AN EXACT CALCULATION OF

THE CAPACITY OF HYPERBOXES

James R. Bitner

Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712

TR-194 February 1982

ABSTRACT:

The capacity of an rth degree polynomial decision function has been defined by Cover [1]. We extend the definition to hyperboxes (the region $\{(x_1,\ldots,x_D) | a_i \leq x_i \leq b_i\}$), show it is a meaningful quantity, and calculate the capacity as 4D/ln 2D where D is the number of dimensions of the feature space.

Section 1. Introduction

A common problem in pattern recognition is the following: we have two classes of objects and are given a number of points (of known class) each corresponding to an object in a D-dimensional "feature space". We are then asked to classify new points whose class is unknown. One method of solving such a problem is to use a decision function. If possible, we can define a geometric figure (such as a hyperplane) such that all the points of a given class are on the same "side" of the figure. (We say the figure achieves a dichotomy.) New points can then be classified by determining on which side of the figure they fall. This method is not always feasible; for some distributions of points there might be no figure of the given type which separates the two classes.

The capacity of the class of rth-degree polynomial decision functions has been defined by Cover ([1], [2], [9], see also [5]) as follows: Given N points in a D-dimensional space, calculate the fraction of the 2^N dichotomies (one for each assignment of classes to the points) that are achievable using polynomials of a given degree (this fraction can also be viewed as the dichotomization probability; i.e., the probability that a random dichotomy is achievable). As N increases and D is held fixed, this fraction is first close to one, then drops sharply to zero. The point where this drop occurs is defined as the capacity; it measures the number of points a function can reasonably handle in a given number of dimensions. For rth-degree polynomials, the capacity was found to be $2\binom{D+r}{D}$. (Note that this is twice the number of adjustable coefficients.) In particular, the hyperplane has capacity 2D+2.

Cover's work raises several questions: Is capacity a meaningful quantity for other classes of decision functions? If so, what is their capacity and how does it compare with that of the class of polynomial functions? We discuss these questions for the class of hyperboxes (or interval complexes [3], [4]) which is the region in D-space defined by $\{x=(x_1,\ldots,x_D)|a_i \leq x_i \leq b_i\}$. (Note that the sides of the box must be parallel to the axes.) The hyperbox is an interesting figure to study because it is simple and natural and because it is an integral part of the variable valued logic system of Michalski [3], [4], [8].

Empirical results were presented in [12] where a program was written to calculate the dichotomization probability for small values of N and D. The results indicated that the probability did drop sharply, and a capacity of 1.4D+7 was conjectured.

In this paper we prove that the dichotomization probability for hyperboxes does, in fact, exhibit the same behavior as it does for hyperplanes. For a fixed D, it is initially one, then when N reaches the capacity, it drops sharply to zero. Hence the capacity is a meaningful quantity for hyperboxes. We show the capacity to be asymptotically 4D/ln 2D (where "ln" is the natural log).

This paper has the following organization: In Section 2 we discuss how to decide when a dichotomy for a given set of points is achievable. We also derive a formula for the dichotomization probability and discuss informally how the capacity will be calculated. Sections 3 and 4 prove some technical lemmas, and Section 5 proves the main result. Section 6 is the conclusion and discusses the significance of the result.

Section 2. An Equation for the Dichotomization Probability

In this section, we discuss how to determine if a given dichotomy of a set of points is achievable. This is followed by some definitions and an equation for the dichotomization probability. We conclude with an informal discussion of how we will calculate the capacity.

We begin by defining a "random" dichotomy of N points in D dimensions. By "random", we mean the following: the classes of the points are independently chosen and each point has equal probability of being in either class. The coordinates of the points are independently chosen from an arbitrary continuous probability density, and the value chosen for one coordinate of a point is independent of its other coordinates.

It is easy to determine if a given dichotomy of a given arrangement is achievable: we index through the dimensions. In a given dimension, a class A (class B) point is said to be an outer A (outer B) if it does not occur between the leftmost B (A) and rightmost B (A) in that dimension. A class-A (class-B) point is said to be active if, in each dimension we have examined, it occurred between the leftmost B (A) and the rightmost B (A) in that dimension. Once we find that a point is an outer A or outer B in a given dimension, it becomes inactive. Since an outer A (or outer B) can clearly be excluded from a hyperbox containing all B's (A's), the dichotomy is achievable iff, after considering all the dimensions, all the A's or all the B's are inactive.

From this procedure it can be seen that only the order of the points (from left to right) in each dimension determines if the dichotomy can be achieved. It is independent of the exact coordinates of the points, as long as their relative order remains the same. Therefore, a distribution of N points can be described by D sequences of length N, with the ith sequence listing the points from left to right in dimension i. Since we assume the N points are independently chosen, each of the N! orderings (sequences) is equally likely. This holds for any probability density. Further, sequences for different dimensions are independent, and since we desire the probability of achieving a random dichotomy, each point is equally likely to belong to a class A or class B.

Notation: N will always denote the number of points in the arrangement, and D will denote the number of dimensions. Nearly all the events and random variables we will define will depend on N and D, but we will omit this to simplify notation.

Definition: The event CANBOX occurs if a random dichotomy of N points in D dimensions is achievable.

Definition: Given an arrangement, let A, OA and OB be, respectively, the number of class A points, outer A's and outer B's. (If a point is "outer" in several dimensions each occurrence is counted.) Let the events BOXAS (and BOXBS) occur if there is a hyperbox containing all class A (class B) points but no class B (class A) points.

Since CANBOX occurs iff BOXAS or BOXBS occurs, we have the following equation.

$$\text{Prob(CANBOX)} = \sum_{n,a,b} \text{Prob}(A=n,\ OA=a,\ OB=b) * \text{Prob}(\text{BOXAS } \underline{\text{or}} \text{ BOXBS} \mid A=n,\ OA=a,\ OB=b)$$

$$= \sum_{n,a,b} \text{Prob}(A=n,\ OA=a,\ OB=b) * [\text{Prob}(\text{BOXAS} \mid A=n,\ OA=a,\ OB=b) + \text{Prob}(\text{BOXBS} \mid A=n,\ OA=a,\ OB=b) - \text{Prob}(\text{BOXAS } \underline{\text{and}} \text{ BOXBS} \mid A=n,\ OA=a,\ OB=b)] \quad (2.1)$$

Now, note that for fixed values of A, OA, and OB, BOXAS and BOXBS are independent events since BOXAS depends on the way the fixed number of class A points are distributed among the fixed number of outer A's and BOXBS depends on the distribution of the fixed number of class B points among the fixed number of outer B's. Clearly, the way in which the class A points are distributed does not affect the manner in which the class B points are distributed. (Note that if A, OA and OB are not fixed, the events are not independent; if BOXAS occurs, it is likely we have a high number of outer A's and, therefore a low number of outer B's, making BOXBS less likely.) Therefore, we have

$$\text{Prob(CANBOX)} = \sum_{n,a,b} \text{Prob}(A = n,\ OA = a,\ OB = b) * [\text{Prob}(\text{BOXAS} \mid A = n,\ OA = a,\ OB = b) + \text{Prob}(\text{BOXBS} \mid A = n,\ OA = a,\ OB = b) - \text{Prob}(\text{BOXAS} \mid A = n,\ OA = a,\ OB = b) * \text{Prob}(\text{BOXBS} \mid A = n,\ OA = a,\ OB = b)] \quad (2.2)$$

We now define an event which will approximate BOXAS and BOXBS but will be more tractable.

Definition: Given n distinct objects, if c choices are made with replacement and each object is chosen at least once, the event <u>USEALL(n,c)</u> occurs.

Lemma 2.1:

Prob(BOXAS | A = n, OA = a, OB = b) $\approx$ Prob(USEALL(N-n, b))

and

Prob(BOXBS | A = n, OA = a, OB = b) $\approx$ Prob(USEALL(n,a))

Proof: We prove the first equation; the second is similar. We are given N-n class B points and b outer B's. We make the simplifying assumption that the b outer B's are chosen with replacement. This allows the possibility of the same point occurring twice in the same dimension. However since there are a very large number of class B points (about N/2) and a very small number of outside B's in a given dimension (on the average 2, proven later) this assumption has only a small effect on the probability. The result then follows because BOXAS occurs iff all N-n class B points are used in b choices. []

From Lemma 2.1 and (2.2) we have

Prob(CANBOX) $\approx$

$$\sum_{n,a,b} \text{Prob}(A = n, OA = a, OB = b) * [\text{Prob}(\text{USEALL}(N-n, b)) + \text{Prob}(\text{USEALL}(n, a)) - \text{Prob}(\text{USEALL}(N-n, b)) * \text{Prob}(\text{USEALL}(n, a))] \quad (2.3)$$

Having derived a tractable equation for the dichotomization probability, we now discuss our strategy for calculating the capacity. (The following is only a sketch of the proof which is given in Sections 3, 4, and 5.) Our proof will rely heavily on the Law of Large Numbers [13, p. 243]. With probability approaching one, we will have about $N/2$ A's and $N/2$ B's. Further, we will show in Section 4 that the expected number of outer A's is 2D. The Law of Large Numbers will also apply in this case and with probability approaching one, we will have about 2D outer A's and 2D outer B's. In Section 3, we will show that Prob(USEALL(n, c)) is very close to zero if $c \ll n \ln n$ and very close to one if $c \gg n \ln n$. Hence we require about $n \ln n$ choices to use all of n objects. Given $N/2$ A's we will need about $(N/2) \ln (N/2)$ choices (i.e., "outer A's") for every A to become inactive. Since we have 2D choices in D dimensions and need $(N/2) \ln (N/2)$, the capacity occurs when $2D = (N/2) \ln (N/2)$ or when $D = (N/4) \ln (N/2)$. In this case $N \simeq 4D/\ln 2D$, giving the capacity.

(Note that since we have a nearly equal number of A's and B's that if we can (cannot) box the A'S, we almost certainly can (cannot) box the B's. Thus, the bracketed term in (2.3) is $0+0-0*0=0$ or $1+1-1*1=1$.)

Section 3. The Behavior of USEALL

In this section, we prove some lemmas about the behavior of USEALL. We begin with an obvious one.

Lemma 3.1: If $n_1<n_2$ and $c_1>c_2$,

$$\mathrm{Prob}(\mathrm{USEALL}(n_1,c_1)) > \mathrm{Prob}(\mathrm{USEALL}(n_2,c_2)) \qquad [\,]$$

For a given n, we will need to know how large c can be before Prob(USEALL(n,c)) drops to approximately zero, and we must also prove that Prob(USEALL(n,c)) drops sharply at this point. The natural candidate for this point is the mean number of choices required to choose all n objects. Let $T_1,T_2,\ldots,T_n$ be the numbers of the choices when an unchosen element is chosen, and let $W_i=T_i-T_{i-1}$, (where $T_0=0$), i.e. the "waiting time" between choosing the i-1st and ith unchosen element. Clearly, $E(W_i)=n/(n-i+1)$. Therefore the expected number of choices required is

$$E(W_1 + \ldots + W_n) = \sum_{i=1}^{n} n/(n-i+1) = \sum_{i=1}^{n} n/i \simeq n \ln n.$$

(We will refer back to this calculation in the conclusion.) We now prove that Prob(USEALL(n,k)) does drop sharply at k=n ln n.

Lemma 3.2:

If $c < 1$ then Prob(USEALL(n,cn ln n)) --> 0
If $c > 1$ then Prob(USEALL(n,cn ln n)) --> 1.

Proof: Given n distinct objects, the number of the n^k sequences of length k containing every object at least once can be easily calculated using the Principle of Inclusion and Exclusion (see [11, p. 101]) as

$$\sum_{i=0}^{n} (-1)^i \binom{n}{i} (n-i)^k$$

Hence $$\text{Prob(USEALL}(n,k)) = \frac{\sum_{i=0}^{n} (-1)^i \binom{n}{i} (n-i)^k}{n^k}$$

Then for any $c > 0$

$$\lim_{n \to \infty} \text{Prob(USEALL}(n, cn \ln n)) = \lim_{n \to \infty} \sum_{i=0}^{n} (-1)^i \binom{n}{i} \left(\frac{n-i}{n}\right)^{cn \ln n} \quad (1)$$

for large n, $\left(\frac{n-i}{n}\right)^n = \left(1 - \frac{i}{n}\right)^n \simeq e^{-i}$

So (1) is approximately

$$\lim_{n \to \infty} \sum_{i=0}^{n} (-1)^i \binom{n}{i} e^{-i\, c \ln n}$$

$$= \lim_{n \to \infty} \sum_{i=0}^{n} (-1)^i \binom{n}{i} n^{-ci}$$

$$= \lim_{n \to \infty} \sum_{i=0}^{n} \binom{n}{i} \left[\frac{-1}{n^c}\right]^i$$

$$= \lim_{n \to \infty} \left(1 - \frac{1}{n^c}\right)^n$$

by the binomial formula. To evaluate this limit, set

$$L = \lim_{n \to \infty} \left(1 - \frac{1}{n^c}\right)^n$$

then

$$\ln L = \lim_{n \to \infty} n \ln \left(1 - \frac{1}{n^c}\right)$$

$$= \lim_{n \to \infty} \frac{\ln(1-n^{-c})}{1/n}$$

$$= \lim_{n \to \infty} \frac{\frac{1}{1-n^{-c}} \, (cn^{-c-1})}{-1/n^2} \quad \text{by l'Hopital's rule}$$

$$= \lim_{n \to \infty} - \frac{cn}{n^c - 1}$$

$$= \lim_{n \to \infty} - \frac{c}{cn^{c-1}} = \lim_{n \to \infty} - \frac{1}{n^{c-1}} \quad \text{again by l'Hopital's rule}$$

Clearly,

if $c < 1$, $\ln L = -\infty$ so $L = 0$

if $c = 1$, $\ln L = -1$ so $L = e^{-1}$

if $c > 1$, $\ln L = 0$ so $L = 1$ []

We will require the following, now general version of Lemma 3.2

Lemma 3.3: Given a function $g(n)$, let

$L = \lim_{n \to \infty} g(n)/(n \ln n)$ then

If $L < 1$ Prob(USEALL(n,g(n))) -> 0

If $L > 1$ Prob(USEALL(n,g(n))) -> 1

Proof: We consider the case where $L < 1$; the other is similar. Let $L = 1 - x$ and choose any ε such that $x > \varepsilon > 0$. Then we have two facts:

(1) There is an N_1 such that for all $n > N_1$

$$g(n)/(n \ln n) < 1 - x + \varepsilon$$

so $g(n) < (1 - x + \varepsilon)(n \ln n)$

(2) There is an N_2 such that for all $n > N_2$

Prob(USEALL(n, $(1 - x + \varepsilon)(n \ln n)$)) < ε

by Lemma 3.2 since $1 - x + \varepsilon < 1$.

Then for all $n > \mathrm{MAX}(N_1, N_2)$

Prob(USEALL(n, g(n)))

< Prob(USEALL(n, $(1 - x + \varepsilon)(n \ln n)$)) by Lemma 3.1

< ε

Hence Prob(USEALL(n, g(n))) -> 0 [ ]

Section 4. <u>The Behavior of A, OA, and OB</u>

In this section, we not only calculate the means of A, OA and OB but show that with probability approaching 1 these random variables are "close" to their means.

Clearly, E(A) = N/2. By the Law of Large Numbers [13, p. 243], for any $\varepsilon>0$

$$\text{Prob}(|A/N - 1/2| < \varepsilon) \rightarrow 1. \qquad (4.1)$$

To determine E(OA), note that OA can be expressed as $X_1+..+X_{2D}$ where X_{2i-1} is the number of A's to the left of the leftmost B in dimension i and X_{2i} is the number to the right of the rightmost B. Clearly for $k\geq 0$

$$\text{Prob}(X_i = k) \simeq (A/N)^k * ((N-A)/N)$$

we start at the leftmost (rightmost) position and choose points until a B is encountered. Since N and A will be very large and $E(X_i)$ will be very small, we assume the points are chosen <u>with replacement</u>. This gives very good approximation. For any given value for A, we have a sequence of Bernoulli trials with probability of "success" (choosing a class B point) equal to (N-A)/N. We require, on the average, N/(N-A) choices to get the first success. Hence $E(X_i) \simeq N/(N-A) - 1 = A/(N-A)$ and $E(OA)=2DE(X_i) \simeq 2D(A/(N-A))$. Since OA is the sum of random variables with common distribution and finite mean, the Law of Large Numbers applies and

$$\text{Prob}(|OA/2D - A/(N-A)| < \varepsilon) \to 1 \qquad (4.2)$$

for all $\varepsilon > 0$. Similarly

$$\text{Prob}(|OB/2D - (N-A)/A| < \varepsilon) \to 1 \qquad (4.3)$$

Note that E(OA) and E(OB) are functions of A and that we cannot simply plug E(A) into A/(N-A) in the formulas for E(OA) and E(OB). We can, however, show that with probability approaching 1, A is "close" to its mean and that E(OA) and E(OB) are in fact "close" to their means when A/(N-A) = (N-A)/A = 1.

Definition: For any $\varepsilon > 0$ let the event <u>T(ε)</u> be

$$\text{"}\left|\frac{A}{N} - \frac{1}{2}\right| < \varepsilon \ \underline{\text{and}}\ \left|\frac{OA}{2D} - 1\right| < \varepsilon \ \underline{\text{and}}\ \left|\frac{OA}{2D} - 1\right| < \varepsilon\text{"}$$

Theorem 4.1: For any $\varepsilon > 0$, Prob(T(ε)) -> 1.

Proof: For any $\varepsilon > 0$, define the following events:

U to be $\text{"}\left|\frac{A}{N} - \frac{1}{2}\right| < \frac{\varepsilon}{8}\text{"}$

V to be $\text{"}\left|\frac{OA}{2D} - \frac{A}{N-A}\right| < \frac{\varepsilon}{4}\text{"}$

W to be $\text{"}\left|\frac{OA}{2D} - 1\right| < \varepsilon\text{"}$

We show Prob(W) -> 1. From (4.1) and (4.2) we have Prob(U) -> 1 and Prob(V) -> 1 hence Prob(U <u>and</u> V) -> 1. We now show Prob(W) $\geq$ Prob(U and V) by showing that if "U <u>and</u> V" occurs, then W occurs.

Given that U occurs,

$$\left(\frac{1}{2}-\frac{\varepsilon}{8}\right)N \le A \le \left(\frac{1}{2}+\frac{\varepsilon}{8}\right)N$$

so

$$\frac{\frac{1}{2}-\frac{\varepsilon}{8}}{\frac{1}{2}+\frac{\varepsilon}{8}} \le \frac{A}{N-A} \le \frac{\frac{1}{2}+\frac{\varepsilon}{8}}{\frac{1}{2}-\frac{\varepsilon}{8}}$$

Given that V occurs

$$\frac{A}{N-A}-\frac{\varepsilon}{4} \le \frac{0A}{2D} \le \frac{A}{N-A}+\frac{\varepsilon}{4}$$

so

$$\frac{8-2\varepsilon}{8+2\varepsilon}-\frac{\varepsilon}{4} \le \frac{0A}{2D} \le \frac{8+2\varepsilon}{8-2\varepsilon}+\frac{\varepsilon}{4}$$

Now W must occur because

$$\frac{8-2\varepsilon}{8+2\varepsilon}-\frac{\varepsilon}{4} = 1-\frac{\varepsilon}{2}+\frac{\varepsilon^2}{8+2\varepsilon}-\frac{\varepsilon}{4} \ge 1-\frac{3\varepsilon}{4} \ge 1-\varepsilon$$

and (assuming $\varepsilon < 1$)

$$\frac{8+2\varepsilon}{8-2\varepsilon}+\frac{\varepsilon}{4} = 1+\frac{\varepsilon}{2}+\frac{\varepsilon^2}{8-2\varepsilon}+\frac{\varepsilon}{4} \le 1+\frac{\varepsilon}{2}+\frac{\varepsilon^2}{8\varepsilon-2\varepsilon}+\frac{\varepsilon}{4} = 1+\frac{11\varepsilon}{12} \le 1+\varepsilon$$

Therefore Prob(U <u>and</u> V) $\le$ Prob(W). Hence Prob(W) -> 1 as desired.

Similarly we can show

$$\text{Prob}\left(\left|\frac{0B}{2D}-1\right| < \varepsilon\right) \to 1$$

hence Prob(T(ε)) -> 1 since the probability of each of the three component events goes to 1. []

Section 5. Calculating the Capacity

In this section we use the results from Sections 3 and 4 to calculate the capacity.

Theorem 5.1: Let D=c(N/4) ln (N/2) then

If c < 1 then $\lim_{N \to \infty}$ Prob(CANBOX) = 0

and

If c > 1 then $\lim_{N \to \infty}$ Prob(CANBOX) = 1

Proof:

Since Prob(CANBOX) =

Prob(T(ε)) * Prob(CANBOX | T(ε)) + Prob(~T(ε)) * Prob(CANBOX | ~T(ε))

(where ~T(ε) is the complement of the event T(ε))

we have

$\lim_{N \to \infty}$ Prob(CANBOX) = $\lim_{N \to \infty}$ Prob(CANBOX | T(ε)) for all ε>0 (1)

since Prob(T(ε)) -> 1 by Lemma 4.1. From equation (2.3) we have

Prob(CANBOX | T(ε)) =

$\sum_{n,a,b}$ Prob(A=n, OA=a, OB=b | T(ε)) *

[Prob(USEALL(N-n,b) | T(ε)) + Prob(USEALL(n,a) | T(ε)) -

Prob(USEALL(N-n,b) | T(ε)) * Prob(USEALL(n,a) | T(ε))]

Given that T(ε) occurs

$$(\tfrac{1}{2}-\varepsilon)N \le n \le (\tfrac{1}{2}+\varepsilon)N$$

and

$$(1-\varepsilon)2D \le a,b \le (1+\varepsilon)2D$$

Hence Prob(USEALL(N-n, b) | T(ε)) and Prob(USEALL(n,a) | T(ε))

$< \text{Prob}(\text{USEALL}((\tfrac{1}{2}-\varepsilon)N, (1+\varepsilon)\ c\ (N/2) \ln (N/2))$

by Lemma 3.1. Let u(N,ε) denote this last quantity. Hence Prob(CANBOX)

$< \sum_{n,a,b} \text{Prob}(A=n,\ OA=a,\ OB=b \mid T(\varepsilon)) * 2u(N,\varepsilon)$

$= 2u(N,\varepsilon) * \sum_{n,a,b} \text{Prob}(A=n,\ OA=a,\ OB=b \mid T(\varepsilon))$

$= 2u(N,\varepsilon)$

Hence

$$\lim_{N\to\infty} \text{Prob}(\text{CANBOX} \mid T(\varepsilon)) < \lim_{N\to\infty} 2u(N,\varepsilon) \text{ for all } \varepsilon > 0 \qquad (2)$$

Now

$$u(N,\varepsilon) = \text{Prob}(\text{USEALL}(1/2-\varepsilon)N,\ (1+\varepsilon)\ c\ (N/2) \ln (N/2)))$$

making the substitution $M = (1/2-\varepsilon)N$ gives

$$u(N,\varepsilon) = \text{Prob}(\text{USEALL}(M,g(M))) \text{ where}$$

$$g(M) = (1+\varepsilon)c\ M\ (\ln M - \ln (1-2\varepsilon))/(1-2\varepsilon)$$

Hence

$$\lim_{M\to\infty} g(M)/(M \ln M) = (1+\varepsilon)c/(1-2\varepsilon)$$

The limit is less than 1 if $\varepsilon < (1-c)/(2+c)$. Therefore

$$\lim_{N\to\infty} u(N,\varepsilon) = 0 \text{ if } \varepsilon < (1-c)/(2+c) \qquad (3)$$

Since $c < 1$, $(1-c)/(2+c) > 0$ hence an $\varepsilon > 0$ exists for which (3) holds. The result then follows from (1), (2) and (3). []

Definition: We say $f(N) = g(N) + o(h(N))$

if $\frac{f(N) - g(N)}{h(N)} \to 0$

Lemma 5.1: If $D = (N/4) \ln (N/2)$ then

$$N = 4D/\ln 2D + o(4D/\ln 2D)$$

Proof:

$$\frac{N - 4D/\ln 2D}{4D/\ln 2D}$$

$$\frac{N}{4D/\ln 2D} - 1 =$$

$$\frac{N}{N \ln (N/2)/(\ln (N/2) + \ln \ln (N/2))} - 1$$

$$\frac{N \ln (N/2) + N \ln \ln (N/2)}{N \ln (N/2)} - 1$$

Clearly, the $N \ln (N/2)$ terms dominate and the limit is $1 - 1 = 0$. [ ]

Theorem 5.2: The capacity of a hyperbox is asymptotically $4D/\ln 2D$. [ ]

Remark: $4D/(\ln 2D - \ln \ln 2D)$ gives an even better approximation to the capacity since it will cancel $N \ln \ln (N/2)$ term in the numerator of the last equation in Lemma 5.1 and leave a lower order term.

Section 6. Conclusion

We have analytically calculated the capacity (as defined by Cover) of the class of hyperboxes, an important component of the variable valued logic system of Michalski and an interesting figure in its own right. The capacity was found to be asymptotically equal to $4D/\ln 2D$ (or, more exactly, $4D/(\ln 2D - \ln \ln 2D)$).

This calculation also revealed that the probability a dichotomy of N points can be achieved using a hyperboxes has the same behavior that for using hyperplanes. That is, it is close to one for a small number of points and remains very close to one until N reaches the capacity, where it drops sharply to zero. This, of course, was not known a priori (the probability might slope very slowly down toward zero). Therefore, capacity is a meaningful quantity for hyperboxes, a figure very different from polynomial functions, indicating the robustness of Cover's definition.

For large D, the capacity is less than the capacity of hyperplanes (which is 2D + 2). The case of small D is studied in [12] where the capacity and efficiency of hyperboxes and hyperplanes is compared and the hyperbox was found to be more efficient. A final point of interest is the "law of diminishing returns" for hyperboxes. For hyperplanes, each new dimension increases the capacity by a constant amount. However, for a hyperboxes, successive dimensions increase the capacity by smaller and smaller amounts (the derivative of the capacity is $[4 \ln 2D - 4]/\ln^2 2D$). This can be explained by our calculation of the mean number of choices required to choose all of n distinct objects (see Section 2). The waiting times between choices of unchosen objects (W_i) get longer as more objects are chosen and it becomes less likely an unchosen element will be chosen. Thus, later dimensions are less and less affective in dichotomizing the points.

REFERENCES

[1] T. M. Cover, "Capacity problems for linear machines," in _Pattern Recognition_, L. Kanal, Ed., Washington, DC: Thompson Book Co, 1968.

[2] ____, "Classification and generalization capabilities of linear threshold logic units," Rome Air Develop. Center Tech. Documentary, Rep. DADC-TDR-64-32, Feb. 1964.

[3] R. S. Michalski and B. H. McCormick, "Interval generalization of switching theory," Univ. of Illinois Dep. Comp. Sci. Rep. 442, May 1971.

[4] R. S. Michalski, "A geometric model for the synthesis of interval covers," Univ. of Illinois Dep. Comput. Sci. Rep. 461, June 1971.

[5] N. J. Nilsson, _Learning Machines_, New York: McGraw-Hill, 1965.

[6] T. M. Cover, "Geometrical and statistical properties of linear threshold devices," Stanford Electron. Lab. Tech. Rep. 6107-1, May 1964.

[7] R. J. Brown, "Adaptive multiple-output threshold systems and their storage capacities," Stanford Electron. Lab., Stanford Univ., Tech. Rep. 667-1, June 1964.

[8] R. S. Michalski, "Variable-valued logic: System VL_1," Presented at the 1974 Int. Symp. Multiple-Valued Logic, West Virginia Univ., Morgantown, WV, May 29-31, 1974.

[9] T. M. Cover, "Geometric and statistical properties of systems of linear inequalities with application to pattern recognition," _IEEE Trans. Electron. Comput._, vol. EC-14, pp. 326-334, June 1965.

[10] ____, "The number of inducible orderings of points in d-space," _SIAM J. Appl. Math._, vol. 15, pp. 434-439, Mar. 1967.

[11] C. L. Liu, _Introduction to Combinatorial Mathematics_, McGraw-Hill, New York, 1968.

[12] J. R. Bitner, "Capacity and efficiency of decision functions," IEEE Transactions on Computers, vol. C-26, November 1977.

[13] W. Feller, _An Introduction to Probability Theory and its Applications_, vol. 1, Wiley and Sons, New York, 1968.