ELSEVIER

Discrete Applied Mathematics 86 (1998) 81–93

DISCRETE APPLIED MATHEMATICS

The degree of approximation of sets in euclidean space using sets with bounded Vapnik–Chervonenkis dimension

Vitaly Maiorov[a,*], Joel Ratsaby[b]

[a]*Department of Mathematics, Haifa 32000, Israel*

[b]*Department of Electrical Engineering, Haifa 32000, Israel*

Received 30 December 1996; received in revised form 24 April 1997; accepted 23 October 1997

Abstract

The degree of approximation of infinite-dimensional function classes using finite n-dimensional manifolds has been the subject of a classical field of study in the area of mathematical approximation theory. In Ratsaby and Maiorov (1997), a new quantity $\rho_n(F, L_q)$ which measures the degree of approximation of a function class F by the best manifold H^n of pseudo-dimension less than or equal to n in the L_q-metric has been introduced. For sets $F \subset \mathbb{R}^m$ it is defined as $\rho_n(F, l_q^m) = \inf_{H^n} dist(F, H^n)$, where $dist(F, H^n) = \sup_{x \in F} \inf_{y \in H^n} \|x - y\|_{l_q^m}$ and $H^n \subset \mathbb{R}^m$ is any set of VC-dimension less than or equal to n where $n < m$. It measures the degree of approximation of the set F by the optimal set $H^n \subset \mathbb{R}^m$ of VC-dimension less than or equal to n in the l_q^m-metric. In this paper we compute $\rho_n(F, l_q^m)$ for F being the unit ball $B_p^m = \{x \in \mathbb{R}^m : \|x\|_{l_p^m} \leqslant 1\}$ for any $1 \leqslant p,\ q \leqslant \infty$, and for F being any subset of the boolean m-cube of size larger than $2^{m\gamma}$, for any $\frac{1}{2} < \gamma < 1$. © 1998 Published by Elsevier Science B.V. All rights reserved.

1. Introduction

We will use the following notation. Let the norm $\|x\|_{l_q^m} = (\sum_{i=1}^m |x_i|^q)^{1/q}$. For two sets $A, B \subset \mathbb{R}^m$ define the distance $dist(A, B, l_q^m) = \sup_{a \in A} \inf_{b \in B} \|a - b\|_{l_q^m}$. Let m be a positive integer. For a vector $x \in \mathbb{R}^m$ denote by $\mathrm{sgn}(x) = [\mathrm{sgn}(x_1), \ldots, \mathrm{sgn}(x_m)]$, where $\mathrm{sgn}(x_i) = 1$ if $x_i > 0$ and $\mathrm{sgn}(x_i) = -1$ if $x_i \leqslant 0$, for $1 \leqslant i \leqslant m$. For a set $A \subset \mathbb{R}^m$ denote by $\mathrm{sgn}(A) = \{\mathrm{sgn}(x) : x \in A\}$. For any finite set B denote by $|B|$ the cardinality of B. The next definition of the VC-dimension of a set $F \subset \mathbb{R}^m$ follows that of Haussler [8].

Definition 1 (*VC-dimension of a set in* $\mathbb{R}^m$). Let $F \subset \mathbb{R}^m$. For an index set $I \subset \{1, 2, \ldots, m\}$ of cardinality k let

$$F_{|I} = \{[x_{i_1}, \ldots, x_{i_k}] : x = [x_1, \ldots, x_m] \in F,\ i_j \in I,\ 1 \leqslant j \leqslant k\}.$$

* Corresponding author. E-mail: maiorov@tx.technion.ac.il.

0166-218X/98/$19.00
PII S0166-218X(98)00015-8

The *Vapnik–Chervonenkis dimension* of F, denoted by $VC(F)$, is the largest k such that there exists an index set I of cardinality k satisfying $|\mathrm{sgn}(F_{|I})| = 2^k$.

Definition 2 (*Pseudo-dimension of a set in* $\mathbb{R}^m$). Let $F \subset \mathbb{R}^m$. For any $y \in \mathbb{R}^k$ and an index set $I \subset \{1,2,\dots,m\}$ of cardinality k let

$$F_{|I,y} = \{[x_{i_1} + y_1, \dots, x_{i_k} + y_k] : x = [x_1,\dots,x_m] \in F,\ i_j \in I,\ 1 \leqslant j \leqslant k\}.$$

The *Pseudo dimension* of F, denoted by $dim_p(F)$, is the largest k such that there exists a $y \in \mathbb{R}^k$ and an index set I of cardinality k satisfying $|\mathrm{sgn}(F_{|I,y})| = 2^k$.

The VC-dimension of classes of indicator functions of sets was first introduced by Vapnik and Chervonenkis [21, 22], who also defined a similar notion of capacity for real-valued function classes. For real-valued functions, Pollard [14] and Haussler [7] later extended the definition of VC-dimension to the pseudo-dimension. By characterizing an important statistical estimation property of a class of functions these dimensions play a central role in the theory of pattern recognition and regression estimation (cf. [20]), empirical processes (cf. [13, 14, 19]) and computational learning theory (cf. [3, 7]). For a class F of functions on X which has a finite VC or pseudo-dimension, it is possible to estimate any $f \in F$ by some $\hat{f} \in F$ to an arbitrary accuracy ε and confidence $1 - \delta$ by just knowing its functional values $f(x_i)$ at a finite number of randomly drawn points $x_i \in X$, $1 \leqslant i \leqslant m < \infty$, where m depends on ε and δ.

Being a measure of capacity, the VC-dimension is related to the more classical notion of ε-entropy of a functional class, cf. [20]. Similarly, the ε-packing number of a Euclidean set F is related to its VC-dimension. Haussler [8] has recently improved this bound for F being any subset of the boolean m-cube having VC-dimension $n < m$. This improved bound takes the form of $O(n/\varepsilon^n)$ and is essential for obtaining tight bounds on the quantities of interest in our work.

In this paper we study the ability of sets of finite VC-dimension or pseudo-dimension in approximating richer sets in Euclidean space. The result is used for determining the degree of approximation of infinite-dimensional classes of functions by finite VC or pseudo-dimensional manifolds of functions, cf. [11], where such manifolds are shown to be powerful in approximating standard functional classes. Together with their statistical estimation property mentioned above, such manifolds prove to be valuable in a framework of learning from examples with partial information, cf. [15, 16]. Before proceeding to describe the main quantity which is estimated in this paper we review some elementary notions in the field of approximation theory. This field deals with calculating the degree of approximation of sets F, in general, normed linear spaces $\mathscr{F}$ by n-dimensional (linear) subspaces H_n of $\mathscr{F}$ and more generally by non-linear n-dimensional manifolds of $\mathscr{F}$.

The classical Kolmogorov width (cf. [12, 9]) measures the degree of approximation of F by the optimal subspace over all n-dimensional subspaces H_n. It is defined as $d_n(F, L_q) = \inf_{H_n} \sup_{f \in F} \inf_{h \in H_n} \|f - h\|_{L_q}$, $q \geqslant 1$. The Gelfand width is similar except it considers approximation of F using subspaces H^n of co-dimension n. It is defined as $d^n(F,L_q) = \inf_{H^n} \sup_{f\in F} \inf_{h\in H^n} \|f-h\|_{L_q}$. The *linear* width is defined as $\delta_n(F,L_q) = \inf_{P_n} \sup_{f\in F} \|f - P_n(f)\|_{L_q}$, where the infimum is taken over all continuous linear operators $P_n : F \to F$ for which the range of P_n is of dimension n. In all three widths, elements of F are approximated by elements of linear n-dimensional manifolds. The problem of non-linear approximation also occupies a significant portion of research in approximation theory. An n-dimensional non-linear manifold $\mathcal{M}_n$ is a class of functions parameterized by a vector $a \in \mathbb{R}^n$ which are, in general, non-linear functions of a. For instance, in the non-linear manifold of functions on $X = \mathbb{R}$ which are represented by single-hidden-layer neural networks, functions take the form $f(x,a) = \sum_{i=1}^{l} c_i\sigma(w_i x + b_i)$, where $\sigma(z) = 1/(1+e^{-z})$, the parameter $a = [c_1,\dots,c_l,w_1,\dots,w_l,b_1,\dots,b_l]$. A general non-linear manifold of functions is the image of a mapping $M_n : \mathbb{R}^n \to \mathcal{M}_n$. If M_n is a linear mapping then $\mathcal{M}_n$ is an n-dimensional subspace.

There are many known function classes F which can be approximated better by non-linear manifolds such as splines, neural networks and radial basis functions, than by linear manifolds such as polynomials. It is therefore of interest to consider the degree of optimal approximation of general classes F by non-linear manifolds. However, the space of all non-linear n-dimensional manifolds $\mathcal{M}_n$ is extremely rich. In order to define the degree of non-linear approximation of F some restriction must be imposed either on the manifolds used for approximation or on the mapping which relates each element $f \in F$ with its approximation element in $\mathcal{M}_n$. Otherwise, as DeVore [4] notes, a one-dimensional non-linear manifold containing a dense subset of F yields an arbitrarily small approximation error for any $f \in F$. This makes the degree of approximation of F by the space of all n-dimensional manifolds be trivially zero.

The classical Alexandrov width of a function class (cf. [17, 4]) is defined as $a_n(F,L_q) = \inf_{S:F\to\mathbb{R}^n,\mathcal{M}_n} \sup_{f\in F} \|f - M_n(S(f))\|_{L_q}$ where S is constrained to be a continuous selection operator mapping F to $\mathbb{R}^n$ and the infimum is taken over all such S and all manifolds $\mathcal{M}_n$. For any element $f \in F$, the best approximation is taken as the optimal element h in the optimal manifold $\mathcal{M}_n$, under the constraint that f is mapped to the parameter a of h through a continuous mapping S. The Alexandrov width differs from the previous widths in permitting not only linear manifolds in the approximation of F. It however introduces a continuity restriction on the selection operator S which in many applications is not natural since it results in the optimal approximating element being not necessarily the closest to the target f among all functions in $\mathcal{M}_n$.

Notions from discrete mathematics are useful in the estimation problems of widths in general sets F of normed linear spaces $\mathscr{F}$. It is often the case that F is the image of the unit ball in $\mathscr{F}$ with respect to the L_q-norm. In such cases it is usually possible to reduce the approximation problem into a finite-dimensional problem where instead of $\mathscr{F}$ and F one has $\mathbb{R}^m$ and $B_p^m = \{x \in \mathbb{R}^m : \|x\|_{l_p^m} \leqslant 1\}$, $p \geqslant 1$, respectively. Distances are measured using the l_q^m-norm, $1 \leqslant q \leqslant \infty$. There are several discretization techniques used for this reduction, cf. [9, p. 451, 12, p. 234]. Once discretized, the calculation of the width of B_p^m leads to an estimate of the width of the original infinite-dimensional function class. For instance, Theorem 3.4 in [12] gives upper bounds on d_n, d^n and δ_n for an infinite-dimensional Sobolev function class directly in terms of d_n, d^n and

δ_n for Euclidean balls, respectively. In [11], we estimated a new width $\rho_n(F, L_q)$ for a function class F using results based on the current work. Thus, the problem of estimating widths of Euclidean sets is central to the problem of estimating widths of more general infinite-dimensional spaces.

The classical widths of the ball B_p^m are well known. For the case $1 \leqslant q \leqslant p \leqslant \infty$, $1 \leqslant n \leqslant m$, all of the three widths above equal $(m-n)^{(1/q)-1/p}$, cf. [9]. The case of $1 \leqslant p \leqslant q \leqslant \infty$ is more involved. For example, in the case of $p=1$, $q=2$, it is known that $d_n(B_1^m, l_2^m) = \delta_n(B_1^m, l_2^m) = \sqrt{1-n/m}$ (for more results cf. [9]). Note that in this case if $m = c_0 n$ for some constant $c_0 > 0$ then the widths d_n, δ_n equal a constant. More generally, there are other cases where the approximation error of B_p^m by the optimal n-dimensional linear manifold does not decrease to zero as n increases. This is representative of the limitation of linear approximation. In contrast, as we will see in this paper, for the same example above, the optimal manifold of VC-dimension n achieves an approximation error of $1/\sqrt{n}$ which asymptotically equals zero as $n \to \infty$, for any $m \geqslant c_1 n$ for some absolute constant $c_1 > 0$.

2. The ρ_n width

We mentioned two independent areas of research the first being approximation theory and the second is VC-theory which mainly studies the statistical estimation properties of classes having a finite VC-dimension or any of the other extended definitions such as the pseudo-dimension (cf. [7]), scale-sensitive dimension (cf. [1]). There are several examples of the cross discipline between these two fields. Warren [23] considered a quantity called the number of connected components of a non-linear manifold of real-valued functions, which closely resembles the growth function of Vapnik and Chervonenkis for set-indicator functions. Using this he determined lower bounds on the degree of approximation by certain non-linear manifolds. Maiorov [10] used these ideas to determine the degree of approximation for the non-linear manifold of ridge functions which include the manifold of neural networks with a single hidden layer. Barron [2] considered the VC-dimension of the dual of a class F of parameterized subsets in Euclidean space which is called the coVC-dimension of F. Using central limit theorem for empirical processes he determined the degree of approximation of a class of functions with bounded variation by the non-linear manifold of neural networks. Gurvits and Koiran [6] used the coVC-dimension to study the approximation degree of the closure of convex hulls of general functional classes by classes of convex combinations of n functions. Girosi [5] considered target classes of functions which are convolutions of some fixed kernel. Using the uniform strong law convergence rate obtained by VC-theory he directly obtained bounds on the approximation degree of such target classes by the non-linear manifold consisting of all linear combinations of n translates of the kernel.

In combining VC-theory and approximation theory, our works [15, 16, 11] differ from the last three above in that the VC-dimension is used to impose a constraint on the non-linear manifolds rather than using VC-theory for converting uniform strong law results into approximation error results. We defined a new width, denoted as $\rho_n(F,L_q)=\inf_{H^n}\sup_{f\in F}\inf_{h\in H^n}\|f-h\|_{L_q}$ where H^n runs over all function classes of pseudo-dimension n which may, of course, be non-linear manifolds. One of its positive attributes when compared to the Alexandrov width is that ρ_n does not restrict the selection operator to be continuous, i.e., the best-approximation mapping, which takes an element $f\in F$ to an element h in some non-linear class H^n, is not restricted. In [11] we estimated $\rho_n(F,L_q)$ for an infinite-dimensional class F of smooth functions with r partial derivatives bounded in the L_p-norm.

As for the classical widths mentioned above, the ρ_n-width is well defined for finite-dimensional spaces. In this paper we obtain a tight estimate on $\rho_n(B_p^m,l_q^m)$, for any $1\leqslant p,\ q\leqslant\infty$, and on $\rho_n(K,l_q^m)$, where $K\subset\{-1,+1\}^m$ is of an exponential cardinality in m. The two main quantities of interest in this paper are defined as follows:

Definition 3 (ρ_n^{VC}*-width*). For any set $F\subset\mathbb{R}^m$ define the ρ_n^{VC}*-width* of F as

$$\rho_n^{\mathrm{VC}}(F,l_q^m)=\inf_{H^n} \mathit{dist}(F,H^n,l_q^m),$$

where H^n runs over all sets in $\mathbb{R}^m$ of VC-dimension less than or equal to n.

We can similarly consider the degree of approximation using sets of pseudo-dimension n.

Definition 4 (ρ_n^P*-width*). For any set $F\subset\mathbb{R}^m$ define the ρ_n^P*-width* of F as

$$\rho_n^P(F,l_q^m)=\inf_{H^n} \mathit{dist}(F,H^n,l_q^m),$$

where H^n runs over all sets in $\mathbb{R}^m$ of Pseudo-dimension less than or equal to n.

3. Statement of results

Since for any set $F\subset\mathbb{R}^m$, $VC(F)\leqslant\dim_p(F)$, it follows that the family of sets of VC-dimension n contains the family of sets of pseudo-dimension n and thus $\rho_n^{\mathrm{VC}}(F,l_q^m)\leqslant\rho_n^P(F,l_q^m)$. As H^n now runs over more than just linear subspaces it is expected that $\rho_n^{\mathrm{VC}}(B_p^m,l_q^m)$ will be less than or equal to the classical widths d_n, d^n and δ_n. This is seen in the next result where for $1\leqslant q\leqslant p\leqslant\infty$, $\rho_n^{\mathrm{VC}}(B_p^m,l_q^m)$ matches the three classical widths while for $1\leqslant p<q\leqslant\infty$, $\rho_n^{\mathrm{VC}}(B_p^m,l_q^m)$ is smaller. The constant $c=\lceil 16\log_2(8e)\rceil$ is used throughout the following results.

Theorem 1. *For any integers* $n\geqslant 1$, $m\geqslant cn$, *we have*

$$\tfrac{1}{16}(m-n)^{1/q-1/p}\leqslant\rho_n^{\mathrm{VC}}(B_p^m,l_q^m)\leqslant(m-n)^{1/q-1/p},\quad \textit{if } 1\leqslant q\leqslant p\leqslant\infty$$

and

$$\frac{c_2}{n^{1/p-1/q}} \leqslant \rho_n^{\mathrm{VC}}(B_p^m, l_q^m) \leqslant \frac{1}{n^{1/p-1/q}}, \quad \text{if } 1 \leqslant p < q \leqslant \infty,$$

where $c_2 = c^{1/q-1/p}/16$.

In the next theorem we consider the approximation of any set $K \subset E$, of cardinality larger than $2^{\gamma m}$, for any constant $\frac{1}{2} < \gamma < 1$, where $E = \{-1, +1\}^m$.

Theorem 2. *For any* $1 \leqslant q \leqslant \infty$, *arbitrary* $\frac{1}{2} < \gamma < 1$, *and* $\eta > 0$ *satisfying* $(1-\gamma)(1+\eta) < \frac{1}{2}$. *For* $n \geqslant 1$, $m \geqslant c_3 n$, *where* $c_3 = \lceil (4/\eta(1-\gamma)) \log_2(8e) \rceil$, *let* $K \subset \{-1, +1\}^m$ *be any set of cardinality* $|K| = 2^{\gamma m}$. *Then*

$$\left(\frac{1}{8} - \sqrt{\frac{(1-\gamma)(1+\eta)}{32}} \right) (m-n)^{1/q} \leqslant \rho_n^{\mathrm{VC}}(K, l_q^m) \leqslant (m-n)^{1/q}.$$

Corollary 1. *As lower and upper bounds on* $\rho_n^P(B_p^m, l_q^m)$ *and* $\rho_n^P(K, l_q^m)$ *we have the lower and upper bounds for* $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m)$ *and* $\rho_n^{\mathrm{VC}}(K, l_q^m)$ *of Theorems* 1 *and* 2, *respectively.*

As a further generalization, let μ be a probability measure on the index set $\{1, 2, \ldots, m\}$ and instead of the l_q^m norm used above, consider the $l_q^m(\mu)$ norm where $\|x\|_{l_q^m(\mu)} = (\sum_{i=1}^m \mu(i)|x_i|^q)^{1/q}$. We have the following corollary.

Corollary 2. *For any fixed probability measure* μ *on* $\{1, 2, \ldots, m\}$ *let* $I = \{i \in \{1, 2, \ldots, m\} : \mu(i) > 0\}$. *Denote by* $\mu_{\min} = \min_{i \in I} \mu(i)$ *and* $\mu_{\max} = \max_{i \in I} \mu(i)$. *Provided that* μ *is such that the cardinality of* I *is greater than* n *then as lower bounds on* $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m(\mu))$ *and* $\rho_n^{\mathrm{VC}}(K, l_q^m(\mu))$, *we have the lower bounds on* $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m)$ *and* $\rho_n^{\mathrm{VC}}(K, l_q^m)$ *multiplied by a factor of* $\mu_{\min}^{1/q}$. *As upper bounds on* $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m(\mu))$ *and* $\rho_n^{\mathrm{VC}}(K, l_q^m(\mu))$, *we have the upper bounds on* $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m)$ *and* $\rho_n^{\mathrm{VC}}(K, l_q^m)$ *multiplied by a factor of* $\mu_{\max}^{1/q}$.

4. Proofs of the results

4.1. Proof of Theorem 1

We first state and prove two auxiliary lemmas.

Lemma 1. *Let* $m \geqslant 16$ *and* $E = \{-1, +1\}^m$. *Then there exists a set* $G \subset E$ *of cardinality* $2^{m/16}$ *such that for any* $v, v' \in G$, *where* $v \neq v'$, *the distance* $\|v - v'\|_{l_1^m} \geqslant m/2$.

Proof. We will construct the set G as follows: take the first point $v^1 \in G$ to be $v^1 = [1, \ldots, 1]$. Fix an $0 < \alpha < \frac{1}{2}$. Define the set $D_{v^1} = \{v \in E : \|v - v^1\|_{l_1^m} > 2\alpha m\}$. The cardinality $|D_{v^1}| \geqslant 2^{(1-\alpha)m} > 1$. We may therefore choose the second point $v^2 \in D_{v^1}$

and thus $\|v^1 - v^2\|_{l_1^m} \geqslant 2\alpha m$. Denote by $\overline{D}_{v^1}$ the complement of the set D_{v^1}. The cardinality $|D_{v^1} \cap D_{v^2}| \geqslant 2^m - |\overline{D}_{v^1}| - |\overline{D}_{v^2}|$ and as an upper bound on both $|\overline{D}_{v^i}|$, $i = 1, 2$, we may use $\sum_{j=0}^{\lfloor \alpha m \rfloor} \binom{m}{j}$ which is bounded from above by $2^m \mathrm{e}^{-2m(1/2 - \lfloor \alpha m \rfloor/m)^2} \leqslant 2^m \mathrm{e}^{-2m(1/2-\alpha)^2} \leqslant 2^{m(1-\beta)}$ where $\beta = 2(\frac{1}{2} - \alpha)^2$, the latter following from a standard application of Chebychev's inequality for the successes of m independent Bernoulli trials with probability $\frac{1}{2}$. Thus $|D_{v^1} \cap D_{v^2}| \geqslant 2^m - 2 \cdot 2^{m(1-\beta)}$ which is greater than 1 for all $m \geqslant 2/\beta$. We may therefore choose a $v^3 \in D_{v^1} \cap D_{v^2}$ where $\|v^3 - v^i\|_{l_1^m} \geqslant 2\alpha m$, $1 \leqslant i \leqslant 2$. We may repeat this for all remaining points v^k, $3 \leqslant k \leqslant N$, picking v^k from $\bigcap_{i=1}^{k-1} D_{v^i}$, as long as $N < 2^{\beta m}$. Letting $\alpha = \frac{1}{4}$ and $N = 2^{\beta m/2}$ and proceeding as above yields a set $G \subset E$ whose points are $m/2$-separated in the l_1^m-norm and whose cardinality is $2^{m/16}$, for all $m \geqslant 16$. $\square$

Lemma 2. *Let $E = \{-1, +1\}^m$. Consider any set $A^n \subset \mathbb{R}^m$ with $VC(A^n) = n$, where $n \leqslant m/c$. Then*

$$dist(E, A^n, l_1^m) = \sup_{v \in E} \inf_{x \in A^n} \|v - x\|_{l_1^m} > \frac{m}{16}.$$

Proof. Consider the set $G \subset E$ defined in Lemma 1. Define the projection operator $P : G \to A^n$ which associates each $v \in G$ with the closest point on A^n to x in the l_1^m-norm. Set

$$\delta = \sup_{v \in G} \inf_{x \in A^n} \|v - x\|_{l_1^m} = dist(G, A^n, l_1^m).$$

Consider any $v \neq v' \in G$. We have

$$\begin{aligned} &\|\mathrm{sgn}(Pv) - \mathrm{sgn}(Pv')\|_{l_1^m} \\ &\quad = \|(\mathrm{sgn}(Pv) - v) + (v' - \mathrm{sgn}(Pv')) + (v - v')\|_{l_1^m} \\ &\quad \geqslant \|v - v'\|_{l_1^m} - \|v' - \mathrm{sgn}(Pv')\|_{l_1^m} - \|v - \mathrm{sgn}(Pv)\|_{l_1^m}. \end{aligned} \tag{1}$$

Now, for any $y \in \mathbb{R}^m$, $\|v - \mathrm{sgn}(y)\|_{l_1^m} = \sum_{i=1}^m |v_i - \mathrm{sgn}(y_i)| \leqslant 2 \sum_{i=1}^m |v_i - y_i| = 2\|v - y\|_{l_1^m}$. Hence, $\|v - \mathrm{sgn}(Pv)\|_{l_1^m} \leqslant 2\|v - Pv\|_{l_1^m} \leqslant 2\delta$. Hence, from (1) we have

$$\begin{aligned} \|\mathrm{sgn}(Pv) - \mathrm{sgn}(Pv')\|_{l_1^m} &\geqslant \|v - v'\|_{l_1^m} - 2\delta - 2\delta \\ &\geqslant \frac{m}{2} - 4\delta. \end{aligned} \tag{2}$$

The set $\mathrm{sgn}(PG)$ has VC-dimension $VC(\mathrm{sgn}(PG)) \leqslant n$ because the set $PG \subset A^n$ and by definition $VC(A^n) = n$. Also, $VC(\mathrm{sgn}(PG)) = VC(PG)$ which follows from Definition 1.

As in the statement of the lemma let us restrict $m \geqslant cn$. Assume that $\delta \leqslant m/16$. The set $\mathrm{sgn}(PG)$ has the following three properties the first two of which follow from the assumption: First, by Lemma 1 it has cardinality $|\mathrm{sgn}(PG)| = 2^{m/16}$. This follows since for any $v \neq v' \in G$ the corresponding vertices $u = \mathrm{sgn}(Pv)$, and $u' = \mathrm{sgn}(Pv')$ satisfy $\|u - u'\|_{l_1^m} \geqslant (m/2) - 4\delta \geqslant m/4 > 0$ and are therefore distinct so $|\mathrm{sgn}(PG)| = |G| = 2^{m/16}$. The second property is that for any $u \neq u' \in \mathrm{sgn}(PG)$, $\|u - u'\|_{l_1^m} \geqslant m/4$. The third property is that $VC(\mathrm{sgn}(PG)) \leqslant n$.

From Theorem 1 of Haussler [8] which states an upper bound on the packing number of a set in E which has VC-dimension n, it follows that the ε-packing number of sgn(PG), in the $1/m\|\cdot\|_{l_1^m}$-norm, is upper bounded by $e(n+1)(8e)^n$, for $\varepsilon=\frac{1}{4}$. Since the set sgn(PG) is itself ε-separated in the $\frac{1}{m}\|\cdot\|_{l_1^m}$-norm then its ε-packing number is lower bounded by its cardinality which is $2^{m/16}$. Then we have the following inequality:

$$2^{m/16} \leqslant e(n+1)(8e)^n. \tag{3}$$

As m was chosen to be larger than or equal to $\lceil 16\log_2(8e)\rceil n$ then (3) reduces to

$$2\log_2(8e) \leqslant \frac{\log_2(e(n+1))}{n} + \log_2(8e) \tag{4}$$

which is false for all $n \geqslant 1$. It follows that the assumption of $\delta \leqslant m/16$ is contradicted for any $n \geqslant 1$. Hence $\delta > m/16$ and using the fact that $G \subset E$ it follows that $dist(E, A^n, l_1^m) \geqslant dist(G, A^n, l_1^m) > m/16$ which completes the proof. $\square$

We now prove the lower bound of Theorem 1.

Proof of Theorem 1 (*Lower bound*). Consider G as defined in Lemma 2. Let $\hat{G} = \{1/m^{1/p}v : v \in G\}$ and $\hat{E} = \{(1/m^{1/p})v : v \in \{-1,+1\}^m\}$. Then $\hat{E} \subset B_p^m$. If in the proof of Lemma 2 we replace the lower bound of $(m/2) - 4\delta$ by $(1/m^{1/p})(m/2) - 4\delta$ and then change the assumption to $\delta \leqslant (1/m^{1/p})m/16$ then we obtain that $dist(\hat{E}, A^n, l_1^m) \geqslant m/16m^{1/p}$. From a well known inequality we have for any vector $x \in \mathbb{R}^m$, for $a \geqslant b \geqslant 1$

$$\frac{1}{m^{1/a}}\|x\|_{l_a^m} \geqslant \frac{1}{m^{1/b}}\|x\|_{l_b^m}. \tag{5}$$

Hence, for any $v \in \hat{E}$, and any $x \in A^n$, $\|v - x\|_{l_q^m} \geqslant (m^{1/q}/m)\|v - x\|_{l_1^m}$. Therefore,

$$dist(B_p^m, A^n, l_q^m) \geqslant dist(\hat{E}, A^n, l_q^m) \geqslant \frac{1}{16} m^{1/q - 1/p} \tag{6}$$

which holds for any set $A^n \subset \mathbb{R}^m$ of $VC(A^n) = n \geqslant 1$, and any $1 \leqslant p,\ q \leqslant \infty$.

For the case that $1 \leqslant q \leqslant p \leqslant \infty$, the right-hand side is bounded from below by $(1/16)(m-n)^{1/q-1/p}$, true for all $n \geqslant 1$, and $m \geqslant cn$ which agrees with the theorem.

Next, we prove the lower bound of the theorem for the case of $1 \leqslant p < q \leqslant \infty$. Using the given condition of $m \geqslant cn$, we have

$$\begin{aligned} \rho_n^{\mathrm{VC}}(B_p^m, l_q^m) &= \inf_{H^n \subset \mathbb{R}^m : VC(H^n) \leqslant n} \ \sup_{x \in B_p^m} \ \inf_{y \in H^n} \|x - y\|_{l_q^m} \\ &\geqslant \inf_{H^n \subset \mathbb{R}^m : VC(H^n) \leqslant n} \ \sup_{x \in B_p^m} \ \inf_{y \in H^n} \|x - y\|_{l_q^{cn}}, \end{aligned} \tag{7}$$

where $\|\cdot\|_{l_q^{cn}}$ is the norm in $\mathbb{R}^{cn}$ which we take as the projection of $\mathbb{R}^m$ onto the first cn coordinates. The expression in (7) may be written as

$$\inf_{H^n \cap \mathbb{R}^{cn}: VC(H^n) \leqslant n} \sup_{x \in B_p^m \cap \mathbb{R}^{cn}} \inf_{y \in H^n} \|x - y\|_{l_q^{cn}}$$
$$= \inf_{H^n \subset \mathbb{R}^{cn}, VC(H^n) \leqslant n} \sup_{x \in B_p^m \cap \mathbb{R}^{cn}} \inf_{y \in H^n} \|x - y\|_{l_q^{cn}}, \tag{8}$$

where in (8) H^n runs over all subsets in $\mathbb{R}^{cn}$ rather than $\mathbb{R}^m$, of VC-dimension no larger than n. The above equality follows since for any $H^n \subset \mathbb{R}^m$, $VC(H^n \cap \mathbb{R}^{cn}) \leqslant VC(H^n) \leqslant n$, i.e., projecting a set in $\mathbb{R}^m$ onto a subspace $\mathbb{R}^{cn}$ cannot increase its VC-dimension. The expression in (8) is precisely $\rho_n^{\mathrm{VC}}(B_p^{cn}, l_q^{cn})$. From (6) it follows that $\rho_n^{\mathrm{VC}}(B_p^{cn}, l_q^{cn}) \geqslant \frac{c^{1/q-1/p}}{16} \frac{1}{n^{1/p-1/q}}$ which completes the proof of the lower bound on $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m)$.

We will need the following simple auxiliary lemma.

Lemma 3. *Define* $\hat{H}^n \subset \mathbb{R}^m$ *as*

$$\hat{H}^n = \{x \in \mathbb{R}^m : \exists i_1, \ldots, i_{m-n} \in \{1, 2, \ldots, m\}, x_{i_1} = \cdots = x_{i_{m-n}} = 0\}.$$

Then $VC(\hat{H}^n) = n$.

Proof. There does not exist an $x \in \hat{H}^n$ and an $I \subset \{1, 2, \ldots, m\}$ with $|I| = n + 1$ such that $\mathrm{sgn}(x_{|I}) = [+1, \ldots, +1]$, where $x_{|I} = [x_{i_1}, \ldots, x_{i_{n+1}}]$, $i_1, \ldots, i_{n+1} \in I$. Hence, there does not exist such I for which $|\mathrm{sgn}(\hat{H}^n_{|I})| = 2^{n+1}$ so $VC(\hat{H}^n) \leqslant n$. Now, consider the subset $A = \{x \in \mathbb{R}^m : x_{n+1} = \cdots = x_m = 0\} \subset \hat{H}^n$. Consider $I' = \{1, 2, \ldots, n\}$. Clearly, $|\mathrm{sgn}(A_{|I'})| = 2^n$. Thus, $VC(A) \geqslant n$. Hence, $VC(\hat{H}^n) \geqslant n$. It follows that $VC(\hat{H}^n) = n$. $\square$

The following proof of the upper bound of Theorem 1 is taken from Tikhomirov [17]. We include it here for completeness.

Proof of Theorem 1 (*Upper bound*). Consider an $x \in B_p^m$. Let the indices $i_1, i_2, \ldots, i_m$ be such that $|x_{i_1}| \leqslant |x_{i_2}| \leqslant \cdots |x_{i_m}|$. Define Q_n as the mapping which takes any $x \in B_p^m$ to a vector $y = Q_n x \in \hat{H}^n$ such that $y_{i_1} = y_{i_2} = \cdots = y_{i_{m-n}} = 0$, and $y_{i_j} = x_{i_j} - |x_{i_{m-n-1}}| \mathrm{sgn}(x_{i_j})$, $m - n + 1 \leqslant j \leqslant n$. For any $x \in B_p^m$,

$$\|x - Q_n x\|_{l_q^m} = \sum_{j=1}^{m-n} |x_{i_j}|^q + |x_{i_{m-n+1}}|^q \sum_{j=m-n+1}^{m} 1 \tag{9}$$
$$= \sum_{j=1}^{m-n} |x_{i_j}|^q + n |x_{i_{m-n+1}}|^q. \tag{10}$$

Set $\delta = \sum_{j=1}^{m-n} |x_{i_j}|^p$. Then $\sum_{j=m-n+1}^{m} |x_{i_j}|^p \leqslant 1 - \delta$. Hence, we have

$$n |x_{i_{m-n+1}}|^p \leqslant \sum_{j=m-n+1}^{m} |x_{i_j}|^p \leqslant 1 - \delta$$

or

$$|x_{i_{m-n+1}}| \leqslant (1-\delta)^{1/p} n^{-1/p}. \tag{11}$$

Also,

$$\sum_{j=1}^{m-n} |x_{i_j}|^q = \sum_{j=1}^{m-n} |x_{i_j}|^p |x_{i_j}|^{q-p} \leqslant \delta |x_{i_{m-n+1}}|^{q-p}. \tag{12}$$

From (10)–(12) we obtain for any $x \in B_p^m$

$$\begin{aligned} \|x - Q_n x\|_{l_q^m}^q &\leqslant \delta |x_{i_{m-n+1}}|^{q-p} + n |x_{i_{m-n+1}}|^q \\ &\leqslant (1-\delta)^{(q-p)/p} n^{(p-q)/p} \leqslant n^{(p-q)/p}, \end{aligned} \tag{13}$$

where the last inequality follows from assuming $1 \leqslant p < q \leqslant \infty$.

It follows that

$$\begin{aligned} \inf_{H^n} \sup_{x \in B_p^m} \inf_{y \in H^n} \|x - y\|_{l_q^m} &\leqslant \sup_{x \in B_p^m} \inf_{y \in \hat{H}^n} \|x - y\|_{l_q^m} \\ &\leqslant \sup_{x \in B_p^m} \|x - Q_n x\|_{l_q^m} \leqslant \frac{1}{n^{1/p - 1/q}}, \end{aligned} \tag{14}$$

where $H^n \subset \mathbb{R}^m$ is any set of VC-dimension less than or equal to n.

Now, we prove the upper bound on $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m)$ for the case of $1 \leqslant q \leqslant p \leqslant \infty$. Here linear approximation using a projection mapping onto a linear subspace is optimal as is seen next since the upper bound obtained differs only by a constant from the lower bound on $\rho_n^{\mathrm{VC}}(B_p^m, l_q^m)$. For any $x \in B_p^m$, let P_n be a projection which maps any $x \in \mathbb{R}^m$ to a $y \in \tilde{H}^n$ as follows: For $m-n+1 \leqslant j \leqslant m$, $y_j = x_j$, while for the remaining coordinates $y_1 = \cdots = y_{m-n} = 0$. Since $\tilde{H}^n$ is merely $\mathbb{R}^n$ then its $VC(\tilde{H}^n) = n$. For any $x \in B_p^m$ the approximation error becomes

$$\|x - P_n x\|_{l_q^m} = \left(\sum_{j=1}^{m-n} |x_{i_j}|^q \right)^{1/q} \leqslant (m-n)^{1/q - 1/p} \left(\sum_{j=1}^{m-n} |x_{i_j}|^p \right)^{1/p} \tag{15}$$

$$\leqslant (m-n)^{1/q-1/p} \|x\|_{l_p^m} \leqslant (m-n)^{1/q-1/p}. \tag{16}$$

This concludes the proof of the upper bound. □

4.2. *Proof of Theorem 2*

Suppose $0 < \alpha < \frac{1}{2}$. Consider any set $K \subseteq E = \{-1, +1\}^m$ of cardinality $2^{\gamma m}$, where $0 < \gamma < 1$. In a similar manner as in Lemma 1, a set $G \subset K$ may be constructed such that for every $u, v \in G$, $u \neq v$, we have $\|u - v\|_{l_1^m} > 2\alpha m$, and $|G| = 2^{c_7 m}$ where $c_7 = (\gamma + 2((1/2) - \alpha)^2 - 1)/2\,m$. In order to maximize the range for γ we introduce another parameter $\eta > 0$ and we choose henceforth $\alpha = \frac{1}{2} - \sqrt{(1-\gamma)(1+\eta)/2}$. This yields a set G which is $2\alpha m$-separated in l_1^m-norm with $|G| = 2^{\eta(1-\gamma)m/2}$, for any $\frac{1}{2} < \gamma < 1$, where the condition on $0 < \alpha < \frac{1}{2}$ implies η must satisfy $(1-\gamma)(1+\eta) < \frac{1}{2}$. We now proceed as in the lower bound proof of Theorem 1 which appeared under the case of $1 \leqslant q \leqslant p \leqslant \infty$.

Fix any set $H^n \subset \mathbb{R}^m$ with $VC(H^n) \leqslant n$ and let $\delta = dist(G, H^n, l_1^m)$, except instead of assuming $\delta \leqslant m/16$ we assume $\delta \leqslant \alpha m/4$ where α is as chosen above. As in (3) this leads to a contradiction of the inequality $2^{\eta(1-\gamma)m/2} \leqslant e(n+1)(8e)^n$ for all $n \geqslant 1$ provided that $m \geqslant (4n/(\eta(1-\gamma))\log_2(8e)$. To convert from the l_1^m-norm to l_q^m-norm we use (5) and obtain

$$\inf_{H^n} dist(K, H^n, l_q^m) \geqslant \frac{\alpha m^{1/q}}{4} \geqslant \frac{\alpha(m-n)^{1/q}}{4}. \tag{17}$$

Substituting for α we then obtain

$$\inf_{H^n} \sup_{v \in K} \inf_{y \in H^n} \|v - y\|_{l_q^m} \geqslant \left(\frac{1}{8} - \sqrt{\frac{(1-\gamma)(1+\eta)}{32}}\right)(m-n)^{1/q} \tag{18}$$

which concludes the proof of the lower bound on $\rho_n^{\mathrm{VC}}(K, l_q^m)$.

The upper bound easily follows from the upper bound of Theorem 1 for the case $1 \leqslant q < p = \infty$ since $K \subset E \subset B_\infty^m$. Thus, we have

$$\rho_n^{\mathrm{VC}}(K, l_q^m) \leqslant \rho_n^{\mathrm{VC}}(B_\infty^m, l_q^m) \leqslant (m-n)^{1/q} \tag{19}$$

which therefore proves Theorem 2. □

4.3. Proof of Corollary 1

Since for any set $F \subset \mathbb{R}^m$, $\rho_n^{\mathrm{VC}}(F, l_q^m) \leqslant \rho_n^P(F, l_q^m)$, the lower bounds of Theorems 1 and 2 hold also for $\rho_n^P(B_p^m, l_q^m)$ and $\rho_n^P(K, l_q^m)$, respectively. For the upper bounds the case is similar since the particular classes $\hat{H}^n$ and $\tilde{H}^n$ of VC-dimension n, which are used for the approximation, have a pseudo-dimension n (proven below). Their rate of approximation of B_p^m upper bounds the degree of approximation by any class of pseudo-dimension n and hence upper bounds $\rho_n^P(B_p^m, l_q^m)$. As in (19), the upper bound on $\rho_n^P(K, l_q^m)$ is $\rho_n^P(B_\infty^m, l_q^m)$.

We now prove that the pseudo-dimension of the set $\hat{H}^n$ defined in Lemma 3 is n. Suppose that it is greater than n. Then there exists an index set $I = \{i_1, \ldots, i_{n+1}\} \subset \{1, 2, \ldots, m\}$ such that $\mathrm{sgn}(\hat{H}^n_{|I,y}) = \{-1, +1\}^{n+1}$. We have $\hat{H}^n_{|I,y} = \bigcup_{j=1}^{n+1} \{[x_{i_1} + y_1, \ldots, x_{i_{j-1}} + y_{j-1}, y_j, x_{i_{j+1}} + y_{j+1}, \ldots, x_{i_{n+1}} + y_{n+1}] : x \in \hat{H}^n\}$ since for any $x \in \hat{H}^n$ there are only n non-zero elements. It follows that the set of boolean vectors $\mathrm{sgn}(\hat{H}^n_{|I,y}) = \bigcup_{j=1}^{n+1} [\pm 1, \ldots, \pm 1, \mathrm{sgn}(y_{i_j}), \pm 1, \ldots, \pm 1]$. Clearly, the boolean vector $-\mathrm{sgn}(y) = -[\mathrm{sgn}(y_1), \ldots, \mathrm{sgn}(y_{n+1})] \notin \mathrm{sgn}(\hat{H}^n_{|I,y})$. Hence, there does not exist an I and y such that $|\mathrm{sgn}(\hat{H}^n_{|I,y})| = 2^{n+1}$. This proves $dim_p(\hat{H}^n) \leqslant n$. We also have $dim_p(\hat{H}^n) \geqslant VC(\hat{H}^n) = n$. It follows that $dim_p(\hat{H}^n) = n$. The same kind of proof is used to show that $dim_p(\tilde{H}^n) = n$. □

4.4. Proof of Corollary 2

For any vectors $u, v \in \mathbb{R}^m$ we have $\|u - v\|_{l_q^m(\mu)} = (\sum_{i \in I} \mu(i)|v_i - u_i|^q)^{1/q} \geqslant \mu_{\min}^{1/q} \|v - u\|_{l_q^m}$. Also, $\|u - v\|_{l_q^m(\mu)} = (\sum_{i \in I} \mu(i)|v_i - u_i|^q)^{1/q}$ which by Holder's inequality is bounded from above by $\mu_{\max}^{1/q} \|v - u\|_{l_q^m}$. Now, provided that the set I, which depends on the probability measure μ, has cardinality greater than n then for any set $A \subset \mathbb{R}^m$, $\rho_n^{\mathrm{VC}}(A, l_q^m(\mu)) = \inf_{H^n} \sup_{x \in A} \inf_{y \in H^n} \|x - y\|_{l_q^m(\mu)} \geqslant \mu_{\min}^{1/q} \inf_{H^n} \sup_{x \in A} \inf_{y \in H^n} \|x - y\|_{l_q^m} = \mu_{\min}^{1/q} \rho_n^{\mathrm{VC}}(A, l_q^m)$ and $\rho_n^{\mathrm{VC}}(A, l_q^m(\mu)) \leqslant \mu_{\max}^{1/q} \rho_n^{\mathrm{VC}}(A, l_q^m)$. If the cardinality of $|I| \leqslant n$ then $\rho_n^{\mathrm{VC}}(A, l_q^m(\mu)) = 0$ since then effectively the set $A \subset \mathbb{R}^{|I|} \subseteq \mathbb{R}^n$, the latter having a VC-dimension n.

Acknowledgements

J. Ratsaby acknowledges the support of a VATAT Post-Doctorate fellowship and the Ollendorff center of the Department of Electrical Engineering at the Technion. V. Maiorov acknowledges the support of the center for absorption in science of the ministry of immigrant absorption, state of Israel. He also acknowledges the support of Allan Pinkus of the Department of Mathematics, Technion. The authors thank the reviewers for interesting and informative comments.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimensions, uniform cinvergence, and learnability, Proc. 34th Annual Symp. on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 1993, pp. 292–301.

[2] A. Barron, Neural net approximation, in: Proc. 7th Yale Workshop on Adaptive and Learning Systems, 1992.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, J. ACM 36(4) (1989) 929–965.

[4] R.A. DeVore, Degree of nonlinear approximation, in: C.K. Chui, L.L. Schumaker, J.D. Ward (Eds.), Approximation Theory VI, vol. 1, pp. 175–201.

[5] F. Girosi, Approximation error bounds that use VC-Bounds, in: Proc. Internat. Conf. Artificial Neural Networks, Paris, 1995.

[6] L. Gurvits, P. Koiran, Approximation and learning of convex superpositions, in: Proc. of Computational Learning Theory: EuroCOLT-95, Springer, Berlin, 1995, pp. 222–236.

[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 (1995) 217–232.

[9] G.G. Lorentz, M.V. Golitschek, Y. Makovoz, Constructive Approximation, Advanced Problems, Springer, Berlin, 1996.

[10] V.E. Maiorov, On best approximation by ridge functions, J. Approx. Theory, June 1996, submitted.

[11] V.E. Maiorov, J. Ratsaby, On the degree of approximation using manifolds of finite pseudo-dimension, J. Constr. Approx., submitted.

[12] A. Pinkus, n-widths in Approximation Theory, Springer, New York, 1985.

[13] D. Pollard, Convergence of Stochastic Processes, Springer Series in Statistics. Springer, Berlin, 1984.

[14] D. Pollard, Empirical processes: theory and applications, NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 2., Institute of Mathmatics, Stat. and Am. Stat. Assoc., providence, RI, 1989.
[15] J. Ratsaby, V. Maiorov, On the value of partial information for learning from examples, J. Complexity (1997) to appear.
[16] J. Ratsaby, V. Maiorov, Generalization of the PAC-model for learning with partial information, in: Proc. 3rd European Conf. on Computational Learning Theory, EuroCOLT 97, Springer, Berlin, 1997.
[17] V.M. Tikhomirov, Some Problems in Approximation Theory, Moscow State University, Moscow, 1976 (In Russian).
[18] L.G. Valiant, A theory of the learnable, Comm. ACM 27 (11) (1984) 1134–1142.
[19] A.W. Van Der Vaart, A.J. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Springer Series in Statistics. Springer, Berlin, 1946.
[20] V.N. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, Berlin, 1982.
[21] V.N. Vapnik, A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (2) (1971) 264–280.
[22] V.N. Vapnik, A.Ya. Chervonenkis, Necessary and sufficient conditions for the uniform convergence of means to their expectations, Theory Probab. Appl. 26 (3) (1981) 532–553.
[23] H.E. Warren, Lower bounds for approximation by non-linear manifolds, Trans. AMS 133 (1968) 167–178.