Discrete Mathematics 312 (2012) 1766–1775

Contents lists available at SciVerse ScienceDirect

Discrete Mathematics

journal homepage: www.elsevier.com/locate/disc

VC bounds on the cardinality of nearly orthogonal function classes

Lee-Ad Gottlieb [a], Aryeh Kontorovich [b,*], Elchanan Mossel [a,c]

[a] Weizmann Institute of Science, Israel
[b] Ben Gurion University, Israel
[c] U.C. Berkeley, United States

ARTICLE INFO

Article history:
Received 2 September 2010
Received in revised form 24 January 2012
Accepted 26 January 2012
Available online 18 February 2012

Keywords:
VC dimension
Packing number
Orthogonal

ABSTRACT

We bound the number of nearly orthogonal vectors with fixed VC-dimension over $\{-1, 1\}^n$. Our bounds are of interest in machine learning and empirical process theory and improve previous bounds by Haussler. The bounds are based on a simple projection argument and they generalize to other product spaces. Along the way we derive tight bounds on the sum of binomial coefficients in terms of the entropy function.

1. Introduction and statement of results

The capacity or "richness" of a function class F is a key parameter which makes a frequent appearance in statistics, empirical processes, and machine learning theory [6,23,10,21,20,22,17,4]. It is natural to consider the metric space (F, ρ), where $F \subseteq \{-1, 1\}^n$ and

$$\rho(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq y_i\}}. \tag{1}$$

A trivial upper bound on the cardinality of F is 2^n. When F has VC-dimension d, the celebrated Sauer–Shelah–Vapnik–Chervonenkis Lemma [19] bounds the cardinality of F as

$$|F| \leq \sum_{i=0}^{d} \binom{n}{i}. \tag{2}$$

The notion of cardinality can be refined by considering the packing numbers of the metric space (F, ρ). These are denoted by $M(\varepsilon, d)$, and defined to be the maximal cardinality of an ε-separated subset of F; in particular $M(1/n, d) = |F|$. For general ε, the best packing bound for a maximal ε-separated subset of F is due to Haussler [12]. (A discussion of the history of this problem may be found therein.) Haussler's upper bound states that

$$M(\varepsilon, d) \leq e(d + 1) \left(\frac{2e}{\varepsilon}\right)^d. \tag{3}$$

* Corresponding author.
E-mail addresses: lee-ad.gottlieb@weizmann.ac.il (L.-A. Gottlieb), karyeh@cs.bgu.ac.il (A. Kontorovich), mossel@stat.berkeley.edu (E. Mossel).

0012-365X/$ – see front matter
doi:10.1016/j.disc.2012.01.030

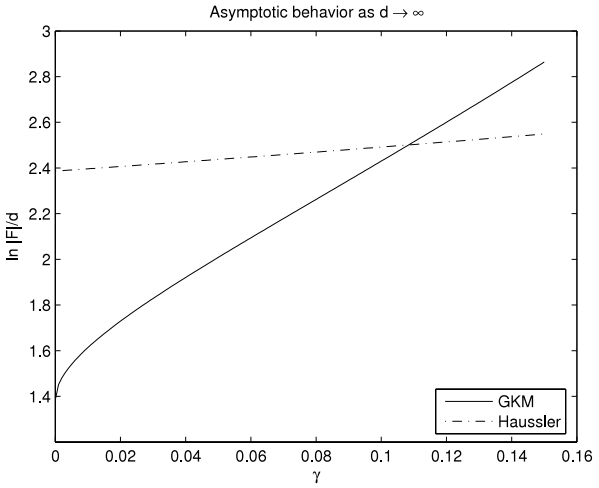

Fig. 1. A comparison of upper bounds.

In this paper, we propose to study the behavior of $M(\varepsilon, d)$ for $\frac{1}{2} - c \leq \varepsilon \leq \frac{1}{2} + c$ (for constant c). As explained below, this corresponds to the case where the vectors of F are close to orthogonal. Our interest in this regime stems from applications in machine learning, where some characterizations and algorithms consider nearly orthogonal or decorrelated function classes [3,7,2]. Our main result is Theorem 1 (Section 3), which sharpens Haussler's estimate of $M(\varepsilon, d)$ as a function of d and $\varepsilon \approx \frac{1}{2}$.

It is convenient to state our results in terms of $\gamma = 1 - 2\varepsilon$ (thus, for $\varepsilon \approx \frac{1}{2}$, we have $\gamma \approx 0$). We will denote Haussler's bound on $|F|$ in (3) by

$$M((1-\gamma)/2, d) \leq \mathrm{DH}(\gamma, d) := e(d+1)\left(\frac{4e}{1-\gamma}\right)^d$$

and our bound in Theorem 1 by

$$M((1-\gamma)/2, d) \leq \mathrm{GKM}(\gamma, d) := 100 \cdot 2^{d\beta(\gamma)},$$

where $\beta : [0, 1] \to [2, \infty)$ is defined in (9).

As $d \to \infty$, our bound asymptotically behaves as

$$\frac{\ln[\mathrm{GKM}(\gamma, d)]}{d} \to (\ln 2)\beta(\gamma)$$

while Haussler's as

$$\frac{\ln[\mathrm{DH}(\gamma, d)]}{d} \to \ln\left(\frac{4e}{1-\gamma}\right).$$

Fig. 1 gives a visual comparison of these bounds, illustrating the significant improvement of our bound over Haussler's for small γ.

Our analysis has the additional advantage of readily extending to k-ary alphabets, while the proof in [12] appears to be strongly tied to the binary case. In Theorem 2 we give what appears to be the first packing bound for alphabets beyond the binary in terms of (a generalized) VC-dimension (but see [1, Lemma 3.3]).

We further wish to understand the relationship between $M(\varepsilon, d)$ and n for fixed ε and d. It is well known [18] that when $\gamma = 1 - 2\varepsilon = O(1/\sqrt{n})$, we have $M(\varepsilon, d) = O(\mathrm{poly}(n))$. Since in many cases of interest [14] the coordinate dimension n may be replaced by its refinement d_{VC}, it is natural to ask whether a poly(n) bound on $M(\varepsilon, d)$ is possible for $\gamma = 1 - 2\varepsilon = O(1/\mathrm{poly}(d))$. We resolve this question in the negative in Theorem 3.

Finally, in Section 6 we give a simple improvement of Haussler's lower bound. Haussler exhibits an infinite family $\left\{F_n \subseteq \{-1, 1\}^n\right\}$ for which $d_{\mathrm{VC}}(F_n) = d$ and

$$M(\varepsilon, d) \geq \left(\frac{1}{2e(\varepsilon + d/n)}\right)^d. \tag{4}$$

He notes that the bounds in (3) and (4) leave "a gap from $1/2e$ to $2e$ for the best universal value of the key constant" and poses the closure of this gap as an "intriguing open problem". The gap has recently been tightened to $[1, 2e]$ by Bshouty et al. [5, Theorem 10], in a rather general and somewhat involved argument. Our lower bound in Theorem 4 achieves the same tightening via a much simpler construction.

2. Definitions and notation

Our basic object is the metric space (F, ρ), with $F \subseteq \{-1, 1\}^n$ and the normalized Hamming distance ρ defined in (1). The inner product

$$\langle x, y\rangle := \frac{1}{n}\sum_{i=1}^{n} x_i y_i, \quad x, y \in F$$

endows F with Euclidean structure. The distance and inner product have a simple relationship:

$$2\rho(x, y) + \langle x, y\rangle = 1. \tag{5}$$

We denote the natural numbers by $\mathbb{N} = \{1, 2, \ldots\}$, and for $n \in \mathbb{N}$, we write $[n] = \{0, 1, \ldots, n-1\}$. For $I = (i_1, i_2, \ldots, i_k) \subseteq [n]$, we denote the projection of F onto I by

$$F|_I = \left\{(x_{i_1}, \ldots, x_{i_k}) : x \in F\right\} \subseteq \{-1, 1\}^k. \tag{6}$$

We say that F *shatters* I if $F|_I = \{-1, 1\}^k$ and define the Vapnik–Chervonenkis dimension of F to be the cardinality of the largest shattered index sequence I:

$$d_{\mathrm{VC}}(F) = \max\left\{|I| : I \subset [n], F|_I = \{-1, 1\}^k\right\}.$$

We define $\gamma = \gamma_{\mathrm{ORT}}(F)$ by

$$\gamma_{\mathrm{ORT}}(F) = \max\{|\langle x, y\rangle| : x \neq y \in F\}. \tag{7}$$

In words, $\gamma_{\mathrm{ORT}}(F)$ is the smallest $\gamma \geq 0$ such that all distinct pairs $x, y \in F$ are "orthogonal to accuracy γ". Whenever (7) holds for some γ, we say that F is γ-orthogonal.

We will use ln to denote the natural logarithm and $\log \equiv \log_2$ and abuse the notation slightly by using H to denote both the binary entropy function (defined in (8)) and the standard entropy function

$$H(Y) = -\sum_{a \in A} P(Y = a) \log P(Y = a)$$

for any random variable Y taking values in the finite set A.

3. Upper estimates on nearly orthogonal sets

3.1. Preliminaries: entropy and β

Recall the binary entropy function, defined as

$$H(x) = -x \log x - (1 - x)\log(1 - x). \tag{8}$$

In the range $[0, 1]$, this function is symmetric about $x = \frac{1}{2}$, where it achieves its maximum value of 1.

Since H is increasing on $[0, \frac{1}{2}]$, it has a well-defined inverse on this domain, which we will denote by $H^{-1} : [0, 1] \to [0, \frac{1}{2}]$. We define the function $\beta : [0, 1] \to [2, \infty)$ by

$$\beta(\gamma) = \frac{1}{H^{-1}[\log(2/(1 + \gamma))]}. \tag{9}$$

Fig. 2 illustrates the behavior of β on $[0, \frac{1}{4}]$.

A sharp bound on $\sum_{i=0}^{d}\binom{n}{i}$ in terms of H is given in Lemma 5.

3.2. Main result

Theorem 1. *Let $F \subseteq \{-1, 1\}^n$ with $1 \leq d = d_{\mathrm{VC}}(F) \leq n/2$ and $\gamma = \gamma_{\mathrm{ORT}}(F)$. Then*

$$|F| \leq 100 \cdot 2^{d\beta(\gamma)}$$

where $\beta : [0, 1] \to [2, \infty)$ is defined in (9).

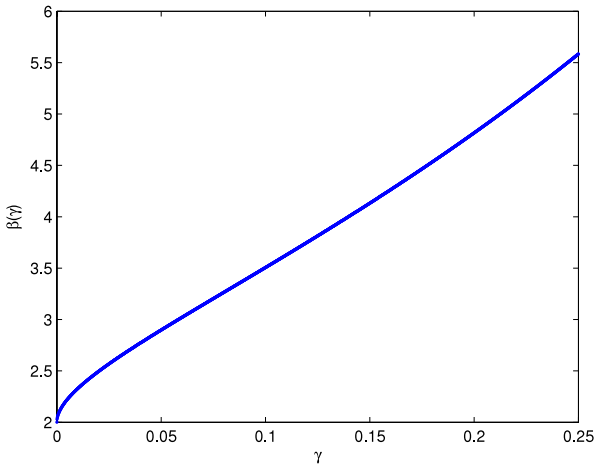

Fig. 2. The function $\beta(\gamma)$.

Proof. Let $r < n$ be unspecified for the moment and choose $I \subset [n]$, $|I| = r$ uniformly at random. Define $\pi = \pi_I$ to be the coordinate projection of F onto I as defined in (6). Let x and y be two uniformly random elements of F, and let A be the event that $\pi(x) = \pi(y)$; thus, $P(A)$ is the probability that x and y are mapped to the same vector. (The probabilities in this proof are with respect to the joint choice of I, x and y.) The latter is upper-bounded by the sum of the probability that x and y are the same vector, and the probability that x and y are distinct vectors but are mapped to the same vector. The first event occurs with probability exactly $|F|^{-1}$. We claim that the second event occurs with probability less than $(\frac{1}{2} + \frac{1}{2}\gamma)^r$. To see this, suppose that the two vectors x, y agree on η fraction of the coordinates. Then $\eta \leq \frac{1}{2} + \frac{1}{2}\gamma$ and the probability that they agree on one random coordinate is exactly η. The probability that they agree on two coordinates is $\eta(n\eta - 1)/(n - 1)$, and so forth. Thus, the probability that they agree on r coordinates is

$$\eta(n\eta - 1)/(n - 1) \cdot \cdots \cdot (n\eta - (r - 1))/(n - (r - 1)) < \eta^r \leq \left(\frac{1}{2} + \frac{1}{2}\gamma\right)^r.$$

By the union bound, we have

$$P(A) < |F|^{-1} + \left(\frac{1}{2} + \frac{1}{2}\gamma\right)^r. \tag{10}$$

As a lower bound on $P(A)$, we claim

$$P(A)^{-1} \leq \sum_{i=0}^{d} \binom{r}{i}. \tag{11}$$

Indeed, if E is any finite set equipped with distribution Q, then the probability of collision (i.e., drawing $e, e' \in E$ independently according to Q and having $e = e'$) is given by $Q(e = e') = \sum_{e \in E} Q(e)^2$. Recall that $\|x\|_2^2 \geq \|x\|_1^2/n$ for all $x \in \mathbb{R}^n$. Thus, viewing Q as a vector in $\mathbb{R}^E$, we have

$$Q(e = e') = \sum_{e \in E} Q(e)^2 = \|Q\|_2^2 \geq \frac{\|Q\|_1^2}{|E|} = \frac{1}{|E|}. \tag{12}$$

Let us denote the event that $\pi(x) = \pi(y)$ conditioned on I by $A \mid I$, and write P_π for the distribution on $F' := F|_I$ induced by π. Then we have

$$\begin{aligned} P(A \mid I) &= \sum_{x' \in F'} P_\pi(x')^2 \\ &\geq |F'|^{-1} \\ &\geq \left(\sum_{i=0}^{d} \binom{r}{i}\right)^{-1}, \end{aligned}$$

where the first inequality is seen by taking $E = F'$ and $Q = P_\pi$ in (12) and the second holds by Sauer's Lemma (2). The claim (11) follows by averaging over all the Is.

Combining (10) and (11) with Lemma 5, we get the key inequality

$$1.02 \cdot 2^{-rH(d/r)} < \frac{1}{|F|} + \left(\frac{1}{2} + \frac{1}{2}\gamma\right)^r, \tag{13}$$

valid for all integer $r \in [2d, n]$. We choose the value

$$r^* = \lceil \beta(\gamma) d \rceil$$

where the function $\beta : [0, 1] \to [2, \infty)$ is defined in (9). It is straightforward to verify from the definition of $\beta(\cdot)$ that for this choice of r^*, we have

$$2^{-r^* H(d/r^*)} \geq \left(\frac{1}{2} + \frac{1}{2}\gamma\right)^{r^*}$$

and therefore

$$0.02 \cdot 2^{-r^*} < |F|^{-1}.$$

Substituting the value of r^* and solving for F, we get

$$\begin{aligned} |F| &\leq 50 \cdot 2^{\lceil \beta(\gamma) d \rceil} \\ &\leq 100 \cdot 2^{\beta(\gamma) d}. \quad \square \end{aligned}$$

4. Generalization to k-ary alphabets

Here we extend our upper bound analysis to k-ary ($k \geq 3$) alphabets. First, we must generalize the notion of orthogonality. Since two vectors x, y drawn uniformly from $[k]^n$ agree in expectation on n/k coordinates, we may define $\gamma_k(x, y)$ by

$$\frac{k}{k-1}\rho(x, y) + \gamma_k(x, y) = 1, \tag{14}$$

where ρ is the normalized Hamming distance defined in (1). Analogously, we define $\gamma^k_{\mathrm{ORT}}(F)$ by

$$\gamma^k_{\mathrm{ORT}}(F) = \max\{|\gamma_k(x, y)| : x \neq y \in F\}. \tag{15}$$

The notion of VC-dimension has various generalizations to k-ary alphabets [11,15–17]. Among these, we consider Pollard's P(seudo)-dimension, Natarajan's G(raph)-dimension, and the GP-dimension; these are defined in Eqs. (13)–(15) of [13], respectively. In the following we continue to write $d_{\mathrm{VC}}(F)$ to denote one of these combinatorial dimensions, without specifying which one we have in mind. This convention is justified by a common generalized Sauer's Lemma shared by these three quantities, due to Haussler and Long [13, Corollary 3]:

$$|F| \leq \sum_{i=0}^{d_{\mathrm{VC}}(F)} \binom{n}{i} k^i. \tag{16}$$

A sharp bound on the rhs of (16) is given in Lemma 6.

Our main result is readily generalized to k-ary alphabets.

Theorem 2. *Let $F \subseteq [k]^n$ with $\frac{6k}{k+1.6} \leq d = d_{\mathrm{VC}}(F) \leq \frac{nk}{k+1.6}$ and $\gamma = \gamma^k_{\mathrm{ORT}}(F)$. Then*

$$|F| \leq 34k^d 2^{d/\delta(\gamma,k)}$$

where $\delta(\gamma, k)$ is the largest $x \in [0, k/(k+1)]$ for which $x \log k + H(x) \leq \log(k/(1 + (k-1)\gamma))$ holds.

Remark. The function $\delta : (0, 1) \times \mathbb{N} \to (0, 1)$ is readily computed numerically.

Proof. Repeating the argument in Theorem 1 (with the generalized Sauer Lemma (16)), we have

$$\left(\sum_{i=0}^{d} \binom{r}{i} k^i\right)^{-1} < |F|^{-1} + \left(\frac{1}{k} + \frac{k-1}{k}\gamma\right)^r.$$

Applying the bound in Lemma 6, we have that for $\frac{6k}{k+1.6} \leq d \leq \frac{rk}{k+1.6}$,

$$1.06 \cdot 2^{-rH(d/r) - d\log k} < |F|^{-1} + \left(\frac{1}{k} + \frac{k-1}{k}\gamma\right)^r.$$

Now we seek the minimum integer $r \in [\frac{k+1.6}{k} d, n]$ that ensures

$$d \log k + rH(d/r) \leq r \log(k/(1 + (k-1)\gamma)).$$

To this end, we consider the following inequality in x:

$$x \log k + H(x) \leq \log(k/(1 + (k-1)\gamma)). \tag{17}$$

Note that the inequality (17) is satisfied at $x = 0$ and define $x^* \equiv \delta(\gamma, k)$ to be the largest $x \in [0, k/(k + 1.6)]$ satisfying it (the proof of Lemma 6 shows that the lhs of (17) is monotonically increasing in this range). Taking $r^* = \lceil d/x^* \rceil$, we have

$$0.06 \cdot 2^{-r^* H(d/r^*) - d \log k} < |F|^{-1},$$

which rearranges to

$$\begin{aligned} |F| &< 17 \cdot 2^{r^* H(d/r^*) + d \log k} \\ &\leq 34 k^d 2^{d/\delta(\gamma,k)}, \end{aligned}$$

as claimed. □

5. Polynomial upper bounds for small γ

The bounds of Haussler (3) and Theorem 1 obscure the dependence of $|F|$ on its coordinate dimension n. It is well known that when $\gamma_{\mathrm{ORT}}(F) = O(1/\sqrt{n})$, we have $|F| = O(\mathrm{poly}(n))$. (In the degenerate case $\gamma_{\mathrm{ORT}}(F) = 0$, linear algebra gives $|F| \leq n + 1$.)

Roth and Seroussi [18] developed a powerful technique for bounding $|F|$ in terms of n and γ. Let $0 < \rho_{\min} \leq \rho_{\max}$ be such that

$$\rho_{\min} \leq n\rho(x, y) \leq \rho_{\max}$$

for all $x, y \in F$. Then [18, Proposition 4.1] shows that

$$1 - |F|^{-1} \leq \left(1 - \frac{1}{n}\right) \left(\frac{\rho_a}{\rho_g}\right)^2$$

where $\rho_a = \frac{1}{2}(\rho_{\min} + \rho_{\max})$ and $\rho_g = \sqrt{\rho_{\min}\rho_{\max}}$. Recalling the relation in (5), we put

$$\rho_{\max} := \frac{n}{2}(1 + \gamma), \qquad \rho_{\min} := \frac{n}{2}(1 - \gamma),$$

which implies $\rho_a = \frac{n}{2}$, $\rho_g = \frac{n}{2}\sqrt{1 - \gamma^2}$, and the following bound on $|F|$:

$$1 - |F|^{-1} \leq \left(1 - \frac{1}{n}\right) \frac{1}{1 - \gamma^2}.$$

Note that when $\gamma^2 \geq n^{-1}$, the right-hand side is at least 1 and the bound is rendered vacuous; thus the nontrivial regime is $\gamma^2 < n^{-1}$. In particular, for $c > 1$, we have

$$\gamma_{\mathrm{ORT}}(F) \leq \frac{1}{\sqrt{cn}} \Longrightarrow |F| \leq \frac{cn - 1}{c - 1}. \tag{18}$$

Since in many situations the VC-dimension d_{VC} is a refinement of the coordinate dimension n, it is natural to ask if a bound similar to (18) holds with d_{VC} replaced by n in both of its occurrences. We resolve this question strongly in the negative.

Theorem 3. *Let $a > 0$ be some constant. Then there infinitely many $n \in \mathbb{N}$ for which there is an $F \subseteq \{-1, 1\}^n$ such that*

(a) $\gamma = d^{-a}$

(b) $|F| = \left\lfloor \exp\left(cn^{\frac{1}{2a+1}}\right) \right\rfloor$

where $\gamma = \gamma_{\mathrm{ORT}}(F)$, $d = d_{\mathrm{VC}}(F)$ and c is an absolute constant.

Proof. Let F be an $m \times n$ matrix whose entries are independent symmetric Bernoulli $\{-1, 1\}$ random variables; we shall identify the rows of F with the functions in F. Then for $f, g \in F$, we have

$$\mathbf{E}\langle f, g\rangle = 0$$

and by Chernoff's bound

$$\mathbf{P}\{|\langle f, g\rangle| > \gamma\} \leq 2\exp(-n\gamma^2/2)$$

for all $n \in \mathbb{N}$ and $\gamma > 0$. The union bound implies that for n large enough there exists an $F \subseteq \{-1, 1\}^n$ with $\gamma_{\mathrm{ORT}}(F) \leq \gamma$ and

$$|F| = \left\lfloor \exp(n\gamma^2/4) \right\rfloor.$$

The claim follows from the relation

$$d = d_{\mathrm{VC}}(F) \leq \log_2 |F| \leq n\gamma^2/4\ln 2$$

and our choice of

$$\gamma = d^{-a}. \quad \square$$

An alternative estimate may be obtained via the Gilbert–Varshamov bound [9,24].

6. A lower bound on the universal constant c_0

Haussler's upper (3) and lower (4) bounds imply the existence of a universal c_0 for which the packing number $M(\varepsilon, d)$ grows as $\Theta((c_0/\varepsilon)^d)$ in ε for constant d. More precisely,

(i) $M(\varepsilon, d) = O(d(c_0/\varepsilon)^d)$ for all n, $F \subseteq \{-1, 1\}^n$ with $d_{\mathrm{VC}}(F) = d$;
(ii) $M(\varepsilon_n, d) = \Omega((c_0/\varepsilon_n)^{d_n})$ for some infinite family $(\varepsilon_n, d_n, F_n \subseteq \{-1, 1\}^n)$ with $d_{\mathrm{VC}}(F_n) = d_n$.

The bounds in (3), (4) peg c_0 at $1/2e \leq c_0 \leq 2e$. An improved lower bound of $c_0 \geq 1$ may be obtained essentially "for free" (cf. [5, Theorem 10]).

Theorem 4. *There exists an infinite family $(\varepsilon_n, d_n, F_n \subseteq \{-1, 1\}^{2^n})$ for which*

(a) $d_{\mathrm{VC}}(F_n) = d_n$,
(b) $M(\varepsilon_n, d) = (1/\varepsilon_n)^{d_n}$.

Proof. For $n = 2^i$, $i = 0, 1, 2, \ldots$, put $\varepsilon_n = \frac{1}{2}$, $d_n = n$, and $F_n \subset \{-1, 1\}^n$ to be the rows of H_n, the Hadamard matrix of order n. The latter may be defined recursively via

$$H_1 = [1]$$

and

$$H_{2n} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}.$$

It is well known (and elementary to verify) that $d_{\mathrm{VC}}(F_n) = n$ and that $\gamma_{\mathrm{ORT}}(F_n) = 0$. Thus F_n is a $\frac{1}{2}$-separated set of size 2^n. $\square$

An immediate consequence of Theorems 3 and 4 is that the condition $\gamma^2_{\mathrm{ORT}}(F) \leq 1/d_{\mathrm{VC}}(F)$ does not imply $|F| \leq \mathrm{poly}(n, d)$.

7. Technical lemmata

Our main result in Theorem 1 requires a sharp estimate on the sum of the binomial coefficients. It is well known [8] that for $d \leq \frac{n}{2}$, $\sum_{i=0}^{d} \binom{n}{i} \leq 2^{nH(d/n)}$, but we need to obtain a slightly tighter bound.

Lemma 5. *For $1 \leq d \leq \frac{n}{2}$, we have*

$$\sum_{i=0}^{d} \binom{n}{i} < \delta \cdot 2^{nH(d/n)},$$

where $\delta = 0.98$.

Remark. The bound δ can be further tightened, at the expense of a more complicated proof. Note, however, that when $d = n/2$ the summation is equal to $\frac{1}{2}2^{nH(d/n)}$, so δ cannot be taken as a constant better than $\frac{1}{2}$.

Proof. Recall Stirling's approximation $i! = \sqrt{2\pi i}\left(\frac{i}{e}\right)^i e^{\lambda_i}$ where $\frac{1}{12i+1} < \lambda_i < \frac{1}{12i}$. Also note that for $0 \leq i \leq n$,

$$\frac{1}{12n} - \frac{1}{12(n-i)+1} - \frac{1}{12i+1} = \frac{-144n^2 + 122ni - 144i^2 - 12n}{(12n)(12n - 12i + 1)(12i + 1)} \leq 0.$$

Thus,

$$\begin{aligned}\binom{n}{i} &= \frac{n!}{i!(n-i)!}\\ &\le e^{\frac{1}{12n}-\frac{1}{12(n-i)+1}-\frac{1}{12i+1}}\cdot\sqrt{\frac{n}{2\pi i(n-i)}}\cdot\frac{n^n}{i^i(n-i)^{n-i}}\\ &< \frac{1}{\sqrt{2\pi i(1-i/n)}}\cdot(i/n)^{-i}(1-i/n)^{-(n-i)}\\ &= \frac{1}{\sqrt{2\pi i(1-i/n)}}\cdot 2^{nH(i/n)}.\end{aligned}$$

We first prove Lemma 5 for small values of d, in particular $1 \le d < n/4$. Note that for $i \le d < n/4$ we have

$$\binom{n}{i-1} = \frac{i}{n-i+1}\binom{n}{i} < \frac{1}{3}\binom{n}{i},$$

and therefore

$$\begin{aligned}\sum_{i=0}^{d}\binom{n}{i} &< 1.5\binom{n}{d}\\ &< \frac{1.5}{\sqrt{2\pi d(1-d/n)}}\cdot 2^{nH(d/n)}\\ &< 0.7\cdot 2^{nH(d/n)}.\end{aligned}$$

We now turn to the case of large d, that is $\frac{n}{4} \le d \le \frac{n}{2}$. If $\sum_{i=0}^{d}\binom{n}{i} < 0.5\cdot 2^{nH(d/n)}$, then Lemma 5 immediately holds, so we may assume that $Z := \sum_{i=0}^{d}\binom{n}{i} \ge 0.5\cdot 2^{nH(d/n)}$. We will show that in this case, much of the weight of the sum is distributed among $\Omega(\sqrt{n})$ coefficients. We will use this fact in conjunction with a standard entropy argument (e.g., [8]) to obtain the desired result.

Now, we have for all $i \le d$ (when $\frac{n}{4} \le d \le \frac{n}{2}$),

$$\binom{n}{i} \le \binom{n}{d} < \frac{1}{\sqrt{2\pi d(1-d/n)}}\cdot 2^{nH(d/n)} < \frac{2}{\sqrt{\pi n}}2^{nH(d/n)} \le \frac{4Z}{\sqrt{\pi n}}.$$

Consider the random vector $(X_1, \ldots, X_n)$ uniformly distributed in $\{x : \{0,1\}^n : \sum_i x_i \le d\}$. Then for all $0 \le r \le d$ we have

$$P\left[\sum_{i=1}^{n} X_i = r\right] = Z^{-1}\binom{n}{r} \le \frac{4}{\sqrt{\pi n}},$$

and therefore

$$P\left[\sum_{i=1}^{n} X_i \ge d - \frac{\sqrt{\pi n}}{8} + 1\right] \le \frac{\sqrt{\pi n}}{8}\frac{4}{\sqrt{\pi n}} \le \frac{1}{2},$$

which implies

$$\begin{aligned}\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] &\le dP\left[\sum_{i=1}^{n} X_i \ge d - \frac{\sqrt{\pi n}}{8} + 1\right] + \left(d - \frac{\sqrt{\pi n}}{8}\right)\left(1 - P\left[\sum_{i=1}^{n} X_i \ge d - \frac{\sqrt{\pi n}}{8} + 1\right]\right)\\ &\le \frac{1}{2}\left(d + d - \frac{\sqrt{\pi n}}{8}\right) = d - \frac{\sqrt{\pi n}}{16}.\end{aligned}$$

Hence, we obtain

$$\begin{aligned}H(X_1, \ldots, X_n) &\le nH(X_i) = nH(\mathbf{E}[X_i])\\ &< nH\left(\frac{d}{n} - \frac{\sqrt{\pi}}{16\sqrt{n}}\right)\\ &= nH\left(\frac{d}{n}\right) - n\left(H\left(\frac{d}{n}\right) - H\left(\frac{d}{n} - \frac{\sqrt{\pi}}{16\sqrt{n}}\right)\right)\\ &< nH\left(\frac{d}{n}\right) - n\left(H\left(\frac{1}{2}\right) - H\left(\frac{1}{2} - \frac{\sqrt{\pi}}{16\sqrt{n}}\right)\right),\end{aligned}$$

where the second inequality uses the monotonicity of the binary entropy function H at $[0, \frac{1}{2}]$, and the third uses the concavity of H. Noting that the Taylor series expansion of $H(x)$ around $\frac{1}{2}$ is equal to $1 - \frac{1}{2\ln 2}\sum_{j=1}^{\infty}\frac{(1-2x)^{2j}}{j(2j-1)} < 1 - \frac{(1-2x)^2}{2\ln 2}$, we have that

$$H\left(\frac{1}{2}\right) - H\left(\frac{1}{2} - \frac{\sqrt{\pi}}{16\sqrt{n}}\right) > \frac{1}{2\ln 2}\frac{\pi}{64n},$$

from which we conclude that

$$H(X_1, \ldots, X_n) < nH(d/n) - \frac{\pi}{128\ln 2}.$$

Hence, we have

$$\begin{aligned}\sum_{i=0}^{d}\binom{n}{i} &= 2^{H(X_1,\ldots,X_n)}\\ &< 2^{-\frac{\pi}{128\ln 2}}2^{nH(d/n)}\\ &< 0.98\cdot 2^{nH(d/n)},\end{aligned}$$

where the first identity holds because $H(Y) = \log|\mathrm{supp}(Y)|$ when Y is uniformly distributed on its support. This completes the proof. □

Our extension to k-ary alphabets requires the corresponding analogue of Lemma 5.

Lemma 6. *For $2 \le d \le \frac{k}{k+1.6}\cdot n$ and $n \ge 6$, we have*

$$\sum_{i=0}^{d}\binom{n}{i}k^i < 0.94\cdot 2^{nH(d/n)+d\log k}.$$

Proof. First note that the derivative of $f(i) = 2^{nH(i/n)+i\log k}$ is $f'(i) = f(i)[\ln(\frac{n}{i} - 1) + \ln k]di$, so $f(i)$ attains its maximum over the range $0 \le i \le n$ at $i = \frac{k}{k+1}\cdot n$. Further note that since $i \le d \le \frac{k}{k+1.6}\cdot n < \frac{k}{k+e^{1/\sqrt{\lfloor n/2\rfloor+1}}}\cdot n$, we have that $\ln(\frac{n}{i} - 1) + \ln k > \frac{1}{\sqrt{\lfloor n/2\rfloor+1}}$.

We break up the analysis into two cases. When $d \le \frac{n}{2}$ we have

$$\begin{aligned}\sum_{i=0}^{d}\binom{n}{i}k^i &< \binom{n}{d}2k^d\\ &< \frac{2k^d}{\sqrt{\pi d}}2^{nH(d/n)}\\ &= \frac{2}{\sqrt{\pi d}}f(d)\\ &< 0.8\cdot f(d).\end{aligned}$$

When $d > \frac{n}{2}$ we have

$$\begin{aligned}\sum_{i=0}^{d}\binom{n}{i}k^i &= \sum_{i=0}^{\lfloor n/2\rfloor}\binom{n}{i}k^i + \sum_{i=\lfloor n/2\rfloor+1}^{d}\binom{n}{n-i}k^i\\ &< \frac{2f(\lfloor n/2\rfloor)}{\sqrt{\pi\lfloor n/2\rfloor}} + \sum_{i=\lfloor n/2\rfloor+1}^{d}\frac{2^{nH(1-i/n)}k^i}{\sqrt{2\pi i(i/n)}}\\ &< \frac{2f(\lfloor n/2\rfloor)}{\sqrt{\pi\lfloor n/2\rfloor}} + \frac{1}{\sqrt{\pi(\lfloor n/2\rfloor+1)}}\sum_{i=\lfloor n/2\rfloor+1}^{d}2^{nH(i/n)+i\log k}\\ &< \frac{2f(\lfloor n/2\rfloor)}{\sqrt{\pi\lfloor n/2\rfloor}} + \frac{f(d)}{\sqrt{\pi(\lfloor n/2\rfloor+1)}} + \frac{1}{\sqrt{\pi(\lfloor n/2\rfloor+1)}}\sum_{i=\lfloor n/2\rfloor+1}^{d-1}f(i)\\ &< \frac{2f(\lfloor n/2\rfloor)}{\sqrt{\pi\lfloor n/2\rfloor}} + \frac{f(d)}{\sqrt{\pi(\lfloor n/2\rfloor+1)}} + \frac{1}{\sqrt{\pi(\lfloor n/2\rfloor+1)}}\int_{\lfloor n/2\rfloor+1}^{d}f(i)\end{aligned}$$

$$\leq \frac{2f(\lfloor n/2\rfloor)}{\sqrt{\pi\lfloor n/2\rfloor}} + \frac{f(d)}{\sqrt{\pi(\lfloor n/2\rfloor+1)}} + \frac{1}{\sqrt{\pi}}\int_{\lfloor n/2\rfloor+1}^{d} f(i)\left[\ln\left(\frac{n}{i}-1\right)+\ln k\right]di$$
$$= \frac{2f(\lfloor n/2\rfloor)}{\sqrt{\pi\lfloor n/2\rfloor}} + \frac{f(d)}{\sqrt{\pi(\lfloor n/2\rfloor+1)}} + \frac{f(d)}{\sqrt{\pi}} - \frac{f(\lfloor n/2\rfloor+1)}{\sqrt{\pi}}$$
$$< 0.94f(d). \quad \square$$

Acknowledgments

We thank Noga Alon for helpful comments and references about Theorem 3 and the anonymous referees for weeding out a number of inaccuracies.

The first author's work was supported in part by The Israel Science Foundation (grant #452/08), and by a Minerva grant.

The third author was partially supported by DMS 0548249 (CAREER) award, by ISF grant 1300/08, by a Minerva Foundation grant and by an ERC Marie Curie Grant 2008 239317.

References

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, J. ACM 44 (4) (1997) 615–631.
[2] Dana Angluin, David Eisenstat, Leonid Kontorovich, Lev Reyzin, Lower bounds on learning random structures with statistical queries, in: ALT, 2010, pp. 194–208.
[3] Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, Steven Rudich, Weakly learning dnf and characterizing statistical query learning using Fourier analysis, in: STOC, 1994, pp. 253–262.
[4] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, Manfred K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, J. Assoc. Comput. Mach. 36 (4) (1989) 929–965.
[5] Nader H. Bshouty, Yi Li, Philip M. Long, Using the doubling dimension to analyze the generalization of learning algorithms, J. Comput. System Sci. 75 (6) (2009) 323–335.
[6] R.M. Dudley, Central limit theorems for empirical measures, Ann. Probab. 6 (6) (1979) 899–929. 1978.
[7] Vitaly Feldman, A complete characterization of statistical query learning with applications to evolvability, in: Symposium on Foundations of Computer Science, FOCS, 2009.
[8] J. Flum, M. Grohe, Parameterized Complexity Theory, in: Texts in Theoretical Computer Science. An EATCS Series, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
[9] E.N. Gilbert, A comparison of signalling alphabets, Bell Syst. Tech. J. 31 (1952) 504–522.
[10] Evarist Giné, Joel Zinn, Some limit theorems for empirical processes, Ann. Probab. 12 (4) (1984) 929–998. With discussion.
[11] David Haussler, Generalizing the PAC model: sample size bounds from metric dimension-based uniform convergence results, in: 30th Annual Symposium on Foundations of Computer Science, 1989, pp. 40–45.
[12] David Haussler, Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension, J. Combin. Theory Ser. A 69 (2) (1995) 217–232.
[13] David Haussler, Philip M. Long, A generalization of Sauer's lemma, J. Combin. Theory Ser. A 71 (2) (1995) 219–240.
[14] S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, Invent. Math. 152 (1) (2003) 37–55.
[15] B.K. Natarajan, On learning sets and functions, Mach. Learn. 4 (1989) 67–97.
[16] David Pollard, Convergence of Stochastic Processes, Springer-Verlag, 1984.
[17] David Pollard, Empirical Processes: Theory and Applications, in: NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 2, Institute of Mathematical Statistics, Hayward, CA, 1990.
[18] Ron M. Roth, Gadiel Seroussi, Bounds for binary codes with narrow distance distributions, IEEE Trans. Inform. Theory 53 (8) (2007) 2760–2768.
[19] Norbert Sauer, On the density of families of sets, J. Combin. Theory Ser. A 13 (1972) 145–147.
[20] Michel Talagrand, Donsker classes and random geometry, Ann. Probab. 15 (4) (1987) 1327–1338.
[21] Michel Talagrand, The Glivenko–Cantelli problem, Ann. Probab. 15 (3) (1987) 837–870.
[22] Michel Talagrand, Donsker classes of sets, Probab. Theory Related Fields 78 (2) (1988) 169–191.
[23] Vladimir N. Vapnik, Estimation of Dependences Based on Empirical Data, Springer-Verlag, 1982.
[24] R.R. Varshamov, Estimate of the number of signals in error correcting codes, Dokl. Akad. Nauk SSSR 117 (1957) 739–741.