**Tight Bounds on Rates of Variable-Basis Approximation via Estimates of Covering Numbers**

Kůrková, Věra
2001

**Institute of Computer Science**
**Academy of Sciences of the Czech Republic**

# Tight bounds on rates of variable-basis approximation via estimates of covering numbers

Věra Kůrková and Marcello Sanguineti

Technical report No. 865

December 2001

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: (+4202) 6605 1111, fax: (+4202) 86 58 57 89,
e-mail:ics@cs.cas.cz

**Institute of Computer Science**
**Academy of Sciences of the Czech Republic**

# Tight bounds on rates of variable-basis approximation via estimates of covering numbers [1]

Věra Kůrková[2] and Marcello Sanguineti[3]

Technical report No. 865

December 2001

Abstract:

Computationally effective approximation schemes for high-dimensional optimization tasks are investigated. For nonlinear approximation of variable-basis type, there are studied sets of multivariable functions that can be approximated without "curse of dimensionality". There are given conditions on a set of basis functions that do not allow a possibility of improving an upper bound on accuracy of approximation of the order of $\mathcal{O}(n^{-(1/2)})$ (where $n$ is the number of basis functions). Tightness of the bounds is derived using estimates of covering numbers. The results are applied to a class nonlinear approximators used in distributed computing.

Keywords:
high-dimensional optimization, curse of dimensionality, rates of approximation, covering numbers, distributed computing, neural networks.

[1]The authors were supported by NATO under Grant PST.CLG.976870 (Project "Approximation and Functional Optimization by Neural Networks"). V. Kůrková was supported in part by GA ČR Grant 201/99/0092. M. Sanguineti was supported in part by the Italian Ministry of University and Research (Project "New Techniques for the Identification and Adaptive Control of Industrial Systems").

[2]Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, P.O. Box 5 – 182 07, Prague 8, Czech Republic – vera@cs.cas.cz

[3]Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy – marcello@dist.unige.it

# 1 Introduction

When an optimization task cannot be solved in an analytical form, one has to search for an approximate solution. Often, such a solution depends on a large number of variables. Feasibility of high-dimensional approximation is limited by the so-called "curse of dimensionality" [3] (i.e., an unfeasibly fast growth of the computational load with the number of variables).

Many high-dimensional problems have been effectively solved using nonlinear approximation methods such as neural networks (see, e.g., [27], [4], [28]). Their efficiency has motivated a theoretical analysis of desirable computational capabilities of nonlinear approximators, guaranteeing that their complexity does not increase too fast with the dimensionality of the problem.

Some insight into properties of sets of multivariable functions admitting such an effective approximation has been obtained by Maurey (see [26]), Jones [12] and Barron [1], [2]. They constructed approximants with rates of convergence of the order of $\mathcal{O}(n^{-1/2})$, which in contrast to curse of dimensionality rates of the order of $\mathcal{O}(n^{-1/d})$, do not depend on the number $d$ of variables of the functions to be approximated. 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 all of $n$-tuples of elements of a given set of basis functions (in contrast to classical linear approximators, which use linear combinations of the first $n$ elements of a basis with a fixed ordering). The variable-basis approximation scheme has been widely investigated (see, e.g., DeVore and Temlyakov [7] and the references therein): it includes splines with free nodes, trigonometric polynomials with free frequencies, as well as feedforward neural networks.

The upper bound of the order of $\mathcal{O}(n^{-1/2})$ on variable-basis approximation has been improved and extended by several authors (see, e.g., Darken et al. [6], Girosi [10], Gurvits and Koiran [11], Makovoz [21], [22], Kůrková, Savický and Hlaváčková [18]). Makovoz [21] improved Maurey's argument (see [26]) by combining it with a concept from metric entropy theory and proved that for a class of neural networks that is widely used in applications, Maurey-Jones-Barron's upper bound cannot be improved to $\mathcal{O}(n^{-\alpha})$ for $\alpha > 1/2 + 1/d$. A similar tightness result was earlier obtained by Barron [1], who used a different proof technique. For the special case of orthonormal variable-bases, Mhaskar and Micchelli [23], Kůrková, Savický and Hlaváčková [18] and Kůrková and Sanguineti [16] have derived tight improvements, the order of $\mathcal{O}((n-1)^{-1/2})$ of Maurey-Jones-Barron's bound.

In this paper, we extend Makovoz's method of comparison of covering numbers, to more general bases. We investigate tightness of the upper bound $\mathcal{O}(n^{-1/2})$ for a basis satisfying two conditions: (1) polynomial growth of its covering number and (2) sufficient "capacity" of the basis, in the sense that its symmetric convex hull has either an orthonormal subset or an orthogonal one that for each positive integer $k$ contains at least $k^d$ functions with norms greater or equal to $1/k$. We show that for such bases, Maurey-Jones-Barron's upper bound cannot be improved beyond $\mathcal{O}(n^{-(1/2)})$ or $\mathcal{O}(n^{-(1/2+1/d)})$, resp. We apply these tightness results to perceptron neural networks.

The paper is organized as follows. Section 2 describes basic concepts and notations concerning approximation by variable-basis functions and presents Maurey-Jones-Barron's upper bound $\mathcal{O}(n^{-1/2})$ in terms of a norm, tailored to a type of basis. Section 3 contains estimates of covering numbers of balls in such norms. Section 4 explores tightness of the upper bound $\mathcal{O}(n^{-1/2})$. In Section 5 the results are applied to neural networks.

# 2 Rates of variable-basis approximation

Let $(X, \|.\|)$ be a normed linear space (when it is clear from the context which norm is used, we shall simply write $X$), $B_r(f, \|.\|)$ denotes the ball in the norm $\|.\|$, with radius $r$ and centered at $f \in X$, i.e., $B_r(f, \|.\|) = \{h \in X : \|h - f\| \leq r\}$. We write shortly $B_r(\|.\|)$ instead of $B_r(0, \|.\|)$. We call a subset of a Hilbert space *orthogonal* (*orthonormal*) if its elements are

pairwise orthogonal (orthonormal, resp.). $\mathcal{R}$ denotes the set of real numbers and $\mathcal{N}_+$ the set of positive integers.

Let $G$ be a subset of $(X, \|.\|)$. For $c \in \mathcal{R}$, $cG = \{cg : g \in G\}$ and, for $c$ positive, $G(c) = \{wg : g \in G, w \in \mathcal{R} \ \& \ |w| \leq c\}$. The *closure* of $G$ is denoted by $cl\, G$. $G$ is *dense* in $(X, \|.\|)$ if $cl\, G = X$.

The *linear span* of $G$ is denoted by $span\, G$; $span_n G$ denotes the set of all linear combinations of at most $n$ elements of $G$, i.e., $span_n G = \{\sum_{i=1}^n w_i g_i : \ w_i \in \mathcal{R}, \ g_i \in G\}$. The *convex hull* of $G$, denoted by $conv\, G$, is the set of all convex combinations of its elements; $conv_n\, G$ denotes the set of all convex combinations of $n$ elements of $G$, i.e., $conv_n\, G = \left\{\sum_{i=1}^n a_i g_i : a_i \in [0,1], \sum_{i=1}^n a_i = 1, g_i \in G\right\}$.

In *linear approximation* the approximating functions belong to a *linear subspace*, which is often generated by the first $n$ elements of a given linearly ordered set. For example, the set of all polynomials of order at most $n-1$ is generated by the first $n$ elements of the set $\{x^{i-1} : \ i \in \mathcal{N}_+\}$. Such an approximation scheme can be called *fixed-basis approximation*, while in *variable-basis approximation* the approximating functions are linear combinations of all $n$-tuples of elements of a given set $G$. In this case, the approximating set is the *union of all finite-dimensional subspaces* generated by all $n$-tuples of elements of $G$, i.e., it corresponds to the set $span_n G$ of all linear combinations of at most $n$ elements of $G$. Variable-basis approximation scheme includes splines with free nodes, trigonometric polynomials with free frequencies and feedforward neural networks (see, e.g., [7] and the references therein).

In practical applications, vectors of coefficients of the linear combinations of basis functions are bounded in some norm $\|.\|_*$ on $\Re^n$. So for $c > 0$, they belong to the set $\{\sum_{i=1}^n w_i\, g_i : g_i \in G, \|w\|_* \leq c\}$, where $w = (w_1, \ldots, w_n) \in \Re^n$. Since all norms on $\Re^n$ are equivalent, there exists $c' > 0$ such that $\|.\|_{l_1} \leq c'\, \|.\|_*$, where $\|.\|_{l_1}$ denotes the $l_1$ norm on $\Re^n$. Hence $\{\sum_{i=1}^n w_i\, g_i : g_i \in G, \|w\|_* \leq c\} \subseteq conv_n\, G(c\, c')$. Indeed, if $f = \sum_{i=1}^n w_i g_i$, where $\|w\|_* \leq c$, then $f$ can be expressed as $f = \sum_{i=1}^n \frac{|w_i|}{\|w\|_{l_1}} \mathrm{sgn}(w_i) \|w\|_{l_1}\, g_i$, where $\mathrm{sgn}(w_i) = \frac{w_i}{|w_i|}$. As $\sum_{i=1}^n \frac{|w_i|}{\|w\|_{l_1}} = 1$ and, for all $i = 1, \ldots, n$, $\frac{|w_i|}{\|w\|_{l_1}} \in [0,1]$ and $|\mathrm{sgn}(w_i)\|w\|_{l_1}| \leq c\, c'$, $f \in conv_n\, G(c\, c')$. Thus it is useful to investigate approximation by sets of convex combinations of variable-basis functions.

Often a class of approximating functions is represented as the union of a nested sequence $\{M_n : \ n \in \mathcal{N}_+\}$ of sets of functions of increasing complexity. The *rate of approximation* of a function $f \in X$ is measured by the decrease of $\{\|f - M_n\| : \ n \in \mathcal{N}_+\}$. Density of $\bigcup_{n \in \mathcal{N}_+} M_n$ in $(X, \|.\|)$ guarantees that, for each $f \in X$, the sequence $\{\|f - M_n\| : \ n \in \mathcal{N}_+\}$ converges to 0. To obtain a desired approximation accuracy for $n$ small enough, the rate of approximation has to be sufficiently fast. A major limitation in multivariable approximation is the "curse of dimensionality" [3], a slow rate of approximation of the order of $\mathcal{O}(n^{-1/d})$, where $d$ is the number of variables (see, e.g., [25, pp. 232-233]).

In fixed-basis approximation, the nested sets $M_n$ are $n$-dimensional subspaces, while in variable-basis approximation, $M_n$ correspond to $span_n\, G$ or $conv_n\, G$. Since $span_n\, G$ is the union of all at most $n$-dimensional subspaces spanned by $n$-tuples of elements of $G$, it is much larger than a single linear subspace and so it might allow faster rates than rates of linear approximation. Similarly, approximation by $conv_n\, G(c)$ might allow better rates for sufficiently large $c$.

Description of sets of functions of $d$ variables that do not exhibit the curse of dimensionality in variable-basis approximation can be derived from the following theorem by Maurey (see [26]), Jones [12] and Barron [2].

**Theorem 2.1** *Let $(X, \|.\|)$ be a Hilbert space, $G$ its bounded subset, $s_G = \sup_{g \in G} \|g\|$, and $f \in cl\, conv\, G$. Then, for every positive integer $n$,*

$$\|f - conv_n\, G\| \leq \sqrt{\frac{s_G^2 - \|f\|^2}{n}}.$$

Note that this theorem gives an upper bound of the order of $\mathcal{O}(n^{1/2})$ for any number $d$ of variables of functions in $X$. Some authors even called it "dimension-independent", which is misleading since with $d$ increasing, sets $cl\,conv\,G$ might be more and more constrained (see [18]).

As $conv_n\,G \subseteq span_n\,G$, the upper bound from Maurey-Jones-Barron's theorem also applies to rates of approximation by $span_n\,G$. However, when $G$ is not closed under multiplication by scalars $cl\,conv\,G$ 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 $X$. However, replacing $G$ by $G(c) = \{wg; w \in \mathcal{R}, |w| \leq c, g \in G\}$, for any $c > 0$, we get $conv_n\,G(c) \subseteq span_n G(c) = span_n G$ and so we can apply Theorem 2.1 to all elements of $\cup_{c\in\mathcal{R}_+} cl\,conv\,G(c)$. This approach can be formulated in terms of a norm tailored to a set $G$.

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 $cl\,conv\,(G \cup -G)$, i.e.,

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

$G$-variation is a norm on $\{f \in X : \|f\|_G < \infty\} \subseteq X$. It has been introduced in [15] as an extension of the concept from [1] of variation with respect to half-spaces (which was motivated by neural networks). $G$-variation depends on the norm $\|.\|$ on $X$ but to simplify the notation we write $\|.\|_G$, assuming that it is clear with respect to which norm $G$-variation is defined.

As $conv\,(G \cup -G) = conv\,G(1)$, we have $\|f\|_G = \inf\{c \in \mathcal{R}_+ : \ f \in cl\,conv\,G(c)\}$ and the unit ball in $G$-variation, $B_1(\|.\|_G)$, is equal to $cl\,conv\,(G \cup -G)$. For $G$ orthonormal, $G$-variation is equal to the $l_1$-norm with respect to $G$ [16]. 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 [1].

The following upper bound is a reformulation of Maurey-Jones-Barron's theorem in terms of $G$-variation [15].

**Theorem 2.2** *Let $(X, \|.\|)$ be a Hilbert space, $G$ its bounded subset, $s_G = \sup_{g\in G} \|g\|$. Then, for every $f \in X$ and every positive integer $n$,*

$$\|f - span_n\,G\| \leq \sqrt{\frac{(s_G\,\|f\|_G)^2 - \|f\|^2}{n}}.$$

This upper bound on distance from $span_n\,G$ can be applied to various types of variable-basis approximation, such as splines with free nodes, trigonometric polynomials with free frequencies, as well as feedforward neural networks (see, e.g., [17] and the references therein).

The worst-case error in approximation of functions from a set $B$ by elements of an approximating set $M$ is formalized by the concept of *deviation of $B$ from $M$*, defined as

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

Maurey-Jones-Barron's theorem implies an upper bound on deviation of a ball in $G$-variation.

**Corollary 2.3** *Let $(X, \|.\|)$ be a Hilbert space, $G$ be its bounded subset, and $s_G = \sup_{g\in G} \|g\|$. Then, for every every positive integer $n$,*

$$\delta(B_1(\|.\|_G), conv_n\,(G \cup -G)) \leq \frac{s_G}{\sqrt{n}}.$$

In Section 4 we shall investigate tightness of the upper bound from Corollary 2.3. Assuming that a faster rate is possible, we shall derive a contradiction with properties of covering numbers of the unit ball in $G$-variation for some bases $G$ that include those frequently used in neural networks.

## 3 Estimates of covering numbers of balls in $G$-variation

In this section we derive estimates of covering numbers of balls in $G$-variation using tools from metric entropy theory and properties of generalized Hadamard matrices.

Recall that, for $\varepsilon > 0$, the *$\varepsilon$-covering number* of a subset $K$ of a normed linear space $(X, \|.\|)$ is defined as

$$cov_\varepsilon(K, \|.\|) = \min\{m \in \mathcal{N}_+ \,:\, K \subseteq \cup_{i=1}^m B_\varepsilon(f_i, \|.\|),\ f_i \in K,\ \ i = 1, \ldots, m\}$$

if the set over which the minimum is taken is nonempty, otherwise $cov_\varepsilon(K, \|.\|) = +\infty$. A subset $\{f_1, \ldots, f_m\}$ of $K$ is called *$\varepsilon$-separated* if for each distinct pair $f_i, f_j$ of its elements, $\|f_i - f_j\| \geq \varepsilon$. The *$\varepsilon$-packing number* of $K$, $pack_\varepsilon(K, \|.\|)$, is defined as the maximal cardinality of a $2\,\varepsilon$-separated subset of $K$. When it is clear from the context which norm is considered, we shall simply write $cov_\varepsilon(K)$ and $pack_\varepsilon(K)$ instead of $cov_\varepsilon(K, \|.\|)$ and $pack_\varepsilon(K, \|.\|)$, resp. By the definitions and the triangle inequality, $pack_\varepsilon(K, \|.\|) \leq cov_\varepsilon(K, \|.\|) \leq pack_{\varepsilon/2}(K, \|.\|)$.

To estimate deviations of balls in $G$-variation, we need the following lemmas. The first one is an elementary estimate of covering numbers of balls in a norm on $\mathcal{R}^d$, following directly from a volume ratio argument.

**Lemma 3.1** *Let $d$ be a positive integer, $\|.\|$ be a norm on $\mathcal{R}^d$ and, $\varepsilon > 0$. Then $(1/\varepsilon)^d \leq cov_\varepsilon(B_1(\|.\|)) \leq (2/\varepsilon)^d$.*

**Proof.** Let *vol* denote the Euclidean volume in $\mathcal{R}^d$, then for every $\varepsilon > 0$, $vol(B_\varepsilon(\|.\|)) = \varepsilon^d\, vol(B_1(\|.\|))$. By the definitions of $\varepsilon$-covering and $\varepsilon$-packing numbers,

$$pack_\varepsilon(B_1(\|.\|))\, vol(B_\varepsilon(\|.\|)) \leq vol(B_1(\|.\|)) \leq cov_\varepsilon(B_1(\|.\|))\, vol(B_\varepsilon(\|.\|)).$$

Hence, $pack_\varepsilon(B_1(\|.\|)) \leq (1/\varepsilon)^d \leq cov_\varepsilon(B_1(\|.\|))$. As for every $K \subset \Re^d$, $pack_\varepsilon(K) \leq cov_\varepsilon(K) \leq pack_{\varepsilon/2}(K)$, we have $cov_\varepsilon(B_1(\|.\|)) \leq pack_{\varepsilon/2}(B_1(\|.\|)) \leq (2/\varepsilon)^d$. Thus, $(1/\varepsilon)^d \leq cov_\varepsilon(B_1(\|.\|)) \leq (2/\varepsilon)^d$. □

The second lemma summarizes some relationships among covering numbers of the sets $G$, $G \cup -G$, and $conv_n\, G$.

**Lemma 3.2** *Let $(X, \|.\|)$ be a normed linear space, $G$ its bounded subset, $s_G = \sup_{g \in G} \|g\|$, and let $l_1^n$ denote the $l_1$ norm on $\Re^n$. Then, for every positive integer $n$ and every $\varepsilon > 0$,*
*(i) $cov_{\varepsilon(1+s_G)}(conv_n\, G) \leq (cov_\varepsilon\, G)^n\, cov_\varepsilon(B_1(\|.\|_{l_1^n}), \|.\|_{l_1^n})$;*
*(ii) $cov_{\varepsilon(1+s_G)}(conv_n\, G) \leq (cov_\varepsilon\, G)^n\ (2/\varepsilon)^n$;*
*(iii) $cov_\varepsilon(G \cup -G) \leq 2\, cov_\varepsilon\, G$.*

**Proof.**
(i) Let $B$ be an $\varepsilon$-net in $B_1(\|.\|_{l_1^n})$ with respect to $l_1^n$ and $A$ be an $\varepsilon$-net in $G$ with respect to $\|.\|$. Let $C$ be a subset of $conv_n\, G$ formed by all elements $\sum_{i=1}^n b_i\, g_i$, where $(g_1, \ldots, g_n) \in A^n$ and $(b_1, \ldots, b_n) \in B$, i.e., $C = \left\{\sum_{i=1}^n b_i\, g_i \,:\, (g_1, \ldots, g_n) \in A^n\,,\, (b_1, \ldots, b_n) \in B\right\}$. We show that $C$ is $\varepsilon\,(1+s_G)$-net in $conv_n\, G$. Let $\sum_{i=1}^n \bar{b}_i\, \bar{g}_i \in conv_n\, G$. Since $B$ is an $\varepsilon$-net in $B_1(\|.\|_{l_1^n})$ with $l_1^n$ norm, there exist $(b_1, \ldots, b_n) \in B$ such that $\sum_{i=1}^n (b_i - \bar{b}_i) \leq \varepsilon$. As $A$ is an $\varepsilon$-net in $G$ with $\|.\|$, there exist $(g_1, \ldots, g_n) \in A^n$ such that for every $i = 1, \ldots, n$, $\|g_i - \bar{g}_i\| \leq \varepsilon$. Thus

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

As $card\, C = (card\, A)^n\, card\, B$, (i) holds.
(ii) follows directly from (i) and Lemma 3.1.
(iii) holds since for $A$ an $\varepsilon$-net in $G$, $-A$ in an $\varepsilon$-net in $-G$. □

In [21, Lemma 3] there was derived a lower bound on covering numbers of balls in $A$-variation, for $A$ satisfying a weakened orthogonality condition. More precisely, it was shown that $cov_{1/\sqrt{m}}\, B_1(\|.\|_A) \geq 2^{c\,m}$, where $m = card\, A$ and $c$ is a positive constant. The proof of this exponential lower bound in [21, Lemma 3] exploits a result from [19] (see also [20, p. 489, Lemma 2.2]) on the exponential growth of the number of vectors in $\{-1,1\}^m$ that differ in more than $m/8$ entries. The following lemmas improve the bound from [21, Lemma 3], allowing more sizes of diameters of covering sets. Their proofs follow similar steps as [21] but they use stronger tools, namely properties of generalized Hadamard matrices.

Recall that a *Hadamard matrix* of order $m$ is a set of pairwise orthogonal vectors in the *Hamming cube* $\{-1,1\}^m$ with a particular ordering. The concept of Hadamard matrix has been generalized in [13] by allowing a certain tolerance in the orthogonality condition: for $\varepsilon \in [0,1]$, an *$\varepsilon$-Hadamard matrix* of order $m$ is 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 $\varepsilon$-Hadamard matrix of order $m$. If $\varepsilon = s/m$, then $|u \cdot v| \leq s$. The weakened orthogonality condition can also be described in terms of *Hamming distance*, denoted by $h$ and defined as the number of coordinates at which two vectors differ. It is equal to $1/2$ of the $l_1$ norm, i.e., $h(u,v) = (1/2)\sum_{i=1}^m |u_i - v_i|$ for $u, v \in \{-1,1\}^m$. It is easy to check that, for each two distinct vectors $u, v$ in an $\varepsilon$-Hadamard matrix of order $m$, the Hamming metric satisfies $h(u,v) \geq m\,(1-\varepsilon)/2$. If $\varepsilon = s/m$, then $h(u,v) \geq (m-s)/2$.

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

**Lemma 3.3** *Let $(X, \|.\|)$ be a Hilbert space, $A$ its orthogonal subset, $card\, A = m$ and $\min_{g\in A} \|g\| \geq a$. Then for any integer $s$ such that $1 \leq s < m$ and $\delta_s = \frac{a}{m}\sqrt{\lceil \frac{m-s}{2} \rceil}$,*

$$cov_{\delta_s}(B_1(\|.\|_A)) \geq R\left(\frac{s}{m}, m\right).$$

**Proof.** Let $A = \{g_1, \ldots, g_m\}$ and $M_s$ be an $(s/m)$-Hadamard matrix of the order $m$ with $R\left(\frac{s}{m}, m\right)$ rows. To verify that the set $A(M_s) = \left\{\frac{1}{m}\sum_{i=1}^m u_i g_i \;:\; u \in M_s\right\}$ is $2\delta_s = \frac{2a}{m}\sqrt{\lceil \frac{m-s}{2} \rceil}$-separated, we have to show that for any two distinct vectors $u, v \in M_s$, $\left\| \frac{1}{m}\sum_{i=1}^m u_i g_i - \frac{1}{m}\sum_{i=1}^m v_i g_i \right\| \geq 2\delta_s$. By the definition of an $(s/m)$-Hadamard matrix, we have $h(u,v) \geq \frac{m-s}{2}$ and so the cardinality of the set $I$ of coordinates at which $u$ and $v$ differ is at least $\lceil (m-s)/2 \rceil$. Thus $\left\| \frac{1}{m}\sum_{i=1}^m (u_i - v_i) g_i \right\| = \frac{2}{m}\left\| \sum_{i\in I} g_i \right\| \geq \frac{2a}{m}\sqrt{\left\lceil \frac{m-s}{2} \right\rceil}$. Since $card\, A(M_s) = R\left(\frac{s}{m}, m\right)$ and $A(M_s) \subset B_1(\|.\|_A)$, we obtain $cov_{\delta_s}(B_1(\|.\|_A)) \geq R\left(\frac{s}{m}, m\right)$. □

A similar lower bound can be derived even if the orthogonality condition on the set $A$ is relaxed to $\varepsilon$-nearly orthogonality. A subset $A = \{g_1, \ldots, g_m\}$ of a Hilbert space $(X, \|.\|)$ is called *$\varepsilon$-nearly orthogonal* if $\sum_{j=1,\, j\neq i}^m |g_i \cdot g_j| \leq \varepsilon, \;\; i = 1, \ldots m$.

**Lemma 3.4** *Let $(X, \|.\|)$ be a Hilbert space, $A$ be its $\varepsilon$-nearly orthogonal subset such that $card A = m$ and $\min_{g\in A} \|g\| \geq a$, and let $\varepsilon \leq \sqrt{a}$. Then for each integer $s$ such that $1 \leq s < m$ and $\varepsilon_s = \frac{\sqrt{a^2 - \varepsilon}}{m}\sqrt{\lceil \frac{m-s}{2} \rceil}$,*

$$cov_{\varepsilon_s}(B_1(\|.\|_A)) \geq R\left(\frac{s}{m}, m\right).$$

**Proof.** As in the proof of Lemma 3.3, define the set $A(M_s) = \left\{\frac{1}{m}\sum_{i=1}^m u_i g_i; u \in M_s\right\}$. To show that it is $2\varepsilon_s = \frac{2\sqrt{a^2-\varepsilon}}{m}\sqrt{\lceil\frac{m-s}{2}\rceil}$-separated, we estimate from below the distance $\|\frac{1}{m}\sum_{i=1}^m u_i g_i - \frac{1}{m}\sum_{i=1}^m v_i\, g_i\|$ for any pair of distinct vectors $u, v \in M_s$. Let $I$ denote the set of coordinates in which $u$ and $v$ differ, $k = card\, I$, and $\zeta_i = \frac{1}{2\sqrt{k}}(u_i - v_i),\ i \in I$. Then $\zeta_i = \pm\frac{1}{\sqrt{k}}$, and $\left\|\frac{1}{m}\sum_{i=1}^m (u_i - v_i) g_i\right\| = \frac{1}{m}\left\|\sum_{i\in I} g_i\right\| = \frac{2\sqrt{k}}{m}\left\|\sum_{i=1}^k \zeta_i\, g_i\right\|$. Moreover, $\left\|\sum_i^k \zeta_i\, g_i\right\|^2 = \left|\sum_{i=1}^k \sum_{j=1}^k \zeta_i\,\zeta_j\, g_i\, g_j\right|$. Since $\sum_{i=1}^k \zeta_i^2 = 1$, it is sufficient to derive a lower bound on the function $\Delta(\zeta_1,\ldots,\zeta_k) = \left|\sum_{i=1}^k \sum_{j=1}^k \zeta_i\,\zeta_j\, g_i\, g_j\right|$ on the unit sphere $S_1$ in $l_2$ norm of $\mathcal{R}^k$. Let $D_I$ be the $k\times k$ matrix defined by $D_{I\,ij} = g_i\, g_j$. Then $\Delta(\zeta_1,\ldots,\zeta_k) \ge \sqrt{|\lambda_{\min}(D_I)|}$ in $S_1$, where $\lambda_{\min}(D_I)$ denotes the minimun eigenvalue of $D_I$. As $|\lambda_{\min}(D_I)| \ge \left|\min_{g_i\in A}\|g_i\|^2 - \sum_{i\in I,\, i\ne j} |g_i \cdot g_j|\right| \ge a^2 - \varepsilon$, we obtain $\frac{1}{m}\left\|\sum_{i=1}^m (u_i - v_i) g_i\right\| \ge \frac{2\sqrt{k\,(a^2-\varepsilon)}}{m} \ge \frac{2\sqrt{a^2-\varepsilon}}{m}\sqrt{\lceil\frac{m-s}{2}\rceil}$. □

The next lemma follows immediately from Lemma 3.3 combined with the lower bound on $R\,(s/m, m)$ from [13, Theorem 3.4].

**Lemma 3.5** *Let $(X, \|.\|)$ be a Hilbert space, $A$ its orthogonal subset such that $card\, A = m$ and $\min_{g\in A}\|g\| \ge a$. Then for any integer $s$ such that $1 \le s \le m-2$,*

$$cov_{\delta_s}(B_1(\|.\|_A)) \ge \frac{2^{m-1}}{B(\lambda_{m,s}, m)},$$

*where $\lambda_{m,s} = \lceil\frac{m-s-2}{2}\rceil,\ B(\lambda, m) = \sum_{i=0}^{\lambda}\binom{m}{i}$ and $\delta_s = \frac{a}{m}\sqrt{\lceil\frac{m-s}{2}\rceil}$.*

Using the binary entropy function $\mathcal{H}(p) = -p\log_2(p) - (1-p)\log_2(1-p),\ 0 < p < 1$, we obtain the following lower bound.

**Lemma 3.6** *Let $(X, \|.\|)$ be a Hilbert space, $A$ be its orthogonal subset of cardinality $m \ge 3$, such that $\min_{g\in A}\|g\| \ge a$. Then*

$$cov_{a/(2\sqrt{m})}(B_1(\|.\|_A)) \ge 2^{b\,m-1},$$

*where $b = 1 - \mathcal{H}(\frac{1}{4})$.*

**Proof.** For $s = \lfloor\frac{m}{2}\rfloor$, $\delta_s$ from Lemma 3.5 is equal to $\frac{a}{m}\sqrt{\lceil\frac{m-s}{2}\rceil} = \frac{a}{m}\sqrt{\left\lceil\frac{m-\lfloor\frac{m}{2}\rfloor}{2}\right\rceil} \ge \frac{a}{m}\sqrt{\left\lceil\frac{m-\frac{m}{2}}{2}\right\rceil} \ge \frac{a}{m}\sqrt{\frac{m}{4}} = \frac{a}{2\sqrt{m}}$. Thus $cov_{a/(2\sqrt{m})}(B_1(\|.\|_A)) \ge cov_{\delta_{\lfloor m/2\rfloor}}(B_1(\|.\|_A))$. By Lemma 3.5, $cov_{\delta_{\lfloor m/2\rfloor}}(B_1(\|.\|_A)) \ge 2^{m-1}\,/\,B\big(\lambda_{m,\lfloor m/2\rfloor}, m\big)$. As $\lambda_{m,\lfloor m/2\rfloor} = \left\lceil\frac{m-\lfloor\frac{m}{2}\rfloor-2}{2}\right\rceil = \left\lceil\frac{\frac{m}{2}-2}{2}\right\rceil \le \frac{m}{4}$, we can use the estimate $B(\lambda, m) \le 2^{m\mathcal{H}(\lambda/m)}$, which is valid for $\lambda < m/2$ [8, p. 44]. As the entropy function $\mathcal{H}$ is increasing on the interval $(0, 1/2)$, we have

$$cov_{a/(2\sqrt{m})}(B_1(\|.\|_A)) \ge \frac{2^{m-1}}{2^{m\,\mathcal{H}\left(\frac{\lambda_{m,\lfloor m/2\rfloor}}{m}\right)}} \ge 2^{m-1}\,2^{-m\,\mathcal{H}(1/4)} = 2^{m\,(1-\mathcal{H}(1/4))-1} = 2^{mb-1},$$

where $b = 1 - \mathcal{H}(1/4) \simeq 0.085$. □

In some cases of interest (see Section 5), the unit ball in $G$ variation has no orthonormal subset, but it has an orthogonal subset that contains for each positive integer $k$ "sufficiently many" elements with norms greater than or equal to $1/k$. In [17] there was defined the concept

of a set *not quickly vanishing with respect to* a positive integer $d$ as a subset $A$ of a normed linear space $(X, \|.\|)$ such that $A = \cup_{k\in\mathcal{N}_+} A_k$, where, for each $k \in \mathcal{N}_+$, $card\, A_k \geq k^d$ and for each $h \in A_k$, $\|h\| \geq 1/k$. The last lemma in this series follows directly from Lemma 3.6.

**Lemma 3.7** *Let $(X, \|.\|)$ be a Hilbert space, $A$ be its orthogonal subset not quickly vanishing with respect to a positive integer $d$, and $r > 0$. Then, for every positive integer $k \geq 3$,*

$$cov_{\varepsilon_k}(B_1(\|.\|_{\frac{1}{r}A})) \geq 2^{b\,k^d - 1},$$

*where $\varepsilon_k = \frac{1}{2\,r\,k^{d/2+1}}$ and $b = 1 - \mathcal{H}(1/4)$.*

# 4 Tightness of the upper bound $O(n^{-1/2})$

We shall investigate tightness of Maurey-Jones-Barron's upper bound from Corollary 2.3 by assuming that there exists a better bound and deriving its consequences on the behavior of certain covering numbers of balls in $G$-variation. These numbers must be "small" when such a hypothetical better bound exists and $\varepsilon$-covering number of $G$ grows at most polynomially with $1/\varepsilon$. On the other hand, when a ball in $G$-variation contains either an infinite orthonormal or an orthogonal set not quickly vanishing with respect to $d$, such covering numbers must be "large". So it is not possible to improve Maurey-Jones-Barron's upper bound when covering numbers of $G$ grow polynomially and $G$ contains an infinite orthogonal subset with "sufficiently large" norms.

For $f, g : \mathcal{N}_+ \to \mathcal{N}_+$, $g(k) = \mathcal{O}(f(k))$ means that there exists $c > 0$ such that, for all but finitely many $k \in \mathcal{N}_+$, $g(k) = c\, f(k)$. Analogously, $g(k) \leq \mathcal{O}(f(k))$ means that there exists $c > 0$ such that, for all but finitely many $k \in \mathcal{N}_+$, $g(k) \leq c\, f(k)$. First, we consider the case when a ball in $G$-variation contains an infinite orthonormal set.

**Theorem 4.1** *Let $(X, \|.\|)$ be a Hilbert space and $G$ be its bounded subset such that*
*(1) there exists a polynomial $p$ such that, for every $\varepsilon > 0$, $cov_\varepsilon(G) \leq \mathcal{O}\big(p\big(1/\varepsilon\big)\big)$;*
*(2) there exists $r > 0$ such that $B_r(\|.\|_G)$ contains an infinite orthonormal set $A$.*
*Then*

$$\delta(B_1(\|.\|_G), conv_n(G \cup -G)) = \mathcal{O}(n^{-\frac{1}{2}}).$$

**Proof.** Assume that there exist $\alpha > 1/2$ and $c > 0$ such that, for all but finitely many $n \in \mathcal{N}_+$,

$$\delta(B_1(\|.\|_G), conv_n\,(G \cup -G)) \leq \frac{c}{n^\alpha}. \tag{4.1}$$

Set $s_G = \sup_{g\in G} \|g\|$ and $\eta = \frac{(1+s_G)c}{n^\alpha}$. We shall derive a contradiction by comparing an upper bound on $cov_\eta B_1(\|.\|_G)$ (obtained from the assumption (1) and this hypothetical upper bound) with a lower bound on the same covering number (obtained from the assumption (2) and Lemma 3.6).

By (4.1), the triangle inequality, and Lemma 3.2 (ii) and (iii), we get

$$cov_\eta B_1(\|.\|_G) \leq cov_{\eta/2} conv_n\,(G \cup -G) \leq (2\, cov_{\eta/(2(1+s_G))} G)^n \left(\frac{4(1+s_G)}{\eta}\right)^n,$$

which gives an upper bound on covering numbers of balls in $G$-variation

$$cov_{(1+s_G)c/n^\alpha}(B_1(\|.\|_G)) \leq \left(\frac{8}{c}\right)^n n^{\alpha n} (cov_{c/2n^\alpha} G)^n. \tag{4.2}$$

On the other hand, using assumption (2) set $A_r = (1/r)A$, where $A$ is orthonormal and $A_r \subset B_1(\|.\|_G)$. By Lemma 3.6, for each positive integer $m \geq 3$ we have

$cov_{1/(2r\sqrt{m})}(B_1(\|.\|_G)) \geq 2^{b\,m-1}$. Set $\bar{m} = (1/2\eta\,r)^2$. If $m \leq \bar{m}$, then $\eta = (1+s_G)c/n^\alpha \leq 1/(2r\sqrt{m})$ and so $cov_\eta(B_1(\|.\|_G)) \geq cov_{1/(2r\sqrt{m})}(B_1(\|.\|_G))$. As for every real number $x \geq 2$, $\lfloor x \rfloor \geq x-1 \geq x/2$, we have $2^{b\lfloor \bar{m} \rfloor - 1} \geq 2^{b\,\bar{m}/2-1}$. Thus, we obtain the following lower bound

$$2^{c_d\,n^{2\alpha}-1} \leq cov_{(1+s_G)c/n^\alpha}(B_1(\|.\|_G)), \tag{4.3}$$

where $c_d = (b/2)\left(2\,(1+s_G)\,c\,r\right)^{-2}$.

The inequalities (4.2) and (4.3) give for all but finitely many $n \in \mathcal{N}_+$

$$2^{c_d\,n^{2\alpha}-1}c \leq \left(\frac{8}{c}\right)^n n^{\alpha n}(cov_{c/2n^\alpha}G)^n. \tag{4.4}$$

Taking the logarithm of both sides of (4.4), we get

$$c_d\,n^{2\alpha} - 1 \leq 3n - n\,\log_2\,c + \alpha\,n\,\log_2 n + n\,\log_2\left(cov_{c/2n^\alpha}G\right). \tag{4.5}$$

For $\alpha > 1/2$, the left-hand side of this inequality has order of infinity $2\alpha > 1$. Hence, the right-hand side must have order larger than 1. Its first two terms have order 1 and its third term has order smaller than every real number larger than 1. Hence, if (4.5) holds, the third term of the right-hand side must satisfy $\mathcal{O}(n^{2\alpha}) \leq n\,\log_2\,(cov_{c/2n^\alpha}\,G)$. Setting $\varepsilon = c/2n^\alpha$, we get $\mathcal{O}\big(2^{(c/2\varepsilon)^\beta}\big) \leq cov_\varepsilon\,G$, where $\beta = (2\alpha-1)/\alpha > 0$ (as $\alpha > 1/2$). This contradicts the assumption (1). □

Even when no ball in $G$-variation is "large enough" to contain an infinite orthonormal subset, there might exist a ball containing an orthogonal subset, not quickly vanishing with respect to a positive integer $d$ (see Section 5 for examples of such sets).

**Theorem 4.2** *Let $d$ a positive integer, $(X, \|.\|)$ be a Hilbert space and $G$ be its bounded subset satisfying the following conditions:*
*(1) there exists a polynomial $p$ such that for every $\varepsilon > 0$, $cov_\varepsilon(G) \leq \mathcal{O}\Big(p\big(\frac{1}{\varepsilon}\big)\Big)$;*
*(2) there exists $r > 0$ such that $B_r(\|.\|_G)$ contains a set of orthogonal elements which is not quickly vanishing with respect to $d$.*
*Then*

$$\delta(B_1(\|.\|_G), conv_n(G \cup -G)) \leq \mathcal{O}(n^{-\alpha}) \quad implies \quad \alpha \leq \frac{1}{2} + \frac{1}{d}.$$

**Proof.** Assume that there exists $\alpha > \frac{1}{2} + \frac{1}{d}$ such that, for all but finitely many $n \in \mathcal{N}_+$,

$$\delta(B_1(\|.\|_G), conv_n(G \cup -G)) \leq \frac{c}{n^\alpha}. \tag{4.6}$$

Set $s_G = \sup_{g\in G}\|g\|$ and $\eta = \frac{(1+s_G)c}{n^\alpha}$. As in the proof of Theorem 4.1 we shall derive a contradiction by comparing an upper and a lower bound on $\eta$-covering number of the unit ball in $G$-variation.

By (4.6), the triangle inequality and Lemma 3.2 (ii) and (iii), we get

$$cov_\eta B_1(\|.\|_G) \leq cov_{\eta/2} conv_n(G \cup -G) \leq (2cov_{\eta/(2(1+s_G))}G)^n \left(\frac{4(1+s_G)}{\eta}\right)^n.$$

Hence we obtain an upper bound

$$cov_{(1+s_G)c/n^\alpha}(B_1(\|.\|_G)) \leq \left(\frac{8}{c}\right)^n n^{\alpha n}(cov_{c/2n^\alpha}G)^n. \tag{4.7}$$

Let $A$ be an orthonogonal set not quickly vanishing with respect to $d$ such that $A_r = (1/r)\,A \subseteq B_1(\|.\|_G)$. By Lemma 3.7, for every positive integer $k \geq 3$ and $\varepsilon_k = \frac{1}{2\,r\,k^{\frac{d}{2}+1}}$,

we have $cov_{\varepsilon_k}(B_1(\|.\|_A)) \geq 2^{b\,k^d-1}$. Hence $cov_{\varepsilon_k}(B_1(\|.\|_G)) \geq cov_{\varepsilon_k}(B_1(\|.\|_{A_r})) \geq 2^{b\,k^d-1}$. If $k \leq \bar{k} = \left(\frac{1}{2\eta\, r}\right)^{\frac{2}{d+2}}$, then $\eta = (1+s_G)c/n^\alpha \leq \varepsilon_r$ and so $cov_\eta(B_1(\|.\|_G)) \geq cov_{\varepsilon_k}(B_1(\|.\|_G))$. As for every real number $x \geq 2$, $\lfloor x \rfloor \geq x-1 \geq x/2$, we have $2^{b\,\lfloor \bar{k}^d \rfloor - 1} \geq 2^{b\,\bar{k}^d/2-1}$. Thus, we obtain the following lower bound

$$2^{c_d\, n^{\alpha\frac{2d}{d+2}}-1} \leq cov_{(1+s_G)c/n^\alpha}(B_1(\|.\|_G)) \tag{4.8}$$

where $c_d = \frac{b}{2}\left(\frac{1}{2(1+s_G)\,r\,c}\right)^{\frac{2d}{d+2}}$.

The inequalities (4.7) and (4.8) give for all but finitely many $n \in \mathcal{N}_+$

$$2^{c_d\, n^{\alpha\frac{2d}{d+2}}-1} \leq \left(\frac{8}{c}\right)^n n^{\alpha n}(cov_{c/2n^\alpha}G)^n. \tag{4.9}$$

Taking the logarithm of both sides of (4.9), we get

$$c_d\, n^{\alpha\frac{2d}{d+2}} - 1 \leq 3n - n\,\log_2\, c + \alpha\, n\,\log_2\, n + n\,\log_2\,(cov_{c/2\,n^\alpha}G) \tag{4.10}$$

If $\alpha > 1/2 + 1/d$, then the left-hand side of (4.10) has order of infinity $\frac{2d\,\alpha}{d+2} > 1$. Its first two terms have order 1 and its third term has order smaller than every real number larger than 1. Hence, if (4.10) holds, the third term of its right-hand side must satisfy $\mathcal{O}(n^{\alpha\frac{2d}{d+2}}) \leq n\,\log_2\, cov_{(c/2n^\alpha)}\,G$. Setting $\varepsilon = c/2n^\alpha$, for all $\alpha > 1/2 + 1/d$ we get $\mathcal{O}\big(2^{(c/2\varepsilon)^\beta}\big) \leq cov_\varepsilon\, G$, where $\beta = \frac{2d}{d+2} - \frac{1}{\alpha} > 0$. This contradicts the assumption (1). □

Inspection of the proofs of Theorems 4.1 and 4.2 shows that the critical value of $\alpha$ is $1/2$, $1/2 + 1/d$, resp. With $d$ increasing, the second critical value, $1/2 + 1/d$, approaches the first one, $1/2$, which, as expected, corresponds to Maurey-Jones-Barron's bound.

# 5 Application to distributed computing

Maurey-Jones-Barron's theorem has been used by many authors to estimate complexity of nonlinear approximators used in distributed computing such as feedforward neural networks. The simplest type of a feedforward network is a *one-hidden-layer network*, whith a single linear output unit, which computes functions of the form $\sum_{i=1}^n w_i\phi(a_i,.)$ , where $n$ is the number of hidden units, $\phi : \mathcal{R}^p \times \mathcal{R}^d \to \mathcal{R}$ is the hidden unit function, and $p$, $d$ denote the dimension of the parameter and the input space, resp..

Denote by $G_\phi = \left\{\phi(a,\cdot) :\ a \in \mathcal{R}^p\right\}$ the parametrized set of functions computable by the unit $\phi$. A single linear output network with $n$ hidden units $\phi$ and $d$ inputs computes functions from $span_n\, G_\phi$ and so approximation by neural networks belongs to the variable-basis approximation scheme. By $\|.\|_{G_\phi}$ is denoted variation with respect to the set $G_\phi$.

Widespread computational units are perceptrons. A *perceptron* with an activation function $\psi : \mathcal{R} \to \mathcal{R}$ computes functions of the form $\phi((v,b),x) = \psi(v\cdot x + b) : \mathcal{R}^{d+1} \times \mathcal{R}^d \to \mathcal{R}$, where $v \in \mathcal{R}^d$ is an input weight vector and $b \in \mathcal{R}$ is a bias. By

$$P_d(\psi) = \left\{f : [0,1]^d \to \mathcal{R}; f(x) = \psi(v\cdot x + b), v \in \mathcal{R}^d, b \in \mathcal{R}\right\}$$

we denote the set of functions on $[0,1]^d$ computable by $\psi$-perceptrons. The most common activation functions are *sigmoidals*, i.e., bounded measurable functions $\sigma : \mathcal{R} \to \mathcal{R}$ 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 *Heaviside function*.

Let $\|.\|_{P_d(\sigma)}$ denote variation with respect to sigmoidal perceptrons with $d$ inputs. As the set $P_d(\vartheta)$ of functions computable by Heaviside perceptrons is equal to the *set of characteristic functions of half-spaces* of $\mathcal{R}^d$ restricted to $[0,1]^d$ (note that $\vartheta(v\cdot.+b)$ restricted to $[0,1]^d$ is

equal to the characteristic function of $\{x \in [0,1]^d : \ v \cdot x + b \geq 0\}$), we shall write $H_d$ instead of $P_d(\vartheta)$. $H_d$-variation will be called *variation with respect to half-spaces*, denoted by $\|.\|_{H_d}$.

Corollary 2.3 implies that all functions with $G_\phi$-variation at most $c$ can be approximated within $s_{G_\phi}\, c\, n^{-1/2}$ by neural networks with $n$ hidden units $\phi$. When $G_\phi$ satisfies the assumptions of Theorems 4.1 or 4.2, this bound is tight. The following proposition shows that the second condition of Theorem 4.2 is satisfied by sigmoidal perceptrons.

**Proposition 5.1** *Let $d, n$ be positive integers and $\sigma : \mathcal{R} \to \mathcal{R}$ be a sigmoidal function. Then in $(\mathcal{L}_2([0,1]^d), \|.\|_2)$, the ball $B_1(\|.\|_{P_d(\sigma)})$ contains an orthogonal subset not quickly vanishing with respect to $d$.*

**Proof.** It follows from [14, Prop. 3.3] that for every sigmoidal function $\sigma$, $B_1(\|.\|_{P_d(\sigma)}) \supseteq B_1(\|.\|_{H_d})$. Thus it is sufficient to show that $B_1(\|.\|_{H_d})$ contains an orthogonal subset not quickly vanishing with respect to $d$. The following construction of such a subset was used in [2], [21], and [17]; we report it here for reader's convenience. For $v = (v_1, \ldots, v_d) \in \mathcal{R}_+^d$, set $h_v(x) = c_v \sin(\pi v \cdot x) : [0,1]^d \to \mathcal{R}$, where $c_v = d\sqrt{2}/(\lceil \sum_{j=1}^d v_k \rceil)$. Let $A_d = \cup_{k \in \mathcal{N}_+} A_{d,k}$, where $A_{d,k} = \{h_v; v \in \{1, \ldots, k\}^d\} \subset (\mathcal{L}_2([0,1]^d), \|.\|_2)$. For any positive integer $d$, $A_d$ is contained in the ball of radius $2\, d\sqrt{2}$ in $H_d$-variation, i.e., $A_d \subseteq B_{2d\sqrt{2}}(\|.\|_{H_d})$, and $A_d$ is orthogonal not quickly vanishing with respect to $d$. □

It was shown in [21, Lemma 2] that for $\sigma$ either Heaviside or Lipschitz sigmoidal that is "similar" to Heaviside, $cov_\varepsilon(P_d(\sigma) = \mathcal{O}(p(1/\varepsilon))$, where $p$ is a polynomial. Theorem 4.2 combined with this estimate of covering numbers gives as a corollary the impossibility of improving Maurey-Jones-Barron's upper bound from Corollary 2.3 for perceptron networks, which was earlier proved in [1] and [21]. In [1], this tightness result was derived using a probabilistic argument, while the proof in [21] is based on covering numbers combined with an analogous result as our Theorem 4.2 stated only for sigmoidal perceptrons satisfying the conditions above.

# 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] Bellman, R.: *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[4] Bertsekas, D. P. and Tsitsiklis, J. N.: *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

[5] Breiman, L.: Hinging hyperplanes for regression, classification, and function approximation, *IEEE Trans. on Information Theory*, vol. 39, no. 3, pp. 993–1013, 1993.

[6] 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.

[7] DeVore, R. A., and Temlyakov, V. N.: Nonlinear approximation by trigonometric sums, *The J. of Fourier Analysis and Applications*, vol. 2, no. 1, pp. 29–48, 1995.

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

[9] Girosi, F.: Regularization theory, radial basis functions and networks, in *From Statistics to Neural Networks. Theory and Pattern Recognition Applications*, J.H. Friedman , V. Cherkassky, H. Wechsler Eds, Subseries F, Computer and System Sciences, pp. 166-187. Springer-Verlag, Berlin, 1993.

[10] 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.

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

[12] 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.

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

[14] V. Kůrková, P. C. Kainen, and V. Kreinovich: Estimates of the number of hidden units and variation with respect to half-spaces, *Neural Networks*, vol. 10, no. 6, pp. 1061–1068, 1997.

[15] 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.

[16] Kůrková, V., and Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation, *IEEE Transactions on Information Theory*, vol. 47, pp. 2659-2665, 2001.

[17] Kůrková, V., and Sanguineti, M.: Comparison of Worst Case Errors in Linear and Neural Network Approximation, *IEEE Transactions on Information Theory*, vol. 48, January 2002.

[18] 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.

[19] Lorentz, G. G.: Metric entropy and approximation. *Bulletin of the American Mathematical Society* 72, pp. 903-937, 1966.

[20] Lorentz, G. G., v. Golitschek, M., and Makovoz, Y.: *Constructive Approximation. Advanced Problems*. Grundlehren der Mathematischen Wissenschaften, vol. 304. Springer-Verlag Berlin Heidelberg, 1996.

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

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

[23] 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.

[24] Parisini, T., Sanguineti, M., and Zoppoli, R.: Nonlinear stabilization by receding-horizon neural regulators, *International J. of Control*, vol. 70, no. 3, pp. 341–362, 1998.

[25] Pinkus, A.: *$n-$Widths in Approximation Theory*. Springer-Verlag, Berlin, 1985.

[26] 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.

[27] Sejnowski, T. J., and Rosenberg, C. R.: Parallel networks that learn to pronounce English text, *Complex Systems*, vol. 1, no. 1, pp. 145–168, 1987.

[28] Zoppoli, R., Sanguineti, M., and Parisini, T.: Approximating Networks and Extended Ritz Method for the Solution of Functional Optimization Problems, Vol. 112, no. 2, February 2002. *Journal of Optimization Theory and Applications*.