On the Complexity of Constrained VC-Classes [1]

Joel Ratsaby

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

Abstract

Sauer's Lemma is extended to classes $\mathcal{H}_N$ of binary-valued functions h on $[n] = \{1, \ldots, n\}$ which have a margin less than or equal to N on all $x \in [n]$ with $h(x) = 1$, where the margin $\mu_h(x)$ of h at $x \in [n]$ is defined as the largest non-negative integer a such that h is constant on the interval $I_a(x) = [x - a, x + a] \subseteq [n]$. Estimates are obtained for the cardinality of classes of binary valued functions with a margin of at least N on a positive sample $S \subseteq [n]$.

Key words: Sauer's lemma, VC-dimension, Boolean functions, integer partitions

1 Introduction

Estimation of the complexity of classes of binary-valued functions has been behind much of recent developments in the theory of learning. In a seminal paper Vapnik and Chervonenkis [1971] applied the law of large numbers uniformly over an infinite class $\mathcal{G}$ 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 randomly drawn samples can be characterized in terms of a combinatorial quantity called the *growth function* of $\mathcal{G}$ which is defined as follows:

Definition 1 *Let X be any domain and $\mathcal{G}$ a class of functions $g : X \to \{0, 1\}$. For a set $A = \{a_1, \ldots, a_k\} \subset X$ denote by $g_A = [g(a_1), \ldots, g(a_k)]$ and $\mathcal{G}_A \equiv$*

URL: `http://www.ratsaby.info/` (Joel Ratsaby).

[1] Present address: P.O. Box 11438, Tel Aviv, ISRAEL. Part of this work was done at and partially supported by the Paul Ivanier Center for Robotics Research and Production Management, Ben-Gurion University of the Negev.

$\{g_A : g \in \mathcal{G}\}$. *The* growth function $\phi_{\mathcal{G}}(k)$ *is defined as*

$$\phi_{\mathcal{G}}(k) \equiv \max_{A \subset X : |A| = k} |\mathcal{G}_A|.$$

The Vapnik-Chervonenkis dimension of $\mathcal{G}$, denoted as $VC(\mathcal{G})$, plays an important role in controlling the rate of increase of $\phi_{\mathcal{G}}(k)$ with respect to k. It is defined as follows:

Definition 2 *The VC-dimension of* $\mathcal{G}$ *is defined as*

$$VC(\mathcal{G}) \equiv \max\{|A| : A \subset X, |\mathcal{G}_A| = 2^{|A|}\}.$$

If the maximum does not exists let $VC(\mathcal{G}) = \infty$.

Consider a finite set $A \subset X$, $|A| = n$ and suppose $\mathrm{VC}(\mathcal{G}) = d < n$. Then $\mathcal{G}_A$ may be viewed as a class $\mathcal{F}$ of binary vectors, i.e., functions $f : [n] \to \{0, 1\}$ where $[n] \equiv \{1, \ldots, n\}$ with $\mathrm{VC}(\mathcal{F}) \leq d$.

Recently there has been interest in learning pattern recognition, e.g., binary classification, by empirical risk minimization under a complexity regularizing constraint, for instance by maximizing the margin that functions have on a sample [Vapnik 1998]. While the class of interest in such problems consists of binary functions the standard analysis approach [Anthony and Bartlett 1999] estimates the growth rate (or more precisely the covering number) of a related class of real-valued functions that have a large sample margin. In this paper we consider a different approach which deals directly with the binary function class. To do that we define a suitable margin parameter for binary functions. We deal with classes $\mathcal{F}$ on $[n]$ (as defined above) and estimate the complexity of subclasses of $\mathcal{F}$ which consist of functions that have a large sample margin. Our approach extends the following result of Vapnik and Chervonenkis [1971], Sauer [1972].

Lemma 1 *Let* $\mathcal{F}$ *be a class of binary functions on* $[n]$ *and suppose* $VC(\mathcal{F}) = d$. *Then*

$$|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k} \equiv \mathbb{S}(d, n).$$

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

Consider the following definition of functional margin which naturally suits binary-valued functions (for definitions of margin of real-valued functions see Vapnik [1998]).

Definition 3 *The* margin $\mu_f(x)$ *of* $f \in \mathcal{F}$ *on an element* $x \in [n]$ *is the*

largest non-negative integer a such that f has a constant value on the interval set $I_a(x) = \{x - a, \ldots, x + a\}$ 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.

2 Motivation

In recent years, in search for better learning algorithms, it has been discovered [see for instance Vapnik 1998, Anthony and Bartlett 1999] that learning classes of real-valued functions that are restricted to have a large margin on a randomly drawn training sample leads to improvements in the generalization error rate. The improved error bounds arise due to tighter bounds on the covering number of such classes (a quantity analogous to the growth-function of binary function classes) which decreases as the sample-margin value increases. For this reason, samples on which the target function (the one to be learned) has a large margin are of considerable worth. In [Ratsaby 2006] estimates of the complexity of such samples as a function of the margin parameter and sample size have been obtained. While the underlying motivation of our work has its roots in statistical learning theory our interest here is in the combinatorial complexity of constrained VC-classes (see also Ratsaby [2007]).

An outline of the paper is as follows: in Section 3 we state and prove several auxiliary lemmas. In Section 4 the main result, Theorem 1, is stated and proved based on an extension of Lemma 1 to a class $\mathcal{H}_N$ of functions constrained to have a margin less than or equal to N. In Section 5, this is used to obtain an estimate (Lemmas 4 and 5) on the cardinality of classes $\mathcal{H}_N(S)$ and $\mathcal{H}_N(S^*)$ of large-margin functions. Being dependent on the margin parameter N, these estimates are analogous to standard bounds on the covering number of real-valued function classes of finite-pseudo-dimension (or fat_γ dimension) under a similar margin constraint [Anthony and Bartlett 1999, Ch. 12].

3 Technical results

Denote by $\mathbb{I}(E)$ the indicator function which equals 1 if the expression E is true and 0 otherwise. We start with an auxiliary lemma:

Lemma 2 *For $0 \le m \le n$, $N \ge 0$, 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*

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

Remark 1 *While our interest is in $[n] = \{1, \ldots, n\}$, 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

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

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

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

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 $m \ge 1$, 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,

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

□

Remark 2 *This expression may alternatively be expressed as*

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

over $m \ge 1$.

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

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

$f(x) = 1$ *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

$$c(k, n-k; m, N) = \binom{n-k}{m-1} \left(w_{m,2N}(k-m+1) + w_{m-1,2N}(k-m-2N) \right). \tag{1}$$

Remark 3 *Note that* $\beta_r^{(N)}(n) = \mathbb{S}(r, n)$ *if* $r < 2N + 1$. *This follows from the standard identity [Graham et al. 1994]*

$$\binom{n}{k} = \sum_{m=1}^{n} \binom{k}{m-1} \binom{n-k}{m-1}.$$

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 valid m-partitions of $[k, n-k]$ are: $\{[0,1][1,1][1,n-4]\}$,$\{[0,1][1,2][1,n-5]\}$, ...,$\{[0,1][1,n-3][1,0]\}$, $\{[0,2][1,1][1,n-5]\}$, $\{[0,2][1,2][1,n-6]\}$, ..., $\{[0,2][1,n-4][1,0]\}$, ..., $\{[0,n-3][1,1][1,0]\}$. 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 $\Pi : F_k \to \mathcal{P}_{n,k}$ 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 $[a_j, b_j]$, $j = 1, \ldots, m$, 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:

$$a_i \leq 2N + 1, \qquad 1 \leq i \leq m. \tag{2}$$

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 $x \in [n]$ such that $f(x) = 1$ implies there does not exist a segment a_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

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

By definition of F it follows that

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

Let us denote by

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

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

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

For $m \geq 1$,

$$G(t_1,t_2) = (t_1^0+t_1^1+\cdots+t_1^{2N+1})(t_2^1+t_2^2+\cdots)^{\mathbb{I}(m\geq 2)} \cdot\left((t_1^1+\cdots+t_1^{2N+1})(t_2^1+t_2^2+\cdots)\right)^{(m-2)_+} \cdot(t_1^1+\cdots+t_1^{2N+1})^{\mathbb{I}(m\geq 2)}(t_2^0+t_2^1+\cdots) \tag{7}$$

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

$$t_1^{m-1}t_2^{m-1}\left(\left(\frac{1-t_1^{2N+1}}{1-t_1}\right)^m + t_1^{2N+1}\left(\frac{1-t_1^{2N+1}}{1-t_1}\right)^{m-1}\right)\left(\frac{1}{1-t_2}\right)^m. \tag{8}$$

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 and denote by $s(n) = \binom{n+m-1}{n}$. So (8) becomes

$$\sum_{\alpha_1,\alpha_2\geq 0} s(\alpha_2)t_2^{\alpha_2+m-1}\left(w_{m,2N}(\alpha_1)t_1^{\alpha_1+m-1} + w_{m-1,2N}(\alpha_1)t_1^{\alpha_1+m+2N}\right). \tag{9}$$

Equating the coefficients of $t_1^{\alpha_1'}t_2^{\alpha_2'}$ in (6) and (9) yields

$$\begin{aligned} c(\alpha_1',\alpha_2';m,N) &= s(\alpha_2'-m+1)\left(w_{m,2N}(\alpha_1'-m+1) + w_{m-1,2N}(\alpha_1'-m-2N)\right) \\ &= \binom{\alpha_2'}{m-1}\left(w_{m,2N}(\alpha_1'-m+1) + w_{m-1,2N}(\alpha_1'-m-2N)\right). \end{aligned}$$

Substituting k for α_1', $n-k$ for α_2', combining (3), (4) and (5) yields the result. □

4 Main result

The next theorem extends Lemma 3 to a class $\mathcal{H}_N$ of VC-dimension no larger than d.

Theorem 1 *Let $n, N \geq 1$, $1 \leq d \leq n$ and $\mathcal{F}$ be a class of binary functions on $[n]$ with VC($\mathcal{F}$) $= d$. Let $\mathcal{H}_N \subseteq \mathcal{F}$ consist of functions h that satisfies*

the margin condition $\mu_h(x) \leq N$ *on any* $x \in [n]$ *such that* $h(x) = 1$*. Then* $|\mathcal{H}_N| \leq \beta_d^{(N)}(n)$*, where* $\beta_d^{(N)}$ *is defined in Lemma 3.*

Remark 4 *As indicated in Remark 3, when* N *is greater than approximately half the VC-dimension* d *the bound* $\beta_d^{(N)}(n)$ *is identical to the bound in Lemma 1 and* N *is ineffective at reducing the size of the class.*

Proof: Let $\mathcal{A}_N$ be the set system corresponding to the function class $\mathcal{H}_N$ which is defined as follows

$$\mathcal{A}_N = \{A_h : h \in \mathcal{H}_N\},\ A_h = \{x \in [n] : h(x) = 1\}.$$

Clearly, $|\mathcal{A}_N| = |\mathcal{H}_N|$. Note that the notion of a bounded margin $\mu_h(x) \leq N$ at x translates to A_h having the property P_N defined as having every subset $E \subseteq A_h$ which consists of consecutive elements $E = \{i, i+1, \ldots, j-1, j\}$ be of cardinality $|E| \leq 2N + 1$. Hence for every element $A \in \mathcal{A}_N$, A satisfies P_N which is denoted by $A \models P_N$. Define $\omega_{\mathcal{A}_N}(k) = \max\{|\{A \cap E : A \in \mathcal{A}_N\}| : E \subseteq [n], |E| = k\}$. The corresponding notion of VC-dimension for a class $\mathcal{A}_N$ of sets is the the so-called *trace number* [Bollobás 1986, p.131] which is defined as $tr(\mathcal{A}_N) = \max\{m : \omega_{\mathcal{A}_N}(m) = 2^m\}$. Clearly, $tr(\mathcal{A}_N) = VC(\mathcal{H}_N) = d$.

The proof proceeds as in the proof of Lemma 1 [Anthony and Bartlett 1999, Theorem 3.6] which is based on the shifting method [see Bollobás 1986, Ch. 17, Theorem 1 & 4] [see also Haussler 1995, Frankl 1987; 1983]. The idea is to transform $\mathcal{A}_N$ into $\mathcal{A}_0$ which is an *ideal* family of sets E, i.e., if $E \in \mathcal{A}_0$ then $S \in \mathcal{A}_0$ for every $S \subset E$, and such that $|\mathcal{A}_N| = |\mathcal{A}_0| \leq \beta_d^{(N)}(n)$.

Start by defining the operator T_x on $\mathcal{A}_N$ which removes an element $x \in [n]$ from every set $A \in \mathcal{A}_N$ provided that this does not duplicate any existing set. It is defined as follows:

$$T_x(\mathcal{A}_N) = \{A \setminus \{x\} : A \in \mathcal{A}_N\} \cup \{A \in \mathcal{A}_N : A \setminus \{x\} \in \mathcal{A}_N\}.$$

Consider now

$$\mathcal{A}_0 = T_1(T_2(\cdots T_n(\mathcal{A}_N)\cdots))$$

and denote the corresponding function class by $\mathcal{H}_0$. Clearly, $|\mathcal{H}_0| = |\mathcal{A}_0|$.

We have $|\mathcal{A}_0| = |\mathcal{A}_N|$ since the only time that the operator T_x changes an element A into a different set $A^* = T_x(A)$ is when A^* does not already exist in the class so no additional element in the new class can be created.

It is also clear that for all $x \in [n]$, $T_x(\mathcal{A}_0) = \mathcal{A}_0$ since for each $E \in \mathcal{A}_0$ there exists a $G \in \mathcal{A}_0$ that differs from it on exactly one element hence it is not possible to remove any element $x \in [n]$ from all sets without creating a duplicate. Applying this repeatedly implies that $\mathcal{A}_0$ is an ideal. Furthermore,

since for all $A \in \mathcal{A}_N$, $A \models P_N$ then removing an element x from A still leaves $A \setminus \{x\} \models P_N$. Hence for all $E \in \mathcal{A}_0$ we have $E \models P_N$.

From Lemma 3 [Bollobás 1986, p.133] we have $\omega_{\mathcal{A}_0}(k) \leq \omega_{\mathcal{A}_N}(k)$, for all $1 \leq k \leq n$. Since $tr(\mathcal{A}_N) = d$ then $tr(\mathcal{A}_0) \leq d$ and since $\mathcal{A}_0$ is an ideal then it follows that for all $E \in \mathcal{A}_0$, $|E| \leq d$. Combined with the fact that for all $E \in \mathcal{A}_0$, $E \models P_N$ then it follows that the corresponding function class $\mathcal{H}_0$ satisfies the following: for all $h \in \mathcal{H}_0$, h has at most d 1's and $\mu_h(x) \leq N$ on every $x \in [n]$ such that $h(x) = 1$. By Lemma 3 above, we therefore have $|\mathcal{H}_0| \leq \beta_d^{(N)}(n)$. From the above, we have $|\mathcal{H}_N| = |\mathcal{A}_N| = |\mathcal{A}_0| = |\mathcal{H}_0|$ and hence $|\mathcal{H}_N| \leq \beta_d^{(N)}(n)$. □

5 Classes of large-margin functions

For any $t \in \mathcal{F}$, h is said to be *consistent* with t on S if $h(x_i) = t(x_i)$, for all $x_i \in S$, $1 \leq i \leq |S|$. Denote by a *positive* sample $S \subseteq [n]$ for $t \in \mathcal{F}$ any set of elements $x \in [n]$ for which $t(x) = 1$. Problems of learning by a positive sample are typical whenever a learner observes an expert act, for instance, as in learning grammatical inference. Let us consider classes of binary-valued functions h consistent with a fixed target $t \in \mathcal{F}$ on a positive sample S such that $\mu_S(h) > N$.

Lemma 4 *Let $n, N, d \geq 1$ and $1 \leq l \leq n$. Let $\mathcal{F}$ be a class of binary-valued functions on $[n]$ with $VC(\mathcal{F}) = d$. For any fixed $t \in \mathcal{F}$, let $S \subseteq [n]$ be a positive sample of cardinality l such that $\mu_S(t) > N$. Consider the subclass $\mathcal{H}_N(S) \subseteq \mathcal{F}$ of all functions h consistent with t on S and having $\mu_S(h) > N$. Then*

$$|\mathcal{H}_N(S)| \leq 1 + e^{-(l+2(N+1))/n}\mathbb{S}(d, n)$$

where $\mathbb{S}(d, n)$ is defined in Lemma 1.

Proof: The condition $\mu_h(x) > N$ for x implies that a function h must take the value 1 over the interval $I_{N+1}(x)$. Consider any $t \in \mathcal{F}$ with S and corresponding $\mathcal{H}_N(S)$ as in the statement of the lemma. Let $R(S) = \{z \in [n] : z \in I_{N+1}(x), x \in S\}$. Since for every $h \in \mathcal{H}_N(S)$, $h(z) = 1$ for all $z \in R(S)$ then the cardinality of the restriction $\mathcal{H}_N(S)_{|R(S)}$ of the class $\mathcal{H}_N(S)$ on the set $R(S)$ is one. Denote by $R^c(S) \equiv [n] \setminus R(S)$ then we have

$$|\mathcal{H}_N(S)| = |\mathcal{H}_N(S)_{|R^c(S)}|.$$

Since $\mathrm{VC}(\mathcal{H}_N(S)) \leq \mathrm{VC}(\mathcal{F}) = d$ then by Lemma 1 it follows that

$$|\mathcal{H}_N(S)_{|R^c(S)}| \leq \mathbb{S}(d, |R^c(S)|). \tag{10}$$

We also have

$$\max\{|R^c(S)| : S \subset [n], |S| = \ell\} = n - \ell - 2(N+1) \tag{11}$$

which is achieved for instance by a set $S' = \{N+2, \ldots, N+l+1\}$ with $R(S') = \{1, \ldots, 2(N+1)+l\}$. Hence for any S as above we have

$$|\mathcal{H}_N(S)| \leq \sum_{k=0}^{d} \binom{n-l-2(N+1)}{k}. \tag{12}$$

Using the standard identity of

$$\binom{k}{m} = \frac{k}{k-m}\binom{k-1}{m}$$

we have for $0 \leq a \leq k$,

$$\binom{k-a}{m} \Big/ \binom{k}{m} = \prod_{i=0}^{a-1} \frac{k-m-i}{k-i} \leq \prod_{i=0}^{a-1} e^{-m/(k-i)} \tag{13}$$

where we used $1 - x \leq \exp(-x)$ which holds for all $x \in \mathbb{R}$. The right side of (13) equals

$$e^{-m\sum_{i=0}^{a-1} 1/(k-i)} \leq e^{-am/k}. \tag{14}$$

Using (13) and (14) the sum on the right side of (12) is bounded from above by

$$\sum_{k=0}^{d} \binom{n}{k} e^{-k(l+2(N+1))/n}.$$

We have

$$\begin{aligned}
\sum_{k=0}^{d} \binom{n}{k} e^{-k(l+2(N+1))/n} &= 1 + \sum_{k=1}^{d} \binom{n}{k} e^{-k(l+2(N+1))/n} \\
&\leq 1 + e^{-(l+2(N+1))/n} \sum_{k=1}^{d} \binom{n}{k} \\
&= (1 - e^{-(l+2(N+1))/n}) + e^{-(l+2(N+1))/n} \sum_{k=0}^{d} \binom{n}{k} \\
&= (1 - e^{-(l+2(N+1))/n}) + e^{-(l+2(N+1))/n} \mathbb{S}(d,n) \\
&\leq 1 + e^{-(l+2(N+1))/n} \mathbb{S}(d,n).
\end{aligned}$$

□

Next we consider an extremal case where S is a *maximal* positive sample S^* on which the target $t \in \mathcal{F}$ has a margin larger than N. The corresponding class

in this case is $\mathcal{H}_N(S^*) \subseteq \mathcal{F}$ which consists of all $h \in \mathcal{F}$ which are consistent with t on S^* and satisfy $\mu_h(x) > N$ if and only if $x \in S^*$. Note that S^* is maximal in the sense that all $x \in [n]$ such that $t(x) = 1$ and $\mu_t(x) > N$ are included in S^*. It thus represents the most informative positive sample for a fixed margin level N and sample size l.

Lemma 5 *Let $n, N, d \geq 1$. Let $\mathcal{F}$ be a class of binary-valued functions on $[n]$ with $VC(\mathcal{F}) = d$. For any fixed $t \in \mathcal{F}$ let $S^* \subseteq [n]$ be a maximal positive sample such that $\mu_{S^*}(t) > N$ and denote by $l = |S^*|$. Consider a subclass $\mathcal{H}_N(S^*) \subseteq \mathcal{F}$ which consists of all functions h consistent with t on S^* and satisfying $\mu_h(x) > N$ iff $x \in S^*$. Then*

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

Proof: The proof follows that of Lemma 4 up to (10) with S^* instead of S. By Theorem 1 we have

$$|\mathcal{H}_N(S^*)_{|R^c(S^*)}| \leq \beta_d^{(N)}(|R^c(S^*)|).$$

With (11) the result follows. □

6 Conclusions

The main result of the paper is a bound on the cardinality of a class of known VC-dimension which consists of binary functions on $[n]$ that have a margin less than or equal to N. This extends a classic result of Sauer (and Vapnik-Chervonenkis) and is subsequently used to obtain estimates on the cardinality of classes of binary-valued functions with a large margin.

References

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.

B. Bollobás. *Combinatorics: Set Systems, Hypergraphs, Families of vectors, and combinatorial probability*. Cambridge University Press, 1986.

P. Frankl. On the trace of finite sets. *Journal of Combinatorial Theory(A)*, 34:41–45, 1983.

P. Frankl. The shifting technique in extremal set theory. In C. Whitehead, editor, *Surveys in Combinatorics*, pages 81–110. Cambridge University Press, 1987.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1994.

D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69:217–232, 1995.

J. Ratsaby. Complexity of hyperconcepts. *Theoretical Computer Science*, 363 (1):2–10, 2006.

J. Ratsaby. Vapnik-Chervonenkis classes of sequences with long runs. *Journal of Discrete Mathematics and Cryptography*, 10(2):205–225, 2007.

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.