

25654



National Library of Canada

Bibliothèque nationale du Canada

CANADIAN THESES ON MICROFICHE

THÈSES CANADIENNES SUR MICROFICHE

NAME OF AUTHOR/NOM DE L'AUTEUR JARL GUNNAR KALLBERG

TITLE OF THESIS/TITRE DE LA THÈSE A GENERALIZED VERSION OF SHANNON'S THEOREM

UNIVERSITY/UNIVERSITÉ SIMON FRASER UNIVERSITY

DEGREE FOR WHICH THESIS WAS PRESENTED/ GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE M. Sc.

YEAR THIS DEGREE CONFERRED/ANNÉE D'OBTENTION DE CE DEGRÉ 1974

NAME OF SUPERVISOR/NOM DU DIRECTEUR DE THÈSE DR. C. KIM

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/DATE Oct. 31 /74 SIGNED/SIGNÉ _____

PERMANENT ADDRESS/RÉSIDENCE FIXE _____

A GENERALIZED VERSION OF
SHANNON'S THEOREM

by

Jarl Gunnar Källberg

B.Sc., University of British Columbia

1972

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
in the Department of
Mathematics

© JARL GUNNAR KALLBERG 1974
SIMON FRASER UNIVERSITY
October 1974

All rights reserved. This thesis may not
be reproduced in whole or in part, by photocopy
or other means, without permission of the author.

APPROVAL

Name: Jarl Gunnar Kallberg
Degree: Master of Science
Title of Thesis: A generalized version of Shannon's
Theorem for Discrete channels.

Examining Committee:

Chairman: S. K. Thomason

C. Kim
Senior Supervisor

R. Russell

D. Mallory

A. R. Freedman

Date Approved: October 17, 1974

ABSTRACT

In this paper we prove a generalization of Shannon's Coding Theorem for Discrete Channels.

Formally, we assume we have X , a finite set of elements, which is the input alphabet. We define \underline{X} to be the doubly-infinite product of copies of X . We let $F_{\underline{X}}$ be the σ -algebra generated by all cylinders in \underline{X} . Then together with λ , a probability on $F_{\underline{X}}$, we have the input message space, $(\underline{X}, F_{\underline{X}}, \lambda)$.

A channel is defined by:

- (i) $(\underline{Y}, F_{\underline{Y}}, \gamma)$ - a probability space which we construct from the finite set Y (the letters to be transmitted) in a manner analagous to the construction of the input message space.
- (ii) $(\underline{W}, F_{\underline{W}})$ - a measurable space constructed from the finite set of output letters W .
- (iii) $v(,)$ - a function on $\underline{Y} \times F_{\underline{W}}$ giving the transition probabilities. Thus $v(\underline{y}, \underline{C})$ is the probability that the transmitted message \underline{y} will be received in the subset \underline{C} of \underline{W} .

The channel is usually assumed to be "noisy", i.e. the message which is received may not be identical to the message which was transmitted. Shannon's Theorem says that for a certain class of channels we can choose a code (a measurable mapping $\underline{X} \rightarrow \underline{Y}$) such that the probability of errors in transmission is small while the rate of transmission is arbitrarily close to the capacity of the channel. Pfaffelhuber in 1971, proved this result for a class of channels (channels with asymptotically decreasing memory and anticipation) that satisfy two conditions. In chapter 3, we show that the second of these is stronger than the concept of strong mixing, and hence by a result

of Adler (which says that any ergodic input is admissible with respect to a channel which is weak mixing) proven in chapter 2, we can replace this condition with the assumption that the channel is weak mixing. We then modify the arguments of Pfaffelhuber and Billingsley to prove Shannon's Theorem for a larger class of channels, namely those that satisfy condition 1) of Pfaffelhuber and are weak mixing.

ACKNOWLEDGEMENTS

I would like to thank Dr. C. Kim for his supervision and for bringing to light some of the errors in the original.

I would also like to thank the Mathematics Department of Simon Fraser and Simon Fraser University for their financial support.

Finally I would like to express my gratitude to Miss Peggy Juzak for her valuable assistance in the typing and proof-reading and to Dr. P. Weinstein for her continued encouragement.

TABLE OF CONTENTS

Chapter I: Introduction

1. History	1
2. Outline	3
3. Notation	4
4. Ergodicity	5
5. Strong Mixing	9
6. Weak Mixing	10
7. The Relationships Between SM, WM, and Erg, With Examples	11
8. Entropy	14
9. The Asymptotic Equipartition Property	17

Chapter II: Ergodic and Mixing Properties of Channels

1. The Channel	19
2. The Compound Message Space	22
3. Properties of Channels	25

Chapter III: Channels With Asymptotically Decreasing
Memory and Anticipation

1. Definitions and Examples	33
2. Indecomposable Finite State Channels	39

3. The Relationship Between Strong Mixing and Asymptotically Decreasing Memory and Anticipation	44
4. Channel Capacity	46
5. Feinstein's Lemma	49

<u>Chapter IV: The Coding Theorem</u>	
1. The Compound Channel	57
2. An Example of Coding Techniques	58
3. Block Codes	61
4. The Coding Theorem by Block Codes	63
<u>Bibliography</u>	67

CHAPTER I: INTRODUCTION

§1 HISTORY

Claude Shannon [1] in 1948 began mathematically to formalize the concepts of production and transmission of information. One of his main results was the coding theorem. Postponing the formal definition of the terms, we can roughly state this theorem.

Suppose we have a discrete channel with a finite number of states. Then it is possible to encode the input messages so that information can be transmitted at a rate approaching the channel capacity and with arbitrarily small probability of error.

Here the channel may be "noisy". By this we mean that due to some imperfection in the device the message received may not be identical with the message sent. However, Shannon's proof was sketchy and proved difficult to carry out with sufficient rigor.

In 1953, Feinstein developed a new approach to the proof of Shannon's theorem. We quote Khinchin [2,p.90]: "Feinstein's idea consists in deriving from the channel itself as much as can possibly be used to prove Shannon's theorems, before coding and even before connecting the channel to any particular source." Khinchin [2], using Feinstein's lemma, proved the theorem for a class of channels with finite memory and zero anticipation.

Later, Blackwell, Breiman and Thomasian [3], proved a coding theorem for indecomposable finite-state channels, thus avoiding some of the pathologies inherent in channels with finite memory. However, the concept of finite-state channels is somewhat unsatisfactory because given any physical channel, it is in general difficult to determine this internal structure.

In Wyner's "Recent Results in the Shannon Theory" [4], he states that the most general version of Shannon's Theorem is by Pfaffelhuber [5]. In [5], Pfaffelhuber introduces the concept of channels with asymptotically decreasing memory and anticipation and outlines a proof of Shannon's Theorem for these channels. He also shows that both channels with finite memory and anticipation, and indecomposable finite-state channels, have asymptotically decreasing memory and anticipation. Thus [5] generalizes the theorems of [2] and [3].

§2 OUTLINE

Chapter one introduces the notation and the basic ergodic theory we will employ. Following Halmos [6], we state the functional forms of ergodicity, weak mixing, and strong mixing. From Billingsley [7] and Feinstein [8], we give two versions of the very important asymptotic equipartition property. Again from [7], we very briefly define and give some properties of the entropy function.

The theorems of chapter two originate with Adler [9]. Here we also introduce the definition of a channel and some of the related concepts, such as asymptotic independence from the remote past. It is Theorem 2.4 from this chapter that enables us to generalize the results of [5].

In chapter three, we introduce the notion of channels with asymptotically decreasing memory and anticipation and we give some examples. We then show that condition 2) in the definition of channels with asymptotically decreasing memory and anticipation is stronger than the notion of strong mixing. This implies that Theorem 2 of [5] is weaker than Theorem 2.2, which Adler had proven more than a decade previously. We then prove the extension of Feinstein's Lemma.

Finally, in chapter four, we prove a version of Shannon's Coding Theorem for channels which satisfy condition 1) in the definition of asymptotically decreasing memory and anticipation, and are weak mixing. The proof involves the extension of Feinstein's Lemma and a modification of the proof of Shannon's Theorem (in the memoryless case) given in [7].

§3 NOTATION

We give a summary of the notation that will be employed throughout this paper. Note especially the use of $(t_1, t_2]$.

S^C is the complement of the set S .

$\#S$ is the number of elements in the set S .

S^m means $\prod_{i=1}^m S$, i.e. the Cartesian product of m copies of S .

$1_S(x)$ is the characteristic (or indicator) function of S .

$\underline{S} = \prod_{i=-\infty}^{\infty} S$, i.e. the doubly-infinite product of copies of S .

If $\underline{s} \in \underline{S}$ then s_t will denote the t -th coordinate of \underline{s} .

$(t_1, t_2] = \{t \text{ an integer: } t_1 < t \leq t_2\}$.

If $\underline{s} \in \underline{S}$ then $s_{(t_1, t_2]} = (s_{t_1+1}, s_{t_1+2}, \dots, s_{t_2})$.

$\lambda.a.a.x$ means for all x , except possibly a set of λ measure zero.

$\text{Clim}_{N \rightarrow \infty} a_n$ is $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=0}^{N-1} a_j$ i.e. the Cesaro limit of a_n .

$S+T = (S-T) \cup (T-S)$ i.e. $+$ is symmetric difference.

L_p is the set of functions f satisfying $\int |f|^p < \infty$.

$\sigma(S)$ is the σ -algebra generated by S .

iff will mean if and only if.

//. will be used to denote the end of a proof.

ERG is a contraction for ergodic,

WM for weak mixing, and

SM for strong mixing.

mpt means measure-preserving transformation.

\mathbb{Z} denotes the set of all integers.

§4 ERGODICITY

Let (X, F, P) be a probability space (i.e. X is a nonempty set, F is a σ -algebra of subsets of X and P is a measure on F such that $P(X)=1$).

Let T be a measurable transformation $:X \rightarrow X$. Recall that T is measurable if $A \in F$ implies that

$$T^{-1}[A] = \{x \in X : Tx \in A\} \in F.$$

We will say that T is invertible if

- (i) T is one-to-one,
- (ii) $TX = X$, and
- (iii) $A \in F$ implies $TA \in F$.

We will call T a measure preserving transformation (mpt) if

$$P(T^{-1}[A]) = P(A) \text{ for all } A \in F.$$

We note that if T is invertible, then T is a mpt iff

$$P(TA) = P(A), \text{ [7, p.2].}$$

An ergodic transformation can be characterized as one in which, for almost all x (i.e. for all $x \in X$ except for a set of measure zero) the orbit of x (which is the set $\{x, Tx, T^2x, \dots\}$) "replicates" X . By this we mean that for all A in F , the orbit of almost all x enters A with asymptotic relative frequency $P(A)$. This is expressed formally by

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} 1_A(T^n x) &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} 1_A(T^n x) \\ &= P(A) \quad \text{P.a.a.x.} \end{aligned}$$

Note that $\sum_{n=0}^{k-1} 1_A(T^n x)$ is the number of elements of the set $\{x, Tx, \dots, T^{k-1}x\}$ that lie in A .

A set $A \in F$ is invariant under T (or T -invariant) if

$$P(A \cap T^{-1}[A]) = 0.$$

T is ergodic (ERG) if every T -invariant set has measure zero or one.

A function $f(x)$ which is F -measurable, is said to be T -invariant if

$$f(Tx) = f(x) .$$

We can now state a theorem of fundamental importance in ergodic theory. For proofs see [7] and for more general versions of the theorem see Foguel [10].

Theorem 1.1 The Pointwise Ergodic Theorem.

If $f \in L_1$, then there exists $\hat{f} \in L_1$ which is T -invariant such that

$$\lim_{n \rightarrow \infty} f(T^n x) = \hat{f}(x) \quad \text{P.a.a.x, and}$$

$E(f) = E(\hat{f})$. (Here $E(f)$ is the expectation of f , i.e. $\int f(x)P(dx)$.)

Furthermore, if T is ergodic, then

$$\hat{f}(x) = E(f) \quad \text{P.a.a.x.}$$

We can obtain one useful application by taking T ergodic and f to be the characteristic function of A , for $A \in F$. We then have

$$(1.1) \quad \lim_{n \rightarrow \infty} 1_A(T^n x) = E(f) \\ = P(A) \quad \text{P.a.a.x.}$$

This is the sense in which the orbits of an ergodic transformation are said to replicate X . Roughly, it says

that for large n , the set $\{x, Tx, \dots, T^n x\}$ has the same statistical properties as X .

If (1.1) holds we say that T respects A . By Theorem 1.1 we can see that an ergodic mpt respects each set A .

We say that F_0 is an algebra if

- (i) F_0 is a nonempty class of sets,
- (ii) $E_1, E_2 \in F_0$ implies that $E_1 \cup E_2 \in F_0$, and
- (iii) $E \in F_0$ implies that $E^c \in F_0$.

If F_0 is closed under countable unions then F_0 is a σ -algebra. The σ -algebra generated by F_0 , denoted $\sigma(F_0)$ is the smallest σ -algebra containing F_0 . Now from [7, p.17] we have

Theorem 1.2 Let F_0 be an algebra generating the σ -algebra F . Then T is ergodic if

$$(1.2) \quad \lim_{n \rightarrow \infty} P(A \cap T^{-n}[B]) = P(A)P(B) \quad \text{for all } A, B \in F_0.$$

This theorem says that to prove T is ergodic on a σ -algebra F , it is sufficient to show (1.2) holds on any algebra which generates F .

A theorem analagous to Theorem 1.1 is the Mean Ergodic Theorem [6, p.16]. One consequence of it is the functional form for ergodicity. See [6, p.36] for proof.

Theorem 1.3 T is ergodic iff

$$\lim_{n \rightarrow \infty} \int f(T^n x) g(x) P(dx) = \int f(x) P(dx) \cdot \int g(x) P(dx)$$

for all $f, g \in L_2$.

§5 STRONG MIXING

A mpt T is said to be strong mixing (SM) if

$$(1.3) \quad P(A \cap T^n[B]) \rightarrow P(A)P(B) \quad \text{for all } A, B \in \mathcal{F}, \text{ as } n \rightarrow \infty.$$

Equivalently if

$$\lim_{n \rightarrow \infty} P(T^n[B] | A) - P(T^n[B]) = 0.$$

Hence mixing is related to the "wearing-off" of the initial conditions. As for ergodic mpts we can prove the following

Theorem 1.4 Let \mathcal{F}_0 be an algebra generating \mathcal{F} . Then if (1.3) holds for all $A, B \in \mathcal{F}_0$, then T is strong mixing.

Theorem 1.5 T is strong mixing iff

$$\lim_{n \rightarrow \infty} \int f(T^n x) g(x) P(dx) = \int f(x) P(dx) \cdot \int g(x) P(dx)$$

for all $f, g \in L_2$.

§6 WEAK MIXING

If T is a mpt, T is weak mixing (WM) iff

$$(1.4) \quad \lim_{n \rightarrow \infty} |P(\bar{T}^n[A] \cap B) - P(A)P(B)| = 0 \text{ for all } A, B \in \mathcal{F}.$$

Analogous to the SM case, we have the following:

Theorem 1.6 Let \mathcal{F}_0 be an algebra generating \mathcal{F} . Then if

(1.4) holds for all $A, B \in \mathcal{F}_0$, then T is weak mixing.

Theorem 1.7 T is weak mixing iff

$$\lim_{n \rightarrow \infty} \left| \int f(T^n x) g(x) P(dx) - \int f(x) P(dx) \cdot \int g(x) P(dx) \right| = 0$$

for all $f, g \in L_2$.

§7 THE RELATIONSHIP BETWEEN SM, WM, AND ERG, WITH EXAMPLES

It is clear from the definitions that SM implies WM, and WM implies ERG. Following [6,p.38] we can give an intuitive interpretation of these. We let T be a particular way of stirring the contents of a vessel full of 90 percent gin and 10 percent vermouth. Let F be some region of the vessel. Ergodicity is expressed by saying on the average F has 10 percent vermouth. Strong mixing is expressed by saying after a while F will have 10 percent vermouth in it. Weak mixing can be interpreted by saying that after a while F will have 10 percent vermouth in it with the exception of a few rare instants during which it may be either too strong or too sweet.

Example 1.1 A Model for a Doubly-Infinite Sequence of Bernoulli Trials.

Let X be a set with r (finite) elements. This set is assumed to be the number of possible outcomes of an experiment. Let F be the set of all subsets of X. We define the probability p on X by assigning to each $x \in X$ a nonnegative p_x such that

$$\sum_{x \in X} p_x = 1 .$$

Let $(\underline{X}, \underline{F}, P)$ be the doubly-infinite product of copies of the probability space (X, F, p) . Now, $\underline{x} \in \underline{X}$ is a doubly-infinite sequence $(\dots, x_{-1}, x_0, x_1, \dots)$ of elements from X . We can interpret \underline{x} as an infinite number of Bernoulli trials, x_n being the outcome of the trial at time n .

The σ -algebra \underline{F} is generated by the (thin) cylinders.

These are the sets of the form

$$\{\underline{x} \in \underline{X} : x_j = i_j \text{ with } i_j \in X \text{ and } j \in (n, n+k]\}.$$

(Recall our nonstandard usage of $(n, n+k]$ to denote the set $\{j \text{ an integer} : n < j \leq n+k\}$.) This set is called the cylinder determined by the interval $(n, n+k]$ and the sequence $(i_{n+1}, i_{n+2}, \dots, i_{n+k})$.

We define P on \underline{F} by its values on cylinders

$$P\{\underline{x} \in \underline{X} : x_j = i_j \text{ with } i_j \in X \text{ and } j \in (n, n+k]\} = \prod_{j=n+1}^{n+k} p_{i_j}.$$

The fact that this uniquely defines P is a consequence of the Kolmogorov Extension Theorem [2, p.3].

We now define T , the shift on \underline{X} , (in this example T is the Bernoulli shift) by

$$T(\dots, x_{-1}, x_0, x_1, \dots) = (\dots, x_0, x_1, x_2, \dots) \quad \text{or} \\ (Tx)_n = x_{n+1}.$$

(Recall that $(Tx)_n$ is the n -th coordinate of Tx .) T shifts each coordinate of \underline{x} by one to the left.

T is invertible, a mpt, and is strong mixing. To prove the last assertion, let \underline{A} and \underline{B} be cylinders in \underline{X} . Then

$\bar{T}^n[B]$ is also a cylinder in X , and if n is taken sufficiently large, the intervals determining A and $\bar{T}^n[B]$ are disjoint. Thus

$$P(\underline{A} \cap \bar{T}^n[B]) = P(\underline{A})P(\underline{B})$$

for large enough n . Thus (1.3) holds for cylinders. Applying Theorem 1.4 we are done.

To indicate one other application of Theorem 1.1, let for $i \in X$

$$f(\underline{x}) = \begin{cases} 0 & \text{if } x_1 \neq i \\ 1 & \text{if } x_1 = i \end{cases} .$$

So that

$$f(\underline{x}) = 1_{\{\underline{x}: x_1 = i\}}(\underline{x}) .$$

Now $\sum_{k=0}^{n-1} f(T^k \underline{x})$ is the number of occurrences of i in $\{x_1, \dots$

$x_n\}$. Hence $\text{Clim}_{n \rightarrow \infty} f(T^k \underline{x})$ is the asymptotic relative frequency

of occurrence of the outcome i in the trials at positive time points. By the Ergodic Theorem this limit exists and is $E(f)$, which is p_i almost everywhere. Note that this is the strong law of large numbers for Bernoulli trials.

See [7] for further examples and applications.

§8 ENTROPY

We begin with a number of definitions. We will say $A = \{A_1, A_2, \dots, A_n\}$ is an F -decomposition of X if A is a finite collection of nonempty elements from F forming a partition of X . The A_i are called the atoms of A . There is a complete duality between F -decompositions and finite subfields (the terms algebra and field are equivalent) of F . By this we mean a finite subfield of F induces a unique F -decomposition, and conversely.

The entropy function $\eta(t)$ for $0 < t < 1$, is defined by

$$\eta(t) = \begin{cases} 0 & \text{if } t=0 \\ -t \log t & \text{if } 0 < t < 1 \end{cases} .$$

Here we let B be the base of the logarithm. It is easy to verify that $\eta(t)$ is continuous except at zero and is non-negative.

We can now define the entropy of a finite field A with atoms $\{A_1, \dots, A_n\}$ as

$$\begin{aligned} H(A) &= \sum_{i=1}^n \eta(P(A_i)) \\ &= -\sum_{i=1}^n P(A_i) \log P(A_i) . \end{aligned}$$

If C is another finite subfield of F with atoms $\{C_1, \dots, C_m\}$, we can define the conditional entropy of A given C by

$$H(A|C) = \sum_{j=1}^m P(C_j) \sum_{i=1}^n \eta(P(A_i|C_j)) \quad .$$

From [7] we can give some intuitive ideas behind these definitions. The expression

$$(1.5) \quad \sum_{i=1}^r \eta(p_i) = - \sum_{i=1}^r p_i \log p_i$$

is a measure of the amount of randomness in a single roll of a die if p_1, \dots, p_r represents the probability of each of the different faces. Kolmogorov derived (1.5) from a set of axioms one feels a measure of randomness should satisfy. A die that one would assume is the most random would be one with each $p_i = 1/r$. Note that this choice of p_i maximizes (1.5). At the other extreme, (1.5) is zero if and only if one $p_i = 1$. In this die the outcome is the least random. Thus (1.5) measures the randomness in the experiment consisting of one roll of the die. This we will call the entropy of the experiment. It also measures the information in the experiment, the amount we learn from the outcome.

Some of the basic properties of the entropy of the finite field A and of the conditional entropy of A given B

where B is another finite field, are the following

$$(A1) \quad H(A \vee B | C) = H(A | C) + H(B | A \vee C),$$

$$(A2) \quad H(A | C) \leq H(B | C) \quad \text{if } A \subset B$$

$$(A3) \quad H(A | C) \leq H(A | B) \quad \text{if } B \subset C.$$

Here C is another finite field and $A \vee B$ denotes the σ -algebra generated by A and B .

§9 THE ASYMPTOTIC EQUIPARTITION PROPERTY

Example 1.1 can easily be extended to obtain a generalized shift on \underline{X} . Here T is assumed to be the shift as before, but now P is any probability preserving T .

Let us denote by $H(x_0, x_1, \dots, x_{n-1})$ the entropy of the finite field having as atoms the r^n sets, (recalling that $\#X=r$) $\{\underline{x} \in \underline{X} : x_{[0, n-1]} = u\}$ for $u \in X^n$. The Kolmogorov-Sinai Theorem then implies that the entropy of T , denoted $h(T)$, is

$$h(T) = \lim_{n \rightarrow \infty} 1/n H(x_0, x_1, \dots, x_{n-1})$$

The asymptotic equipartition property (AEP) is a consequence of the following theorem; see [7, p.128] for proof.

Theorem 1.8 The Shannon-McMillan-Breiman Theorem.

If T is an ergodic shift then

$$\lim_{n \rightarrow \infty} \{-1/n \log p_0(x_0, x_1, \dots, x_{n-1})\} = h(T) \quad \text{P.a.a. } \underline{x}.$$

Here $p_0(x_0, \dots, x_{n-1})$ is the probability of the sequence (x_0, \dots, x_{n-1}) being observed. More specifically, for any positive integer n , the mapping \underline{x} into (x_0, \dots, x_{n-1}) induces a probability on X^n . The probability of such an n -tuple $u = (u_1, \dots, u_n)$ is

$$p_0(u) = P(\underline{x} \in \underline{X} : x_{[0, n-1]} = u)$$

The AEP says that for large n , X^n can be decomposed into two subsets. The first subset has low total probability and the second subset consists of n -tuples with probabilities near $B^{-nh(T)}$. Thus we state from [7,p.135]

Theorem 1.9 The Asymptotic Equipartition Property.

Let T be an ergodic shift with entropy h . Then for any positive ε there exists a positive integer $b_0(\varepsilon)$ such that if $b \geq b_0(\varepsilon)$ then X^b decomposes into two sets H and L such that

$$\sum_{u \in L} p_0(u) = P\{x_{(0,b]} \in L\} \leq \varepsilon$$

and such that

$$B^{-b(h+\varepsilon)} < p_0(u) = P\{x_{(0,b]} = u\} < B^{-b(h-\varepsilon)}$$

for any b -tuple $u \in H$.

Another version of the AEP is given by [8,p.88].

Theorem 1.10 Let T be an ergodic shift with entropy h . Then for all positive ε and δ there exists an integer $n(\varepsilon, \delta)$ such that if S is the set of $u \in X^n$ for which

$$|1/n \log p_0(u) + h| < \varepsilon$$

does not hold, then

$$p_0(S) < \delta$$

when

$$n \geq n(\varepsilon, \delta)$$

CHAPTER II: ERGODIC AND MIXING PROPERTIES OF CHANNELS

§1 THE CHANNEL

Let (X, F_X) be a measurable space (i.e. a set X and a σ -algebra F_X of subsets of X such that $\bigcup_{F \in F_X} F = X$). We will denote by $(\underline{X}, F_{\underline{X}})$, as before, the doubly-infinite product space

$$(\underline{X}, F_{\underline{X}}) = \prod_{i=-\infty}^{\infty} (X, F_X).$$

If λ is a probability on F_X we call $(\underline{X}, F_{\underline{X}}, \lambda)$ a message space or information source.

Let T denote the shift on \underline{X} . We shall say the message space is stationary if T is a λ -mpt and we shall say $(\underline{X}, F_{\underline{X}}, \lambda)$ is ergodic if T is an ergodic λ -mpt.

In addition, we assume we have the following doubly-infinite product space

$$(\underline{Y}, F_{\underline{Y}}) = \prod_{i=-\infty}^{\infty} (Y, F_Y),$$

where (Y, F_Y) is a measurable space.

We can now define a channel as a triple $((\underline{X}, F_{\underline{X}}, \lambda), (\underline{Y}, F_{\underline{Y}}), v(\cdot, \cdot))$, where $v(\cdot, \cdot)$ is a function on $\underline{X} \times F_{\underline{Y}}$ such that

- (i) $v(\underline{x}, \cdot)$ is a probability on $F_{\underline{Y}}$ λ -a.a. \underline{x} ,
- (ii) $v(\cdot, \underline{E})$ is a measurable function of \underline{x} , for

all fixed $\underline{E} \in F_Y$.

The function $v(\cdot, \cdot)$ is called the kernel of the channel.

For brevity, we will write the channel as $(\underline{X}, \underline{Y}, v)$.

The channel will be called stationary if

$$v(T\underline{x}, T\underline{E}) = v(\underline{x}, \underline{E}) \quad \lambda.a.a.\underline{x} \text{ and for all } \underline{E} \in F_Y.$$

Here we have defined T on $\underline{X} \times \underline{Y}$ as the direct product of the shifts on \underline{X} and \underline{Y} , viz.

$$T(\underline{x}, \underline{y}) = (T\underline{x}, T\underline{y}) .$$

In the above, X corresponds to the input alphabet, a set (usually finite) of letters to be transmitted. Hence, $(\underline{X}, F_X, \lambda)$ describes the input to the channel. Similarly, Y corresponds to the output alphabet and (\underline{Y}, F_Y) describes the output. A message is $\underline{x} \in X$ with the interpretation that x_i is the letter sent at time i . The kernel $v(\underline{x}, \underline{E})$ is the probability that the message \underline{x} after being fed through the channel will lie in a subset \underline{E} of F_Y . Finally, channel stationarity means that the structure of the channel is time-invariant.

Example 2.1 The Channel Without Memory.

Let (c_{jk}) be a stochastic matrix (i.e. each row consists of nonnegative entries which sum to one). The rows and columns are indexed by elements of X and Y respectively. We can now define the kernel by

$$v(\underline{x}, \{y_l = k_l : l = m, \dots, n\}) = \prod_{l=m}^n c_{x_l k_l}.$$

Recall that v is uniquely defined by specifying its values on cylinders.

This channel is stationary since each letter is treated independently. Here, c_{jk} is the probability that the letter k is received given that the letter j has been fed into the channel.

Take $X=Y=\{0,1\}$; this is the binary symmetric channel.

Let

$$(c_{jk}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this case we have a "noiseless" channel. It has the property that if the message \underline{x} is sent, \underline{x} is received with perfect accuracy.

Usually, however, we expect the message to be affected by the presence of noise in the channel. For example, with the binary symmetric channel we may have

$$(c_{jk}) = \begin{bmatrix} .9 & .1 \\ .05 & .95 \end{bmatrix}.$$

In this instance, the received message may not necessarily correspond exactly to the transmitted message. The presence of noise in the channel may lead to a loss of information.

Coding theory deals with the problem of "coding" messages so that the transmission of information is effective despite the presence of noise.

§2 THE COMPOUND MESSAGE SPACE

$(\underline{X}, F_{\underline{X}}, \lambda)$ with $(\underline{X}, \underline{Y}, v)$ defines a compound message space $(\underline{X} \times \underline{Y}, F_{\underline{X}} \times F_{\underline{Y}}, \omega)$, where ω is a probability on $F_{\underline{X}} \times F_{\underline{Y}}$ which describes the joint distribution of the input and output of the channel. We define

$$(2.1) \quad \omega(\underline{C} \times \underline{E}) = \int_{\underline{C}} v(\underline{x}, \underline{E}) \lambda(d\underline{x})$$

for $\underline{C} \in F_{\underline{X}}$ and $\underline{E} \in F_{\underline{Y}}$.

We now prove this is the correct expression. The integral is well-defined since v is a kernel. For $\underline{M} \in F_{\underline{X}} \times F_{\underline{Y}}$ take more generally

$$(2.2) \quad \omega(\underline{M}) = \int_{\underline{X}} v(\underline{x}, \{\underline{y}: (\underline{x}, \underline{y}) \in \underline{M}\}) \lambda(d\underline{x}).$$

Let A denote the class of sets in $F_{\underline{X}} \times F_{\underline{Y}}$ for which the integrand in (2.2) is measurable, and let B denote the class of measurable rectangles. If $\underline{M} \in B$ then $\underline{M} \in A$ since $\{\underline{y}: (\underline{x}, \underline{y}) \in \underline{M}\}$ is a section, thus measurable since v is a kernel, Halmos [11, p.143].

If \underline{M} is a finite disjoint union from B , then $\underline{M} \in A$ since we can then express the integrand as a sum of measurable functions. Thus A contains the ring (A ring is a class of sets closed under finite unions and finite intersections.) generated by B , [11, p.139].

A is a monotone class since a limit of measurable functions is measurable. Hence, by the monotone class theorem, [11,p.27], A contains $\sigma(B)$, which is $F_{\underline{X}} \times F_{\underline{Y}}$. Hence the integral in (2.2) is well-defined. Since v is a kernel, and $(\underline{X}, F_{\underline{X}}, \lambda)$ is a probability space, it follows that ω is a probability.

Furthermore, ω is the unique probability satisfying (2.1) since any two probabilities that agree on a ring must agree on the generated σ -ring [11,p.54].

From the compound message space can be constructed the output message space $(\underline{Y}, F_{\underline{Y}}, \mu)$ with

$$\begin{aligned} \mu(\underline{E}) &= \omega(\underline{X} \times \underline{E}) \\ &= \int_{\underline{X}} v(\underline{x}, \underline{E}) \lambda(d\underline{x}) \quad \text{for } \underline{E} \in F_{\underline{Y}}. \end{aligned}$$

We have the following result.

Theorem 2.1 If $(\underline{X}, F_{\underline{X}}, \lambda)$ is a stationary message space and $(\underline{X}, \underline{Y}, v)$ is a stationary channel, then the compound message space and the output message space are both stationary.

Proof. Let $\underline{A} = \underline{C} \times \underline{E}$ with $\underline{C} \in F_{\underline{X}}, \underline{E} \in F_{\underline{Y}}$ (We note that it is sufficient to show the result for measurable rectangles.).

$$\begin{aligned} \omega(\underline{TA}) &= \omega(\underline{TC} \times \underline{TE}) \\ &= \int_{\underline{TC}} v(\underline{x}, \underline{TE}) \lambda(d\underline{x}) \end{aligned}$$

Now using a change of variable $\underline{z} = T^{-1}\underline{x}$,

$$= \int_{\underline{C}} v(\underline{Tz}, \underline{TE}) \lambda(\underline{Tdz}).$$

Since T is a λ -mpt, and by channel stationarity

$$= \int_{\underline{C}} v(\underline{z}, \underline{E}) \lambda(\underline{dz})$$

$$= \omega(\underline{C} \times \underline{E}).$$

The result for the output message space follows by setting

$$\underline{C} = \underline{X}. \quad //.$$

§3 PROPERTIES OF CHANNELS

We let $K=(\underline{X}, \underline{Y}, \nu)$ be a channel.

(a) (Khinchin) K is said to have finite memory if there exists a positive integer m , such that if $\underline{E} \in \mathcal{F}_{\underline{Y}, (t, t+n]}$

(Recall that $\mathcal{F}_{\underline{Y}, (t, t+n]}$ denotes the σ -algebra generated by the cylinders in \underline{Y} determined by $(t, t+n]$.) and if

$$\begin{aligned} \underline{x}(t-m, t+n] &= \underline{x}'(t-m, t+n] \\ &= (x'_{t-m+1}, \dots, x'_{t+n}) \end{aligned}$$

for t an integer and n a nonnegative integer, then

$$\nu(\underline{x}, \underline{E}) = \nu(\underline{x}', \underline{E}).$$

Here the message segment $\underline{x}(t, t+n]$ depends also on the m letters immediately preceding x_t . The smallest such m is called the length of the memory; if no such m exists the channel is said to have infinite memory. In example 1.1 we have $m=0$.

(b) (Takano) We will say K is independent from the remote past if there exists a positive integer m , such that if $\underline{E} \in \mathcal{F}_{\underline{Y}, (i, j]}$ and $\underline{F} \in \mathcal{F}_{\underline{Y}, (l, k]}$ with $i < j$, $l < k$, and $j+m < l$ then

$$\nu(\underline{x}, \underline{E} \wedge \underline{F}) = \nu(\underline{x}, \underline{E}) \nu(\underline{x}, \underline{F}) \quad \lambda.a.a.x.$$

The smallest such m gives the order of remoteness. In example 1.1 we have $m=0$.

(c) (Adler) K is asymptotically independent from the remote past if for any two cylinders $\underline{E}, \underline{F} \in \mathcal{F}_Y$

$$\lim_{n \rightarrow \infty} [v(\underline{x}, T^n \underline{E} \wedge \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})] = 0 \quad \lambda.a.a.\underline{x}.$$

Khinchin stated that if an ergodic message space is fed through a channel with finite memory then the output message space and the compound message space are both ergodic, i.e. using the terminology of [8, p.87], an ergodic input is "admissible" with respect to a channel with finite memory. However definition (a) is insufficient to give the result.

It was first proved by Takano by strengthening the definition to include (b) as well. The result we prove, due to Adler [9], uses only (c) of which (b) is obviously a special case. It says that any ergodic input is admissible with respect to a channel which is asymptotically independent from the remote past.

Theorem 2.2 Let $(\underline{X}, \mathcal{F}_X, \lambda)$ be an ergodic stationary message space and $(\underline{X}, \underline{Y}, v)$ a stationary channel asymptotically independent from the remote past. Then the compound message space and the output message space are both ergodic.

Proof. By Theorem 1.2, to prove ω is ergodic it suffices to show

$$\lim_{n \rightarrow \infty} \omega(T^n \underline{A} \wedge \underline{B}) = \omega(\underline{A})\omega(\underline{B})$$

for $\underline{A}=\underline{C}\times\underline{E}$, $\underline{B}=\underline{D}\times\underline{F}$ with \underline{C} , \underline{D} cylinders in $F_{\underline{X}}$ and \underline{E} , \underline{F} cylinders in $F_{\underline{Y}}$.

$$\begin{aligned} & \frac{1}{N} \sum_{n=0}^{N-1} \int_{T^n \underline{C} \cap \underline{D}} v(\underline{x}, T^n \underline{E}) v(\underline{x}, \underline{F}) \lambda(d\underline{x}) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} \int_{\underline{X}} v(\underline{x}, T^n \underline{E}) 1_{\underline{C}}(T^{-n} \underline{x}) v(\underline{x}, \underline{F}) 1_{\underline{D}}(\underline{x}) \lambda(d\underline{x}). \end{aligned}$$

Now, by the stationarity of the channel

$$= \frac{1}{N} \sum_{n=0}^{N-1} \int_{\underline{X}} v(T^{-n} \underline{x}, \underline{E}) 1_{\underline{C}}(T^{-n} \underline{x}) v(\underline{x}, \underline{F}) 1_{\underline{D}}(\underline{x}) \lambda(d\underline{x}),$$

which

$$\begin{aligned} & \rightarrow \int_{\underline{X}} v(\underline{x}, \underline{E}) 1_{\underline{C}}(\underline{x}) \lambda(d\underline{x}) \int_{\underline{X}} v(\underline{x}, \underline{F}) 1_{\underline{D}}(\underline{x}) \lambda(d\underline{x}) \quad \text{as } N \rightarrow \infty \\ &= \omega(\underline{A}) \omega(\underline{B}). \end{aligned}$$

This follows from the functional form of ergodicity, Theorem 1.3, with

$$\begin{aligned} f(T^{-n} \underline{x}) &= v(T^{-n} \underline{x}, \underline{E}) 1_{\underline{C}}(T^{-n} \underline{x}) \quad \text{and} \\ g(\underline{x}) &= v(\underline{x}, \underline{F}) 1_{\underline{D}}(\underline{x}). \end{aligned}$$

Both are functions in L_2 . We note that T is invertible so that we can take T^{-n} as well as T^n . Now by the asymptotic independence from the remote past we have

$$[v(\underline{x}, T^n \underline{E} \cap \underline{F}) - v(\underline{x}, \underline{F}) \cdot v(\underline{x}, T^n \underline{E})] \rightarrow 0 \quad \lambda.a.a.\underline{x}$$

as $n \rightarrow \infty$. A fortiori

$$[v(\underline{x}, T^n \underline{E} \cap \underline{F}) - v(\underline{x}, \underline{F}) \cdot v(\underline{x}, T^n \underline{E})] 1_{T^n \underline{C} \cap \underline{D}}(\underline{x}) \rightarrow 0$$

$\lambda.a.a.\underline{x}$ as $n \rightarrow \infty$.

From the Lebesgue Dominated Convergence Theorem

$$\lim_{n \rightarrow \infty} \int_{T^n \underline{C} \cap \underline{D}} [v(\underline{x}, T^n \underline{E} \cap \underline{F}) - v(\underline{x}, T^n \underline{E}) v(\underline{x}, \underline{F})] \lambda(d\underline{x}) = 0,$$

and since convergence is stronger than Cesaro convergence

$$C \lim_{n \rightarrow \infty} \int_{T^n \underline{C} \cap \underline{D}} [v(\underline{x}, T^n \underline{E} \cap \underline{F}) - v(\underline{x}, T^n \underline{E}) v(\underline{x}, \underline{F})] \lambda(d\underline{x}) = 0$$

λ .a.a. \underline{x} . Combining these limits we obtain

$$\begin{aligned} & \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} \omega(T^n \underline{A} \cap \underline{B}) - \omega(\underline{A}) \omega(\underline{B}) \\ &= \lim_{N \rightarrow \infty} \left[\frac{1}{N} \sum_{n=0}^{N-1} \int_{T^n \underline{C} \cap \underline{D}} v(\underline{x}, T^n \underline{E} \cap \underline{F}) \lambda(d\underline{x}) \right. \\ & \quad \left. - \frac{1}{N} \sum_{n=0}^{N-1} \int_{T^n \underline{C} \cap \underline{D}} v(\underline{x}, T^n \underline{E}) v(\underline{x}, \underline{F}) \lambda(d\underline{x}) \right. \\ & \quad \left. + \frac{1}{N} \sum_{n=0}^{N-1} \int_{T^n \underline{C} \cap \underline{D}} v(\underline{x}, T^n \underline{E}) v(\underline{x}, \underline{F}) \lambda(d\underline{x}) - \omega(\underline{A}) \omega(\underline{B}) \right] \\ &= 0. \end{aligned}$$

The result for the output message space follows by letting $\underline{C} = \underline{D} = \underline{X}$. //.

We now make some further definitions. For the following we assume that all measures and channels are stationary. We let $K = (\underline{X}, \underline{Y}, \nu)$ be the channel.

(1) K is SM if it is asymptotically independent from the remote past.

(2) K is WM if for any two cylinders $\underline{E}, \underline{F} \in \mathcal{F}_{\underline{Y}}$ there is a sequence of positive integers J , with density zero such that for all $n \in J^c$

$$(2.3) \quad [v(\underline{x}, T^n \underline{E} \cap \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})] \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \lambda.\text{a.a.}\underline{x}.$$

(the density of J is the limit of the ratio $1/n \cdot \#(J \cap (0, n])$)

By for all $n \in J^c$, we mean that if n is restricted to being in J^c then (2.3) holds. A condition equivalent to (2.3) is

$$(2.4) \quad \text{Clim}_{n \rightarrow \infty} |v(\underline{x}, T^n \underline{E} \cap \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})| = 0 \quad \lambda.\text{a.a.}\underline{x}.$$

This equivalence is an immediate consequence of the following lemma; see [6] for proof.

Lemma 2.3 If $\{a_n\}$ is a bounded sequence of real numbers then

$$\text{Clim}_{n \rightarrow \infty} |a_n - a| = 0$$

iff there exists a sequence J , of positive integers such that the density of J is zero and such that if $n \in J^c$ then

$$\lim_{n \rightarrow \infty} a_n = a.$$

(3) K is ergodic if for all $\underline{E}, \underline{F} \in \mathcal{F}_{\underline{Y}}$

$$\text{Clim}_{n \rightarrow \infty} [v(\underline{x}, T^n \underline{E} \cap \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})] = 0 \quad \lambda.\text{a.a.}\underline{x}.$$

We can formulate the definitions of chapter one for the input measure λ as well.

(1') λ is SM if for all $\underline{C}, \underline{D} \in \mathcal{F}_{\underline{X}}$

$$\lim_{n \rightarrow \infty} [\lambda(T^n \underline{C} \cap \underline{D}) - \lambda(\underline{C})\lambda(\underline{D})] = 0.$$

$$(2') \lambda \text{ is WM if for all } \underline{C}, \underline{D} \in \mathcal{F}_{\underline{X}}$$

$$\lim_{n \rightarrow \infty} |\lambda(T^n \underline{C} \cap \underline{D}) - \lambda(\underline{C})\lambda(\underline{D})| = 0.$$

Equivalently, if there exists J , a sequence of positive integers with density zero, such that if $n \in J^c$

$$\lambda(T^n \underline{C} \cap \underline{D}) \rightarrow \lambda(\underline{C})\lambda(\underline{D}).$$

$$(3') \lambda \text{ is ergodic if for all } \underline{C}, \underline{D} \in \mathcal{F}_{\underline{X}}$$

$$\lim_{n \rightarrow \infty} [\lambda(T^n \underline{C} \cap \underline{D}) - \lambda(\underline{C})\lambda(\underline{D})] = 0.$$

The theorem that follows is an extension of our main result (Theorem 2.2). We employ the same technique and notation used in the proof of the latter.

Theorem 2.4 If $(\underline{X}, \mathcal{F}_{\underline{X}}, \lambda)$ is a stationary ergodic message space, and if $(\underline{X}, \underline{Y}, \nu)$ is a stationary WM channel, then the compound message space and the output message space are both ergodic. Thus, any ergodic input is admissible with respect to a channel which is WM.

Proof. For an ERG input, we have from the proof of Theorem 2.2 that

$$\lim_{n \rightarrow \infty} \int_{T^n \underline{C} \cap \underline{D}} \nu(\underline{x}, T^n \underline{E}) \nu(\underline{x}, \underline{F}) \lambda(d\underline{x}) = \omega(\underline{A})\omega(\underline{B}).$$

By the WM property of the channel, we have that there exists J , a sequence of positive integers with density zero, such

that for all $n \in J^c$

$$\lim_{n \rightarrow \infty} [v(\underline{x}, T^n \underline{E} \wedge \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})] = 0 \quad \lambda.a.a.\underline{x}.$$

Then for all $n \in J^c$,

$$\lim_{n \rightarrow \infty} \int_{T^n \underline{C} \wedge \underline{D}} [v(\underline{x}, T^n \underline{E} \wedge \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})] \lambda(d\underline{x}) = 0$$

$\lambda.a.a.\underline{x}$. Now by Lemma 2.3

$$\text{Clim}_{n \rightarrow \infty} \left| \int_{T^n \underline{C} \wedge \underline{D}} [v(\underline{x}, T^n \underline{E} \wedge \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})] \lambda(d\underline{x}) \right| = 0$$

$\lambda.a.a.\underline{x}$. Now

$$\text{Clim}_{n \rightarrow \infty} \int_{T^n \underline{C} \wedge \underline{D}} [v(\underline{x}, T^n \underline{E} \wedge \underline{F}) - v(\underline{x}, T^n \underline{E})v(\underline{x}, \underline{F})] \lambda(d\underline{x}) = 0$$

$\lambda.a.a.\underline{x}$. Here we have used the fact that $\text{Clim}_{n \rightarrow \infty} |a_n - a| = 0$

implies that $\text{Clim}_{n \rightarrow \infty} a_n = a$, [6, p.38]. To complete the proof we

combine limits as in the proof of Theorem 2.2. //.

To complete this chapter we summarize some of the results we have proven together with related results from [9]. For the latter, the proofs mimic the proof of Theorem 2.4, using the appropriate functional forms equivalent to SM, WM, or ERG.

	Channel	
Input $(\underline{X}, \underline{Y}, \nu)$		
$(\underline{X}, \underline{F}_{\underline{X}}, \lambda)$	WM	SM
ERG	ERG	ERG
WM	WM	WM
SM	WM	SM

Here the entry inside the table gives the property of the compound (and output) message space, given the properties of the input (row entry) and of the channel (column entry).

Note that the method of proof we have employed is not applicable when the channel is only assumed to be ergodic. This is because in the ergodic case we have no characterization analogous to (2.3).

These statements summarized in the table, are the strongest possible in the general case. This can be seen by setting

$\nu(\underline{x}, \underline{E}) = \mu(\underline{E})$ for all $\underline{x} \in \underline{X}$, with μ some probability on $F_{\underline{Y}}$. Then

$\omega(\underline{D} \times \underline{F}) = \lambda(\underline{D}) \mu(\underline{F})$ for $\underline{D} \in F_{\underline{X}}$ and $\underline{F} \in F_{\underline{Y}}$; that is, ω is the direct product measure $\lambda \times \mu$. It can be shown that ω will take on the weaker of the properties (SM, WM, or ERG) of λ or of μ .

CHAPTER III: CHANNELS WITH ASYMPTOTICALLY DECREASING
MEMORY AND ANTICIPATION

§1 DEFINITIONS AND EXAMPLES

Let $K=(\underline{X}, \underline{Y}, v)$ be a channel. We consider the following conditions on m_1, m_2 (integers or $\pm\infty$):

a) for all finite intervals $(t_1, t_2]$ and each cylinder $\underline{D} \in \underline{Y}$ determined by this interval, we have

$$v(\underline{x}, \underline{D}) = v(\underline{x}', \underline{D}) \quad \text{whenever}$$

$$\underline{x}(t_1 - m_1, t_2 + m_2] = \underline{x}'(t_1 - m_1, t_2 + m_2].$$

b) For all pairs $(t_1, t_2]$ and $(t_1', t_2']$ of finite intervals and all pairs \underline{D} and \underline{D}' of cylinders in \underline{Y} that are determined by these intervals (respectively), we have for all $\underline{x} \in \underline{X}$ (or λ .a.a. \underline{x} if there is a probability λ on \underline{X})

$$v(\underline{x}, R^k \underline{D} \cap R^{k'} \underline{D}') = v(\underline{x}, R^k \underline{D}) v(\underline{x}, R^{k'} \underline{D}')$$

(Here R is the right shift, i.e. the inverse of the shift T as defined in chapter 1.)

whenever k and k' are integers such that

$$k' - k > t_2 - t_1' + m_1.$$

That is, whenever the separation between the intervals determining the cylinders $R^k \underline{D}$ and $R^{k'} \underline{D}'$ is greater than m_1 .

Now the infimum m_a of all m_2 satisfying a) for some

m_1 , is called the anticipation of the channel K . The infimum m_m of all m_1 satisfying a) and b) for some m_2 , is called the memory of K . Thus a pair m_1, m_2 satisfies a) and b) iff $m_1 > m_m$ and $m_2 > m_a$.

We will say K has asymptotically decreasing memory and anticipation iff

Condition 1) for all finite intervals $(t_1, t_2]$ and each output event $\underline{D} \in \mathcal{F}_Y, (t_1, t_2]$ that occurs in this interval, we have

$$\lim_{\substack{m_1 \rightarrow \infty \\ m_2 \rightarrow \infty}} v(x'(-\infty, t_1 - m_1] \times (t_1 - m_1, t_2 + m_2] \times (t_2 + m_2, \infty], \underline{D})) \\ = v(\underline{x}, \underline{D})$$

Uniformly with respect to $t_1, t_2, \underline{E}, \underline{x}$ and \underline{x}' , for all $\underline{x}, \underline{x}' \in \underline{X}$ (or $\lambda.a.a. \underline{x}$ and \underline{x}'). Here

$$x'(-\infty, t_1 - m_1] \times (t_1 - m_1, t_2 + m_2] \times (t_2 + m_2, \infty] \\ = (\dots, x'_{t_1 - m_1}, x'_{t_1 - m_1 + 1}, \dots, x'_{t_2 + m_2}, x'_{t_2 + m_2 + 1}, \dots)$$

Another formulation of condition 1) is the following: for any positive ε there exist integers m_1, m_2 such that for any finite interval $(t_1, t_2]$ and every output event \underline{D} occurring in this interval, we have

$$|v(\underline{x}, \underline{D}) - v(\underline{x}', \underline{D})| \leq \varepsilon \quad \text{whenever}$$

$$x(t_1 - m_1, t_2 + m_2] = x'(t_1 - m_1, t_2 + m_2]$$

Condition 2) For each pair of cylinders $\underline{C}, \underline{C}'$ in \underline{Y}

$$\lim_{k' - k \rightarrow \infty} [v(\underline{x}, R^k \underline{C} \cap R^{k'} \underline{C}') - v(\underline{x}, R^k \underline{C}) v(\underline{x}, R^{k'} \underline{C}')] = 0$$

for all $\underline{x} \in \underline{X}$ (or $\lambda.a.a.\underline{x}$).

This condition requires that two output messages that occur in different intervals given (almost) any \underline{x} , are almost independent if these intervals are sufficiently far apart.

It is clear that any channel with finite memory and anticipation satisfies conditions 1) and 2).

We will now construct examples.

Let X and Y be finite nonempty sets. For all $x \in X$ let $p(y|x)$ be a probability over Y . Furthermore, assume $p(y|x)$ is dependent on x . Otherwise the output would be independent of the input and no information could be passed through the channel.

Let $\varepsilon(\tau) \geq 0$ for all $\tau \in \mathbb{Z}$ (Recall that \mathbb{Z} is the set of all integers.), and

$$\sum_{\tau \in \mathbb{Z}} \varepsilon(\tau) = 1. \text{ Define}$$

$$v_t(\underline{x}, \underline{D}) = \sum_{\tau} p(D_t | x_{t-\tau}) \varepsilon(\tau) \text{ for } t \in \mathbb{Z}.$$

This is a probability over \underline{Y} for all \underline{x} in \underline{X} . Now let

$$v(\underline{x}, \underline{D}) = \prod_{t \in \mathbb{Z}} v_t(\underline{x}, \underline{D}) .$$

This is also a probability over \underline{Y} for all $\underline{x} \in X$; see Neveu [12, p.165]. We will now specialize this example.

Example 3.1 Take

$$\varepsilon(\tau) = \begin{cases} 1 & \text{if } \tau=0 \\ 0 & \text{elsewhere.} \end{cases}$$

$$\begin{aligned} v(\underline{x}, \underline{C}) &= \prod_t \left\{ \sum_{\tau} p(C_t | x_{t-\tau}) \varepsilon(\tau) \right\} \\ &= \prod_t p(C_t | x_t) , \end{aligned}$$

for $\underline{x} \in X$ and $\underline{C} \in F_{\underline{Y}}$.

In this case each letter is treated independently;

$$v_t(\underline{x}, \underline{C}) = p(C_t | x_t)$$

is the probability that C_t is received given that x_t is sent through the channel. This is clearly the channel without memory, example 2.1; it has memory and anticipation equal to zero.

Example 3.2 Take

$$\varepsilon(\tau) = \begin{cases} 1/3 & \text{if } \tau=0, \pm 1 \\ 0 & \text{elsewhere.} \end{cases}$$

$$v(\underline{x}, \underline{C}) = \prod_t \frac{1}{3} (p(C_t | x_{t-1}) + p(C_t | x_t) + p(C_t | x_{t+1})) .$$

We claim that this channel (with X and Y as before) has memory and anticipation equal to one.

Proof. (i) To show $m_a=1$. Let $(t_1, t_2]$ be finite; let \underline{C} be a cylinder in \underline{Y} determined by $(t_1, t_2]$. It suffices to show

$$v(\underline{x}, \underline{C}) = v(\underline{x}', \underline{C}) \quad \text{whenever}$$

$$\underline{x}(t_1-1, t_2+k] = \underline{x}'(t_1-1, t_2+k]$$

for k a positive integer. From the definition of v_t it follows that for all $\underline{x} \in X$

$$v_t(\underline{x}, \underline{C}) = 1$$

if $t \notin (t_1-1, t_2+k]$ and it depends on \underline{x}_t if $t \in (t_1-1, t_2+k]$.

Thus the result follows.

(ii) To show $m_m=1$. This is clear since the condition

$$k' - k \geq t_2 - t_1' + 1$$

guarantees that the intervals determining the cylinders $R^k \underline{C}$ and $R^{k'} \underline{C}'$ are separated by at least two. The result then follows from the definition of v_t . //.

From this type of argument we can see that the anticipation corresponds to the number of $\varepsilon(\tau) > 0$ for positive τ and the memory corresponds to the number of $\varepsilon(\tau) > 0$ for negative τ . In this manner we can construct channels with any finite or infinite memory and anticipation.

Example 3.3 Suppose

$$\sum_{\tau} \epsilon(\tau) = 1$$

as before, and furthermore that

$$\sum_{i=0}^{\infty} \sum_{|\tau| > i} \epsilon(\tau) < \infty .$$

With X and Y as before, $K=(\underline{X}, \underline{Y}, v)$ has asymptotically decreasing memory and anticipation. The proof of this fact is lengthy but reasonably straightforward.

§2 INDECOMPOSABLE FINITE STATE CHANNELS

For the remainder of this paper we will assume X and Y are finite, nonempty sets.

Let S be a finite, nonempty set and for all $x \in X$, $s \in S$, let $\gamma(\cdot | s, x)$ be a probability over $S \times Y$. Then $\hat{K} = (X, Y, S, \gamma)$ is called a finite state channel. The elements of S correspond to the possible states of the channel and $\gamma(s' \times y | s, x)$ represents the probability of going to state s' and outputting y , given the input x and present state s .

For $t \in \mathbb{Z}$, $s \in S$ let

$$\sigma(s, t) = \{ \underline{s}' \in \underline{S} : s'_t = s \} \times \underline{Y}.$$

This is in $F_{\underline{S} \times \underline{Y}}(t', \infty]$ for all $t' < t$.

This is the event corresponding to being in state s at time t . If t is an integer and m a positive integer, let \underline{C} be a (thin) cylinder in \underline{S} determined by $(t, t+m]$ and $s(m) = (s_1, \dots, s_m) \in S^m$; let \underline{D} be a cylinder in \underline{Y} determined by $(t, t+m]$ and $y(m) = (y_1, \dots, y_m) \in Y^m$. We define for $s_0 \in S$ and $\underline{x} \in \underline{X}$

$$(3.1) \quad \gamma_t(\underline{C} \times \underline{D} | \sigma(s_0, t), \underline{x}) = \prod_{i=1}^m \gamma(s_i \times y_i | s_{i-1}, x_{t+1}).$$

It follows from the Kolmogorov Extension Theorem [7, p.3] that for all $t \in \mathbb{Z}$, $s \in S$, $\underline{x} \in \underline{X}$, (3.1) induces a unique

probability $\gamma_t(\cdot | \sigma(s,t), \underline{x})$ on $F_{S \times Y, (t, \infty]}$. This probability has the property that if $\underline{E} \in F_{S \times Y, (t_1, t_2]}$ with t_1 finite, then $\gamma_{t_1}(\underline{E} | \sigma(s, t_1), \underline{x})$ depends only on $X_{(t_1, t_2]}$. This follows immediately from (3.1) and the fact that \underline{E} is generated by cylinders determined by $(t_1, t_2]$.

Secondly, if $\underline{D} \in F_{Y, (t, \infty]}$ with t finite then

$$(3.2) \quad \gamma_t(\underline{D} | \sigma(s,t), \underline{x}) = \gamma_{t+1}(\underline{RD} | \sigma(s, t+1), R\underline{x})$$

Again this follows easily from (3.1).

A finite state channel $\hat{K} = (X, Y, S, \gamma)$ is called indecomposable if the following holds:

Condition ID: for all positive ϵ , there exists a positive integer N_0 , such that for some integer t (Hence for all t by (3.2).) and every $s, s', s'' \in S$, $\underline{x} \in X$ (or $\lambda.a.a. \underline{x}$), $m \geq N_0$, we have

$$|\gamma_{t-m}(\sigma(s,t) | \sigma(s', t-m), \underline{x}) - \gamma_{t-m}(\sigma(s,t) | \sigma(s'', t-m), \underline{x})| \leq \epsilon$$

Roughly, this means that the effect of the initial state wears off as time increases.

Given an indecomposable finite state channel $\hat{K} = (X, Y, S, \gamma)$, we can construct a channel $K = (\underline{X}, \underline{Y}, \nu)$ which

is stationary and has zero anticipation. For $\underline{D} \in \mathcal{F}_{\underline{Y}}$, let

$$v(\underline{x}, \underline{D}) = \lim_{m \rightarrow \infty} \gamma_{t-m}(\underline{S} \times \underline{D} | \sigma(s', t-m), \underline{x}) .$$

It can be shown (see [5]) that this limit exists independently of s' and is uniform in $t, \underline{D}, \underline{x}$ and s' .

Now, from [5] we can state the following result:

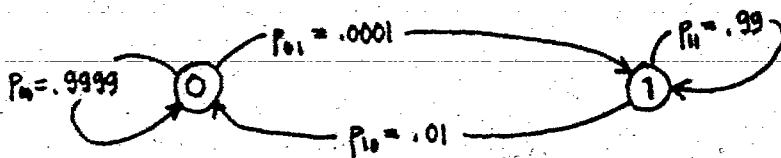
Theorem 3.1 Let $\hat{K} = (X, Y, \hat{S}, \gamma)$ be an indecomposable finite state channel and let $K = (X, Y, v)$ be the channel constructed from \hat{K} . Then K has asymptotically decreasing memory and anticipation.

We can now give examples from Gallager [13], of finite state channels which can be shown to be indecomposable.

Example 3.4 A Simple Model of a Burst-error Channel.

This is an example of a channel where errors tend to cluster together in "bursts". State 0 corresponds to relatively good transmission of data, while state 1 corresponds to the error-prone phase of the channel.

State transition diagram: p_{ij} represents the probability of going from state i to state j .



Input-output probabilities: q_{ij} is the probability of receiving j given that i was sent.

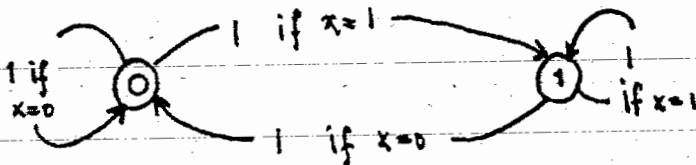
In state 0	In state 1
$q_{00} = .99999$	$q_{00} = .75$
$q_{10} = .00001$	$q_{10} = .25$
$q_{11} = .99999$	$q_{11} = .70$
$q_{01} = .00001$	$q_{01} = .30$

Note that in this example the state transition probabilities do not depend on the input.

Example 3.5 A Simple Model of Intersymbol Interference.

In this example the state at a given time is the same as the input at that time. The probability of an error is greater if $x_n \neq x_{n-1}$ than when $x_n = x_{n-1}$.

State transition diagram



Input-output probabilities:

In state 0

$$q_{00} = .999$$

$$q_{10} = .001$$

$$q_{11} = .990$$

$$q_{01} = .010$$

In state 1

$$q_{00} = .990$$

$$q_{10} = .010$$

$$q_{11} = .999$$

$$q_{01} = .001$$

§3 THE RELATIONSHIP BETWEEN STRONG MIXING AND ASYMPTOTICALLY
DECREASING MEMORY AND ANTICIPATION

Theorem 3.2 Let $K=(\underline{X}, \underline{Y}, \nu)$ be a channel with asymptotically decreasing memory and anticipation. Then K is SM with respect to R and with respect to T .

Proof. By hypothesis, for any two cylinders \underline{C} and \underline{D} in \underline{Y}

$$\lim_{k' \rightarrow \infty} [\nu(\underline{x}, R^k \underline{C} \wedge R^{k'} \underline{D}) - \nu(\underline{x}, R^k \underline{C}) \nu(\underline{x}, R^{k'} \underline{D})] = 0 .$$

Take $k=0$ and let $k' \rightarrow \infty$. Then

$$\lim_{k' \rightarrow \infty} [\nu(\underline{x}, \underline{C} \wedge R^{k'} \underline{D}) - \nu(\underline{x}, \underline{C}) \nu(\underline{x}, R^{k'} \underline{D})] = 0 .$$

This says that K is SM with respect to R . Now take $k'=0$ and let $k \rightarrow -\infty$.

$$\lim_{k \rightarrow -\infty} [\nu(\underline{x}, R^k \underline{C} \wedge \underline{D}) - \nu(\underline{x}, R^k \underline{C}) \nu(\underline{x}, \underline{D})] = 0$$

$$\lim_{k \rightarrow \infty} [\nu(\underline{x}, T^k \underline{C} \wedge \underline{D}) - \nu(\underline{x}, T^k \underline{C}) \nu(\underline{x}, \underline{D})] = 0 .$$

This says that K is SM with respect to T . //.

Furthermore, it appears that asymptotically decreasing memory and anticipation is a strictly stronger property than strong mixing. The former allows the cylinders \underline{C} and \underline{D} to be shifted arbitrarily, only requiring that the distance between the intervals determining these shifted cylinders becomes large. Strong mixing, on the other hand, allows

only one of the cylinders to be shifted, the other being fixed.

Theorem 2 of [5], says that any ergodic input is admissible with respect to a channel with asymptotically decreasing memory and anticipation. In chapter II, Theorem 2.2 (2.4) says that any ergodic input is admissible with respect to a SM (WM) channel. The fact that these two theorems are formulated for the left shift T , rather than the right shift R (as in [5]) presents no problem. This is because if T is invertible, T is ergodic iff T^{-1} is; see [7,p.9].

By the above remarks and Theorem 3.2, we can see that Theorem 2.2, and a fortiori Theorem 2.4, are stronger than Theorem 2 of [5].

§4 CHANNEL CAPACITY

Let $(\underline{Y}, F_{\underline{Y}}, \lambda)$ be an input message space and let $K = (\underline{Y}, \underline{W}, \nu)$ be a channel. We have three shifts

- (i) $R_{\underline{Y} \times \underline{W}}$ with the compound measure ω on $\underline{Y} \times \underline{W}$,
- (ii) $R_{\underline{Y}}$ with the probability λ on \underline{Y} , and
- (iii) $R_{\underline{W}}$ with the output measure μ on \underline{W} .

We define the rate of transmission $R(\lambda)$, of $(\underline{Y}, F_{\underline{Y}}, \lambda)$ over K by

$$R(\lambda) = h(R_{\underline{Y}}) + h(R_{\underline{W}}) - h(R_{\underline{Y} \times \underline{W}})$$

From [7, p.156] we can obtain another formulation.

$$\begin{aligned} R(\lambda) &= \lim_{n \rightarrow \infty} 1/n \{ H(y_0, \dots, y_{n-1}) + H(w_0, \dots, w_{n-1}) \\ &\quad - H(y_0, \dots, y_{n-1}, w_0, \dots, w_{n-1}) \} \\ &= \lim_{n \rightarrow \infty} 1/n \{ H(w_0, \dots, w_{n-1}) \\ &\quad - H(w_0, \dots, w_{n-1} | y_0, \dots, y_{n-1}) \} \\ &= \lim_{n \rightarrow \infty} 1/n \{ H(y_0, \dots, y_{n-1}) \\ &\quad - H(y_0, \dots, y_{n-1} | w_0, \dots, w_{n-1}) \} \end{aligned}$$

This is the amount of information received per letter.

We can make the following interpretation: the mean amount of information gained by receiving the message (y_0, \dots, y_{n-1}) is $H(y_0, \dots, y_{n-1})$. However, the message actually received is (w_0, \dots, w_{n-1}) , from which we try to deduce the original message. The amount of uncertainty

involved in this is $H(y_0, \dots, y_{n-1} | w_0, \dots, w_{n-1})$. So that the amount of information we receive is the amount of information in the message sent, minus the uncertainty about the message sent.

The quantity

$$\begin{aligned} \lim_{n \rightarrow \infty} 1/n H(y_0, \dots, y_{n-1} | w_0, \dots, w_{n-1}) \\ = h(R_Y) - R(\lambda) \\ = h(R_{Y \times W}) - h(R_W) \end{aligned}$$

is called the equivocation per letter. Clearly a low rate of equivocation is desirable; this corresponds to a high rate of transmission.

We can now define the stationary capacity of the channel K by

$$C_s(K) = \sup \{R(\lambda) : \lambda \text{ is such that } R \text{ is a } \lambda\text{-mpt}\} .$$

The ergodic capacity is

$$C_e(K) = \sup \{R(\lambda) : \lambda \text{ is such that } R \text{ is ergodic with respect to } \lambda\} .$$

The rate of transmission and capacity are difficult to compute except for some simple channels. For example, we can compute them for the channel without memory, example

2.1.

$$\text{Let } (c_{jk}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix} .$$

Here it can be shown (see [7,p.159]) that the transmission rate equals the entropy of the input. This channel, although noisy, is called "lossless". It has stationary and ergodic capacity equal to $\log 2$.

For a second example, suppose the rows of (c_{jk}) are all identical. So that c_{jk} is independent of j . Here

$$H(w_0|y_0) = H(w_0) ,$$

and it can be shown that the transmission rate is zero for any input. Of course no information can be sent through this channel. The stationary and ergodic capacity are both zero.

§5 FEINSTEIN'S LEMMA

This section deals with Pfaffelhuber's version (from [5]) of Feinstein's lemma. The definitions are from [8] and the proof is a variation of Feinstein's proof.

We define a maximal set $M[\epsilon]$, (for $0 < \epsilon < 1/2$) in a channel (X, Y, p) as a set x_1, \dots, x_N ($N > 1$) of elements of X such that:

(i) To each x_i there corresponds a B_i in Y such that

$$p(x_i, B_i) \geq 1 - \epsilon .$$

(ii) The B_i are disjoint.

(iii) There exist no x_{N+1} and B_{N+1} such that x_1, \dots, x_{N+1} and B_1, \dots, B_{N+1} satisfy conditions (i) and (ii).

Given a channel (X, Y, p) with input measure λ , a set $M[\epsilon, K]$ (for $0 < \epsilon, K < 1$) of elements of X , $\{x_1, \dots, x_N\}$, ($N > 1$) will be called an enlarging set with respect to λ if to each x_i there corresponds an A_i in Y such that

(i) $p(x_i, A_i) \geq 1 - \epsilon$ and

(ii) $\lambda(A_i) < K$,

for $i=1, \dots, N$.

A maximal set $M[e]$ will be said to be L -bounded (for $0 < L < 1$) with respect to an input measure λ if the B_i satisfy

$$\lambda(B_i) < L ,$$

for $i=1, \dots, N$ and $N \geq 1$.

Theorem 3.3 Let $K=(X, Y, v)$ be a stationary channel which has ergodic capacity C . Assume further that the input measure λ being ergodic implies that ω and μ are ergodic as well. If H and e are positive and $H < C$, then there exists a positive integer $n(e, H)$ such that if

$$n \geq n(e, H)$$

there exist U_1, \dots, U_N in X^n with

$$N \geq B^{nH}$$

and disjoint sets A_1, \dots, A_N in Y^n such that

$$p(A_i | u_i) \geq 1 - e$$

for $i=1, \dots, N$.

(The meaning of $p(A_i | u_i)$ will be clarified in the proof. The following proof is a summary of the proof given by Feinstein; see [8] for the details.)

Proof. If $H < C$, there exists an ergodic input measure λ such that

$$h(R_X) + h(R_Y) - h(R_{X \times Y}) > H .$$

Let us denote the elements of X^n and Y^n by u and v

respectively.

Given positive ε and δ , let V' be the set of (u,v) satisfying

$$|1/n(\log \omega(u \times v)) + h(R_{X \times Y})| < \varepsilon/2 .$$

Let S' be the set of v satisfying

$$|1/n(\log \mu(v)) + h(R_Y)| < \varepsilon/2 .$$

By assumption ω and μ are ergodic since λ is. We can now apply Theorem 1.10. Thus for sufficiently large n , given positive δ_1

$$\omega(V') > 1 - \delta_1/2$$

and

$$\begin{aligned} \omega(X^n \times S') &= \mu(S') \\ &> 1 - \delta_1/2 . \end{aligned}$$

Let

$$V = V' \cap (X^n \times S') .$$

$$(3.3) \quad \omega(V) > 1 - \delta_1 .$$

Hence for any $(u,v) \in V$

$$|1/n \log p(u|v) + h(R_{X \times Y}) - h(R_Y)| < \varepsilon .$$

Here

$$p(u|v) = \omega(u \times v) / \mu(v) .$$

Given $\varepsilon', \delta_2 > 0$, let

$$S_0 = \{u \in X^n : |1/n(\log \lambda(u)) + h(R_X)| < \varepsilon'\} .$$

Applying Theorem 1.10, we obtain that for sufficiently large n

$$(3.4) \quad \lambda(U_0) > 1 - \delta_2 .$$

Now take n large enough so that (3.3) and (3.4) hold.

To each $u \in U_0$, let

$$A_u = \{v \in Y^n : (u, v) \in V\} .$$

For $0 < a < 1$, let

$$U_1 = \{u \in U_0 : p(A_u | u) \geq 1 - a\} .$$

By Lemma 2 [8, p.48],

$$\lambda(U_1) > 1 - \delta_2 - \delta_1 / a .$$

It is easy to see that for $u \in U_0$,

$$\lambda(u) < B^{-n(h(R_X) - \epsilon')} ,$$

and for $(u, v) \in V$,

$$p(u | v) > B^{-n(h(R_X | R_Y) + \epsilon)} .$$

(Here $h(R_X | R_Y)$ is the equivocation $h(R_{X \times Y}) - h(R_Y)$.)

Let $u \in U_1$ and let $v \in A_u$, then

$$\begin{aligned} p(u | v) / \lambda(u) &> B^{-n(h(R_X | R_Y) + \epsilon)} / B^{-n(h(R_X) - \epsilon')} \\ &= B^{n(R(\lambda) - \epsilon - \epsilon')} . \end{aligned}$$

For notational convenience, let

$$d = n(R(\lambda) - \epsilon - \epsilon') .$$

We have now that

$$\omega(u \times v) / \lambda(u) > B^d \mu(v) .$$

Summing over A_u ,

$$(3.5) \quad \omega(\bar{u} \times A_u) / \lambda(u) > B^d \mu(A_u) .$$

The lefthand side of (3.5) is bounded by 1, hence

$$\mu(A_u) < B^{-d} .$$

Hence, U_1 is an enlarging set $M[a, B^{-d}]$ relative to λ , assuming U_1 is not empty. Let

$$M[e] = \{x_1, \dots, x_N\}$$

be a (B^{-d}) -bounded maximal set relative to λ . Then $U_1 - U_1 \cap M[e]$ is still an enlarging set $M[a, B^{-d}]$ disjoint from $M[e]$. Now the inequality of [8, p.46] yields

$$\begin{aligned} NB^{-d} &> (e-a)\lambda(U_1 - U_1 \cap M[e]) + (1-e)\lambda(M[e]) \\ &\geq \min(e-a, 1-e)(1-\delta_2 - \delta_1/a) \end{aligned}$$

Given e and H , choose λ such that

$$C > R(\lambda) > H$$

Let (taking n sufficiently large):

$$\begin{aligned} a &= e/2 ; \\ \epsilon + \epsilon' &= (R(\lambda) - H)/2 ; \\ 1 - \delta_2 - (2\delta_1)/e &\geq 1/2 ; \\ m &= \min(e/2, 1-e) \end{aligned}$$

Then $0 < m < 1$, and

$$\begin{aligned} N &> (m/2) B^{n(H + (R(\lambda) - H)/2)} \\ &= B^{n(H + (R(\lambda) - H)/2 + (\log(m/2))/n)} \\ &\geq B^{nH} \quad // \end{aligned}$$

Theorem 3.4 Let $K = (\underline{X}, \underline{Y}, \nu)$ be a stationary channel which satisfies condition 1). Let $C > 0$ be its ergodic capacity.

Let λ be such that λ ergodic implies ω and μ are. Then for any $\epsilon > 0$, there exist $m_0, m_1, m_2 \in \mathbb{Z}$ such that if $m \geq m_0$, then there exists an integer N satisfying

$$N > B^{m(C-\epsilon)}$$

and N distinct input messages $x_M^{(1)}, \dots, x_M^{(N)}$ of length

$$M = m + m_1 + m_2$$

and N disjoint groups $V^{(1)}, \dots, V^{(N)}$ of output messages y_m of length m , such that for any integers t_1, t_2 with

$$m = t_2 - t_1$$

and for all $i = 1, \dots, N$, we have

$$P(\underline{x}'_{(t_1, t_2]} \in V^{(i)}) \geq 1 - \epsilon$$

whenever

$$\underline{x}'_{(t_1 - m_1, t_2 + m_2]} = x_M^{(i)}.$$

Proof. By Theorem 3.3, we can choose a probability λ on F_X making $(\underline{X}, F_X, \lambda)$ an ergodic message space. Furthermore, we can choose an integer m_0 such that for any $m \geq m_0$ there exists an integer N satisfying

$$N > B^{m(C-\epsilon)},$$

such that there exist N input messages of length m , $\{u_1, \dots, u_N\}$, and N disjoint groups $V^{(1)}, \dots, V^{(N)}$ of output messages of length m . These have the property that for $i = 1, \dots, N$

$$P(\underline{E}^{(i)} | \underline{\xi}^{(i)}) \geq 1 - \epsilon/2.$$

Here

$$\underline{\xi}^{(i)} = \{\underline{x} \in X : x_{(0, m]} = u_i\},$$

i.e. the cylinder determined by $(0, m]$ and u_i . Similarly

$$\underline{E}^{(i)} = \{\underline{y} \in Y : y_{(0, m]} \in V^{(i)}\}.$$

$$(3.6) \quad p(\underline{E}^{(i)} | \underline{\xi}^{(i)}) = \int_{\underline{\xi}^{(i)}} v(\underline{x}, \underline{E}^{(i)}) \lambda(d\underline{x}) / \mu(\underline{\xi}^{(i)})$$

$$\geq 1 - \epsilon/2.$$

Hence, there exists $\underline{x}^{(i)} \in \underline{\xi}^{(i)}$ such that

$$(3.7) \quad v(\underline{x}^{(i)}, \underline{E}^{(i)}) \geq 1 - (3\epsilon)/4,$$

otherwise (3.6) cannot hold.

By condition 1), we can choose m_1, m_2 such that

$$(3.8) \quad |v(\underline{x}^{(i)}, \underline{E}^{(i)}) - v(\underline{x}', \underline{E}^{(i)})| < \epsilon/4$$

whenever

$$\underline{x}^{(i)}(t_1 - m_1, t_2 + m_2] = \underline{x}'(t_1 - m_1, t_2 + m_2]$$

for all $i=1, \dots, N$.

Now define

$$\underline{x}_M^{(i)} = \underline{x}^{(i)}(-m_1, m+m_2].$$

Then (3.7) and (3.8) yield that

$$v(\underline{x}, \underline{y}(0, m] \in V^{(i)}) \geq 1 - \epsilon$$

if

$$\underline{x}'(-m, m+m_2] = \underline{x}_M^{(i)}.$$

The result follows by the stationarity of the channel. //.

Using the terminology of [2], Theorem 3.4 says that there exists a distinguishable group of

$$N > B^m(C - \epsilon)$$

input sequences $\underline{x}_M^{(1)}, \dots, \underline{x}_M^{(N)}$ of length

$$M = m + m_1 + m_2.$$

We conclude this chapter with one application of Theorem 3.4; see [7,p.174] for proof.

Theorem 3.5 We assume the hypotheses of Theorem 3.4 and furthermore that the input measure λ satisfies

$$(3.9) \quad \lambda\{(x_1, \dots, x_m) \in \{x_M^{(1)}, \dots, x_M^{(N)}\}\} = 1$$

(i.e. with probability one the transmitted message is one of the $x_M^{(i)}$)

then

$$H(x_1, \dots, x_m | y_1, \dots, y_m) < \eta(\epsilon) + \eta(1-\epsilon) + \epsilon \log(\#X)^m$$

if

$$\epsilon < 1/B \quad .$$

CHAPTER IV: THE CODING THEOREM

§1 THE COMPOUND CHANNEL

We now define formally the concept of a code ϕ , as a measurable mapping $\underline{X} \rightarrow \underline{Y}$. A code is said to be stationary if

$$\phi R_{\underline{X}} = R_{\underline{Y}} \phi$$

This means that the structure of the coding device is time-invariant.

A code is nonanticipating if for all $i \in \underline{Y}$

$$\phi^{-1}\{\underline{y}: y_n = i\} = \{\underline{x}: (\phi \underline{x})_n = i\}$$

is in the σ -field generated by $\{\dots, x_{n-1}, x_n\}$. This says that the coding device need not be clairvoyant.

If $K = (\underline{Y}, \underline{W}, \nu)$ is a channel, then a code $\phi: \underline{X} \rightarrow \underline{Y}$ gives rise to a compound channel.

$$K_C = (\underline{X}, \underline{W}, \nu(\phi(\cdot), \cdot))$$

This follows from the fact that $\nu(\phi(\cdot), \cdot)$ is a kernel if ν is. The measure P on $\underline{X} \times \underline{Y} \times \underline{W}$ is given by

$$P(\underline{A} \times \underline{B} \times \underline{C}) = \int_{[\underline{B}]} \nu(\phi(\underline{x}), \underline{C}) \lambda(d\underline{x})$$

with $\underline{A} \in \mathcal{F}_{\underline{X}}$, $\underline{B} \in \mathcal{F}_{\underline{Y}}$, and $\underline{C} \in \mathcal{F}_{\underline{W}}$.

§2 AN EXAMPLE OF CODING TECHNIQUES

Shannon's first version of the coding theorem in [1], in part says that for a certain class of channels we can choose a sequence of codes giving us increasingly small probability of error. To clarify this we will give a simple example of how we could encode input messages in this manner. This question of choosing codes to fit a certain channel is one of considerable practical importance.

In our example we take the binary symmetric channel without memory. So $Y=W=\{0,1\}$ and

$$(c_{jk}) = \begin{bmatrix} .95 & .05 \\ .05 & .95 \end{bmatrix} .$$

Suppose further that the input alphabet is $\{a,b,c,d\}=X$; F_X is the set of all subsets of X ; F_Y and F_W both equal the set of all subsets of $\{0,1\}$; the input measure λ , is such that $\lambda(x) > 0$ for all $x \in X$.

We now construct the following codes:

$C_1: X \rightarrow Y^2$ such that

$$C_1(a) = 00$$

$$C_1(b) = 01$$

$$C_1(c) = 10$$

$$C_1(d) = 11 .$$

With this code the probability of the incorrect transmission of one input letter is clearly

$$(.95)^2 \approx .90 .$$

Let $C_2: X \rightarrow Y^6$ such that

$$C_2(a) = 000000 = d_1$$

$$C_2(b) = 010101 = d_2$$

$$C_2(c) = 101010 = d_3$$

$$C_2(d) = 111111 = d_4 .$$

The code C_2 is an example of an "error-correcting" code. Whereas using C_1 we would lose information if one error in transmission occurred, this code C_2 , gives us the capability of recovering from errors. Using the definitions of Van Lint [14] we will call the number of ones in a particular finite sequence from Y the Hamming weight, HW. The Hamming distance HD, between two elements of Y^n is the number of coordinates where they do not agree. For example, if $x = 100000$ and $y = 011110$ then $HW(x) = 1$, $HW(y) = 4$, and $HD(x, y) = 5$.

Upon reception of a particular output sequence we use maximum likelihood decoding to determine what the most probable input sequence was. For example, if we receive $d = 001000$ we will infer that the letter a was sent. That is we determine the d_i for which $HD(d_i, d)$ is minimal. In our

example the Hamming distance between any two of the d_i is at least three. From this it follows that if any one error occurs in the transmission we can still decode correctly, in some cases even if two errors have occurred. Hence, the probability of error in the transmission of one letter is

$$\begin{aligned}
 &< 1 - (\text{probability of no errors}) - (\text{probability of} \\
 &\quad \text{exactly one error}) \\
 &= 1 - [(.95)^6 + 6(.05)(.95)^5] \\
 &\approx .96
 \end{aligned}$$

Hence with this second coding scheme we can reduce the probability of error. It is of course possible to create longer codes which have a smaller probability of error. But note that in this way we are making the transmission error smaller by employing redundancy in the coding. At the same time we are making the transmission rate lower. The strength of Shannon's Theorem is that one can reduce the number of errors in transmission while retaining a rate of transmission arbitrarily close to channel capacity.

We now prove this result for a wide class of channels.

§3 BLOCK CODES

We assume the following:

- (i) $(\underline{X}, F_{\underline{X}}, \lambda)$ is an ergodic, stationary input message space with entropy h ;
- (ii) (\underline{Y}, W, ν) is a stationary, WM channel which satisfies condition 1) and has ergodic capacity C ;
- (iii) X, Y and W are finite alphabets;
- (iv) $\#X = r$;
- (v) $\#Y = t$.

For all positive integers b and all integers n , let

$$\bar{x}_n = (x_{nb+1}, \dots, x_{nb+b}) \quad \text{and}$$

$$\bar{x} = (\dots, \bar{x}_{-1}, \bar{x}_0, \bar{x}_1, \dots) .$$

Here \bar{x} is an element of the space \bar{X} of doubly-infinite sequences from X^b . We will similarly define \bar{y} , \bar{w} , \bar{Y} and \bar{W} .

We define a b -block code $\phi: \underline{X} \rightarrow \underline{Y}$ determined by a stationary code $\bar{\phi}: \bar{X} \rightarrow \bar{Y}$ with

$$((\phi x)_{nb+1}, \dots, (\phi x)_{nb+b}) = (\bar{\phi} \bar{x})_n .$$

Note that ϕ is not strictly stationary, but is stationary in blocks of b , i.e.

$$\phi R^b \underline{x} = R^b \phi \underline{x} .$$

A b -block code ϕ is nonanticipating if $\bar{\phi}$ is.

Since ϕ is not stationary the original definitions of rate, equivocation, etc., are not immediately applicable.

We will define these quantities for \underline{x} , \underline{y} and \underline{w} by dividing by b the corresponding quantities for \bar{x} , \bar{y} and \bar{w} .

§4 THE CODING THEOREM BY BLOCK CODES

Theorem 4.1 We assume conditions (i) to (v) of §4.3 hold. If $h < C$ and $\delta > 0$ there exists for some positive integer b , a b -block code ϕ , such that if \underline{x} is transmitted through the compound channel then the rate of transmission exceeds $h - \delta$.

Proof. Choose ϵ such that:

$$(4.1) \quad h + \epsilon < C - \epsilon$$

$$(4.2) \quad \eta(\epsilon) + \eta(1 - \epsilon) + \epsilon \log t < \delta/2$$

(This is possible since $\eta(0) = \eta(1) = 0$.)

$$(4.3) \quad \epsilon \log r < \delta/2$$

$$(4.4) \quad \epsilon < 1/B$$

In what follows, P is the probability on the compound message space as defined in §4.1, and ϕ is the b -block code to be constructed.

Take $b > b_0(\epsilon)$, with $b_0(\epsilon)$ as given in Theorem 1.9. Hence X^b can be partitioned into two sets H (the high probability group) and L (the low probability group) such that

$$(4.5) \quad P\{(x_1, \dots, x_b) \in L\} < \epsilon$$

and

$$(4.6) \quad P\{(x_1, \dots, x_b) = u\} > B^{-b(h+\epsilon)}$$

for all $u \in H$.

Take $b > M$, with M as given in Theorem 3.4 and Theorem 3.5. Now given that ϕ is such that with probability one $\bar{y} = (y_1, \dots, y_b)$ is among the distinguishable group $\{u_1, \dots, u_N\}$ of Y^b , we have that

$$(4.7) \quad H(y_1, \dots, y_b | w_1, \dots, w_b) < \eta(\epsilon) + \eta(1-\epsilon) + \epsilon b \log t.$$

Also

$$(4.8) \quad N \geq B^{b(C-\epsilon)}.$$

Fix $b \geq \max(b_0(\epsilon), M)$.

$$\#_{H < B^{b(h+\epsilon)}} < B^{b(C-\epsilon)}$$

from (4.1) and (4.6).

This implies that there exists $\psi: X^b \rightarrow Y^b$ such that H is carried one-to-one, onto $\psi(H)$, a proper subset of $\{u_1, \dots, u_N\}$. All elements of L are mapped into some $u_i \notin \psi(H)$.

We define $\bar{\phi}$ (and ϕ) by

$$(\bar{\phi}\bar{x})_n = \psi(\bar{x}_n).$$

Now

$$\bar{y}_1 = \psi(\bar{x}_1)$$

is in the set $\{u_1, \dots, u_N\}$. So by (4.7)

$$(4.9) \quad H(\bar{y}_1 | \bar{w}_1) < \eta(\epsilon) + \eta(1-\epsilon) + \epsilon b \log t.$$

Since $b \geq 1$ and $\eta(\alpha) \geq 0$ for $0 \leq \alpha \leq 1$, the right-hand side of (4.9) is

$$\leq b\eta(\epsilon) + b\eta(1-\epsilon) + \epsilon b \log t$$

$$< b\delta/2.$$

by (4.2).

Since ψ is one-to-one and onto, if $\bar{y}_1 \in \psi(H)$ then it completely determines \bar{x}_1 . Hence for $u \in \psi(H)$

$$(4.10) \quad \sum_{v \in X^b} \eta(P(\bar{x}_1=v | \bar{y}_1=u)) = 0 \quad .$$

For any $u \in Y^b$, since $\#X^b = r^b$, the left-hand side of (4.10) is dominated by $\log r^b$; see [7, p.61].

Since

$$P\{\bar{y}_1 \in \psi(H)\} = P\{\bar{x}_1 \in H\} \quad ,$$

we have

$$(4.11) \quad H(\bar{x}_1 | \bar{y}_1) \leq \epsilon \log r^b \\ < b\delta/2$$

by (4.10) and (4.2).

By (A2) of §1.8

$$H(\bar{x}_1 | \bar{w}_1) \leq H(\bar{x}_1, \bar{y}_1 | \bar{w}_1) \quad .$$

Now by (A1)

$$= H(\bar{y}_1 | \bar{w}_1) + H(\bar{x}_1 | \bar{y}_1, \bar{w}_1) \quad .$$

Using (A3)

$$\leq H(\bar{y}_1 | \bar{w}_1) + H(\bar{x}_1 | \bar{y}_1) \quad .$$

Now using (4.9) and (4.11)

$$H(\bar{x}_1 | \bar{w}_1) < b\delta \quad .$$

But by (A1) and (A3)

$$H(\bar{x}_1, \dots, \bar{x}_n | \bar{w}_1, \dots, \bar{w}_n) \leq \sum_{i=1}^n H(\bar{x}_i | \bar{w}_1, \dots, \bar{w}_n) \quad .$$

By (A3)


$$\begin{aligned} H(\bar{x}_1, \dots, \bar{x}_n | \bar{w}_1, \dots, \bar{w}_n) &< \sum_{i=1}^n H(\bar{x}_i | \bar{w}_i) \quad . \\ &= nH(\bar{x}_1 | \bar{w}_1) \quad . \end{aligned}$$

Thus

$$1/n H(\bar{x}_1, \dots, \bar{x}_n | \bar{w}_1, \dots, \bar{w}_n) < b\delta \quad .$$

Hence in the limit the equivocation (transmitting $\bar{x} \rightarrow \bar{w}$ through the compound channel) is $< b\delta$. Thus the equivocation for \bar{x} is less than δ and it then follows that the rate of transmission is $> 1 - \delta$. //.

Bibliography

- [1] C.E.Shannon and W.Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, (1949).
- [2] A.I.Khinchin, Mathematical Foundations of Information Theory, Dover Publications Inc., New York, (1957).
- [3] D.Blackwell, L.Breiman and A.Thomasian, Proof of Shannon's Transmission Theorem for Finite State Indecomposable Channels, Ann. Math. Stats. (29), (1958), pp. 1209-1220.
- [4] A.Wyner, Recent Results in the Shannon Theory, IEEE Trans. on IT, Vol. IT-20, No.1, January 1974, pp. 2-10.
- [5] E.Pfaffelhuber, Channels With Asymptotically Decreasing Memory and Anticipation, IEEE Trans. on IT, Vol. IT-17, No.4, July 1971, pp. 379-385.
- [6] P.Halmos, Lectures on Ergodic Theory, Chelsea Publishing Co., New York, (1956).
- [7] P.Billingsley, Ergodic Theory and Information, John Wiley and Sons, New York, (1965).
- [8] A.Feinstein, Foundations of Information Theory, McGraw-Hill, New York, (1958).
- 

- [9] R.Adler, Ergodic and Mixing Properties of Infinite Memory Channels, Proc. Amer. Math. Soc. 12, (1961), pp. 924-930.
- [10] S.Foguel, The Ergodic Theory of Markov Processes, Van Nostrand Reinhold Co., New York, (1969).
- [11] P.Halmos, Measure Theory, Van Nostrand, Princeton, (1950).
- [12] J.Neveu, Mathematical Foundations of the Calculus of Probability, Holden-Day Inc., San Fransico, (1965).
- [13] R.Gallagher, Information Theory and Reliable Communication, John Wiley and Sons, New York, (1968).
- [14] J.H.Van Lint, Coding Theory, Springer-Verlag, Berlin, (1971).