Geometric Upper Bounds on Rates of Variable-Basis Approximation

Věra Kůrková and Marcello Sanguineti

Abstract—**In this paper, approximation by linear combinations of an increasing number n of computational units with adjustable parameters (such as perceptrons and radial basis functions) is investigated. Geometric upper bounds on rates of convergence of approximation errors are derived. The bounds depend on certain parameters specific for each function to be approximated. The results are illustrated by examples of values of such parameters in the case of approximation by linear combinations of orthonormal functions.**

Index Terms—**Approximation from a dictionary, model complexity, neural networks, rates of approximation, variable-basis approximation.**

I. Introduction

MANY computational models currently used in soft computing can be formally described as devices producing input–output functions in the form of linear combinations of simple computational units corresponding to the model (e.g., free-node splines, wavelets, trigonometric polynomials with free frequencies, sigmoidal perceptrons, and radial basis functions). Coefficients of linear combinations as well as inner parameters of computational units are adjustable by various learning algorithms (see, e.g., [1]).

Such models have been successfully used in many pattern recognition, optimization, and classification applications, some of them high-dimensional (see, e.g., [2]–[8] and the references therein). It seems that these computational models are more suitable for high-dimensional tasks than traditional linear ones (such as algebraic and trigonometric polynomials). It is known that complexity (measured by the number of computational units) of linear models needed to guarantee accuracy ε in approximation of multivariable functions defined by classical smoothness conditions may grow with the number of variables d as $\mathcal{O}(\varepsilon^{-d})$ [9, pp. 232–233].

Linear models can be called *fixed-basis models* as merely coefficients of linear combinations of the first n elements chosen from an ordered basis (or a set of functions called "dictionary" [10]) are adjustable. In contrast, the computational models mentioned above are sometimes called *variable-basis models* [11] because in addition to coefficients of linear combinations, one can also choose a suitable n-tuple of elements from the dictionary. Often such dictionaries are formed by parameterized families of functions computable by computational units of various types. In such cases, the choice of an n-tuple of suitable functions corresponds to finding the optimal parameters of such units.

Some insights into why, in solving high-dimensional tasks, model complexity requirements of variable-basis models may be considerably smaller than the ones of linear models, can be obtained from an estimate of rates of variable-basis approximation by Maurey [12], Jones [13], and Barron [14]. They derived an upper bound on the square of the error in approximation of a function f from the closure of the convex hull of a set G (dictionary) by a convex combination of n elements of G. The bound has the form $\frac{1}{n}(s_G^2 - \|f\|^2)$, where s_G is the supremum of the norms of elements of G and $\|f\|$ is the norm in the ambient Hilbert space of the function f to be approximated. This bound can be extended to all functions in the Hilbert space, using the concept of a norm tailored to the set G [15]–[17].

Maurey–Jones–Barron's theorem received a lot of attention because it implies an estimate of model complexity of the order $\mathcal{O}(\varepsilon^{-2})$. Several authors derived tight improvements of this bound for various sets G (e.g., G orthogonal [18], [19], G formed by functions computable by sigmoidal perceptrons, and G with certain properties of covering numbers [20], [21]). However, all these tightness results are worst-case estimates (i.e., they give upper bounds holding for all functions from the closure of the symmetric convex hull of G). Thus, one can expect that for suitable subsets of such hull better rates may hold.

A step towards a description of subsets with better rates was made by Lavretsky [22]. He noticed that when in the iterative construction derived by Jones [13] and improved by Barron [14], in each step, only functions satisfying a certain angular condition are chosen, then the term $\frac{1}{n}$ can be replaced with $(1-\delta^2)^{n-1}$, where $\delta \in (0,1]$ is the cosine corresponding to the angular constraint. However, Lavretsky left open the problem of characterization of functions satisfying such an angular condition. He illustrated his result by one example, in which, as he

Manuscript received February 21, 2008; revised July 30, 2008. Current version published November 21, 2008. The work of V. Kůrková was supported in part by Project 1ET100300517 of the program "Information Society" of the National Research Program of the Czech Republic and the Institutional Research Plan AV0Z10300504. The work of M. Sanguineti was supported in part by a PRIN grant from the Italian Ministry for University and Research, project "Models and Algorithms for Robust Network Optimization." Collaboration between V. Kůrková and M. Sanguineti was supported in part by the 2007–2009 Scientific Agreement among University of Genova, National Research Council of Italy, and Academy of Sciences of the Czech Republic, project "Learning from Data by Neural Networks and Kernel Methods as an Approximate Optimization."

V. Kůrková is with the Institute of Computer Science, Academy of Sciences of the Czech Republic, 182 07 Prague 8, Czech Republic (e-mail: vera@cs.cas.cz).

M. Sanguineti is with the Department of Communications, Computer, and System Sciences (DIST), University of Genova, 16145 Genova, Italy (e-mail: marcello@dist.unige.it).

Communicated by A. Krzyzak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2008.2006383

himself remarked [22, p. 280], "the geometric convergence ... is evident and easy to establish without the use of" his result.

In this paper, we show that for every function f in the convex hull of a bounded subset G of a Hilbert space there exists $\delta(f) \in (0,1]$ such that the rate of approximation of f by convex combinations of n functions from G is bounded from above by $(1-\delta(f)^2)^{n-1}(s_G^2 - \|f\|^2)$. However, we do not claim that every function in the convex hull satisfies the angular condition implying Lavretsky's estimate. Instead, we derive the geometric rate by modifying the incremental construction originally used by Jones [13], later improved by Barron [14], and refined by Lavretsky [22]. We also show that a similar estimate holds for all functions from the linear span of G in approximation by linear combinations of n elements of G. We illustrate our results by estimating values of parameters of geometric rates when G is an orthonormal basis. We exploit these estimates to derive some insights into the structure of sets of functions with fixed values of parameters of such rates. A preliminary version of some results appeared in the conference proceedings [23].

This paper is organized as follows. In Section II, Maurey–Jones–Barron's theorem and its improvements are stated, together with definitions of concepts used in the paper. In Section III, our main theorem on geometric rates of approximation for functions in convex hulls and its corollary for linear spans are proven. In Section IV, examples of functions in convex hulls of orthonormal sets with estimates of the values of parameters determining geometric rates are given. In Section V, properties of sets with fixed values of such parameters are investigated. Section VI is a brief discussion.

II. Maurey–Jones–Barron's Theorem and Its Improvement

Many computational models used in soft computing can be mathematically described as *variable-basis* schemes. Such models compute functions from sets of the form

$$\mathrm{span}_n G = \left\{ \sum_{i=1}^{n} w_i g_i | w_i \in \mathbb{R},\ g_i \in G \right\}$$

where G is a set of functions, which is sometimes called a *dictionary*, and $\mathbb{R}$ is the set of real numbers. Typically, G is a parameterized set of functions that can be computed by computational units of a given type, such as perceptrons and radial basis functions units. Note that for G linearly independent and such that $\mathrm{card}\, G > n$, $\mathrm{span}_n\, G$ is not convex.

A useful tool for investigation of rates of decrease of errors in approximation by $\mathrm{span}_n\, G$ with n increasing is Maurey–Jones–Barron's theorem [12]–[14]. This theorem is formulated for approximation of functions from the closure of

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

where G is a given bounded subset of a Hilbert space, by elements of

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

The following upper bound is a version of Jones' result [13] as improved by Barron [14] (see also an earlier estimate by Maurey in [12]). For a subset M of a normed linear space $(X, \|\cdot\|)$ and $f \in X$, we denote by

$$\|f - M\| = \inf_{g \in M} \|f - g\|$$

the *distance* of f from M and by cl the *closure* with respect to the topology induced by $\|\cdot\|$.

Theorem 1 (Maurey–Jones–Barron): Let $(X, \|\cdot\|)$ be a Hilbert space, G its bounded nonempty subset, $s_G = \sup_{g \in G} \|g\|$, and $f \in \mathrm{clconv} G$. For every positive integer n

$$\|f - \mathrm{conv}_n G\|^2 \le \frac{s_G^2 - \|f\|^2}{n}.$$

In [16] (see also [17]), Theorem 1 was extended using the concept of *G-variation*, defined for all functions $f \in X$ as

$$\|f\|_G = \min\{c > 0 | f/c \in \mathrm{clconv}(G \cup -G)\}$$

where

$$-G = \{-g | g \in G\}.$$

Note that $\|\cdot\|_G$ is the Minkowski functional[1] of the set $\mathrm{cl}\,\mathrm{conv}(G \cup -G)$ and so it is a norm on the subspace of X containing those $f \in X$ for which $\|f\|_G < \infty$. It is easy to check that Theorem 1 implies that for a Hilbert space $(X, \|\cdot\|)$, its bounded subset G with $s_G = \sup_{g \in G} \|g\|$, and $f \in X$

$$\|f - \mathrm{span}_n G\|^2 \le \frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}. \tag{1}$$

Lavretsky [22] noticed that the argument used by Jones [13] and Barron [14] can yield better rates when applied to functions satisfying a certain angular relationship with respect to G. For $\delta > 0$, he defined the set

$$\begin{aligned} F_\delta(G) = \Big\{ f \in \mathrm{cl}\,\mathrm{conv}\, G | (\forall h \in \mathrm{conv}\, G),\\ \big(f \ne h \Rightarrow \exists g \in G : (f-g)\cdot(f-h)\\ \le -\delta \|f-g\|\,\|f-h\|\big)\Big\}. \end{aligned} \tag{2}$$

Note that for all $\delta > 0, G \subseteq F_\delta$. Indeed for every $f \in G$, setting $g = f$, we get

$$(f-f)\cdot(f-h) \le -\delta\|f-h\|\,\|f-f\|.$$

Lavretsky [22] realized that the incremental construction developed by Jones and Barron uses in each step a certain property of functions from the convex hulls, which restated in Lavretsky's terminology says that $\mathrm{cl}\,\mathrm{conv}\, G = F_0(G)$. Strengthening the condition on the function f to be approximated by assuming that $f \in F_\delta(G)$ for some $\delta > 0$, he derived the following geometric upper bound on rates of approximation by sets $\mathrm{conv}_n G$ [22, Th. 1].

[1]The *Minkowski functional* of a set $M \subseteq X$ is the functional $p_M : X \to [0, +\infty]$ defined for every $f \in X$ as $p_M(f) = \inf\{c > 0 | f/c \in M\}$[24, p. 131]

Theorem 2 (Lavretsky): Let $(X, \|\cdot\|)$ be a Hilbert space, G its bounded symmetric subset containing 0, $s_G = \sup_{g\in G}\|g\|$, and $\delta > 0$. Then, for every $f \in F_\delta(G)$ and every positive integer n

$$\|f - \mathrm{conv}_n\, G\|^2 \leq (1-\delta^2)^{n-1}\left(s_G^2 - \|f\|^2\right).$$

Unfortunately, the definition (2) does not enable an easy verification whether a function is in $F_\delta(G)$. Also, it is not clear which functions are contained in sets $F_\delta(G)$. Only for finite-dimensional Hilbert spaces and G satisfying certain conditions, Lavretsky [22] described subsets of $\mathrm{cl\,conv}\, G$ with the property that for each element f, there exists $\delta_f \in (0,1]$ such that $f \in F_{\delta_f}(G)$. He defined the *affine interior* of a convex subset M of a normed linear space $(X, \|\cdot\|)$ with $0 \in M$ as

$$I_{\mathrm{aff}}(M) = \{f \in M | (\exists \varepsilon_f > 0)(\forall h \in \mathrm{span}\, M)(\|h\| < \varepsilon_f \Rightarrow f + h \in M)\}$$

and proved the following theorem about the relationships between $I_{\mathrm{aff}}(\mathrm{cl\,conv}\, G)$ and $F_\delta(G)$.

Theorem 3 (Lavretsky): Let $(X, \|\cdot\|)$ be a finite-dimensional Hilbert space, G its bounded symmetric subset such that $0 \in G$ and $\mathrm{card}\, G \geq \dim X$, and $s_G = \sup_{g\in G}\|g\|$. Then, $I_{\mathrm{aff}}(\mathrm{cl\,conv} G) \neq \emptyset$ and for every $f \in I_{\mathrm{aff}}(\mathrm{cl\,conv}\, G)$, there exists $\delta_f \in (0,1]$ such that $f \in F_{\delta_f}(G)$ and

$$\|f - \mathrm{conv}_n\, G\|^2 \leq (1-\delta_f^2)^{n-1}\left(s_G^2 - \|f\|^2\right).$$

However, in [22], no examples illustrating possible values of parameters δ_f were given.

III. Geometric Rates of Variable-Basis Approximation

In this section, we show that the estimate from Theorem 3 holds for all functions in the convex hull of any bounded subset of any Hilbert space.

Theorem 4: Let $(X, \|\cdot\|)$ be a Hilbert space, G its bounded nonempty subset, and $s_G = \sup_{g\in G}\|g\|$. For every $f \in \mathrm{conv}\, G$, there exists $\delta_f \in (0,1]$ such that for every positive integer n

$$\|f - \mathrm{conv}_n\, G\|^2 \leq \left(1-\delta_f^2\right)^{n-1}\left(s_G^2 - \|f\|^2\right).$$

Proof: Let $f = \sum_{j=1}^m a_j\, g_j$ be a representation of f as a convex combination of elements of G with all $a_j > 0$ and let

$$G' = \{g_1, \ldots, g_m\}.$$

We will construct a sequence of functions $\{f_n | n = 1, \ldots, m\}$ and a sequence of positive real numbers $\{\delta_n | n = 1, \ldots, m\}$ such that for each $n = 1, \ldots, m, f_n \in \mathrm{conv}_n G$ and

$$\|f - f_n\|^2 \leq (1-\delta_n)^{n-1}\left(s_G^2 - \|f\|^2\right).$$

We start with choosing some $g_{j_1} \in G'$ satisfying

$$\|f - g_{j_1}\| = \min_{g\in G'}\|f - g\|$$

and we set $f_1 = g_{j_1}$. As

$$\sum_{j=1}^m a_j\|f - g_j\|^2 = \|f\|^2 - 2f\cdot\sum_{i=1}^m a_j g_j + \sum_{j=1}^m a_j\|g_j\|^2 \leq s_G^2 - \|f\|^2$$

we get $\|f - f_1\|^2 \leq s_G^2 - \|f\|^2$ and so the statement holds for $n = 1$ with any $\delta \in (0,1)$.

Assuming that we have f_{n-1}, we define f_n. When $f_{n-1} = f$, we set $f_n = f_{n-1}$ and the estimate holds trivially.

When $f_{n-1} \neq f$, we define f_n as the convex combination

$$f_n = \alpha_n f_{n-1} + (1-\alpha_n) g_{j_n} \tag{3}$$

with $g_{j_n} \in G'$ and $\alpha_n \in [0,1]$ chosen in such a way that for some $\delta_n > 0$

$$\|f - f_n\|^2 \leq \left(1-\delta_n^2\right)^{n-1}\|f - f_{n-1}\|^2.$$

First, we choose a suitable g_{j_n} and then we find α_n depending on our choice of g_{j_n}. Denoting $e_n = \|f - f_n\|$, by (3) we get

$$e_n^2 = \alpha_n^2 e_{n-1}^2 + 2\alpha_n(1-\alpha_n)(f - f_{n-1})\cdot(f - g_{j_n}) + (1-\alpha_n)^2\|f - g_{j_n}\|^2. \tag{4}$$

For all $j \in \{1, \ldots, m\}$, set

$$\eta_j = -\frac{(f - f_{n-1})\cdot(f - g_j)}{\|f - f_{n-1}\|\,\|f - g_j\|}$$

(both terms in the denominator are nonzero: the first one because we consider the case when $f \neq f_{n-1}$ and the second one because we assume that for all $j, a_j > 0$ and thus $f \neq g_j$). Note that for all $j, \eta_j \in [-1,1]$ as it is the cosine of the angle between the vectors $f - f_{n-1}$ and $f - g_j$.

As $f = \sum_{j=1}^m a_j\, g_j$, we have

$$\sum_{j=1}^m a_j(f - f_{n-1})\cdot(f - g_j) = (f - f_{n-1})\cdot\left(f - \sum_{j=1}^m a_j g_j\right) = 0.$$

Thus

i) either there exists $g \in G'$, for which $(f - f_{n-1})\cdot(f - g) < 0$;
ii) or for all $g \in G'$, $(f - f_{n-1})\cdot(f - g) = 0$.

We show that the case ii) implies that $f = f_{n-1}$. Indeed, $f_{n-1} \in \mathrm{conv}_{n-1} G'$, thus it can be expressed as

$$f_{n-1} = \sum_{k=1}^{n-1} b_k g_{j_k}$$

with all $b_k \in [0,1]$ and $\sum_{k=1}^{n-1} b_k = 1$. If for all $g \in G', (f - f_{n-1})\cdot(f - g) = 0$, then

$$\|f - f_{n-1}\|^2 = (f - f_{n-1})\cdot(f - \sum_{k=1}^{n-1} b_k g_{j_k}) = \sum_{k=1}^{n-1} b_k(f - f_{n-1})\cdot(f - g_{j_k}) = 0.$$

So in the case now considered, i.e., $f \neq f_{n-1}$, i) holds and thus the subset

$$G'' = \{g \in G' | (f - f_{n-1}) \cdot (f - g) < 0\}$$

is nonempty. Let $g_{j_n} \in G''$ be chosen so that

$$\eta_{j_n} = \max_{j=1,\ldots,m} \eta_j$$

and set $\delta_n = \eta_{j_n}$. As $G'' \neq \emptyset$, we have $\delta_n > 0$.

Set $r_n = \|f - g_{j_n}\|$. By (4), we get

$$e_n^2 = \alpha_n^2 e_{n-1}^2 - 2\alpha_n(1-\alpha_n)\delta_n e_{n-1} r_n + (1-\alpha_n)^2 r_n^2. \quad (5)$$

To define f_n as a convex combination of f_{n-1} and g_{j_n}, it remains to find $\alpha_n \in [0,1]$ for which e_n^2 is minimal as a function of α_n. By (5), we have

$$e_n^2 = \alpha_n^2 \left(e_{n-1}^2 + 2\delta_n e_{n-1} r_n + r_n^2\right) - 2\alpha_n \left(\delta_n e_{n-1} r_n + r_n^2\right) + r_n^2. \quad (6)$$

Thus

$$\frac{\partial e_n^2}{\partial \alpha_n} = 2\alpha_n \left(e_{n-1}^2 + 2\delta_n e_{n-1} r_n + r_n^2\right) - 2\left(\delta_n e_{n-1} r_n + r_n^2\right)$$

and

$$\frac{\partial^2 e_n^2}{\partial^2 \alpha_n} = 2\left(e_{n-1}^2 + 2\delta_n e_{n-1} r_n + r_n^2\right).$$

As now we are considering the case when $f \neq f_{n-1}$, we have $e_{n-1} > 0$ and hence $\frac{\partial e_n^2}{\partial^2 \alpha_n} > 0$. So the minimum is achieved at

$$\alpha_n = \frac{\delta_n e_{n-1} r_n + r_n^2}{e_{n-1}^2 + 2\delta_n e_{n-1} r_n + r_n^2}. \quad (7)$$

Plugging (7) into (6), we get

$$e_n^2 = \frac{\left(1-\delta_n^2\right) e_{n-1}^2 r_n^2}{e_{n-1}^2 + 2\delta_n e_{n-1} r_n + r_n^2} < \frac{\left(1-\delta_n^2\right) e_{n-1}^2 r_n^2}{r_n^2} = (1-\delta_n^2) e_{n-1}^2.$$

Let

$$k = \max\{n \in \{1,\ldots,m\} | f_n \neq f_{n-1}\}.$$

Setting

$$\delta_f = \min\{\delta_n | n = 1,\ldots,k\}$$

by induction, we get the upper bound

$$\|f - \mathrm{conv}_n\, G\|^2 \leq \left(1 - \delta_f^2\right)^{n-1} \left(s_G^2 - \|f\|^2\right)$$

holding for all n (for $n > m$, it holds trivially with $f_n = f$). $\square$

Note that Theorem 4 guarantees the upper bound $(1-\delta_f^2)^{n-1}(s_G^2 - \|f\|^2)$ for all functions $f \in \mathrm{conv} G$, including those that are not elements of $F_{\delta_f}(G)$. We achieve this result using a proof technique that needs a suitable angular property of only one element of $\mathrm{conv}\, G$ (the one constructed as the approximant in the previous step). This is a much weaker angular condition than the one required in the definition (2) of $F_{\delta_f}(G)$ and it is weak enough to hold for all functions of $\mathrm{conv}\, G$.

The proof of Theorem 4 is constructive. Schematically, it can be described as an incremental procedure, which constructs, for a function $f = \sum_{j=1}^m a_j g_j \in \mathrm{conv}\, G$, a sequence of approximants $f_n \in \mathrm{conv}_n G$. The next table describes this procedure.

- CHOOSE $g_{j_1} \in \{g_j | j = 1,\ldots,m\}$
  SUCH THAT $\|f - g_{j_1}\| = \min_{j=1,\ldots,m} \|f - g_j\|$;
- $f_1 := g_{j_1}$;
- FOR $n = 2,\ldots,m-1$:

  BEGIN

  FOR $j = 1,\ldots,m$,

  COMPUTE $\eta_j := -\frac{(f-f_{n-1})\cdot(f-g_j)}{\|f-f_{n-1}\|\,\|f-g_j\|}$

  IF FOR $j = 1,\ldots,m$ ONE HAS $\eta_j = 0$, THEN

  BEGIN

  $f^* := f_{n-1}$;

  $n^* := n-1$;

  END

  ELSE

  BEGIN

  $\delta_n := \max\{\eta_j > 0 | j = 1,\ldots,m\}$;

  CHOOSE g_{j_n} SUCH THAT $\delta_n = \eta_{j_n}$;

  COMPUTE $e_{n-1} := \|f - f_{n-1}\|$;

  COMPUTE $r_n := \|f - g_{j_n}\|$;

  COMPUTE $\alpha_n := \frac{\delta_n e_{n-1} r_n + r_n^2}{e_{n-1}^2 + 2\delta_n e_{n-1} r_n + r_n^2}$;

  $f_n := \alpha_n f_{n-1} + (1-\alpha_n)\, g_n$;

  $n := n+1$.

  END

  END

In the next section, we will prove that Theorem 4 cannot be extended to all functions in the closure of the convex hull of G. However, the geometric bound can be extended to all functions from $\mathrm{span}\, G$, as the next corollary shows.

Corollary 1: Let $(X, \|\cdot\|)$ be a Hilbert space, G its bounded nonempty subset, $s_G = \sup_{g\in G} \|g\|$, and $f \in \mathrm{span}\, G$. Then, there exists $b > 0$ and $\delta_{f,b} \in (0,1]$ such that for every positive integer n

$$\|f - \operatorname{span}_n G\|^2 \le \left(1 - \delta_{f,b}^2\right)^{n-1} \left((s_G\, b)^2 - \|f\|^2\right).$$

Proof: For every $f \in \operatorname{span} G$, there exists $b > 0$ such that $f \in \operatorname{conv}(b\,G \cup -b\,G)$, where $b\,G = \{b\,g \,|\, g \in G\}$. By Theorem 4, there exists $\delta_{f,b} \in (0,1]$ such that for every positive integer n

$$\|f - \operatorname{conv}_n(b\,G \cup -b\,G)\|^2 \le \left(1 - \delta_{f,b}^2\right)^{n-1} \left((s_G\, b)^2 - \|f\|^2\right).$$

As

$$\|f - \operatorname{span}_n G\| \le \|f - \operatorname{conv}_n(b\,(G \cup -G))\|$$

the statement follows. □

Note that Corollary 1 cannot be extended to

$$b_f = \inf\{b > 0 | f \in \operatorname{conv}(b\,(G \cup -G))\}$$

as

$$\inf\{\delta_{f,b} | b > 0, f \in \operatorname{conv}(b\,(G \cup -G))\}$$

might be equal to zero.

IV. Parameters of Geometric Rates

Theorem 4 implies that for all functions f in $\operatorname{conv} G$, the upper bound $\frac{1}{n}(s_G^2 - \|f\|^2)$ from Maurey–Jones–Barron's theorem can be replaced with $(1 - \delta_f^2)^{n-1}\,(s_G^2 - \|f\|^2)$. The speed of decrease of this geometric estimate depends on δ_f, which has some value in $(0,1]$ corresponding to the smallest cosine of the angles between functions used in the construction of approximants. In this section, we illustrate this result with some estimates of values of the parameters δ_f.

Inspection of the proof of Theorem 4 shows that the parameter δ_f is not defined uniquely. It depends on the choice of a representation of $f = \sum_{j=1}^m a_j g_j$ as a convex combination of elements of G and on the choice of g_{j_n} for those positive integers n, for which there exist more than one g_j with the same cosine δ_n. However, the *maximal parameter*, for which the geometric upper bound from Theorem 4 holds, is unique. Define

$$\delta(f) = \max\bigl\{\delta > 0 |\, \|f - \operatorname{conv}_n G\| \le (1 - \delta^2)^{n-1} \left(s_G^2 - \|f\|^2\right) \bigr\}. \quad (8)$$

By Theorem 4, for each $f \in \operatorname{conv} G$, the set on the right-hand side of (8) is nonempty and bounded. It follows from the definition of this set that its supremum is achieved, i.e., it is its maximum. So

$$\|f - \operatorname{conv}_n G\| \le ((1 - \delta(f)^2)^{n-1} \left(s_G^2 - \|f\|^2\right).$$

For each $\delta \in (0,1]$, define

$$A_\delta(G) = \{f \in \operatorname{conv} G | \delta(f) = \delta\}.$$

It follows from the definition of $A_\delta(G)$ that $G \subseteq \bigcap_{\delta > 0} A_\delta(G)$ and if $\delta_1 \le \delta_2$, then $A_{\delta_1}(G) \supseteq A_{\delta_2}(G)$. Theorem 4 implies that $\operatorname{conv} G = \bigcup_{\delta \in (0,1]} A_\delta(G)$. Some insight into properties of the sets $A_\delta(G)$ can be obtained from the following estimate of maximal parameters for the case when G is orthonormal.

Proposition 1: Let $(X, \|\cdot\|)$ be an infinite-dimensional separable Hilbert space and G its orthonormal basis. Then, for every positive integer $n \ge 3$, there exists $h_{2n} \in \operatorname{conv} G$ such that $\|h_{2n}\| = \frac{1}{\sqrt{2n}}$ and

$$\delta(h_{2n})^2 \le 1 - 5^{-\frac{1}{n-1}}\, e^{-\frac{\ln(n-1)}{n-1}}.$$

Proof: Let $G = \{g_i\}$. For each positive integer n, define $h_n = 1/n \sum_{i=1}^n g_i$. Then, $h_n \in \operatorname{conv}_n G$ and $\|h_n\| = 1/\sqrt{n}$. It is easy to see that

$$\|h_{2n} - \operatorname{span}_n G\| = 1/(2\sqrt{n}).$$

Thus, by Theorem 4

$$\frac{1}{4n} = \|h_{2n} - \operatorname{span}_n G\|^2 \le \|h_{2n} - \operatorname{conv}_n G\|^2 \le (1 - \delta(h_{2n})^2)^{n-1} \left(1 - \frac{1}{2n}\right).$$

So for all $n \ge 3$

$$\delta(h_{2n})^2 \le 1 - \left(\frac{1}{2(2n-1)}\right)^{\frac{1}{n-1}} \le 1 - \left(\frac{1}{5(n-1)}\right)^{\frac{1}{n-1}} = 1 - 5^{-\frac{1}{n-1}}\, e^{-\frac{\ln(n-1)}{n-1}}.$$

□

Proposition 1 shows that for the sequence $\{h_{2n}\}$ of barycenters of an increasing number n of elements of an orthonormal basis, the sequence of corresponding parameters $\{\delta(h_{2n})\}$ converges to zero exponentially fast. Thus, for n large enough, the upper bound

$$(1 - \delta(h_{2n})^2)^{n-1}\,(1 - \|h_{2n}\|^2)$$

from Theorem 4 guarantees sufficiently small errors of approximation by $\operatorname{conv}_n G$.

The sequence of the barycenters $\{h_{2n}\}$ from the proof of Proposition 1 can also be used as follows to show that for G an orthonormal basis of an infinite-dimensional separable Hilbert space, Theorem 4 cannot be extended to all functions from the closure of the convex hull of G. Let h_0 be the element with all coefficients equal to zero. Then

$$\lim_{n\to\infty} \|h_n - h_0\| = \lim_{n\to\infty} \|h_n\| = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0$$

and so $h_0 \in \operatorname{cl}\operatorname{conv} G$. It is easy to see that for all $n, \|h_0 - \operatorname{conv}_n G\| = \|h_0 - h_n\| = \frac{1}{\sqrt{n}}$. If Theorem 4 held for all functions from $\operatorname{cl}\operatorname{conv} G$, then there would exist $\delta(h_0) > 0$ such that for all n

$$\|h_0 - \operatorname{conv}_n G\| \le (1 - \delta(h_0)^2)^{n-1}(1 - \|h_0\|^2) = (1 - \delta(h_0)^2)^{n-1}.$$

This would imply that for all n

$$\frac{1}{\sqrt{n}} \le (1 - \delta(h_0)^2)^{n-1}.$$

However, the right-hand side of this inequality converges to zero faster than the left-hand side, so such $2\delta(h_0) > 0$ cannot exist.

Proposition 1 implies that for G orthonormal, the sets $A_\delta(G)$ are neither convex nor open in the topology induced on X by the norm $\|\cdot\|$, as stated in the next corollary. For $r>0$, we denote by $B_r(\|\cdot\|)$ the ball of radius r centered at 0, i.e.,

$$B_r(\|\cdot\|)=\{f\in X \mid \|f\|\le r\}.$$

Corollary 2: Let $(X,\|\cdot\|)$ be an infinite-dimensional separable Hilbert space, G its orthornormal basis, and $\bar{G}$ a bounded subset of X containing G. Then, for every $\delta\in(0,1]$, $A_\delta(\bar{G})$ is not convex and for every $r>0$, $B_r(\|\cdot\|)\subseteq A_\delta(\bar{G})$.

Proof: If $A_\delta(\bar{G})$ were convex, then we would get $A_\delta(\bar{G})=\operatorname{conv}\bar{G}$, because $\bar{G}\subseteq A_\delta(\bar{G})$. However, this would contradict Proposition 1. Inspection of its proof shows that for each $r>0$ and $\delta\in(0,1]$, there exists $h_{2n}\in\operatorname{conv}G\subseteq\operatorname{conv}\bar{G}$ with $\|h_{2n}\|=\frac{1}{\sqrt{2n}}<r$ and $\delta(h_{2n})<\delta$, so $B_r(\|\cdot\|)\subseteq A_\delta(\bar{G})$. □

The next proposition shows that for G orthonormal, $A_\delta(G)$ cannot contain even a sphere of a radius $r\in(0,1]$.

Proposition 2: Let $(X,\|\cdot\|)$ be an infinite-dimensional separable Hilbert space and $G=\{g_i\}$ its orthornormal basis. For every $r\in(0,1]$ and every odd positive integer $k>\frac{1}{r^2}$, there exists $h_k\in\operatorname{conv}_k G$ such that $\|h_k\|=r, \lim_{k\to\infty}\|h_k-rg_1\|=0$, and

$$\delta(h_k)^2\le 1-8^{-\frac{1}{k-1}}e^{-\frac{\ln(k-1)}{k-1}}.$$

Proof: Set $c_k=\sqrt{\frac{kr^2-1}{k-1}}$ and define

$$h_k=\frac{1}{k}\left((1+(k-1)c_k)g_1+\sum_{i=2}^{k}(1-c_k)g_i\right).$$

As $k>\frac{1}{r^2}$, we have $c_k\in[0,1]$ and thus $h_k\in\operatorname{conv}_k G$. It is easy to check that $\|h_k\|=r$ and $\lim_{k\to\infty}\|h_k-rg_1\|=0$. Let $n_k=\frac{k+1}{2}$. It was shown in [18, p. 2664] that

$$\|h_k-\operatorname{span}_{n_k}G\|\ge\frac{1-r^2}{4\sqrt{n_k-1}}.$$

On the other hand, by Theorem 4

$$\|h_k-\operatorname{conv}_{n_k}G\|\le(1-\delta(h_k)^2)^{n_k-1}(1-r^2).$$

Thus, $\frac{1}{4\sqrt{n_k-1}}\le(1-\delta(h_k)^2)^{n_k-1}$ and so

$$\frac{1}{2\sqrt{2(k-1)}}\le(1-\delta(h_k)^2)^{\frac{k-1}{2}}.$$

Hence

$$\begin{aligned}\delta(h_k)^2&\le 1-\left(\frac{1}{2\sqrt{2}\sqrt{k-1}}\right)^{\frac{2}{k-1}}\\&=1-\left(\frac{1}{8}\right)^{\frac{1}{k-1}}\left(\frac{1}{k-1}\right)^{\frac{1}{k-1}}\\&=1-8^{-\frac{1}{k-1}}e^{-\frac{\ln(k-1)}{k-1}}.\end{aligned}$$ □

Proposition 2 shows that the limit of a sequence of parameters $\{\delta(h_k)\}$ needs not to be equal to the parameter of the limit. Indeed, let $\bar{G}=G\cup\{0\}$. Then, $\lim_{k\to\infty}h_k=r\,g_1\in A_1(\bar{G})$, so $\delta(rg_1)=1$, but the sequence $\{\delta(h_k)\}$ converges to zero (the convergence is even exponentially fast). Thus, the mapping $\delta:\operatorname{conv}G\to(0,1]$ is not continuous.

V. Properties of Sets of Functions With Geometric Rates

In Section IV, we proved that when G is an orthonormal basis of a separable Hilbert space, the sets $A_\delta(G)$ are neither convex, nor open, and they do not contain any sphere of any radius $r\in(0,1]$ and δ, as a mapping from $\operatorname{conv}G$ to $(0,1]$, is not continuous. In this section, we describe some properties of the sets $A_\delta(G)$ of functions with the maximal parameter $\delta(f)=\delta$. To derive such properties, we extend Lavretsky's result [22, Th. 2] (here stated as Theorem 3) on a relationship between affine interiors of certain subsets of finite-dimensional Hilbert spaces and sets $F_\delta(G)$.

For a subset M of a normed linear space $(X,\|\cdot\|)$ such that $X=\operatorname{span}M$ and $\varepsilon\ge0$, we define the *ε-interior* of M in $(X,\|\cdot\|)$ as

$$I_\varepsilon(M)=\{f\in M\,|(\forall h\in X)\,(\|h\|\le\varepsilon\Rightarrow f+h\in M)\}. \quad (9)$$

It is easy to see that $\cup_{\varepsilon>0}I_\varepsilon(M)=I_{\text{aff}}(M)$. The following proposition states some properties of $I_\varepsilon(M)$, which follow easily from the definition.

Proposition 3: Let $(X,\|\cdot\|)$ be a normed linear space, M its bounded subset such that $\operatorname{span}M=X$, and $\varepsilon>0$.

i) If M is symmetric, then $I_\varepsilon(M)$ is symmetric;
ii) if M is convex, then $I_\varepsilon(M)$ is convex;
iii) if M is convex and symmetric and $I_\varepsilon(M)\ne\emptyset$, then $B_\varepsilon(\|\cdot\|)\subseteq M$.

Proof:

i) Let $f\in I_\varepsilon(M)$. For every $h\in X$ with $\|h\|\le\varepsilon$, $f-h\in M$ and as M is symmetric, $-f+h\in M$.
ii) Let $f,g\in I_\varepsilon(M)$ and $h\in X$ with $\|h\|\le\varepsilon$. Then, $f+h\in M, g+h\in M$, and as M is convex
$$\frac{f+h}{2}+\frac{g+h}{2}=\frac{f+g}{2}+h\in M.$$
So $\frac{f+g}{2}\in I_\varepsilon(M)$.
iii) By i), there exists f such that both f and $-f$ are in $I_\varepsilon(M)$. By ii), $I_\varepsilon(M)$ is convex, and thus, $0\in I_\varepsilon(M)$. Hence, the statement follows by the definition of $I_\varepsilon(M)$. □

The next proposition shows that the *ε-interior* of $\operatorname{conv}G$ is contained in $A_\delta(G)$ with $\delta=\varepsilon/2$.

Proposition 4: Let $(X,\|\cdot\|)$ be a Hilbert space, G its bounded subset with $\operatorname{span}G=X$, and $\varepsilon,\delta>0$ such that $\varepsilon=2s_G\delta$. Then, $I_\varepsilon(\operatorname{conv}G)\subseteq F_\delta(G)\subseteq A_\delta(G)$.

Proof: To prove the first inclusion by contradiction, suppose that $F_\delta(G)\subseteq A_\delta(G)$. Assume that there exists $f\in I_\varepsilon(\operatorname{conv}G)$ such that $f\notin F_\delta(G)$. Then, there exists $h\in\operatorname{conv}G$ such that for all $g\in G$

$$(f-g)\cdot(f-h)>-\delta\|f-g\|\|f-h\|.$$

Hence, $h\ne f$ and

$$(g-f)\cdot\frac{f-h}{\|f-h\|}<\delta\|f-g\|\le 2s_G\delta=\varepsilon$$

for all $g \in G$. So for all $g = \sum_{i=1}^{m} a_i g_i \in \operatorname{conv} G$, we get

$$\left(\sum_{i=1}^{m} a_i g_i - f\right) \cdot \frac{f-h}{\|f-h\|} = \sum_{i=1}^{m} a_i (g_i - f) \cdot \frac{f-h}{\|f-h\|} < \varepsilon \sum_{i=1}^{m} a_i = \varepsilon.$$

Thus, for every $g \in \operatorname{conv} G$

$$(g-f) \cdot \frac{f-h}{\|f-h\|} < \varepsilon. \tag{10}$$

Because $f \in I_\varepsilon(\operatorname{conv} G)$, $f + \varepsilon \frac{f-h}{\|f-h\|} \in \operatorname{conv} G$. Setting

$$g = f + \varepsilon \frac{f-h}{\|f-h\|}$$

we get

$$(g-f) \cdot \frac{f-h}{\|f-h\|} = \left(f + \varepsilon \frac{f-h}{\|f-h\|} - f\right) \cdot \frac{f-h}{\|f-h\|} = \varepsilon$$

which contradicts (10).

The second inclusion follows by Theorem 2 and the definition of $A_\delta(G)$. □

Proposition 4 implies that for G a bounded subset spanning a finite-dimensional Hilbert space, the sets $A_\delta(G)$ are open.

Corollary 3: Let $(X, \|\cdot\|)$ be a finite-dimensional Hilbert space and G its bounded subset with $\operatorname{span} G = X$. Then, there exist $r > 0$ and $\delta \in (0,1]$ such that $B_r(\|\cdot\|) \subseteq A_\delta(G)$.

Proof: Any symmetric convex set is the unit ball of the norm defined by its Minkowski functional, so in particular $\operatorname{conv}(G \cup -G)$ is the unit ball of a norm on X. As in a finite-dimensional Hilbert space all norms are equivalent, there exists $c > 0$ such that $B_c(\|\cdot\|) \subseteq \operatorname{conv}(G \cup -G)$. Hence, by Proposition 4, for all $r > 0$ and all $\varepsilon > 0$ such that

$$r + 2 s_G \delta = r + \varepsilon \le c$$

we get

$$B_r(\|\cdot\|) \subseteq I_\varepsilon(\operatorname{conv}(G \cup -G)) \subseteq F_\delta(G) \subseteq A_\delta(G). \quad \square$$

By Corollary 3, for every bounded subset of a finite-dimensional Hilbert space, there exist $r > 0$ and $\delta > 0$ such that for all functions f in the ball $B_r(\|\cdot\|)$

$$\|f - \operatorname{conv}_n G\|^2 \le (1-\delta^2)^{n-1} \left(s_G^2 - \|f\|^2\right).$$

For orthonormal bases of finite-dimensional Hilbert spaces, Corollary 3 combined with the construction from the proof of Proposition 2 gives some insights into the structure of the sets $A_\delta(G)$.

Corollary 4: Let $(X, \|\cdot\|)$ be a finite-dimensional Hilbert space, $\dim X = m$, and $G = \{g_1, \ldots, g_m\}$ its orthonormal basis. Then

i) for every $f \in \operatorname{conv} G$ such that $\|f\| < \frac{1}{\sqrt{m}}$

$$\delta(f) \ge \frac{1}{2}\left(\frac{1}{\sqrt{m}} - \|f\|\right);$$

ii) for all $r \ge \frac{1}{\sqrt{m}}$, there exists $h_r \in \operatorname{conv} G$ such that $\|h_r\| = r$ and

$$\delta(h_r) \le 1 - 8^{-\frac{1}{m-1}}\, e^{-\frac{\ln(m-1)}{m-1}}.$$

Estimates from Corollary 4 can be applied to the space of real-valued Boolean functions of d-variables with either Euclidean or Fourier basis [19], [25].

VI. Discussion

Our results contribute to the investigation of variable-basis approximation, which is an important tool in estimation of model complexity of neural networks, radial and kernel basis functions, splines, and other computational models with flexible parameters of functions from suitable dictionaries.

Theorem 4 and Corollary 1 show that for all functions from convex hulls, Maurey–Jones–Barron's [12]–[14] estimate on rates of variable-basis approximation can be improved to a geometric rate. However, such a rate depends on a parameter $\delta(f) \in (0,1]$ specific for the function f to be approximated. Our results show that Lavretsky's geometric upper bound proven in [22] for functions in affine interiors of closures of convex hulls of "dictionaries" in finite-dimensional spaces holds for *all functions in convex hulls in general Hilbert spaces*. Moreover, our theorems imply a geometric rate for all functions in linear spans of "dictionaries."

Although theoretically a geometric rate holds quite generally, its real effect depends on the size of the parameter $\delta(f)$. As our examples in Section IV illustrate, for some functions, the values of these parameters may be so close to 1 that, unless n is sufficiently large, Maurey–Jones–Barron's bound is better. This is not very surprising, because we considered functions that are the worst cases also for Maurey–Jones–Barron's theorem [18]. However, these examples provide some insights into the structure of sets $A_\delta(G)$ formed by functions with the same $\delta(f) = \delta$. They also show that generally Theorem 4 cannot be extended from the convex hull to its closure.

The proof of our main theorem is constructive and can be described as an incremental procedure, constructing a sequence of approximants of a function belonging to the convex hull of a given finite dictionary. However, the procedure requires in each step a search through the whole dictionary, whose feasibility for large dictionaries might require some heuristics enabling to fasten the search.

References

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Clifss, NJ: Prentice-Hall, 1998.

[2] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg, and Q. Zhang, "Nonlinear black-box models in system identification: Mathematical foundations," *Automatica*, vol. 31, pp. 1725–1750, 1995.

[3] P. C. Kainen, V. Kůrková, and M. Sanguineti, "Minimization of error functionals over variable-basis functions," *SIAM J. Optim.*, vol. 14, pp. 732–742, 2003.

[4] K. A. Smith, "Neural networks for combinatorial optimization: A review of more than a decade of research," *INFORMS J. Comput.*, vol. 11, no. 1, pp. 15–34, 1999.

[5] V. Kůrková and M. Sanguineti, "Error estimates for approximate optimization by the extended Ritz method," *SIAM J. Optim.*, vol. 15, pp. 261–287, 2005.

[6] K. S. Narendra and S. Mukhopadhyay, "Adaptive control using neural networks and approximate models," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 475–485, May 1997.

[7] R. Zoppoli, T. Parisini, and M. Sanguineti, "Approximating networks and extended Ritz method for the solution of functional optimization problems," *J. Optim. Theory Appl.*, vol. 112, pp. 403–439, 2002.

[8] R. Zoppoli, T. Parisini, M. Baglietto, and M. Sanguineti, *Neural Approximations for Optimal Control and Decision*. London, U.K.: Springer-Verlag, to be published.

[9] A. Pinkus, *N-Widths in Approximation Theory*. New York: Springer-Verlag, 1986.

[10] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[11] V. Kůrková and M. Sanguineti, "Comparison of worst-case errors in linear and neural network approximation," *IEEE Trans. Inf. Theory*, vol. 28, no. 1, pp. 264–275, Jan. 2002.

[12] G. Pisier, "Remarques sur un résultat non publié de B. Maurey," in *Séminaire d'Analyse Fonctionnelle 1980–81*, Palaiseau, France, Exposé no. V, pp. V-1–V-12, École Polytechnique, Centre de Mathématiques.

[13] 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.*, vol. 20, pp. 608–613, 1992.

[14] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[15] A. R. Barron, "Neural net approximation," in *Proc. 7th Yale Workshop Adapt. Learn. Syst.*, K. Narendra, Ed., 1992, pp. 69–72, Yale Univ. Press.

[16] V. Kůrková, "Dimension-independent rates of approximation by neural networks," in *Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality*, K. Warwick and M. Kárný, Eds. Boston, MA: Birkhäuser, 1997, pp. 261–270.

[17] V. Kůrková, "High-dimensional approximation and optimization by neural networks," in *Advances in Learning Theory: Methods, Models and Applications*, J. Suykens, G. Horváth, S. Basu, C. Micchelli, and J. Vandewalle, Eds. Amsterdam, The Netherlands: IOS Press, 2003, ch. 4, pp. 69–88.

[18] V. Kůrková and M. Sanguineti, "Bounds on rates of variable-basis and neural-network approximation," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2659–2665, Nov. 2001.

[19] V. Kůrková, P. Savický, and K. Hlaváčková, "Representations and rates of approximation of real-valued boolean functions by neural networks," *Neural Netw.*, vol. 11, pp. 651–659, 1998.

[20] V. Kůrková and M. Sanguineti, "Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets," *Discrete Appl. Math.*, vol. 155, pp. 1930–1942, 2007.

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

[22] E. Lavretsky, "On the geometric convergence of neural approximations," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 274–282, Jan. 2002.

[23] V. Kůrková and M. Sanguineti, "Geometric rates of approximation by neural networks," in *Lecture Notes in Computer Science*, V. Geffert, J. Karhumaki, A. Bertoni, B. Preneel, P. Návrat, and M. Bieliková, Eds. Berlin, Germany: Springer-Verlag, 2008, vol. 4910, pp. 541–550.

[24] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*. New York: Dover, 1975.

[25] P. C. Kainen, V. Kůrková, and M. Sanguineti, "Rates of approximate minimization of error functionals over Boolean variable-basis functions," *J. Math. Model. Algorithms*, vol. 4, pp. 355–368, 2005.