is interpreted as a 0, and \( β \) is the probability that a 1 transmitted at time \( t_0 \) is interpreted as a 0. Similarly, \( α_o, β_o \) and \( α_1, β_1 \) are the corresponding probabilities at time \( t_0 + γ \), under the hypothesis that a 0 or a 1 was transmitted at time \( t_0 \), respectively (see below). Clearly we have

\[
α = \text{Prob}(ξ > -A) = 1 - F(-A),
β = \text{Prob}(ξ > A) = 1 - F(A),
\]

where

\[
F(x) = \frac{1}{\sqrt{2πσ'}} \int_{-∞}^{x} e^{-u^2/2σ'^2} \, du
\]

is the distribution function of the random variable \( ξ \). It follows from the symmetry of \( F(x) \) that \( α + β = 1 \), i.e. that \( C \) is a symmetric channel. At the time \( t_0 + γ \), there are two possibilities. If a 0 was transmitted at time \( t_0 \), then the channel is described by the matrix

\[
C_0 = \begin{bmatrix}
α_o & 1-α_o \\
β_o & 1-β_o
\end{bmatrix},
\]

where

\[
α_o = \text{Prob}[f_0(0) + f_0(γ) + ξ > 0] = 1 - F[-A(1+ρ)],
\]

\[
β_o = \text{Prob}[f_0(0) + f_1(γ) + ξ > 0] = 1 - F[-A(1-ρ)],
\]

and \( ρ = \exp(-γ/υ) \). Clearly we have \( α_o + β_o > 1 \) and \( C_0 \) is asymmetric. Similarly, if a 1 was transmitted at \( t_0 \), then at time \( t_0 + γ \) the channel is
described by the matrix

\[ C_1 = \begin{pmatrix} a_1 & 1-a_1 \\ 1-a_1 & a_1 \\ \beta_1 & 1-\beta_1 \\ 1-\beta_1 & \beta_1 \end{pmatrix}, \]

where

\[ a_1 = \text{Prob} \left[ f_1(o) + f_0(y) + \xi > 0 \right] = 1 - F[A(1-p)] , \]
\[ \beta_1 = \text{Prob} \left[ f_1(o) + f_1(y) + \xi > 0 \right] = 1 - F[A(1+p)] . \]

This time \( a_1 + \beta_1 < 1 \) and again \( C_1 \) is asymmetric.

To find the capacity of the channel with memory, it is simplest to proceed on a doublet basis and enlarge the channel to a new \( 4 \times 4 \) memoryless channel \( C' \), with inputs and outputs consisting of the four pairs 00, 01, 10 and 11. Thus, we consider the \( 4 \times 4 \) channel matrix

\[ C' = \begin{pmatrix} aC_0 & (1-a)C_0 \\ (1-a)C_0 & aC_0 \\ \beta C_1 & (1-\beta)C_1 \\ (1-\beta)C_1 & \beta C_1 \end{pmatrix}, \]

whose elements are themselves \( 2 \times 2 \) matrices. Following Chang [16], we find the capacity of \( C' \) by using Muroga's method (see Section 4). The row-entropy vector \( \vec{H}' \) of \( C' \) is easily seen to be

\[ \vec{H}' = \begin{pmatrix} H(a) + H(a_o) \\ H(a) + H(\beta_o) \\ H(\beta) + H(a_i) \\ H(\beta) + H(\beta_i) \end{pmatrix} = \begin{pmatrix} H(a) \vec{U} + \vec{H}_0 \\ H(\beta) \vec{U} + \vec{H}_1 \end{pmatrix}, \]
where \( \mathbf{U} \) denotes the vector \( \begin{bmatrix} 1 \\ 1 \end{bmatrix} \), and \( \mathbf{H}_0, \mathbf{H}_1 \) are the two-dimensional row-entropy vectors of the channels \( \mathcal{C}_0, \mathcal{C}_1 \), respectively. Now we must find the auxiliary vector \( \mathbf{x}' \) satisfying

\[
\mathbf{x}' = -\mathbf{C}^{-1}_1 \mathbf{H}'.
\]

Since \( \mathbf{C}' \) can be written as

\[
\mathbf{C}' = \begin{bmatrix} \mathcal{C}_0 & 0 \\ 0 & \mathcal{C}_1 \end{bmatrix} \begin{bmatrix} \alpha & 1-\alpha \\ \beta & 1-\beta \end{bmatrix},
\]

we have

\[
\mathbf{C}'^{-1} = \frac{1}{\alpha-\beta} \begin{bmatrix} 1-\beta & \alpha-1 \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} \mathcal{C}_0^{-1} & 0 \\ 0 & \mathcal{C}_1^{-1} \end{bmatrix},
\]

and

\[
\mathbf{x}' = -\frac{1}{\alpha-\beta} \begin{bmatrix} 1-\beta & \alpha-1 \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} \mathcal{C}_0^{-1} & 0 \\ 0 & \mathcal{C}_1^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{H}(\alpha) \mathbf{U} + \mathbf{H}_0 \\ \mathbf{H}(\beta) \mathbf{U} + \mathbf{H}_1 \end{bmatrix}
\]

\[
= -\frac{1}{\alpha-\beta} \begin{bmatrix} (1-\beta) \mathcal{C}_0^{-1}(\mathbf{H}(\alpha) \mathbf{U} + \mathbf{H}_0) + (\alpha-1) \mathcal{C}_1^{-1}(\mathbf{H}(\beta) \mathbf{U} + \mathbf{H}_1) \\ -\beta \mathcal{C}_0^{-1}(\mathbf{H}(\alpha) \mathbf{U} + \mathbf{H}_0) + \alpha \mathcal{C}_1^{-1}(\mathbf{H}(\beta) \mathbf{U} + \mathbf{H}_1) \end{bmatrix}.
\]

Since \( \mathcal{C}_0^{-1} \mathbf{U} = \mathcal{C}_1^{-1} \mathbf{U} = \mathbf{U} \), we have

\[
\begin{align*}
\mathbf{x}' &= -\frac{1}{\alpha-\beta} \begin{bmatrix} (1-\beta)(\mathbf{H}(\alpha) \mathbf{U} - \mathbf{X}_0) + (\alpha-1)(\mathbf{H}(\beta) \mathbf{U} - \mathbf{X}_1) \\ -\beta (\mathbf{H}(\alpha) \mathbf{U} - \mathbf{X}_0) + \alpha (\mathbf{H}(\beta) \mathbf{U} - \mathbf{X}_1) \end{bmatrix} \\
&= -\frac{1}{\alpha-\beta} \begin{bmatrix} 1-\beta & \alpha-1 \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} \mathbf{H}(\alpha) \mathbf{U} - \mathbf{X}_0 \\ \mathbf{H}(\beta) \mathbf{U} - \mathbf{X}_1 \end{bmatrix}.
\end{align*}
\]
where we have introduced the two-dimensional auxiliary vectors $X_0$, $X_1$ corresponding to the channels $C_0$, $C_1$, respectively, i.e., satisfying the equations $X_0 = -C_0^{-1} H_0$ and $X_1 = -C_1^{-1} H_1$. Then, if $X$ is the auxiliary vector corresponding to the channel $C$, i.e. satisfying the equation

$$\begin{vmatrix} X_0 \\ X_1 \end{vmatrix} = \begin{vmatrix} C^{-1} \\ H(\alpha) \end{vmatrix}$$

it follows from (29) that

$$X = \begin{vmatrix} x_0 \\ x_1 \end{vmatrix} = \begin{vmatrix} 0 & \frac{1}{\alpha - \beta} \\ \beta & \alpha \end{vmatrix} \begin{vmatrix} x_0 \\ x_1 \end{vmatrix}$$

Eq. (30) expresses the four-dimensional auxiliary vector $X'$ of the expanded $4 \times 4$ channel $C'$ in terms of the two-dimensional auxiliary vectors $X_0$, $X_1$ and $\bar{X}$ of the binary channels $C$, $C_0$ and $C_1$. Finally, the capacity of $C'$ in bits/doublet is obtained from (30) by using Eq. (17). Thus, the calculation of $X'$ and $c(C')$ is a simple matter, provided one has available a table of the values of the auxiliary vector $X$ corresponding to the general binary channel. Such a table is given in Chang's paper [16], to which the reader is referred for further details.

The simple example just given, where isolated doublets are transmitted, illustrates the general approach to the problem of determining the capacity of channels with memory of the type under consideration. More generally, one can
consider the case of isolated groups of n equally spaced digits, and then the case of a channel driven at a uniform rate with no pauses for recovery from the effects of previously transmitted signals. (Of course, the latter case corresponds to passing to the limit n → ∞ in the case of transmitting isolated groups of n equally spaced digits.) One can also consider m × m channels with memory, where m > 2. Such problems have been studied by Chang and co-workers [16], [17]. Finally, it is natural to study the case where the noise itself has appreciable memory or even the case where there is statistical dependence between the noise and the signals. It appears that much remains to be done along these lines.
References


<table>
<thead>
<tr>
<th>Institution</th>
<th>Address</th>
<th>Contact Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal Communications Commission</td>
<td>Washington, D.C.</td>
<td>Mrs. Barbara O. Grimes</td>
</tr>
<tr>
<td>Upper Atmosphere Research Section</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Radio Propagation Laboratory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Bureau of Standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boulder, Colorado</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argonne National Laboratory</td>
<td>P.O. Box 279</td>
<td>Dr. Roylan De Young</td>
</tr>
<tr>
<td>Lemont, Illinois</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bell Telephone Labs.</td>
<td>Murray Hill, New Jersey</td>
<td>Dr. S. O. Nise</td>
</tr>
<tr>
<td>Carnegie Institute of Washington</td>
<td>Dept. of Terrestrial Magnetism</td>
<td></td>
</tr>
<tr>
<td>S201 Broad Branch Road, N.W.</td>
<td>Washington 15, D.C.</td>
<td></td>
</tr>
<tr>
<td>Attn: Library</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Georgia Tech Research Institute</td>
<td>225 N. Avenue, N.W.</td>
<td>Dr. James B. Boyd</td>
</tr>
<tr>
<td>University of Maryland</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College Park, Maryland</td>
<td></td>
<td>Dr. A. Weinstain</td>
</tr>
<tr>
<td>Institute of Fluid Dynamics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Massachusetts Institute of Technology</td>
<td>Lincoln Laboratory</td>
<td></td>
</tr>
<tr>
<td>P. O. Box 73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexington, Massachusetts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Willow Run Research Center</td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Michigan</td>
<td>Willow Run Airport</td>
<td>Dr. C. L. Dolph</td>
</tr>
<tr>
<td>Temiskamin, Michigan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School of Engineering</td>
<td>New York University</td>
<td></td>
</tr>
<tr>
<td>University Heights</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New York, New York</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shell Fellowship Committee of the Shell Companies Foundation, Inc.</td>
<td>50 West 57th Street</td>
<td>Mr. J. J. Janassen</td>
</tr>
<tr>
<td>New York 20, N.Y.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essex Research and Engineering Co.</td>
<td></td>
<td>Mr. C. L. Brown, Managing</td>
</tr>
<tr>
<td>P.O. Box 21</td>
<td></td>
<td>Engineer</td>
</tr>
<tr>
<td>Linden, New Jersey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Union Carbide and Carbon Corp.</td>
<td>30 E. 4th Street, New York 17, New York</td>
<td>Mr. L. E. Danielson</td>
</tr>
<tr>
<td>New York, New York</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convair</td>
<td>San Diego 12, California</td>
<td>Mr. Martin Stern</td>
</tr>
<tr>
<td>Bell Telephone Labs., Inc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New York 13, N.Y.</td>
<td></td>
<td>Dr. Marvin J. Kelly</td>
</tr>
<tr>
<td>Engineering Library</td>
<td>University of California</td>
<td></td>
</tr>
<tr>
<td>Los Angeles 26, California</td>
<td>L025 Hilgard Avenue</td>
<td></td>
</tr>
<tr>
<td>Convair, A Division of General Dynamics Corp.</td>
<td>College of Engineering</td>
<td></td>
</tr>
<tr>
<td>Daingerfield, Texas</td>
<td>Dept. of Electrical Engineering</td>
<td></td>
</tr>
<tr>
<td>San Diego 12, California</td>
<td>University of Florida</td>
<td></td>
</tr>
<tr>
<td>Convair, A Division of General Dynamics Corp.</td>
<td>Ocala, Florida</td>
<td></td>
</tr>
<tr>
<td>San Diego 12, California</td>
<td>Ft. Lauderdale, Florida</td>
<td></td>
</tr>
<tr>
<td>Convair, A Division of General Dynamics Corp.</td>
<td>Providence, R.I.</td>
<td></td>
</tr>
<tr>
<td>San Diego 12, California</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convair, A Division of General Dynamics Corp.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Diego 12, California</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convair, A Division of General Dynamics Corp.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Diego 12, California</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convair, A Division of General Dynamics Corp.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Diego 12, California</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>