Covering Numbers and Rates of Neural-Network Approximation

Kůrková, Věra
2001

Institute of Computer Science
Academy of Sciences of the Czech Republic

Covering numbers and rates of neural-network approximation

Věra Kůrková Marcello Sanguineti

Technical report No. 830

January, 2001

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: 6605 3231, fax: (+4202) 86 58 57 89,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Covering numbers and rates of neural-network approximation

Věra Kůrková Marcello Sanguineti[1]

Technical report No. 830

January, 2001

Abstract:

Tightness of dimension-independent upper bounds on neural network approximation is investigated in the framework of variable-basis approximation. Conditions are given on a variable basis that do not allow a possibility of improving such bounds beyond $\mathcal{O}(n_{-(\frac{1}{2}+\frac{1}{d})})$, where d is the number of variables of the functions to be approximated. Such conditions are satisfied by sigmoidal perceptrons.

Keywords:
nonlinear approximation, rates of approximation, variable-basis approximation, feedforward neural networks, covering numbers.

[1]Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genoa, Italy, E-mail: marcello@dist.unige.it.

1 Introduction

Feedforward networks are mostly simulated on classical computers; for such simulations, one of the limiting factors is the *number n of hidden units*. Jones [9] has obtained insight into the reason that some high-dimensional tasks can be performed efficiently by neural networks with a moderate number of hidden units. He constructed incremental approximants with rates of convergence of the order of $\mathcal{O}(n^{-\frac{1}{2}})$. The same estimates had earlier been proved by Maurey using a probabilistic argument (see Pisier [18] and also Barron [2]). Barron [2] improved Jones's [9] upper bound and applied it to neural networks. Using a weighted Fourier transform, he described sets of multivariable functions that can be approximated by perceptron networks having n hidden units within an accuracy of the order of $\mathcal{O}(n^{-\frac{1}{2}})$. Such bounds are sometimes called "dimension-independent" as they do not depend on the number of variables. However, such a term can be misleading, as sets of multivariable functions to which such estimates apply become more and more constrained as the number of variables increases.

The Maurey-Jones-Barron upper bound is quite general, as it applies to *nonlinear approximation of the variable-basis type*, i.e., approximation by linear combinations of n-tuples of elements of a given set of basis functions. This approximation scheme has been widely investigated (see, e.g., DeVore and Temlyakov [**?**] and the references therein): it includes splines with free nodes, trigonometric polynomials with free frequencies, sums of wavelets and feedforward neural networks.

Several authors have further improved or extended these dimension-independent bounds. An extension to $\mathcal{L}_p$-spaces, with $p \in (1, \infty)$, has been derived by Darken et al. [3] (with a rate of approximation of the order of $\mathcal{O}(n^{-\frac{1}{q}})$, where $q = \max(p, \frac{p}{p-1})$), and an extension to $\mathcal{L}_\infty$-spaces has been obtained by Barron [1], Girosi [7], Gurvits and Koiran [8], Makovoz [16] and Kůrková, Savický and Hlaváčková [14].

Makovoz [15] improved Maurey's probabilistic argument by combining it with a concept from metric entropy theory, which he also used to show that in the case of Lipschitz sigmoidal perceptron networks, the upper bound cannot be improved to $\mathcal{O}(n^{-\alpha})$ for $\alpha > \frac{1}{2} + \frac{1}{d}$, where d is the number of variables of the functions to be approximated. A similar tightness result for perceptron networks was earlier obtained by Barron [1], who used a more complicated proof technique. For the special case of orthonormal variable-basis, Mhaskar and Micchelli [17], Kůrková, Savický and Hlaváčková [14] and Kůrková and Sanguineti [13] have derived tight improvements of Maurey-Jones-Barron's bound.

In this paper, we extend tightness results derived by Barron [1] and Makovoz [15] for approximation by convex combinations of functions computable by sigmoidal perceptrons to combinations of more general basis functions satisfying certain conditions, that are fulfilled by standard neural-network hidden units. These conditions are defined in terms of (i) polynomial growth of the number of sets of a given diameter needed to cover such basis and (ii) sufficient "capacity" of the basis, in the sense that its convex hull has an orthogonal subset that for each positive integer k contains at least k^d functions with norms greater or equal to $\frac{1}{k}$. The proofs of our results, which are only sketched here, are given in [**?**].

2 Approximation by neural networks and by variable-basis functions

Approximation by feedforward neural networks can be studied in a more general context of *approximation by variable-basis functions*. In this approximation scheme, elements of a real normed linear space $(X, \|.\|)$ are approximated by linear combinations of at most n elements of a given subset G. The set of such combinations is denoted by $span_n G = \{\sum_{i=1}^n w_i g_i ; w_i \in \mathcal{R}, g_i \in G\}$; it is equal to the union of n-dimensional subspaces generated by all n-tuples of elements of G. G can represent the set of functions computable by *hidden units* in neural networks. Such units compute functions of the form $\phi : \mathcal{R}^p \times \mathcal{R}^d \to \mathcal{R}$, where $\mathcal{R}$ denotes the set of real numbers, ϕ corresponds to the type of unit, and p and d to the dimension of a *parameter space* and an *input space*, resp.. The set of input/output functions of a network with a single linear output unit and n hidden units computing the function ϕ is equal to $span_n G_\phi$, where $G_\phi = \{\phi(\mathbf{a}, \cdot); \mathbf{a} \in \mathcal{R}^p\}$. Also multilayer networks with a single linear output unit and n units in the last hidden layer belong to this approximation scheme; they compute functions from $span_n G$ with G depending on the number of units in the previous hidden layers.

Recall that a *perceptron* with an activation function $\psi : \mathcal{R} \to \mathcal{R}$ computes functions of the form $\phi((\mathbf{v}, b), \mathbf{x}) = \psi(\mathbf{v}\cdot\mathbf{x}+b) : \mathcal{R}^{d+1} \times \mathcal{R}^d \to \mathcal{R}$, where $\mathbf{v} \in \mathcal{R}^d$ is an *input weight* vector and $b \in \mathcal{R}$ is a *bias*. By $P_d(\psi) = \{f : [0,1]^d \to \mathcal{R}; f(\mathbf{x}) = \psi(\mathbf{v}\cdot\mathbf{x}+b), \mathbf{v} \in \mathcal{R}^d, b \in \mathcal{R}\}$ we denote the set of functions on $[0,1]^d$ computable by ψ-perceptrons. The most common activation functions are *sigmoidals*, i.e., functions $\sigma : \mathcal{R} \to [0,1]$ such that $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$; the discontinuous sigmoidal defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$ is called *the Heaviside function*. A function $\sigma : \mathcal{R} \to \mathcal{R}$ is Lipschitz if there exists $M > 0$ such that $|\sigma(t) - \sigma(t')| \leq M|t - t'|$ for all $t, t' \in \mathcal{R}$.

Rates of approximation of functions from a set Y by functions from a set M can be studied in terms of the *worst-case error* formalized by the concept of *deviation of* Y *from* M and defined as $\delta(Y, M) = \delta(Y, M, (X, \|.\|)) = \sup_{f\in Y} \|f - M\| = \sup_{f\in Y} \inf_{g\in M} \|f - g\|$. To formulate estimates of deviation from $span_n G$ we need to introduce a few more concepts and notations. If G is a subset of $(X, \|.\|)$ and $c \in \mathcal{R}$, then we define $c\,G = \{c\,g; g \in G\}$ and $G(c) = \{wg; g \in G, w \in \mathcal{R}\,\&\,|w| \leq c\}$. The *closure* of G is denoted by $cl\,G$ and defined as $cl\,G = \{f \in X; (\forall\varepsilon > 0)(\exists g \in G)(\|f - g\| < \varepsilon)\}$. G is *dense* in $(X, \|.\|)$ if $cl\,G = X$. The *convex hull* of G, denoted by $conv\,G$, is the set of all convex combinations of its elements, i.e., $conv\,G = \{\sum_{i=1}^n a_i g_i; a_i \in [0,1], \sum_{i=1}^n a_i = 1, g_i \in G, n \in \mathcal{N}_+\}$. $conv_n\,G$ denotes the set of all convex combinations of n elements of G, i.e., $conv_n\,G = \{\sum_{i=1}^n a_i g_i;\ a_i \in [0,1], \sum_{i=1}^n a_i = 1, g_i \in G\}$. $B_r(x, \|.\|)$ denotes the ball of radius r with respect to the norm $\|.\|$ centered at $x \in X$, i.e., $B_r(x, \|.\|) = \{y \in X; \|y - x\| \leq r\}$. We write shortly $B_r(\|.\|)$ instead of $B_r(0, \|.\|)$.

The following estimate is a version of Jones' result as improved by Barron [2] and also of earlier result of Maurey. Recall that a Hilbert space is a normed linear space with the norm induced by an inner product.

Theorem 2.1 *Let* $(X, \|.\|)$ *be a Hilbert space,* b *a positive real number,* G *a subset of* X *such that for every* $g \in G$ $\|g\| \leq b$*, and let* $f \in cl\,conv\,G$*. Then, for every positive integer* n*,* $\|f - conv_n G\| \leq \sqrt{\frac{b^2 - \|f\|^2}{n}}$.

In the following, we shall sometimes refer to Theorem 2.1 and to its bound as Maurey-Jones-Barron's theorem and bound, resp. As $conv_n G \subseteq span_n G$, the upper bound from Theorem 2.1 also applies to rates of approximation by $span_n G$. However, when G is not closed up to multiplication by scalars, $conv\,G$ is a proper subset of $span\,G$, and hence also $cl\,conv\,G$ is a proper subset of $cl\,span\,G$. Thus density of $span\,G$ in $(X, \|.\|)$ does not guarantee that Theorem 2.1 can be applied to all elements of X. As $conv_n G(c) \subset span_n G(c) = span_n G$ for any $c \in \mathcal{R}$, by replacing the set G by $G(c) = \{wg; w \in \mathcal{R}, |w| \leq c, g \in G\}$ we can apply Theorem 2.1 to all elements of $\cup_{c\in\mathcal{R}_+} cl\,conv\,G(c)$. This approach can be mathematically formulated in terms of a norm tailored to a set G (in particular, to sets G_ϕ corresponding to various computational units ϕ in neural networks). Let $(X, \|.\|)$ be a normed linear space and G be its subset, then *G-variation* (variation with respect to G) denoted by $\|.\|_G$ is defined as the Minkowski functional of the set $cl\,conv\,G(1) = cl\,conv(G \cup -G)$, i.e.,

$$\|f\|_G = \inf\{c \in \mathcal{R}_+; f \in cl\,conv\,G(c)\}.$$

G-variation has been introduced by Kůrková [11] as an extension of Barron's [1] concept of variation with respect to half-spaces (more precisely, variation with respect to characteristic functions of half-spaces) corresponding to perceptrons with Heaviside activation function. For functions of one variable, variation with respect to half-spaces coincides, up to a constant, with the notion of total variation studied in integration theory; for G orthonormal, it is equal to the l_1-norm with respect to G (see [13]). The following theorem is a corollary of Theorem 2.1 formulated in terms of G-variation (see [11]). Recall that for any G, the unit ball in G-variation is equal to $cl\,conv(G \cup -G)$.

Theorem 2.2 *Let* $(X, \|.\|)$ *be a Hilbert space and* G *be its subset. Then, for every* $f \in X$ *and every positive integer* n*,* $\delta(B_1(\|.\|_G), span_n G) \leq \frac{s_G}{\sqrt{n}}$*, where* $s_G = \sup_{g\in G} \|g\|$.

Thus all functions from the unit ball in G_ϕ-variation can be approximated within $\frac{s_{G_\phi}}{\sqrt{n}}$ by ϕ-networks with n hidden units independently on the number d of variables. However, with increasing number of variables, the condition of being in the unit ball in G_ϕ-variation becomes more and more constraining (see [14] for examples of functions with variations depending exponentially on d).

3 Covering numbers

Recall that for $\varepsilon > 0$, the *ε-covering number* of a subset K of a normed linear space $(X, \|.\|)$ is defined as $cov_\varepsilon K = cov_\varepsilon(K, \|.\|) = \min\{n \in \mathcal{N}_+ ; K \subseteq \cup_{i=1}^n B_\varepsilon(x_i, \|.\|), x_i \in K\}$ if the set over which the minimum is taken is nonempty, otherwise $cov_\varepsilon(K) = +\infty$. The *$\varepsilon$-metric entropy* of K is defined as $H_\varepsilon(K) = log_2\, cov_\varepsilon K$.

The *n-covering diameter* of K is defined as $diam_n(K) = \inf\{\varepsilon \in \mathcal{R}_+ ; K \subseteq \cup_{i=1}^n B_\varepsilon(x_i, \|.\|)\}$. When the covering sets are open or closed balls of radius $\frac{\varepsilon}{2}$, then $diam_n(K)$ is the *n-th entropy number* $\epsilon_n(K)$ (see [4, p.7]).

A subset $\{x_1, \dots, x_m\}$ of K is called *ε-distinguishable* if for each distinct pair x_i, x_j of its elements, $\|x_i - x_j\| > \varepsilon$. The *$\varepsilon$-packing number* of K, $pack_\varepsilon K$, is defined as the maximal cardinality of an ε-distinguishable subset of K. The *ε-capacity* of K is defined as $C_\varepsilon(K) = log_2\, pack_\varepsilon K$.

It follows directly from the definitions and the triangle inequality that $pack_{2\varepsilon}(K) \le cov_\varepsilon(K) \le pack_\varepsilon(K)$. Obviously, the same relationships hold between $H_\varepsilon(K)$ and $C_\varepsilon(K)$.

The following lemma gives an elementary estimate of covering numbers of balls in a norm on $\mathcal{R}^n$.

Lemma 3.1 *Let n be a positive integer, $\|.\|$ be a norm on $\mathcal{R}^n$ and $\varepsilon > 0$, then $\left(\frac{1}{\varepsilon}\right)^n \le cov_\varepsilon B_1(\|.\|) \le \left(\frac{2}{\varepsilon}\right)^n$.*

Proof. Let vol denotes the Euclidean volume in $\mathcal{R}^n$. For every $\varepsilon > 0$, we have $vol(B_\varepsilon(\|.\|)) = \varepsilon^n\, vol(B_1(\|.\|))$. It follows from It follows directly from the definitions that $cov_\varepsilon B_1(\|.\|)\, vol(B_\varepsilon(\|.\|)) \ge vol(B_1(\|.\|))$ and $pack_{2\varepsilon} B_1(\|.\|)\, vol(B_\varepsilon(\|.\|)) \le vol(B_1(\|.\|))$. Hence, $pack_{2\varepsilon} B_1(\|.\|) \le \varepsilon^{-n} \le cov_\varepsilon B_1(\|.\|)$. Since $pack_{2\varepsilon} K \le cov_\varepsilon K \le pack_\varepsilon K$ we have $cov_\varepsilon B_1(\|.\|) \le pack_\varepsilon B_1(\|.\|) \le \left(\frac{2}{\varepsilon}\right)^n \le cov_{\frac{\varepsilon}{2}} B_1(\|.\|)$, and hence $\left(\frac{1}{\varepsilon}\right)^n \le cov_\varepsilon B_1(\|.\|) \le \left(\frac{2}{\varepsilon}\right)^n$. □

Lemma 3.2 *Let $(X, \|.\|)$ be a Hilbert space, G its subset and $s_G = \sup_{g \in G} \|g\|$. Then, for every $\varepsilon > 0$,*
(i) $cov_{\varepsilon(1+s_G)}(conv_n\, G, \|.\|) \le (cov_\varepsilon\, (G, \|.\|))^n\, cov_\varepsilon(B_1(\|.\|_{l_1^n}), \|.\|_{l_1^n})$;
(ii) $cov_{\varepsilon(1+s_G)}(conv_n\, G) \le (cov_\varepsilon\, G)^n\, \left(\frac{2}{\varepsilon}\right)^n$;
(iii) $cov_\varepsilon(G \cup -G) \le 2 cov_\varepsilon\, G$.

Proof. (i) Let B be an ε-net in $B_1(\|.\|_{l_1^n})$ with respect to l_1^n and A be an ε-net in G with respect to the norm $\|.\|$. Let C be a subset of $conv_n\, G$ formed by all expressions $\sum_{i=1}^n b_i\, g_i$, where $(g_1, \dots, g_n) \in G^n$ and $(b_1, \dots, b_n) \in B$. We have $card\, C = (card\, A)^n\, card\, B$. Since $\|\sum_{i=1}^n b_i\, g_i - \sum_{i=1}^n \bar{b}_i\, \bar{g}_i\| \le \|\sum_{i=1}^n b_i\, g_i - \sum_{i=1}^n b_i\, \bar{g}_i\| + \|\sum_{i=1}^n b_i\, \bar{g}_i - \sum_{i=1}^n \bar{b}_i\, \bar{g}_i\| = \|\sum_{i=1}^n b_i\, (g_i - \bar{g}_i)\| + \|\sum_{i=1}^n (b_i - \bar{b}_i)\, \bar{g}_i\| \le \sum_{i=1}^n |b_i|\, \varepsilon + \sum_{i=1}^n |b_i - \bar{b}_i|\, \|g_i\| \le \varepsilon + \varepsilon\, s_G = \varepsilon(1 + s_G)$, C is an $\varepsilon(1+s_G)$-net in $conv_n\, G$ with respect to $\|.\|$.
(ii) follows directly from (i).
(iii) If C in an ε-net in G, then $-C$ in an ε-net in $-G$ and hence $C \cup -C$ is an ε-net in $G \cup -G$. □

4 Quasiorthogonal dimension

The cube $\{-1,1\}^m$ is called the *Hamming cube*. Let h denotes a metric on $\{-1,1\}^m$ defined as the the number of coordinates at which two vectors differ; usually called the *Hamming metric*, it is just the l_1-norm.

A *Hadamard matrix* of order m is a set of pairwise orthogonal vectors in the Hamming cube $\{-1,1\}^m$ with a particular ordering. It is well-known that, except for $m = 1$ and $m = 2$, a Hadamard matrix can only exist when m is divisible by 4 and this condition is believed to be sufficient. Kainen and Kůrková [10] have generalized the concept of Hadamard matrix by allowing a certain tolerance in the orthogonality condition. For $\varepsilon \in [0,1]$, they have defined an *ε-Hadamard matrix* of order m as an ordered set of vectors in $\{-1,1\}^m$ with all inner products of any two distinct rows in absolute value 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. Since the absolute value of the inner product of a pair of vectors in $\{-1,1\}^m$ is equal to an integer between 0 and

m, it follows that $R(\varepsilon, m) = R(\frac{\lfloor\varepsilon\rfloor}{m}, m)$ for each $\varepsilon \in [0,1]$. When $\varepsilon = \frac{k}{m}$, then $|\mathbf{u} \cdot \mathbf{v}| \leq k$. It is easy to check that, for each two distinct vectors $\mathbf{u}, \mathbf{v}$ in an ε-Hadamard matrix of order m, $h(\mathbf{u}, \mathbf{v}) \geq m(\frac{1-\varepsilon}{2})$. When $\varepsilon = \frac{k}{m}$, then $h(\mathbf{u}, \mathbf{v}) \geq \frac{m-k}{2}$.

The following lemma gives a lower bound on certain covering numbers of the unit ball in variation with respect to an orthogonal set.

Lemma 4.1 *Let $(X, \|.\|)$ be a Hilbert space, A be its orthogonal subset such that $card A = m$ and $\min_{g \in A} \|g\| = a$. Then for each integer k such that $1 \leq k < m$, $cov_{\delta_k} B_1(\|.\|_A) \geq R\left(\frac{k}{m}, m\right)$, where $\delta_k = \frac{a}{m}\sqrt{\lceil \frac{m-k}{2} \rceil}$.*

Proof. Let $A = \{g_1, \ldots, g_m\}$ and let M_k be a $\frac{k}{m}$-Hadamard matrix of order m with $R\left(\frac{k}{m}, m\right)$ rows. We shall show that the set $A(M_k) = \{\frac{1}{m}\sum_{i=1}^m u_i g_i; \mathbf{u} \in M_k\}$ is $2\delta_k = \frac{2a}{m}\sqrt{\lceil \frac{m-k}{2} \rceil}$-separated. For any two distinct vectors $\mathbf{u}, \mathbf{v} \in M_k$, we have $h(\mathbf{u}, \mathbf{v}) \geq \frac{m-k}{2}$. Thus the cardinality of the set I of indices, representing the coordinates where $\mathbf{u}$ and $\mathbf{v}$ differ, satisfies $\lceil \frac{m-k}{2} \rceil \leq card I \leq \lfloor \frac{m-k}{2} \rfloor$. Hence $\|\frac{1}{m}\sum_{i=1}^m (u_i - v_i) g_i\| = \frac{2}{m}\|\sum_{i \in I} g_i\| \geq \frac{2a}{m}\sqrt{\lceil \frac{m-k}{2} \rceil}$. Finally, $card A(M_k) = R\left(\frac{k}{m}, m\right)$ and $A(M_k) \subset B_1(\|.\|_A)$, imply that $cov_{\delta_k}(B_1(\|.\|_A)) \geq R\left(\frac{k}{m}, m\right)$. □

Lemma 4.1 gives a lower bound on certain covering numbers of balls in variation with respect to an orthogonal set. For a smaller value of δ_k, a similar lower bound on $cov_{\delta_k} B_1(\|.\|_A)$ can be obtained even if the orthogonality condition on the set A is relaxed to ε-nearly orthogonality.

A subset $A = \{g_1, \ldots . g_m\}$ of a Hilbert space $(X, \|.\|)$ is called *ε-nearly orthogonal* if $\sum_{j=1,\, j \neq i}^m |g_i \cdot g_j| \leq \varepsilon,\ \ i = 1, \ldots m$.

Hech-Nielsen introduced the concept of quasiortogonality. For $\varepsilon \in (0,1)$, two vectors $\mathbf{u}, \mathbf{v} \in \mathcal{R}^n$ are called *ε-quasiorthogonal* if $|\mathbf{u} \cdot \mathbf{v}| \leq \varepsilon \|u\|\, \|\mathbf{v}\|$. If $A = \{g_1, \ldots, g_m\}$ is a set of pairwise ε-quasiorthogonal vectors in $\mathcal{R}^n$, then A is $(m-1)\varepsilon$-nearly orthogonal (as $\sum_{i \neq j} |g_i \cdot g_j| \leq (m-1)\varepsilon$).

Lemma 4.2 *Let $(X, \|.\|)$ be a Hilbert space, A be its ε-nearly orthogonal subset such that $card A = m$ and $\min_{g \in A} \|g\| = a$, and let $\varepsilon \leq \sqrt{a}$. Then for each integer k such that $1 \leq k < m$, $cov_{\delta_k}(B_1(\|.\|_A)) \geq R\left(\frac{k}{m}, m\right)$, where $\delta_k = \frac{\sqrt{|a^2 - \varepsilon|}}{m}\sqrt{\lceil \frac{m-k}{2} \rceil}$.*

Proof. Analogously as in the proof of Lemma 4.1 we derive that the set $A(M_k) = \{\frac{1}{m}\sum_{i=1}^m u_i g_i; \mathbf{u} \in M_k\}$ is $2\delta_k = \frac{2\sqrt{|a^2-\varepsilon|}}{m}\sqrt{\lceil \frac{m-k}{2} \rceil}$-separated. A lower bound on $\|\frac{1}{m}\sum_{i=1}^m (u_i - v_i) g_i\|$ is calculated as follows: Let $x_i = \frac{1}{2\sqrt{r}}(u_i - v_i)$, $i \in I$. Then $x_i = \pm\frac{1}{\sqrt{r}}$, and $\|\frac{1}{m}\sum_{i=1}^m (u_i - v_i) g_i\| = \frac{1}{m}\|\sum_{i \in I} g_i\| = \frac{2\sqrt{r}}{m}\|\sum_{i=1}^r x_i\, g_i\|$, where $r = card I$. Moreover, $\|\sum_i^r x_i\, g_i\|^2 = |\sum_{i=1}^r \sum_{j=1}^r x_i\, x_j\, d_{ij}|$, where $d_{ij} = x_i\, x_j$. Since $\sum_{i=1}^r x_i^2 = 1$, it is sufficient to estimate from below the function $f(x_1, \ldots, x_r) = |\sum_{i=1}^r \sum_{j=1}^r x_i\, x_j\, d_{ij}|$ on the unit sphere of $\mathcal{R}^r$. Let D_I be a matrix defined by $D_{I\,ij} = d_{ij}$. Then $f(x_1, \ldots, x_r) \geq \frac{2\sqrt{r}}{m}\sqrt{|\lambda_{\min}(D_I)|}$, where $\lambda_{\min}(D_I)$ denotes the minimun eigenvalue of D_I. As $\lambda_{\min}(D_I)| \geq |\min_{g_i \in A} \|g_i\|^2 - \sum_{i \in I,\, i \neq j} |g_i \cdot g_j|| \geq |a^2 - \varepsilon|$, we get $\frac{1}{m}\|\sum_{i=1}^m (u_i - v_i) g_i\| \geq \frac{2\sqrt{r|a^2-\varepsilon|}}{m} \geq \frac{2\sqrt{|a^2-\varepsilon|}}{m}\sqrt{\lceil \frac{m-k}{2} \rceil}$.
□

Combining Lemma 4.1 with a lower bound on $R\left(\frac{k}{m}, m\right)$ we get a lower bound on ε-covering number of balls in G-variation containing an orthogonal subset for ε defined in terms of the cardinality of such an orthogonal subset and the minimum of norms of its elements. The proof of the next lemma is based on the exponentional growth of quasiorthogonal dimension studied in [10]. $H(p) = p\log(p) + (1-p)\log(1-p)$ denotes the entropy function.

Lemma 4.3 *Let $(X, \|.\|)$ be a Hilbert space, G, A be its subsets such that $A \subseteq B_1(\|.\|_G)$, A is a set of m orthogonal elements and $\min_{h \in A} \|h\| = a$.*
Then $cov_{\frac{a}{2\sqrt{m}}} B_1(\|.\|_G) \geq 2^{b\,m}$, where $b = H(\frac{1}{4})$.

Proof. By [10, Theorem 3.4] for every positive integer m and $k \in \{1, \ldots, m-1\}$, $R\left(\frac{k}{m}, m\right) \geq \frac{2^{m-1}}{B(\lambda_{m,k}, m)}$, where $\lambda_{m,k} = \lceil \frac{m-k-2}{2} \rceil$ and $B(\lambda, m) = \sum_{i=0}^{\lambda} \binom{m}{i}$ is a partial sum of binomials.

As $\lambda_{m,\frac{k}{m}} = \lceil \frac{\frac{m}{2}-2}{2} \rceil < \frac{m}{2}$ we can use the estimate $B(\lambda, m) \leq 2^{mH\left(\frac{\lambda}{m}\right)}$, that is valid for $\lambda < \frac{m}{2}$ (see [6]).

Thus, $R\left(\frac{k}{m}, m\right) \geq \frac{2^{m-1}}{B(\lambda_{m,k}, m)} \geq 2^{m-1} 2^{-mH\left(\frac{\lambda_{m,\frac{k}{m}}}{m}\right)} = 2^{m\left[1-H\left(\frac{\lambda_{m,\frac{k}{m}}}{m}\right)\right]-1}$.

As the entropy function increasing and $\lambda_{m,\frac{k}{m}} = \lceil \frac{m-k-2}{2} \rceil = \lceil \frac{m-k}{2} - 1 \rceil \leq \frac{m-k}{2}$, we get $R\left(\frac{k}{m}, m\right) \geq 2^{mH\left(\frac{m-k}{2m}\right)}$. Setting $k = \lfloor \frac{m}{2} \rfloor$ we have $H\left(\frac{m-k}{2m}\right) = H\left(\frac{m-\lfloor \frac{m}{2} \rfloor}{2m}\right) \geq H\left(\frac{1}{4}\right)$.

By Lemma 4.1, $cov_{\delta_k}(B_1(\|.\|_A)) \geq R\left(\frac{k}{m}, m\right) \geq 2^{mH\left(\frac{1}{4}\right)} = 2^{bm}$, where $b = H(\frac{1}{4})$. □

5 Tightness of the bound $\mathcal{O}(n^{-\frac{1}{2}})$ on variable-basis approximation

To disprove for certain sets G the possibility of an improvement of Maurey-Jones-Barron's upper bound beyond $\mathcal{O}(n^{-(\frac{1}{2}+\frac{1}{d})})$, we shall assume that such an improvement is possible and derive a contradiction by considering its consequences on the growth of certain covering numbers of the unit ball in G-variation.

We shall apply Lemma 4.3 to a ball containing a sequence of subsets with increasing cardinality, that contain orthogonal elements with norms that do not vanish "too quickly". More precisely, for a positive integer d (corresponding, in the following, to the number of variables of functions in X), we call a subset A of a normed linear space $(X, \|.\|)$ *not quickly vanishing with respect to* d if $A = \cup_{k \in \mathcal{N}_+} A_k$, where, for each $k \in \mathcal{N}_+$, $card\, A_k \geq k^d$ and for each $h \in A_k$, $\|h\| \geq \frac{1}{k}$ (see [12]).

Recall that for $f, g : \mathcal{N}_+ \to \mathcal{N}_+$, $g(n) \leq \mathcal{O}(f(n))$ if there exists $c \in \mathcal{R}_+$ such that for all but finitely many $n \in \mathcal{N}_+$, $g(n) \leq c\, f(n)$. Makovoz [15] proved that when σ is a Lipschitz sigmoidal, then the rate of the order of $\mathcal{O}(n^{-\frac{1}{2}})$ in approximation of elements of the unit ball in $P_d(\sigma)$-variation by $conv_n(P_d(\sigma) \cup -P_d(\sigma))$, that is guaranteed by Maurey-Jones-Barron's theorem, cannot be improved to $\mathcal{O}(n^{-\alpha})$ for $\alpha > \frac{1}{2} + \frac{1}{d}$. Our main theorem extends this Makovoz's result to sets G of functions of d variables that have covering numbers depending only polynomially on the number of variables d and for which the unit ball in G-variation contains an orthogonal subset that is not quickly vanishing with respect to d.

Theorem 5.1 *Let $(X, \|.\|)$ be a Hilbert space of functions of d variables and G be its bounded subset satisfying the following conditions:*
(i) there exists a polynomial $p(d)$ and $b \in \mathcal{R}_+$ such that, for every $\varepsilon > 0$, $cov_\varepsilon(G) \leq b\left(\frac{1}{\varepsilon}\right)^{p(d)}$;
(ii) there exists $r \in \mathcal{R}_+$ for which $B_r(\|.\|_G)$ contains a set of orthogonal elements which is not quickly vanishing with respect to d.
Then $\delta(B_1(\|.\|_G), conv_n(G \cup -G)) \leq \mathcal{O}(n^{-\alpha})$ implies $\alpha \leq \frac{1}{2} + \frac{1}{d}$.

Proof. Assume that there exists $\alpha > \frac{1}{2} + \frac{1}{d}$ such that, for all but finitely many $n \in \mathcal{N}_+$, $\delta(B_1(\|.\|_G), conv_n(G \cup -G)) \leq \frac{c}{n^\alpha}$. Set $\delta = \frac{2c}{n^\alpha}$. We shall derive a contradiction by comparing an upper bound on $cov_\delta B_1(\|.\|_G)$ (obtained from the assumption (i) and this hypothetical upper bound) with a lower bound on the same covering number (obtained from the assumption (ii) and Lemma 4.3). Without loss of generality assume $s_G = 1$. By the triangle inequality, Lemma 3.2 and the assumption (i), we get $cov_\delta B_1(\|.\|_G) \leq cov_{\delta/2} conv_n(G \cup -G) \leq (2cov_{\delta/4} G)^n (\frac{8}{\delta})^n \leq a^n 4^{n(2+p(d))} \delta^{-n(1+p(d))} = a(n,d) n^{\alpha n(1+p(d))}$, where $a(n,d) = a^n 4^{n(2+p(d))} (2c)^{-n(1+p(d))}$. On the other hand, using the assumption (ii) set for each positive integer k, $A_{r,k} = \frac{1}{r} A_k$. We have $A_{r,k} \subset B_1(\|.\|_G)$ and by Lemma 4.3, $cov_{\varepsilon_k} B_1(\|.\|_G) \geq cov_{\varepsilon_k} B_1(\|.\|_{A_{r,k}}) \geq 2^{bk^d}$, where $b = H(\frac{1}{4})$ and $\varepsilon_k = \frac{1}{2rk^{d/2+1}}$. If $k \leq \bar{k} = \frac{n^\alpha}{4cr}^{\frac{2}{d+2}}$, then $\delta \leq \varepsilon_k$. So for $\bar{k}$ an integer, set $k = \bar{k}$. Then we get $cov_\delta B_1(\|.\|_G) \geq cov_{\varepsilon_k} B_1(\|.\|_G) \geq 2^{bk^d} \geq 2^{c_d n^{\frac{\alpha}{1/2+1/d}}}$, where $c_d = b\left(\frac{1}{4cr}\right)^{\frac{1}{1/2+1/d}}$, which gives for large n a contradiction. If $\bar{k}$ is not integer, set $k = \lfloor \bar{k} \rfloor \geq \bar{k} - 1 \geq \frac{\bar{k}}{2}$ for $\bar{k} \geq 2$, and get a contradiction in a similar way. □

Since both assumptions of Theorem 5.1 are satisfied by sets of functions computable by perceptrons with Lipschitz sigmoidal activation, we get the following corollary.

Corollary 5.2 *Let d, n be positive integers and let $\sigma : \mathcal{R} \to \mathcal{R}$ be a Lipschitz sigmoidal function. Then in* $(\mathcal{L}^2([0,1]^d), \|.\|_2)$,
$\delta(B_1(\|.\|_{P_d(\sigma)}), span_n(P_d(\sigma) \cup -P_d(\sigma)) \leq \mathcal{O}(n^{-\alpha})$ *implies* $\alpha \leq \frac{1}{2} + \frac{1}{d}$.

Proof. It is sufficient to check that both conditions (i) and (ii) from Theorem 5.1 are satisfied by $P_d(\sigma)$. For the condition (i), see [15, Lemma 2]. The condition (ii) is guaranteed by the following construction from [12]: set $A_d = \cup_{k \in \mathcal{N}_+} A_{d,k}$, where $A_{d,k} = \{h_{\mathbf{v}}; \mathbf{v} \in \{1, \ldots, k\}^d\} \subset (\mathcal{L}_2([0,1]^d), \|.\|_2)$, with $h_{\mathbf{v}}(\mathbf{x}) = c_{\mathbf{v}} \cos(2\pi \mathbf{v} \cdot \mathbf{x}) : [0,1]^d \to \mathcal{R}$, $c_{\mathbf{v}} = d/(\sqrt{2} \lceil \sum_{j=1}^d v_k \rceil)$, and $\mathbf{v} = (v_1, \ldots, v_d)$. It is shown in [12] that for any positive integer d, $A_d \subset B_{d/\sqrt{8}}(\|.\|_{P_d(\sigma)})$ and that $A = \cup_{d \in \mathcal{N}_+} A_d$ is orthogonal not quickly vanishing with respect to d. □

6 Discussion

We have stated conditions that prevent an improvement of Maurey-Jones-Barron's upper bound to $\mathcal{O}(n^{-\alpha})$, for $\alpha > \frac{1}{2} + \frac{1}{d}$. As sets of functions computable by Lipschitz sigmoidal perceptrons satisfy these conditions, it follows that one cannot improve the upper bound on the approximation rate for one-hidden-layer networks with such perceptrons when the sum of the absolute values of the output weights is kept below a certain fixed bound. It is an open problem whether Theorem 5.1 can be generalized to approximation by linear instead of convex combinations (a special case of this problem concerning one-hidden layer perceptron networks with a Lipschitz sigmoidal activation function and unconstrained output weights was stated by Makovoz in [15]).

Better rates than $\mathcal{O}(n^{-(\frac{1}{2}+\frac{1}{d})})$ might be achievable using networks with more than one hidden layer since for some of such networks, sets of basis functions might be are much larger than in teh case of one-hidden-layer networks, and thus they might not satisfy condition (i) on polynomial growth of covering numbers.

7 Acknowledgments

The authors were partially supported by NATO Grant PST.CLG.976870. V. Kůrková was also partially supported by grant GA ČR 201/00/1489. M. Sanguineti was also partially supported by the Italian Ministry for the University and Research (MURST) and by grant D.R.42 of the University of Genoa.

Bibliography

[1] Barron, A.R.: Neural net approximation. *Proc. 7th Yale Workshop on Adaptive and Learning Systems* (K. Narendra, Ed.), pp. 69-72. Yale University Press, 1992.

[2] Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory* 39, pp. 930-945, 1993.

[3] Darken, C., Donahue, M., Gurvits, L., and Sontag, E.: Rate of approximation results motivated by robust neural network learning. *Proc. Sixth Annual ACM Conference on Computational Learning Theory.* The Association for Computing Machinery, New York, N.Y., pp. 303-309, 1993.

[4] Carl, B. and Stephani, I.: *Entropy, Compactness and the Approximation of Operators.* Cambridge University Press, 1990.

[5] DeVore, R.A., and Temlyakov, V.N.: Nonlinear approximation by trigonometric sums. *The J. of Fourier Analysis and Applications* 2, pp. 29-48, 1995.

[6] Fine, T.L.: *Feedforward Neural Network Methodology.* Springer-Verlag, New York, 1999.

[7] Girosi, F.: Approximation error bounds that use VC-bounds. *Proc. International Conference on Artificial Neural Networks ICANN'95.* Paris: EC2 & Cie, pp. 295–302, 1995.

[8] Gurvits, L., and Koiran, P.: Approximation and learning of convex superpositions. *J. of Computer and System Sciences* 55, pp. 161–170, 1997.

[9] Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics* 20, pp. 608–613, 1992.

[10] Kainen, P.C., and Kůrková, V.: Quasiorthogonal dimension of Euclidean spaces. *Applied Math. Lett.* 6, pp. 7–10, 1993.

[11] Kůrková, V.: Dimension-independent rates of approximation by neural networks. In *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality* (K. Warwick, M. Kárný, Eds.). Birkhauser, Boston, pp. 261-270, 1997.

[12] Kůrková, V., and Sanguineti, M.: Tools for comparing neural network and linear approximation. *Submitted to IEEE Trans. on Information Theory.*

[13] Kůrková, V., and Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation. To appear in *IEEE Trans. on Information Theory.*

[14] Kůrková, V., Savický, P., and Hlaváčková, K.: Representations and rates of approximation of real–valued Boolean functions by neural networks. *Neural Networks* 11, pp. 651-659, 1998.

[15] Makovoz, Y.: Random approximants and neural networks. *J. of Approximation Theory* 85, pp. 98–109, 1996.

[16] Makovoz, Y.: Uniform approximation by neural networks. *J. of Approximation Theory* 95, pp. 215–228, 1998.

[17] Mhaskar, H.N. and Micchelli, C.A.: Dimension-independent bounds on the degree of approximation by neural networks. *IBM J. of Research and Development* 38, n. 3, pp. 277–283, 1994.

[18] Pisier, G.: Remarques sur un resultat non publié de B. Maurey. *Seminaire d'Analyse Fonctionelle*, vol. I, no. 12. École Polytechnique, Centre de Mathématiques, Palaiseau, 1980-81.