COMBINATORIAL ENTROPY AND UNIFORM LIMIT LAWS

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

John Michael Steele

May 1975

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Kai Lai Chung
(Principal Adviser)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

P van Moerbeke

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Sidney I Resnick
(Statistics)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Thomas M Cover
(Statistics)

Approved for the University Committee
on Graduate Studies:

J. Merrill Carlsmith
Dean of Graduate Studies

ACKNOWLEDGMENTS

I wish to thank my advisor, Professor Kai Lai Chung, for the assistance he has given me in bringing this work to fruition and also for sharing with me his wide interest in mathematics.

Further, I wish to thank Dr. Tom Cover of the Department of Statistics, Stanford University, Dr. Richard Dudley of M.I.T., and Dr. Paul Erdős of the Hungarian Academy of Sciences for their interest in this work and their very constructive comments.

For the excellent typing of this thesis, I gratefully acknowledge the craftsmanship of Sheila Hill.

Most fundamentally I wish to thank my mother for having been a lifelong inspiration of creativity and hard work.

TABLE OF CONTENTS

INTRODUCTION

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent identically distributed random variables defined on probability space $(\Omega, P, \mathcal{F})$ taking values in $\mathbb{R}^d$. If A is a Borel subset of $\mathbb{R}^d$, then, writing 1_A for the indicator function of A, we have the primordial fact that

$$\lim_{\ell \to \infty} \ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) = P(X_1 \in A) \quad \text{a.s.}$$

Indeed this consequence of the law of large numbers rests as one of the basic means of expressing what is meant by probability. Directly associated with this ancient result is a question almost as basic:

If S is a class of Borel subsets of $\mathbb{R}^d$, what is a necessary and sufficient condition that

$$\lim_{\ell \to \infty} \sup_{A \in S} \left| \ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \in A) \right| = 0 \quad \text{a.s. ?}$$

If S is a class of sets which satisfies the preceding equation, we call S a uniformity class. There are several such classes which have been discovered and found to be useful. The first such class to be widely used is of course $S = \{A : (-\infty, x] = A,\ x \in \mathbb{R}\}$ where $d = 1$, and the statement that this S is a uniformity class is exactly the classical Glivenko-Cantelli Theorem. The study of uniformity classes really acquires depth only when $d \geq 2$.

In this case, special uniformity classes have been studied by numerous authors including Fortet and Mourier [10], Wolfowitz [24] and [25], Blum [2], R. R. Rao [15], and S. Ahmad [1]. The general question as we pose it was first studied by Vapnik and Chervonenkis [22] and independently by Topsøe [20].

It is the direction of Vapnik and Chervonenkis which we pursue in this work. The fundamental idea is to introduce a type of combinatorial entropy of the class S which will lead to a complete characterization of the uniformity classes.

If $x_1, x_2, \ldots, x_\ell$ are ℓ points in $\mathbb{R}^d$ we define a class of subsets of these points by

$$I_S(x_1, x_2, \ldots, x_\ell) = \{\{x_1, x_2, \ldots, x_\ell\} \cap A : A \in S\} ,$$

and define

$$\Delta^S(x_1, x_2, \ldots, x_\ell) = |I_S(x_1, x_2, \ldots, x_\ell)|$$

where $|\mathcal{C}|$ denotes the cardinality of the class $\mathcal{C}$. As we will be concerned with these functions at great length, it is worthwhile to verify one's understanding by checking a simple case. To provide such a case let $d = 1$ and take $S = \{A : (-\infty,x] = A, x \in \mathbb{R}\}$. The class $I_S(x_1, x_2, \ldots, x_\ell)$ is seen to consist of the sets $\emptyset$, $\{x_1\}$, $\{x_1, x_2\}, \ldots, \{x_1, x_2, \ldots, x_\ell\}$ and these are aptly called the subsets of $\{x_1, x_2, \ldots, x_\ell\}$ which are <u>induced</u> by S. From this enumeration of $I_S(x_1, x_2, \ldots, x_\ell)$ one obtains $\Delta^S(x_1,x_2,\ldots,x_\ell) = \ell+1$.

The combinatorial entropy which concerns us is a function $h(F,S)$ of the class S and the distribution F of the X_i. It is defined by

$$h(F,S) = \lim_{\ell \to \infty} \ell^{-1} E \log \Delta^S(X_1, X_2, \ldots, X_\ell) .$$

This definition allows us to state the Main Theorem of Vapnik and

Chervonenkis:

A necessary and sufficient condition that the random variables

$$\sup_{A\in S} \Big| \ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \in A) \Big|$$

converge to zero in probability is that

$$h(F,S) = 0 \ .$$

In the first chapter of this thesis we strengthen this result to yield a corresponding uniform strong law and thus provide a complete answer to the question posed in the first paragraph of this introduction. In succeeding chapters we focus on results of second order and are able to obtain results which go beyond the law of large numbers for several large and useful classes. Finally we conduct a study of the particular classes given by $\{A, TA, T^2A, \ldots\}$ where T is a bimeasurable, measure preserving transformation and A is a Borel set. This study allows a complete comparison of the entropy of Vapnik and Chervonenkis with the entropy of Kolmogorov.

CHAPTER I

THE GENERAL CONVERGENCE THEOREMS

The principal objective of this chapter is to prove the strong limit law corresponding to the main theorem of Vapnik and Chervonenkis. In order to make this proof reasonably self-contained, we first give a brief proof of a technical lemma used by Vapnik and Chervonenkis. This repetition is justified by the numerous applications we make of the lemma and also by the understanding its proof fosters on the manner in which Δ^S enters our estimates. As an economy in expression we employ the following notation:

$$\nu_A'(\ell) = \ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) \qquad \nu_A''(\ell) = \ell^{-1} \sum_{i=\ell+1}^{2\ell} 1_A(X_i)$$

$$\rho_A^\ell = |\nu'_A(\ell) - \nu_A''(\ell)| \qquad \rho^\ell = \sup_{A \in S} \rho_A^\ell$$

$$\pi_A^\ell = |\nu_A'(\ell) - P(X_1 \in A)| \qquad \pi^\ell = \sup_{A \in S} \pi_A^\ell .$$

One quickly finds these six notations are significantly less cumbersome than the lengthy formulas their avoidance would require.

It is significant to note that the functions ρ^ℓ and π^ℓ are not necessarily measurable as certain pathological choices of S will easily show. In almost all cases, it is quite easy to show ρ^ℓ is measurable, but it happens sometimes that the measurability of π^ℓ requires more work. In Chapter I we will always assume that ρ^ℓ and π^ℓ are measurable, but some effort will be made to point out those cases when it suffices to assume only the measurability of ρ^ℓ. We are now in a position to state the last of our preliminaries.

Vapnik-Chervonenkis Lemma. For all ω we have

$$\frac{1}{(2\ell)!}\sum_{\sigma} 1\{\omega : \rho^{\ell}(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2\ell)}) > \epsilon\}$$

$$\leq 3\Delta^{S}(X_1, X_2, \dots , X_{2\ell})\exp(-\epsilon^2(\ell-1))$$

where the summation is over all permutations of $\{1, 2, \dots , 2\ell\}$.

Proof. For each ω there is a finite subset S' of S such that $\Delta^{S'}(X_1(\omega), X_2(\omega), \dots , X_{2\ell}(\omega)) = \Delta^{S}(X_1(\omega), X_2(\omega), \dots , X_{2\ell}(\omega))$ and for such an S' we have

$$1\{\omega : \rho^{\ell}(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2\ell)}) > \epsilon\}$$

$$\leq \sum_{A\in S'} 1\{\omega : \rho^{\ell}_{A}(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2,\ell)}) > \epsilon\} .$$

Hence we have

$$\frac{1}{(2\ell)!}\sum_{\sigma} 1\{\omega : \rho^{\ell}(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2\ell)}) > \epsilon\}$$

$$\leq \sum_{A\in S'} \frac{1}{(2\ell)!}\sum_{\sigma} 1\{\omega : \rho^{\ell}_{A}(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2\ell)}) > \epsilon\} .$$

We have to estimate

$$\frac{1}{(2\ell)!}\sum_{\sigma} 1\{\omega : \rho^{\ell}_{A}(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2\ell)}) > \epsilon\}$$

so we begin by assuming that m of the samples $X_1(\omega), X_2(\omega), \dots , X_{2\ell}(\omega)$ are in A. If the samples are then so distributed that k of $X_{\sigma(1)}(\omega), X_{\sigma(2)}(\omega), \dots , X_{\sigma(\ell)}(\omega)$ are in A, $m - k$ of $X_{\sigma(\ell+1)}(\omega)$, $X_{\sigma(\ell+2)}(\omega), \dots , X_{\sigma(2\ell)}(\omega)$ are in A and $|k/\ell - (m-k)/\ell| > \epsilon$, then

$$1\{\omega : \rho_A^\ell(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2\ell)}) > \epsilon\} = 1 .$$

The number of permutations σ which perform the above feat is seen to be

$$\sum_{k; |\frac{k}{\ell} - \frac{m-k}{\ell}| > \epsilon} \binom{m}{k}\binom{2\ell-m}{\ell-k}\ell!\ell!$$

If we now let

$$H(m,n,\epsilon) = \sum_{k; |\frac{k}{\ell} - \frac{(m-k)}{\ell}| > \epsilon} \binom{m}{k}\binom{2\ell-m}{\ell-k}/\binom{2\ell}{\ell}$$

we have proved that

$$\frac{1}{(2\ell)!} \sum_\sigma 1\{\omega : \rho_A^\ell(X_{\sigma(1)}, X_{\sigma(2)}, \dots , X_{\sigma(2\ell)}) > \epsilon\} = H(m,n,\epsilon) .$$

Here we recognize that $H(m,n,\epsilon)$ is the tail of a hypergeometric distribution, and we note that $H(m,n,\epsilon)$ can be estimated by classical methods (see [23], p. 273) to obtain $H(m,n,\epsilon) \leq 3 \exp(-\epsilon^2(\ell-1))$. Finally

$$\frac{1}{(2\ell)!} \sum_\sigma 1\{\omega : \rho^\ell(X_{\sigma(1)}(\omega), X_{\sigma(2)}(\omega), \dots , X_{\sigma(2\ell)}(\omega)) > \epsilon\}$$

$$\leq \sum_{A\in S'} 3 \exp(-\epsilon^2(\ell-1)) = 3\Delta^S(X_1(\omega), X_2(\omega), \dots , X_{2\ell}(\omega))\exp(-\epsilon^2(\ell-1)) .$$

Theorem 1.1. Let $Y_{tv} = \log_2\Delta^S(X_t, X_{t+1}, \dots , X_v)$, then we have the following properties of the process $\{Y_{tv}, t > 0, v > t\}$:

(1) Y_{tv} is stationary, that is the process defined by $Y'_{tv} = Y_{t+1,v+1}$ has the same finite-dimensional distributions as Y_{tv}.

(2) Y_{tv} is subadditive, that is $Y_{tv} \leq Y_{tu} + Y_{ut}$ for any $t < u < v$.

(3) Each Y_{tv} has finite expectation, in fact $0 \leq Y_{tv} \leq v - t + 1$.

(4) $\lim\limits_{\ell \to \infty} \ell^{-1}E(Y_{1\ell}) = \lim\limits_{\ell \to \infty} \ell^{-1}E(Y_{1\ell}) = c \geq 0$.

(5) $\lim\limits_{\ell \to \infty} \ell^{-1}Y_{1\ell} = c$ a.s.

Proof. Since the random variables $X_1, X_2, X_3, \ldots$ are i.i.d. the first conclusion is immediate. Since

$$\Delta^S(X_t, \ldots , X_u, \ldots , X_v) \leq \Delta^S(X_t, \ldots , X_u)\Delta^S(X_{u+1}, \ldots , X_v)$$

we have $Y_{tv} \leq Y_{tu} + Y_{u+1,v}$, and since $\Delta^S(X_{u+1}, \ldots , X_v) \leq \Delta^S(X_u, \ldots , X_v)$ this verifies $Y_{tv} \leq Y_{tu} + Y_{uv}$. Next we note that conclusion (3) follows direct from the fact that

$$1 \leq \Delta^S(X_t, \ldots , X_v) \leq 2^{v-t+1} .$$

Now if we let $a_n = E(Y_{1n})$, $n \geq 2$ we obtain from conclusions (1) and (2) that $0 \leq a_{n+m} \leq a_n + a_m$. For such a subadditive sequence it is well-known that $\lim\limits_{\ell \to \infty} \ell^{-1}a_\ell = \underline{\lim}_{\ell \to \infty} \ell^{-1}a_\ell = c \geq 0$. The only nontrivial conclusion to be considered is the last. The conditions of stationarity and subadditivity are exactly those of Kingman's subadditive ergodic theorem [14]. The conclusion of Kingman's result is that $\lim\limits_{\ell \to \infty} \ell^{-1}Y_{1\ell} = \xi$ a.s. for a random variable ξ. We will show that ξ is a.s. a constant by an application of the Hewitt-Savage zero-one law. Assuming that ξ is a constant it can be identified as c by an application of conclusion (4) and the dominated convergence theorem.

Let σ be a finite permutation of the sequence $(X_1, X_2, \ldots)$. If k is the largest index such that $\sigma(X_k) \neq X_k$ then we have

$$\Delta^S(X_1, X_2, \ldots, X_\ell) = \Delta^S(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(\ell)})$$

for all $\ell \geq k$, and hence

$$\lim_{\ell \to \infty} \ell^{-1} \log \Delta^S(X_1, X_2, \ldots, X_\ell) = \lim_{\ell \to \infty} \ell^{-1} \log \Delta^S(X_{\sigma(1)}, \ldots, X_{\sigma(\ell)}) .$$

This shows ξ is a random variable which is in the permutable field of the stationary independent process $(X_1, X_2, \ldots)$, so ξ is indeed a.s. a constant.

Theorem 1.2. A sufficient condition that

$$(1) \quad P(\lim_{\ell \to \infty} \sup_{A \in S} \ell^{-1} | \sum_{i=1}^{\ell} 1_A(X_i) - \sum_{i=\ell+1}^{2\ell} 1_A(X_i) | = 0) = 1$$

is that

$$(2) \quad \lim_{\ell \to \infty} \ell^{-1} E \log_2 \Delta^S(X_1, X_2, \ldots, X_\ell) = 0 .$$

Proof. Defining $D(\ell,\delta) = \{\omega : \log_2 \Delta^S(X_1, X_2, \ldots, X_{2\ell}) \geq \ell\delta\}$, $\delta > 0$, we have by Theorem 1.1 and the hypothesis (2) that

$$P(\bigcap_{L=1}^{\infty} \bigcup_{\ell=L}^{\infty} D(\ell,\delta)) = 0 .$$

Now setting $C(\ell,\epsilon) = \{\omega : \rho^\ell > \epsilon\}$, $\epsilon > 0$, we have

$$\bigcap_{L=1}^{\infty} \bigcup_{\ell=L}^{\infty} C(\ell,\epsilon) = \Big(\bigcap_{L=1}^{\infty} \bigcup_{\ell=L}^{\infty} C(\ell,\epsilon) \cap D(\ell,\delta)^c\Big) \bigcap \Big(\bigcap_{L=1}^{\infty} \bigcup_{\ell=1}^{\infty} C(\ell,\epsilon) \cap D(\ell,\delta)\Big),$$

and so to show $P(\bigcap_{L=1}^{\infty} \bigcup_{\ell=L}^{\infty} C(\ell,\epsilon)) = 0$ we need only show for some $\delta = \delta(\epsilon)$ that $P(\bigcap_{L=1}^{\infty} \bigcup_{\ell=L}^{\infty} C(\ell,\epsilon) \cap D(\ell,\delta)^c) = 0$. We will accomplish this by obtaining an estimate of $P(C(\ell,\epsilon) \cap D(\ell,\delta)^c)$ suitable for an application of the Borel-Cantelli lemma. For any permutation σ of $1, 2, \ldots , 2\ell$ we have

$$P(C(\ell,\epsilon) \cap D(\ell,\delta)^c) = \int_{D(\ell,\delta)^c} 1\{\omega : \rho^{\ell}(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(2\ell)}) > \epsilon\} dP$$

since in fact both ρ^{ℓ} and $D(\ell,\delta)^c$ are invariant under permutation. Now by summing over all permutations σ of $1, 2, \ldots , 2\ell$ we have

$$P(C(\ell,\epsilon) \cap D(\ell,\delta)^c)$$

$$= \int_{D(\ell,\delta)^c} \frac{1}{(2\ell)!} \sum_{\sigma} 1\{\omega : \rho^{\ell}(X_{\sigma(1)}, X_{\sigma(2)}, \ldots , X_{\sigma(2\ell)}) > \epsilon\} dP .$$

By the Vapnik-Chervonenkis lemma we therefore have

$$P(C(\ell,\epsilon) \cap D(\ell,\delta)^c) \leq \int_{D(\ell,\delta)^c} 3\Delta^{\delta}(X_1, X_2, \ldots , X_{2\ell})\exp(-\epsilon^2(\ell-1))dP$$

$\leq 3\cdot 2^{\ell\delta}\exp(-\epsilon^2(\ell-1))$, since on $D(\ell,\delta)^c$ we have $\Delta^{S}(X_1, X_2, \ldots , X_{\ell}) \leq 2^{\ell\delta}$. Finally by choosing $\delta < \epsilon^2/2$ our estimate shows that

$$\sum_{\ell=1}^{\infty} P(C(\ell,\epsilon) \cap D(\ell,\delta)^c) < \infty$$

and the theorem is proved.

Theorem 1.3. A necessary and sufficient condition that

$$P(\lim_{\ell \to \infty} \sup_{A \in S} |\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \in A)| = 0) = 1$$

is that $\lim_{\ell \to \infty} \ell^{-1} E \log_2 \Delta^S(X_1, X_2, \ldots , X_\ell) = 0$.

Proof. One-half of the result follows directly from the main theorem of Vapnik and Chervonenkis given in the introduction, since $\pi^\ell \to 0$ a.s. implies $\pi^\ell \to 0$ in probability and hence

$$\lim_{\ell \to \infty} \ell^{-1} E \log_2 \Delta^S(X_1, X_2, \ldots , X_\ell) = 0 .$$

The proof of sufficiency is the main issue, so we assume that $\lim_{\ell \to \infty} \ell^{-1} E \log_2 \Delta^S(X_1, X_2, \ldots , X_\ell) = 0$. As before, we let $D(\ell,\delta) = \{\omega : \log_2 \Delta^S(X_1, X_2, \ldots , X_{2\ell}) \geq \delta\ell\}$, and now we work to obtain an inequality relating $1(\rho^\ell > \epsilon)$ and $1(\sup_{A \in S} |\nu_A'(\ell) - P(X_1 \in A)| \geq 2\epsilon)$.

$$E(1_{D(\ell,\delta)^c}\, 1(\rho^\ell > \epsilon) | X_1, X_2, \ldots , X_\ell)$$

$$\geq \sup_{A \in S} E(1_{D(\ell,\delta)^c}\, 1(|\nu_A'(\ell) - \nu_A''(\ell)| > \epsilon) | X_1, X_2, \ldots , X_\ell)$$

$$\geq \sup_{A \in S} E(1_{D(\ell,\delta)^c}\, 1(|\nu_A'(\ell) - P(X_1 \in A)| > 2\epsilon) 1(|\nu_A''(\ell) - P(X_1 \in A)| < \epsilon) | X_1, X_2, \ldots, X_\ell).$$

Now we note that

$$1_{D(\ell,\delta)^c} \geq 1(\log_2 \Delta^S(X_1, X_2, \ldots, X_\ell) \leq \delta\ell/2) \cdot 1(\log_2 \Delta^S(X_{\ell+1}, X_{\ell+2}, \ldots, X_{2\ell}) \leq \delta\ell/2).$$

Since $1(|\nu_A'(\ell) - P(X_1 \in A)| > 2\epsilon) 1(\log_2 \Delta^S(X_1, X_2, \ldots, X_\ell) \leq \delta\ell/2)$ is contained in the sigma field $\sigma(X_1, X_2, \dots , X_\ell)$ and

$$1(|\nu''_A(\ell) - P(X_1 \in A)| < \epsilon)\cdot 1(\log_2 \Delta^S(X_{\ell+1},X_{\ell+2},\dots,X_{2\ell}) \leq \delta\ell/2)$$

is independent of $\sigma(X_1, X_2, \dots , X_\ell)$, we then have

$$E(1_{D(\ell,\delta)^c}\, 1(\rho^\ell > \epsilon)|X_1, X_2, \dots , X_\ell)$$

$$\geq \sup_{A\in S}\{1(|\nu'_A(\ell) - P(X_1 \in A)| > 2\epsilon)\cdot 1(\log_2 \Delta^S(X_1,X_2,\dots,X_\ell) \leq \delta\ell/2)\cdot$$

$$E(1(|\nu''_A(\ell) - P(X_1 \in A)| < \epsilon)\cdot 1(\log_2 \Delta^S(X_{\ell+1},X_{\ell+2},\dots,X_{2\ell}) \leq \delta\ell/2))\}.$$

Now by Chebyshev's inequality we note that

$$P(|\nu''_A(\ell) - P(X_1 \in A)| \geq \epsilon) \leq \epsilon^{-2}\ell^{-1}P(X_1 \in A)P(X_1 \notin A) \leq \epsilon^{-2}\ell^{-1}$$

and we also note that

$$\sup_{A\in S} 1(|\nu'_A(\ell) - P(X_1 \in A)| > 2\epsilon) = 1(\sup|\nu'_A(\ell) - P(X_1 \in A)| > 2\epsilon) .$$

Applying these observations we obtain the required inequality

$$(1) \quad E(1_{D(\ell,\delta)^c}\, 1(\rho^\ell > \epsilon)|X_1, X_2, \dots , X_\ell)$$

$$\geq 1(\sup_{A\in S}|\nu'_A(\ell) - P(X_1 \in A)| > 2\epsilon; \ \log_2 \Delta^S(X_1,X_2,\dots,X_\ell) \leq \delta\ell/2)\cdot$$

$$(1 - \epsilon^{-2}\ell^{-1} - P(\log_2 \Delta^S(X_{\ell+1},X_{\ell+2},\dots,X_{2\ell}) \geq \delta\ell/2) .$$

If we now let $g(\omega) = \sum_{\ell=1}^{\infty} E(1_{D(\ell,\delta)^c}\, 1(\rho^\ell > \epsilon)|X_1, X_2, \dots , X_\ell)$, then for $\delta < \epsilon^2/2$ we can check that $g(\omega)$ is integrable by the estimate given in Theorem 1.2. Specifically, we have

$$(2) \qquad E(E(1_{D(\ell,\delta)^c} 1(\rho^\ell > \epsilon)|X_1, X_2, \ldots , X_\ell))$$

$$= P(D(\ell,\delta)^c \cap \{\rho^\ell > \epsilon\}) < 3\cdot 2^{\ell\delta}\exp(-\epsilon^2(\ell-1)) \;.$$

For the concluding argument, we now let

$$\psi(\delta,\epsilon,\ell) = 1 - \epsilon^{-2}\ell^{-1} - P(\log_2 \Delta^S(X_1,X_2,\ldots,X_\ell) \geq \delta\ell/2)$$

and

$$h(\omega) = \sum_{\ell=1}^{\infty} \psi(\delta,\epsilon,\ell)1(\sup_{A\in S}|\nu'_A(\ell)-P(X_1\in A)|> 2\epsilon;\log_2\Delta^S(X_1,X_2,\ldots,X_\ell)\leq \delta\ell/2) \;.$$

By inequalities (1) and (2) we see $h(\omega)$ is integrable. On setting $F_\ell = \{\omega : \psi(\delta,\epsilon,\ell)1(\log_2 \Delta^S(X_1, X_2, \ldots , X_\ell) \leq \delta\ell/2) \geq 1/2\}$ and $G_k = \bigcap_{\ell=k}^{\infty} F_\ell$ we have G_k increases to an event of probability one. Further we see that

$$G_k \cap \{\omega : \sum_{\ell=1}^{\infty} 1(\sup_{A\in S}|\nu'_A(\ell) - P(X_1 \in A)| > 2\epsilon) = \infty\}$$

is contained in $\{\omega : h(\omega) = \infty\}$ and hence has probability zero. This last fact immediately yields the theorem.

As a consequence of Theorem 1.3 we note that the sufficient condition given in Theorem 1.2 is in fact also necessary. This stems from the easily established fact that $\pi^\ell \rightarrow 0$ a.s. if and only if $\rho^\ell \rightarrow 0$ a.s.

Under the very general hypothesis that

$$\lim_{\ell \rightarrow \infty} \ell^{-1}E \log \Delta^S(X_1, X_2, \ldots , X_\ell) = 0 \;,$$

the conclusions of Theorems 1.2 and 1.3 cannot likely be sharpened to give rates of convergence. Nevertheless, we can provide sharper results in case more is required on Δ^S. In fact, for many classes we have the condition that there exist constants C and τ such that

$$\Delta^S(X_1, X_2, \ldots , X_\ell) \leq C\ell^\tau .$$

The following result provides a more quantitative complement to Theorem 1.3 in this instance.

<u>Theorem 1.4</u>. If $\Delta^S(X_1, X_2, \ldots , X_\ell) \leq C\ell^\tau$ for some constants C and τ, then

$$P(\sup_{A\in S}|\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \in A)| > (c(\tau)\ell^{-1} \log \ell)^{1/2} \text{ i.o.}) = 0$$

where $c(\tau) = 4(2+\tau)^{1/2}$.

<u>Proof</u>. We let $\delta = 2\ell^{-1}(\tau \log_2(2\ell) + C)$, then

$$D(\ell,\delta)^c = \{\omega : \log \Delta^S(X_1, X_2, \ldots , X_{2\ell}) \leq \ell\delta\} = \Omega$$

and also $\{\omega : \log \Delta^S(X_1, X_2, \ldots , X_\ell) \leq \ell\delta/2\} = \Omega$. By inequality (1) of Theorem 1.3 we have

$$E(1_{D(\ell,\delta)^c} 1(\rho^\ell > \epsilon)|X_1, X_2, \ldots , X_\ell)$$

$$\geq 1(\sup_{A\in S}|\nu_A'(\ell) - P(X_1 \in A)| > 2\epsilon; \log_2 \Delta^S(X_1, X_2, \ldots , X_\ell) \leq \delta\ell/2)\cdot$$

$$(1 - \epsilon^{-2}\ell^{-1} - P(\log_2 \Delta^S(X_1, X_2, \ldots , X_\ell) \geq \delta\ell/2) .$$

This simplifies to yield

$$(1) \qquad E(1(\rho^{\ell} > \epsilon)|X_1, X_2, \dots , X_{\ell})$$

$$\geq (1 - \epsilon^{-2}\ell^{-1}) 1(\sup_{A \in S} |\nu_A'(\ell) - P(X_1 \in A)| > 2\epsilon) .$$

We also have

$$EE(1(\rho^{\ell} > \epsilon)|X_1, X_2, \dots , X_{\ell}) = P(\rho^{\ell} > \epsilon)$$

and

$$P(\rho^{\ell} > \epsilon) \leq 3C\ell^{\tau}\exp(-\epsilon^2(\ell-1))$$

by the Vapnik-Chervonenkis lemma. Now letting

$$\epsilon^2 = \epsilon(\ell)^2 = (\ell-1)^{-1}(\tau+2)\log \ell$$

we have

$$\sum_{\ell=1}^{\infty} P(\rho^{\ell} > \epsilon(\ell)) < \infty .$$

Thus by inequality (1) we have

$$\sum_{\ell=1}^{\infty} 1(\sup_{A \in S} |\nu_A'(\ell) - P(X_1 \in A)| > 2\epsilon)$$

is bounded by an integrable function, and consequentially

$$P(\sup_{A \in S} |\nu_A'(\ell) - P(X_1 \in A)| > 2\epsilon(\ell) \text{ i.o.}) = 0 .$$

Since $2\epsilon(\ell) \leq 4(2+\tau)^{1/2}(\ell^{-1}\log \ell)^{1/2}$ the proof is complete.

CHAPTER II

SECOND ORDER LIMIT THEOREMS

We will first focus on obtaining estimates for the function $\Delta^S(X_1, X_2, \dots, X_\ell)$ for certain classical classes S. These estimates are then put to work to yield strong and "very strong" limit theorems like the law of the iterated logarithm. The main result is in fact a law of the iterated logarithm for π^ℓ when the class S is taken to be the class of polynomial regions of bounded degree. In passing, a result is obtained which sharpens Wolfowitz' Glivenko-Cantelli theorem for half-spaces in $\mathbb{R}^d$. Further we show that our methods give the "upper" part of the laws of iterated logarithm for the empirical distribution function due to Chung [4] when $d = 1$ and Kiefer [13] when $d \geq 2$, but our results are not so precise in this instance as these earlier works.

In the following we use $P_{n,d}$ to denote the set of all real polynomials in d variables which have degree not greater than n. If $g \in P_{d,n}$, then the subset of $\mathbb{R}^d$ defined by $A_g = \{x : g(x) \geq 0\}$ is called a polynomial region of degree n. Our immediate task is to estimate $\Delta^S(X_1, X_2, \dots, X_\ell)$ where $S = \{A_g : g \in P_{d,n}\}$. Two proofs of the estimate are given, the first being the longer and more geometrical with the second being briefer and more algebraic.

Before stating the next result, we emphasize that $P_{n,d}$ will be considered always as a real vector space and $\dim_R P_{n,d}$ just denotes the dimension of $P_{n,d}$ as a vector space. Specifically, we know by a classical counting that $\tau = \binom{n+d}{n}$.

Theorem 2.1. If $S = \{A_g : g \in P_{n,d}\}$ and $\tau = \binom{n+d}{d}$, then $\Delta^S(x_1, x_2, \ldots, x_\ell) \leq \binom{\ell}{\tau-1}2^{\tau-1}$ for all $\ell \geq \tau$ and all $x_i \in \mathbb{R}^d$.

First Proof. Since two proofs will be given, we will first content ourselves with giving a sketch of a geometrical proof. We first will show that we may assume that the x_i are in general position, i.e., for any $g \in P_{n,d}$ we have $g(x_i) = 0$ for at most τ of the x_i. If we suppose that $x_1, x_2, \ldots, x_\ell$ are given, it then suffices to show that there are $x_1', x_2', \ldots, x_\ell'$ such that

$$\Delta^S(x_1, x_2, \ldots, x_\ell) \leq \Delta^S(x_1', x_2', \ldots, x_\ell') .$$

We do this by choosing polynomials g_i, $i = 1,2,\ldots,\Delta^S(x_1,x_2,\ldots,x_\ell)$ such that the sets A_{g_i} induce all the subsets of $x_1, x_2, \ldots, x_\ell$ which are induced by S. A moments thought shows that there is a constant $\alpha > 0$ such that

$$\{x_1,x_2,\ldots,x_\ell\} \cap \{x:g_i(x) \geq 0\} = \{x_1,x_2,\ldots,x_\ell\} \cap \{x:g_i(x) + \alpha \geq 0\}$$

for all $i = 1, 2, \ldots, \Delta^S(x_1, x_2, \ldots, x_\ell)$, and additionally one can require of α that $g_i(x_j) \neq 0$ for all i and j. Since we have that each x_j is in the interior of $\{x : g_i(x) + \alpha \geq 0\}$ or else in the open set $\{x : g_i(x) + \alpha < 0\}$, we can choose an $\epsilon > 0$ such that the ball $B(x_j, \epsilon)$ of radius ϵ about x_j is contained in either $\{x : g_i(x) + \alpha \geq 0\}$ or $\{x : g_i(x) + \alpha < 0\}$ for all i and j. Now choose x_j', $j = 1, 2, \ldots, \ell$ in general position such that $x_j' \in B(x_j, \epsilon)$. Since the polynomial domains $\{x : g_i(x) + \alpha \geq 0\}$ partition the balls $B(x_j, \epsilon)$ into $\Delta^S(x_1, x_2, \ldots, x_\ell)$ distinct subsets, we have that the x_j' are partitioned by $\{x : g_i(x) + \alpha \geq 0\}$ into $\Delta^S(x_1, x_2, \ldots , x_\ell)$ distinct subsets. A fortiori we have that

$$\Delta^S(x_1', x_2', \ldots , x_\ell') \geq \Delta^S(x_1, x_2, \ldots , x_\ell) .$$

Next we observe that one can assume that $x_i \neq 0$ for all i, and that $S_o = \{A_g : g \in P_{n,d}; g(0) = 1\}$ induces the same partitions on any set of $x_i \neq 0$ that S would induce.

These remarks complete the preliminaries. Now let $g \in P_{n,d}$, $g(0) = 1$ be given. Intuitively what we wish to do is to deform g (i.e., change its coefficients continuously) until we obtain a extremal polynomial $g^* \in P_{n,d}$, $g^*(0) = 1$, such that (1) $g^*(x_i) = 0$ for exactly $\tau - 1$ of the x_i, $i = 1, 2, \ldots , \ell$ and (2)

$$g^*(x_i) \geq 0 \quad \text{if} \quad g(x_i) \geq 0$$

and

$$g^*(x_i) \leq 0 \quad \text{if} \quad g(x_i) < 0 .$$

We now do some counting to show that the theorem is proved once we show that to each g we can associate a g^* as above. Since $g^*(0) = 1$, it is uniquely determined by the choice of $\tau - 1$ of the x_i such that $g^*(x_i) = 0$. Hence there are at most $\binom{\ell}{\tau-1}$ polynomials g^* which may arise in our association. By the second property of g^* there are at most $2^{\tau-1}$ distinct partitions which can give rise to the same g^*, since g^* determines the same partition of the x_i as g except possibly when $g^*(x_i) = 0$ which occurs exactly $\tau - 1$ times.

These two observations immediately give us that

$$\Delta^S(x_1, x_2, \dots, x_\ell) \leq 2^{\tau-1}\binom{\ell}{\tau-1} .$$

To complete the proof we need to show that the g^* exist. This can be done in several ways, but as these are messy and do not contribute further to the understanding of the estimate, we go on to our second proof.

The second proof of Theorem 2.1 depends on the following:

Schläfli's Theorem. Let $x_1, x_2, \dots, x_\ell$ be elements of $\mathbb{R}^d$, $d \geq 2$ in general position (i.e., any d elements of $x_1, x_2, \dots, x_\ell$ are linearly independent). Further, let H be the class of all $A_h = \{x : x \cdot h > 0\}$ with $h \in \mathbb{R}^d$, then

$$\Delta^H(x_1, x_2, \dots, x_\ell) = \sum_{k=o}^{d-1} \binom{\ell}{k} .$$

This result goes back to the work of Schläfli [18] and its direct bearing on Theorem 2.1 is explicitly pointed out in Cover [6]. This later reference contains a generalization of Schläfli's theorem, and applies the generalization to calculate functions related to $\Delta^S(x_1, x_2, \dots, x_\ell)$ for several classes such as hypercones and hyperspheres.

The second ingredient of the second proof is provided by another piece of ancient mathematics. We define (after the early algebraic geometer Veronesi) the mapping $\varphi(x) : \mathbb{R}^d \to \mathbb{R}^\tau$ given by

$$x = (x(1), x(2), \dots, x(d))$$

and

$$\varphi(x) = (1, x(1), \dots , x(d), x(1)^2, \dots , x(i)x(j), \dots , x(d)^n) .$$

Finally the proof of Theorem 2.1 can be made very brief, since

$$(1) \qquad \Delta^S(x_1, x_2, \dots , x_\ell) = \Delta^H(\varphi(x_1), \varphi(x_2), \dots , \varphi(x_\ell))$$

where

$$H = \{A_h : A_h = \{x \in \mathbb{R}^\tau : x \cdot h \geq 0\}\} .$$

By the perturbation result given at the beginning of our first proof we may assume $\varphi(x_i)$, $i = 1, 2, \dots , \ell$ are in general position in $\mathbb{R}^\tau$. By Schläfli's theorem we therefore have

$$\Delta^H(\varphi(x_1), \varphi(x_2), \dots , \varphi(x_\ell)) \leq \sum_{k=o}^{\tau-1} \binom{\ell}{k} .$$

We have thus proved by (1) that

$$\Delta^S(x_1, x_2, \dots , x_\ell) \leq \sum_{k=o}^{\tau-1} \binom{\ell}{k} .$$

Now

$$\sum_{k=o}^{\tau-1} \binom{\ell}{k} \leq 2^{\tau-1}\binom{\ell}{\tau-1}$$

for $1 \leq \tau \leq \ell$ so the second proof of the theorem is complete.

We note that the two proofs of Theorem 2.1, though methodically quite different, both provide bounds which are of degree $\tau - 1$ in ℓ. The two estimates are thus equivalent in most applications.

To provide one immediate such application we state the following:

Corollary. If $S = \{A_g : g \in P_{n,d}\}$ and $\tau = \dim_{\mathbb{R}} P_{n,d}$, then there is a constant $c = c(\tau)$ such that

$$P(\sup_{A \in S} \pi_A^\ell > c(\tau)(\ell^{-1} \log \ell)^{1/2} \text{ i.o.}) = 0 .$$

This side product of our efforts is, of course, immediate from Theorem 1.4 and Theorem 2.1. It is noteworthy that i.i.d. is the only assumption made on random variables X_i, $i = 1, 2, \ldots$ which define $\pi_A^\ell = |\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \in A)|$. As a consequence, we see that this corollary contains the Glivenko-Cantelli theorem of Wolfowitz [25] for half-spaces in $\mathbb{R}^d$. In fact, we see the corollary goes farther both in that the supremum is taken over a larger class and in that the provision of a rate of convergence estimate is given.

Our method for proving the required estimates on the tail of the distribution of π^ℓ consists in decomposing S into subclasses which can be more efficiently estimated. First we require a known result ([23], p. 273) which is obtained by a means quite similar to the Vapnik-Chervonenkis lemma of Chapter I.

One-sided estimates of Vapnik and Chervonenkis: For any class S we have

$$(1)\quad P\left(\sup_{A\in S} \frac{P(X_1 \in A) - \nu_A(\ell)}{\sqrt{P(X_1 \in A)}} > \epsilon\right) < 16m^S(2\ell)\exp(-\epsilon^2\ell/4)$$

and

$$(2)\quad P\left(\sup_{A\in S} \frac{-P(X_1 \in A) + \nu_A(\ell)}{\sqrt{1 - P(X_1 \in A)}} > \epsilon\right) \leq 16m^S(2\ell)\exp(-\epsilon^2\ell/4)$$

where $m^S(2\ell) = \max \Delta^S(X_1, X_2, \ldots , X_{2\ell})$ where the maximum is taken over all $\omega \in \Omega$.

In order to pursue the result of the corollary to the level of best possible rate of convergence, we are forced to develop sharper estimates on the tail of the distribution of π^ℓ. Once this is done, we then connect the function π^ℓ to related submartingales. This then makes Kolmogorov's inequality available and allows us a rather classical path to the LIL. In order to pinpoint our argument, we prove a general LIL for submartingales which is much more convenient for our purposes than the related result of E. Csaki [5]. This general result and our tail estimates then provide a unified proof of the LIL for π^ℓ for the polynomial regions and the rectangular regions of the classical $(d \geq 1)$ empirical distribution function.

To make good use of these estimates we show that each element of S is the union of a small simple set and an element from a moderate subclass of S. We begin by making S into a metric space by defining the distance $\lambda(A,B) = P(X_1 \in (A \cup B)\setminus(A \cap B))$ for $A,B \in S$. The required result is then essentially a lemma in metric entropy.

<u>Theorem 2.2</u>. Let S be the class of half-spaces in $\mathbb{R}^d$ and let λ be the metric defined above. If X_1 has bounded density and finite

expectation, then there is a constant $\beta = \beta(\lambda)$ such that we have:

For any integer $k > 0$ there is a subset S_k of S with $|S_k| \leq \beta k^{d^2}$ such that any element of S has λ distance at most $1/k$ from an element of S_k.

Proof. Since X_1 has finite expectation we have by Chebyshev's inequality that there is a constant γ such that for all k, $P(X_1 \notin C(\gamma k)) \leq 1/2k$ where $C(k\gamma)$ is the cube of side $k\gamma$ and center 0.

To accommodate half spaces A whose boundary does not intersect $C(\gamma k)$ we need only include in S_k a half space which contains $C(\gamma k)$ and one which does not intersect $C(\gamma k)$. One of these half spaces will surely have distance less than $1/k$ from A.

The remaining necessary elements of S_k are given by forming all the hyperplanes determined by a finite subset of the boundary of $C(\gamma k)$ and then including in S_k the closed half spaces determined by these hyperplanes. We first partition each of the edges of the cube $C(\gamma k)$ into segments of length less than or equal to $\alpha > 0$. Then we let H denote the class of all hyperplanes determined by the endpoints of these segments. We note that an arbitrary hyperplane which intersects the cube $C(\gamma k)$ will pass between two elements of H in the interior of $C(\gamma k)$. We therefore need to calculate the upper bound of the probability that X_1 be between two consecutive elements of H. Since the diagonal of the cube is of length $\sqrt{d}\gamma k$ we have for $\lambda(A) = P(X_1 \in A)$ $A \in B$ that the λ measure between any two consecutive hyperplanes is less than $(\sqrt{d}\gamma k)^{d-1} \cdot \alpha \cdot M$ where M is the bound on the density of X_1. Finally, taking $\alpha = [2kd(\gamma k)^{d-1}M]^{-1}$ we obtain the fact that

$$|H| \leq \binom{d2^{d-1}}{d}(2kd(\gamma k)^{d-1}M)^d \leq Ck^{d^2}$$

where C is a constant depending only on λ. By our previous estimate on the measure between consecutive elements of H and the measure of $C(\gamma k)^c$ we immediately see that S_k has the required approximation property. Finally, we have $|S_k| \leq 2 + 2|H|$ so the theorem is proved.

Theorem 2.3. Let S be the class of half-spaces in $\mathbb{R}^d$ and suppose the random variables X_i have finite expectation and bounded density. Then there is a constant C depending only on the distribution of X_1 such that

$$P(\sqrt{\ell}\,\pi^{\ell} > r) \leq C(\log \ell)^{d^2} \exp(-\frac{r^2}{8}) \text{ .}$$

Proof. Let $\varphi(\ell)$ be an integer valued function which will be specified later in the proof, and let $\lambda(A) = P(X_1 \in A)$ for $A \in (\mathbb{R}^d)$. By Theorem 2.2 we select a finite subset $S(\ell)$ of S such that $|S(\ell)| \leq \beta\varphi(\ell)^{d^2}$ with the property that any element of S is at most a distance of $1/\varphi(\ell)$ from $S(\ell)$ in the λ metric. Next denoting by D the class of subsets of $\mathbb{R}^d$ which are given as the symmetric difference of pairs of elements of S, we have

$$(1)\quad P(\sup_{A\in S}|\lambda(A)-\nu_A'(\ell)| > r) \leq P(\sup_{A\in D,\lambda(A)\varphi(\ell)\leq 1}|\lambda(A)-\nu_A'(\ell)| \geq r/2)$$

$$+ P(\sup_{A\in S(\ell)}|\lambda(A) - \nu_A'(\ell)| \geq r/2) \text{ .}$$

Next we note that

$$(2) \qquad P\Big(\sup_{A\in D;\lambda(A)\varphi(\ell)\leq 1} -\lambda(A) + \nu_A'(\ell) \geq r/2\Big)$$

$$\leq P\Big(\sup_{A\in D;\lambda(A)\geq 1-1/\varphi(\ell)} -\lambda(A) + \nu_A'(\ell) \geq r/4\Big)$$

$$+ P\Big(\sup_{A\in S(\ell)} |\lambda(A) - \nu_A'(A)| \geq r/4\Big)$$

and consequently

$$(3) \quad P\Big(\sup_{A\in S}|\lambda(A) - \nu_A'(\ell)| \geq r\Big) \leq 2P\Big(\sup_{A\in S(\ell)} |\lambda(A) - \nu_A'(\ell)| \geq r/4\Big)$$

$$+ P\Big(\sup_{A\in D;\lambda(A)\geq 1-1/\varphi(\ell)} - \lambda(A) + \nu_A'(\ell) \geq r/4\Big)$$

$$+ P\Big(\sup_{A\in D;\lambda(A)\leq 1/\varphi(\ell)} \lambda(A) - \nu_A'(\ell) \geq r/4\Big) .$$

By the classical estimate on the tail of the binomial distribution we have a constant C such that

$$(4) \quad P\Big(\sup_{A\in S(\ell)} |\lambda(A)-\nu_A'(\ell)| > r\Big) \leq |S(\ell)|\cdot C\cdot e^{-2r^2\ell} \leq C\beta\varphi(\ell)d^2 e^{-2r^2\ell} .$$

Also $m^D(2\ell) \leq (16)^d\ell^{2d}$ so we apply Theorem 2.3 to (3) and obtain our basic estimate for $\ell \geq 2r^{-2}$

$$(5) \qquad P(\sqrt{\ell}\,\pi^\ell \geq r) \leq 2\beta\varphi(\ell)^{d^2}\cdot C \exp(-r^2/8)$$

$$+ 2(16)^{d+1}\ell^{2d}\exp(-r^2\varphi(\ell)/4\cdot 16) .$$

Now let $\varphi(\ell) = 1 + [(2 + 2d \log \ell)4\cdot 16]$ and note for $r \geq 1$ $\ell^{2d}\exp(-r^2\varphi(\ell)/4\cdot 16) \leq \exp(-2r^2)$, and $\varphi(\ell)^{d^2} \leq \gamma(\log \ell)^{d^2}$ for a constant $\gamma = \gamma(d)$. The theorem is thus proved.

The role of the class of half spaces is brought out by its immediate applicability to the study of more general regions by means of the

Veronesi (or other) mappings.

Theorem 2.4. Let $S = \{A_g : g \in P_{d,n}\}$ be the class of polynomial regions in $\mathbb{R}^d$ of degree not greater than n. Suppose further that the random variables X_i have finite n^{th} moment and bounded density. Then there is a constant C such that

$$P(\sqrt{\ell}\,\pi^{\ell} > r) \leq C(\log \ell)^{\tau^2} \exp(-r^2/8)$$

where $\tau = \binom{n+d}{n}$.

Proof. We let $\psi : \mathbb{R}^d \to \mathbb{R}^\tau$ be the Veronesi map defined in the second proof of Theorem 2.1. By our hypothesis, we obtain that $\psi(X_i)$, $i = 1, 2, \ldots$ are i.i.d. random variables with finite expectation and bounded densities. If A_ω is a half space in $\mathbb{R}^\tau$ given by $\{z : z\cdot\omega > 0\}$ where $\omega \in \mathbb{R}^\tau$ is the vector formed by the coefficients of $g \in P_{n,d}$ given in lexicographical order, then we have the basic observation:

$$\psi(X_i) \in A_\omega \text{ if and only if } X_i \in A_g .$$

With this identification the present result follows directly from Theorem 2.3.

Now we embark on the second part of our program; we establish the connection between our inequalities and the theory of submartingales. This is done completely in the next two results.

Theorem 2.5. Let $\{Y(\ell), \mathcal{F}_\ell\}$ be a submartingale such that $P\{Y(\ell) > u\} \leq C(\log \ell)^\tau \exp(-\alpha u^2/\ell)$, then

$$P(\limsup_{\ell \to \infty} Y(\ell)(\log\log \ell/\ell)^{1/2} \leq D) = 1$$

where $(\tau+1)^{-1/2}D = (4\alpha)^{1/2} + (4\alpha)^{-3/2}$.

Proof. We first estimate the moment generating function of $Y(\ell)$.

$$E(\exp(tY(\ell))) = 1 + t\int_0^\infty \exp(tu)P(Y(\ell) > u)du$$

$$\leq 1 + ct(\log \ell)^\tau \int_{-\infty}^\infty \exp(tu - \alpha u^2/\ell)du$$

$$\leq 1 + ct(\log \ell)^\tau \sqrt{\ell} \exp(t^2\ell/4\alpha) .$$

Now since $Y(\ell)$ is a submartingale so is $\exp(tY(\ell))$ and Kolmogorov's inequality can be applied. We let n_k be an increasing sequence of integers, and observe for $\epsilon > 0$

$$P_k = P(\sup_{n_k < \ell \leq n_{k+1}} Y(\ell) \geq (D+\epsilon)(n_k/\log\log n_k)^{1/2})$$

$$\leq P(\sup_{n_k < \ell \leq n_{k+1}} \exp(tY(\ell)) \geq \exp(t(D+\epsilon)(n_k/\log\log n_k)^{1/2})$$

$$\leq E(\exp(tY(n_{k+1}))\cdot\exp(-t(D+\epsilon)(n_k/\log\log n_k)^{1/2}) .$$

It remains to choose n_k and t in such a way that $\sum_{k=1}^{\infty} P_k < \infty$ which then would complete the proof.

Let A and B denote positive constants and set $n_k = [1+\epsilon A]^k$ and $t_k = B(\log\log n_{k+1}/n_{k+1})^{1/2}$. By our previous estimates $P_k \leq \gamma_k\gamma'_k$ where

$$\gamma_k = 1+cB \text{ loglog } n_{k+1}(\log n_k)^{\tau}(n_k/n_{k+1})^{1/2}\exp(B^2(\text{loglog } n_{k+1})\cdot n_k/4\alpha n_{k+1})$$

$$\gamma_k' = \exp(-B(\text{loglog } n_{k+1} \text{ loglog } n_k)^{1/2}\cdot(D+\epsilon)(n_k/n_{k+1})^{1/2}) \text{ .}$$

For any $\delta > 0$ we can choose $A = A(\epsilon)$ such that $1-\delta \leq n_k/n_{k+1} \leq 1+\delta$ for all k. Further we note

$$(\text{loglog } n_{k+1} \text{ loglog } n_k)^{1/2} \leq \text{loglog}(n_{k+1}) + \lambda \text{ ,}$$

λ a constant. We therefore have

$$\gamma_k \leq 1 + cB(1+\delta)^{1/2} \text{ loglog } n_{k+1}(\log n_k)^{\tau}(\log n_{k+1})^{B^2(1+\delta)/4\alpha}$$

$$\gamma_k' \leq \exp(-B\lambda(D+\epsilon)(1-\delta)^{1/2})(\log n_{k+1})^{-B(D+\epsilon)(1-\delta)^{1/2}} \text{ .}$$

The remaining choices are easy; by taking B and D such that $\tau + B^2/4\alpha - BD = -1$ and then choosing A to make δ small we see that $\sum P_k < \infty$. Finally calculus suggest that $B = \sqrt{(\tau+1)/4\alpha}$ is the optimal choice of B, so $D = (1+\tau)^{1/2}((4\alpha)^{1/2} + (4\alpha)^{-3/2})$ is a sufficient choice to prove the theorem.

Theorem 2.6. Let S denote the class of open half spaces and suppose the random variables X_i have absolutely continuous distribution. Then the random variables $Y_n = n \sup_{A\in S}(\lambda(A) - \nu_A'(n))$ and $Z_n = n \sup_{A\in S}(\nu_A'(n) - \lambda(A))$ are submartingales with respect to the fields $\mathcal{F}_n = \sigma(X_1, X_2, \ldots , X_n)$.

Proof. If $p = (s, p_1, p_2, \ldots , p_d) \in \mathbb{Z}_2 \times \mathbb{R}^d$, then p is associated in a natural way with an element A_p of S as follows: the p_i, $i = 1, 2, \ldots , d$ denote the intercepts of ∂A_p the boundary of A_p

with the axes and $s = 1$ if A_p is the region above ∂A_p (and $s = 0$ if A_p is the region below ∂A_p). We see then by the absolute continuity of λ that $\lambda(A_p) - \nu'_{A_p}(n)$ will for each ω attain it's supremum on a set $A_p(n,\omega)$. Also by continuity and the fact that $\nu'_A(n)$ assumes only finitely many values, we can show that $p(n,\omega)$ is a random variable measurable with respect to $\mathcal{F}_n$. It is now easy to check that Y_n is a submartingale:

$$E(Y_{n+1}|\mathcal{F}_n) \geq E((n+1)\lambda(A_{p(n,\omega)} - \sum_{i=1}^{n+1} 1_{A(p(n,\omega))}(X_i)| \quad _n)$$

$$\leq Y_n + E(\lambda(A_{p(n,\omega)}) - 1_{A_{p(n,\omega)}}(X_{n+1})$$

$$\geq Y_n \; .$$

Hence we obtain that $(Y_n, \mathcal{F}_n)$ is a submartingale. The same proof also shows $(Z_n, \mathcal{F}_n)$ is a submartingale, so the proof is complete.

The preceding work can now be brought together to provide the principal result of this chapter.

Theorem 2.7. Let $S = \{A_g : g \in P_{n,d}\}$ be the class of polynomial regions in $\mathbb{R}^d$ of degree not greater than n. For random variables X_i, $i = 1, 2, \ldots$ with bounded density and finite n^{th} moment we have

$$P(\sqrt{\ell}\, \pi^{\ell} > D(\text{loglog } \ell)^{1/2} \text{ i.o.}) = 0$$

where $D = (2^{-1/2} + 2^{3/2})(1 + \binom{n+d}{d}^2)^{1/2}$.

Proof. All that remains is to assemble the collected pieces. By Theorem 2.6 and the Veronesi mapping we see that

$Y_n = n \sup_{A\in S}(P(X_1 \in A) - \nu'_A(n))$ and $Z_n = n \sup_{A\in S}(\nu'_A(n) - P(X_1 \in A))$ are submartingales. The estimate of Theorem 2.4 is applicable to both Y_n and Z_n which are then subject to the conclusion of Theorem 2.5 and then yields the present result.

The assumptions of bounded density and the existence of moments in the preceding results may at first seem like a significant restriction compared to the corresponding results for the classical empirical distribution function. To understand how our result differs from the classical case, consider the class

$$S = \{A : A = \{y \in \mathbb{R}^d : y_i \leq x_i, \ i = 1,2,\ldots,d\}; \ x = (x_1,x_2,\ldots,x_d) \in \mathbb{R}^d\}$$

which provides the correspondence

$$\pi^\ell = \sup_{x\in\mathbb{R}^d} |F_n(x) - F(x)|$$

between π^ℓ and we note that the empirical distribution function $F_n(x)$ has some very special properties. In particular, this class is invariant under monotone transformations of the coordinates x_i, $i = 1, 2, \ldots, d$. By a familiar device, this reduces the study of distributions F with absolute continuity to the study of the uniform distribution on the cube. This observation together with the trival fact that $\Delta^S(x_1, x_2, \ldots, x_\ell) \leq (\ell+1)^d$ shows that the proof of Theorem 2.7 is sufficient to yield the following result:

Theorem 2.8. Let $F(x)$ be an absolutely continuous distribution function on $\mathbb{R}^d$ and let $F_n(x)$ be the corresponding empirical distribution function. We have

$$P(\sqrt{n} \sup_{x\in\mathbb{R}^d} |F_n(x) - F(x)| \geq D\sqrt{\log\log n} \text{ i.o.}) = 0$$

for $D = (2^{-1/2} + 2^{3/2})(1 + d^2)^{1/2}$.

We note that our constant D is not best possible, and indeed the best possible value of D has already been shown in this case to be $2^{1/2}$ by Kiefer [13]. Since our method attacks a much more general problem than this last result, there is good reason to expect some imprecision in its application to some special cases. In fact, it seems remarkable that one can come so close to such a delicate result as Kiefer's by such a general attack. In this regard it is of interest to note that in the case $d = 1$ Chung [4] was able to determine completely the upper and lower classes of functions for the law of the iterated logarithm so at each stage of generalization there has been a corresponding loss of precision.

CHAPTER III

COMBINATORIAL RESULTS AND APPLICATIONS

The combinatorial function $\Delta^S(x_1, x_2, \dots, x_\ell)$ has been seen in the preceding chapters to have a critical relationship to the probabilistic behavior of ρ^ℓ and π^ℓ. Experience quickly leads one to discover that $\Delta^S(x_1, x_2, \dots, x_\ell)$ may be difficult to compute, and for this reason it is of importance to discover other functions which are easier to compute yet which provide information about $\Delta^S(x_1, x_2, \dots, x_\ell)$. We therefore introduce a new combinatorial function $K^S(x_1, x_2, \dots, x_\ell)$ defined by

$$K^S(x_1, x_2, \dots, x_\ell) = \max\{k : \Delta^S(x_{i_1}, x_{i_2}, \dots, x_{i_k}) = 2^k ;$$

$$\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\} \subset \{x_1, x_2, \dots, x_\ell\}\} .$$

This definition can be otherwise expressed by saying that $K^S(x_1, x_2, \dots, x_\ell)$ is equal to the order of the largest subset of $\{x_1, x_2, \dots, x_\ell\}$ such that S partitions the subset in all possible ways. After some purely combinatorial work (which entails a new proof and generalization of a theorem of N. Sauer [17]) we obtain the rather surprising result that the limit behaviors of $\log \Delta^S(X_1, X_2, \dots, X_\ell)$ and $K^S(X_1, X_2, \dots, X_\ell)$ are essentially equivalent. We then provide an explicit calculation of $\lim_{\ell \to \infty} K^S(X_1, X_2, \dots, X_\ell)$ when S is the class of convex subsets of $\mathbb{R}^d$. This calculation then provides as a corollary a result of Glivenko-Cantelli type for the class of convex sets which was proved earlier by R. Ranga Rao [15]. Next we are able to estimate

$K^S(X_1, X_2, \ldots, X_\ell)$ where S is the class of lower layers in $\mathbb{R}^d$. This enables us to make a positive step toward the conjecture of Robertson and Wright [16] concerning the law of the iterated logarithm for lower layers. We do not prove the conjecture but we are able to provide the first result which goes beyond the level of the law of large numbers. Finally we show that the theory of lower layers can be applied to obtain results for the class of convex sets.

We now begin our combinatorial work.

Theorem 3.1. Let M be an $\ell \times c$ matrix with entries from $\{1, 2, \ldots, s\}$ and let Δ be the number of distinct columns of M. If k is the largest integer such that there are k rows which form a matrix with s^k columns, then we have

$$\Delta \leq s^\ell - \sum_{j=k}^{\ell} \binom{\ell}{j}(s-1)^{\ell-j} .$$

Further we note that for any values of s, ℓ, and k there is a matrix M such that equality holds.

Proof. We first provide an example which shows this result is best possible and which suggests the specific form of the inequality. We define M by deleting from the $\ell \times s^\ell$ matrix all columns with k or more 1's. The resulting matrix is readily seen to have

$$s^\ell - \sum_{j=k}^{\ell} \binom{\ell}{j}(s-1)^{\ell-j}$$

distinct columns, yet we see that no matrix formed by k of its rows can have s^k distinct columns since any such matrix lacks the column consisting of all 1's.

We proceed to prove the theorem by showing that the only way equality can hold is if a consistency analogous to that in the example holds among the missing column vectors.

We let C_i, $i = 1, 2, \ldots, \binom{\ell}{k} = \tau$ be a list of the k subsets of $\{1, 2, \ldots, \ell\}$ and we write $M(C_i)$ for the matrix formed by the corresponding rows of M. By the hypothesis of the theorem we may assume that none of the matrices $M(C_i)$ has s^k distinct columns, and for each i we select a vector $v_i = (v_i(1), v_i(2), \ldots, v_i(k))$ which is not a column vector of $M(C_i)$. This allows us to define a function $f_i : C_i \to \{1, 2, \ldots, s\}$ by letting $f_i(j) = v_i(r)$ if j is the r^{th} element of C_i. Next we define $Z_i = Z_i(f_i)$ to be the set of all column vectors $w = (w_1, w_2, \ldots, w_\ell)$ of the $\ell \times s^\ell$ matrix such that $\omega_\alpha = f_i(\alpha)$ for $\alpha \in c_i$. We can assume that M consists of the matrix formed by all column vectors of the $\ell \times s^\ell$ matrix except those in $Z_1 \cup Z_2 \cup \cdots \cup Z_\tau$, and the observation allows us to assume that $\Delta = s^\ell - |Z_1 \cup Z_2 \cup \cdots \cup Z_\tau|$. We will say that Z_i, $i = 1, 2, \ldots, \tau$ are consistent if $f_i = f_j$ on $C_i \cap C_j$; the theorem is then seen to be a consequence of the following lemma.

<u>Lemma</u>. $|Z_1 \cup Z_2 \cup \cdots \cup Z_\tau| \geq \sum_{j=k}^{\ell} \binom{\ell}{j}(s-1)^{\ell-j}$ for all choices of f_i, $i = 1, 2, \ldots, \tau$ and equality holds if and only if the Z_i are consistent.

Proof. If the Z_i are consistent, there is a column vector $v = (v_1, v_2, \dots , v_\ell)$ of the $\ell \times s^\ell$ matrix such that $f_i(j) = v_j$ for $j \in C_i$ and all $i = 1, 2, \dots , \tau$. We see then that $Z_1 \cup Z_2 \cup \cdots \cup Z_\tau$ is the collection of all column vectors of the $\ell \times s^\ell$ matrix which agree with v in k or more places. This then proves that $|Z_1 \cup Z_2 \cup \cdots \cup Z_\tau| = \sum_{j=k}^{\ell} \binom{\ell}{j}(s-1)^{\ell-j}$ if the Z_i are consistent.

Suppose now the Z_i are not consistent. For any $\beta \in \{1,2,\dots,s\}$ we define $\Phi_\beta(Z)$ to be the collection of columns of the $\ell \times s^\ell$ matrix given by $\hat{f}_i(j)$ where

$$\hat{f}_i(j) = \begin{cases} f_i(j) & j \neq \beta \quad i \in c_i \\ 1 & j = \beta \quad i \in c_i \, . \end{cases}$$

and where $Z = Z_1 \cup Z_2 \cup \cdots \cup Z_\tau$. It is immediate that $|\Phi_\beta(Z)| \leq |Z|$ and that $\Phi_\tau \Phi_{\tau-1} \cdots \Phi_1(Z)$ is a consistent system. We can assume therefore that the original Z_i, $i = 1, 2, \dots , \tau$ fail to be consistent in exactly one place β. With this assumption we will prove $|\Phi_\beta(Z)| < |Z|$ and thus prove the lemma. To carry this plan out let $A_\beta = \bigcup_{i:\beta \in C_i} Z_i$ and $A'_\beta = \bigcup_{i:\beta \notin C_i} Z_i$ and write $Z = A_\beta \cup A'_\beta$. Next we have

$$A_\beta = \Big(\bigcup_{\substack{i:\beta\in C_i \\ f_i(\beta)=1}} Z_i\Big) \cup \Big(\bigcup_{\substack{i:\beta\in C_i \\ f_i(\beta)=2}} Z_i\Big) \cup \cdots \cup \Big(\bigcup_{\substack{i:\beta\in C_i \\ f_i(\beta)=S}} Z_i\Big)$$

is a disjoint decomposition of A_β. Further we note that $\Phi_\beta(A'_\beta) = A'_\beta$

and $\Phi_\beta(A'_\beta \setminus A_\beta) = A'_\beta \setminus A_\beta$. The critical point is that

$$|\Phi_\beta(A_\beta)| < |A_\beta|$$

since Φ_β maps a disjoint union into a non-disjoint union. This proves $|\Phi_\beta(Z)| < |Z|$ and completes the lemma.

Corollary. $\Delta^S(X_1, X_2, \ldots , X_\ell) \leq \sum_{j=o}^{\beta} \binom{\ell}{j}$ where

$$\beta = K^S(X_1, X_2, \ldots , X_\ell) .$$

Proof. We define a matrix M by taking as columns all of the vectors $v = (v_1, v_2, \ldots , v_\ell)$ where $v_i = 1$ if $X_i \in A$ and $v_i = 0$ if $X_i \notin A$ and $A \in S$. The corollary is the direct consequence of Theorem 3.1.

This corollary shows the usefulness of Theorem 3.1 and will be the only way Theorem 3.1 enters into the succeeding work. The corollary has been proved before and as stated is due to N. Sauer [17] who provided the result in response to a question of P. Erdős. Quite independently the logician, S. Shelah, had proved the corollary but did not publish his proof, as the paper of N. Sauer would appear at essentially the same time. Thirdly, we come to this theorem via the work of Vapnik and Chervonenkis [21] and [22] where a similar result is proved but which is not quite as sharp as the above. Regarding the possible merits of Theorem 3.1, we note that our method is the only one which seems to work for matrices with other than $\{0,1\}$ entries, and that our proof gives a complete understanding of the circumstance when equality can occur. Even in the known case of

$\{0,1\}$ entries our proof is perhaps of interest as it provides a proof without the use of recurrence relations and inductions, but instead is an explicit constructive analysis.

In the next theorem we collect the basic properties of $K^S(X_1, X_2, \ldots, X_\ell)$ in much the same way as we collected those of $\Delta^S(X_1, X_2, \ldots, X_\ell)$ in Theorem 1.1. In fact, the methods are so similar and easy we omit the proof.

Theorem 3.2. Let $K^S(X_t, X_{t+1}, \ldots, X_v) = Y_{t,v}$, then we have the following properties of the process $\{Y_{t,v} \quad t > 0,\ v > t\}$:

(1) $Y_{t,v}$ is stationary, that is, the process defined by $Y'_{t,v} = Y_{t+1,v+1}$ has the same finite dimensional distributions as $Y_{t,v}$.

(2) $Y_{t,v}$ is subadditive, that is, $Y_{t,v} \leq Y_{t,u} + Y_{u,v}$ for any $t < u < v$.

(3) Each $Y_{t,v}$ has finite expectation, in fact, $1 \leq Y_{t,v} \leq v-t$.

(4) $\lim_{\ell \to \infty} \ell^{-1}E(Y_{1,\ell}) = \lim_{\ell \to \infty} \ell^{-1}E(Y_{1,\ell}) = k \geq 0$, k constant.

(5) $\lim_{\ell \to \infty} \ell^{-1}EY_{1,\ell} = k \geq 0$ a.s.

We can now show explicitly how the limit behavior of $K^S(X_1,X_2,\ldots,X_\ell)$ and that of $\log \Delta^S(X_1, X_2, \ldots, X_\ell)$ can be considered as equivalent. This gives the basic link to applications of $K^S(X_1, X_2, \ldots, X_\ell)$ to our original probabilistic concerns.

Theorem 3.3. Let

$$c = \lim_{\ell \to \infty} \ell^{-1} \Delta^S(X_1, X_2, \ldots , X_\ell)$$

and

$$k = \lim_{\ell \to \infty} \ell^{-1} K^S(X_1, X_2, \ldots , X_\ell) .$$

Then $c = 0$ if and only if $k = 0$. Further, if $k > 0$, then

$$k \leq c \leq -\log(k^k(1-k)^{1-k}) .$$

Proof. Since $2^{K^S(X_1,X_2,\ldots,X_\ell)} \leq \Delta^S(X_1, X_2, \ldots , X_\ell)$, we have immediately that $k \leq c$ for any value of c. Next suppose $\alpha > 0$ and that $K^S(X_1, X_2, \ldots , X_\ell) \leq \alpha\ell$ on a set E. By Theorem 3.1 we have $\Delta^S(X_1, X_2, \ldots , X_\ell) \leq \sum_{j=o}^{[\alpha\ell]} \binom{\ell}{j}$ on E. Also $\sum_{j=o}^{[\alpha\ell]} \binom{\ell}{j} \leq (\alpha\ell)\binom{\ell}{[\alpha\ell]}$ and by Stirling's formula we have $\log\binom{\rho}{[\alpha\ell]} \leq \ell \log\left(\frac{1}{\alpha^\alpha(1-\alpha)^{1-\alpha}}\right) + \gamma$ for a constant $\gamma > 0$ and all ℓ. Hence we obtain $\lim_{\ell \to \infty} \ell^{-1}\log \Delta^S(X_1, X_2, \ldots , X_\ell) \leq -\log(\alpha^\alpha(1-\alpha)^{1-\alpha})$. This estimate and the definition of k then complete the proof of the theorem.

As promised in the beginning of this chapter, we show that $\lim_{\ell \to \infty} \ell^{-1}K^S(X_1, X_2, \ldots , X_\ell)$ can be explicitly calculated for several important classes S. The first such calculation is provided in the following result.

Theorem 3.4. Let X_i, $i = 1, 2, \ldots$ be a stationary ergodic process with values in $\mathbb{R}^d$. If S is the class of convex Borel subsets of $\mathbb{R}^d$, then $\lim_{\ell \to \infty} \ell^{-1} K^S(X_1, X_2, \ldots, X_\ell) = \sup_{A \in S} P(X_1 \in \partial A)$.

Proof. For any convex set A we have

$$K^S(X_1, X_2, \ldots, X_\ell) \geq \sum_{i=1}^{\ell} 1_{\partial A}(X_i)$$

since if $\{y_1, y_2, \ldots, y_k\} \subset \partial A$ we have $K^S(y_1, y_2, \ldots, y_k) = 2^k$. By the ergodic theorem we then have

$$\lim_{\ell \to \infty} \ell^{-1} K^S(X_1, X_2, \ldots, X_\ell) \geq P(X_i \in \partial A)$$

for all $A \in S$.

To prove an inequality in the opposite direction, we employ a compactness argument based on the famous Blaschke selection theorem (see Eggleston [9], p. 64). To make this result available, we first require a truncation lemma which will allow us to focus on a compact subset of $\mathbb{R}^d$.

Lemma. Suppose $Xi = 1, 2, \ldots$ is a stationary ergodic process which takes values in $\mathbb{R}^d$. Let B_m be the ball about 0 of radius m. Then let

$$c_m = \lim_{\ell \to \infty} \ell^{-1} \log \Delta^S(\{X_1, X_2, \ldots, X_\ell\} \cap B_m)$$

and

$$k_m = \lim_{\ell \to \infty} K^S(\{X_1, X_2, \ldots, X_\ell\} \cap B_m) .$$

Then we have

$$\lim_{m \to \infty} c_m = c = \lim_{\ell \to \infty} \ell^{-1} \log \Delta^S(X_1, X_2, \dots , X_\ell)$$

and

$$\lim_{m \to \infty} k_m = k = \lim_{\ell \to \infty} \ell^{-1} K^S(X_1, X_2, \dots , X_\ell) .$$

Proof of Lemma. We note that

$$c_m \leq c \leq \lim_{\ell \to \infty} \ell^{-1} \log \Delta^S(\{X_1, X_2, \dots , X_\ell\} \cap B_m) + \lim_{\ell \to \infty} \ell^{-1} \sum_{i=1}^{\ell} 1_{B_m^c}(X_i).$$

Hence $c_m \leq c \leq c_m + P(X_1 \in B_m^c)$ and we thus obtain $\lim_{m \to \infty} c_m = c$.

The proof that $\lim_{m \to \infty} k_m = k$ is similar.

We now return to the proof of the theorem. For each integer $r > 0$ we obtain by Blaschke's theorem a finite set $S(r)$ of convex subsets of B_m such that the following holds:

If C is a convex subset of B_m, then there is an element A of $S(r)$ such that

$$\partial C \subset \{x : \inf_{y \in \partial A} (s-y) < 1/r\} \equiv T(A, 1/r) .$$

(Here we introduce the thickened boundary $T(A,\delta) = \{x : \inf_{y \in \partial A} (x-y) < \delta\}$.)

By the definition of $S(r)$ we have

$$K^S(\{X_1, X_2, \dots , X_\ell\} \cap B_m) = \max_{\substack{C \in S \\ C \subset B_m}} \sum_{i=1}^{\ell} 1_{\partial C}(X_i)$$

$$\leq \max_{A \in S(r)} \sum_{i=1}^{\ell} 1_{T(A,1/r)}(X_i) .$$

On applying the ergodic theorem we obtain

$$(1) \qquad k_m \leq \max_{A \in S(r)} P(X_1 \in T(A, 1/r)) .$$

To take advantage of (1) we choose for each r an $A_r \in S(r)$ such that

$$(2) \qquad k_m \leq P(X_1 \in T(A_r, 1/r)) \text{ for all } r .$$

Now again by Blaschke's theorem, there is a convex subset A' of B_m such that we have the following:

Given $\delta > 0$, there is an $R = R(\delta)$ such that for $r \geq R$ we have $T(A, 1/r) \subset T(A',\delta)$.

This proves by (2) that

$$k_m \leq P(X_1 \in T(A',\delta))$$

which shows by an application of the dominated convergence theorem that

$$(3) \qquad k_m \leq P(X_1 \in \partial A') \leq \sup_{A \in C} P(X_1 \in \partial A) .$$

Since the last inequality is independent of m, the theorem follows on letting m go to infinity.

We now consider an immediate consequence of the preceding theorem working with the earlier Theorems 1.3 and 3.1.

Corollary. (R. Ranga Rao) Let X_i, $i = 1, 2, \ldots$ be i.i.d. with values in $\mathbb{R}^d$ and let S be the class of convex Borel sets in $\mathbb{R}^d$. We have $\pi^\ell \to 0$ a.s. if and only if

$$\sup_{C \in S} P(X_1 \in \partial C) = 0 .$$

The virtue of Theorem 3.4 does not rest so much in the new proof of the theorem of R. Ranga Rao ([15], p. 674) as in its relation to subadditive ergodic theory. In fact, the above corollary was provided by R. Ranga Rao in the ergodic stationary case where we require independence. What Theorem 3.4 particularly accomplishes is the specific identification of a limit obtained in general by the theory of subadditive processes, and according to J. F. C. Kingman such identifications hold the "pride of place among the unsolved problems of subadditive ergodic theory" (see Kingman [14], p. 897).

A result which is completely analogous to Theorem 3.4 can also be given for the class S of "lower layers" in $\mathbb{R}^d$. This class is certainly less well known than the class of convex sets, yet it enters naturally into several probabilistic and statistical contexts. We say A is a lower layer in $\mathbb{R}^d$ if A is a Borel subset of $\mathbb{R}^d$ such that $y = (y_1, y_2, \ldots, y_d) \in A$ implies $x = (x_1, x_2, \ldots, x_d) \in A$ if $x_i \leq y_i$ for $i = 1, 2, \ldots, d$. In particular, we see that if $d = 2$, the class of lower layers is exactly the class of sets $B = \{(x_1, x_2) : x_2 \leq f(x_1)\}$ where f is a monotone decreasing function. One should note that there are many lower layers which are not convex and vice versa. We now state our result concerning this class.

Theorem 3.5. Let X_i, $i = 1, 2, \ldots$ be a stationary ergodic process with values in $\mathbb{R}^d$. If S is the class of lower layers in $\mathbb{R}^d$, then

$$\lim_{\ell \to \infty} \ell^{-1} K^S(X_1, X_2, \ldots, X_\ell) = \sup_{A \in S} P(X_1 \in \partial A) .$$

We omit the proof of Theorem 3.5 since it exactly duplicates our proof of the corresponding result for convex sets. The crucial ingredients of the proof are (1) the fact that $K^S(\{y_1, \ldots, y_k\}) = 2^K$ if and only if y_i, $i = 1, 2, \ldots, k$ are on the boundary of a lower layer and (2) there is a compactness result due to Brunk, H. D., Ewing, G. M. and Utz, W. R. [3] which performs the same function for the lower layers that Blaschke's theorem does for convex sets. The truncation lemma and the assembling of the pieces is easily checked to follow in the manner of Theorem 3.4. Further we note that we obtain a corollary to Theorem 3.5 just as we obtained one from Theorem 3.4. This time the result is essentially a Glivenko-Cantelli theorem due to J. Dehardt [7].

Our main concern with the theory of lower layers does not rest so much in Theorem 3.5 as in the conjecture of T. Robertson and F. T. Wright [16] that the law of the iterated logarithm holds for π^ℓ. Explicitly it is conjectured that

$$P\left(\overline{\lim_{\ell \to \infty}} \left(\frac{\ell}{\operatorname{loglog} \ell} \right)^{1/2} \sup_{A \in S} \left| \ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \in A) \right| = \sqrt{2} \right) = 1 .$$

In our approach to this problem we will consider only the case $d = 2$, and our principal objective will be to obtain good estimates on the tails of π^ℓ and ρ^ℓ. Such estimates are very likely to have value in any attempt to prove the above conjecture. Although we make only modest progress in this direction, it is nonetheless the first step to go beyond Theorem 3.4.

Our tools are actually best designed for the study of ρ^ℓ so we first point out a general relationship between π^ℓ and ρ^ℓ. If S is <u>any</u> class such that π^ℓ and ρ^ℓ are measurable, then $\pi^\ell \leq E(\rho^\ell | X_1, X_2, \ldots, X_\ell)$ a.s. To see this fact, we note that

$$\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \epsilon A) = E(\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - \ell^{-1} \sum_{i=\ell+1}^{2\ell} 1_A(X_i) | X_1, X_2, \ldots, X_\ell)$$

so we obtain immediately that

$$\sup_{A \epsilon S}(\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \epsilon A))$$

$$\leq E(\sup_{A \epsilon S}(\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - \ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i)) | X_1, X_2, \ldots, X_\ell)$$

and

$$\inf_{A \epsilon S}(\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - P(X_1 \epsilon A))$$

$$\geq E(\inf_{A \epsilon S}(\ell^{-1} \sum_{i=1}^{\ell} 1_A(X_i) - \ell^{-1} \sum_{i=\ell+1}^{2\ell} 1_A(X_i)) | X_1, X_2, \ldots, X_\ell) .$$

These two inequalities are in fact stronger than our ascertion that $\pi^\ell \leq E(\rho^\ell | X_1, X_2, \ldots, X_\ell)$.

From this observation we see that any upper law of the iterated logarithm we prove for ρ^ℓ carries over to π^ℓ with the constant unchanged. In general, however, to carry an upper law for π^ℓ to one for ρ^ℓ one has to double the constant.

We can now state our basic estimate on lower layers.

Theorem 3.6. Let X_i, $i = 1, 2, \ldots$ be i.i.d. random variables with absolutely continuous distributions. If S is the class of lower layers in $\mathbb{R}^2$, then there are constants C_1, C_2, and $\alpha > 0$ such that

$$P(\rho^\ell > \lambda) \leq C_1 \exp(\alpha \log \ell \sqrt{\ell} - \lambda^2 \ell) + C_2 \exp(-\sqrt{\ell}) .$$

Proof. We begin by showing that the X_i may be assumed to be uniformly distributed in the cube $[0,1]^d$. This is done by noting that the X_i have absolutely continuous distribution; we can define the conditional distribution functions

$$F_1(x_1), F_2(x_2|x_1), \ldots , F_d(x_d|x_1, x_2, \ldots , x_{d-1})$$

and then define the one-one mapping of $\mathbb{R}^d$ onto $\mathbb{R}^d$ by

$$\begin{aligned} z_1 &= F_1(x_1) \\ z_2 &= F_2(x_2|x_1) \\ &\vdots \\ z_d &= F_d(x_d|x_1, x_2, \ldots , x_d) . \end{aligned}$$

We call the map obtained in this way $\Phi : \mathbb{R}^d \to \mathbb{R}^d$. One can check by direct calculation that the random variables defined by $Z_i = \Phi(X_i)$, $i = 1, 2, \ldots$ are independent and uniformly distributed. So far we have only repeated a trick which is familiar in the study of the classical empirical distribution function. The useful observation is that the class of lower layers is invariant under coordinate-wise monotone mappings. This provides us with the fact that $P(\rho^{\ell}(X_1, X_2, \ldots , X_{\ell}) > \lambda) = P(\rho^{\ell}(Z_1, Z_2, \ldots , Z_{\ell}) > \lambda)$ and completes the proof of the claim that it suffices to consider uniformly distributed random variables which take values in the cube.

Now we begin our estimations by observing that

$$(1) \qquad P(\rho^{\ell} > \lambda) = P(\rho^{\ell} > \lambda \,;\, \Delta^S(X_1, X_2, \ldots , X_{\ell}) \leq 2^{\delta\ell})$$

$$+ P(\Delta^S(X_1, X_2, \ldots , X_{\ell}) \geq 2^{\delta\ell})$$

$$\leq 3\cdot 2^{\delta\ell}\exp(-\lambda^2(\ell-1)) + P(\sum_{j=o}^{K^S(X_1,X_2,\ldots,X_{\ell})} \binom{\ell}{j} \geq 2^{\delta\ell})$$

where δ is any positive real number. This last inequality follows from the Vapnik-Chervonenkis lemma in the first part and the corollary to Theorem 3.1 in the second.

Next since $\sum_{j=o}^{k} \binom{\ell}{j} \leq 1 + \ell^k$, we obtain

$$(2) \quad P(\sum_{j=o}^{K^S(X_1,X_2,\ldots,X_{\ell})} \binom{\ell}{j} \geq 2^{\delta\ell}) \leq P(K^S(X_1,X_2,\ldots,X_{\ell})\log \ell \geq (\delta\ell-1)\log 2).$$

Thus we have a need to compute the tail probability of the random variable $K^S(X_1, X_2, \ldots , X_{\ell})$.

The crucial fact is that one can prove that

$$P(K^S(X_1, X_2, \ldots , X_{\ell}) \geq 2r) \leq \binom{\ell}{r}(r!\binom{2r}{r})^{-1}$$

where r is any positive integer. This inequality is essentially equivalent to a result of Hammersley, but it is as brief to use his argument as to apply his result (this method is used and clearly given in Kingman [14]). We let $X_{(i)}$, $i = 1, 2, \ldots$ be the X_i reordered in such a way that the first coordinates are increasing. Next denote by y_i the second coordinate of $X_{(i)}$, $i = 1, 2, \ldots$. We define a random variable ν as the number of sequences $i_1 < i_2 < \cdots < i_r \leq \ell$ such that $y_{i_1} > y_{i_2} > \cdots > y_{i_r}$ and where r is a fixed integer. By the uniform distribution of the X_i we have

$$E(\nu) = \sum_{i_1 < i_2 < \cdots < i_k} P(y_{i_1} > y_{i_2} > \cdots > y_{i_r})$$

$$= \binom{n}{r}(r!)^{-1} .$$

If $K^S(X_1, X_2, \ldots , X_\ell) \geq 2r$, then $\nu \geq \binom{2r}{r}$ so by Chebyshov's inequality

$$P(K^S(X_1, X_2, \ldots , X_\ell) \geq 2r) \leq \binom{n}{r}\left(r!\binom{2r}{r}\right)^{-1} .$$

We then obtain by Stirling's formula that

$$(3) \qquad P(K^S(X_1, X_2, \ldots , X_\ell) \geq 2r) \leq (4\pi r)^{-1/2}(e\sqrt{\ell}/2r)^{2r} .$$

Finally, the choice of $\delta = 4e \log \ell / \sqrt{\ell}$ and the inequalities (1), (2), and (3) directly provide an inequality which is somewhat stronger than the one claimed in this theorem.

The usual Borel-Cantelli argument provides one immediate consequence of our estimate.

Corollary. If X_i, $i = 1, 2, \ldots$ have density and take values in $\mathbb{R}^2$, then on letting S be the class of lower layers, we have $P(\rho^\ell > \beta(\log \ell)^{1/2}\ell^{-1/4}$ i.o.$) = 0$ for some constant β.

This corollary is, of course, only the crudest means of employing the inequality given by Theorem 3.6. There is a martingale which can be directly related to ρ^{ℓ} and thus make available the more elaborate methods of Theorem 2.5. By this attack I have proved that

$$P(\rho^{\ell} > \beta(\log\log \ell)^{1/2}\ell^{-1/4} \text{ i.o.}) = 0 ,$$

but since the factor $\ell^{-1/4}$ is probably not near the best (recall $\ell^{-1/2}$ is conjectured) we do not pursue the point here.

Since the class of lower layers does not at first seem as natural as some of the classes we have considered, it is of value to note that many results for lower layers have a direct application to the class of convex sets. The source of the connection is easily seen in two dimensions: If a convex set has n extreme points, then there is a subset of $[n/4]$ of these whose graph is the graph of a monotonic function. This translation immediately provides the following result.

Theorem 3.7. If S is the class of convex Borel subsets of $\mathbb{R}^2$, then $\overline{\lim_{\ell \to \infty}} \ell^{-1/2}K^S(X_1, X_2, \ldots , X_{\ell}) = C \geq 0$ for a constant C provided that the X_i, $i = 1, 2, \ldots$ are i.i.d. with densities.

Further, since $P(K^S(X_1, X_2, \ldots , X_{\ell}) \geq r)$ can be estimated by first looking only for monotone subsets, the procedure of Theorem 3.6 can be used to estimate the tail of ρ^S for the class of convex Borel sets in $\mathbb{R}^2$. Here we remark that we do not need that the class of convex sets is invariant under the map Φ of Theorem 3.6. (Indeed, it is not invariant.)

This sketch and the preceding work on lower layer thus provide the following results which we state for the record.

Theorem 3.8. Let X_i, $i = 1, 2, \ldots$ be i.i.d. random variables with absolutely continuous distribution. If S is the class of convex Borel subsets of $\mathbb{R}^2$, then

(1) $P(K^S(X_1, X_2, \ldots, X_\ell) \geq 8r) \leq \binom{\ell}{r}(r!\binom{2r}{r})^{-1}$.

(2) There are constants C_1, C_2, and α such that
$$P(\rho^\ell > \lambda) \leq C_1 \exp(\alpha \log \ell \sqrt{\ell} - \lambda^2 \ell) + C_2 \exp(-\sqrt{\ell}).$$

(3) There is a constant β such that
$$P(\rho^\ell > \beta(\log \ell)^{1/2}\ell^{-1/4} \text{ i.o.}) = 0 \text{ and}$$
$$P(\pi^\ell > \beta(\log \ell)^{1/2}\ell^{-1/4} \text{ i.o.}) = 0.$$

Here is perhaps an appropriate place to remark that this last result is the only result for even so basic a class as that of the convex sets which goes beyond the level of the law of large numbers. The fact that the result is given only in $\mathbb{R}^2$ is a reflection of the cleverness of Hammersly's argument given in the proof of Theorem 3.6. It would be by generalization of that argument that the above result could be extended to $\mathbb{R}^d$ $d > 2$.

CHAPTER IV

ANALYSIS OF TWO ENTROPIES

The more one regards $h(F, S) = \lim_{\ell \to \infty} \ell^{-1} E \log \Delta^S(X_1, X_2, \ldots , X_\ell)$ as an entropy, the more compelling it is to compare $h(S,X_1)$ with the entropy of Kolmogorov. The most direct means of pursuing this comparison is to let $S = \{A, TA, T^2A, \ldots\}$ where A is a measurable set and T is a bimeasurable transformation. Once this class is considered, there is an immediate question: If T is taken to be rotation on the unit circle by an irrational multiple of 2π, do we have that S is a uniformity class for all choices of A? In the first theorem of this chapter we prove in a very strong way that the answer to this question is no. Although the construction given in Theorem 4.1 is quite elementary, the result is so strong as to be surprising.

The vein opened by Theorem 4.1 is then worked until its limits in generality (Theorem 4.5) and in precision (Theorems 4.3 and 4.4) are reached. The result given in Theorem 4.4 is of a very classical type although as best one can tell it is new. It is, in any case, the product of P. Erdős' curiosity in the limits of my Theorem 4.1 and his kind insistence that I find those limits.

Finally, in the last part of this chapter the motivation for considering $S = \{A, TA, T^2A, \ldots\}$ is revisited and Kolmogorov's entropy is compared with that of Vapnik and Chervonenkis.

<u>Theorem 4.1</u>. Let $T(x) = (x + \alpha) \bmod 1$ for $x \in [0,1)$ and α irrational. For any $\delta > 0$ there is a Lebesque measurable set A

with $\lambda(A) < \delta$ and the following property: For any finite set J there is an integer $j = j(J)$ such that $J \subset T^j A$.

Proof. We construct sets A_k which are the union of finitely many intervals such that for any set S of k elements there is an n so that $S \subset T^n A_k$. Also, the A_k have measure less than $\delta 2^{-k}$ so the theorem follows by setting $A = \bigcup_{k=1}^{\infty} A_k$. Since the construction of the A_k for $k > 3$ is analogous to the construction of A_3, we restrict ourselves at first to this case.

Let n_1, n_2, n_3 be integers such that $6 < n_1 < n_2 < n_3$ which will have just one further restriction placed on them. Now we can define the basic set,

$$A_3 = [0,6/n_1] \cup \{ \bigcup_{k=o}^{n_1-1} [k/n_1, k/n_1+6/n_1n_2]\} \cup \{ \bigcup_{k=o}^{n_1n_2-1} [k/n_1n_2, k/n_1n_2+6/n_1n_2n_3]\}.$$

We note that $\lambda(A_3) < 6(1/n_1 + 1/n_2 + 1/n_3)$, so choosing n_1 large guarantees that $\lambda(A_3) < \delta 2^{-3}$, and allows n_1, n_2, n_3 to be fixed.

Let $\{x_1, x_2, x_3\}$ denote a three element subset of $[0,1)$. We choose a real number τ so that $(x_1 + \tau) \bmod 1 = 3/n_1$. We have now three more choices:

(1) Choose τ_1 such that $(x_2 + \tau + \tau_1) \bmod 1 = k/n_1 + 3/n_1n_2$ and with $|\tau_1| \leq 1/n_1$.

(2) Choose τ_2 such that $(x_3 + \tau + \tau_1 + \tau_2) \bmod 1 = k^1/n_1n_2 + 3/n_1n_2n_3$ and with $|\tau_2| \leq 1/n_1n_2$.

(3) Choose τ_3 such that $(\tau + \tau_1 + \tau_2 + \tau_3) \bmod 1 = -n\alpha$ and with $|\tau_3| \leq 1/n_1n_2n_3$.

Finally, the fact that $\{x_1, x_2, x_3\} \subset T^n A$ follows from the inequalities $|\tau_1 + \tau_2 + \tau_3| < 3/n_1$, $|\tau_2 + \tau_3| < 3/n_1 n_2$, and $|T_3| < 3/n_1 n_2 n_3$ together with the definition of A, τ, and τ_i.

For the general case we have the representation

$$A_k = [0, 2k/n_1] \cup \{ \bigcup_{j=o}^{n_1-1} [j/n_1, j/n_1 + 2k/n_1 n_2\}$$

$$\cdots \cup \{ \bigcup_{j=o}^{(n_1 n_2 \cdots n_{k-1})-1} [j/n_1 \cdots n_{k-1}, j/n_1 \cdots n_{k-1} + 2k/n_1 \cdots n_k]$$

where the restriction on the n_i are that

$$2k < n_1 < \cdots < n_k \quad \text{and} \quad 2k(1/n_1 + \cdots + 1/n_k) < \delta 2^{-k} \text{ .}$$

The verification that A_k has the desired properties proceeds just as it did with A_3.

The higher dimensional analog of Theorem 4.1 is valid and so is a converse. This is made explicit in the following:

Corollary. Let T be a translation on the n-torus $\mathcal{T}^n$. T is a periodic if and only if for each $\delta > 0$ there is a measurable set A such that $m(A) < \delta$ and with the property that for any finite set $J \subset \mathcal{T}^n$ there is a $j = j(J)$ such that $J \subset T^j A$.

We note that Theorem 4.1 is best possible in the sense that we certainly can choose a countable set J such that J is not contained in any of the sets $T^j A$, no matter what A provided $m(A) < 1$. To give such a J one just chooses $x_i \in (T^i A)^c$ and sets $J = \{x_1, x_2, \ldots\}$. It is nevertheless possible to give a strengthened version of Theorem 4.1 which is potentially useful.

Theorem 4.2. Given any $\epsilon > 0$ there is a Lebesque measurable set A with $m(A) < \epsilon$ such that we have the following:

For any α irrational and any F with only a finite number of limit points, then $F \subset (A + n\alpha) \bmod 1$ for infinitely many integers n.

Proof. We first observe that given $\epsilon > 0$ there exist a set B_k which is a finite union of intervals, $m(B_k) < \epsilon$, and there exist a constant δ_k so that we have the following:

For any k set $F = \{f_1, f_2, \dots, f_k\}$ and any α irrational there exist an n such that $(f_i - \delta_k, f_i + \delta_k) \subset (B_k + n\alpha) \bmod 1$ for $i = 1, 2, \dots, k$. To prove this observation we note that as in Theorem 4.1 we can construct a set B_k' so that $\{f_1, \dots, f_k\}$ is contained in $(B_k' + n\alpha) \bmod 1$. Let δ_k be the length of the length of the smallest interval in the construction of B_k'. Now define B_k as the set obtained by expanding each interval in the construction of B_k' by a factor of 3 about the midpoint of that interval. This B_k then has the property asserted by the first observation.

The first observation will help us catch the limit points; now we observe how to catch the remaining point once the limits are caught. More explicitly, we observe that given an $\epsilon > 0$ and any $\delta > 0$, there is a set $C_{\delta,j}$, a union of intervals, such that $m(C_{\delta,j}) < \epsilon$ and with the following property:

For any R and any j set $P = \{p_1, p_2, \dots, p_j\}$ there is an interval $(r_1, r_2) \subset (0, \delta)$ such that for any $r \in (r_1, r_2)$ we have $P \subset (C_{\delta,j} + r + R) \bmod 1$. The proof of this second observation is already contained in the proof of Theorem 4.1.

Now to conclude the proof of Theorem 4.2 we first choose sets B_k and intervals δ_k with $m(B_k) \leq \epsilon_1 2^{-k}$ and such that we have the characteristic property of our first observation. Now by the second observation we choose sets $C(\delta_k/2, j)$ such that $m(C(\delta_k/2, j)) \leq \epsilon_2 2^{-k-j}$ and finally let $A = \bigcup_{i,j} B_k \cup C(\delta_k/2, j)$. We have $m(A) < \epsilon_1 + \epsilon_2$ so $m(A) < \epsilon$ if $\epsilon_1 < \epsilon/2$, $\epsilon_2 < \epsilon/2$. By the observations, $B_k(f_i - \delta_k, f_i - \delta_k) \subset (B_k + n\alpha)\text{mod } 1$ for some n, and all $i = 1, 2, \ldots , k$ and where $\{f_1, f_2, \ldots , f_k\}$ is the set of limit points of F. Let $P = \{p_1, p_2, \ldots , p_j\}$ be the points of F not contained in any of the sets $(f_i - \delta_k/2, f_i + \delta_k/2)$, $i = 1, 2, \ldots , k$. Now by the second observation $P \subset (C(j, \delta_k/2) + n\alpha + r)\text{mod } 1$ for all $r \in (r_1, r_2) \subset (0, \delta_k/2)$, and hence also $F \subset P \cup (\bigcup_{i=1}^{k} (f_i - \delta_k/2, f_i - \delta_k/2)) \subset (B_k \cup C(j, \delta_k/2) + n\alpha + r)\text{mod } 1$ since $r \leq \delta_k/2$. Choosing $r \in (r_1, r_2)$ such that $r = m\alpha \text{ mod } 1$, and noting there are infinitely many such m, we have $F \subset (B_k \cup C(j, \delta_k/2) + (n+m)\alpha)\text{mod } 1 \subset (A + (n+m)\alpha)\text{mod } 1$ as required. Theorems 4.1 and 4.2 also have their analog on the line $\mathbb{R}$, which we state without proof since the same method is involved as in Theorem 4.1.

Theorem 4.3. Given any $\epsilon > 0$ there is a measurable set A such that $m(A) < \epsilon$ which has the following property:

Given any dense set D of $\mathbb{R}$ and any set F with only a finite number of limit points, then there is a $d \in D$ such that $F \subset A + d$.

We remark that the set D could, of course, be $\mathbb{R}$, and in fact an interesting situation arises with this choice. It is no longer clear that one could not by some other method construct a measurable A with $m(A) < \epsilon$ and such that for any countable set F there is an $r \in \mathbb{R}$ such that $F \subset A + r$. This presents a problem: Given a measurable set $A \subset [0,1]$ construct a countable set F such that $F \not\subset (A+r) \bmod 1$ for all r.

This problem is not solved. The result we provide in this direction is given in the corollary to Theorem 4. We are indebted to P. Erdős for his suggestion that a construction like the one in Theorem 4.4 should be tried.

Theorem 4.4. Let $\alpha_i > 0$ such that $\sum_{i=1}^{\infty} \alpha_i = \alpha < 1$ be given. Then there is a closed subset F of $[0,1]$ such that $m(F) = 0$ and $F \not\subset \bigcup_{i=1}^{\infty} I_i$ for any set of open intervals I_i with $mI_i = \alpha_i$.

Proof. We will construct sets F_k, $k = 1, 2, \ldots$ such that each F_k is closed and $F_k \supset F_{k+1}$. We will require that F_k satisfy the property that $F_k \not\subset \bigcup_{i=1}^{\infty} I_i$ for any set of open interval with $mI_i = \alpha_i$. By making $m(F_k)$ converge to zero, we have by taking $F = \bigcap_{k=1}^{\infty} F_k$ that $m(F) = 0$. We now consider any union $\bigcup_{i=1}^{\infty} I_i = \mathcal{O}$ where $mI_i = \alpha_i$. Supposing the F_k have been constructed, we have $G_k = F_k \cap \mathcal{O}^c \neq \emptyset$ and the G_k are nested and compact, so $F \supset \bigcap_{k=1}^{\infty} G_k \neq 0$, so $F \not\subset \bigcup_{i=1}^{\infty} I_i$. It remains to construct the F_k.

The F_k will be defined inductively and their definition will require an increasing sequence of integers $n_1, n_2, n_3, \ldots$. We define F_1 by $\bigcup_{j=o}^{n_1-1} [j/n_1, j/n_1 + 1/2n_1]$. Then F_2 is constructed by dividing each of the intervals of F_1 into $2n_2$ equal sections of length $1/2^2 n_1 n_2$ and then taking F_2 to be the union formed by taking every other one of these sections. In general, $F_{\ell+1}$ is formed from F_ℓ by dividing each interval of F_ℓ into $2n_{\ell+1}$ equal prices and then taking $F_{\ell+1}$ to be the union of every other one of these.

We list now the important properties of the F_ℓ:

(1) F_ℓ consists of $n_1 n_2 \cdots n_\ell$ intervals of length $(2^\ell n_1 n_2 \cdots n_\ell)^{-1}$.

(2) If $g_\ell(X)$ denotes the maximum number of sections of F_ℓ which can be covered by an interval of length X, then we have

$$g_\ell(X) \leq X n_1 n_2 \cdots n_\ell + 1 .$$

(3) $F_\ell > F_{\ell+1}$.

These properties are immediately verified and all that we are required to show is that $n_1, n_2, \ldots , n_\ell$ can be chosen so that $F_\ell \not\subset \bigcup_{i=1}^{\infty} I_i$ for any I_i such that $mI_i = \alpha_i$. The proof depends on the observations that small intervals and large intervals are used with different effectiveness to cover F_ℓ. To make this explicit let $N(\ell)$ be the maximum number of intervals of F_ℓ which can be covered by I_i with $m(I_i) = \alpha_i$. We have

$$N(\ell) < \Big(\sum_{\alpha_i \leq (2^\ell n_1 n_2 \cdots n_\ell)^{-1/2}} \alpha_i \Big) \cdot 2^\ell n_1 n_2 \cdots n_\ell$$

$$+ \sum_{\alpha_i > (2^\ell n_1 n_2 \cdots n_\ell)^{-1/2}} (\alpha_i n_1 n_2 \cdots n_\ell + 1) .$$

This estimate follows from the consideration that the most efficient use of an interval of length α_i less than $(2^{\ell} n_1 n_2 \cdots n_2)^{-1/2}$ would be made by having all such intervals to be of length $(2^{\ell} n_1 \cdots n_{\ell})^{-1}$ and using each one to cover an interval of F_{ℓ}. The second term in the estimate is of course from property (2) above.

In the case $\ell = 1$, the estimate reduces to

$$N(1)/n_1 < 2 \sum_{\alpha_i < (2n_1)^{-1/2}} \alpha_i + \sum_{\alpha_i > (2^{\ell} n_1)^{-1/2}} (\alpha_i + 1/n_1)$$

$$\leq 2 \sum_{\alpha_i < (2n_1)^{-1/2}} + (1 + (n_1)^{-1/2})\alpha$$

where we used the fact that $\sum_{i=1}^{\infty} \alpha_i = \alpha$ and that $1/n_1 < n_1^{-1/2}\alpha_i$ for all α_i in the second sum. Now it is evident that n_1 can be chosen so large that $N(1)/n_1 < 1$. Hence, n_1 can be chosen so that F_1 is not covered by intervals I_i with $m(I_i) = \alpha_i$.

The general case is not so different. Our estimate yields

$$N(\ell)/n_1 n_2 \cdots n_{\ell} < 2^{\ell} \sum_{\alpha_i < (2^{\ell} n_1 n_2 \cdots n_{\ell})^{-1/2}} \alpha_i + \alpha(1+(n_1 n_2 \cdots n_{\ell})^{-1/2} .$$

Hence, we can choose n_{ℓ} so that $N(\ell)/n_1 \cdots n_{\ell} < 1$, hence constructing F_{ℓ} which cannot be covered.

<u>Corollary</u>. Given any measurable set A with $m(A) < 1$, then there is a closed set F of measure zero such that $F \not\subset (A+r) \bmod 1$ for any $r \in \mathbb{R}$.

Proof. Since $A \subset \bigcup_{i=1}^{\infty} I_i$ with $m(\bigcup_{i=1}^{\infty} I_i) < 1$ and I_i disjoint, it suffices to construct F so that $F \not\subset (\bigcup_{i=1}^{\infty} I_i + r) \bmod 1$. A much stronger statement has been proved in Theorem 4.4, since $m(I_i) = \alpha_i$ with $\sum_{i=1}^{\infty} \alpha_i < 1$ and $(\bigcup_{i=1}^{\infty} I_i + r) \bmod 1$ is a union of intervals I_i' with $m(I_i') = \alpha_i$.

Since the conclusions of the preceding theorems concern specific countable sets, our results on rotations and translations cannot possibly carry over to a measure preserving transformation which may not be defined except up to null sets. It is possible, however, to introduce a notion of "almost every finite subset of Ω" which allows a general complement to the preceding work. To define this notion we suppose that X_i, $i = 1, 2, 3, \ldots$ is a sequence of measurable functions from a probability space $(\bar{\Omega}, \bar{\sum}, P)$ to a probability space $(\Omega, \sum, \mu)$, such that $P(X_i^{-1}(A)) = \mu(A)$ for $A \in \sum$ and such that the functions X_i, $i = 1, 2, \ldots$ are independent. Now for $\bar{\omega} \in \bar{\Omega}$ we have that $\{X_1(\bar{\omega}), X_2(\bar{\omega}), \ldots, X_k(\bar{\omega})\}$ is a k element subset of Ω which is intuitively a random sample from Ω. We say that a property holds for <u>almost every finite sample of Ω</u> if the property holds for $\{X_1(\bar{\omega}), X_2(\bar{\omega}), \ldots, X_k(\bar{\omega})\}$ for almost every $\bar{\omega}$ and every $k \geq 1$.

Before proving the next result, it is necessary to recall some definitions and results from the theory of measure preserving transformations.

A bi-measurable, measure preserving transformation T on $(\Omega, \sum, \mu)$ is <u>aperiodic</u> if for any $B \in \sum$ with $m(B) \neq 0$ there is a $B' \subset B$ such that $P(B' \Delta T^{-n} B') \neq 0$ for some n.

We note that any ergodic transformation is aperiodic and thus irrational rotations are aperiodic. For an example of an aperiodic transformation which is not ergodic, take the transformation $T(x,y) = ((x + \alpha) \bmod 1, y)$ on the square where α is irrational.

We also recall that a measure space $(\Omega, \sum, \mu)$ is called a Lebesque space if it is measure theoretically isomorphic to $[0,1]$. An enormous number of the measure spaces in probability theory are Lebesque spaces.

We require one lemma from the general theory.

<u>Lemma</u>. Let T be an aperiodic transformation on a Lebesque space $(\Omega, \sum, \mu)$. Given $\epsilon > 0$ and an integer $n \geq 1$, there is an $E \in \sum$ such that $E, TE, T^2E, \ldots, T^nE$ are disjoint and $\mu(\bigcup_{i=o}^{n} T^iE) > 1 - \epsilon$.

This lemma is often called Rohlin's theorem. As stated, it is due to Jones-Kringle [12]. A proof of a closely related result (for antiperiodic transformations) is given in Halmos [11, p. 71].

<u>Theorem 4.5</u>. Let T be a measure preserving transformation on a Lebesque probability space $(\Omega, \sum, \mu)$. T is aperiodic if and only if for each $\delta > 0$ there is an $A \in \sum$ with $\mu(A) < \delta$ and the following property:

For almost every finite sample J of Ω there is an integer $j = j(J)$ such that $J \subset T^{-j}A$.

Proof. Suppose first that T is aperiodic. We will construct sets A_k and B_k in $\sum$ such that for each $\bar{\omega} \in \bar{\Omega}$ with $X_i(\bar{\omega}) \in B_k$ for $i = 1, 2, \dots, k$ there is a j such that $T^j X_i(\bar{\omega}) \in A_k$ for $i = 1, 2, \dots, k$. Further, we will have $\mu(A_k) \leq \delta 2^{-k}$ and $\mu(B_k) \geq 1 - 2^{-k}$, so by setting $A = \bigcup_{k=1}^{\infty} A_k$, the first half of the theorem is proved.

The sets A_k are defined via a sequence of zeros and ones which are used to label the levels of a Rohlin tower. To define this sequence, we will use $k + 1$ integers denoted by $\ell_2, \ell_3, \dots, \ell_k$, L and N whose properties will be specified later. If S is a finite string of zeros and ones, then $|S|$ will denote the length of that string. Also, if S_1 and S_2 are two strings, then $(S_1)(S_2)$ denotes the string of length $|S_1| + |S_2|$ obtained by following S_1 with S_2. Similarly, $(S)^{\ell}$ denotes the string of length $\ell|S|$ given by S repeated ℓ times. Now we define nested strings as follows:

$$b_2 = (11)(01)^{\ell_2}$$

$$b_3 = (kN \cdot |b_2| \text{ ones})[(N \cdot |b_2| \text{ zeros})(b_2)]^{\ell_s}$$

and in general

$$b_t = (kN \cdot |b_{t-1}| \text{ ones})[(N \cdot |b_{t-1}| \text{ zeros})(b_{t-1})]^{\ell_t} .$$

This allows the definition $\text{label}(A_k) = (b_k)^L$.

By Rohlin's theorem we select a set B such that $T^{-1}B, T^{-2}B, \dots, T^{-n}B$ are disjoint with $n = |\text{label}(A_3)|$ and $\mu(\bigcup_{i=1}^{n} T^{-i}B)^c < \delta_1$. Now A_k is defined by $A_k = \bigcup_{i \in S} T^i B$ where S is the set of integers such that the i^{th} element of the string label(A_k) is one.

The construction of A_k is complete once the selection of the constants appearing in its definition are made. We note first that $\mu(A_k)$ is less than the percentage of ones appearing in b_k, so $\mu(A_k) < 1/N + k/\ell_k$. To begin, we choose N such that $1/N < 1/2(\delta 2^{-k})$, and we then construct $\ell_2, \ell_3, \ldots, \ell_k$ sequentially in such a way as to assure that most samples are well separated with respect to the block structure of label (A_k). To make this explicit, let $\ell(x) = i$ if $x \in T^{-i}B$ and let $\ell(x)$ be infinity if $x \notin A_k$. Consider then the following conditions:

$$(1)\quad |\ell(X_i(\bar{\omega})) - \ell(X_j(\bar{\omega}))| \bmod (2\ell_2) > 1 \quad \text{for} \quad i \neq j, \quad 1 \leq i,j \leq k$$

$$(2)\quad |\ell(X_i(\bar{\omega})) - \ell(X_j(\bar{\omega}))| \bmod |b_t| > kN|b_{t-1}|$$

for $i \neq j$, $1 \leq i,j \leq k$, and $3 \leq t \leq k$.

We first choose ℓ_2 so large that (1) holds for all $\bar{\omega}$ except a set of measure $\delta_2 + \delta_1$. The choice of ℓ_2 determines $|b_2|$ so then ℓ_3 is chosen so large that (2) holds for $t = 3$ for all $\bar{\omega}$ except a set of measure $\delta_3 + \delta_1$. Continuing this procedure, we obtain $\ell_2, \ell_3, \ldots, \ell_k$ so that (1) and (2) hold for all $\bar{\omega}$ except a set of measure $\delta \leq k\delta_1 + \delta_2 + \cdots + \delta_k$. Since ℓ_k can be chosen arbitrarily large, we choose ℓ_k so that $k/\ell_k \leq 1/2(\delta 2^{-k})$ and hence guarantee that $\mu(A_k) \leq \delta 2^{-k}$. The last choice we make is of L, and we choose L so large that

$$(3)\quad \ell(X_i(\omega)) < |\text{label}(A_k)| - K|b_k|$$

for all $\bar{\omega}$ except a set of measure $\delta_1 + \delta_2$. Now we specify that $(K+1)\delta_1 + \delta_2 + \cdots + \delta_k + \delta_L < 2^{-k}$, and it remains only to show that for each $\bar{\omega}$ such that conditions (1), (2), and (3) are fulfilled we have a j such that $T^j X_i(\bar{\omega}) \in A_k$ for $i = 1, 2, \ldots, k$.

Let $X_1'(\bar{\omega}), \ldots, X_k'(\bar{\omega})$ denote the sample $X_1(\bar{\omega}), \ldots, X_k(\bar{\omega})$ resubscripted in such a way that $\ell(X_1'(\bar{\omega})) < \ell(X_2'(\bar{\omega})) < \cdots < \ell(X_k'(\bar{\omega}))$. First choose $j(1)$ so that $T^{j(1)}X_1'(\bar{\omega})$ is contained in the first level of a b_k block; this is possible by condition (3). Next choose $j(2)$ so that $T^{j(2)}(T^{j(1)}X_2'(\bar{\omega}))$ is contained in the first level of a b_{k-1} block and such that $j(2) < N|b_{k-1}|$; this is possible by conditions (2) and (3). Continue inductively for $3 \leq t \leq k-1$ choosing $j(t)$ so that $T^{j(t)}(T^{j(1)+j(2)+\cdots+j(t-1)}X_t'(\bar{\omega}))$ is in the first level of a b_{k-t+1} block and $j(t) < N|B_{k-t+1}|$. Finally, choose $j(k)$ so that

$$T^{j(k)}(T^{j(k-1)+\cdots+j(1)}X_k'(\omega)) \text{ is in } A_k \text{ and } j(k) \leq 2 ;$$

This is possible by condition (1).

Now set $j = j(1) + \cdots + j(k)$. We already have that $T^j(X_k'(\bar{\omega})) \in A_k$, so suppose $1 \leq t \leq k-1$ and consider $T^j(X_t'(\bar{\omega}))$. By the definition of $j(t)$ we have $T^{j(1)+\cdots+j(t)}X_t'(\omega)$ is in the first level of a b_{k-t+1} block. Such a block begins with $kN|b_{k-t}|$ levels labeled with a 1, and by the bounds on the $j(S)$ we have $j(t+1) + j(t+2) + \cdots + j(k) < kN|b_{k-t}|$. This proves that $T^{j(1)+\cdots+j(k)}(X_t(\bar{\omega}))$ is in A_k, and completes the first half of the theorem. The converse is much easier, and, as it is not involved with our basic construction, is pointed out as a consequence of the following lemma.

<u>Lemma</u>. If for any $\delta > 0$ there is an $A \in \sum$ such that $\mu(\bigcup_{i=1}^{\infty} T^{-i}A) = 1$, then T is aperiodic.

<u>Proof</u>. Suppose T is not aperiodic and let $B \in \sum$ denote an invariant set with $0 < \mu(B) \leq 1$ for which there is an n such that $\mu(B' \Delta T^{-n}B') = 0$ for all $B' \subset B$, $B' \in \sum$. If A is chosen as in the hypothesis, then $\mu(B \Delta \bigcup_{i=o}^{\infty} T^{-i}(A \cap B)) = 0$ since B is invariant. But $\mu(A \cap B \Delta T^{-n}(A \cap B)) = 0$ so $\mu(\bigcup_{i=o}^{\infty} T^{-i}(A \cap B)) \leq n\mu(A \cap B)$, and consequentially $\mu(B) \leq n\mu(A \cap B) \leq n\mu(A)$. Since $\mu(A)$ can be chosen arbitrarily small this gives a contradiction.

We are now able to show quite simply that the Kolmogorov entropy can differ in almost every qualitative combination from the Vapnik-Chervonenkis entropy.

We first consider the case of a transformation with zero Kolmogorov entropy. The simplest such ergodic transformation is irrational rotation. We note that by Theorem 4.1 there is measurable set A such that if X_i, $i = 1, 2, \ldots$ are i.i.d. with <u>any</u> distribution that the Vapnik-Chervonenkis entropy $h(F, S) \neq 0$, where $S = \{A, TA, T^2A, \ldots\}$. Moreover, we see that if T is any ergodic (hence aperiodic) transformation, it is still possible to have $h(F, S) \neq 0$ by taking X_i to be uniformly distributed in the sense described in the preliminaries to Theorem 4.5.

In the other direction, we provide an example where the Vapnik-Chervonenkis entropy is zero, and so is the Kolmogorov entropy. Again the transformation is taken to be irrational rotation on the circle, but now A is chosen to be an interval. Setting $S\{A, TA, T^2A, \ldots\}$ it is easy to see that $\Delta^S(X_1, X_2, \ldots , X_\ell) \leq \ell + 1$ and consequentially $h(F, S) = \lim_{\ell \to \infty} \ell^{-1} E \log \Delta^S(X_1, X_2, \ldots , X_\ell) = 0$.

Only one qualitative comparison remains between the two entropies: Is there an ergodic transformation T and a measurable set A such that

(1) the Kolmogorov entropy of the transformation T with the partition $P = (A,A^c)$ is positive

and

(2) the Vapnik-Chervonenkis entropy of the class $S = \{A,TA,T^2A,\ldots\}$ with some i.i.d. random variables X_i, $i = 1, 2, \ldots$ is equal to zero.

If the random variables X_i, $i = 1, 2, \ldots$ are taken to assume only countably many distinct values, then the Vapnik-Chervonenkis entropy is seen to be zero no matter what the class S is taken to be. This provides an answer to the question asked and completes the comparison, but it does not strike the heart of the matter.

In fact, we close with the conjecture: If $T : [0,1] \to [0,1]$ is ergodic, and $P = (A,A^c)$ is a partition of positive Kolmogorov entropy under T, then the class $S = (A, TA, T^2A, \ldots)$ has positive Vapnik-Chervonenkis entropy when X_i, $i = 1, 2, \ldots$ are uniformly distributed on $[0,1]$.

BIBLIOGRAPHY

[1] S. Ahmad, "Sur le theorem de Glivenko-Cantelli," C. R. Acad. Sci. Paris, 252 (1961), 1413-1414.

[2] J. R. Blum, "On convergence of empirical distribution functions," Ann. Math. Stat., 26 (1955), 527-529.

[3] H. D. Brunk, G. M. Ewing, and W. R. Utz, "Some Helley theorems for monotone functions," Proc. Amer. Math. Soc., 7 (1956), 776-783.

[4] K. L. Chung, "An estimate concerning the Kolmogoroff limit distributions," AMS Transactions, 67 (1949), 36-50.

[5] E. Csaki, "An iterated logarithm law for semimartingales and its application to empirical distribution functions," Studia Sc. Math. Hung., 3 (1968), 287-292.

[6] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with application to pattern recognition," IEEE Trans. on Elec. Comp., EC-14 (1965), 326-334.

[7] J. DeHardt, "Generalizations of the Glivenko-Cantelli theorem," Ann. Math. Stat., 42 (1971), 2050-2055.

[8] ________, "A necessary condition for Glivenko-Cantelli convergence in E_n," Ann. Math. Stat., 41 (1970), 2177-2178.

[9] H. G. Eggleston, Convexity, Cambridge University Press, Cambridge, 1958.

[10] R. Fortet and E. Mourier, "Convergence de la repartition empirique vers la repartition theorique," Ann. Sci. Ecole Norm. Sup. III, Ser. 60 (1953), 267-285.

[11] P. R. Halmos, Lectures on Ergodic Theory, Chelsea Publishing Co., New York, 1956.

[12] L. K. Jones and U. Krengel, "On transformations without invariant measure," Adv. in Math., 12 (1974), 275-295.

[13] J. Kiefer, "On the large deviation of the empiric d.f. of vector chance variables and a law of the iterated logarithm," Pacific J. of Mathematics, 11 (1961), 649-660.

[14] J. F. C. Kingman, "Subadditive ergodic theory," Ann. of Prob., 1 (1973), 883-909.

[15] R. R. Rao, "Relations between weak and uniform convergence of measures with applications," Ann. Math. Stat., 33 (1962), 659-680.

[16] T. Robertson and F. T. Wright, "On the maximum likelihood estimation of stochastically ordered random variates, Ann. Stat., 2 (1974), 528-534.

[17] N. Saver, "On the density of families of sets," J. Comb. Theory (A), 13 (1972), 145-147.

[18] L. Schläfli, Gesammelte Mathematische Abhandlungen I, Verlag Birkhäuser, Born, 1950, 209-212.

[19] W. Schlee, "Zwei Glivenko-Cantelli-Theoreme für endliche dimension," Z. für Wahr, 28 (1973), 1-4.

[20] F. Topsøe, "On the Glivenko-Cantelli theorem," Z. für Wahr, 14 (1970), 239-250.

[21] A. G. Tucker, "A generalization of the Glivenko-Cantelli theorem," Ann. Math. Stat., 30 (1959), 828-830.

[22] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," Theory Prob. and Appl., 16 (1971), 264-280.

[23] _______, Teoriya Raspoznavaniya Obrazov, Navka, Moscow, 1974.

[24] J. Wolfowitz, "Generalization of the theorem of Glivenko-Cantelli," Ann. Math. Stat., 25 (1954), 131-138.

[25] _______, Consequence of the Empiric Distribution Function on Half-Spaces, Contributions to Probability and Statistics, Stanford University Press, Stanford, 1960.