

A Generalization of Sauer's Lemma to Classes of Large-Margin Functions^{*} ^{**}

Joel Ratsaby

University College London
Gower Street, London WC1E 6BT, United Kingdom
J.Ratsaby@cs.ucl.ac.uk,

WWW home page: <http://www.cs.ucl.ac.uk/staff/J.Ratsaby/>

Abstract. We generalize Sauer's Lemma to classes H of binary-valued functions on $[n] = \{1, \dots, n\}$ which have a margin of at least N on every element in a sample $S \subseteq [n]$ of cardinality l , where the margin $\mu_h(x)$ of $f \in F$ on a point $x \in [n]$ is defined as the largest non-negative integer a such that h is constant on the interval $I_a(x) = [x - a, x + a] \subseteq [n]$.

1 Introduction

Estimation of the complexity of classes of binary-valued functions has been behind much of recent developments in the of theory learning. In a seminal paper Vapnik and Chervonenkis [1971] applied the law of large numbers uniformly over an infinite class \mathcal{F} of binary functions, i.e., indicator functions of sets A in a general domain X , and showed that the complexity of the problem of learning pattern recognition from samples of n randomly drawn examples can be characterized in terms of a combinatorial complexity of \mathcal{F} .

This complexity, known as the *growth function* of \mathcal{F} and denoted by $\phi_{\mathcal{F}}(n)$, counts the maximal number of dichotomies, i.e., binary vectors corresponding to the restriction of functions $f \in \mathcal{F}$ on a finite subset $S \subset X$ of cardinality n , where the maximum runs over all such S . The Vapnik-Chervonenkis dimension of \mathcal{F} , denoted as $VC(\mathcal{F})$, plays a crucial role in controlling the rate of the growth of $\phi_{\mathcal{F}}(n)$ with respect to n . Such binary vectors may be viewed as binary-valued functions on a finite domain $[n] \equiv \{1, \dots, n\}$ and hence form a finite class \mathcal{H} of the same VC-dimension as \mathcal{F} .

In this paper we consider classes of binary-valued functions on $[n]$ which satisfy a constraint of having a large margin on any one set $S \subset X$ of cardinality l . We obtain an estimate on the cardinality of such a class.

Recently there has been interest in learning classes via maximizing the margin [see for instance Vapnik, 1998, Cristianini and Shawe-Taylor, 2000]. The usual approach analyzes the growth rate (or more precisely the covering number) of

^{*} Paper appeared in the *Proc. of Third Colloq. on Math. and Comp. Sci. Alg., Trees, Combin. and Probb. (MATHINFO 2004)*, Vienna Austria, Sept. 2004.

^{**} University College London, Computer Science Department, Technical Report RN/03/13

classes of real-valued functions with a large-margin on some set (sample) S . So to the best of the author's knowledge, the approach taken in the current paper of estimating the complexity of classes of large-margin binary-valued functions by a generalization of Sauer's lemma is novel.

Before discussing this further we first introduce some needed notation.

2 Some notations, definitions and existing results

Let $\mathbb{I}(E)$ denote the indicator function which equals 1 if the expression E is true and 0 otherwise. Let F be a class of functions $f : [n] \rightarrow \{0, 1\}$. For a set $A = \{a_1, \dots, a_k\} \subseteq [n]$ denote by $f|_A = [f(a_1), \dots, f(a_k)]$. F is said to *shatter* A if

$$|\{f|_A : f \in F\}| = 2^k.$$

The Vapnik-Chervonenkis dimension of F , denoted as $VC(F)$, is defined as the cardinality of the largest set shattered by F .

Sauer [1972] obtained the following result:

Lemma 1. [Sauer, 1972] *If the VC-dimension of F is d then*

$$|F| \leq \sum_{i=0}^d \binom{n}{i}.$$

We note that the bound is tight as for all $d, n \geq 1$ there exist classes $F \subseteq 2^{[n]}$ of VC-dimension d which achieve the equality.

Consider the following definition of functional margin¹ which naturally suits binary-valued functions.

Definition 1. *The margin $\mu_f(x)$ of $f \in F$ on an element $x \in [n]$ is the largest non-negative integer a such that f has a constant value of either 0 or 1 on the interval set $I_a(x) = \{x - a, \dots, x + a\}$ provided that $I_a(x) \subseteq [n]$.*

The sample-margin $\mu_S(f)$ of f on a subset $S \subseteq [n]$ is defined as

$$\mu_S(f) \equiv \min_{x \in S} \mu_f(x).$$

More generally, this definition applies also to classes on other domains X if there is a linear ordering on X .

3 Motivation and aim of the paper

In recent years, in search for better learning algorithms, it has been discovered [see for instance Vapnik, 1998, Cristianini and Shawe-Taylor, 2000] that learning

¹ For other definitions of margin see for instance Cristianini and Shawe-Taylor [2000].

classes \mathcal{F} of real-valued functions that are restricted to have a large margin on a training sample leads to a faster learning-error convergence².

The crucial reason for such improved error bounds is the fast decrease of the covering number of \mathcal{F} , which is analogous to the growth-number in the case of classes of binary-valued functions, with respect to an increase in sample-margin value. In this paper, we consider the latter case, restricting to a finite class $\mathcal{H}(S)$ of binary valued functions h on $[n]$ with the constraint that for a fixed subset $S \subset [n]$ of size l , for all $h \in \mathcal{H}(S)$, $\mu_S(h) \geq N$.

By generalizing Sauer's result (Lemma 1), we obtain an estimate on the cardinality of $\mathcal{H}(S)$. Being dependent on the margin parameter N , our estimate may be viewed as being analogous to existing results that bound the covering number of classes of finite-pseudo-dimension (or fat_γ dimension) which consists of real-valued functions under a similar margin constraint [see for instance Anthony and Bartlett, 1999, Ch. 12].

4 Technical results

We start with an auxiliary lemma:

Lemma 2. *For $N \geq 0$, $n \geq 0$, $0 \leq m \leq n$, let $w_{m,N}(n)$ be the number of standard (one-dimensional) ordered partitions of a nonnegative integer n into m parts each no larger than N . Then*

$$w_{m,N}(n) = \begin{cases} \mathbb{I}(n=0) & \text{if } m=0 \\ \sum_{i=0, N+1, 2(N+1), \dots}^n (-1)^{i/(N+1)} \binom{m}{i/(N+1)} \binom{n-i+m-1}{n-i} & \text{if } m \geq 1. \end{cases}$$

Remark 1. While our interest is in $[n] = \{1, \dots, n\}$, we allow $w_{m,N}(n)$ to be defined on $n=0$ for use by Lemma 3.

Proof: The generating function (g.f.) for $w_{m,N}(n)$ is

$$W(x) = \sum_{n \geq 0} w_{m,N}(n) x^n = \left(\frac{1 - x^{N+1}}{1 - x} \right)^m.$$

When $m=0$ the only non-zero coefficient is of x^0 and it equals 1 so $w_{0,N}(n) = \mathbb{I}(n=0)$. Let $T(x) = (1 - x^{N+1})^m$ and $S(x) = \left(\frac{1}{1-x} \right)^m$. Then

$$T(x) = \sum_{i=0}^m (-1)^i \binom{m}{i} x^{i(N+1)}$$

² Hence samples on which the target function (the one to be learnt) has a large margin are of considerable worth. Ratsaby [2003] estimated the complexity of such samples as a function of the margin parameter and sample size.

which generates the sequence $t_N(n) = \binom{m}{n/(N+1)}(-1)^{n/(N+1)}\mathbb{I}(n \bmod (N+1) = 0)$. Similarly, for $m \geq 1$, it is easy to show $S(x)$ generates $s(n) = \binom{n+m-1}{n}$. The product $W(x) = T(x)S(x)$ generates their convolution $t_N(n) \star s(n)$, namely,

$$w_{m,N}(n) = \sum_{i=0, N+1, 2(N+1), \dots, n} (-1)^{i/(N+1)} \binom{m}{i/(N+1)} \binom{n-i+m-1}{n-i}. \square$$

Remark 2. By an alternate proof one obtains a slightly simpler form of

$$w_{m,N}(n) = \sum_{k=0}^m (-1)^k \binom{m}{k} \binom{n+m-1-k(N+1)}{m-1},$$

over $m \geq 1$.

Before proceeding to the main theorem we have two additional lemmas.

Lemma 3. *Let the integer $1 \leq N \leq n$ and consider the class F consisting of all binary-valued functions f on $[n]$ which take the value 1 on no more than $r \leq n$ elements of $[n]$ and whose margin on any element $x \in [n]$ satisfies $\mu_f(x) \leq N$. Then*

$$|F| = \sum_{k=0}^r \sum_{m=1}^n c(k, n-k; m, N) \equiv \beta_r^{(N)}(n)$$

where

$$\begin{aligned} & c(k, n-k; m, N) \\ &= \sum_{i,j=0}^1 w_{m-i, 2N}(k-m+1-i(2N+1)) w_{m-j, 2N}(n-k-m+1-j(2N+1)). \end{aligned} \tag{1}$$

Proof: Consider the integer pair $[k, n-k]$, where $n \geq 1$ and $0 \leq k \leq n$. A two-dimensional ordered m -partition of $[k, n-k]$ is an ordered partition into m two-dimensional parts, $[a_j, b_j]$ where $0 \leq a_j, b_j \leq n$ but not both are zero and where $\sum_{j=1}^m [a_j, b_j] = [k, n-k]$. For instance, $[2, 1] = [0, 1] + [2, 0] = [1, 1] + [1, 0] = [2, 0] + [0, 1]$ are three partitions of $[2, 1]$ into two parts (for more examples see Andrews [1998]).

Suppose we add the constraint that only a_1 or b_m may be zero while all remaining $a_j, b_k \geq 1$, $2 \leq j \leq m$, $1 \leq k \leq m-1$. Denote any partition that satisfies this as *valid*. For instance, let $k=2$, $m=3$ then the m -partitions of $[k, n-k]$ are: $\{[0, 1][1, 1][1, n-4]\}, \{[0, 1][1, 2][1, n-5]\}, \dots, \{[0, 1][1, n-3][1, 0]\}, \{[0, 2][1, 1][1, n-5]\}, \{[0, 2][1, 2][1, n-6]\}, \dots, \{[0, 2][1, n-4][1, 0]\}, \dots, \{[0, n-3][1, 1][1, 0]\}$. For $[k, n-k]$, let $\mathcal{P}_{n,k}$ be the collection of all valid partitions of $[k, n-k]$.

Let F_k denote all binary functions on $[n]$ which take the value 1 over exactly k elements of $[n]$. Define the mapping $\Pi : F_k \rightarrow \mathcal{P}_{n,k}$ where for any $f \in F_k$

the partition $\Pi(f)$ is defined by the following procedure: Start from the first element of $[n]$, i.e., 1. If f takes the value 1 on it then let a_1 be the length of the constant 1-segment, i.e., the set of all elements starting from 1 on which f takes the constant value 1. Otherwise if f takes the value 0 let $a_1 = 0$. Then let b_1 be the length of the subsequent 0-segment on which f takes the value 0. Let $[a_1, b_1]$ be the first part of $\Pi(f)$. Next, repeat the following: if there is at least one more element of $[n]$ which has not been included in the preceding segment, then let a_j be the length of the next 1-segment and b_j the length of the subsequent 0-segment. Let $[a_j, b_j]$, $j = 1, \dots, m$, be the resulting sequence of parts where m is the total number of parts. Only the last part may have a zero valued b_m since the function may take the value 1 on the last element n of $[n]$ while all other parts, $[a_j, b_j]$, $2 \leq j \leq m-1$, must have $a_j, b_j \geq 1$. The result is a valid partition of $[k, n-k]$ into m parts.

Clearly, every $f \in F_k$ has a unique partition. Therefore Π is a bijection. Moreover, we may divide $\mathcal{P}_{n,k}$ into mutually exclusive subsets V_m consisting of all valid partitions of $[k, n-k]$ having exactly m parts, where $1 \leq m \leq n$. Thus

$$|F_k| = \sum_{m=1}^n |V_m|.$$

Consider the following constraint on components of parts:

$$a_i, b_i \leq 2N + 1, \quad 1 \leq i \leq m. \quad (2)$$

Denote by $V_{m,N} \subset \mathcal{P}_{n,k}$ the collection of valid partitions of $[k, n-k]$ into m parts each of which satisfies this constraint.

Let $F_{k,N} = F \cap F_k$ consist of all functions satisfying the margin constraint in the statement of the lemma and having exactly k ones. Note that f having a margin no larger than N on any element of $[n]$ implies there does not exist a segment a_i or b_i of length larger than $2N + 1$ on which f takes a constant value. Hence the parts of $\Pi(f)$ satisfy (2). Hence, for any $f \in F_{k,N}$, its unique valid partition $\Pi(f)$ must be in $V_{m,N}$. We therefore have

$$|F_{k,N}| = \sum_{m=1}^n |V_{m,N}|. \quad (3)$$

By definition of F it follows that

$$|F| = \sum_{k=0}^r |F_{k,N}|. \quad (4)$$

Let us denote by

$$c(k, n-k; m, N) \equiv |V_{m,N}| \quad (5)$$

the number of valid partitions of $[k, n-k]$ into exactly m parts whose components satisfy (2). In order to determine $|F|$ it therefore suffices to determine $c(k, n-k; m, N)$.

We next construct the generating function

$$G(t_1, t_2) = \sum_{\alpha_1 \geq 0} \sum_{\alpha_2 \geq 0} c(\alpha_1, \alpha_2; m, N) t_1^{\alpha_1} t_2^{\alpha_2}. \quad (6)$$

For $m \geq 1$,

$$\begin{aligned} G(t_1, t_2) &= (t_1^0 + t_1^1 + \cdots + t_1^{2N+1})(t_2^0 + t_2^1 + \cdots + t_2^{2N+1})^{\mathbb{I}(m \geq 2)} \\ &\quad \cdot ((t_1^1 + \cdots + t_1^{2N+1})(t_2^1 + \cdots + t_2^{2N+1}))^{(m-2)_+} \\ &\quad \cdot (t_1^1 + \cdots + t_1^{2N+1})^{\mathbb{I}(m \geq 2)} (t_2^0 + t_2^1 + \cdots + t_2^{2N+1}) \end{aligned} \quad (7)$$

where the values of the exponents of all terms in the first and second factors represent the possible values for a_1 and b_1 , respectively. The values of the exponents in the middle $m-2$ factors are for the values of $a_j, b_j, 2 \leq j \leq m-1$ and those in the factor before last and last are for a_m and b_m , respectively. Equating this to (6) implies the coefficient of $t_1^{\alpha_1} t_2^{\alpha_2}$ equals $c(\alpha_1, \alpha_2; m, N)$ which we seek.

The right side of (7) equals

$$\begin{aligned} &t_1^{m-1} t_2^{m-1} \left(\left(\frac{1-t_1^{2N+1}}{1-t_1} \right)^m + t_1^{2N+1} \left(\frac{1-t_1^{2N+1}}{1-t_1} \right)^{m-1} \right) \\ &\quad \cdot \left(\left(\frac{1-t_2^{2N+1}}{1-t_2} \right)^m + t_2^{2N+1} \left(\frac{1-t_2^{2N+1}}{1-t_2} \right)^{m-1} \right). \end{aligned} \quad (8)$$

Let $W(x) = \left(\frac{1-x^{2N+1}}{1-x} \right)^{m-1}$ generate $w_{m-1, 2N}(n)$ which is defined in Lemma 2. So (8) becomes

$$\begin{aligned} &\sum_{\alpha_1, \alpha_2 \geq 0} \left(w_{m, 2N}(\alpha_1) w_{m, 2N}(\alpha_2) t_1^{\alpha_1+m-1} t_2^{\alpha_2+m-1} \right. \\ &\quad + w_{m-1, 2N}(\alpha_1) w_{m, 2N}(\alpha_2) t_1^{\alpha_1+m+2N} t_2^{\alpha_2+m-1} \\ &\quad + w_{m, 2N}(\alpha_1) w_{m-1, 2N}(\alpha_2) t_1^{\alpha_1+m-1} t_2^{\alpha_2+m+2N} \\ &\quad \left. + w_{m-1, 2N}(\alpha_1) w_{m-1, 2N}(\alpha_2) t_1^{\alpha_1+m+2N} t_2^{\alpha_2+m+2N} \right). \end{aligned} \quad (9)$$

Equating the coefficients of $t_1^{\alpha'_1} t_2^{\alpha'_2}$ in (6) and (9) yields

$$\begin{aligned} &c(\alpha'_1, \alpha'_2; m, N) \\ &= \sum_{i, j=0}^1 w_{m-i, 2N}(\alpha'_1 - m + 1 - i(2N+1)) w_{m-j, 2N}(\alpha'_2 - m + 1 - j(2N+1)). \end{aligned}$$

Substituting k for $\alpha'_1, n-k$ for α'_2 , combining (3), (4) and (5) yields the result. \blacksquare

The next lemma extends the result of Lemma 3 to classes \mathcal{H} of finite VC-dimension.

Lemma 4. *Let $n \geq 1$ and $0 \leq d \leq n$. Let \mathcal{H} be a class of binary-valued functions h on $[n]$ satisfying $\mu_h(x) \leq N$ on any $x \in [n]$ and let $VC(\mathcal{H}) \leq d$. Then*

$$|\mathcal{H}| \leq \beta_d^{(N)}(n)$$

where $\beta_d^{(N)}$ is defined in Lemma 3.

Proof: The proof builds on that of Lemma 7 in Haussler & Long [1995] which considered generalizations of the VC-dimension and is done by double induction on n and d .

Start with the case $d = 0$, the bound reduces to $|\mathcal{H}| \leq 1$ since $\beta_0^{(N)}(n) \leq 1$ when $n \geq 1$. The bound is correct since if $|\mathcal{H}| > 1$ then it implies there are two distinct functions h, g . Let $k \in [n]$ be the element on which they differ. Then the singleton $\{k\}$ is shattered by \mathcal{H} hence the VC-dimension of \mathcal{H} is at least 1 which contradicts the assumption that $d = 0$ hence $|\mathcal{H}| \leq 1$ and the lemma holds.

Next, suppose $d = n$. Consider the class F in Lemma 3 with $r = n$. Such F consists of all binary-valued functions f on $[n]$ which satisfy the margin constraint $\mu_f(x) \leq N$ on every $x \in [n]$. By Lemma 3, $|F| \leq \beta_n^{(N)}(n)$. Clearly by definition, $\mathcal{H} \subseteq F$. Hence $|\mathcal{H}| \leq \beta_n^{(N)}(n)$ as claimed.

Next, suppose $0 < d < n$. Define $\pi : \mathcal{H} \rightarrow \{0, 1\}^{n-1}$ by $\pi(h) = [h(1), \dots, h(n-1)]$. Define $\alpha : \pi(\mathcal{H}) \rightarrow \{0, 1\}$ by $\alpha(u_1, \dots, u_{n-1}) = \min\{v : \exists h \in \mathcal{H}, h(i) = u_i, h(n) = v, 1 \leq i \leq n-1\}$. Define $A = \{h \in \mathcal{H} : h(n) = \alpha(h(1), \dots, h(n-1))\}$ and denote by $A^c = \mathcal{H} \setminus A$. Consider any $h \in \mathcal{H}$. If $\alpha(h(1), \dots, h(n-1)) = 1$ then A^c does not contain h . Otherwise, A^c contains the function g which agrees with h on $1, \dots, n-1$ and $g(n) = 1$. Hence for all $h \in A^c$, $h(n) = 1$.

Make the inductive assumption that the claimed bound holds for all classes \mathcal{H} on any subset of $[n]$ having cardinality $n-1$ and satisfying the margin constraint. Then we claim the following:

Claim 1

$$|A| \leq \beta_d^{(N)}(n-1).$$

This is proved next: the mapping π is one-to-one on A and the set $\pi(A)$ has VC-dimension no larger than d since any subset of $[n]$ shattered by $\pi(A)$ is also shattered by A which is in \mathcal{H} and $VC(\mathcal{H}) \leq d$. Hence by the induction hypothesis

$$|\pi(A)| \leq \beta_d^{(N)}(n-1)$$

and since π is one-to-one then $|A| = |\pi(A)|$. □

Next, under the same induction hypothesis, we have:

Claim 2

$$|A^c| \leq \beta_{d-1}^{(N)}(n-1).$$

We prove this next: First we show that $VC(A^c) \leq d-1$. Let $E \subset [n]$ be shattered by A^c and let $|E| = l$. Note that $n \notin E$ since as noted earlier $h(n) = 1$ for all

$h \in A^c$. For any $b \in \{0, 1\}^{l+1}$ let $h \in A^c$ be such that $h|_E = [b_1, \dots, b_l]$. If $b_{l+1} = 1$ then $h(n) = b_{l+1}$ since all functions in A^c take the value 1 on n . If $b_{l+1} = 0$ then there exists a $g \in A$ which satisfies $g(i) = h(i)$, $1 \leq i \leq n-1$ and $g(n) = \alpha(h(1), \dots, h(n-1))$, the latter being $g(n) = 0$. It follows that $E \cap \{n\}$ is shattered by \mathcal{H} . But by assumption $VC(\mathcal{H}) \leq d$ and $n \notin E$ hence $|E| \leq d-1$. Since E was chosen arbitrarily then $VC(A^c) \leq d-1$. The same argument as in the proof of Claim 1 applied to A^c using $d-1$ to bound its VC-dimension, obtains the statement of Claim 2. \square

From Claims 1 and 2 and recalling the definition of $c(k, n-k; m, N)$ from Lemma 3, it follows that

$$\begin{aligned}
|\mathcal{H}| &\leq \beta_d^{(N)}(n-1) + \beta_{d-1}^{(N)}(n-1) \\
&= \sum_{k=0}^d \sum_{m=1}^{n-1} c(k, n-k-1; m, N) + \sum_{k=0}^{d-1} \sum_{m=1}^{n-1} c(k, n-k-1; m, N) \\
&= \mathbb{I}(n/2 \leq N) + \sum_{k=1}^d \sum_{m=1}^{n-1} c(k, n-k-1; m, N) + \sum_{k=1}^d \sum_{m=1}^{n-1} c(k-1, n-k; m, N) \\
&= \mathbb{I}(n/2 \leq N) + \sum_{k=1}^d \sum_{m=1}^{n-1} (c(k, n-k-1; m, N) + c(k-1, n-k; m, N))
\end{aligned} \tag{10}$$

where the indicator $\mathbb{I}(n/2 \leq N)$ enters here since in case $k=0$ the only valid function h is the constant-0 on $[n]$ satisfying $\mu_h(x) \leq n/2$, for all $x \in [n]$. We now have:

Claim 3

$$\sum_{m=1}^n c(k, n-k; m, N) = \sum_{m=1}^{n-1} (c(k, n-k-1; m, N) + c(k-1, n-k; m, N)). \tag{11}$$

Note that this is a recurrence formula for the number (left hand side of (11)) of valid partitions of $[k, n-k]$ (excluding the case $k=0$) into parts that satisfy (2).

We prove the claim next: given any such partition π_n there is exactly one of four possible ways that it can be constructed by adding a part to a valid two dimensional partition π_{n-1} of $[n-1]$ while still satisfying (2). The first two amount to starting from a partition π_{n-1} of $[k-1, n-k]$ and: (i) adding the part $[1, 0]$ algebraically to any existing part in π_{n-1} , e.g., $[x, y] + [1, 0] = [x+1, y]$, to obtain a π_n with $[x+1, y]$ as one of the parts (provided that (2) is still satisfied) which yields a total number of parts no larger than $n-1$ or (ii) adding $[1, 0]$ to π_{n-1} as a new last part to obtain a π_n (provided it is still valid) with no more than n parts. The remaining two ways amount to starting from a partition π_{n-1} of $[k, n-k-1]$ and acting as before except now adding the part $[0, 1]$ instead of

$[1, 0]$ either algebraically or as a new first part. There are

$$\sum_{m=1}^{n-1} c(k, n - k - 1; m, N)$$

valid partitions of $[k, n - k - 1]$ and there are

$$\sum_{m=1}^{n-1} c(k - 1, n - k; m, N)$$

valid partitions of $[k - 1, n - k]$, all satisfying (2). Doing the aforementioned construction to each one of these partitions yields all valid partitions of $[k, n - k]$ that satisfy (2). \square

Continuing, the right hand side of (10) becomes

$$\mathbb{I}(n/2 \leq N) + \sum_{k=1}^d \sum_{m=1}^n c(k, n - k; m, N) = \sum_{k=0}^d \sum_{m=1}^n c(k, n - k; m, N)$$

which is precisely $\beta_d^{(N)}(n)$. This completes the induction. \blacksquare

Theorem 1. *Let $n \geq 1$, $1 \leq l \leq n$ and $0 \leq d < n - l$. Let \mathcal{H} be a class of binary-valued functions h on $[n]$ having $VC(\mathcal{H}) \leq d$. Let $S \subseteq [n]$ be a sample of cardinality l and consider the subclass $\mathcal{H}(S) \subseteq \mathcal{H}$ which consists of all functions $h \in \mathcal{H}$ with a margin $\mu_h(x) > N$ iff $x \in S$. Then*

$$|\mathcal{H}(S)| \leq \beta_d^{(N)}(n - l - 2(N + 1)).$$

Proof: The condition $\mu_h(x) > N$ implies only two types of functions h are allowed, those which take either a constant-0 value or a constant-1 value over all elements in the interval $I_{N+1}(x)$. The condition $\mu_h(x) \leq N$ implies that any function is possible except one that takes a constant-0 or a constant-1 value over $I_{N+1}(x)$ (see Definition 1). Hence clearly the first condition is significantly more restrictive.

Since we seek an upper bound on $|\mathcal{H}(S)|$ then we consider among all sets of $[n]$ of cardinality l a set S with the least restrictive constraint, namely, causing as few elements $x \in [n]$ as possible (except those in S) to have $\mu_h(x) > N$. This is achieved by a maximally-packed set $S^* \subset [n]$ of l elements, for instance $S^* = \{N + 2, \dots, N + l + 1\}$. It yields a minimal-size region $R = \{1, \dots, 2(N + 1) + l\}$ on which every candidate h must take either a constant-0 or constant-1 value, i.e., have a margin larger than N for every $x \in S^*$.

This leaves a maximal-size region $[n] \setminus R$ on which the less stringent constraint of having a margin no larger than N must hold. By Lemma 4, there are no more than $\beta_d^{(N)}(n - l - 2(N + 1))$ functions in \mathcal{H} that satisfy the latter. Hence for any $S \subseteq [n]$ of cardinality $|S| = l$, $|\mathcal{H}(S)| \leq \beta_d^{(N)}(n - l - 2(N + 1))$. \square

5 Conclusions

The main result of the paper is a bound on the cardinality of a class of finite VC-dimension consisting of binary functions on $[n]$ which have a margin greater than N on a set S of cardinality l . This result generalizes the well known Sauer's Lemma and is analogous to existing bounds on the covering number of classes of *real-valued* functions that have a large-margin on a sample S . The result may be used for obtaining the sample complexity of PAC-learning a class of boolean hypotheses while maximizing the margin on a given training sample.

Bibliography

- G. E. Andrews. *The Theory of Partitions*. Cambridge University Press, 1998.
- M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
- N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based learning methods*. Cambridge University Press, 2000.
- D. Haussler and P.M. Long. A generalization of sauer's lemma. *Journal of Combinatorial Theory (A)*, 71(2):219–240, 1995.
- J. Ratsaby. On the complexity of good samples for learning. Technical Report RN/03/12, Department of Computer Science, University College London, September 2003.
- N. Sauer. On the density of families of sets. *J. Combinatorial Theory (A)*, 13:145–147, 1972.
- V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
- V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Apl.*, 16:264–280, 1971.