AN APPLICATION OF THE ERGODIC THEOREM
TO INFORMATION THEORY

A THESIS

Presented to

The Faculty of the Division of

Graduate Studies and Research

By

Lon Day Hadden

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Mathematics

Georgia Institute of Technology

March, 1973

AN APPLICATION OF THE ERGODIC THEOREM

TO INFORMATION THEORY

Approved:

James W. Walker, Chairman

Samuel H. Coleman

Pranas Zunde

Date approved by Chairman: 9 Jun. 71

ACKNOWLEDGMENTS

I gratefully acknowledge my debt to Dr. James W. Walker for his assistance and encouragement in the preparation of this paper. I also express my thanks to Dr. Samuel H. Coleman and Dr. Pranas Zunde for their valuable suggestions. My thanks also go to Mrs. Betty Sims for rapid and accurate typing of the manuscript. Finally, I wish to publicly thank my wife, Beverly, for without her encouragement and determination this work would have never been completed.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Ergodic theorems are concerned with convergence of averages of iterations of an operator acting on a function space or more generally on a topological linear space.

The first result of ergodic theory was proved by J. Von Neumann about 1930 and published in 1932. The von Neumann mean ergodic theorem states that if T is a measure preserving transformation on a measure space (X,A,μ), then for every $f \varepsilon L_2(X,A,\mu)$ there is a function $f^* \varepsilon L_2$ such that

$$\lim_n \int |f^*(x) - \frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)|^2 d\mu = 0.$$

At about the same time G. D. Birkhoff proved under additional restrictions on the transformation T and the space X that for $f \varepsilon L_1$ the sequence $\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)$ is pointwise convergent to f^* for almost all x. These supplementary restrictions were later shown to be superfluous. The general theorem is known as the Birkhoff pointwise ergodic theorem.

Many generalizations of these theorems have followed. Specifically, S. Kakutani, K. Yosida and F. Riesz proved various assertions concerning mean convergence of operator averages in an abstract Banach space during the period 1935-1945.

Notable extensions of the Birkhoff theorem have been provided by

E. Hopf, N. Dunford and J. T. Schwartz, and R. V. Chacon and D. S. Ornstein.

The theory of information originated in the work of C. E. Shannon in 1948. In his fundamental paper, Shannon set up a mathematical scheme in which the concepts of an information source and of information transmission could be defined quantitatively. He then formulated and proved a number of very general results which showed the importance and usefulness of these definitions. Since 1948 a number of papers have been published which simplify and extend Shannon's original work.

In particular, in 1953 McMillan proved a very general result which states that for any stationary source, information may be transmitted at any rate less than channel capacity with arbitrarily small probability of error. This result is known as the McMillan theorem or the Asymptotic Equi-partition Property (AEP).

In Chapter II of this paper, after developing the necessary machinery from functional analysis, we prove an extension of the Von Neumann mean ergodic theorem. This result is then used to arrive at the Birkhoff pointwise ergodic theorem.

In Chapter III we turn our attention to information theory. The object of study here is a "communication system." This chapter is devoted to developing the theory of information to provide the background for Chapter IV.

In Chapter IV we use the Birkhoff theorem proved in Chapter II to extend the results of Chapter III. Specifically, we prove the McMillan theorem and hence establish a relationship between ergodic theory and information theory.

CHAPTER II

ERGODIC THEORY

In ergodic theory, one studies transformations that preserve the structure of measure spaces. In this chapter we shall discuss some concepts of ergodic theory and prove the Birkhoff point-wise ergodic theorem. This theorem will then be used in Chapter IV to prove the McMillan theorem. First, we need a few definitions.

In all that follows let $(\Omega, \mathcal{F}, P)$ be a probability space.

DEFINITION: Let T be a transformation of Ω into itself. Then T is *measurable* if $A \in \mathcal{F}$ implies $T^{-1}A = \{\omega\colon T\omega \in A\} \in \mathcal{F}$.

DEFINITION: Let T be a measurable transformation. If T is one-to-one, if $T\Omega = \Omega$, and if $A \in \mathcal{F}$ implies $TA = \{T\omega\colon \omega \in A\} \in \mathcal{F}$, then T is *invertible*.

DEFINITION: Let T be a measurable transformation. Then T is *measure preserving* in case $P(T^{-1}A) = P(A)$ for every $A \in \mathcal{F}$.

Let us now turn to a specific probability space of the type with which we will be concerned. Let X be a random variable with finite range, $\rho = \{s_1, s_2, \ldots, s_r\}$. Let $p_i = P[s_i]$ be the associated probability measure. Let $(\Omega, \mathcal{F}, P)$ be the product of a doubly infinite sequence of copies of the resulting measure space. Then the general element of Ω is a doubly infinite sequence

$$\omega = (\ldots,\omega_{-1},\omega_0,\omega_1,\ldots)$$

of elements of ρ. Let x_n be the nth coordinate function; that is, x_n is the mapping from Ω to ρ whose value $x_n(\omega)$ at the point ω is the nth coordinate ω_n of ω. We wish to characterize the probability measure, P, on $\mathcal{F}$. For this, we appeal to the Product Probability Theorem.

THEOREM 2.1. (PRODUCT PROBABILITY THEOREM). Let $(\Omega_t,\mathcal{A}_t,P_t)$, $t \in T$, be probability spaces. Let $\mathcal{C}_T$ be the class of all measurable cylinders of the form

$$\mathrm{Cyl}_{\Omega^T}[\underset{t\in T_N}{\times} A_t], \quad A_t \in \mathcal{A}_t.$$

That is, $\mathcal{C}_T$ is the class of all measurable cylinders in Ω^T based on the Cartesian products $\underset{t\in T_N}{\times} A_t$ for $A_t \in \mathcal{A}_t$. Define P_T on the class $\mathcal{C}_T$ by

$$P_T(\mathrm{Cyl}_{\Omega^T} \underset{t\in T_N}{\times} A_t) = \prod_{t\in T_N} P_t A_t.$$

Then, the product probability, P_T, on $\mathcal{C}_T$ is σ-additive and determines its extension to a probability, P_T, on the product σ-algebra $\mathcal{A}_T$.

Proof: See Loève pg. 91.

Hence, P is specified in our example by its values on what may be called "thin" cylinders of the form

$$\{\omega\colon x_\ell(\omega) = i_\ell, \quad n\le\ell<n+k\}$$

in the following manner

$$P\{\omega\colon x_\ell(\omega) = i_\ell, \quad n \le \ell < n+k\} = \prod_{\ell=n}^{n+k-1} p_{i_\ell}.$$

Let $T\colon \Omega \to \Omega$ be the mapping that carries $(\ldots,\omega_{-1},\omega_0,\omega_1,\ldots)$ into $(\ldots,\omega_0,\omega_1,\omega_2,\ldots)$, that is, T is defined by

$$x_n(T\omega) = x_{n+1}(\omega).$$

Note that $x_n(\omega) = x_0\big(T^n(\omega)\big)$ and consequently any statement about the random variables x_n can be converted into a statement about x_0 and T. If A is any cylinder of the form

$$\{\omega\colon \big(x_n(\omega),\ldots,x_{n+k-1}(\omega)\big) \in E\}$$

with E a subset of the Cartesian product ρ^k of k copies of ρ, then $T^{-1}A$ is also a cylinder and $T^{-1}A \in F$, and $P(T^{-1}A) = P(A)$. The following theorem shows that T is both measurable and measure preserving.

THEOREM 2.2. Let F_0 be a field generating F. If $T^{-1}A \in F$ and $P(T^{-1}A) = P(A)$ for every $A \in F_0$, then T is a measure preserving transformation.

Proof: See Billingsley, pg. 4.

We turn now to the proof of the Birkhoff point-wise ergodic theorem. The theorem will be proved in three steps. We first prove a slight generalization of the von Neumann mean ergodic theorem, then the maximal ergodic theorem, and finally the Birkhoff theorem itself. In the course of this development we shall need some results from the theory of Hilbert spaces. For completeness and to introduce notation, we include these results.

Let $(X,\mathcal{A},\mu)$ be a σ-finite measure space. Any measurable transformation T on X into X, measure preserving or not, induces a transformation V_T on M (the space of complex measurable functions defined a.e. on X) as follows: Letting $F \in M$, then for any $x \in X$ define

$$(V_T f)(x) = f(Tx)$$

provided the right-hand side of this equation is defined. The next lemma is central to the ergodic convergence theorems for measure preserving transformations.

<u>LEMMA 2.1</u>. Let T be measure preserving on X, and let V_T be the induced transformation on M. Then V_T is linear and positive (i.e. $f \geq 0$ a.e. implies that $V_T f \geq 0$ a.e.). Moreover,

$$\int V_t f d\mu = \int f d\mu \quad (f \in L_1);$$

and

$$\|V_T f\|_p = \|f\|_p \quad (f \in L_p,\ 1 \leq p \leq \infty)$$

that is, V_T is a linear isometry on each L_p.

<u>Proof</u>: That V_T is linear and positive is clear from its definition. To prove (i) suppose first that f is an integrable simple function, say $f = \Sigma\ C_k I_{A_k}$; then

$$(V_T f)(x) = \Sigma\ C_k I_{A_k}(Tx) = \Sigma\ C_k I_{T^{-1}A_k}(x) \tag{1}$$

and hence

$$\int V_T f d\mu = \sum C_k \mu(T^{-1}A_k) \tag{2}$$

$$= \sum C_k \mu(A_k) = \int f d\mu.$$

Now let f be non-negative and integrable on X. We may choose a sequence $\{f_n\}$ of non-negative integrable simple functions such that $f_n \leq f$ and $f_n(x) \to f(x)$ a.e. It follows from (1) and (2) above that $\{V_T f_n\}$ is a sequence of non-negative integrable functions. Moreover, $\int V_T f_n d\mu = \int f_n d\mu$ and

$$(V_T f_n)(x) = f_n(Tx) \uparrow f(Tx) = (V_T f)(x) \text{ a.e.}$$

Applying the monotone convergence theorem it follows that

$$\int f d\mu = \lim \int f_n d\mu = \lim \int V_T f_n d\mu = \int V_T f d\mu. \tag{3}$$

That (i) holds for an arbitrary $f \in L_1$ may now be seen by writing $f = f_1 - f_2 + i(f_3 - f_4)$ where $f_j \geq 0$ a.e., $f_j \in L_1$ $(j=1,2,3,4)$ and applying (3) to each f_j.

To prove (ii) we consider two cases.

(I) Assume $f \in L_p$ for some $p \in [1,\infty)$.

Then for $x \in X$

$$|(V_t f)(x)|^p = |f(Tx)|^p = (V_T |f|^p)(x),$$

whereupon by (i) $(|f|^p \in L_1)$ we have

$$\|V_T f\|_p^p = \int |V_T f|^p d\mu = \int V_T |f|^p d\mu = \int |f|^p d\mu = \|f\|_p^p,$$

and hence $\|V_T f\|_p = \|f\|_p$.

(II) Assume $f \in L_\infty$. Then for any $a>0$

$$\mu[|V_T f| \geq a] = \mu(T^{-1}[|f| \geq a]) = \mu[|f| \geq a];$$

and therefore

$$\|V_T f\|_\infty = \inf\{a\colon \mu[|V_T f| \geq a] = 0\}$$

$$= \inf\{a\colon \mu[|f| \geq a] = 0\} = \|f\|_\infty. \quad \square$$

We shall use the following notation. The inner product of two elements f and g of a Hilbert space will be denoted by (f,g). The adjoint of the operator U will be denoted U^* and is characterized by the equation $(Uf,g) = (f,U^*g)$, for all f and g.

LEMMA 2.2. If U is an isometry, then a necessary and sufficient condition that $Uf = f$ is that $U^*f = f$.

Proof. See Halmos [2], pg. 15.

We now come to the generalization of the von Neumann theorem.

THEOREM 2.3 (MEAN ERGODIC THEOREM). If U is an isometry on a complex Hilbert space, $\mathcal{H}$, and if P is the projection on the space of all invariant elements of $\mathcal{H}$ under U, then $\frac{1}{n}\sum_{j=0}^{n-1} U^j f$ converges to Pf for every $f \in \mathcal{H}$.

Proof. Let

$$S_1 = \{f \in H\colon Uf = f\}.$$

be the set of all invariant elements of H. Then, for $f \in S_1$,

$$\frac{1}{n}\sum_{j=0}^{n-1} U^j f = \frac{1}{n}\,[f+Uf+U^2f+\ldots+U^{n-1}f]$$

$$= \frac{1}{n}\,[f+f+f+\ldots+f] = \frac{1}{n}[nf] = f.$$

Hence, if $f \in S_1$, the theorem is true.

Let

$$S_2 = \{f \in H\colon f = g - Ug \text{ for some } g \in H\}.$$

Then, for $f \in S_2$

$$\frac{1}{n}\sum_{j=0}^{n-1} U^j f = \frac{1}{n}\sum_{j=0}^{n-1} U^j(g-Ug) = \frac{1}{n}\,[g-Ug+Ug-U^2g+\ldots+U^{n-1}g-U^ng]$$

$$= \frac{1}{n}\,(g-U^ng).$$

Therefore, for $f \in S_2$,

$$\left\|\frac{1}{n}\overset{n-1}{\underset{j=0}{}} U^j f\right\| = \left\|\frac{1}{n}\,(g-U^ng)\right\|$$

$$\leq \frac{1}{n}\,\|g\| + \frac{1}{n}\,\|U^ng\| = \frac{2}{n}\,\|g\|$$

Hence, for $f \in S_2$,

$$\lim_{n\to\infty} \left\| \frac{1}{n} \sum_{j=0}^{n-1} U^j f \right\| \le \lim_{n\to\infty} \frac{2}{n} \|g\| = 0.$$

We show next that if f is an element of the closure of S_2 then

$$\lim_{n\to\infty} \left\| \frac{1}{n} \sum_{j=0}^{n-1} U^j f \right\| = 0.$$

First we must establish a relation between $\left\| \frac{1}{n} \sum_{n=0}^{n-1} U^j f \right\|$ and $\|f\|$ for any $f \in H$. Let $f \in H$. Consider

$$\left\| \frac{1}{n} \sum_{j=0}^{n-1} U^j f \right\| \le \frac{1}{n} \sum_{j=0}^{n-1} \|U^j f\| = \frac{1}{n} \sum_{j=0}^{n-1} \|f\|$$

$$= \|f\|.$$

Therefore, for every $f \in H$,

$$\left\| \frac{1}{n} \sum_{j=0}^{n-1} U^j f \right\| \le \|f\|.$$

Now let $A_n = \frac{1}{n} \sum_{j=0}^{n-1} U^j$ and let f be any element in the closure of S_2. Then there is a sequence $\{f_k\} \subset S_2$ such that given $\epsilon>0$ there exists M such that $k>M$ implies $\|f_k - f\| < \frac{\epsilon}{2}$. Also, since each $f_k \in S_2$, for each k there exists N_k such that $n>N_k$ implies $\|A_n f_k\| < \frac{\epsilon}{2}$. Let $\epsilon>0$ be given and consider

$$\|A_n f\| \le \|A_n(f-f_k)\| + \|A_n f_k\|$$

$$\le \|f-f_k\| + \|A_n f_k\|.$$

Fix $k>M$. Then $\|f-f_k\| < \frac{\epsilon}{2}$. For this k choose $n>M_k$. Then $\|A_n f_k\| < \frac{\epsilon}{2}$. Hence given $\epsilon>0$ there exists N such that if $n>N$ then $\|A_n f\| < \varepsilon$ or that $\lim_{n\to\infty} \|A_n f\| = 0$. Therefore $\lim_{n\to\infty} \|A_n f\| = 0$ for every $f \in \bar{S}_2$.

We now establish the fact that the orthogonal complement of S_2 is the same as the orthogonal complement of $\bar{S}_2$. We shall denote the orthogonal complement of a set S by $S^{\perp}$.

LEMMA 2.3. For any set S in a Hilbert space H

$$S^{\perp} = \bar{S}^{\perp}.$$

Proof. If $f \in \bar{S}^{\perp}$, then $(f,g) = 0$ for every $g \in \bar{S}$. Hence since $S \subset \bar{S}$ $(f,g) = 0$ for every $g \in S$. Therefore $f \in S^{\perp}$ and $S^{\perp} \subset \bar{S}^{\perp}$.

Now let $f \in S^{\perp}$. Then $(f,g) = 0$ for every $g \in S$. Let $g^* \in \bar{S}$. Then there is a sequence $\{g_k\} \subset S$ such that $\lim_{k\to\infty} g_k = g^*$. Hence,

$$(f,g^*) = (f,\lim_{k\to\infty} g_k) = \lim_{k\to\infty} (f,g_k) = \lim_{k\to\infty} 0 = 0$$

using the continuity of the inner product. Therefore $\bar{S}^{\perp} \subset S^{\perp}$. Combining this with the previous inclusion we have the result $\bar{S}^{\perp} = S^{\perp}$.

Using this fact let us determine $\bar{S}_2^{\perp}$ by considering $S_2^{\perp}$. Let $h \in S_2^{\perp}$. Then $(h,g-Ug) = 0$ for all $g \in H$. Hence,

$$(h,g) - (h,Ug) = 0$$

or

$$(h,g) - (U^*h,g) = 0$$

or

$$(h-U^*h,g) = 0 \qquad \text{for every } g \in H.$$

Therefore $h - U^*h = 0$. Then $h = U^*h$ and by Lemma 2.2 $h = Uh$. Thus if $h \in S_2^\perp$ (hence $h \in \bar{S}_2^\perp$), then $Uh = h$.

Now let h be such that $Uh = h$. Then by reversing the previous argument $h \in S_2^\perp$ and hence $h \in \bar{S}_2^\perp$. Therefore

$$\bar{S}_2^\perp = S_1.$$

Now by the projection theorem every $f \in H$ can be expressed as a sum $f_1 + f_2$ where $f_1 \in S_1$ and $f_2 \in \bar{S}_2$. □

We need one definition and a lemma before moving to the Maximal Ergodic Theorem.

DEFINITION. Suppose that $\{a_i\}$, $i=1,2,\ldots,n$ is a finite sequence of real numbers and that m is a positive integer, $m \le n$. A term a_k of the sequence is an *m-leader* if there exists a positive integer p, $1 \le p \le m$, such that $a_k + \ldots + a_{k+p-1} \ge 0$.

LEMMA 2.4. The sum of the m-leaders is non-negative.

Proof. If there are no m-leaders, the assertion is true since an empty sum is 0 by convention. Let a_k be the first m-leader and let p be the smallest integer such that $p \le m$ and $a_k + \ldots + a_{k+p-1} \ge 0$. We shall show that a_h, $k \le h \le k+p-1$, is also an m-leader and that the sum $a_h + \ldots + a_{k+p-1} \ge 0$. Suppose not; i.e. suppose $a_h + \ldots + a_{k+p-1} < 0$. Then $a_k + \ldots + a_{k-1} > 0$. But this contradicts the choice of p. Now consider the sequence $a_{k+p},\ldots,a_n$. If this sequence has no m-leaders, then we have shown the theorem to be true. If there is at least one m-leader, let $a_{k'}$ be the first one and let p' be the smallest integer such that $p' \le m$ and $a_{k'} + \ldots + a_{k'+p'-1} \ge 0$. As before, we can show that each of these terms is also an m-leader. We proceed in this manner until there are no more m-leaders in the remaining sequence or we have exhausted the sequence. Observe that at this point we have some number, say N, of non-negative sums of length $p, p', p'', p^{(3)}, \ldots, p^{(N-1)}$. Each of these sums is non-negative and the only elements in these sums are m-leaders. Conversely, each m-leader is included in exactly one of the sums. Hence, the sum of the m-leaders is non-negative.

We now state and prove the Maximal Ergodic Theorem.

THEOREM 2.4 (MAXIMAL ERGODIC THEOREM). Let f be real valued and $f \in L_1$. Let T be a measure-preserving transformation of a space X. Denote $f(T^j x)$ by $f_j(x)$. If E is the set of points x such that $f_o(x) + \ldots + f_{n-1}(x) \ge 0$ for some n, then $\int_E f(x)d\mu \ge 0$.

Proof. Let E_m be the set of those points x for which at least one of the sums $f_o(x) + \ldots + f_p(x)$ is non-negative with $p \le m$. Note that the sequence $\{E_m\}$ is increasing and the union of the E_m's is E. Hence it will be sufficient to show that $\int_{E_m} f(x)d\mu \ge 0$ for each m.

Let n be an arbitrary positive integer and consider for each point x the m-leaders of the sequence $f_o(x), \ldots, f_{n+m-1}(x)$. Let s(x) be their sum. Let D_k be the set of those points x for which $f_k(x)$ is an m-leader of the sequence $f_o(x), \ldots, f_{n+m-1}(x)$ and let I_k be its indicator function. Note that each $f_j(x)$ is a measurable function and hence each

D_k is a measurable set. Note also that $s(x) = \sum_{k=0}^{n+m-1} f_k(x)I_k(x)$. Hence $s(x)$ is both measurable and integrable. By the lemma

$$\sum_{k=0}^{n+m-1} f_k(x)I_k(x) \geq 0$$

and hence

$$\int \sum_{k=0}^{n+m-1} f_k(x)I_k(x)d\mu = \sum_{k=0}^{n+m-1} \int_{D_k} f_k(x)d\mu \geq 0.$$

Observe that if $Tx \in D_{k-1}$ then $f_{k-1}(Tx) + \ldots + f_{k-1+p-1}(Tx) \geq 0$ for some $p \leq m$. This implies that $f_k(x) + \ldots + f_{k+p-1}(x) \geq 0$ for some $p \leq m$. This in turn means that $x \in D_k$. Since each of these steps is reversible, the four conditions are equivalent. Hence, $D_k = T^{-1} D_{k-1}$ for $k = 1,2,\ldots,n-1$, or $D_k = T^{-k} D_o$ for $k = 1,2,\ldots,n-1$. Therefore

$$\int_{D_k} f_k(x)d\mu = \int_{T^{-k}D_o} f(T^k x)d\mu = \int_{D_o} f(x)d\mu.$$

Hence

$$\sum_{k=0}^{n-1} \int_{D_k} f_k(x)d\mu = n \int_{D_o} f(x)d\mu.$$

Now, D_o is the set of those points x such that $f_o(x)$ is an m-leader of the sequence $f_o(x),\ldots,f_{n+m-1}(x)$. That is, $x \in D_o$ if and only if there is an integer p" such that the sum $f_o(x) + \ldots + f_{p'} \geq 0$, $1 \leq p \leq m$. But this is exactly the set E_m. Therefore $\sum_{k=0}^{n-1} \int_{D_k} f_k(x)d\mu = n \int_{E_m} f(x)d\mu$.
Note that

$$\int_{D_k} f_k(x)d\mu \le \int_{D_k} |f_k(x)|d\mu = \int_{T^{-k}D_o} |f(x)|d\mu \le \int |f(x)|d\mu.$$

Hence,

$$\sum_{k=n}^{n+m-1} \int_{D_k} f_k(x)d\mu \le m \int |f(x)|d\mu.$$

Therefore, we have

$$0 \le \sum_{k=0}^{n+m-1} \int_{D_k} f_k(x)d\mu \le n \int_{E_m} f(x)d\mu + m \int |f(x)|d\mu$$

and dividing by n

$$\int_{E_m} f(x)d\mu + \frac{m}{n} \int |f(x)|d\mu \ge 0$$

for every m and n. Now let n tend to infinity. This yields

$$\int_{E_m} f(x)d\mu \ge 0 \qquad \text{for every } m.$$

Thus, $\int_E f(x)d\mu \ge 0$. $\square$

We come now to the major point of this chapter.

<u>THEOREM 2.5</u> (BIRKHOFF POINTWISE ERGODIC THEOREM). Let (X,A,μ) be a σ-finite measure space and T a measure-preserving transformation on X. If $f \in L_1$, then $\frac{1}{n}\sum_{j=0}^{n-1} f\left(T^j(x)\right)$ converges almost everywhere. The limit function f^* is integrable and invariant in the sense that $f^*(Tx) = f^*(x)$ almost everywhere. If in addition $\mu(X)<\infty$, then

$$\int f^*(x)d\mu = \int f(x)d\mu.$$

Proof. Let a and b be real numbers with a<b. Define the set

$$Y(a,b) = \{x: \liminf_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f_j(x) < a < b < \limsup_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f_j$$

By the definitions of lim inf and lim sup $Y(a,b)$ is measurable and invariant under T in the sense that $Y(a,b) = T^{-1}Y(a,b)$. We shall show first that $\mu\big(Y(a,b)\big)$ is finite and then that $\mu\big(Y(a,b)\big) = 0$.

We first assume that $b>0$. Let C be any subset of $Y(a,b)$ such that C is measurable and $\mu(C)<\infty$. Let I_C be the indicator function of C. Then the Maximal Ergodic Theorem applies to $f-bI_C$ since $\mu(C)<\infty$ implies $bI_C \in L_1$ and hence $f - bI_C \in L_1$. Let E be the set as described in the Maximal Ergodic Theorem but for $f - bI_C$ rather than f. Then we have

$$\int_E (f-bI_C)(x)d\mu \geq 0.$$

Now if $x \in Y(a,b)$, then $b < \limsup \frac{1}{n} \sum_{j=0}^{n-1} f_j(x)$. But this means that at least one of the averages $\frac{1}{n} \sum_{j=0}^{n-1} f_j(x)$ must be greater then b. Hence $\frac{1}{n} \sum_{j=0}^{n-1} f_j(x) - b > 0$ for at least one n. Thus, we have the following inequalities

$$0 < \frac{1}{n} \sum_{j=0}^{n-1} f_j(x) - b$$

$$\leq \frac{1}{n} \sum_{j=0}^{n-1} f_j(x) - bI_C(x)$$

$$\leq \sum_{j=0}^{n-1} f_j(x) - bI_C(x).$$

Therefore, for $x \in Y(a,b)$ at least one of the sums $\sum_{j=0}^{n-1} \big(f_j(x)-bI_C(x)\big) \geq 0$. But this means that $x \in E$. Hence $Y \subset E$. Now by the Maximal Ergodic Theorem

$$\int_E \big(f(x)-bI_C(x)\big)d\mu \geq 0.$$

Therefore,

$$\int |f(x)|d\mu \geq \int_E |f(x)|d\mu \geq \int_E bI_C(x)d\mu = b\mu(C).$$

We have shown thus far that if $C \subset Y(a,b)$ is measurable and has finite measure then

$$\mu(C) \leq \frac{1}{b} \int |f(x)|d\mu.$$

Now, since X is of σ-finite measure, there is a decomposition of X, call it $\{C_i\}$, such that

$$C_i \cap C_j = \phi \qquad i \neq j$$

$$\mu(C_i) < \infty \qquad i=1,2,\ldots$$

$$X = \bigcup_{i=1}^{\infty} C_i.$$

The sequence of sets $\{C_i \cap Y\}$ then forms a decomposition of Y. Since for each i $C_i \cap Y \subset Y(a,b)$, the $\mu[C_i \cap Y] \le \frac{1}{b}\int |f(x)|d\mu$. Note now that the sequence of sets $\{[\bigcup_{i=1}^{r} C_i]\cap Y\} = \{\bigcup_{i=1}^{r} [C_i \cap Y]\}$ is monotone increasing and that for every r

$$[\bigcup_{i=1}^{r} C_i] \cap Y \subset Y.$$

Since $\mu[\bigcup_{i=1}^{r} C_i] < \infty$ for every r, then

$$\mu\{[\bigcup_{i=1}^{r} C_i]\cap Y\} \le \frac{1}{b}\int |f(x)|d\mu$$

for every r.

Therefore

$$\lim_{r\to\infty} \mu\{[\bigcup_{i=1}^{r} C_i]\cap Y\} \le \frac{1}{b}\int |f(x)|d\mu$$

But

$$\lim_{r\to\infty} \mu\{[\bigcup_{i=1}^{r} C_i]\cap Y\} =$$

$$\mu \lim_{r\to\infty} \{[\bigcup_{i=1}^{r} C_i]\cap Y\} =$$

$$\mu[\lim_{r\to\infty} \bigcup_{i=1}^{r} C_i \cap Y] =$$

$$\mu[X \cap Y] = \mu[Y]$$

Hence

$$\mu[Y] \le \frac{1}{b}\int |f(x)|d\mu < \infty.$$

Now consider the space Y and the function f-b. Since

$$\int_Y |f-b|d\mu < \int_Y |f|d\mu \leq \int |f|d\mu < \infty$$

f-b is an integrable function.

Let E_{f-b} be the set defined in the Maximal Ergodic Theorem. Then

$$E_{f-b} = \{x: \ f_0(x) - b + f_1(x) - b + \ldots + f_{n-1}(x) - b \geq 0 \text{ for some } n\}.$$

$$E_{f-b} = \{x: \ \sum_{j=0}^{n-1} f_j(x) - nb \geq 0 \quad \text{for some } n\}$$

$$E_{f-b} = \{x: \ \frac{1}{n}\sum_{j=0}^{n-1} f_j(x) - b \geq 0 \text{ for some } n\}.$$

Note that if $x \in Y$ then $x \in E_{f-b}$. Hence $Y \subset E_{f-b}$. Also, since we are treating Y as the whole space (it is invariant), $E_{f-b} \subset Y$. Therefore $E_{f-b} = Y$ and hence

$$\int_Y \big(f(x)-b\big)d\mu = \int_{E_{f-b}} \big(f(x)-b\big)d\mu \geq 0$$

Applying the maximal ergodic theorem to a-f in a similar fashion we have

$$\int_Y \big(a-f(x)\big)d\mu \geq 0.$$

Combining these two inequalities we have

$$\int_Y (a-b)d\mu \geq 0$$

$$(a-b)\mu(Y) \geq 0.$$

But a<b and hence $\mu(Y) = 0$. Hence for every pair of rational numbers and such that a<b, the measure of the set Y such that

$$\liminf \sum_{j=0}^{n-1} f_j(x) < a < b < \limsup \sum_{j=0}^{n-1} f_j(x)$$

is zero. Therefore

$$\liminf \sum_{j=0}^{n-1} f_j(x) = \limsup \sum_{j=0}^{n-1} f_j(x)$$

Hence, the limit function f* does exist almost everywhere.

In our argument we have relied heavily on the assumption that b>0. If this were not the case, then a would have to be negative and the same argument could be carried through with -f and -a in place of f and b, respectively. Hence no generality has been lost. Note now that

$$\int \left|\frac{1}{n} \sum_{j=0}^{n-1} f_j(x)\right| d\mu \leq \frac{1}{n} \int \sum_{j=0}^{n-1} |f_j(x)| d\mu$$

$$= \frac{1}{n} \sum_{j=0}^{n-1} \int |f_j(x)| d\mu$$

$$= \frac{1}{n} \sum_{j=0}^{n-1} \int_{T^{-j}(X)} |f(x)| d\mu$$

$$= \frac{1}{n} \sum_{j=0}^{n-1} \int_X |f(x)| d\mu$$

$$= \int |f(x)| d\mu < \infty.$$

Therefore

$$\int \left|\frac{1}{n} \sum_{j=0}^{n-1} f_j(x)\right| d\mu \leq \infty \quad \text{for every } n.$$

Now by Fatou's Lemma we have

$$\int |f^*(x)| d\mu = \int \lim\inf \left|\frac{1}{n} \sum_{j=0}^{n-1} f_j(x)\right| d\mu$$

$$\leq \lim\inf \int \left|\frac{1}{n} \sum_{j=0}^{n-1} f_j(x)\right| d\mu < \infty$$

Therefore,

$$\int |f^*(x)| d\mu < \infty$$

and hence $f^*(x)$ is finite almost everywhere.

We now wish to show that f^* is invariant.

$$f^*(Tx) = \lim_{t\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f\left(T^j(Tx)\right)$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} f(T^j x)$$

$$= \lim_{n\to\infty} \left[\frac{1}{n} \sum_{j=0}^{n-1} f(T^j x) + \frac{1}{n} f(T^n x) - \frac{1}{n} f(x)\right]$$

$$= \lim_{n\to\infty} \frac{1}{n} f(T^n x) - \lim_{n\to\infty} \frac{1}{n} f(x) + f^*(x)$$

Now since $\int |f(x)| d\mu < \infty$, then

$$f(x) < \infty \quad \text{almost everywhere}$$

and hence

$$\lim_{n\to\infty} \frac{1}{n} f(x) = 0 \quad \text{almost everywhere.}$$

Also

$$\lim_{n\to\infty} \frac{f(T^n x)}{n} = 0 \quad \text{almost everywhere.}$$

since $\frac{1}{n} \sum_{j=0}^{n-1} f(T^j x)$ converges almost everywhere.

Hence

$$f^*(Tx) = f^*(x) \quad \text{almost everywhere}$$

and hence f^* is invariant.

We must now show that if $\mu(X) < \infty$, then

$$\int f d\mu = \int f^* d\mu.$$

Suppose that f^* is such that $f^*(x) \geq a$ for all x. Then at least one of the sums $\sum_{j=0}^{n-1} \left(f_j(x) - a + \varepsilon\right)$ must be non-negative for each ε. Then by the maximal ergodic theorem

$$\int f(x)d\mu \geq (a-\varepsilon)\mu(X) \quad \text{for each } \varepsilon>0.$$

Hence

$$\int f(x)d\mu \geq a\mu(X).$$

In a similar manner if $f^*(x) \leq b$ for every x, then

$$\int f(x)d\mu \leq b\mu(X).$$

Fix n and let

$$X(k,n) = \{x: \frac{k}{2^n} \leq f^*(x) \leq \frac{k+1}{2^n}\} \ .$$

Each $X(k,n)$ is invariant and so the above inequalities apply, so that

$$\frac{k}{2^n}\,\mu\big(X(k,n)\big) \leq \int_{X(k,n)} f(x)d\mu \leq \frac{k+1}{2^n}\,\mu\big(X(k,n)\big)$$

and

$$\frac{k}{2^n}\,\mu\big(X(k,n)\big) \leq \int_{X(k,n)} f^*(x)d\mu \leq \frac{k+1}{2^n}\,\mu\big(X(k,n)\big).$$

Thus, combining these two inequalities, we have

$$-\frac{1}{2^n}\,\mu\big(X(k,n)\big) \leq \int_{X(k,n)} f(x)d\mu - \int_{X(k,n)} f^*(x)d\mu \leq \frac{1}{2^n}$$

Or,

$$\Big|\int_{X(k,n)} f(x)d\mu - \int_{X(k,n)} f^*(x)d\mu\Big| \leq \frac{1}{2^n}\,\mu\big(X(k,n)\big)$$

Now, summing over k, we have

$$\left|\int f(x)d\mu - \int f^*(x)d\mu\right| \leq \frac{1}{2^n}\,\mu(X)$$

and since n is arbitrary

$$\int f(x)d\mu = \int f^*(x)d\mu. \quad \square$$

This completes the proof of the Birkhoff Theorem.

The results obtained in this chapter have been generalized to large classes of operators on large classes of abstract vector spaces. For other versions of the Von Neumann theorem see Dunford and Schwartz [2], Yosida, or Kakutani and Yosida. Generalizations of the Birkhoff Theorem may be found in Dunford and Schwartz [2] and Chacon and Ornstein.

CHAPTER III

INFORMATION AND UNCERTAINTY

Information theory is concerned with the analysis of a "communication system," which may be described as follows: A person or machine, called a *source*, produces a message to be communicated. An *encoder* then associates with each message an "object" called a *code word* which is suitable for transmission. The *code word* is presented to a *channel*, the medium over which the coded message is transmitted. A *decoder* then receives the output from the channel and attempts to reconstruct the original message for delivery to the *destination*. In general, the decoder cannot function with complete reliability because of *noise*, which is a general term for anything which tends to produce transmission errors.

It will be the purpose of this chapter to give meaning to the various terms "uncertainty," "information," "channel," "noisy," "code word," "rate," and "capacity." The development here will follow ASH. However, it will be our intent to illumine the concepts of uncertainty and information rather than to detail the mathematics involved. For this reason we will include many results without proof. For a different development of these concepts see Pinsker. We proceed by taking an intuitive view of

Let X be a random variable which takes on the values $x_1,x_2,\ldots,x_m$, with probabilities $p_1,p_2,\ldots,p_m$, respectively. We will require that

$p_i > 0$ for each $i=1,2,\ldots,M$, and, of course, that $\sum_{i=1}^{M} p_i = 1$. Then we say that we have a *finite scheme*

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_M \\ p_1 & p_2 & \cdots & p_M \end{pmatrix} .$$

Every finite scheme describes a state of uncertainty. It appears obvious that the "uncertainty" is different in different schemes. Consider the three schemes below

$$\begin{pmatrix} x_1 & x_2 \\ 0.5 & 0.5 \end{pmatrix}, \quad \begin{pmatrix} x_1 & x_2 \\ 0.9 & 0.1 \end{pmatrix}, \quad \begin{pmatrix} x_1 & x_2 \\ 0.7 & 0.3 \end{pmatrix}$$

In the second case it is almost certain that X will have the value x_1. In the first case the chances are equal that the value of X will be x_1 or x_2. The third case represents an amount of uncertainty between the other two.

We now attempt to arrive at a number that will measure the uncertainty associated with X. We shall do this by imposing certain reasonable requirements on the uncertainty associated with X and then showing that this leads us to an essentially unique function. For each M we define a function H_M of the M variables $p_1,p_2,\ldots,p_M$. The function $H_M(p_1,p_2,\ldots,p_M)$ will be interpreted as the average uncertainty associated with the events $\{X=x_i\}$. We will write $H_M(p_1,\ldots,p_M)$ as $H(p_1,\ldots,p_M)$ or as $H(X)$.

We now proceed to impose requirements on H. First suppose that all values of X are equally likely. We denote by f(M) the average uncertainty associated with M equally probable outcomes, that is, $f(M) = H(Y_M,\dots,1/M)$. For example, f(2) would be the uncertainty associated with the toss of a fair coin, and f(6) would be the uncertainty associated with the roll of an unbiased die. It seems reasonable that there should be a greater amount of uncertainty associated with rolling the die than with tossing the coin. Hence we arrive at our first requirement on the uncertainty function.

CONDITION I: $f(M) = H(1/M,\dots,1/M)$ is a monotonically increasing function of M.

Now consider an experiment involving two independent random variables X and Y. Let $X = \{x_1,\dots,x_M\}$, $Y = \{y_1,\dots,y_N\}$ and suppose that both X and Y have equally likely outcomes. Let $Z = X \times Y$ be the Cartesian product space. Then Z has equal probabilities at each of the MN points. Hence, the uncertainty associated with the joint experiment is f(MN). If the value of X is revealed, the uncertainty about Y should not be changed since X and Y are independent. Therefore, we expect that the uncertainty associated with Z minus the uncertainty associated with X should equal the uncertainty associated with Y. Now the uncertainty associated with X is just f(M). Hence, we have the second requirement on the uncertainty function H.

CONDITION II: $H\left(\frac{1}{MN},\dots,\frac{1}{MN}\right) = H\left(\frac{1}{M}\right) + H\left(\frac{1}{N}\right)$ or $f(MN) = f(M) + f(N)$.

We now drop the requirement of equally likely outcomes and turn to the general case. Let the random variable X take on the values $x_1, x_2, \ldots, x_M$ with probabilities $p_1, p_2, \ldots, p_M$, respectively. We divide the outcomes into two groups, A and B, where $A = \{x_1, \ldots, x_r\}$ and $B = \{x_{r+1}, \ldots, x_M\}$. Now consider the compound experiment which consists of first choosing one of the groups, A or B, and then picking one of the elements, x_i, from that group. The probability of choosing group A is exactly $p_1+p_2+\ldots+p_r$, and the probability of choosing group B is $p_{r+1}+\ldots+p_M$. Letting $p = P[A]$ and $1-p = p[B]$ we have

$$p = p[A] = \sum_{i=1}^{r} p_i$$

$$1 - p = p[B] = \sum_{i=r+1}^{M} p_i.$$

Then, if group A is selected, the probability that x_i, $i=1,2,\ldots,r$, will be chosen is $P[x_i/A]$. Now, for $i=1,2,\ldots,r$,

$$p[x_i/A] = \frac{P[x_i \cap A]}{P[A]} = \frac{P[x_i]}{P[A]} = \frac{p_i}{p}$$

Similarly, if group B is chosen, then the probability that x_i, $i=r+1, \ldots, M$, will be picked is

$$P[x_i/B] = \frac{p_i}{1 - p}.$$

The compound experiment described is equivalent to the original experiment of picking one of the elements x_i, $i=1,2,\ldots,M$. To establish this let Y be the outcome of the compound experiment. Then, if $x_i \in A$,

$$P[Y=x_i] = P[A]P[x_i/A]$$

$$= p\,\frac{p_i}{p} = p_i.$$

If $x_i \varepsilon B$, then

$$P[Y=x_i] = P[B]P[x_i/B]$$

$$= p\,\frac{p_i}{p} = p_i.$$

Hence $P[Y=x_i] = p_i = P[X=x_i]$ for $i=1,2,\ldots,M$. Before the compound experiment is performed, the uncertainty associated with the outcome is $H(p_1,\ldots,p_M)$. Revealing which group is selected removes on the average an amount of uncertainty $H(p,1-p)$. If group A is chosen, the uncertainty remaining is $H\left(\frac{p_1}{p}, \frac{p_2}{p}, \ldots, \frac{p_r}{p}\right)$. If group B is chosen, the uncertainty remaining is $H\left(\frac{p_{r+1}}{1-p}, \frac{p_{r+2}}{1-p}, \ldots, \frac{p_M}{1-p}\right)$. Now, since group A is chosen with probability, p, and B is chosen with probability, 1-p, the average uncertainty remaining after specifying the group is

$$pH\left(\frac{p_1}{p}, \frac{p_2}{p}, \ldots, \frac{p_r}{p}\right) + (1-p)H\left(\frac{p_{r+1}}{1-p}, \frac{p_{r+2}}{1-p}, \ldots, \frac{p_M}{1-p}\right)$$

Since the original experiment and the compound experiment are equivalent, we expect that the average uncertainty of the compound experiment minus the average uncertainty removed by specifying the group equals the average uncertainty remaining after the group is specified. Hence, we have the third requirement that we will impose on the uncertainty function.

<u>CONDITION III</u>: $H(p_1,p_2,\ldots,p_M) = H(p,1-p) + pH\left(\frac{p_1}{p},\ldots,\frac{p_r}{p}\right)$

$$+ (1-p)H\left(\frac{p_{r+1}}{1-p},\ldots,\frac{p_M}{1-p}\right)$$

where

$$p = \sum_{i=1}^{r} p_i, \qquad 1 - p = \sum_{i=r+1}^{M} p_i.$$

Finally, we expect that a small change in probabilities should cause only a small change in uncertainty and hence we require as our fourth condition:

<u>CONDITION IV</u>: $H(p,1-p)$ is a continuous function of p.

We now recapitulate the four requirements which we impose on the uncertainty function:

I. $f(M) = H(1/M,\ldots,1/M)$ is a monotonically increasing function of M.

II. $H\left(\frac{1}{MN},\ldots,\frac{1}{MN}\right) = H\left(\frac{1}{M}\right) + H\left(\frac{1}{N}\right)$ or $f(MN) = f(M) + f(N)$.

III. $H(p_1,\ldots,p_M) = H(p,1-p) + pH\left(\frac{p_1}{p},\ldots,\frac{p_r}{p}\right)$

$$+ (1-p)H\left(\frac{p_{r+1}}{1-p},\ldots,\frac{p_M}{1-p}\right)$$

where

$$p = \sum_{i=1}^{r} p_i, \qquad 1 - p = \sum_{i=r+1}^{M} p_i.$$

IV. $H(p,1-p)$ is a continuous function of p.

We now state and outline the proof of the following theorem which yields the uncertainty function.

THEOREM 1. The only function which satisfies the four conditions given above is

$$H(p_1,p_2,\ldots,p_M) = -C \sum_{i=1}^{M} p_i \log p_i,$$

where C is an arbitrary positive number and the logarithm base is any number greater than 1.

Proof. (Sketch). It is easily verified that the function

$$H(p_1,p_2,\ldots,p_M) = -C \sum_{i=1}^{M} p_i \log p_i$$

satisfies the four conditions imposed on the uncertainty function. In order to show that any function which satisfies the four conditions is of the specified form, we proceed as follows. First, using induction we show that $f(M^k) = kf(M)$. Again using induction, we show that $f(M) = C \log M$. We next establish that for any rational number p such that $0<p<1$, then $H(p,1-p) = -c[p \log p + (1-p) \log (1-p)]$. Using this result and the condition of continuity, we have

$$H(p,1-p) = -c[p \log p + (1-p) \log (1-p)]$$

for all real $p \in (0,1)$. Using this result and Condition III, we proceed by induction to prove the theorem.

Having arrived at a measure of the uncertainty associated with a random variable, we will now note some of the important properties of the uncertainty function. Although we shall state the properties as lemmas, theorems, and corollaries, we shall give only a few comments on the proofs. The details of these proofs may be found in any standard text on information theory such as Pinsker, Ash, or Khinchine.

We note first that since $p_i \log p_i \geq 0$ for all i then $H(X) \geq 0$.

LEMMA 1. Let $p_1, p_2, \ldots, p_M$ and $q_1, q_2, \ldots, q_M$ be arbitrary numbers such that

$$p_i > 0, \quad i=1,2,\ldots,M$$

$$q_i > 0, \quad i=1,2,\ldots,M$$

$$\sum_{i=1}^{M} p_i = \sum_{i=1}^{M} q_i = 1.$$

Then

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log q_i$$

with equality if and only if

$$p_i = q_i, \quad i=1,2,\ldots,M.$$

Proof. The proof of this lemma is based on the convexity of the function $f(x) = \log x$.

THEOREM 2. $H(p_1, p_2, \ldots, p_M) \leq \log M$ with equality if and only if $p_i = 1/M$, $i=1,2,\ldots,M$.

Proof. Apply Lemma 1 with $g_i = 1/M$.

Thus far, we have been concerned only with the uncertainty associated with a single random variable. We turn now to the case of two random variables and their joint and conditional uncertainty. The results here generalize to any finite number of random variables but we shall not discuss these generalizations. We first include some results from probability theory. Although familiarity with these results is assumed, we include them for completeness.

Suppose we have a space Z with a probability P_Z defined. Let each point of Z be expressed as an ordered pair (x,y) and write

$$P_Z[\{(x,y)\}] = p(x,y).$$

Note that the spaces X and Y are projections of Z. If we define

$$P_X[A] = p_Z[A\times Y] \text{ for } A \subset X$$

and

$$P_Y[B] = P_Z[X\times B] \text{ for } B \subset Y,$$

then it is easily verified that P_X and P_Y are probability measures on X and Y, respectively. In particular,

$$P_X(x) = P_Z[\{x\}\times Y] = \sum_{y\varepsilon Y} p_Z(x,y).$$

The measures P_X and P_Y are called *marginal* probability measures. Let $C_0 \subset Z$ be such that $P_Z[C_0] > 0$ and define

$$P_Z^{C_0}[C] = \frac{p_Z[C \cap C_0]}{p_Z[C_0]}, \quad \text{for } C \subset Z.$$

Again it is easily verified that $P_Z^{C_0}$ is a probability measure on the sets of Z. In the same manner as before $P_Z^{C_0}$ induces marginal probabilities on the sets of X and Y. In particular, let $C_0 = A \times Y$ and consider the marginal measure on the sets of Y induced by $P_Z^{A\times Y}$. We shall write

$$P_Z^{A\times Y}[X\times B] \text{ as } P_{Y/X}[B/A]$$

Then

$$P_{Y/X}[B/A] = P_Z^{A\times Y}[X\times B] = \frac{P_Z[(A\times Y)\cap(X\times B)]}{P_Z[A\times Y]}$$

$$= \frac{P_Z[A\times B]}{P_X[A]} .$$

The measure $P_{y/X}$ we shall call the conditional probability of Y given X. In particular, we write

$$P_{Y/X}(y/x) = P_{Y/X}[\{y\}/\{x\}] = \frac{P_Z[\{(x,y)\}]}{P_X[\{x\}]}$$

$$= \frac{p_Z(x,y)}{p_X(x)}$$

and

$$p_{X/Y}(x/y) = \frac{p_Z(x,y)}{p_y(y)} .$$

We say that the random vectors X and Y are independent in case

$$p_Z(x,y) = p_X(x)p_Y(y), \quad \text{for all } (x,y) \varepsilon Z$$

or

$$p_Z[A \times B] = (P_X[A])(P_Y[B]),$$

for all $A \subset X$, $B \subset Y$ and such that $A \times B \subset Z$.

DEFINITION. Let $P_Z = P_{X,Y}$ be a probability measure on the sets of $Z = X \times Y$. We define the *joint uncertainty* of X and Y by

$$H(X,Y) = - \sum_{(x,y)\varepsilon Z} p_Z(x,y) \log p_Z(x,y).$$

THEOREM 3. $H(X,Y) \leq H(X) + H(Y)$ with equality if and only if X and Y are independent.

Proof. Use the defining equations to compute $H(X) + H(Y)$ and then apply Lemma 1.

DEFINITION. We define the *conditional uncertainty* of Y given x by

$$H(Y/x) = - \sum_{y\varepsilon Y} p_{Y/X}(y/x) \log p_{Y/X}(y/x).$$

Furthermore, the average conditional uncertainty of Y given X is defined as the weighted averages of $H(Y/x)$ taken over all $x \varepsilon X$. That is,

$$H(Y/X) = \sum_{x\varepsilon X} p_X(x)H(Y/x).$$

THEOREM 4. $H(X,Y) = H(X) + H(Y?X) = H(Y) + H(X/Y)$.

Proof. This may be verified by direct calculation using the defining equations.

This last theorem justifies the intuitive idea that if the two random variables are observed but only the value of X is revealed, then the remaining uncertainty about Y should be the conditional uncertainty $H(Y/X)$.

THEOREM 5. $H(X) \geq H(X/Y)$ with equality if and only if X and Y are independent.

Proof. This follows directly from Theorems 3 and 4.

We are now ready to define a measure of information.

DEFINITION. The *information* about X conveyed by Y is given by

$$I(X/Y) = H(X) - H(X/Y).$$

Note that $I(X/Y)$ is always non-negative, and is zero if and only if X and Y are independent.

We have shown that

$$H(X/Y) = H(X,Y) - H(Y);$$

hence

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

But $H(X,Y) = H(Y,X)$ and hence

$$I(X/Y) = I(Y/X).$$

Therefore,

$$H(X) - H(X/Y) = I(X/Y) = I(Y/X) = H(Y) - H(Y/X)$$

and the information may be computed by either formula depending on the problem posed.

The fundamental significance of the information measure comes from its application to the reliable transmission of messages through noisy communications channels. We shall discuss this topic later. At this point however we turn our attention to describing the noiseless coding problem, that is, the problem of efficient coding of messages to be sent over a channel which allows perfect transmission. Any channel with this property will be called *noiseless*. We shall formally define an information channel later; but for now an intuitive idea will suffice.

Let $X = \{x_1,x_2,\ldots,x_M\}$ be a space with a probability measure defined on the points of X. We may think of the points, x_i, as words of a language. A message is constructed by sampling X. Thus $x_3x_1x_5x_1x_1$ would be a message. The channel is a device which accepts input from a code alphabet $\{a_1,a_2,\ldots,a_D\}$. Since the channel is assumed to be noiseless, the letters of the code alphabet are transmitted without error. A "word" $x_i \in X$ will be represented by a finite sequence of letters of the alphabet. This representation will be called the *code word* for x_i. The collection of all the code words will be called a *code*. The noiseless coding problem is then to minimize the average code word length, $\bar{n}$, by using different coding techniques. We define $\bar{n}$ by the following equation:

$$\bar{n} = \sum_{i=1}^{M} p_i n_i,$$

where

$p_i = p(x_i)$,

n_i = the length of the codeword associated with x_i.

We note immediately that there are some restrictions to be placed on the code words. For example, suppose that the alphabet is the set $\{0,1\}$, $X = \{x_1,x_2,x_3\}$ and code words were assigned as follows:

Word	Code Word
x_1	0
x_2	1
x_3	01

If the sequence 01 were received, we would be unable to determine whether x_3 was sent or the sequence x_1x_2. We wish to avoid such problems and are led to the following definition.

DEFINITION. A code is *uniquely decipherable* if every finite sequence of code characters corresponds to at most one message.

We now state another definition and note a theorem showing the relation of the two.

DEFINITION. A code is *instantaneous* if no code word has a prefix which is also a code word. By a prefix here we mean some initial string of letters from the code alphabet.

For clarification we give an example of a non-instantaneous code.

Word	Code Word
x_1	0
x_2	01

Note that although this code is uniquely decipherable, it is not instantaneous since the code word for x_1 is a prefix of the code word for x_2. This leads us to the following theorem.

THEOREM 6. If a code is instantaneous, then it is uniquely deciperable. The converse is false.

Proof. Given a finite sequence of code letters of an instantaneous code, proceed from left to right until a code word is formed. Since the code is instantaneous, this code word cannot be just the prefix of a larger code word and hence must represent the first word of the message. This process may be repeated until the sequence of code letters is exhausted. Hence every instantaneous code is uniquely decipherable. The previous example shows that the converse is not true.

Later in this chapter we shall state a result which guarantees that, for the purpose of solving the noiseless coding problem, we may restrict our attention to instantaneous codes. For this reason we now investigate the properties of such a code. First, we pose the following problem. Suppose we have a language $x_1,x_2,\ldots,x_M$, an alphabet $a_1,a_2,\ldots,$ a_D, and a set of positive integers $n_1,n_2,\ldots,n_M$. Under what conditions is it possible to construct an instantaneous code such that n_i is the length of the code word associated with x_i for $i=1,2,\ldots,M$. The following theorem provides the answer.

THEOREM 7. An instantaneous code with code word lengths $n_1, n_2, \ldots, n_M$ exists if and only if

$$\sum_{i=1}^{M} D^{-n_i} \leq 1,$$

where

D = the size of the code alphabet.

Proof. The proof rests on the construction of a probability tree of order D and size n_M, i.e., the Cartesian product of the alphabet space with itself n_M times, and noting that a code word of length n_k excludes $D^{n_M - n_k}$ paths through the tree or $D^{n_M - n_k}$ points or vectors in the Cartesian product space.

Theorem 7 may be strengthened to include not only instantaneous codes but also the class of uniquely decipherable codes. We will not prove this result but we will use it later.

We proceed now to solve the noiseless coding problem; that is, to find a uniquely decipherable code which minimizes the average code-word length $\bar{n}$. There are three steps in the solution. First, we establish a lower bound on $\bar{n}$; then we find out how close we can come to this lower bound. The third step is to construct the "best" code. We shall not pursue the third step of the problem in this work. To establish the lower bound on $\bar{n}$, we appeal to the following theorem.

THEOREM 8 (NOISELESS CODING THEOREM). If $\bar{n} = \sum_{i=1}^{M} p_i n_i$ is the average code-word length of a uniquely decipherable code for the random variable

X, then $\bar{n} \geq \frac{H(X)}{\log D}$ with equality if and only if $p_i = D^{-n_i}$, $i=1,2,\ldots,M$.

<u>Proof</u>. The condition $\bar{n} \geq \frac{H(X)}{\log D}$ may be rewritten as

$$\sum_{i=1}^{M} p_i n_i \log D \geq -\sum_{i=1}^{M} p_i \log p_i$$

or

$$-\sum_{i=1}^{M} p_i \log D^{-n_i} \geq -\sum_{i=1}^{M} p_i \log p_i$$

Hence all we must do is establish this last inequality. Recall that if $\sum_{i=1}^{M} p_i = 1$ and $\sum_{i=1}^{M} q_i = 1$, then

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log q_i$$

by Lemma 1. Define $q_i = \dfrac{D^{-n_i}}{\sum_{j=1}^{M} D^{-n_j}}$.

Hence

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log\left(\frac{D^{-n_i}}{\sum_{j=1}^{M} D^{-n_j}}\right)$$

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log D^{-n_i} + \left(\sum_{i=1}^{M} p_i\right) \log\left(\sum_{j=1}^{M} D^{-n_i}\right)$$

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log D^{-n_i} + \log \sum_{j=1}^{M} D^{-n_j}.$$

But, since the code is uniquely decipherable,

$$\sum_{j=1}^{M} D^{-n_j} \leq 1.$$

Hence

$$\log \sum_{j=1}^{M} D^{-n_j} \leq 0.$$

Therefore,

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log D^{-n_i}.$$

This last inequality guarantees that

$$\bar{n} = \frac{H(X)}{\log D} \quad \text{if} \quad p_i = D^{-n_i}.$$

Conversely, suppose

$$-\sum_{i=1}^{M} p_i \log p_i = -\sum_{i=1}^{M} p_i \log D^{-n_i}.$$

We wish to show that this implies $p_i = D^{-n_i}$, $i=1,2,\ldots,M$.

Rewriting the above equality we have

$$-\sum_{i=1}^{M} p_i \log p_i = -\sum_{i=1}^{M} p_i \log D^{-n_i} + 0$$

$$\geq -\sum_{i=1}^{M} p_i \log D^{-n_i} + \log\left(\sum_{j=1}^{M} D^{-n_j}\right)$$

$$= -\sum_{i=1}^{M} p_i \log\left(\frac{D^{-n_i}}{\sum_{j=1}^{M} D^{-n_j}}\right).$$

But we have already shown that

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log\left(\frac{D^{-n_i}}{\sum_{j=1}^{M} D^{-n_j}}\right).$$

Therefore,

$$-\sum_{i=1}^{M} p_i \log p_i = -\sum_{i=1}^{M} p_i \log D^{-n_i} = -\sum_{i=1}^{M} p_i \log\left(\frac{D^{-n_i}}{\sum_{j=1}^{M} D^{-n_j}}\right).$$

Hence $\sum_{j=1}^{M} D^{-n_j} = 1.$

Then

$$-\sum_{i=1}^{M} p_i \log p_i = -\sum_{i=1}^{M} p_i \log\left(\frac{D^{-n_i}}{\sum_{j=1}^{M} D^{-n_j}}\right).$$

Hence, applying Lemma 1 again,

$$p_i = \frac{D^{-n_i}}{\sum_{j=1}^{M} D^{-n_j}} = D^{-n_i}, \qquad i=1,2,\ldots,M.$$

In general, we will not be able to construct a code for a given set of probabilities which will achieve this minimum, since if we choose n_i so that $p_i = D^{-n_i}$, then $n_i = \frac{-\log p_i}{\log D}$ and this may not be an integer. The next theorem shows that although we may not achieve this minimum, we can come close.

THEOREM 9. Given a random variable X with uncertainty H(X), there exists a base D instantaneous code for X whose average code-word length satisfies

$$\frac{H(X)}{\log D} \leq \bar{n} \leq \frac{H(X)}{\log D} + 1.$$

Proof. Choose n_i such that

$$-\frac{\log p_i}{\log D} \leq n_i < \frac{\log p_i}{\log D} + 1.$$

We wish to show that an instantaneous code can be constructed with code-word lengths n_i defined above. Since

$$-\frac{\log p_i}{\log D} \leq n_i \qquad \text{for all } i$$

then

$$-\log p_i \leq n_i \log D$$

or

$$\log p_i \geq \log D^{-n_i}.$$

Hence

$$p_i \geq D^{-n_i}.$$

Therefore,

$$\sum_{i=1}^{M} D^{-n_i} \leq \sum_{i=1}^{M} p_i = 1.$$

Hence by Theorem 6, an instantaneous code with code-word lengths n_i does exist. We must show now that for this code

$$\frac{H(X)}{\log D} \leq \bar{n} \leq \frac{H(X)}{\log D} + 1.$$

We had

$$-\frac{\log p_i}{\log D} \leq n_i \leq \frac{\log p_i}{\log D} + 1.$$

If we multiply each term in this inequality by p_i and sum over all i we have

$$\frac{H(X)}{\log D} \leq \bar{n} \leq \frac{H(X)}{\log D} + 1.$$

We have thus completed the first two steps in solving the noiseless coding problem. The only remaining step is to construct the required code. Most texts on information theory discuss this topic. In particular, a celebrated construction is given in Huffman. Although we shall not pursue this issue, we include one theorem (without proof) which will allow us to restrict our search for such a code to the realm of instantaneous codes. First we need a definition.

DEFINITION. A code C, relative to a probability space, is *optimal* in a class of codes in case

$$\bar{n}_C \le \bar{n}_{C'},$$

where c' is any other code in the class.

THEOREM 10. If C is an optimal code within the class of instantaneous codes, then C is optimal within the class of uniquely decipherable codes.

Thus far, we have considered a channel as that portion of a communications system which carries the coded message from the sender to the receiver. We now attempt to present a mathematical model of a channel and define several types of channels.

DEFINITION. A triple $(X,Y,p(y/x))$ is called a *channel*. X is the space of "sendable" symbols and Y is the space of "receivable" symbols.

We define the information content of a channel in the same manner as before. That is,

$$I(X/Y) = H(X) - H(X/Y).$$

DEFINITION. If $H(X/Y) = 0$, then we say that the channel is *lossless*. Let A_{x_i} be a partition of Y such that $P[A_{x_i}/x_j] = 1$ for $i=j$ and

$P[A_{x_i}/x_j] = 0$ for $i \neq j$. Then, if in addition to being lossless, the channel has the property that for each i A_{x_i} is a singleton set, then we say that the channel is *noiseless.*

Define

$$c(p) = H(X) - H(X/Y)$$

where $p = \{p_1, p_2, \ldots, p_M\}$ is a probability measure on X so that $c(p)$ is defined on the simplex

$$s = \{p\colon \sum_{i=1}^{M} p_i = 1, \quad p_i \geq 0\}.$$

We define the *channel capacity* C by

$$C = \max_{p} c(p).$$

We remark here that this is a true maximum since

$$C(p) = I(X/Y)$$

is a continuous function on a compact set.

In general, we wish to transmit several successive elements, x_i, through a channel rather than just one. Although it is not a mathematical necessity, it may help the intuitive feeling to view the x_i's as being sent sequentially in time. This leads us to the definition of the extended channel.

DEFINITION. Given $\left(X,Y,p(y/x)\right)$, we define the triple $\left(U,V,p(V/u)\right)$ where

$$u = \{(x_1,x_2,\ldots,x_7): x_i \in X\}$$

$$V = \{(y_1,y_2,\ldots,y_n): y_j \in Y\}$$

$$p(v/u) = p\ \left((y_1,y_2,\ldots,y_n)/(x_1,x_2,\ldots,x_n)\right)$$

as an *extension* of length n of the channel $\left(X,Y,p(y/x)\right)$. We say that the extended channel is *memoryless* in case

$$p\left((y_1,y_2,\ldots,y_n)/(x_1,x_2,\ldots,x_n)\right) = p(y_1/x_1)p(y_2/x_2)\ \ldots\ p(y_n/x_n).$$

That is, the extended channel is memoryless in case the signal transmitted at time i is dependent only on the signal received at time i, i.e. independent of signals sent or received before time i.

THEOREM 11. Let $\left(X,Y,p(y/x)\right)$ be a discrete channel without memory, having capacity C. Then the capacity of its extension of length n is nC.

Proof. Show first that if

$$p(u) = p(x_1,x_2,\ldots,x_n) = p(x_1),p(x_2)\ldots p(x_n),$$

then

$$I(U/V) = H(u) - H(U/V) = n[H(X)-H(X/Y)].$$

Next show that $H(u) - H(u/v)$ for any probability distribution is bounded above by $H(U) - H(U/V)$ in the special case where $p(x_1,x_2,\ldots,x_n) = p(x_1)p(x_2)\ldots p(x_n)$. For details of the proof, see Feinstein.

We turn now to the problem of defining the "decoder." The purpose of the decoder is to translate the output of the channel into one of the possible input symbols. The decoder makes use of a decision scheme to perform this function. A decision scheme is nothing more than a partition of the space Y into M subsets $A_1,A_2,\ldots,A_M$ and a rule which assumes that x_i was transmitted if A_i was observed. To put the definition in negative terms, we say that if x_i is sent and the output y falls into A_j, $j\neq i$, then we have an error. Hence the probability of an error is

$$p(e) = \sum_{Y} p(y)[\left(1-p(x_y/y)\right]$$

where $p(x_y/y)$ is used to denote the probability that x_y was sent given that y was received. We now define one type of decision scheme.

<u>DEFINITION</u>. Let $\left(X,Y,p(y,x)\right)$ be a given channel. Then the partition of Y_3 into the sets $\{A_1,A_2,\ldots,A_M\}$ is called a *uniform error bounding* decision scheme with bound e in case

$$p(A_i/x_i) \geq 1 - 4, \quad i=1,2,\ldots,M.$$

It should be noted that it is not always possible to construct such a decision scheme for a given channel. In particular, Feinstein shows that in the case of a non-lossless channel such a construction is impossible.

We come now to a most important result in information theory. This theorem, known as the coding theorem for discrete memoryless channels and as the fundamental theorem of information theory, was first stated by C. E. Shannon in 1948.

THEOREM 12. Let $X,Y,p(/x)$ be a discrete memoryless channel with capacity C. Let H and e be given, with $0<H<C$ and $e>0$: then there exists a positive integer $n(e,H)$ such that in every extension of the channel $\left(X,Y,p(/x)\right)$ of length $n \geq n(e,H)$, there exists a set u_i, $i=1,2,\ldots,N$, $N \geq 2^{nH}$, to each of which is associated a v-set A_i, $i=1,2,\ldots,N$, such that the sets $\{A_i\}$ are disjoint and $p(A_i/u_i) \geq 1 - e$.

Since this theorem is not directly pertinent to the main investigations of this work, we shall omit the proof. The proof is presented in great detail in both Ash and Feinstein. We shall, however, discuss the importance of the result. Note, first, that an immediate result is the existence of a uniform error bounding decision scheme with bound e for all $e>0$.

Another important result is that by coding the messages to be sent with codes of sufficient length, we may transmit the coded messages at any rate less than channel capacity with arbitrarily small probability of error.

As previously noted, the development of the first part of this chapter follows Ash. The portion on the discrete memoryless channel follows Feinstein. It should be noted that the four conditions imposed on the uncertainty function may be replaced by three. These somewhat weaker conditions are given in Feinstein. Lee presents a development based on an even weaker set of conditions. Developments and results in the area of coding theory are discussed in Feinstein, Abamson, and Fano. The problem of determining the capacity of a given channel is dealt with in Muroga, Fano, and Ash.

CHAPTER IV

THE McMILLAN THEOREM

In the previous chapters we have discussed some aspects of ergodic theory and information theory. It will be the aim of this chapter to use the Birkhoff ergodic theorem to prove one of the major results in information theory, the McMillan theorem, and so tie the two concepts together. The concept of a source is central to the study of information and hence to the McMillan theorem. We begin by formulating this concept.

A *source* is that portion of an information system which creates the output or signal to be transmitted. Underlying the definition of a source is the set A of symbols used by it. We shall call A the *alphabet* of the source and refer to individual elements of A as *letters*. The alphabet A will be assumed to be finite. We shall denote by A^I the set of all doubly infinite sequences of the form $x = (\ldots, x_{-1}, x_0, x_1, x_2, \ldots)$. We define a set $Z \subset A^I$ to be a *cylinder set*, or briefly a *cylinder*, if it may be expressed in the form

$$Z = \{x: x_{t_i} = \alpha_i, \quad n \leq \ell < n+k\}.$$

Let F be the minimal Borel field over all the cylinders of the alphabet A. Then, as we have shown in Chapter II, the probability of any set $S \varepsilon F$ is uniquely determined by knowing the probabilities on all cylinder sets. Hence we can completely describe a source by specifying its alphabet A and the probability measure P on each of the cylinders of A. Hence we shall denote a source by [A,P]. Note that (A^I, F, P) is a probability space.

<u>DEFINITION.</u> The transformation T which carries the sequence $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$ into the sequence $Tx = (\ldots, x'_{-1}, x'_0, x'_1, \ldots)$ where $x_k = x'_{k+1}$ will be called the *shift operator.* (Notice that this operator is measurable).

<u>DEFINITION.</u> If $P(TS) = P(S)$ for every set $S \in F$, then the source is called *stationary.* Recalling the definition of a measure-preserving transformation we see that a source is stationary if and only if the shift operator is measure-preserving.

In the study of information, the prime characteristic of a source is the rate at which it emits information, i.e., the average amount of information given by each symbol produced. In the following we shall formulate an exact definition of this quantity. Let $C = \{x_t, x_{t+1}, \ldots, x_{t+n-1}\}$ be a sequence of length n of letters of A. If A consists of a letters, then there are exactly a^n such sequences. Each C so defined is a cylinder in A^I and hence has a definite probability P(C). Therefore we have a finite probability space consisting of a^n elements C. In Chapter III we arrived at the following measure of the information contained in this space

$$H_n = - \sum_C P(C) \log P(C).$$

Since we are assuming stationarity, the probabilities P(C) are uniquely determined by the nature of the source and by the number n. The same is obviously true for the entropy H_n. Therefore, the average amount of information per symbol emitted by the source is H_n/n. We would like to define the source entropy as the limit of H_n/n if this limit exists. Hence we are led to the following theorem.

<u>THEOREM 4.1</u>. If [A,P] is a stationary source, then

$$\lim_{n\to\infty} \frac{H_n}{n} \quad \text{exists and is finite.}$$

<u>Proof</u>. Let A_{n+m} be the space of sequences of length n+m. As was noted in Chapter III, A_{n+m} can be regarded as the product of the two spaces A_n and A_m. By the results of that chapter we have

$$H(A_{n+m}) = H(A_n) + H_{A_n}(A_m)$$

and

$$H_{A_n}(A_m) \leq H(A_m).$$

Combining these two results and using our new notation we have

$$H_n \leq H_{n+m} \leq H_n + H_m$$

for all integers m and n. Letting m=1 in the first of these inequalities we have

$$H_n \le H_{n+1}.$$

By induction the second inequality may be extended to yield

$$H_{n_1+n_2+\ldots+n_k} \le H_{n_1} + H_{n_2} + \ldots + H_{n_k}.$$

Then, taking $n_1=n_2=\ldots=n_k=n$,

$$H_{kn} \le k\,H_n \qquad \text{for all integer } K \text{ and } n.$$

In particular set $n=1$, then for any integer $k\ge 1$

$$H_k \le k\,H_1$$

Therefore

$$\frac{H_k}{k} \le H_1 \qquad \text{for every } k\ge 1.$$

Hence

$$\liminf_{n\to\infty} \frac{H_n}{n} < \infty .$$

Let $a = \liminf_{n\to\infty} \frac{H_n}{n}$. We now show that $\lim_{n\to\infty} \frac{H_n}{n}$ exists and is a.

Let $\varepsilon>0$ be given. Since $a = \liminf_{n\to\infty} H_n/n$, there is an index q such that

$$\frac{H_q}{q} < a + \varepsilon.$$

Note that for any $n>q$, there is an integer $k>1$ such that

$$(k-1)q < n \le kq.$$

Since we have shown that H_n is a monotonically nondecreasing function, we have, for n, k, and q as above,

$$H_n \le H_{kq}.$$

Then, since $(k-1)q < n$,

$$\frac{H_n}{n} \le \frac{H_{kq}}{(k-1)q} \, .$$

But $H_{kq} \le kHq$ and $\frac{H_q}{q} < a + \varepsilon$. Hence,

$$\frac{H_n}{n} \quad \frac{H_{kq}}{(k-1)q} \le \frac{kH_q}{(k-1)q} < \frac{k}{k-1}(a+\varepsilon) < a + \varepsilon.$$

Let n' be chosen such that $n>n'$ implies

$$\frac{H_n}{n} > a - \varepsilon.$$

Then we have

$$a - \varepsilon < \frac{H_n}{n} < a + \varepsilon.$$

or

$$\left|\frac{H_n}{n} - a\right| < \varepsilon.$$

But this is simply the defining inequality for

$$\lim_{n\to\infty} \frac{H_n}{n} = a. \quad \square$$

Consider now the random variable

$$\frac{1}{n} \log P(C)$$

where C is the cylinder $x_t, x_{t+1}, \ldots, x_{t+n-1}$. Obviously, $f_n(x)$ has the same value for all x belonging to the cylinder C. Hence, the mathematic expectation of $f_n(x)$ can be computed by elementary means. Therefore, letting $M\big(f(x)\big)$ denote the expectation of the random variable $f(x)$, we have

$$M\left(-\frac{1}{n} \log P(C)\right) = -\frac{1}{n} \sum_C P(C) \log P(C).$$

Recall that

$$-\sum_C P(C) \log P(C)$$

is the entropy of n-term sequences from the given source which we denoted by H_n. Since we are assuming the source to be stationary, we set t=0, so that C denotes the sequence $x_0, x_1, \ldots, x_{n-1}$. Then the random variable $-\frac{1}{n} \log P(C)$ is a function of x and n, which we denote by

$f_n(x)$; thus

$$Mf_n(x) = \frac{H_n}{n}.$$

We have shown that $\lim_{n\to\infty} \frac{H_n}{n} = H$, the entropy of the source. Hence for any stationary source

$$\lim_{n\to\infty} Mf_n(x) = H.$$

We now introduce the concept of a martingale which will facilitate the proof of the McMillan theorem. Since we need only one theorem, due to Doob, we will pursue the theory only as far as is required for the statement and later use of this theorem.

DEFINITION. Let $\{\xi_m\}$, $m=1,2,\ldots$, be a sequence of random variables defined on the space of elementary events $x \in A^I$. We shall denote the conditional expectation of ξ_m given that $\xi_1 = a_1$, $\xi_2 = a_2, \ldots, \xi_{m-1} = a_{m-1}$, by $M_{a_1 a_2 \ldots a_{m-1}}(\xi_m)$. The sequence $\{\xi_m\}$ is called a *martingale* if for any $m>1$

$$M_{a_1 a_2 \ldots a_{m-1}}(\xi_m) = a_{m-1}.$$

We shall deal only with bounded martingales, i.e., martingales, $\{\xi_m\}$, such that $|\xi_m| < C$ for every $x \in A^I$ and every index m.

THEOREM 4.2 (DOOB'S THEOREM). Every bounded martingale converges almost everywhere on A^I.

Proof. See Loève.

In order to prove the McMillan theorem we need to prove a few preliminary lemmas. We begin by establishing the notation to be used. We have already noted that every quantity which can be uniquely determined by the sequence $x_t,\dots,x_{t+n-1}$ of letters of the alphabet A can be regarded as a random variable on the space A^I. If C is the sequence $x_0,x_1,\dots,x_{n-1}$, then the function

$$f_n(x) = -\frac{1}{n}\log P(C)$$

is such a random variable. Let C_n be the sequence $x_{-n},\dots,x_{-1}$ and $c_n + x_0$ the sequence $x_{-n},\dots,x_{-1},x_0$. Each of these sequences is also a cylinder of the space A^I, as is the sequence

$$c_n + \alpha = x_{-n},\dots,x_{-1},\alpha$$

where α is any letter of the alphabet A. Now define the two random variables $p_n(x)$ and $p_n(x,\alpha)$ by

$$p_n(x) = \frac{P(C_n+x_0)}{P(C_n)}$$

$$p_n(x,\alpha) = \frac{P(C_n+\alpha)}{P(C_n)}\,.$$

We shall agree that $p_0(x) = P(x_0)$. These two random variables represent the conditional probability that x_0 will appear after the sequence C_n

and the conditional probability that α will appear after the sequence C_n, respectively.

LEMMA 4.1. The sequence $p_n(x,\alpha)$, $n=0,1,2,\ldots$ is a martingale.

Proof. We shall write $p_n(x,\alpha)$ as ξ_n. Let $a_{-1},\ldots,a_{-(n-1)}$ be any sequence of $n-1$ letters of A and denote by B_{n-1} the cylinder $x_{-i} = a_{-i}$, $i=1,2,\ldots,n-1$. Then $B_{n-1} \subset A^I$. Let Γ_β be the cylinder $x_{-n} = \beta$, $\beta \in A$. Now $\sum_{\beta\in A} \Gamma_\beta = A^I$. Hence

$$\int_{B_{n-1}} \xi_n dP = \sum_{\beta\in A} \int_{B_{n-1}\cap\Gamma_\beta} \xi_n dP.$$

If $x \in B_{n-1} \cap \Gamma_\beta$,

$$\xi_n = \frac{P(B_{n-1}\cap\Gamma_\beta+\alpha)}{P(B_{n-1}\cap\Gamma_\beta)}$$

Therefore

$$\int_{B_{n-1}} \xi_n dP = \sum_{\beta\in A} \int_{B_{n-1}\cap\Gamma_\beta} \frac{P(B_{n-1}\cap\Gamma_\beta+\alpha)}{P(B_{n-1}\cap\Gamma_\beta)} dP$$

$$= \sum_{\beta\in A} \frac{P(B_{n-1}\cap\Gamma_\beta+\alpha)}{P(B_{n-1}\cap\Gamma_\beta)} P(B_{n-1}\cap\Gamma_\beta)$$

$$= \sum_{\beta\in A} P(B_{n-1}\cap\Gamma_\beta+\alpha)$$

$$= P(B_{n-1}+\alpha)$$

Therefore

$$\int_{B_{n-1}} \xi_n dP = P(B_{n-1}+\alpha).$$

Denote by $[\xi_{n-1}]_{B_{n-1}}$ the value of the random variable ξ_{n-1} at $C_{n-1} = B_{n-1}$. Then

$$[\xi_{n-1}]_{B_{n-1}} = \frac{P(B_{n-1}+\alpha)}{P(B_{n-1})}$$

Hence

$$\int_{B_{n-1}} \xi_n dP = [\xi_{n-1}]_{B_{n-1}} P(B_{n-1}).$$

Now, let k_{n-1} be the set of all x for which $\xi_1,\ldots,\xi_{n-1}$ take on the given values $\xi_i = \pi_i$ $(i\leq i\leq n-1)$. The numbers π_i, $1\leq i\leq n-1$, are uniquely determined by specifying the cylinder B_{n-1}. Hence the set k_{n-1} is the union of several cylinders B_{n-1} and $[\xi_i]_{B_{n-1}} = [\xi_i]_{k_{n-1}} = \pi_i$, $1\leq i\leq n-1$, for all B_{n-1} in k_{n-1}. Therefore

$$\int_{k_{n-1}} \xi_n dP = \sum_{B_{n-1}\subset k_{n-1}} \int_{B_{n-1}} \xi_n dP = \sum_{B_{n-1}\subset k_{n-1}} [\xi_{n-1}]_{B_{n-1}} P(B_{n-1})$$

$$= \sum_{B_{n-1}\subset k_{n-1}} \pi_i P(B_{n-1})$$

$$= \pi_i P(k_{n-1})$$

Hence,

$$\pi_i = \frac{1}{P(k_{n-1})} \int_{k_{n-1}} \xi_n dP = M_{\pi_1\pi_2\ldots\pi_{n-1}}(\xi_n)$$

Therefore the sequence $\{p_n(x,\alpha)\}$ is a martingale. ▯

LEMMA 4.2. The sequence $\{p_n(x)\}$, n=0,1,... converges almost everywhere.

Proof. Let $x \in A^I$ be fixed. Then there exists $\alpha \in A$ such that

$$p_n(x) = p_n(x,\alpha) \qquad \text{for } n=0,1,\ldots$$

For α chosen in this manner

$$|p_n(x)-p_m(x)| = |p_n(x,\alpha)-p_m(x,\alpha)| \leq \sum_{\alpha\in A} |p_n(x,\alpha)-p_m(x,\alpha)$$

Now $\{p_n(x,\alpha)\}$ is a martingale and is obviously bounded by 1. Hence by Lemma 4.1 $\{p_n(x,\alpha)\}$ converges almost everywhere. Hence given $\varepsilon>0$, there exists N such that $n,m>N$ implies $\sum_{\alpha\in A} |p_n(x,\alpha)-p_m(x,\alpha)| < \varepsilon$. For n and m chosen this way then

$$|p_n(x)-p_m(x)| < \varepsilon.$$

But this means that $\{p_n(x)\}$ is a Cauchy sequence of real numbers and hence, converges. We, seemingly, have proved that the sequence $\{p_n(x)\}$ converges everywhere; however, recall that $p_n(x)$ is defined only for those x such that $P(C_n)>0$. The set of x such that $P(C_n) = 0$ is obviously of measure 0 and hence we have the conclusion almost everywhere. ▯

LEMMA 4.3. Let $g_n(x) = -\log p_n(x)$, n=1,2,... and let $E_{n,k}$, $n\geq 0$, $k\geq 0$ be defined by

$$E_{n,k} = \{x\colon k \le g_n(x) < k+1\}.$$

Then,

$$\int_{E_{n,k}} g_n(x)dP \le N(k+1)2^{-k},$$

where N is the number of letters in A.

Proof. Let B_n be defined as in Lemma 4.1, and let Z_α be the cylinder $x_0 = \alpha$, for $\alpha \in A$. For $x \in B_n \cap Z_\alpha$,

$$q_n(x) = -\log \frac{P(B_n+\alpha)}{P(B_n)} = -\log \frac{P(B_n \cap Z_\alpha)}{P(B_n)} .$$

Hence the value of $g_n(x)$ is determined uniquely by specifying B_n and α. Clearly,

$$B_n \cap E_{n,k} = \sum_{\alpha\in A^*} B_n \cap Z_\alpha$$

where A^* is the set

$$A^* = \{\alpha\in A\colon k \le g_n(x) < k+1,\ x \in B_n \cap Z\}.$$

Therefore

$$\int_{B_n \cap E_{n,k}} g_n(x)dP = \sum_{\alpha\in A^*} \int_{B_n \cap Z\alpha} g_n(x)dP. \tag{1}$$

In each of the integrals on the right

$$k \le g_n(x) = -\frac{\log P(B_n \cap Z_\alpha)}{P(B_n)} < k + 1.$$

Recalling that the logarithm base is 2, we have

$$\log \frac{P(B_n \cap Z_\alpha)}{P(B_n)} \le -k$$

$$\frac{P(B_n \cap Z_\alpha)}{P(B_n)} \le 2^{-k}$$

or

$$P(B_n \cap Z_\alpha) \le 2^{-k} P(B_n)$$

and substituting in (1)

$$\int_{B_n \cap E_{n,k}} g_n(x)dP \le \sum_{\alpha \varepsilon A^*} \int_{B_n \cap Z} (k+1)dP = \sum_{\alpha \varepsilon A^*} (k+1)P(B_n \cap Z_\alpha)$$

$$\le N(k+1)2^{-k}P(B_n).$$

Now summing over all cylinders B_n yields

$$\int_{E_{n,k}} g_n(x)dP \le N(k+1)2^{-k}. \quad \square$$

<u>LEMMA 4.4</u>. Given $L>0$, let $A_{n,L}$ be the set

$$A_{n,L} = \{x \varepsilon A^I : g_n(x) \ge L\}.$$

Then, given $\varepsilon>0$, there exists L_0 such that, for $L \ge L_0$ and all $n=1,2,...$

$$\int_{A_{n,L}} g_n(x)dP < \varepsilon.$$

Proof. Note first that for every n and L,

$$A_{n,L} = \sum_{k=L}^{\infty} E_{n,k}$$

and that $E_{n,k} \cap E_{n,j} = \phi$ for $j \neq k$.

Therefore

$$\int_{A_{n,L}} g_n(x)dP = \sum_{k=L}^{\infty} \int_{E_{n,k}} g_n(x)dP \leq \sum_{k=L}^{\infty} N(k+1)2^{-k}.$$

Now $\lim_{k\to\infty} \sum_{k=L}^{\infty} (k+1)2^{-k} < \infty$. Hence there is an L_0 such that $L \geq L_0$ implies

$$\sum_{k=L}^{\infty} (k+1)2^{-k} < \frac{\varepsilon}{N}.$$

Therefore, for $L > L_0$

$$\int_{A_{n,L}} g_n(x)dP < \varepsilon. \quad \square$$

LEMMA 4.5. Given $\varepsilon>0$, there is a $\delta>0$ such that if $E \in F$ and $P(E) < \delta$, then

$$\int_E g_n(x)dP < \delta, \; n=1,2,\ldots$$

Proof. By Lemma 4.4 given $\varepsilon>0$ there is an L such that

$$\int_{A_{n,L}} g_n(x)dP < \frac{\varepsilon}{2}, \qquad n=1,2,\ldots$$

Set $\delta = \frac{\varepsilon}{2L}$ and let $P(E) < \delta$. Then

$$\int_E g_n(x)dP = \int_{E\cap A_{n,L}} g_n(x)dP + \int_{(E\cap A_{n,L})^c} g_n(x)dP$$

Now for $x \in (E\cap A_{n,L})^c$, $g_n(x)<L$.

Therefore,

$$\int_E g_n(x)dP \le \int_{A_{n,L}} g_n(x)dP + LP(E)$$

$$< \frac{\varepsilon}{2} + L\,\frac{\varepsilon}{2L} = \varepsilon. \quad \square$$

Notice that $g_n(x) < \infty$ almost everywhere on A^I as a result of this lemma.

LEMMA 4.6. Let $g(x) = \lim_{n\to\infty} g_n(x)$. Then this limit exists almost everywhere on A^I and

$$\int_{A^I} g(x)dP < \infty.$$

Proof. That $g(x)$ exists almost everywhere, allowing the value $+\infty$, is an immediate consequence of Lemma 4.2. For $L>0$ set

$$g_n^L(x) = \min\{L, g_n(x)\}.$$

Then, since $g_n(x)\to g(x)$ almost everywhere, $g_n^L(x)\to g^L(x)$. Recall that the functions $g_n^L(x)$ are uniformly bounded for all n. Using this fact, Lemma 4.3, and the Lebesque Dominated Convergence Theorem, we have

$$\int_{A^I} g^L(x)dP = \int_{A^I} \lim_{n\to\infty} g_n^L(x)dP$$

$$= \lim_{n\to\infty} \int_{A^I} g_n^L(x)dP$$

$$\leq \limsup_{n\to\infty} \int_{A^I} g_n(x)dP$$

$$= \limsup_{n\to\infty} \sum_{k=0}^{\infty} \int_{E_{n,k}} g_n(x)dP$$

$$< N \sum_{k=0}^{\infty} (k+1)2^{-k}.$$

Therefore

$$\int_{A^I} g^L(x)dP < N \sum_{k=0}^{\infty} (k+1)2^{-k}$$

for every L>0. Hence

$$\int_{A^I} g(x)dP \leq N \sum_{k=0}^{\infty} (k+1)2^{-k} < \infty.$$

g(x) is finite almost everywhere, since

$$\int_{A^I} g(x)dP < \infty. \quad \square$$

LEMMA 4.7.

$$\lim_{n\to\infty} \int_{A^I} |g_n(x)-g(x)| dP = 0$$

Proof. Let $\varepsilon>0$ be given. Let E_n be defined by

$$E_n = \{x\varepsilon A^I \colon |g_n(x)-g(x)|>\varepsilon\}.$$

Then

$$\int_{A^I} |g_n(x)-g(x)|dP = \int_{E_n} |g_n(x)-g(x)|dP + \int_{E_n^c} |g_n(x)-g(x)|dP$$

$$\leq \int_{E_n} g_n(x)dP + \;_{E_n} g(x)dP + \varepsilon\, P(E_n^c).$$

By Lemma 4.4 there is a $\delta>0$ such that if $P(E_n)<\delta$, then

$$\int_{E_n} g_n(x)dP < \varepsilon.$$

Since $g_n(x)\to g(x)$ almost everywhere there is an n' such that $n>n'$ implies $P(E_n) < \delta$. Note also that by Lemma 4.6 $g(x)$ is summable over A^I and hence there exists a $\delta'>0$ such that if $P(E_n)<\delta'$

$$\int_{E_n} g(x)dP < \varepsilon.$$

Let $\delta^* = \min\{\delta,\delta'\}$ and let n^* be such that $P(E_n)<\delta^*$ for $n>n^*$. Then for any such n

$$\int_{E_n} g_n(x)dP < \varepsilon$$

and

$$\int_{E_n} g(x)dP < \varepsilon.$$

Also note that since $P(\cdot)$ is a probability measure $P(E_n^c) \leq 1$ for every set $E_n^c \in \mathcal{F}$. Hence

$$\int_{A^I} |g_n(x)-g(x)|dP < 3\varepsilon, \qquad \text{for } n>n^*.$$

Therefore

$$\lim_{n\to\infty} \int_{A^I} |g_n(x)-g(x)|dP = 0. \quad \square$$

We are almost ready to move to the McMillan Theorem. However, in that theorem we will be concerned with the function

$$f_n(x) = -\frac{1}{n} \log P(C),$$

where C is the cylinder $x_0, x_1, \ldots, x_{n-1}$. In order to use the results of the lemmas we have proved we must relate the functions $f_n(x)$ to the functions $g_n(x)$ we have been studying.

<u>LEMMA 4.8.</u> For all $x \in A^I$ and $n \geq 1$

$$f_n(x) = \frac{1}{n} \sum_{k=0}^{n-1} g_k(T^k x)$$

where T is the shift operator.

Proof. We shall use the following notation. The probability of the sequence $x_r,\ldots,x_{r+s}$ will be denoted by $P[x_r,\ldots,x_{r+s}]$. Using this notation we have

$$f_n(x) = -\frac{1}{n}\log P[x_0,\ldots,x_{n-1}]$$

and

$$p_n(x) = \frac{P[x_{-n},\ldots,x_0]}{P[x_{-n},\ldots,x_{-1}]}\,.$$

For $k\geq 0$ it is obvious that

$$p_n(T^k x) = \frac{P[x_{k-n},\ldots,x_k]}{P[x_{k-n},\ldots,x_{k-1}]}\,,$$

and for $n=k$

$$p_k(T^k x) = \frac{P[x_0,\ldots,x_k]}{P[x_0,\ldots,x_{k-1}]}\,.$$

This equality holds for all $k\geq 1$. Recall that $p_0(x) = P[x_0]$ by definitiontion. Hence

$$p_0(T^o x) = p_0(x) = P[x_0].$$

Therefore

$$\prod_{k=0}^{n-1} p_k(T^k x) = P[x_0]\cdot\frac{P[x_0,x_1]}{P[x_0]}\cdot\frac{P[x_0,x_1,x_2]}{P[x_0,x_1]}\cdots\frac{P[x_0,x_1,\ldots,x_n]}{P[x_0,x_1,\ldots,x_{n-2}]}$$

$$= P[x_0,x_1,\ldots,x_{n-1}].$$

Taking logarithms yields

$$\sum_{k=0}^{n-1} \log p_k(T^k x) = \log P[x_0,x_1,\ldots,x_{n-1}]$$

Recalling now that $g_n(x) = -\log p_n(x)$ and that $f_n(x) = -\frac{1}{n} \log P[x_0,\ldots,x_{n-1}]$, we have

$$\sum_{k=0}^{n-1} g_n(T^k x) = nf_n(x). \quad \square$$

We have defined a set S to be invariant under a transformation T if TS = S. In our present work we shall let T be the shift operator. The set A^I is always invariant as is the set $\{\ldots,T^{-1}x,x,Tx,T^2x,\ldots\}$.

DEFINITION. The source [A,P] is called *ergodic* if the probability P(S) of every invariant set $S \in \mathcal{F}$ is either 0 or 1.

THEOREM 4.3 (McMILLAN'S THEOREM). For any stationary source [A,P] the sequence $f_n(x)$ converges in L^1-mean to some invariant function h(x). In the case of an ergodic source, h(x) coincides almost everywhere in A^I with the entropy H of the source.

Proof. The function g(x) which we have defined is summable over A^I, i.e., $g(x) \in L^1$, by Lemma 7.6. Hence the Birkhoff ergodic theorem may be applied to g(x) and we have the result that

$$\frac{1}{n}\sum_{k=0}^{n-1} g(T^k x)$$

converges in L^1-mean to some invariant function $h(x)$. (We have noted previously that the shift operator T is measure-preserving if [A,P] is stationary.) By Lemma 4.8

$$\begin{aligned}\int_{A^I} |f_n(x)-h(x)|\,dP &= \int_{A^I} \left|\frac{1}{n}\sum_{k=0}^{n-1} g_k(T^k x)-h(x)\right| dP \\ &\le \int_{A^I} \left|\frac{1}{n}\sum_{k=0}^{n-1} [g_k(T^k x)-g(T^k x)]\right| dP \\ &+ \int_{A^I} \left|\frac{1}{n}\sum_{k=0}^{n-1} g(T^k x)-h(x)\right| dP \\ &\le \frac{1}{n}\sum_{k=0}^{n-1} \int_{A^I} |g_k(T^k x)-g(T^k x|\,dP + \int_{A^I} \left|\frac{1}{n}\sum_{k=0}^{n-1} g(T^k x)-h(x)\right| dP\end{aligned}$$

Now by stationarity

$$\int_{A^I} |g_k(T^k x)-g(T^k x)|\,dP = \int_{A^I} |g_k(x)-g(x)|\,dP.$$

Since $g_k(x)\to g(x)$ as $k\to\infty$, $|g_k(x)-g(x)|\to 0$ and hence

$$\lim_{n\to\infty}\left[\frac{1}{n}\sum_{k=0}^{n-1}\int_{A^I} |g_k(x)-g(x)|\,dP\right] = 0.$$

Therefore, given $\varepsilon>0$ there is an index n' such that $n>n'$ implies

$$\frac{1}{n}\sum_{k=0}^{n-1}\int_{A^I}|g_k(x)-g(x)|dP < \frac{\varepsilon}{2}\,.$$

By the definition of h(x), given $\varepsilon>0$ there is an index n" such that $n>n''$ implies

$$\int_{A^I}\left|\frac{1}{n}\sum_{k=0}^{n-1} g(T^k x)-h(x)\right|dP < \frac{\varepsilon}{2}\,.$$

Let $\varepsilon>0$ be given and choose n' and n" as above. Let $n^* = \max\{n',n''\}$. Then for $n>n^*$ we have

$$\begin{aligned}\int_{A^I}|f_n(x)-h(x)|dP &\le \frac{1}{n}\sum_{k=0}^{n-1}\int_{A^I}|g_k(x)-g(x)|dP \\ &+ \int_{A^I}\left|\frac{1}{n}\sum_{k=0}^{n-1} g(T^k x)-h(x)\right|dP \\ &< \varepsilon.\end{aligned}$$

Hence $f_n(x)$ converges in L^1-mean to h(x) and the first part of the theorem is proved.

In the case of an ergodic source, the corollary to the Birkhoff theorem states that the function h(x) is almost everywhere a constant h. Thus, to prove the second part of the theorem, we must show that h=H. The fact that $f_n(x)$ converges in L^1-mean to h implies that

$$\lim_{n\to\infty}\int_{A^I} f_n(x)dP = \int_{A^I} h dP = hP(A^I) = h$$

Now

$$\int_{A^I} f_n(x)dP$$

is just the mathematical expectation of the random variable $f_n(x)$ which we have shown has limit H. Hence

$$h = H. \qquad \square$$

This theorem, also called the *asymptotic equipartition property* (AEP) allows us to draw the following conclusion about the encoding of information produced by an ergodic source with uncertainty H. Suppose that the information produced by such a source is to be transmitted through a discrete memoryless channel with capacity C. Suppose $H<C$, and choose R such that $H<R<C$. Then, for sufficiently large n we can divide the sequences of length n into two classes S_1 and S_2 such that S_1 has at least $2^{n(H-\delta)}$ and at most $2^{n(H+\delta)}$ sequences for any $\delta>0$. In particular then, we may choose n so that S_1 has fewer than 2^{nR} sequences. Since the total probability of the sequences in S_2 can be made $\leq\varepsilon/2$, we can find a code with 2^{nR} input sequences of length n whose maximum probability of error is $\leq \varepsilon/2$ by assigning a code word of this code to each sequence in S_1 and assigning an arbitrary input sequence of length n to each sequence of S_2. Hence a source with uncertainty H can be handled by a channel with capacity C provided $H<C$.

For additional results in this area see Ash. Ash, Feinstein and Billingsley offer other developments of the topics treated in this chapter. Billingsley also discusses additional connections between the theories of ergodicity and information.

BIBLIOGRAPHY

ABRAMSON, N., *Information Theory and Coding,* New York: McGraw-Hill, 1963.

ASH, R. B., *Information Theory,* New York: Interscience Publishers, 1965.

BILLINGSLEY, P., *Ergodic Theory and Information,* New York: Wiley, 1965.

CHACON, R. V. and Ornstein, D. S., "A General Ergodic Theorem," *Illinois Journal of Mathematics,* Vol. 4, 1960.

DUNFORD, N. and Schwartz, J. T. [1], "Convergence Almost Everywhere of Operator Averages," *Journal of Rational Mechanics and Analysis,* Vol. 5, 1956.

_______ [2], *Linear Operators,* Vol. I, New York: Interscience, 1958.

FANO, R. M., *Transmission of Information; A Statistical Theory of Communications,* New York: M.I.T. Press, 1961.

FEINSTEIN, A., *Foundations of Information Theory,* New York: McGraw-Hill, 1958.

HALMOS, P. R. [1], *Measure Theory,* Princeton, N. J.: Van Nostrand, 1950.

_______ [2], *Lectures on Ergodic Theory,* New York: Chelsea, 1956.

_______ [3], "Recent Progress in Ergodic Theory," *Bulletin of the American Mathematical Society,* Vol. 67, 1961.

HUFFMAN, D., "A Method for the Construction of Minimum Redundancy Codes," *Communication Theory,* London: Butterworth's Scientific Publications, 1953.

KAKUTANI, S. and Yosida, K., "Operator-Theoretical Treatment of Markhoff Process and Mean Ergodic Theorem," *Annals of Mathematics,* Vol. 42, 1941.

KHINCHINE, A., *Mathematical Foundations of Information Theory,* New York: Dover, 1958.

LOÈVE, M., *Probability Theory,* 3rd ed., Princeton, N. J.: Van Nostrand, 1963.

McMILLAN, B., "The Basic Theorems of Information Theory," *Annals of Mathematical Statistics*, vol. 24, 1953.

MUROGA, S., *Threshold Logic and Its Applications*, New York: Wiley-Interscience, 1971.

PINSKER, M., *Information and Information Stability of Random Variables and Processes*, San Francisco: Holden-Day, 1964.

RUDIN, W., *Real and Complex Analysis*, New York: McGraw-Hill, 1966.

SHANNON, C. E. and Weaver, W., *The Mathematical Theory of Communication*, Urbana, Ill.: University of Illinois Press, 1949.

YOSIDA, K., *Functional Analysis*, Academic Press, 1965.