Available online at www.sciencedirect.com
ScienceDirect
Discrete Applied Mathematics 155 (2007) 1930–1942
DISCRETE APPLIED MATHEMATICS
www.elsevier.com/locate/dam

Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets

Věra Kůrková[a,1,2], Marcello Sanguineti[b,*,1,3]

[a] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic
[b] Department of Communications, Computer, and System Sciences (DIST), University of Genova, Via Opera Pia 13, 16145 Genova, Italy

Received 25 July 2006; received in revised form 16 November 2006; accepted 11 April 2007
Available online 18 April 2007

Abstract

Covering numbers of precompact symmetric convex subsets of Hilbert spaces are investigated. Lower bounds are derived for sets containing orthogonal subsets with norms of their elements converging to zero sufficiently slowly. When these sets are convex hulls of sets with power-type covering numbers, the bounds are tight. The arguments exploit properties of generalized Hadamard matrices. The results are illustrated by examples from machine learning, neurocomputing, and nonlinear approximation.

Keywords: Symmetric convex hulls; Lower bounds on covering numbers; Power-type covering numbers; Generalized Hadamard matrices; Minkowski functional

1. Introduction

Covering numbers, introduced by Kolmogorov [24], play an important role in a variety of areas, such as density estimation [6,14], empirical processes [36], machine learning [1,17,40,42,45,46], eigenvalue estimation [8,11,16], and Gaussian processes [28,31].

Covering numbers have been studied in ambient spaces with various metrics. For example, with the metrics induced by the supremum norm [2, Chapter 10, 13] and the $\mathscr{L}_1$-norm [2, Chapter 17], they were used in statistical learning theory to estimate sample errors. With the metric induced by the $\mathscr{L}_2$-norm, covering numbers were investigated in machine learning [2, Section 18.5], probability [15], approximation [32], convex geometry [34], mathematical theory of neural networks [32], and to derive bounds on $\mathscr{L}_1$-covering numbers [5]. (The list of references in this paragraph is by no means complete.)

* Corresponding author.
E-mail addresses: vera@cs.cas.cz (V. Kůrková), marcello@dist.unige.it (M. Sanguineti).
[1] Collaboration between V.K. and M.S. was supported by the 2007–2009 Scientific Agreement among University of Genova, National Research Council of Italy, and Czech Academy of Sciences.
[2] Partially supported by the Ministry of Education of the Czech Republic, project Center of Applied Cybernetics 1M684077004 (1M0567).
[3] Partially supported by a PRIN Grant from the Italian Ministry of University and Research, Project "New Techniques for the Identification and Adaptive Control of Industrial Systems".

0166-218X/$ - see front matter
doi:10.1016/j.dam.2007.04.007

Various authors studied the dependence of covering numbers of convex hulls on covering numbers of sets generating them (e.g., [7,9,10,12,19,20,29,33,41]) and derived estimates via entropy numbers of operators (e.g., [38,39,45]).

In contrast, our approach is based on exploitation of suitable properties of orthogonal subsets of convex sets. A precompact subset of a Hilbert space cannot contain an infinite orthogonal subset with the magnitudes of the norms of its elements bounded from below. But it may contain an infinite orthogonal subset with the magnitudes of the norms converging to zero rather slowly. We show that the slower the rate of convergence, the larger the lower bound on covering numbers of the convex hull of the precompact set. Even when a precompact set does not contain such an orthogonal subset, it may contain a sequence of finite orthogonal subsets of increasing cardinality with minima of norms of their elements converging to zero. Also in this case, we show that the faster the increase of cardinality of the orthogonal sets in the sequence, the larger the lower bound on covering numbers of the convex hull of the precompact set. For the symmetric convex hulls of sets with power-type covering numbers (in particular, sets of finite Vapnik–Chervonenkis (VC)-dimension), the bounds that we derive are tight.

We illustrate our results by examples from machine learning, neurocomputing, and nonlinear approximation. We show that balls in certain variational norms generated by computational units called perceptrons are precompact and satisfy assumptions implying our tight estimates. This allows us to extend a result by Makovoz [32] disproving the possibility of a substantial improvement of a bound on approximation rates by certain perceptron neural networks. Makovoz's [32] estimate is based on a result by Lorentz [30], while our proofs take advantage of the exponential growth of the size of generalized Hadamard matrices [23] (which differ from the classical ones in allowing a tolerance in the orthogonality condition).

The paper is organized as follows. Section 2 introduces notations and definitions. Section 3 gives lower bounds on covering numbers of symmetric convex precompact subsets of Hilbert spaces in terms of rates of decay of norms of their orthogonal subsets and includes examples of such sets. It is also shown that for symmetric convex hulls of sets with power-type covering numbers (such as sets with finite VC-dimension) our lower bounds are tight. Proofs of the bounds are given in Section 4. Section 5 applies estimates from the previous sections to neurocomputing and Section 6 uses them to derive tightness results on rates of nonlinear approximation. Section 7 is a brief discussion.

2. Preliminaries

By $\mathbb{R}$ and $\mathbb{R}_+$ are denoted the sets of real and positive real numbers, resp., and by $\mathbb{N}$ and $\mathbb{N}_+$ the sets of natural numbers and positive integers, resp. For a positive integer d, ℓ_1^d and ℓ_2^d denote the ℓ_1- and ℓ_2-norms on $\mathbb{R}^d$, resp. Sequences are denoted by $\{s_i\} = \{s_i | i \in \mathbb{N}_+\}$. For $f, g : \mathbb{R}_+ \to \mathbb{R}$, we write

$$g(\varepsilon) \leqslant f(\varepsilon) \quad \text{for } \varepsilon \downarrow 0$$

when there exists $c > 0$ such that for every decreasing sequence $\{\varepsilon_i\}$ of positive real numbers with $\lim_{i\to+\infty}\varepsilon_i = 0$ one has $g(\varepsilon_i) \leqslant c\, f(\varepsilon_i)$ for all positive integers i. When both $g(\varepsilon) \leqslant f(\varepsilon)$ for $\varepsilon \downarrow 0$ and $f(\varepsilon) \leqslant g(\varepsilon)$ for $\varepsilon \downarrow 0$, we write

$$g(\varepsilon) \sim f(\varepsilon) \quad \text{for } \varepsilon \downarrow 0.$$

Let $(X, \|.\|)$ be a normed linear space, $f \in X$, and $r > 0$. By $B_r(f, \|.\|)$ is denoted the closed ball of radius r in the norm $\|.\|$ centered at $f \in X$, i.e.,

$$B_r(f, \|.\|) = \{h \in X | \|h - f\| \leqslant r\}.$$

We write $B_r(\|.\|)$ instead of $B_r(0, \|.\|)$.

For a positive integer d and a set $\Omega \subseteq \mathbb{R}^d$, $(\mathscr{L}_2(\Omega), \|.\|_2)$ denotes the Hilbert space of real-valued, square-integrable functions on Ω with the $\mathscr{L}_2$-norm denoted by $\|.\|_2$.

For a subset G of $(X, \|.\|)$, $\operatorname{cl} G$ denotes its *closure* with respect to the topology generated by the norm $\|.\|$ and $\operatorname{conv} G$ is its *convex hull*, i.e.,

$$\operatorname{conv} G = \left\{ \sum_{i=1}^{n} a_i g_i \,\middle|\, a_i \in [0, 1], \sum_{i=1}^{n} a_i = 1, g_i \in G, n \in \mathbb{N}_+ \right\}.$$

For a positive integer n we denote

$$\mathrm{conv}_n\, G = \left\{ \sum_{i=1}^{n} a_i g_i \,\middle|\, a_i \in [0,1],\ \sum_{i=1}^{n} a_i = 1,\ g_i \in G \right\}.$$

For $G \subseteq (X, \|.\|)$ and $\varepsilon > 0$, $\{g_1, \ldots, g_m\} \subseteq G$ is called an *ε-net in G* if the family of closed balls of radii ε centered at g_i covers G, i.e., if $G \subseteq \bigcup_{i=1}^m B_\varepsilon(g_i, \|.\|)$, and $\{g_1, \ldots, g_m\}$ is called *ε-separated* if for each distinct pair $i, j \in \{1, \ldots, m\}$, $\|g_i - g_j\| \geqslant \varepsilon$. If a set G contains a 2ε-separated subset of size m, then every ε-net in G must contain at least m elements.

The *ε-covering number* of a subset G of $(X, \|.\|)$ is the cardinality of a minimal ε-net in G, i.e.,

$$\mathcal{N}(G, \|.\|, \varepsilon) = \min \left\{ m \in \mathbb{N}_+ \mid \exists f_1, \ldots, f_m \in G \text{ such that } G \subseteq \bigcup_{i=1}^{m} B_\varepsilon(f_i, \|.\|) \right\}.$$

If the set over which the minimum is taken is empty, then $\mathcal{N}(G, \|.\|, \varepsilon) = +\infty$. Note that we consider covering numbers defined in terms of closed balls as in [10,44], but some authors (e.g., [2, p. 148]) use open balls.

When we use covering numbers of balls in another norm than the one on the ambient normed linear space, we include the norm into the notation $\mathcal{N}(G, \|.\|, \varepsilon)$, otherwise we write merely $\mathcal{N}(G, \varepsilon)$.

When there exists $\beta > 0$ such that $\mathcal{N}(G, \varepsilon) \leqslant (1/\varepsilon)^\beta$ for $\varepsilon \downarrow 0$, G is said to have *power-type covering numbers*.

The closed symmetric convex hull of a bounded subset G of a normed linear space $(X, \|.\|)$ generates a norm via its Minkowski functional [37, p. 25]. This norm, called *G-variation* and denoted by $\|.\|_G$, is defined as

$$\|f\|_G = \inf \left\{ c \in \mathbb{R}_+ \,\middle|\, \frac{f}{c} \in \ \mathrm{cl}\ (\mathrm{conv}(G \cup -G)) \right\},$$

where the closure is taken with respect to the ambient space norm $\|.\|$. G-variation was used in [25] as an extension of the concept of variation with respect to characteristic functions of half-spaces from [3].

Balls in G-variation play an important role in machine learning. For their elements, rates of approximation by linear combinations of n elements of G are bounded from above by $rn^{-1/2}$ [3,4,22,35], where r is the radius of the ball. By the definition, the unit ball in G-variation is the closure in the norm $\|.\|$ of the symmetric convex hull of G, i.e.,

$$B_1(\|.\|_G) = \mathrm{cl}\ (\mathrm{conv}\ (G \cup -G)). \tag{1}$$

It is easy to check that for every G and every $\varepsilon > 0$

$$\mathcal{N}\ (B_1(\|.\|_G), \varepsilon) = \mathcal{N}(\mathrm{conv}\ (G \cup -G), \varepsilon), \tag{2}$$

where the covering number is considered with respect to the norm $\|.\|$ of the ambient space.

By $\mathcal{H}$ is denoted the *binary entropy function*, defined for every $p \in (0, 1)$ as

$$\mathcal{H}(p) = -p \log_2(p) - (1-p) \log_2(p-1).$$

3. Lower bounds

For a subset A of a normed linear space $(X, \|.\|)$ and a positive integer r, we denote

$$A_r = \left\{ f \in A \,\middle|\, \|f\| \geqslant \frac{1}{r} \right\}.$$

The larger the sets A_r, the slower the decrease of the norms of the elements of A.

Definition 3.1. When A_r is finite for all positive integers r, the function $\alpha_A : \mathbb{N}_+ \to \mathbb{N}_+$ defined as

$$\alpha_A(r) = \mathrm{card}\ A_r$$

is called the *decay function of A*.

Definition 3.2. A set A such that A_r is finite for all positive integers r is called *slowly decaying with respect to* γ if there exists $\gamma > 0$ such that $\alpha_A(r) = r^\gamma$.

Note that if A is a precompact subset of a Hilbert space and A_r is orthogonal, then A_r must be finite. Thus decay functions are defined for all precompact orthogonal subsets of Hilbert spaces and also for subsets $A = \bigcup_{r=1}^\infty A_r$ with all A_r orthogonal but A not necessarily orthogonal.

Definition 3.3. A set A formed by d-variable functions with the decay function $\alpha_A(r) = r^d$ is called *slowly decaying*.

Under a slightly different name, the concept of a slowly decaying set was introduced in [27] to compare worst-case errors in linear and neural-network approximation.

Example 3.4. The set $A = \{n^{-1/\gamma} e_n\}$, where $\{e_n\}$ is the standard orthonormal basis of ℓ_2 and $\gamma > 0$ (investigated in [10, p. 886]), is an orthogonal precompact subset of ℓ_2 and its decay function is $\alpha_A(r) = r^\gamma$, so A is slowly decaying with respect to γ.

Example 3.5. Let $A = \bigcup_{r=1}^\infty A_r$ with $A_r = \{n^{-1/\gamma} \rho_r(e_n) | n = 1, \ldots, r^\gamma\}$, where $\{\rho_r\}$ is a sequence of distinct rotations of ℓ_2. This subset of ℓ_2 is slowly decaying with respect to γ but it is not orthogonal as soon as one of the rotations is not the identity.

Example 3.6. The precompact subset $A = \bigcup_{r=1}^\infty A_r$ of $(\mathscr{L}_2([0,1]^d), \|.\|_2)$, where $A_r = \{h_v | v = (v_1, \ldots, v_d) \in \{1, \ldots, r\}^d\}$, $h_v = c_v \sin(\pi v \cdot x)$, and $c_v = d\sqrt{2} / \sum_{k=1}^d v_k$, is slowly decaying with respect to the number d of variables (its decay function is r^d). Indeed, it is easy to check that for each $h_v = c_v \sin(\pi v \cdot x) \in A_r$, $\|h_v\|_2 \geqslant \frac{c_v}{\sqrt{2}} = d\sqrt{2}/\sqrt{2}\sum_{k=1}^d v_k \geqslant \frac{1}{r}$.

The following theorem estimates from below covering numbers of symmetric convex precompact subsets of infinite-dimensional Hilbert spaces in terms of decay functions of their nearly orthogonal (and in particular orthogonal) subsets.

Definition 3.7. For $\delta \geqslant 0$, a finite subset $A = \{g_1, \ldots, g_m\}$ of a Hilbert space $(X, \|.\|)$ with inner product $\langle \cdot, \cdot \rangle$ is called *δ-nearly orthogonal* if

$$\sum_{i,j=1, j \neq i}^{m} |\langle g_i, g_j \rangle| \leqslant \delta.$$

Note that for $\delta = 0$ the set A is orthogonal.

Theorem 3.8. *Let $(X, \|.\|)$ be a Hilbert space, F a symmetric convex subset containing an infinite set $A = \bigcup_{r=1}^\infty A_r$ with the decay function α_A such that for every positive integer r, $\alpha_A(r) \geqslant 3$ and A_r is δ_r-nearly orthogonal with $\delta_r \leqslant 1/r^2$, and $b = 1 - \mathscr{H}(\frac{1}{4}) \simeq 0.085$, where $\mathscr{H}$ denotes the binary entropy function. Then for every positive integer r*

$$b\alpha_A(r) - 1 \leqslant \log_2 \mathscr{N}\left(F, \frac{1}{2r}\sqrt{\frac{1 - r^2\delta_r}{\alpha_A(r)}}\right).$$

The proof of Theorem 3.8 is based on properties of generalized Hadamard matrices, is given in Section 4.

Example 3.9. Let $F = \mathrm{conv}(A \cup -A)$, where $A = \{n^{-1/\gamma} e_n\}$ is the subset of ℓ_2 considered in Example 3.4 with the decay function $\alpha_A(r) = r^\gamma$. By Theorem 3.8 with $\delta_r = 0$ for all r

$$br^\gamma - 1 \leqslant \log_2 \mathscr{N}(F, \tfrac{1}{2} r^{-(\gamma+2)/2}). \tag{3}$$

Covering numbers of the set F were investigated in [10, p. 886], where for all positive integers γ the tight bounds

$$\log_2 \mathcal{N}(F, c_1 r^{-(\gamma+2)/2}) \leqslant r^\gamma - 1 \quad \text{and} \quad r^\gamma - 1 \leqslant \log_2 \mathcal{N}(F, c_2 r^{-(\gamma+2)/2}), \tag{4}$$

with c_1 and c_2 constants, were derived. So for the set F the lower bound (3) is up to constants the same as the asymptotically tight bound (4).

Example 3.10. Let $F = \text{conv}(A \cup -A)$, where $A = \bigcup_{r=1}^{\infty} A_r$ is a subset of $(\mathcal{L}_2([0,1]^d), \|.\|_2)$ with $A_r = \{h_v | v = (v_1, \ldots, v_d) \in \{1, \ldots, r\}^d\}$, $h_v = c_v \sin(\pi v \cdot x)$, and $c_v = d\sqrt{2}/\sum_{k=1}^d v_k$. By Theorem 3.8 with $\delta_r = 0$ for all r

$$br^d - 1 \leqslant \log_2 \mathcal{N}\left(F, \frac{1}{2r^{d/2+1}}\right).$$

For the special case of sets containing subsets slowly decaying with respect to $\gamma > 0$, the next asymptotic estimate holds.

Corollary 3.11. *Let $(X, \|.\|)$ be a Hilbert space, F its symmetric convex subset containing for some $t > 0$ a set tA, where $A = \bigcup_{r=1}^{\infty} A_r$ with all A_r orthogonal, A slowly decaying with respect to $\gamma > 0$, and $b = 1 - \mathcal{H}(\frac{1}{4})$, where $\mathcal{H}$ denotes the binary entropy function. Then*

$$\left(\frac{1}{\varepsilon}\right)^{2\gamma/(\gamma+2)} - 1 \leqslant \log_2 \mathcal{N}(F, \varepsilon) \quad \textit{for } \varepsilon \downarrow 0.$$

The next theorem exploits the upper bound derived in [10, Proposition 5.1] to show that the estimate from Corollary 3.11 is tight for convex hulls of sets with power-type covering numbers.

Theorem 3.12. *Let $(X, \|.\|)$ be a Hilbert space, G a precompact subset of its unit ball such that there exist $t, \gamma, \beta > 0$ with $\mathcal{N}(G, \varepsilon) \leqslant (1/\varepsilon)^\beta$ for $\varepsilon \downarrow 0$, and $\text{conv}(G \cup -G) \supseteq tA$, where $A = \bigcup_{r=1}^{\infty} A_r$ with all A_r orthogonal and A slowly decaying with respect to γ. Then*

$$\left(\frac{1}{\varepsilon}\right)^{2\gamma/(\gamma+2)} \leqslant \log_2 \mathcal{N}(\text{conv}(G \cup -G), \varepsilon) \leqslant \left(\frac{1}{\varepsilon}\right)^{2\beta/(\beta+2)} \quad \textit{for } \varepsilon \downarrow 0.$$

Theorem 3.12 shows that if G has power-type covering numbers with an exponent β, then its symmetric convex hull cannot contain an orthogonal set slowly decaying with respect to $\gamma > \beta$. When β and γ are close to each other, Theorem 3.12 gives a tight estimate. In particular, when $\beta = \gamma$ we get

$$\log_2 \mathcal{N}(\text{conv}(G \cup -G), \varepsilon) \sim \left(\frac{1}{\varepsilon}\right)^{2\gamma/(\gamma+2)} \quad \text{for } \varepsilon \downarrow 0.$$

Sets of functions with finite VC-dimension have power-type covering numbers [43]. For a set G of $\{0, 1\}$-valued functions defined on a set Ω and $S \subset \Omega$, we denote by $G|_S$ the set of functions from G restricted to S. Functions from S to $\{0, 1\}$ are called *dichotomies*. If $G|_S$ contains all dichotomies, then G is said to *shatter* S. The *VC-dimension* of G, denoted by $VC(G)$, is the cardinality of the largest subset S of Ω that is shattered by G; if the largest set is infinite, then $VC(G) = \infty$.

The next corollary shows that symmetric convex hulls of sets of finite VC dimension cannot contain orthogonal subsets slowly decaying with respect to the VC-dimension of the generating set.

Corollary 3.13. *Let $(X, \|.\|)$ be a Hilbert space and G a precompact subset of its unit ball such that G contains only $\{0, 1\}$-valued functions, $VC(G) = v < \infty$, and $t, \gamma > 0$ such that $\text{conv}(G \cup -G) \supseteq tA$, where A is an orthogonal set slowly decaying with respect to γ. Then*

$$\left(\frac{1}{\varepsilon}\right)^{2\gamma/(\gamma+2)} \leqslant \log_2 \mathcal{N}(\text{conv}(G \cup -G), \varepsilon) \leqslant \left(\frac{1}{\varepsilon}\right)^{2v/(v+1)} \quad \textit{for } \varepsilon \downarrow 0.$$

4. Proofs of the lower bounds

To prove Theorem 3.8 and Corollary 3.11, we construct ε-separated subsets of symmetric convex hulls of orthogonal sets using coefficient vectors obtained from "large" sets of quasiorthogonal vectors from the *Hamming cube* $\{-1,+1\}^m$. Recall that a *Hadamard matrix of order m* is a matrix with m columns, entries equal to $+1$ or -1, and each pair of distinct rows orthogonal. The concept of Hadamard matrix has been generalized in [23] by allowing a tolerance in the orthogonality condition.

Definition 4.1. For $\varepsilon \in (0,1]$, an *ε-Hadamard matrix of order m* is a matrix with m columns, entries equal to $+1$ or -1, and the inner products of any two distinct rows less than or equal to $m\varepsilon$.

Let

$$R(\varepsilon, m)$$

denote the *maximal number of rows of an ε-Hadamard matrix of order m*. If $\varepsilon = s/m$ for a positive integer s, M is the matrix for which the maximum is reached, and T_M is the set of its row vectors, then for each pair of distinct vectors $u, v \in T_M$,

$$|u \cdot v| \leqslant \varepsilon m = s,$$

where "·" denotes the Euclidean inner product. The weakened orthogonality condition can also be described in terms of *Hamming distance*, denoted by h and defined on $\{-1,1\}^m$ as the number of coordinates at which two vectors differ. The Hamming distance of two vectors $u, v \in \{-1,1\}^m$ is equal to $\frac{1}{2}$ of the ℓ_1^m-norm of the vector $u - v$, i.e.,

$$h(u,v) = (1/2)\sum_{i=1}^{m} |u_i - v_i|.$$

It is easy to check that the Hamming distance of two vectors $u, v \in T_M$, where M is an ε-Hadamard matrix of order m, satisfies

$$h(u,v) \geqslant m(1-\varepsilon)/2.$$

In particular, for $\varepsilon = s/m$ one has

$$h(u,v) \geqslant (m-s)/2. \tag{5}$$

The next lemma gives lower bounds on covering numbers of convex symmetric sets in terms of the cardinality of their nearly orthogonal or orthogonal subsets with minima of magnitudes of norms of their elements bounded from below. For a real number s, we denote by $\lceil s \rceil$ the smallest integer $n \geqslant s$ and by $\lfloor s \rfloor$ the largest integer $n \leqslant s$. We also denote

$$B(\lambda, m) = \frac{\lambda!}{m!(\lambda - m)!}.$$

Lemma 4.2. *Let F be a convex symmetric subset of a Hilbert space* $(X, \|.\|)$ *such that F contains for some* $\delta \geqslant 0$ *a δ-nearly orthogonal subset A with* card $A = m$, $\min_{g\in A}\|g\| = a$, *and* $\delta \leqslant a^2$. *Then the following estimates hold*:

(i) *for every positive integer s such that* $1 \leqslant s < m$,

$$R\left(\frac{s}{m}, m\right) \leqslant \mathcal{N}\left(F, \frac{\sqrt{a^2-\delta}}{m}\sqrt{\left\lceil \frac{m-s}{2} \right\rceil}\right);$$

(ii) *for every positive integer s such that* $1 \leqslant s \leqslant m-2$,

$$\frac{2^{m-1}}{B(\lambda_{m,s}, m)} \leqslant \mathcal{N}\left(F, \frac{\sqrt{a^2-\delta}}{m}\sqrt{\left\lceil \frac{m-s}{2} \right\rceil}\right);$$

(iii) *for* $m \geqslant 3$,

$$b\,m - 1 \leqslant \log_2 \mathcal{N}\left(F, \frac{1}{2}\sqrt{\frac{a^2-\delta}{m}}\right),$$

where $b = 1 - \mathcal{H}(\frac{1}{4}) \simeq 0.085$ *and* $\mathcal{H}$ *denotes the binary entropy function.*

Proof. (i) Let $A = \{g_1, \ldots, g_m\}$, M be an (s/m)-Hadamard matrix of order m with $R(s/m, m)$ rows, T_M the set of its row vectors, $A(M) = \{\frac{1}{m}\sum_{i=1}^{m} u_i g_i | u_i \in T_M\}$, and $\varepsilon_s = \frac{\sqrt{a^2-\varepsilon}}{m}\sqrt{\lceil \frac{m-s}{2} \rceil}$. We show that $A(M_s)$ is $2\varepsilon_s$-separated. For any pair of distinct vectors $u, v \in T_M$, we first estimate from below the distance $\|\frac{1}{m}\sum_{i=1}^{m} u_i g_i - \frac{1}{m}\sum_{i=1}^{m} v_i\, g_i\|$. Let I denote the set of coordinates at which u and v differ, $k = \mathrm{card}\, I$, and $\zeta_i = \frac{1}{2\sqrt{k}}(u_i - v_i)$, $i \in I$. Then $\zeta_i = \pm\frac{1}{\sqrt{k}}$, $\|\frac{1}{m}\sum_{i=1}^{m}(u_i - v_i) g_i\| = \frac{1}{m}\|\sum_{i \in I} g_i\| = \frac{2\sqrt{k}}{m}\|\sum_{i=1}^{k}\zeta_i\, g_i\|$, and $\|\sum_{i=1}^{k}\zeta_i\, g_i\|^2 = |\sum_{i=1}^{k}\sum_{j=1}^{k}\zeta_i\zeta_j g_i \cdot g_j|$. Since $\sum_{i=1}^{k}\zeta_i^2 = 1$, it is sufficient to derive a lower bound on the function $\Delta(\zeta_1, \ldots, \zeta_k) = |\sum_{i=1}^{k}\sum_{j=1}^{k}\zeta_i\zeta_j g_i \cdot g_j|$ on the unit sphere S_1 in the l_2-norm on $\mathbb{R}^k$. Let D_I be the $k \times k$ matrix defined by $D_{I\,ij} = g_i \cdot g_j$. Then $\Delta(\zeta_1, \ldots, \zeta_k) \geqslant \sqrt{|\lambda_{\min}(D_I)|}$ in S_1, where $\lambda_{\min}(D_I)$ denotes the minimum eigenvalue of D_I. As $|\lambda_{\min}(D_I)| \geqslant \left|\min_{g_i \in A}\|g_i\|^2 - \sum_{i \in I, i \neq j}|g_i \cdot g_j|\right| \geqslant a^2 - \delta$, we get $\frac{1}{m}\|\sum_{i=1}^{m}(u_i - v_i) g_i\| \geqslant \frac{2\sqrt{k(a^2-\delta)}}{m} \geqslant \frac{2\sqrt{a^2-\delta}}{m}\sqrt{\lceil \frac{m-s}{2} \rceil}$.

(ii) follows from (i) combined with the lower bound $R\,(s/m, m) \geqslant 2^{m-1}/B(\lambda_{m,s}, m)$ from [23, Theorem 3.4].

(iii) Let $\varepsilon_s = \frac{\sqrt{a^2-\delta}}{m}\sqrt{\lceil \frac{m-s}{2} \rceil}$. From (ii) with $s = \lfloor \frac{m}{2} \rfloor$, we get

$$\varepsilon_s = \frac{\sqrt{a^2-\delta}}{m}\sqrt{\left\lceil \frac{m - \lfloor \frac{m}{2} \rfloor}{2} \right\rceil} \geqslant \frac{\sqrt{a^2-\delta}}{m}\sqrt{\left\lceil \frac{m - \frac{m}{2}}{2} \right\rceil} \geqslant \frac{\sqrt{a^2-\delta}}{m}\sqrt{\frac{m}{4}} = \frac{\sqrt{a^2-\delta}}{2\sqrt{m}}$$

and

$$\mathcal{N}\left(F, \frac{1}{2}\sqrt{\frac{a^2-\delta}{m}}\right) \geqslant \mathcal{N}\left(F, \delta_{\lfloor m/2 \rfloor}\right) \geqslant 2^{m-1}/B(\lambda_{m,\lfloor m/2 \rfloor}, m).$$

As

$$\lambda_{m,\lfloor m/2 \rfloor} = \left\lceil \frac{m - \lfloor \frac{m}{2} \rfloor - 2}{2} \right\rceil = \left\lceil \frac{\frac{m}{2} - 2}{2} \right\rceil \leqslant \frac{m}{4},$$

we can use the estimate $B(\lambda, m) \leqslant 2^{m\mathcal{H}(\lambda/m)}$ from [18, p. 44], which is valid for $\lambda < m/2$. Finally, as the entropy function $\mathcal{H}$ is increasing over the interval $(0, \frac{1}{2})$ we get

$$\mathcal{N}\left(F, \frac{1}{2}\sqrt{\frac{a^2-\delta}{m}}\right) \geqslant \frac{2^{m-1}}{2^{m\mathcal{H}\left(\frac{\lambda_{m,\lfloor m/2 \rfloor}}{m}\right)}} \geqslant 2^{m-1}\,2^{-m\mathcal{H}(1/4)} = 2^{m(1-\mathcal{H}(1/4))-1} = 2^{mb-1}. \qquad \square$$

Using Lemma 4.2 we now prove Theorem 3.8 and Corollary 3.11.

Proof of Theorem 3.8. For every positive integer r, by Lemma 4.2(iii) with $A = A_r$, $a = 1/r$, and $m = \alpha_A(r)$ we get $\frac{1}{2}\sqrt{\frac{a^2-\delta_r}{m}} = \frac{1}{2r}\sqrt{\frac{1-r^2\delta_r}{\alpha_A(r)}}$. Thus, $b\,\alpha_A(r) - 1 \leqslant \log_2 \mathcal{N}\left(F, \frac{1}{2r}\sqrt{\frac{1-r^2\delta_r}{\alpha_A(r)}}\right)$. $\square$

Proof of Corollary 3.11. By Lemma 4.2(iii) with $A = A_r$, $a = t/r$, $m = r^\gamma$, and $\delta = 0$, for every positive integer r such that $r^\gamma \geqslant 3$ we have $b\,r^\gamma - 1 \leqslant \log_2 \mathcal{N}\left(F, \frac{t}{2\,\gamma/(\gamma+2)}\right)$. So $c(1/\varepsilon)^{2\gamma/(\gamma+2)} - 1 \leqslant \log_2 \mathcal{N}(F, \varepsilon)$, where $c = b\,(t/2)^{2\gamma/(\gamma+2)}$. Hence $(1/\varepsilon)^{2\gamma/(\gamma+2)} - 1 \leqslant \log_2 \mathcal{N}(F, \varepsilon)$ for $\varepsilon \downarrow 0$. $\square$

Proof of Theorem 3.12. The upper bound follows from [10, Proposition 5.1], which states that $\mathcal{N}(G,\varepsilon) \leqslant (\frac{1}{\varepsilon})^{\beta}$ for $\varepsilon \downarrow 0$ implies

$$\log_2 \mathcal{N}(\mathrm{conv}(G \cup -G), \varepsilon) \leqslant \left(\frac{1}{\varepsilon}\right)^{2\beta/(\beta+2)} \quad \text{for } \varepsilon \downarrow 0.$$

The lower bound follows from Corollary 3.11. □

Proof of Corollary 3.13. By [33, Theorem 2.6], there exists an absolute constant c such that for all $\varepsilon > 0$, $\mathcal{N}(G,\varepsilon) \leqslant cv(4e)^{v}\varepsilon^{-2v}$. So the estimate follows from Theorem 3.12. □

5. Application to neurocomputing

An important class of sets with power-type covering numbers in $(\mathcal{L}_2(\Omega), \|, \|_2)$, with $\Omega \subset \mathbb{R}^d$ bounded, consists of sets of functions computable by *perceptrons* with various types of *activation functions* $\psi : \mathbb{R} \to \mathbb{R}$. Such sets are of the form

$$P_d(\psi) = \{f : \Omega \to \mathbb{R} | f(x) = \psi(a \cdot x + b), x \in \Omega, a \in \mathbb{R}^d, b \in \mathbb{R}\}. \tag{6}$$

Widely used activation functions are *sigmoidals*, i.e., measurable functions $\sigma : \mathbb{R} \to \mathbb{R}$ such that

$$\lim_{t\to-\infty} \sigma(t) = 0 \quad \text{and} \quad \lim_{t\to+\infty} \sigma(t) = 1.$$

An important type of sigmoidal is the *Heaviside function* ϑ, defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geqslant 0$. We say that a sigmoidal is *polynomially quickly approximating the Heaviside* if there exist $\eta,\ C > 0$ such that for all $t \in \mathbb{R}$,

$$|\sigma(t) - \vartheta(t)| \leqslant C\,|t|^{\eta}.$$

The set $P_d(\vartheta)$ is the set of *characteristic functions of half-spaces of* $\mathbb{R}^d$ *restricted to* Ω. We denote it by H_d, i.e.,

$$H_d = P_d(\vartheta) = \{f : \Omega \to \mathbb{R} \mid f(x) = \vartheta(a \cdot x + b), a \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

Gurvits and Koiran [21] proved that for every d and every $\Omega \subset \mathbb{R}^d$ bounded, the set H_d is compact in $(\mathcal{L}_2(\Omega), \|.\|_2)$ (inspection of their proof shows that compactness also holds in $\mathcal{L}_p$-spaces with $p \in [1,\infty)$). Makovoz [32] estimated from above its covering numbers; he proved that for every positive integer d

$$\mathcal{N}(H_d, \varepsilon) \leqslant \left(\frac{1}{\varepsilon}\right)^{2d} \quad \text{for } \varepsilon \downarrow 0. \tag{7}$$

Moreover, he showed that for σ a Lipschitz continuous sigmoidal polynomially quickly approximating the Heaviside, $P_d(\sigma)$ has power-type covering numbers, i.e., there exists $\beta > 0$ such that

$$\mathcal{N}(P_d(\sigma), \varepsilon) \leqslant \left(\frac{1}{\varepsilon}\right)^{\beta} \quad \text{for } \varepsilon \downarrow 0. \tag{8}$$

So, for such sigmoidals the set $P_d(\sigma)$ is precompact. The next proposition shows that precompactness of $P_d(\sigma)$ holds even for Lipschitz continuous non-decreasing sigmoidals.

Proposition 5.1. *Let d be a positive integer, $\Omega \subset \mathbb{R}^d$ bounded, and σ a Lipschitz continuous non-decreasing sigmoidal. Then $P_d(\sigma)$ is precompact in $(\mathcal{L}_2(\Omega), \|.\|_2)$.*

Proof. For $\varepsilon > 0$, we decompose $P_d(\sigma)$ into three sets, in each of which we construct an ε-net. To simplify the notation, we write $\sigma_{a,b}(x)$ and $\vartheta_{a,b}(x)$ instead of $\sigma(a \cdot x + b)$ and $\vartheta(a \cdot x + b)$, resp. Let

$$P_d(\sigma) = P_d^{1,\varepsilon} \cup P_d^{2,\varepsilon} \cup P_d^{3,\varepsilon},$$

where

$$P_d^{1,\varepsilon}(\sigma)=\{\sigma_{a,b} \mid \|a\|_{l_2}\geqslant a_\varepsilon, b\in\mathbb{R}\},$$

$$P_d^{2,\varepsilon}(\sigma)=\{\sigma_{a,b} \mid \|a\|_{l_2}<a_\varepsilon, |b|\geqslant b_\varepsilon\},$$

and

$$P_d^{3,\varepsilon}(\sigma)=\{\sigma_{a,b} \mid \|a\|_{l_2}<a_\varepsilon, |b|<b_\varepsilon\}.$$

As Ω is bounded, for every $\varepsilon>0$ we can choose $a_\varepsilon\in\mathbb{R}^d_+$ such that for every $a\in\mathbb{R}^d$ with $\|a\|_{l_2}\geqslant a_\varepsilon$.

$$\left\|\sigma_{a,b}-\vartheta_{a,b}\right\|_2=\left(\int_\Omega(\sigma(a\cdot x+b)-\vartheta(a\cdot x+b))^2\,\mathrm{d}x\right)^{1/2}\leqslant\frac{\varepsilon}{3}.$$

As $\vartheta_{a,b}=\vartheta_{a/\|a\|_{l_2},b/\|a\|_{l_2}}$, we get

$$\left\|\sigma_{a,b}-\vartheta_{a/\|a\|_{l_2},b/\|a\|_{l_2}}\right\|_2\leqslant\frac{\varepsilon}{3}. \tag{9}$$

Since σ is sigmoidal, $\lim_{t\to\pm\infty}(\sigma(t)-\vartheta(t))=0$. So for every $\varepsilon>0$, we can choose $a_\varepsilon, b_\varepsilon>0$ such that for every $a\in\mathbb{R}^d$ with $\|a\|_{l_2}<a_\varepsilon$ and $b\in\mathbb{R}$ with $|b|\geqslant b_\varepsilon$:

$$\left\|\sigma_{a,\,b}-\vartheta_{a,\,b}\right\|_2\leqslant\frac{\varepsilon}{3}. \tag{10}$$

As σ is Lipschitz continuous, for every $a, a'\in\mathbb{R}^d$ and every $b, b'\in\mathbb{R}$ there exist $M_1, M_2>0$ such that

$$\|\sigma_{a,\,b}-\sigma_{a',b'}\|_2\leqslant M_1\,|a\cdot x-b-a'\cdot x+b'|\leqslant M_2\,(\|a-a'\|_{l_2}+|b-b'|). \tag{11}$$

If $\{\vartheta_{e_i^1,\,c_i^1}\}$ is an $\varepsilon/3$-net in H_d, then $\{\sigma_{a_i^1,b_i^1}\}:=\{\sigma_{a_\varepsilon e_i^1,a_\varepsilon c_i^1}\}$ is an ε-net in $P_d^{1,\varepsilon}(\sigma)$. Indeed, (9) gives for every $a\in\mathbb{R}^d$ with $\|a\|_{l_2}\geqslant a_\varepsilon$

$$\left\|\sigma_{a,b}-\sigma_{a_i^1,b_i^1}\right\|_2\leqslant\left\|\sigma_{a,b}-\vartheta_{a/\|a\|_{l_2},b/\|a\|_{l_2}}\right\|_2+\left\|\vartheta_{a/\|a\|_{l_2},b/\|a\|_{l_2}}-\vartheta_{e_i^1,\,c_i^1}\right\|_2$$
$$+\left\|\vartheta_{e_i^1,c_i^1}-\sigma_{a_i^1,b_i^1}\right\|_2\leqslant\varepsilon.$$

If $\{\vartheta_{e_i^2,b_i^2}\}$ is an $\varepsilon/3$-net in H_d, then $\{\sigma_{e_i^2,b_i^2}\}$ is an ε-net in $P_d^{2,\varepsilon}(\sigma)$. Indeed, for every $a\in\mathbb{R}^d$ with $\|a\|_{l_2}<a_\varepsilon$ and every $b\in\mathbb{R}$ with $|b|\geqslant b_\varepsilon$, by (10) we have

$$\left\|\sigma_{a,b}-\sigma_{a_i^2,b_i^2}\right\|_2\leqslant\left\|\sigma_{a,b}-\vartheta_{a,b}\right\|_2+\left\|\vartheta_{a,b}-\vartheta_{a_i^2,b_i^2}\right\|_2+\left\|\vartheta_{a_i^2,b_i^2}-\sigma_{a_i^2,b_i^2}\right\|_2\leqslant\varepsilon.$$

For $M_2>0$, if $\{a_i^3\}$ is an $\varepsilon/(2M_2)$-net in $[0, a_\varepsilon]$ and $\{b_i^3\}$ is an $\varepsilon/(2M_2)$-net in $[0, b_\varepsilon]$, then $\{\sigma_{a_i^3,b_i^3}\}$ is an ε-net in $P_d^{3,\varepsilon}(\sigma)$. Indeed, by (11) we get

$$\left\|\sigma_{a,b}-\sigma_{a_i^3,b_i^3}\right\|_2\leqslant M_2(\|a-a_i^3\|_{l_2}+|b-b_i^3|)\leqslant M_2\left(\frac{\varepsilon}{2M_2}+\frac{\varepsilon}{2M_2}\right)\leqslant\varepsilon.$$

As $P_d(\sigma)=P_d^{1,\varepsilon}(\sigma)\cup P_d^{2,\varepsilon}(\sigma)\cup P_d^{3,\varepsilon}(\sigma)$, the set $\{\sigma_{a_i^1,b_i^1}\}\cup\{\sigma_{a_i^2,b_i^2}\}\cup\{\sigma_{a_i^3,b_i^3}\}$ is an ε-net in $P_d(\sigma)$. $\square$

It was shown in [26, Propositions 3.3 and 3.4] that in $(\mathscr{L}_2(\Omega), \|.\|_2)$ with $\Omega\subset\mathbb{R}^d$ compact, for every continuous non-decreasing sigmoidal σ, $P_d(\sigma)$-variation is equal to H_d-variation and so the unit balls $B_1(\|.\|_{H_d})$ and $B_1(\|.\|_{P_d(\sigma)})$ are equal. The next theorem gives a tight estimate for the covering numbers of these balls.

Theorem 5.2. *Let d be a positive integer and* $\sigma : \mathbb{R} \to \mathbb{R}$ *either the Heaviside function or a continuous non-decreasing sigmoidal. Then in* $(\mathscr{L}_2([0,1]^d), \|.\|_2)$:

$$\log_2 \mathscr{N}(B_1(\|.\|_{H_d}), \varepsilon) = \log_2 \mathscr{N}(B_1(\|.\|_{P_d(\sigma)}), \varepsilon) \sim \left(\frac{1}{\varepsilon}\right)^{2d/(d+1)} \quad \textit{for } \varepsilon \downarrow 0.$$

Proof. By (7) and the upper bound from Theorem 3.12 with $\beta = 2d$, we get $\log_2 \mathscr{N}(B_1(\|.\|_{H_d}), \varepsilon) \leqslant (1/\varepsilon)^{2d/(d+1)}$ for $\varepsilon \downarrow 0$.

To prove the lower bound, we recall the construction that we made in [27] extending an idea from [3]. Let $A_d = \bigcup_{r=1}^{\infty} A_{d,r}$, where $A_{d,r} = \{h_v \mid v = (v_1, \ldots, v_d) \in \{1, \ldots, r\}^d\} \subset (\mathscr{L}_2([0,1]^d), \|.\|_2)$, $h_v(x) = c_v \sin(\pi v \cdot x) : [0,1]^d \to \mathbb{R}$, and $c_v = d\sqrt{2}/\sum_{j=1}^d v_j$. The sets $A_{d,r}$ are orthogonal and $B_{d\sqrt{8}}(\|.\|_{H_d}) \supset A_d$. So A_d is orthogonal slowly decaying with respect to d and is contained in the ball of radius $d\sqrt{8}$ in H_d-variation. Thus $B_1(\|.\|_{H_d}) \supset \frac{1}{d\sqrt{8}} A_d$ and by the lower bound from Theorem 3.12 with $\beta = 2d$ we get

$$\left(\frac{1}{\varepsilon}\right)^{2d/(d+1)} \leqslant \log_2 \mathscr{N}(B_1(\|.\|_{H_d}), \varepsilon) \quad \text{for } \varepsilon \downarrow 0. \qquad \square$$

6. Application to nonlinear approximation

In this section, we extend Makovoz's [32] result on tightness of an upper bound on rates of approximation of elements of the closed symmetric convex hulls of sets $P_d(\sigma)$, which was derived by Maurey (see [35]), Jones [22] and Barron [4].

Given two subsets S and T of a normed linear space $(X, \|.\|)$, we denote by $\delta(S, T)$ the *deviation of S from T*, which is the worst-case error in the approximation of elements of S by elements of T, i.e.,

$$\delta(S, T) = \delta(S, T, (X, \|.\|)) = \sup_{f \in S} \inf_{g \in T} \|f - g\|.$$

Reformulated in terms of G-variation [25], Maurey–Jones–Barron's estimates states that for a bounded subset G of a Hilbert space $(X, \|.\|)$ with $s_G = \sup_{g \in G} \|g\|$ and every positive integer n,

$$\delta(B_1(\|.\|_G), \mathrm{conv}_n(G \cup -G)) \leqslant \frac{s_G}{n^{1/2}}. \tag{12}$$

For perceptron networks with certain sigmoidal functions, the impossibility of improving the exponent $\frac{1}{2}$ in the bound (12) over $\frac{1}{2} + 1/d$ was proven by Barron [3] via a probabilistic argument and by Makovoz [32] via estimates of covering numbers. Exploiting Makovoz's [32] method of proof, we establish the tightness of the upper bound (12) for a set G with (i) power-type covering numbers and (ii) a sufficient "capacity" of its symmetric convex hull $\mathrm{conv}(G \cup -G)$, in the sense that $\mathrm{conv}\,(G \cup -G)$ contains a subset slowly decaying with respect to some $\gamma > 0$. The next theorem shows that for sets satisfying these two conditions, the exponent $\frac{1}{2}$ cannot be improved over $\frac{1}{2} + 1/\gamma$.

Theorem 6.1. *Let* $(X, \|.\|)$ *be a Hilbert space, G its bounded precompact subset with* $s_G = \sup_{g \in G} \|g\|$ *and power-type covering numbers,* $t, \gamma > 0$, *and* $B_1(\|.\|_G) \supseteq t\,A$, *where A is slowly decaying with respect to* γ. *If* $\tau > 0$ *is such that for some* $c > 0$ *and all positive integers n one has*

$$\delta(B_1(\|.\|_G), \mathrm{conv}_n(G \cup -G)) \leqslant c/n^{\tau}, \textit{ then } \tau \leqslant \tfrac{1}{2} + 1/\gamma.$$

To prove this theorem, we need the following lemma.

Lemma 6.2. *Let* $(X, \|.\|)$ *be a normed linear space and G be a bounded subset with* $s_G = \sup_{g \in G} \|g\|$. *For every* $\varepsilon > 0$ *and every positive integer n*,

$$\mathscr{N}(\mathrm{conv}_n\, G, \varepsilon(1 + s_G)) \leqslant (\mathscr{N}(G, \varepsilon))^n (2/\varepsilon)^n.$$

Proof. *Let B be an* ε-net in $B_1(\|.\|_{\ell_1^n})$ with respect to the ℓ_1^n-norm and A an ε-net in G with respect to the norm $\|.\|$ of X. Let $C \subset \mathrm{conv}_n\, G$ be defined as $C = \{\sum_{i=1}^n b_i\, g_i \mid (g_1, \dots, g_n) \in A^n, (b_1, \dots, b_n) \in B\}$. We show that C is an $\varepsilon(1+s_G)$-net in $\mathrm{conv}_n\, G$. Let $\sum_{i=1}^n \bar{b}_i\, \bar{g}_i \in \mathrm{conv}_n\, G$. Since B is an ε-net in $B_1(\|.\|_{l_1^n})$ with the l_1^n-norm, there exist $(b_1, \dots, b_n) \in B$ such that $\sum_{i=1}^n (b_i - \bar{b}_i) \leqslant \varepsilon$. As A is an ε-net in G with the norm $\|.\|$ of X, there exist $(g_1, \dots, g_n) \in A^n$ such that for every $i = 1, \dots, n$, $\|g_i - \bar{g}_i\| \leqslant \varepsilon$. Thus,

$$\left\| \sum_{i=1}^n b_i\, g_i - \sum_{i=1}^n \bar{b}_i \bar{g}_i \right\| \leqslant \left\| \sum_{i=1}^n b_i g_i - \sum_{i=1}^n b_i \bar{g}_i \right\| + \left\| \sum_{i=1}^n b_i \bar{g}_i - \sum_{i=1}^n \bar{b}_i \bar{g}_i \right\|$$
$$= \left\| \sum_{i=1}^n b_i (g_i - \bar{g}_i) \right\| + \left\| \sum_{i=1}^n (b_i - \bar{b}_i) \bar{g}_i \right\|$$
$$\leqslant \sum_{i=1}^n |b_i| \varepsilon + \sum_{i=1}^n |b_i - \bar{b}_i| \|g_i\| \leqslant \varepsilon + \varepsilon s_G = \varepsilon(1+s_G).$$

As card $C = (\mathrm{card}\, A)^n \mathrm{card}\, B$, we get

$$\mathcal{N}(\mathrm{conv}_n\, G, \|.\|, \varepsilon(1+s_G)) \leqslant (\mathcal{N}(G, \|.\|, \varepsilon))^n\, \mathcal{N}(B_1(\|.\|_{\ell_1^n}), \|.\|_{\ell_1^n}, \varepsilon).$$

It is well-known (see, e.g., [11, 1.1.10]) and easy to check that for a positive integer d, a norm $|.|$ on $\mathbb{R}^d$, and $\varepsilon > 0$, one has $(1/\varepsilon)^d \leqslant \mathcal{N}(B_1(|.|), |.|, \varepsilon) \leqslant (2/\varepsilon)^d$. So $\mathcal{N}(\mathrm{conv}_n\, G, \|.\|, \varepsilon(1+s_G)) \leqslant (\mathcal{N}(G, \|.\|, \varepsilon))^n\, (2/\varepsilon)^n$. □

Using Corollary 3.11 and Lemma 6.2 we now prove Theorem 6.1.

Proof of Theorem 6.1. Suppose *ab absurdo* that $\tau > \frac{1}{2} + 1/\gamma$ is such that for some $c > 0$ and every positive integer n one has $\delta(B_1(\|.\|_G), \mathrm{conv}_n(G \cup -G)) \leqslant c/n^\tau$.

For $\varepsilon > 0$, let $n_\varepsilon = \lceil (2c/\varepsilon)^{1/\tau} \rceil$, so $c/n_\varepsilon^\tau \leqslant \varepsilon/2$. Let Φ_{n_ε} be an ε/2-net in $\mathrm{conv}_{n_\varepsilon}(G \cup -G)$. As for every $f \in B_1(\|.\|_G)$ there exist $h_{n_\varepsilon} \in \mathrm{conv}_{n_\varepsilon}(G \cup -G)$ and $\phi_{n_\varepsilon} \in \Phi_{n_\varepsilon}$ such that $\|f - h_{n_\varepsilon}\| \leqslant c/n^\tau$ and $\|h_{n_\varepsilon} - \phi_{n_\varepsilon}\| \leqslant \varepsilon/2$, by the triangle inequality $\|f - \phi_{n_\varepsilon}\| \leqslant c/n_\varepsilon^\tau + \varepsilon/2 \leqslant \varepsilon$. So, Φ_{n_ε} is an ε-net in $B_1(\|.\|_G)$.

Since for an ε-net in G, $-A$ is an ε-net in $-G$, we get $\mathcal{N}(G \cup -G, \varepsilon) \leqslant 2\, \mathcal{N}(G, \varepsilon)$. This together with Lemma 6.2, implies that the cardinality of Φ_{n_ε} is bounded from above by $(\frac{4(1+s_G)}{\varepsilon} \mathcal{N}(G, \frac{\varepsilon}{1+s_G}))^{n_\varepsilon}$. As G has power-type covering numbers, there exists $\beta > 0$ such that $\mathcal{N}(G, \varepsilon) \leqslant (1/\varepsilon)^\beta$ for $\varepsilon \downarrow 0$ and so $\mathcal{N}(B_1(\|.\|_G), \varepsilon) \leqslant ((\frac{1+s_G}{\varepsilon})^\beta \frac{4(1+s_G)}{\varepsilon})^{n_\varepsilon} = (4\frac{1+s_G}{\varepsilon})^{n_\varepsilon(\beta+1)}$. Thus, $\log_2 \mathcal{N}(B_1(\|.\|_G), \varepsilon) \leqslant n_\varepsilon(\beta+1) \log_2(4\frac{1+s_G}{\varepsilon})$. As $\varepsilon \geqslant 2c/n_\varepsilon^\tau$, we get

$$\log_2 \mathcal{N}(B_1(\|.\|_G), \varepsilon) \leqslant n_\varepsilon(\beta+1) \log_2\left(4\frac{1+s_G}{\varepsilon}\right)$$
$$\leqslant \left\lceil \left(\frac{2c}{\varepsilon}\right)^{1/\tau} \right\rceil (\beta+1) \log_2\left(4\frac{1+s_G}{\varepsilon}\right). \tag{13}$$

On the other hand, by Corollary 3.11

$$\left(\frac{1}{\varepsilon}\right)^{2\gamma/(\gamma+2)} \leqslant \log_2 \mathcal{N}(B_1(\|.\|_G), \varepsilon) \quad \text{for } \varepsilon \downarrow 0. \tag{14}$$

Combining the bounds (13) and (14), we obtain

$$\left(\frac{1}{\varepsilon}\right)^{2\gamma/(\gamma+2)} \leqslant \log_2 \mathcal{N}(B_1(\|.\|_G), \varepsilon) \leqslant \left\lceil \left(\frac{2c}{\varepsilon}\right)^{1/\tau} \right\rceil (\beta+1) \log_2\left(4\frac{1+s_G}{\varepsilon}\right) \quad \text{for } \varepsilon \downarrow 0. \tag{15}$$

When $\tau > \frac{1}{2} + 1/\gamma$, we get $\frac{1}{\tau} < \frac{2\gamma}{\gamma+2}$ and so for ε small enough, (15) gives a contradiction (as the lower bound is larger than the upper bound). □

Thus, the exponent τ in the bound from Theorem 6.1 can be at most $\frac{1}{2} + 1/\gamma$ when G has power-type covering numbers and its symmetric convex hull contains an infinite set with orthogonal subsets of increasing cardinalities and magnitudes of the norms of their elements slowly decaying with respect to some $\gamma > 0$. The critical value of the exponent τ in the denominator is $\frac{1}{2} + 1/\gamma$. When γ increases, $\frac{1}{2} + 1/\gamma$ approaches $\frac{1}{2}$, which is the exponent in the bound (12).

Example 6.3. The set $A = \{n^{-1/\gamma} e_n\}$ considered in Example 3.4 satisfies the assumptions of Theorem 6.1. Indeed, for all $\varepsilon > 0$ and all positive integers $n \geqslant (1/\varepsilon)^\gamma$ we have $n^{-1/\gamma} e_n \in B_\varepsilon(\|.\|_2)$. So A has power-type covering numbers. As A is also slowly decaying with respect to γ, by Theorem 6.1 the term $n^{-\tau}$ in the upper bound on approximation of elements of $\mathrm{cl\,conv}(A \cup -A) = B_1(\|.\|_A)$ by $\mathrm{conv}_n\, A$ cannot be improved over $n^{-1/2-1/\gamma}$.

For every $\Omega \subset \mathbb{R}^d$ compact and every non-decreasing sigmoidal σ, in $(\mathscr{L}_2(\Omega), \|.\|_2)$ $P_d(\sigma)$-variation is equal to H_d-variation [26, Propositions 3.3 and 3.4] and $B_1(\|.\|_{\mathscr{H}_d})$ contains a set that is slowly decaying with respect to d (see the second part of the proof of Theorem 5.2). So we can apply Theorem 6.1 to the set $P_d(\sigma)$ of functions computable by perceptrons (see (6)), where σ is either the Heaviside function or a Lipschitz continuous sigmoidal polynomially quickly approximating the Heaviside. This implies Makovoz's result [32, Theorem 4, (11)]. Hence, in the upper bound (12) on approximation of elements of $\mathrm{cl\,conv}(P_d(\sigma) \cup -P_d(\sigma)) = B_1(\|.\|_{P_d(\sigma)})$ by $\mathrm{conv}_n\, P_d(\sigma)$, the term $n^{-1/2}$ cannot be improved over $n^{-1/2-1/d}$.

7. Discussion

We have derived lower bounds on covering numbers of precompact symmetric convex sets in terms of rates of decay of the magnitudes of the norms of the elements of their orthogonal subsets. The slower the rate of decay, the larger the lower bound. For symmetric convex hulls of sets with power-type covering numbers, by comparing our lower bounds with upper bounds we have obtained tight estimates of covering numbers. In particular, we have derived estimates for sets with finite VC-dimension.

Our results extend an estimate derived by Makovoz [32, Lemma 3], who using a result from [30] showed that for an orthogonal set A with cardinality m:

$$cm \leqslant \log_2 \mathscr{N}\left(\mathrm{conv}(A \cup -A), \frac{1}{\sqrt{m}}\right),$$

where c is an unspecified positive absolute constant. We have used a different proof technique (based on generalized Hadamard matrices) that provides more general results and allows one to specify the constant.

Applying our estimates to sets G of functions used in neurocomputing, we have obtained tight power-type bounds on covering numbers of $\mathrm{conv}\,(G \cup -G)$. Functions from such convex hulls can be approximated by convex combinations of n elements of G at rates $n^{1/2}$ [3,22,35]. We have shown that the exponent $\frac{1}{2}$ cannot be improved over $\frac{1}{2} + 1/\gamma$, where $\gamma > 0$ depends on the rate of decay of the magnitude of the norms of the elements of orthogonal subsets of $\mathrm{conv}\,(G \cup -G)$. This extends a result from [32] for perceptron neural networks with certain sigmoidals as activation functions. We have also shown that in $\mathscr{L}_2$-norm, sets of functions computable by perceptrons with more general sigmoidals (non-decreasing Lipschitz continuous) are precompact .

Acknowledgement

We thank a reviewer for useful references and for suggesting Example 3.4 and the comparison in Example 3.9.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, J. Assoc. Comput. Mach. 44 (1997) 615–631.
[2] M. Anthony, P.L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, Cambridge, UK, 1999.
[3] A.R. Barron, Neural net approximation, in: K. Narendra (Ed.), Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems, Yale University Press, New Haven, CT, 1992, pp. 69–72.
[4] A.R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inform. Theory 39 (1993) 930–945.

[5] P.L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Trans. Inform. Theory 44 (1998) 525–536.
[6] L. Birgé, Estimating a density under order restrictions: nonasymptotic minimax risk Ann. Statist. 15 (1987) 995–1012.
[7] O. Bousquet, V. Koltchinskii, D. Panchenko, Some local measures of complexity of convex hulls and generalization bounds, in: Proceedings of the 15th Annual Conference on Computational Learning Theory, Springer, London, 2002, pp. 59–73.
[8] B. Carl, Entropy numbers of diagonal operators with an application to eigenvalue problems, J. Approx. Theory 32 (1981) 135–150.
[9] B. Carl, Metric entropy of convex hulls in Hilbert spaces, Bull. London Math. Soc. 29 (1997) 452–458.
[10] B. Carl, I. Kyrezi, A. Pajor, Metric entropy of convex hulls in Banach spaces, J. London Math. Soc. 60 (1999) 871–896.
[11] B. Carl, I. Stephani, Entropy, Compactness, and the Approximation of Operators, Cambridge University Press, Cambridge, UK, 1990.
[12] J. Creutzig, I. Steinwart, Metric entropy of convex hulls in type p spaces—the critical case, Proc. Amer. Math. Soc. 130 (2002) 733–743.
[13] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[14] L. Devroye, G. Lugosi, Combinatorial Methods in Density Estimation, Springer, Berlin, 2001.
[15] R.M. Dudley, Uniform Central Limit Theorems, Cambridge Studies in Advanced Mathematics, vol. 63, Cambridge University Press, Cambridge, UK, 1999.
[16] D.E. Edmunds, H. Triebel, Function Spaces, Entropy Numbers, and Differential Operators, Cambridge University Press, Cambridge, UK, 1996.
[17] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[18] T.L. Fine, Feedforward Neural Network Methodology, Springer, New York, 1999.
[19] F. Gao, Metric entropy of convex hulls, Israel J. Math. 123 (2001) 359–364.
[20] F. Gao, Entropy of absolute convex hulls in Hilbert spaces, Bull. London Math. Soc. 36 (2004) 460–468.
[21] L. Gurvits, P. Koiran, Approximation and learning of convex superpositions, J. Comput. System Sci. 55 (1997) 161–170.
[22] L.K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Statist. 20 (1992) 608–613.
[23] P.C. Kainen, V. Kůrková, Quasiorthogonal dimension of Euclidean spaces, Appl. Math. Lett. 6 (1993) 7–10.
[24] A.N. Kolmogorov, Asymptotic characteristics of some completely bounded metric spaces, Dokl. Akad. Nauk. SSSR 108 (1956) 585–589.
[25] V. Kůrková, Dimension-independent rates of approximation by neural networks, in: K. Warwick, M. Kárný (Eds.), Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, Birkhauser, Boston, 1997, pp. 261–270.
[26] V. Kůrková, P.C. Kainen, V. Kreinovich, Estimates of the number of hidden units and variation with respect to half-spaces, Neural Networks 10 (1997) 1061–1068.
[27] V. Kůrková, M. Sanguineti, Comparison of worst case errors in linear and neural network approximation, IEEE Trans. Inform. Theory 48 (2002) 264–275.
[28] W.V. Li, W. Linde, Approximation, metric entropy and small ball estimates for Gaussian measures, Ann. Probab. 27 (1999) 1556–1578.
[29] W.V. Li, W. Linde, Metric entropy of convex hulls in Hilbert spaces, Studia Math. 139 (2000) 29–45.
[30] G.G. Lorentz, Metric entropy and approximation, Bull. Amer. Math. Soc. 72 (1966) 903–937.
[31] H. Luschgy, G. Pagés, Sharp asymptotics of the Kolmogorov entropy for Gaussian measures, J. Funct. Anal. 212 (2004) 89–120.
[32] Y. Makovoz, Random approximants and neural networks, J. Approx. Theory 85 (1996) 98–109.
[33] S. Mendelson, On the size of convex hulls of small sets, J. Mach. Learning Res. 2 (2001) 1–18.
[34] S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, Inven. Math. 152 (2003) 37–55.
[35] G. Pisier, Remarques sur un résultat non publié de B. Maurey. Séminaire d'Analyse Fonctionnelle 1980–81, Exposé no. V, École Polytechnique, Centre de Mathématiques, Palaiseau, France, pp. V.1–V.12.
[36] D. Pollard, Convergence of Stochastic Processes, Springer, New York, 1984.
[37] W. Rudin, Functional Analysis, 2nd ed., McGraw-Hill, USA, 1991.
[38] A.J. Smola, A. Elisseeff, B. Schölkopf, R.C. Williamson, Entropy numbers for convex combinations and MLPs, in: A.J. Smola, P.L. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 2000, pp. 369–387.
[39] A.J. Smola, R.C. Williamson, B. Schölkopf, Generalization bounds for convex combinations of kernel functions, NeuroCOLT Technical Report NC-TR-98-022, 1998.
[40] I. Steinwart, Entropy numbers of convex hulls and an application to learning algorithms, Arch. Math. 80 (2003) 310–318.
[41] I. Steinwart, Entropy of convex hulls—some Lorentz norm results, J. Approx. Theory 128 (2004) 42–52.
[42] I. Steinwart, C. Scovel, Fast rates for support vector machines using Gaussian kernels, Ann. Statist. 35 (2) (2007).
[43] V.N. Vapnik, Statistical Learning Theory, Wiley, USA, 1998.
[44] M. Vidyasagar, A Theory of Learning and Generalization, Springer, Berlin, 1997.
[45] R.C. Williamson, A.J. Smola, B. Schölkopf, Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators, IEEE Trans. Inform. Theory 47 (2001) 2516–2532.
[46] D.-X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.