THE λ-CAPACITY OF AN INFORMATION CHANNEL

THE λ-CAPACITY OF AN INFORMATION CHANNEL

By

IBRAHIM DEIF ABDOU, B.Sc.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University
August 1971

MASTER OF SCIENCE (1971)
(Mathematics)

McMASTER UNIVERSITY
Hamilton, Ontario.

TITLE: The λ-Capacity of an Information Channel

AUTHOR: Ibrahim Deif Abdou, B.Sc. (Cairo University, U.A.R.)

SUPERVISOR: Dr. M. Behara

NUMBER OF PAGES: 40

SCOPE AND CONTENTS: Wolfowitz's strong converse of the coding theorem holds for a discrete information channel, iff the channel capacity is independent of λ, the probability of transmission errors. Here we prove the existence of a class of channel capacities for which Wolfowitz's strong converse does not hold.

ACKNOWLEDGEMENTS

I am indebted to Dr. M. Behara for his valuable help in the preparation of this thesis.

I would also like to acknowledge the financial support given me by McMaster University.

TABLE OF CONTENTS

CHAPTER I

PRELIMINARIES ON INFORMATION SOURCES AND CHANNELS

1.1 Introduction

Let X and Y be input and output random variables taking the values $x_1,\dots,x_n$ and $y_1,y_2,\dots,y_m$, respectively, such that

$$P(X=x_k) = p_k,\ 0 \leq p_k \leq 1,\quad \sum_{k=1}^{n} p_k = 1$$

$$P(Y=y_j) = q_j,\ 0 \leq q_j \leq 1,\quad \sum_{j=1}^{m} q_j = 1$$

$$P(x_n,y_j) = P(X=x_k,\ Y=y_j) = r_{kj},\ 0 \leq r_{kj} \leq 1,$$

$$\sum_{k=1}^{n} \sum_{j=1}^{m} r_{kj} = 1$$

$$P(y_j/x_k) = P(Y=y_j/X=x_k) = q_{j/k},\quad \sum_{j=1}^{m} q_{j/k} = 1$$

$$P(x_k/y_j) = P(X=x_k/Y=y_j) = p_{k/j},\quad \sum_{k=1}^{n} p_{k/j} = 1$$

Let $P = (p_1,p_2,\dots,p_n)$ and $Q = (q_1,q_2,\dots,q_m)$ be the probability distributions of X and Y respectively.

The marginal, joint and conditional entropies are defined as:

$$H(P) = H(X) = - \sum_{k=1}^{n} p_k \log_2 p_k \quad .$$

$$H(Q) = H(Y) = - \sum_{j=1}^{m} q_j \log_2 q_j$$

$$H(P,Q) = H(X,Y) = - \sum_{k=1}^{n} \sum_{j=1}^{m} r_{kj} \log_2 r_{kj}$$

$$H(P/Q) = H(X/Y) = - \sum_{k=1}^{n} \sum_{j=1}^{m} P(x_k,y_j) \log_2 P(x_k/y_j)$$

$$= - \sum_{k=1}^{n} \sum_{j=1}^{m} r_{kj} \log_2 P_{k/j}$$

$$H(Q/P) = H(Y/X) = - \sum_{k=1}^{n} \sum_{j=1}^{m} P(x_k,y_j) \log_2 P(y_j/x_k)$$

$$= - \sum_{k=1}^{n} \sum_{j=1}^{m} r_{kj} \log_2 q_{j/k}$$

The average amount of information conveyed by the random variable X about the random variable Y and conversely is defined as:

$$I(P:Q) = I(X:Y) = \sum_{k=1}^{n} \sum_{j=1}^{m} P(x_k,y_j) \log_2 \frac{P(x_k,y_j)}{P(X=x_k)\ P(Y=y_j)}$$

The quantity I(X:Y) is also called the transinformation of the channel.

$$I(X:Y) = H(X) + H(Y) - H(X,Y)$$
$$= H(X) - H(X/Y)$$
$$= H(Y) - H(Y/X)$$

The above definitions and results can be generalized to the case of random vectors and stochastic processes.

1.2 <u>Information Source</u>

By a source, we shall understand any object which gives us information. Hence the output of the source or equivalently the source itself is similar to the space of a random experiment. In other words, a source is the assemblage of all possible events associated with the sample space of a complete random experiment (an experiment whose all events are observable).

Each outcome of the experiment corresponds to an elementary output of the source; and is called a symbol or a character or a letter.

A finite sequence of characters may be called a word or a message, i.e. we think of the source as emitting a sequence of symbols from a fixed finite source. Given an alphabet $S = \{s_1, s_2, \ldots, s_n\}$ successive letters from S are selected according to some fixed probability law.

A binary source is associated with the sample space of a random binary experiment when the experiment is repeated over and over.

The following three steps summarize the information - theoretic performance of a binary source:

(i) Alphabet = {letters}

(ii) Probability matrix

$$[P]=[p,1-p]=[p,q], \; q = 1-p$$

(iii) (1) Self-Information matrix

$$[I]=[-\log_2 p, \; -\log_2(1-p)]$$

(2) Average Information per letter

$$H = \bar{I} = -p \log_2 p - (1-p) \log_2(1-p)$$

The communication entropy for such a binary source is nothing but the average information per letter i.e.,

$$H(P) = -p \log_2 p - (1-p) \log_2(1-p)$$

1.2.1 The Complete Description of the Source

Let A be a finite source - alphabet. If the source under consideration is stochastic, then from the theoretical point of view, its output can be regarded as an aggregate of doubly infinite sequences of the form

$$x = (\ldots x_{-1}, x_0, x_1 \ldots), \; x_i \varepsilon A, \; i = 0, \pm 1, \pm 2, \ldots$$

All the doubly infinite sequences having specified letters at specified positions constitue a cylinder with respect to those specified letters. Hence to describe the stochastic source, it is sufficient to know the probabilities $\mu(Z)$ of all the cylinders Z.

Let us consider the set of all cylinders and its Borel extension F_A. Then if the probabilities $\mu(Z)$ of all cylinders Z are known, the corresponding probabilities $\mu(S)$ of all $S \varepsilon F_A$ can be easily determined. Thus a complete description of the source as a random process is achived by specifying:

(1) An alphabet A

(2) A probability measure $\mu(S)$ defined for all $S \varepsilon F_A$

(3) In particular, we also have $\mu(A^I) = 1$ where A^I is the set of all elementary events of the given space i.e., A^I is the set of all doubly infinite sequences. Therefore we can denote the source by the pair $[A,\mu]$.

1.2.2 Stationary Source

Let $Tx = (\ldots x'_{-1}, x'_0, x'_1 \ldots)$

where

(i) $x = (\ldots x_{-1}, x_0, x_1 \ldots)$

(ii) $x'_k = x_{k+1} \quad (-\infty < k < \infty)$

(iii) The operator T denotes the "shift" by one time unit.

Clearly, if S is any set of elements x, then

$$TS = \{Tx : x \varepsilon S\}$$

It is easy to see that if $S \varepsilon F_A$, then $TS \varepsilon F_A$, also the operator T maps the set A^I onto itself i.e. $TA^I = A^I$.

If $\mu(TS) = \mu(S)$ for any set $S \varepsilon F_A$, then the source is called stationary.

1.3 Information Channel

1.3.1 Concept of Channel

The Mathematical elements which can be used to characterize a given channel are:

(1) A finite input alphabet A which represents the symbols or letters which the channel is capable of transmitting.

(2) A finite output alphabet B which represents the symbols or letters emerging from the channel. It is not necessary that the number of letters in B is the same as the number of letters in A.

If, to every transmitted symbol 'a', a letter of input alphabet A , there is given at the output a unique letter $b = b(a)$ of the output alphabet B then the channel is called a noiseless channel. But in general the interference (noise) causes different letters $b \varepsilon B$ to be obtained at the output in different cases when the same letter $a \varepsilon A$ is transmitted.

Therefore we can speak of probabilities of obtaining the letter $b \varepsilon B$ at the channel output given that the letter $a \varepsilon A$ was transmitted and this probability sometimes depends not only on 'a' but also on the sequence of signals transmitted earlier than 'a'. If it does not depend upon the letters transmitted earlier than 'a', the channel is said to be without memory.

Thus a channel is completely specified, if we know the following three facts:

(1) The input alphabet A

(2) The output alphabet B

(3) The probability $\nu_x(S)$ that (y) is received when a given (x) is transmitted where $S \varepsilon F_B$ for any $x \varepsilon A^I$.

Thus a channel can be represented by the triplet $[A, \nu_x, B]$.

1.3.2 Stationary Channel

We shall call the channel $[A, \nu_x, B]$ stationary, if, for all $x \varepsilon A^I$ and $S \varepsilon F_B$

$$\nu_{Tx}(TS) = \nu_x(S)$$

where T is the shift operator.

1.4 The Discrete Channel

The random nature of the channel may, in many cases, be described by giving a probability distribution over the set of possible outputs.

The distribution of the output will in general depend on the particular input chosen for transmission and in addition may depend on the internal structure of the channel at the time the input is applied.

This means that if we apply a sequence $r_1, r_2, \ldots, r_n$ at the input of a channel, then at the output, perhaps after an appropriate delay, we will receive a sequence $t_1, \ldots, t_n$. It is reasonable to describe the action of the channel by giving a probability distribution over the output sequences $t_1, t_2, \ldots, t_n$ for each input sequence $r_1, r_2, \ldots, r_n$.

The family of distributions should also reflect the fact that the internal state of the channel at the time the input is applied, will affect the transmission of information.

<u>Definition</u>: Given two finite sets τ, τ' to be called the input and output alphabet respectively and an arbitrary set S, called the set of states, a discrete channel is defined as a system of probability functions.

$$P_n(t_1, t_2, \ldots, t_n / r_1, r_2, \ldots, r_n; s)$$

where

$$r_1, r_2, \ldots, r_n \varepsilon \tau$$

and

$$t_1, t_2, \ldots, t_n \varepsilon \tau' \quad \text{and } s \varepsilon S;\ n = 1, 2, \ldots$$

which satisfy:

(1) $P_n(t_1,t_2,\ldots,t_n/r_1,r_2,\ldots,r_n;s) \geq 0$

for all n; $r_1,r_2,\ldots,r_n$,

$t_1,t_2,\ldots,t_n;s$

(2) $\Sigma P_n(t_1,\ldots,t_n/r_1,\ldots,r_n;s) = 1$

for all n; $r_1,r_2,\ldots,r_n$;

$t_1,t_2,\ldots,t_n;s$.

Physcially, we interpret $P_n(t_1,\ldots,t_n/r_1,\ldots,r_n;s)$ as the probability that the sequence $t_1,\ldots,t_n$ will appear at the output if the input sequence $r_1,\ldots,r_n$ is applied and the initial state of the channel, that is, the state just prior to the sequence of (r_i) is (s).

1.4.1 The Discrete Memoryless Channel

The discrete channel is memoryless if

(1) The function $P_n(t_1,\ldots,t_n/r_1,\ldots,r_n;s)$ does not depend on s, and therefore may be written as $P_n(t_1,\ldots,t_n/r_1,\ldots,r_n)$,

(2) $P_n(t_1,\ldots,t_n/r_1,\ldots,r_n) = P_1(t_1/r_1)\ P_1(t_2/r_2) \ldots P_1(t_n/r_n)$

for all $t_1,\ldots,t_n \varepsilon \tau'$; $r_1,\ldots,r_n \varepsilon \tau$; $n = 1,2,\ldots$

1.4.2 Discrete Channel with Finite Memory

Let the input and output alphabets be A and B respectively. In memoryless channels, the noise structure is generally specified by a conditional probability matrix

$$P\{b_j/a_k\} \text{ for all } a_k \varepsilon A \text{ and } b_j \varepsilon B.$$

When the channel has no memory the noise probability matrix is independent of the life history of the channel.

When the channel has a finite memory the noise probability depends on the life history of the transmitted sequences up to the finite memory time prior to the emission of the signal. For this purpose, consider a member x of an input ensemble, and its corresponding member y at the output that is, if x is transmitted, then y is received.

Input Alphabet	Output Alphabet
A	B
$x = (\ldots x_{-2}, x_{-1}, x_0, x_1, \ldots)$	$y = (\ldots y_{-1}, y_0, y_{-1} \ldots)$

and let A^I and B^I be all possible source and received sequences, respectively.

In A^I, let us focus attention on a cylinder $x^{4,1}$ which has a specific letter, say a_1, at a specific position, say x_4, then

$$x^{4,1} = \ldots x_{-1}, x_0, x_1, x_2, x_3, a_1, x_5, \ldots$$

Similarly for a moment, concentrate on a particular cylinder at the output $y^{1,2}$, which has a specified letter, b_2, at the position y_1, then

$$y^{1,2} = \ldots y_{-1}, y_0, b_2, y_2, \ldots$$

To know the noise characteristic, we must know the conditional probability of cylinder $y^{1,2}$ being received when $x^{4,1}$ is transmitted, i.e.

$$P\{y^{1,2}/x^{4,1}\}$$

More specifically, for all possible cylinders $S_A \varepsilon A^I$ at the input, we must have the conditional probability corresponding to any possible cylinder for messages at the output $S_B \varepsilon B^I$. To sum up, the following requirements are necessary in order to specify a general channel:

(1) Input Alphabet A

(2) Output Alphabet B

(3) $P\{S_B/S_A\} = \nu_x$ for all $S_A \varepsilon A^I$ and $S_B \varepsilon B^I$

Thus a discrete channel is specified by the triple $[A, \nu_x, B]$.

1.5 Connection of the Channel $[A, \nu_x, B]$ to the Source $[A, \mu]$

When the letters from some message $x = \ldots x_{-1}, x_0, x_1, \ldots$ from the given source A are fed into the channel one at a time, we obtain at the output the corresponding sequences $y = \ldots y_{-1}, y_0, y_1, \ldots$ of letters from alphabet B.

Let us consider the probability space in which the elementary events are all possible pairs (x,y); $x\varepsilon A^I$ and $y\varepsilon B^I$.

Let C be the set of all pairs (a,b) where $a\varepsilon A$ and $b\varepsilon B$ i.e., we can regard C as a new alphabet and we denote by C^I the set of pairs (x,y) of which we just spoke i.e., specification of $(x,y)\varepsilon C^I$ is equivalent to specification of $x\varepsilon A^I$ and $y\varepsilon B^I$.

We must now introduce probabilities into the space C^I. Let $D\subset A^I$ and $E\subset B^I$ i.e., let D be some set of elements x and E be some set of elements y.

Let $D\varepsilon F_A$, $E\varepsilon F_B$ so that $\mu(D)$ and $\mu(E)$ have finite values. Let $S = \{(x,y):x\varepsilon D,\ y\varepsilon E\}$; clearly $S\subset C^I$, we shall write $S = D\otimes E$ and call it the direct product of D and E.

The probability W(S) of this set of the space C^I should naturally be understood to be the probability of the joint event $x\varepsilon D$ and $y\varepsilon E$.

But the distribution in the space of elementary events $x\varepsilon A^I$ is determined by the μ-law, and for a given x, the distribution in the space of elementary events $y\varepsilon B^I$ is determined by the ν_x-law.

Therefore:

$$W(S) = W(D\otimes E) = \int_D \nu_x(E)\,d\mu(x)$$

1.5.1 Compound Source

Connecting the channel $[A, \nu_x, B]$ to the source $[A, \mu]$ and driving it uniquely determines a new source $[C, W]$ which is called the compound source, where its alphabet is the direct product $A \otimes B$; the set C^I is the elementary events (x,y) and it is the direct product $A^I \otimes B^I$, and the probability measure is

$$W(S) = W(D \otimes E) = \int_D \nu_x(E)\, d\mu(x) \quad \ldots (1)$$

1.6 The Channel Output

Let us put $D = A^I$ in equation (1), while leaving $E \varepsilon F_B$ arbitrary, the quantity $W(D \otimes E)$ is then the probability of the joint event $x \varepsilon A^I$ and $y \varepsilon E$.

Since the first of these two events is certain, $W(D \otimes E)$ is simply the probability $\eta(E)$ of obtaining a sequence y belonging to the set $E \varepsilon F_B$ at the channel output.

Thus we see that the distribution $\eta(E)$ plays the same rule for the space B^I as $\mu(D)$ does for the space A^I. Therefore, for $D = A^I$, equation (1) becomes,

$$\eta(E) = W(A^I \otimes E) = \int_{A^I} \nu_x(E)\, d\mu(x) \quad \ldots (2)$$

Therefore we can speak of the source $[B, \eta]$ as the channel output. This source with the sequence

$y = \ldots y_{-1}, y_0, y_1 \ldots$ of letters from the alphabet B as its output, is uniquely determined, by using equation (2), by the data of our problem, i.e., by the source A and the channel $[A, \nu_x, B]$.

Theorem 1.1: If the source $[A,\mu]$ and the channel $[A,\nu_x,B]$ are stationary, then the source $[C,W]$ is also stationary.

Proof: Suppose $S \subset C^I$ and $S = D \otimes E$ where $D \varepsilon F_A$ and $E \varepsilon F_B$.

It is obvious that $TS = TD \otimes TE$, therefore equation (1) gives:

$$W(TS) = W(TD \otimes TE) = \int_{x \varepsilon TD} \nu_x(TE)\, d\mu(x)$$

which is equivalent to

$$W(TS) = \int_{Tz \varepsilon TD} \nu_{Tz}(TE)\, d\mu(Tz)$$

Since the source $[A,\mu]$ and the channel $[A,\nu_x,B]$ are both stationary, we have:

$$d\mu(Tz) = d\mu(z)$$

$$\nu_{Tz}(TE) = \nu_z(E)$$

Thus

$$W(TS) = \int_{z \varepsilon D} \nu_z(E)\, d\mu(z) = W(S)$$

$$\text{i.e.} \quad W(TS) = W(S)$$

Therefore the compound source $[C,W]$ is also stationary.

Theorem 1.2: If the source $[A,\mu]$ and the channel $[A,\nu_x,B]$ are stationary, then the output alphabet source $[B,\eta]$ is also stationary.

Proof: We know that

$$\eta(E) = W(A^I \otimes E) = \int_{A^I} \nu_x(E)\, d\mu(x)$$

Thus,

$$\eta(TE) = W(A^I \otimes TE) = W(TA^I \otimes TE).$$

It has been proved earlier that, the compound source is stationary if both the source $[A,\mu]$ and the channel $[A,\nu_x,B]$ are stationary.

Thus,

$$\eta(TE) = W(TA^I \otimes TE) = W(A^I \otimes E) = \eta(E)$$

Therefore the source $[B,\eta]$ is also stationary.

Remark: The description of Entropy is given in Reza [8], Khinchin [7], Ash [4] and Wolfowitz [10]. The description of stationary source and channels are given in Kinchin [7], Adler [2], Ash [4] and Wolfowitz [10].

The specification of discrete channel is given in Abramson [1], Ash [4], Feinstein [6] and Wolfowitz [10].

The descriptions of compound sources and channels are given in Kinchin [7] and Feinstein [6].

CHAPTER II

ERGODIC SOURCES AND CHANNELS

2.1 Introduction

The state space of the experiment is the set E of all possible outcomes; for example, E comprises the faces of the coin or the sides of the die.

Suppose the experiment is performed once each minute (say), and that has been going on forever and will continue forever. We can regard the whole doubly infinite sequence of experiments as one grand experiment. An outcome is represented by a doubly infinite sequence $x = \ldots x_{-1}, x_0, x_1 \ldots$ of elements of E.

The probability structure of this grand experiment is described by a probability measure μ in the space A^I of such sequences x.

We want to reflect mathematically the idea that passage of time does not affect the set of joint probability laws governing the experimentations.

Shifting the sequence x to the left, by one place, produces a new sequence $x' = \ldots x_0, x_1, x_2 \ldots$, where x_1 stands now in the 0th place.

Thus x and x' are identical realizations of the grand experiment, apart from a change in the origin of the time scale.

If probability laws are to be constant in time, then μ should assign the same probability to x' as to x.

Actually x and x' will generally have probability '0', and what we must require is that μ be preserved by the transformation T, which carries x to x', in the sense that

$$P(A) = P(TA) \quad \text{for all sets } A$$

This leads us to study measure-preserving transformation in Ergodic Theory.

2.2 Ergodic Source

An ergodic source is a source which, if observed a very long time, will with probability 1 emit a sequence of source symbols which is typical, [by a typical sequence, we mean a sequence, in which, each symbol occurs approximately with its expected frequency].

The source $[A,\mu]$ is called ergodic if the probability $\mu(M)$ of every invariant set $M \varepsilon F_A$ is either '0' or '1'. In other words $TM = M \Rightarrow \mu(M) = 0$ or 1, T being a measure preserving transformation. Let $g_M(x)$ be the indicator function of the set $M \varepsilon F_A$, i.e., $g_M(x) = 1$ if $x \varepsilon M$ and $g_M(x) = 0$ if $x \notin M$.

Birkhoff's ergodic theorem states that in the case of an ergodic source we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} g_M(T^k x) = \mu(M)$$

Where $\sum_{k=0}^{n-1} g_M(T^k x)$ is the number of terms of the series $x, Tx, T^2x, \ldots T^{n-1}x$ which belong to the set M.

The source $[A,\mu]$ is said to reflect the set $M \varepsilon F_A$ if almost everywhere (with probability 1),

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} g_M(T^k x) = \mu(M)$$

i.e., The ergodic source reflects the set $M \varepsilon F_A$. Conversely, if the stationary source $[A,\mu]$ reflects any set $M \varepsilon F_A$ then the source is ergodic.

2.3 The Ergodic Channel

2.3.1 Notation

The probability that $y_n = b$, that is y_n coincides with a given letter $b \varepsilon B$ does not depend on all letters of the transmitted sequence $x = \ldots x_{-1}, x_0, x_1 \ldots$, but only on those with indices rather close to n.

First, we shall always assume that the distribution of y_n is independent of the transmitted signals that are transmitted after x_n i.e., y_n is independent of x_k for $k > n$.

This means $y_n = b$ has the same value for all transmitted messages x for which the signals $\ldots . x_{n-1}, x_n$ are identical.

In this case we speak of a channel without anticipation.

As regards the signals $x_{n-1}, x_{n-2}, \ldots$ preceding x_n, in the majority of cases, only a limited number of them, (e.g. $x_{n-1}, x_{n-2}, \ldots, x_{n-m}$) can influence the distribution of y. This means that the probability that $y_n = b$ is the same for all x with identical $x_{n-m}, \ldots, x_{n-1}, x_n$.

In this case we speak of a channel with a finite memory. The smallest number m for which the above holds, is called the memory of the channel. In particular, the distribution of y_n for a channel without memory $(m = 0)$ depends only on x_n.

<u>Lemma 2.1</u>: Let μ be a stationary probability measure on A^I, then a necessary and sufficient condition

for μ to be ergodic with respect to the shift transformation T, is that for all measurable sets E and $D \subset A^I$, we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \mu(T^{-i}E \cap D) = \mu(E)\,\mu(D) \quad \ldots(2.1.1)$$

Proof: (i) Sufficiency:

Suppose that E is invariant under T; evidently $T^{-i}E \cap E = E$ for all i, then for E = D, the equality which is asserted to be sufficient for the ergodicity of μ reduces to $\mu(E) = \mu(E)^2$ i.e., $\mu(E) = 0$ or 1, thus μ is ergodic.

(ii) Necessity:

Let f and g be the indicator functions of E and D respectively and assume μ to be ergodic. Then the Birkhoff ergodic theorem and its immediate corollaries imply that:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) = \mu(E)$$

Then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x)\, g(x) = \mu(E)\, g(x) \quad \text{a.e. } \mu.$$

Now $f(T^i x)\, g(x)$ is clearly the indicator function of $(T^{-i}E \cap D)$ and by using the following lemmas:

1) $$\int_\Lambda f \, d\mu = \lim_{i \to \infty} \int_\Lambda f_i \, d\mu$$

Where $f = \lim_{i \to \infty} f_i$ and $|f_i| \leq g$, g is summable over Λ_0.

$$2)\quad \int_\Lambda f \, d\mu = \sum_{i=1}^{\infty} \int_\Lambda f_i \, d\mu$$

Where $f = \sum_{i=1}^{\infty} f_i$ and f_i's are measurable and non-negative on Λ_0, $\Lambda \subset \Lambda_0$.

Then we have

$$\mu(E)\,\mu(D) = \int_{A^I} \mu(E)\, g(x)\, \mu(dx)$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \int_{A^I} f(T^i x)\, g(x)\, \mu(dx)$$

$$\mu(E)\,\mu(D) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \mu(T^{-i}E \cap D)$$

Lemma 2.2: Lemma 2.1 remains true if E and D are restricted to be cylinder sets in A^I.

Proof: Clearly, the sufficient assertian of lemma 2.1 is the only part affected. We will show that the validity of equation (2.1.1) for all E and D implies its validity for arbitrary measurable sets E and D.

Let D be a fixed cylinder set, and let m_D be the family of measurable sets for which $L \epsilon m_D$.

Then for $L_1 \subset L_2$ ... where $L = \bigcup_{j=1}^{\infty} L_j, L_i \epsilon m_D$

or for $L_1 \supset L_2 \ldots$ where $L = \bigcap_{i=1}^{\infty} L_i$, $L_i \epsilon m_D$

$$\mu(L)\,\mu(D) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} (T^{-i} L \cap D)$$

Thus m_D is a monotonic class of measurable sets which contain all finite disjoint unions of cylinder sets and hence coincides with the family of measurable sets in A^I. In other words,

$$\mu(L)\,\mu(D) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} (T^{-i} L \cap D)$$

Where D is a cylinder and L is measurable.

2.4 The Ergodic Compound Source

The compound source is ergodic, if the joint measure W on $(A \otimes B)^I$ satisfies the condition that for every pair E and D of cylinder sets in $(A \otimes Y)^I$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} W(T^{-i} E \cap D) = W(E)\, W(D).$$

Let $[A, \nu_x, B]$ be a channel with finite memory m, and suppose that μ is an ergodic input to this channel.

Now any cylinder set in $(A \otimes B)^I$ is of the form $U \otimes V$ where U and V are cylinders in A^I and B^I respectively, both determined by conditions on the same components $t, \ldots t+k$ of their respective product space.

Let then $E = U_1 \otimes V_1$ and $D = U_2 \otimes V_2$ be two cylinder sets in $(A \otimes B)^I$, where E is determined by conditions on components $t, \ldots t+k$ and D by $s, \ldots s+r$, and denote by L_1, L_2 arbitrary cylinders in A^I determined by conditions on the components $t-m, \ldots t-1$; $s-m, \ldots s-1$, respectively.

Now it is clear that there is a positive integer i_0 such that for $i \geq i_0$ the cylinders $T^{-i}[L_1, U_1]$ and $[L_2, U_2]$ in A^I are separated and, therefore, the cylinders $T^{-i}V_1$ and V_2 in B^I are separated by more than m spaces.

Since we know from the definition of the channel with a finite memory that, for any two cylinder sets, $y_i, \ldots, y_j$ and $y'_k, \ldots, y'_n$ such that $j + m < k$ where m is a fixed positive integer, we have

$$\nu\left[(y_i, \ldots y_j) \cap (y'_k, \ldots, y'_n)/x_\infty\right] =$$

$$\nu\left[(y_i, \ldots, y_j/x_\infty\right] \cdot \nu\left[(y'_k, \ldots, y'_n)/x_\infty\right]$$

Thus

$$\nu(T^{-i}V_1 \cap V_2/T^{-i}\left[L_1, U_1\right] \cap \left[L_2, U_2\right]) =$$

$$\nu(T^{-i}V_1/T^{-i}\left[L_1, U_1\right])\ \nu(V_2/\left[L_2, U_2\right])$$

$$= \nu(V_1/[L_1,U_1])\ \nu(V_2/[L_2,U_2])$$

where $i \geq i_0$.

Thus

$$W(T^{-i}E \cap D) = \sum_{L_1,L_2} \Big[\mu(T^{-i}[L_1,U_1] \cap [L_2,U_2])$$

$$\nu(V_1/[L_1,U_1])\ \nu(V_2/[L_2,U_2])\Big]$$

for all $i \geq 0$.

Therefore

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=i_0}^{n+i_0-1} W(T^{-i}E \cap D)$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=i_0}^{n+i_0-1} \sum_{L_1,L_2} \Big[\mu(T^{-i}[L_1,U_1] \cap [L_2,U_2])\ \nu(V_1/[L_1,U_1])\cdot$$

$$\nu(V_2/[L_2,U_2])\Big]$$

$$= \sum_{L_1,L_2} \Big[\mu([L_1,U_1])\ \mu([L_2,U_2])\ \nu(V_1/[L_1,U_1])\ \nu(V_2/[L_2,U_2])\Big]$$

$$\therefore \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} W(T^{-i}E \cap D) = W(U_1 \otimes V_1)\ W(U_2 \otimes V_2) = W(E)\ W(D)$$

Therefore, the compound source is ergodic.

Remark: The description of Ergodicity, Ergodic source, and ergodic channel are given in Khinchin [7], Adler [2], Feinistein [6].

CHAPTER III

THE λ-CAPACITY

3.1 Channel Capacity

Consider an information channel with input alphabet A, output alphabet B and conditional probabilities $P(b_j/a_i)$.

In order to calculate the mutual information

$$I(A,B) = \sum_{A,B} P(a,b) \log_2 \frac{P(a,b)}{P(a)\, P(b)}$$

it is necessary to know the input symbol probabilities $P(a_i)$. Thus the mutual information, depends not only upon the channel, but also upon how we use the channel.

Definition: The capacity of the channel C is,

$$C = \underset{P(a_i)}{\text{Max.}}\ I(A,B) = \text{Max.} \left[H(A) - H(A|B) \right]$$

Note: (1) The capacity of an information channel is a function only on the conditional probabilies defining that channel, i.e., it does not depend upon the input probabilities .

(2) The concepts of channel capacity has been introduced by Shannon.

(3) The maximization is with respect to all possible sets of probabilites that could be assigned to the source alphabet; that is, all discrete memoryless sources.

(4) Sometimes we call the quantity C, the ergodic capacity of the channel. [see, Khinchin [7]].

3.2 The λ-Capacity

Introduction

Let $\underline{X} = (X_1,\ldots,X_n)$ and $\underline{Y} = (Y_1,\ldots,Y_n)$ denote the random vectors of input and output n-sequences respectively.

A message of length N of n-sequences may be written as $\underline{X}^1,\ldots,\underline{X}^N$. Then the probability that ith sequence $(\underline{X}^i)$ will be incorrectly received is denoted by $\lambda(\underline{X}^i)$.

Then the average probability of error of the message is given by $\lambda = \frac{1}{N}\sum_{1}^{N}\lambda(\underline{X}^i)$ and the probability of error by max. $\lambda(\underline{X}^i)$.

Therefore we can define a code as a triple (N,n,λ), where N is the length of the code consisting of sequences of length n, and probability of error max. $\lambda(\underline{X}^i)$.

Therefore, given any discrete channel and let

$N(n,\lambda)$ be the maximum possible number of code words in a code that uses sequences of length n and that has maximum probability of error, at most λ. Therefore the λ-capacity of the given channel is defined by,

$$C(\lambda) = \lim_{n \to \infty} \sup \frac{1}{n} \log N(n,\lambda), \quad 0 < \lambda < 1.$$

If we used codes whose probability of error is $\leq \lambda$, then $C(\lambda)$ may be regarded as the least upper bound of all permissible rates of transmission. We call R a λ-permissible rate of transmission if given any positive integer n_0, there exist a code $(2^{nR}, n, \lambda_n)$ with $n \geq n_0$ and $\lambda_n \leq \lambda$ wherefrom it follows that $C(\lambda)$ is the least upper bound of all λ-permissible transmission rates.

Shannon-Wolfowitz theorem (S-W for short) will now be stated for a discrete channel with memory as follows.

For sufficiently large n and for any $\varepsilon > 0$ there exists a code $(2^{n(C-\varepsilon)}, n, \lambda)$ if $0 < \lambda \leq 1$ and there does not exist a code $(2^{n(C+\varepsilon)}, n, \lambda)$ if $0 \leq \lambda < 1$ where C, is the channel capacity.

The first part of the theorem is due to Shannon and is known as the fundamental theorem of information theory, here we get a lower bound for N.

The second part of the theorem is due to Wolfowitz and is known as the strong converse which gives an upper bound of N. The strong converse assures that $\lambda \to 1$ as $n \to \infty$, when the rate of transmission exceeds the capacity of the channel.

The bounds may be made rather accurate or may be improved by deriving appropriate value of ε.

It can be further shown that the strong converse is equivalent to the fact that C is a constant, independent of λ. Consequently a channel capacity can not be defined if C varies along with λ.

Lemma 3.1: The channel capacity is given by

$$C = \lim_{\lambda \to 0} C(\lambda)$$

Proof: Let $C_0 = \lim_{\lambda \to 0} C(\lambda)$ if $R < C_0$, $0 < \lambda < 1$, then $R < C(\lambda)$. Hence R is λ-permissible.

Thus given any n_0, we can find a code $(2^{nR}, n, \lambda_n)$ with $n \geq n_0$ and $\lambda_n \leq \lambda$.

Since λ may be chosen arbitrarily small it follows that $C \geq C_0$.

If $R > C_0$ there is a $\lambda \varepsilon (o,1)$ such that $R > C(\lambda)$. Since R is not λ-permissible for sufficiently large n, there does not exist a code $(2^{nR}, n, \lambda_n)$ such that $\lambda_n \leq \lambda$.

Therefore no number greater than C_0 can be a permissible transmission rate, hence $C \leq C_0$; then

$$C = \lim_{\lambda \to 0} C(\lambda),$$

Lemma 3.2: The strong converse holds for a discrete channel if and only if

$$C(\lambda) = C \text{ for all } \lambda \varepsilon (0,1)$$

Proof: If strong converse holds, then given any $\lambda \varepsilon (0,1)$ and $R > C$ and for sufficiently large n; no code (2^{nR} ,n,λ_n) can exist with $\lambda_n \leq \lambda$.

Therefore R is not λ-permissible for any λ; therefore:

$$C(\lambda) \leq R \quad \text{for all } \lambda$$

then

$C(\lambda) \leq C$ for all λ; for any $R > C$ and since $C(\lambda)$ is non-decreasing [lemma 3.1], therefore

$$C(\lambda) = C \text{ for all } \lambda,\ \lambda \varepsilon (0,1).$$

Conversely, suppose $C(\lambda) = C$ for all $\lambda \varepsilon (0,1)$, and given $R > C$, then $R > C(\lambda)$, therefore R is not λ-permissible. Therefore, if we have any sequence of codes (2^{nR}.,n,λ_n), we must have $\lambda_n > \lambda$, in other words $\lambda_n \to 1$ as $n \to \infty$, and the strong converse holds.

3.3.1 Formulation of the Main Theorem

Consider a unit square in a coordinate plane, whose abscissa and ordinate are denoted by λ and Y respectively. The sides of the square are given by $\lambda = 0$, $Y = 0$, $\lambda = 1$, $Y = 1$.

Let this square be intersected by the set Σ of straight lines. Next, we will consider two different subsets of the above set Σ.

(1) Let $\Gamma \subset \Sigma$ be a "half-open" set of lines which contain the line $Y = \frac{1}{2}(1+\lambda)$; not the line $\lambda = 1$, and all the lines between them meet at the point (1,1).

(2) Let $\Delta \subset \Sigma$ be a "half-open" set of lines which contains the line $Y = \frac{1}{2}(1-\lambda)$ and not the line $\lambda = 1$, and all lines of the set Δ between them meet at the point (1,0).

Now, we will be in a position to state and prove our main theorem showing the channels which do not have capacity when the elements from the sets Γ and Δ are taken as the arguements of the corresponding entropy functions.

Theorem 3.1: (1) For a given $\gamma_0 \varepsilon \Gamma$, $\exists$ a $\lambda_0 = \lambda_0(\gamma_0)$ such that

$$\lim_{\lambda_0 < \lambda \uparrow 1} C_{\gamma_0}(\lambda) = 1 \ , \ \gamma_0 \varepsilon \Gamma$$

where $C_{\gamma_0}(\lambda) \geq 1 - H(\gamma_0)$.

(2) For a given $\delta_0 \varepsilon \Delta$, $\exists$ a $\lambda_0 = \lambda_0(\delta_0)$ such that

$$\lim_{\lambda \downarrow \lambda_0} C_{\delta_0}(\lambda) = 0 \;, \; \delta_0 \varepsilon \Delta$$

where

$$C_{\delta_0}(\lambda) \leq 1 - H(\delta_0).$$

<u>Proof</u>: Part (1) and part (2) of our theorem holds for any elements of Γ and Δ respectively as is proved below.

To prove the first part of the theorem we have:

(1) The Varsharmov-Gilbert-Sacks condition which states, that if

$$2^{n-k} > \sum_{i=0}^{\frac{1}{2}t-1} \binom{n-1}{i} \quad \ldots.(3.1)$$

it is possible to construct a binary code with 2^k words of length n which will correct $\frac{1}{2}$-tuple and all smaller errors.

(2) Chernoff's exponential bounding technique which gives

$$\sum_{i=0}^{an} \binom{n}{i} \leq 2^{nH(a)} \quad \text{if } 0 \leq a < \tfrac{1}{2}.$$

Now subsituting $a = (\frac{1}{2}t-1)/(n-1)$, we get

$$\sum_{i=0}^{\frac{1}{2}t-1} \binom{n-1}{i} \leq 2^{(n-1)\ H(\frac{1}{2}t-1/n-1)} \quad \ldots.(3.2)$$

where

$$0 \leq (\tfrac{1}{2}t-1)/(n-1) < \tfrac{1}{2}$$

If (3.1) and (3.2) are such that, we get

$$2^{n-1} > 2^{(n-1)\ H(\frac{1}{2}t-1/n-1)}$$

In particular,

$$2^{n-k} > 2^{nH(\frac{1}{2}t/n)} \quad \text{Where} \quad \tfrac{1}{2}t/n < \tfrac{1}{2} \quad \ldots.(3.3)$$

We can find a binary codes with 2^k words of length n which corrects all errors of weight $\frac{1}{2}t$ or less. If we put $R = k/n$ and $\alpha = \frac{1}{2}t/n$ in equation (3.3), we get

$$R < 1 - H(\alpha) \quad \ldots.(3.4),$$

which means that we can construct a code with 2^{nR} words of length n which corrects $(2n\alpha)$-tuple and all smaller errors.

Next, we use the above code in a given channel. If $P(e')$ is the probability of correct decoding and Z is the number of transmission errors, then

$$\begin{aligned} P(e') &= P(z \leq 2n\alpha)\ P(e'/z \leq 2n\alpha) + P(z > 2n\alpha)\ P(e'/z > 2n\alpha) \\ &\geq P(z \leq 2n\alpha)\ P(e'/z \leq 2n\alpha) \\ &\geq \frac{2n\alpha + 1}{n + 1} \cdot 1 \\ &\geq 2\alpha \end{aligned}$$

so that $\lambda = 1 - P(e') \leq 1 - 2\alpha$.

Since $N(n,\lambda)$ is the length of the largest code of words of length n and probability of error λ, it follows from using equation (3.4) that:

$$N(n, 1-2\alpha) \geq 2^{nR}$$

where R may be chosen arbitrarily close to $1 - H(\alpha)$.

Therefore,

$$C_{\gamma_0}(1-2\alpha) \geq 1 - H(\alpha) = 1 - H(1-\alpha), \ (\alpha<\tfrac{1}{2})$$

i.e., $C_{\gamma_0}(\lambda) \geq 1 - H(\frac{1}{2}(1+\lambda)), \ \lambda \varepsilon (0,1)$

i.e., $C_{\gamma_0}(\lambda) \geq 1 - H(\gamma_0)$

where $\frac{1}{2}(1+\lambda) = \gamma_0 \varepsilon \Gamma$.

Consequently,

$$\lim_{\lambda_0 \leq \lambda \uparrow 1} C\gamma_0(\lambda) = 1$$

Since $Y = \frac{1}{2}(1+\lambda)$ is an element of Γ and is choosen arbitrarily, the first part of the theorem holds for any element of Γ.

<u>Note:</u> For $\gamma_o = \frac{1}{2}(1+\lambda)$, $C_{\gamma_0}(\lambda)$ depends on $\lambda \varepsilon (0,1)$. For $\gamma_0 \neq \gamma \varepsilon \Gamma$, $C_\gamma(\lambda)$ depends on $\lambda \varepsilon (\lambda_0, 1)$ such that λ_0 moves among from 0 and γ moves further from γ_0.

To prove the second part of the theorem:

We consider any code $N(n,\lambda)$ for a given channel. Since each error pattern of weight i or n-i has probability

$$\left[(n+1)\binom{n}{i}\right]^{-1}$$

which decreases as i varies from 0 to $\frac{1}{2}n$; it follows that if r is the smallest integer such that,

$$\tfrac{1}{2}r/(n+1) \geq 1 - \lambda \quad \ldots .(3.5)$$

each decoding set of the code has at least $2 \cdot \sum_{j=0}^{n-2} \binom{n}{j}$ sequences; therefore the Hamming upper bound on the number of code words is given in:

$$N \leq \frac{2^n}{2 \cdot \sum_{j=0}^{r-2} \binom{n}{j}} \leq \frac{2^{n-1}}{\sum_{j=0}^{2(n+1)(1-\lambda)-2} \binom{n}{j}} \quad \ldots .(3.6),$$

If we let $N(n,\lambda)$ to be the maximum number of code words in a code that uses sequences of length n and that has maximum probability of error at most λ; we may write

$$N(n,\lambda) \leq \frac{2^{n-1}}{\sum_{j=0}^{2(n+1)(1-\lambda)-2} \binom{n}{j}}$$

Therefore,

$$\log (n,\lambda) \leq \log_2 2^{n-1}\, 1 - \log_2\left[\sum_{j=0}^{k-2} \binom{n}{j}\right] \ldots\ldots(3.7)$$

where $k = 2(n+1)(1-\lambda)$

$$\log (n,\lambda) \leq n-1 - \log_2\left[\sum_{j=0}^{k-2} \binom{n}{j}\right] \ldots(3.8)$$

And by taking the limits on both sides of equation (3.8) and by using Chernoff's exponential bounding technique, we get

$$\lim_{n\to\infty} \frac{1}{n} \log N(n,\lambda) \leq 1 - H(2(1-\lambda))$$

$$\lim_{n\to\infty} \sup \frac{1}{n} \log_2 N(n,\lambda) \leq 1 - H(2(1-\lambda)$$

$$C_{\delta_0}(\lambda) \leq 1 - H(\delta_0)$$

where $2(1-\lambda) = \delta_0 \varepsilon \Delta$.

Consequently,

$$\lim_{\lambda \downarrow \lambda_0} C_{\delta_0}(\lambda) = 0\ ,\ \delta_0 \varepsilon \Delta.$$

Since $Y = 2(1-\lambda)$ is an element of Δ and is chosen arbitrarily, the second part of the theorem holds for any element of Δ.

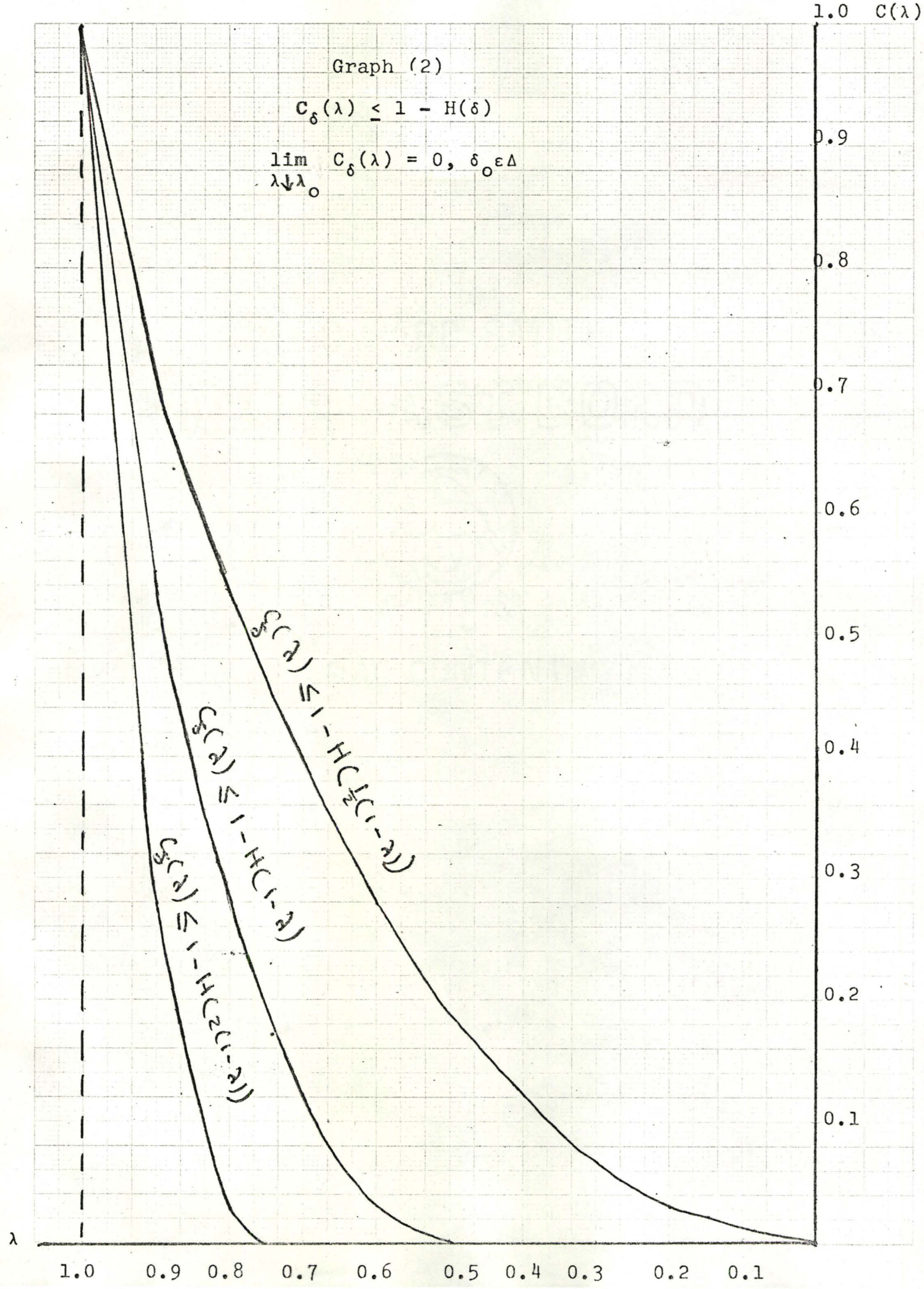
Note: For $\delta = 2(1-\lambda)$, $C_\gamma(\lambda)$ depends on $\lambda \varepsilon (\lambda_0, 1)$ such that $\lambda_0 \downarrow 0$ as δ approaches the boundary $Y = \frac{1}{2}(1-\lambda)$ of Δ.

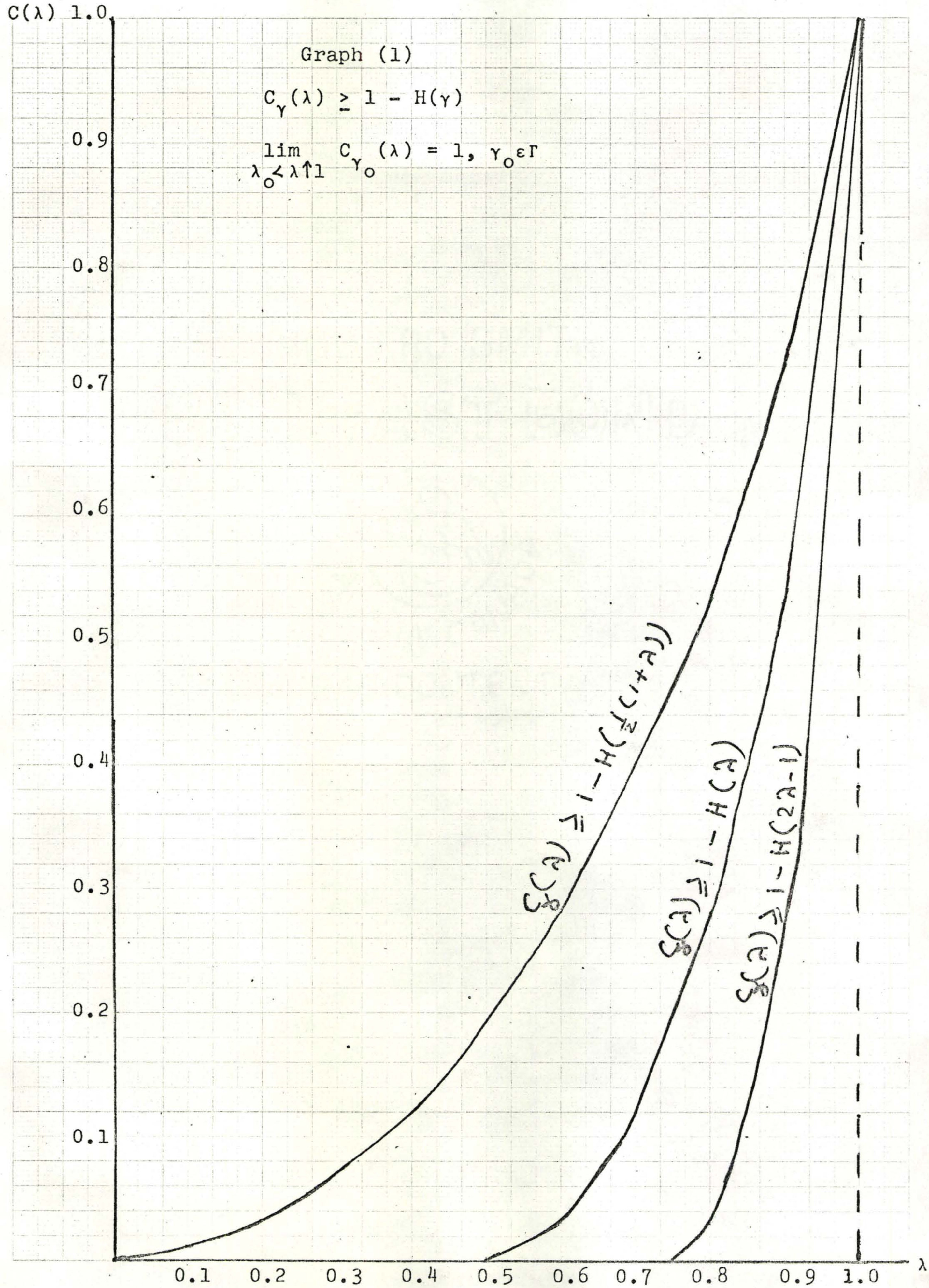
Remark I: (i) It may be noted that the theorem in Ash [3] is only a special case of our theorem 3.1.

(ii) Since the definiton of the capacity given here is independent of any specific channel, the Theorem 3.1 holds for the various channels described in Chapter (1) and (2).

Remark II: The description of mutual information and channel capacity are given in Khinchin [7], Reza [8], Ash [4], Abramson [1] and Wolfowitz [10].

The specification of Shannon-Wolfowitz theorem is given in Behara [5]. Leamma 3.1 and 3.2 are given in Ash [4]. Our theorem is illustrated graphically in the following two graphs.


Graph (2)
Cδ(λ) ≤ 1 - H(δ)
lim λ↓λo Cδ(λ) = 0, δo ε Δ
C(λ)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
λ
1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
C(λ) ≤ 1-H(½(1-λ))
C(λ) ≤ 1-H(1-λ)
C(λ) ≤ 1-H(2(1-λ))



C(λ) 1.0
Graph (1)
$C_\gamma(\lambda) \geq 1 - H(\gamma)$
$\lim_{\lambda_0 < \lambda \uparrow 1} C_{\gamma_0}(\lambda) = 1, \gamma_0 \varepsilon \Gamma$
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
$C_\gamma(\lambda) \geq 1 - H(\frac{1}{2}(1+\lambda))$
$C_\gamma(\lambda) \geq 1 - H(\lambda)$
$C_\gamma(\lambda) \geq 1 - H(2\lambda-1)$
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
λ

References

[1.] Abramson, N. M. (1963), Information Theory and Coding, McGraw-Hill Book Co., New York.

[2.] Adler, R. (1961), Ergodic and Mixing properties of Infinite memory Channels, Proc. Amer. Math. Soc. Vol. 12.

[3.] Ash, R. B. (1965), A simple example of a Channel for which the strong converse fails, IEEE Tans. Inform. Theory.

[4.] Ash, R. B. (1967), Information Theory, John Wiley & Sons, New York.

[5.] Behara, M. (1970), Shannon-Wolfowitz Coding Theorem of Inform. Theory, McMaster University, Canada.

[6.] Feinstein, A. (1958), Foundation of Information Theory, McGraw-Hill Book Co., New York.

[7.] Khinchin, A. (1957), Mathematical foundation of Inform. Theory, Dover Publications Inc., New York.

[8.] Reza, F. (1961), An introduction to Information Theory, McGraw-Hill Book Co., New York.

[9.] Wolfowitz, J. (1963), On Channels without Capacity, Information and Control, 6, 49-54.

[10.] Wolfowitz, J. (1964), Coding Theorems of Inform. Theory, Springer-Verlag, New York, Inc.