ELSEVIER

Discrete Applied Mathematics 86 (1998) 27–35

DISCRETE APPLIED MATHEMATICS

A graph-theoretic generalization of the Sauer–Shelah lemma

Nicolò Cesa-Bianchi [a,*], David Haussler [b]

[a] *DSI, University of Milan, Via Comelico, 39, 20135 Milan, Italy.*
[b] *Department of Computer Science, University of California, Santa Cruz, CA 95064, USA.*

Received 17 December 1996; received in revised form 16 April 1997; accepted 23 October 1997

Abstract

We show a natural graph-theoretic generalization of the Sauer–Shelah lemma. This result is applied to bound the ℓ_∞ and L_1 packing numbers of classes of functions whose range is an arbitrary, totally bounded metric space.

Keywords: Vapnik–Chervonenkis dimension; Packing number; Metric space

1. Definitions and statement of the main result

Let $|S|$ denote the cardinality of an arbitrary set S. For any $n \geqslant 1$, the nth power of an undirected and irreflexive graph $G = \langle V, E\rangle$ is the graph $G^n = \langle V^n, E^n\rangle$, where V^n is the n-fold product of V and $\{(v_1,\ldots,v_n),(w_1,\ldots,w_n)\} \in E^n$ if and only if $\{v_i, w_i\} \in E$ for at least one $1 \leqslant i \leqslant n$. For all $A = \{i_1,\ldots,i_\ell\} \subseteq \{1,\ldots,n\}$ and $F \subseteq V^n$, the *projection of F onto A* is $F_{|A} = \{(v_{i_1},\ldots,v_{i_\ell}) : (v_1,\ldots,v_n) \in F\}$. A set $C \subseteq V^n$ is a *cube in* G^n if $C = \{v_1, w_1\} \times \cdots \times \{v_n, w_n\}$, where $\{v_i, w_i\} \in E$, $i = 1,\ldots,n$. We say that $\langle A, C\rangle$ is a *d-dimensional projected cube* (*d-P-cube*) of a set $F \subseteq V^n$ if $A \subseteq \{1,\ldots,n\}$, $|A| = d > 0$, and $C \subseteq F_{|A}$ is a cube in G^d. Recall that a set of vertices in a graph is a *clique* if any two of them are connected by an edge. Finally, for an undirected, irreflexive graph G, let $h(G,n,d)$ be the smallest nonnegative integer h such that every clique F in G^n with $|F| > h$ contains a $(d+1)$-P-cube.

Theorem 1.1. *For any undirected, and irreflexive graph $G = \langle V, E\rangle$ and any $n > d \geqslant 0$,*

$$h(G,n,d) < 2(2n|E|)^{\left\lceil \log_2 \sum_{i=0}^{d} \binom{n}{i} |E|^i \right\rceil}.$$

This result, which is proven in Section 3, goes toward solving an open problem stated in [9].

* Corresponding author. E-mail: cesabian@dsi.unimi.it

0166-218X/98/$19.00
PII S0166-218X(98)00012-2

2. Related results and a corollary

The problem of calculating the largest size $N(G,n)$ of a clique in the nth power of a graph G was first proposed, in an information-theoretic context, by Shannon [14]. In Shannon's original formulation, one wants to calculate the limit $\lim_{n\to\infty} n^{-1}\log N(G,n)$ for a given (arbitrary) graph G (see [6] for a survey in this area.) Our motivation is different from Shannon's. We are interested in obtaining bounds on packing numbers for classes of functions that take values in a metric space, like the bounds for packing numbers of classes of real-valued functions given in [1, 5, 7, 12]. This leads to the alternate question studied here: what is the size of the largest clique in G^n that does not contain a $(d+1)$-dimensional projected cube. Bounds on this can be obtained directly from Theorem 1.1. As can be seen, these bounds grow subexponentially in n, in contrast to the size $N(G,n)$ of the largest (unrestricted) clique.

Special cases of Theorem 1.1, albeit sometimes with better bounds than those given here, have been obtained before for particular graphs G. Let $G=\langle V,E\rangle$ be the complete graph on $V=\{v_1,v_2\}$. Then it can be shown that $h(G,n,d)=\sum_{i=0}^{d}\binom{n}{i}$. To see this, note that every set $F\subseteq\{v_1,v_2\}^n$ is a clique. So, in this case, $h(G,n,d)\leqslant\sum_{i=0}^{d}\binom{n}{i}$ reduces to the statement that for every subset $F\subseteq\{v_1,v_2\}^n$ with $|F|>\sum_{i=0}^{d}\binom{n}{i}$ there is a set $A\subseteq\{1,\dots,n\}$ with $|A|=d+1$ such that $F_{|A}=\{v_1,v_2\}^{d+1}$, which is the Sauer–Shelah lemma [13, 15] (independently proven, even if in a slightly weaker form, also by Vapnik and Chervonenkis [16].) The lower bound $h(G,n,d)\geqslant\sum_{i=0}^{d}\binom{n}{i}$ follows from an easy and well-known construction, wherein F is taken to be all elements of $\{v_1,v_2\}^n$ with at most d occurrences of v_1.

Now let G be the complete graph with $r\geqslant 2$ vertices. Then

$$\sum_{i=0}^{d}\binom{n}{i}(r-1)^i\leqslant h(G,n,d)<\sum_{i=0}^{d}\binom{n}{i}\binom{r}{2}^i. \tag{1}$$

This generalization of the Sauer–Shelah lemma was shown in [9].

For $r\geqslant 2$ let $G=\langle V,E\rangle$ where $V=\{v_1,\dots,v_r\}$ and, for each pair $1\leqslant i,j\leqslant r$, $\{v_i,v_j\}\in E$ if and only if $|i-j|>1$. The bound

$$h(G,n,d)<2(nr^2)^{\left\lceil\log_2\sum_{i=1}^{d}\binom{n}{i}r^i\right\rceil}$$

was shown, using a different terminology, in [1, Lemma 3.2].

Finally, for any $r\geqslant 2$ and $n>d\geqslant 0$, let $\boldsymbol{h}(r,n,d)$ be the maximum of $h(G,n,d)$ over all graphs G with r vertices. Using the lower bound in (1) above, and the facts that for $n\geqslant d\geqslant 1$, $(n/d)^d\leqslant\binom{n}{d}\leqslant\sum_{i=0}^{d}\binom{n}{i}\leqslant(en/d)^d$ and $\binom{r}{2}\leqslant r^2/2$, we have the following corollary of Theorem 1.1.

Corollary 2.1. *For all $r\geqslant 2$ and all $n>d\geqslant 1$.*

$$\left(\frac{n(r-1)}{d}\right)^d\leqslant\boldsymbol{h}(r,n,d)<2(nr^2)^{\lceil d\log_2(enr^2/2d)\rceil}.$$

Hence for fixed r and d, the function $\boldsymbol{h}(n) = \boldsymbol{h}(r,n,d)$ is $\Omega(n^d)$ and $O(n^{c \log n})$ for some positive constant c. We conjecture that the lower bound is the more accurate approximation. However, we presently know very little about this. It is still open whether or not $\boldsymbol{h}(n)$ is in fact polynomial in n.

3. Proof of Theorem 1.1

The proof is based on an adaptation of [1, Lemma 3.2]. Fix any undirected and irreflexive graph $G = \langle V, E\rangle$. For $|E| = 0$ or $d = 0$ the theorem is easily verified. Hence assume $|E| > 0$ and $d > 0$. For all integers $h \geqslant 2$ and $n \geqslant 1$, let $t(h,n)$ denote the maximum integer t such that every clique F in G^n with $|F| = h$ contains at least t distinct P-cubes (P-cubes of any dimension $d > 0$ are allowed.) If for some h and n no such an F exists, then $t(h,n)$ is infinite.

Note that for $1 \leqslant |A| \leqslant d$ the number of P-cubes $\langle A, C\rangle$ in F is at most $\sum_{i=1}^{d} \binom{n}{i}|E|^i$, and hence strictly less than $y \stackrel{\text{def}}{=} \sum_{i=0}^{d} \binom{n}{i}|E|^i$. Thus, if $t(h,n) \geqslant y$ for some h, then every clique F in G^n of size h has a $(d+1)$-P-cube. Hence $h(G,n,d) < h$.

Let $k = |E|$. We now show that $t(H(n,k,d),n) \geqslant y$ for all $n > d \geqslant 1$, where

$$H(n,k,d) \stackrel{\text{def}}{=} 2(2nk)^{\lceil \log_2 \sum_{i=0}^{d} \binom{n}{i} k^i \rceil}.$$

We will use the following properties of the function t:

$$t(2,m) = 1 \qquad \text{for all } m \geqslant 1, \qquad (P-1)$$

$$t(h,1) \geqslant \binom{h}{2} \qquad \text{for all } h \geqslant 2, \qquad (P-2)$$

$$t(2m\cdot(2nk),n) \geqslant 2 \cdot t(2m, n-1) \qquad \text{for all } n \geqslant 2 \text{ and all } m,k \geqslant 1. \qquad (P-3)$$

Property (P-1) is readily verified. To show (P-2), fix an arbitrary $h \geqslant 2$ and assume, without loss of generality, there exists a clique F in G with $|F| = h$. Fix any $\{f,g\} \subseteq F$. Then $\{f,g\} \in E$, implying that $\langle \{1\}, \{f,g\}\rangle$ is a P-cube in G. As this holds for each choice of $\{f,g\} \subseteq F$, there are at least $\binom{h}{2}$ P-cubes in G and we conclude $t(h,1) \geqslant \binom{h}{2}$.

To show (P-3) assume, again without loss of generality, there exists a clique F in G^n with $|F| = 2m\cdot(2nk)$. Split F arbitrarily into $2m\cdot nk$ unordered pairs. For each pair $\{\boldsymbol{v},\boldsymbol{w}\}$ pick a coordinate i such that $\{v_i, w_i\} \in E$. Then, the same coordinate i is picked for at least $2m\cdot k$ pairs, and for at least $2m$ of these pairs the set $\{v_i, w_i\}$ is the same for this fixed i. But then F contains two subsets F' and F'', with $|F'| = |F''| = 2m$, such that for each $\boldsymbol{f}' \in F'$, $f'_i = v_i$, and for each $\boldsymbol{f}'' \in F''$, $f''_i = w_i$. Let $T = \{1,\ldots,n\} \setminus i$. As G is irreflexive, $F'_{|T}$ and $F''_{|T}$ are both cliques in G^{n-1}. Hence, by definition of the function t, both F' and F'' contain at least $t(2m, n-1)$ P-cubes. Also, if for some $A \subseteq T$, F' and F'' have the same P-cube $\langle A, C\rangle$, then F also contains the P-cube $\langle A \cup \{i\}, C \times \{v_i, w_i\}\rangle$. This implies that $t(2m\cdot(2nk),n) \geqslant 2 \cdot t(2m, n-1)$, concluding the proof of (P-3).

The proof of the theorem is completed by a simple case analysis. Let $r = \lceil \log_2 y \rceil$ (recall that $y = \sum_{i=0}^{d} \binom{n}{i} k^i$.)

Case 1: $n > r$. Let $h = 2(2nk)(2(n-1)k)\cdots(2(n-r+1)k)$. By applying (P-3) r times and then using (P-1), we find that $t(h,n) \geqslant 2^r \geqslant y$. As $2(2nk)^r \geqslant h$, and since t is clearly monotone in its first argument, we get $t\,(2(2nk)^r, n) \geqslant t(h,n) \geqslant y$.

Case 2: $n \leqslant r$. Let $h = 2(2nk)^{r-n+1}(2(n-1)k)\cdots(2k)$. We apply (P-3) $n-1$ times and find that $t(h,n) \geqslant 2^{n-1} \cdot t\left(4k(2nk)^{r-n}, 1\right)$. As $r-n \geqslant 0$ and $k \geqslant 1$, we have $4k(2nk)^{r-n} \geqslant 4$. Applying (P-2), we find that $t(h,n) \geqslant 2^{n-1} \cdot \binom{4k(2nk)^{r-n}}{2} > 2^{n-1}4k(2nk)^{r-n} = 2^r 2k(nk)^{r-n} \geqslant y \cdot 2k(nk)^{r-n} > y$. As $2(2nk)^r \geqslant h$, again since t is monotone in its first argument it follows that $t\,(2(2nk)^r, n) \geqslant t(h,n) \geqslant y$. □

4. Applications

Theorem 1.1 leads to packing number bounds for families of functions taking values in arbitrary metric spaces. We first recall the definition of packing numbers for a metric space.

A set $T \subseteq Y$ is ε-separated in a metric space $\langle Y, \rho\rangle$ if $\rho(y, y') > \varepsilon$ for any distinct $y, y' \in T$. The space $\langle Y, \rho\rangle$ is totally bounded if, for all $\varepsilon > 0$, the cardinality of its largest ε-separated subset, denoted by $\mathscr{M}_\varepsilon(Y, \rho)$, is finite. The numbers $\mathscr{M}_\varepsilon(Y, \rho)$ are called *packing numbers*.

To derive bounds on packing numbers for families of functions mapping into a metric space, we use generalizations of the notions of shattering and VC dimension commonly used in the literature on empirical processes. Let $\boldsymbol{F} \subseteq Y^n$. For any $\gamma > 0$ and $\alpha \geqslant 2$, we say that $\boldsymbol{F}$ (α, γ)-shatters a nonempty set $\{i_1, \ldots, i_d\} \subseteq \{1, \ldots, n\}$ if there exists $(\boldsymbol{v}, \boldsymbol{w}) \in Y^d \times Y^d$ such that, $\rho(v_j, w_j) > \alpha\gamma$ for each $j = 1, \ldots, d$ and

$$(\forall \boldsymbol{y} \in \{v_1, w_1\} \times \cdots \times \{v_d, w_d\})(\exists \boldsymbol{f} \in \boldsymbol{F}) \quad \rho(y_j, f_{i_j}) \leqslant \gamma \quad \text{for each } j = 1, \ldots, d.$$

Let $\mathscr{F}$ be a family of functions $f : X \to Y$, where X is an arbitrary set and $\langle Y, \rho\rangle$ is a totally bounded metric space. Define, for each $(x_1, \ldots, x_n) \in X^n$,

$$\mathscr{F}_{|(x_1,\ldots,x_n)} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathscr{F}\}.$$

For any $\gamma > 0$ and $\alpha \geqslant 2$, the (α, γ)-dimension of $\mathscr{F}$, denoted by $\mathrm{DIM}_{\alpha,\gamma}(\mathscr{F})$, is defined by

$$\max\left\{d : (\exists \boldsymbol{x} \in X^d)\, \mathscr{F}_{|\boldsymbol{x}} \ (\alpha, \gamma)\text{-shatters } \{1, \ldots, d\}\right\}.$$

If for each $d > 0$ there exists $\boldsymbol{x} \in X^d$ such that $\mathscr{F}_{|\boldsymbol{x}}$ (α, γ)-shatters $\{1, \ldots, d\}$, then we define $\mathrm{DIM}_{\alpha,\gamma}(\mathscr{F}) = \infty$.

The notion of (α, γ)-shattering defined here generalizes the notion of γ-shattering given in [1] (originally introduced by Kearns and Schapire in [10]), which is defined only for the case when Y is a bounded interval on the real line and $\rho(u, v) = |u - v|$. In particular, for this metric space, if $\boldsymbol{x}$ is $(4, \gamma)$-shattered then $\boldsymbol{x}$ is γ-shattered in the sense of [1]. This implies that $\mathrm{DIM}_{\alpha,\gamma}$ is smaller than or equal to the P_γ-dimension defined in [1] for all $\alpha \geqslant 4$. As pointed out in [1], the P_γ-dimension is less than or equal to the pseudo-dimension defined by Pollard [12] (see also [7]) for all $\gamma > 0$.

Our packing bounds for function classes will depend on a quantity directly related to the metric structure of $\langle Y, \rho \rangle$. An (α, γ)-*packed graph* for $\langle Y, \rho \rangle$ is any undirected and irreflexive graph $G = \langle V, E \rangle$ such that: (i) V is a maximal γ-separated set in $\langle Y, \rho \rangle$, (ii) $\{v, v'\} \in E$ if and only if $\rho(v, v') > \alpha\gamma$, and (iii) $\kappa_{\alpha,\gamma}(Y, \rho) \stackrel{\text{def}}{=} |E|$ is minimized over all graphs $G = \langle V, E \rangle$ satisfying (i) and (ii).

Note that since $|V| = \mathscr{M}_\gamma(Y, \rho)$, $\kappa_{\alpha,\gamma}(Y, \rho) \leqslant \binom{\mathscr{M}_\gamma(Y,\ \rho)}{2}$.

Finally, for any metric space $\langle Y, \rho \rangle$ and any $n > 0$, we associate with Y^n the metric ρ_n defined by $\rho_n(\boldsymbol{u}, \boldsymbol{v}) = \max_{1 \leqslant i \leqslant n} \rho(u_i, v_i)$ for all $\boldsymbol{u}, \boldsymbol{v} \in Y^n$.

Theorem 4.1. *Let $\mathscr{F}$ be an arbitrary family of functions $f : X \to Y$, where X is a set and $\langle Y, \rho \rangle$ is a totally bounded metric space. If $\mathrm{DIM}_{\alpha,\gamma}(\mathscr{F}) = d < \infty$, then for all $n > d$, for all $\boldsymbol{x} \in X^n$, and for all $\gamma > 0$, $\alpha \geqslant 2$,*

$$\mathscr{M}_{(\alpha+2)\gamma}(\mathscr{F}_{|\boldsymbol{x}}, \rho_n) \leqslant 2(2nk)^{\left\lceil \log_2 \sum_{i=0}^{d} \binom{n}{i} k^i \right\rceil},$$

where $k = \kappa_{\alpha,\gamma}(Y, \rho)$.

The packing numbers $\mathscr{M}_\varepsilon(\mathscr{F}_{|\boldsymbol{x}}, \rho_n)$ for $\varepsilon > 0$ will be called ℓ_∞ *packing numbers* for (restrictions of) $\mathscr{F}$. To get the best bounds on these packing numbers from the above theorem, one must explore different settings for $\alpha \geqslant 2$ and $\gamma > 0$ such that $\varepsilon = (\alpha+2)\gamma$. For example, note that for fixed γ, as α grows, $\mathrm{DIM}_{\alpha,\gamma}$ can only get smaller, since the conditions for (α, γ)-shattering get stricter. Hence the value d in the above theorem gets smaller as α grows, giving a smaller upper bound. However, to balance out an increase in α, one must reduce γ, and by similar reasoning one sees that this has the effect of increasing the bound.

The proof of Theorem 4.1 is based on the following lemma. Recall from Section 3 that

$$H(n, k, d) = 2(2nk)^{\left\lceil \log_2 \sum_{i=0}^{d} \binom{n}{i} k^i \right\rceil}.$$

Lemma 4.1. *Let $\boldsymbol{F}$ be $(\alpha+2)\gamma$-separated in $\langle Y^n, \rho_n \rangle$, where $\langle Y, \rho \rangle$ is a totally bounded metric space. If $|\boldsymbol{F}| > H\left(n, \kappa_{\alpha,\gamma}(Y, \rho), d\right)$, then $\boldsymbol{F}$ (α, γ)-shatters a set $A \subseteq \{1, \ldots, n\}$ with $|A| = d + 1$.*

Proof. Choose any (α, γ)-packed graph $G = \langle V, E \rangle$ for $\langle Y, \rho \rangle$ and define a Voronoi tessellation of Y through any mapping $\mu : Y \to V$ satisfying $\rho(y, \mu(y)) = \min_{v \in V} \rho(y, v)$ for each $y \in Y$.

Pick any two distinct $\boldsymbol{f}, \boldsymbol{g} \in \boldsymbol{F}$ and find a coordinate i, $1 \leqslant i \leqslant n$, such that $\rho(f_i, g_i) > (\alpha+2)\gamma$. Note that, as V is a maximal γ-separated set, $\rho(f_i, \mu(f_i)) \leqslant \gamma$ and $\rho(g_i, \mu(g_i)) \leqslant \gamma$. Thus by the triangle inequality $\rho(\mu(f_i), \mu(g_i)) > \alpha\gamma$, implying $\{\mu(f_i), \mu(g_i)\} \in E$. Hence $\boldsymbol{\mu}(\boldsymbol{F}) \subseteq V^n$, defined by

$$\boldsymbol{\mu}(\boldsymbol{F}) = \{(\mu(f_1), \ldots, \mu(f_n)) : \boldsymbol{f} \in \boldsymbol{F}\},$$

has cardinality $|\boldsymbol{\mu}(\boldsymbol{F})| > H\left(n, \kappa_{\alpha,\gamma}(Y,\rho), d\right)$ and is a clique in G^n. Therefore, since $|E| = \kappa_{\alpha,\gamma}(Y,\rho)$ by definition of G, by Theorem 1.1 there exists a set $A = \{i_1,\ldots,i_{d+1}\}$ such that a subset $C = \{v_{i_1}, w_{i_1}\} \times \cdots \times \{v_{i_{d+1}}, w_{i_{d+1}}\}$ of $\boldsymbol{\mu}(\boldsymbol{F})_{|A}$ is a cube in G^{d+1}.

Since C is a cube in G^{d+1}, $\{v_{i_j}, w_{i_j}\}$ is an edge in G for all $1 \leqslant j \leqslant d+1$. Hence, $\rho(v_{i_j}, w_{i_j}) > \alpha\gamma$, $j = 1,\ldots,d+1$. Choose any $\boldsymbol{y} \in C$. Find $\boldsymbol{f} \in \boldsymbol{F}$ such that $\mu(f_{i_j}) = y_{i_j}$ for $j = 1,\ldots,d+1$. As V is a maximal γ-separated set in (Y,ρ), we have $\rho(f_{i_j}, y_{i_j}) \leqslant \gamma$ for $j = 1,\ldots,d+1$. Hence A is (α,γ)-shattered by $\boldsymbol{F}$. □

Proof of Theorem 4.1. By contradiction. Choose $\boldsymbol{x} \in X^n$ and let $\boldsymbol{F} = \mathscr{F}_{|\boldsymbol{x}}$ be $(\alpha+2)\gamma$-separated in $\langle Y^n, \rho_n\rangle$ with $|\boldsymbol{F}| > H(n,k,d)$. By Lemma 4.1, there exists $A \subseteq \{1,\ldots,n\}$ with $|A| = d+1$ that is (α,γ)-shattered by $\mathscr{F}_{|\boldsymbol{x}}$. This contradicts the assumption that $\mathrm{DIM}_{\alpha,\gamma}(\mathscr{F}) = d$. □

Now let $\mathscr{F}$ be a family of functions from a set X into a metric space (Y,ρ) as above and let P be a probability distribution on X. Define the distance $d_{L_1(P)}$ on $\mathscr{F}$ by $d_{L_1(P)}(f,g) = \int \rho(f(x), g(x))dP(x)$. Using a trick from [4], we can apply Theorem 4.1 to bound the quantity $\mathscr{M}_\varepsilon(\mathscr{F}, d_{L_1(P)})$ as well, which we refer to as the L_1 *packing numbers* for $\mathscr{F}$.

The *diameter* of a totally bounded metric space (Y,ρ) is $\sup_{y,y' \in Y} \rho(y,y')$. Note that from the triangle inequality, the diameter is at most ε times the size of its largest ε-separated subset plus 1, for any $\varepsilon > 0$.

Theorem 4.2. *Let $\mathscr{F}$ be an arbitrary family of functions $f : X \to Y$, where X is a set and $\langle Y,\rho\rangle$ is a totally bounded metric space with diameter R. If $\mathrm{DIM}_{\alpha,\gamma}(\mathscr{F}) = d < \infty$, then there exists a constant $c > 0$ such that for all $\gamma > 0$ and for all $\alpha \geqslant 2$,*

$$\sup_P \mathscr{M}_{2(\alpha+2)\gamma}(\mathscr{F}, d_{L_1(P)}) \leqslant \left\lceil \left(\frac{kdR}{\gamma}\right)^{cd\ln(kdR/\gamma)} \right\rceil, \tag{2}$$

where $k = \kappa_{\alpha,\gamma}(Y,\rho)$ and the supremum is taken over all probability distributions P on X.

This is complemented by the following result by Bartlett et al. (for completeness, we repeat their proof using our terminology) showing that any function class of high $(4,\gamma)$-dimension must include a large set that is $(\gamma/2)$-separated in the sense of Theorem 4.2.

Theorem 4.3 (Bartlett et al. [3])**.** *Let $\mathscr{F}$ be a family of functions $f : X \to Y$, where X is a set and (Y,ρ) is a metric space. Then for any $\gamma > 0$*

$$\sup_P \mathscr{M}_{\gamma/2}(\mathscr{F}, d_{L_1(P)}) \geqslant \left\lceil \mathrm{e}^{d/8} \right\rceil,$$

where $d = \mathrm{DIM}_{4,\gamma}(\mathscr{F})$.

To prove Theorem 4.2 we use a "probabilistic method" that goes back to Dudley [4] (Dudley's trick also inspired Bartlett et al. in [3].) The basic tool in our proof is the

following Chernoff–type bound (proven in [2] in a slightly less general form) on the sum of independent random variables with bounded range.

Lemma 4.2. *Let $\xi_1,\ldots,\xi_n$ be a sequence of mutually independent random variables such that $0 \leqslant \xi_i \leqslant M$, $M < \infty$, and $\mathbf{E}[\xi_i] = \mu$, $i = 1,\ldots,n$. Then, for all $\delta \geqslant 0$,*

$$\Pr\left\{\sum_{i=1}^{n} \xi_i \leqslant (1-\delta)\mu n\right\} \leqslant \mathrm{e}^{-\delta^2 \mu n/(2M)}.$$

Proof of Theorem 4.2. Let P be a distribution on X and let $\mathscr{G} \subseteq \mathscr{F}$ be any maximal set which is $2(\alpha+2)\gamma$-separated with respect to $d_{L_1(P)}$. As $\langle Y, \rho\rangle$ is totally bounded, we have $\sup_{x\in X} \rho(f(x), g(x)) \leqslant R$, where $0 < R < \infty$ is the diameter of $\langle Y, \rho\rangle$. Let $x_1,\ldots,x_n$ be mutually independent random draws from P. For each $\{f,g\} \subseteq \mathscr{G}$ we apply Lemma 4.2, with $\delta = 1/2$, to the random variables $\xi_i = \rho(f(x_i), g(x_i))$. Noting that $\mathbf{E}[\xi_i] > 2(\alpha+2)\gamma$, we get

$$P^n\left\{\min_{\{f,g\}\subseteq\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \rho(f(x_i), g(x_i)) \leqslant (\alpha+2)\gamma\right\} \leqslant \binom{|\mathscr{G}|}{2} \exp\left(-\frac{(\alpha+2)\gamma n}{4R}\right). \quad (3)$$

Therefore, for $n > (4R/(\alpha+2)\gamma)\ln\binom{|\mathscr{G}|}{2}$ we can find $\boldsymbol{x} = (x_1,\ldots,x_n) \in X^n$ such that for any $\{\boldsymbol{f},\boldsymbol{g}\} \subseteq \mathscr{G}$ it holds that $n^{-1}\sum_{i=1}^{n} \rho(f_i, g_i) > (\alpha+2)\gamma$. This clearly implies that, for this $\boldsymbol{x}$, $\mathscr{G}_{|\boldsymbol{x}}$ is $(\alpha+2)\gamma$-separated in $\langle Y^n, \rho_n\rangle$.

Let $N = |\mathscr{G}| = |\mathscr{G}_{|\boldsymbol{x}}|$ and assume (i) $n > (4R/(\alpha+2)\gamma)\ln\binom{N}{2}$ and (ii) $N > H(n,k,d)$ both hold, where $k = \kappa_{\alpha,\gamma}(Y,\rho)$. Then, using Lemma 4.1, we conclude that $\mathscr{G}_{|\boldsymbol{x}}$ (α,γ)-shatters a set of cardinality $d+1$, contradicting $\mathrm{DIM}_{\alpha,\gamma}(\mathscr{G}) \leqslant \mathrm{DIM}_{\alpha,\gamma}(\mathscr{F}) = d$.

As (i) is implied by $n \geqslant (2R/\gamma)\ln N$, for (i) and (ii) to hold it is sufficient that

$$H(n,k,d) < N \leqslant \mathrm{e}^{n\cdot\gamma/(2R)}. \quad (4)$$

Using $2^{d\log_2^2(enk)+1}$ to upper bound $H(n,k,d)$ — see discussion before Corollary 2.1, a positive constant c can be found such that $n \geqslant (2R/\gamma)cd\ \ln^2(kdR/\gamma)$ implies $\mathrm{e}^{n\cdot\gamma/(2R)} > H(\lceil n\rceil,k,d)$ for all $k \geqslant 1$ and all $d \geqslant 1$. Hence, for each integer $N \geqslant \mathrm{e}^{cd\ \ln^2(kdR/\gamma)}$ some integer $n \geqslant (2R/\gamma)cd\ \ln^2(kdR/\gamma)$ can be found such that (4) holds, leading to a contradiction. It follows that

$$|\mathscr{G}| = N < \left\lceil \mathrm{e}^{cd\ \ln^2(kdR/\gamma)} \right\rceil.$$

Since $\mathscr{G}$ was an arbitrary maximal $2(\alpha+2)\gamma$-separated subset of $\mathscr{F}$ with respect to $d_{L_1(P)}$, the result follows. □

Proof of Theorem 4.3. Choose $\mathscr{F}$ and choose $\gamma > 0$. Let $d = \mathrm{DIM}_{4,\gamma}(\mathscr{F})$. Let $\boldsymbol{x} \in X^d$ be a sequence that is $(4,\gamma)$-shattered by some $\boldsymbol{F} \subseteq \mathscr{F}_{|\boldsymbol{x}}$ of cardinality 2^d. Let $C(\gamma/2)$ be the minimum integer c such that

$$|\{\boldsymbol{g} \in \boldsymbol{F} : \ell_1(\boldsymbol{f},\boldsymbol{g}) \leqslant \gamma/2\}| \leqslant c \quad \text{for all } \boldsymbol{f} \in \boldsymbol{F},$$

where we define $\ell_1(\boldsymbol{f},\boldsymbol{g}) = d^{-1}\sum_{i=1}^{d}\rho(f_i,g_i)$. For any two $\boldsymbol{f},\boldsymbol{g}\in\boldsymbol{F}$ let

$$e(\boldsymbol{f},\boldsymbol{g}) = \{i : 1 \leqslant i \leqslant d,\ \rho(f_i,g_i) > 2\gamma\}.$$

Note that, by definition of $(4,\gamma)$-shattering and by our choice of $\boldsymbol{F}$, $e(\boldsymbol{f},\boldsymbol{g}) = e(\boldsymbol{f},\boldsymbol{g}')$ if and only if $\boldsymbol{g}=\boldsymbol{g}'$ for any $\boldsymbol{f},\boldsymbol{g},\boldsymbol{g}'\in\boldsymbol{F}$. Furthermore, $\ell_1(\boldsymbol{f},\boldsymbol{g}) \leqslant \gamma/2$ implies $|e(\boldsymbol{f},\boldsymbol{g})| \leqslant d/4$. Hence,

$$C(\gamma/2) \leqslant \sum_{k=0}^{d/4}\binom{d}{k}.$$

Using the Chernoff bound (see [3])

$$\sum_{k=0}^{m}\binom{d}{k}p^k(1-p)^{d-k} \leqslant \exp\left\{-\frac{(dp-m)^2}{2dp(1-p)}\right\} \quad \text{for all } p\leqslant 1/2 \text{ and } m\leqslant dp$$

and letting $p=1/2$ and $m=d/4$ we get

$$\sum_{k=0}^{d/4}\binom{d}{k} \leqslant 2^d e^{-d/8}.$$

Hence,

$$\begin{aligned}\sup_P \mathscr{M}_{\gamma/2}\left(\mathscr{F}, d_{L_1(P)}\right) &\geqslant \mathscr{M}_{\gamma/2}(\boldsymbol{F},\ell_1)\\ &\geqslant \left\lceil\frac{2^d}{C(\gamma/2)}\right\rceil\\ &\geqslant \left\lceil e^{d/8}\right\rceil\end{aligned}$$

and this concludes the proof. □

5. Conclusions

We have given bounds on the ℓ_∞ and L_1 packing numbers for sets of functions mapping into a totally bounded metric space. These are based on certain combinatorial notions of shattering and dimension that generalize earlier related notions, which have proved useful in establishing strong and uniform laws of large numbers and for investigating the learnability of function classes in some formal learning models as well (see e.g. [7, 10, 12].)

Our results extend to metric spaces previous results shown for the case when Y is the interval $[0,1]$ and $\rho(u,v) = |u-v|$. For sets of real-valued functions, L_1 packing number bounds were derived in [7, 8, 12] using Pollard's notion of pseudo-dimension. Further bounds, based on the notion of γ-shattering (closely related to our notion of (α,γ)-dimension), were later shown in [1] for the ℓ_∞ norm and in [3, 11] for the L_1 norm. For a discussion about the relationships between these bounds see [3].

Acknowledgements

Part of this research was done while Nicolò Cesa-Bianchi was visiting UC Santa Cruz. Partial support by the ESPRIT working group 8556 NeuroCOLT is gratefully acknowledged.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, J. ACM 44 (1997) 615–631.
[2] D. Angluin, L.G. Valiant, Fast probabilistic algorithms for Hamiltonian circuits and matchings, J. Comput. Systems Sci. 18 (1979) 155–193.
[3] P.L. Bartlett, S.R. Kulkarni, S.E. Posner, Covering numbers for real-valued function classes, IEEE Trans. Inform. Theory, 43(1997) 1721–1725.
[4] R.M. Dudley, Central limit theorems for empirical measures, Ann. Probab. 6 (1979) 899–929, Correction in 7 (1979) 909–911.
[5] R.M. Dudley, E. Giné, J. Zinn, Uniform and universal Glivenko–Cantelli classes, J. Theoret. Probab. 4 (1991) 485–510.
[6] L. Gargano, J. Körner, U. Vaccaro, Capacities: from information theory to extremal set theory, J. Combin. Theory Ser. A 68 (1994) 296–316.
[7] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, Inform. Comput. 100 (1) (1992) 78–150.
[8] D. 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.
[9] D. Haussler, P.M. Long, A generalization of Sauer's lemma, J. Combin. Theory Ser. A 71 (1995) 219–240.
[10] M. Kearns, R.E. Schapire, Efficient distribution-free learning of probabilistic concepts, J. Comput. Systems Sci. 48 (3) (1994) 464–497. An extended abstract appeared in the proc. 30th Ann. Symp. on the Foundations of Computer Science.
[11] W.S. Lee, P. Bartlett, R.C. Williamson, On efficient agnostic learning of linear combinations of basis functions. In Proceedings of the 8th Annual Conference on Computational Learning Theory, pages 369–376. ACM Press, 1995.
[12] D. Pollard, Empirical Processes : Theory and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 2, Institute of Math. Stat. and Am. Stat. Assoc., 1990.
[13] N. Sauer, On the density of families of sets, J. Combin. Theory Ser. A 13 (1972) 145–147.
[14] C.E. Shannon, The zero-error capacity of a noisy channel, IRE Trans. Inform. Theory 2 (1956) 8–19.
[15] S. Shelah, A combinatorial problem: Stability and order for models and theories in infinitary languages, Pacific J. Math. 41 (1972) 247–261.
[16] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (2) (1971) 264–280.