Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2011 Special Issue

Can dictionary-based computational models outperform the best linear ones?

Giorgio Gnecco[a,*], Věra Kůrková[b], Marcello Sanguineti[a]

[a] *Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy*

[b] *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic*

ARTICLE INFO

Keywords:
Dictionary-based approximation
Linear approximation
Rates of approximation
Worst-case error
Kolmogorov width
Perceptron networks

ABSTRACT

Approximation capabilities of two types of computational models are explored: dictionary-based models (i.e., linear combinations of n-tuples of basis functions computable by units belonging to a set called "dictionary") and linear ones (i.e., linear combinations of n fixed basis functions). The two models are compared in terms of approximation rates, i.e., speeds of decrease of approximation errors for a growing number n of basis functions. Proofs of upper bounds on approximation rates by dictionary-based models are inspected, to show that for individual functions they do not imply estimates for dictionary-based models that do not hold also for some linear models. Instead, the possibility of getting faster approximation rates by dictionary-based models is demonstrated for worst-case errors in approximation of suitable sets of functions. For such sets, even geometric upper bounds hold.

1. Introduction

Many computational models used in neurocomputing belong to the class of "*dictionary-based computational models*". They consist of one "*hidden layer*" of computational units (such as perceptrons, kernel units, and radial units) and one linear output. The set of functions computable by hidden units is called a *dictionary* (Gribonval & Vandergheynst, 2006). Dictionary-based models with n hidden units compute all linear combinations of arbitrary n-tuples of functions from the dictionary. Thus they are sometimes called "*variable-basis models*" (Kůrková & Sanguineti, 2002), in contrast to linear models, where one has merely linear combinations of first n elements of a set of basis functions with a *fixed linear ordering*.

Dictionary-based models have been extensively used in machine learning and optimization tasks (see, e.g., Alessandri & Sanguineti, 2005; Giulini & Sanguineti, 2009; Gnecco & Sanguineti, 2009, 2010; Kůrková & Sanguineti, 2008a, and references therein). These models contain more adjustable parameters than linear models with the same number of units. Indeed, in addition to output weights (i.e., coefficients of linear combinations of hidden-unit functions) also inner parameters of hidden units (such as input weights, biases, widths, and centers) are adjustable during learning. So, it seems obvious that dictionary-based models should be able to achieve better rates in approximation of multivariable functions. This has been confirmed experimentally. For example, in Zoppoli, Sanguineti, and Parisini (2002) simulation results on high-dimensional optimization tasks were presented, showing that linear combinations of basis functions with adjustable parameters perform better than linear schemes expressed as linear combinations of the same basis functions with fixed parameter values—the overall number of adjustable parameters (i.e., coefficients of linear combinations and parameters in the basis functions) being fixed.

Yet theoretically, proving that dictionary-based models are more powerful than the linear ones performing the same multivariable approximation tasks is not easy. The comparison of linear models with variable-basis ones was initiated by Barron (1993). He proposed to compare upper bounds on dictionary-based approximation of certain sets of functions with lower bounds on worst-case errors in approximation of the same sets by optimal linear approximators, formalized by the concept of "*n-width*" (Kolmogorov, 1936). The n-width of a set, introduced by Kolmogorov and later called the "*Kolmogorov width*", is defined as the worst-case error in approximation of elements of the set by an optimal n-dimensional subspace; it measures "how far" the set is from being n-dimensional.

Using the "big O" and "big Ω" notations (see, e.g., Knuth, 1976), Barron (1993) showed that certain dictionary-based models can achieve a worst-case approximation error of order $O\left(n^{-1/2}\right)$, while Pinkus (1985, pp. 232–233) proved that asymptotically linear methods cannot do better than $\Omega\left(n^{-1/d}\right)$, where d denotes the number of variables, i.e., the number of inputs to the computational model. But such a comparison must be considered with care, as the "big O" and "big Ω" notations hide coefficients that often depend on the number d of variables (Kainen, Kůrková, & Sanguineti, 2009b) and, in addition, the two bounds apply to different sets of functions. The $O\left(n^{-1/2}\right)$ bound from Barron (1993) applies to

[*] Corresponding author. Tel.: +39 010 353 2071.
E-mail addresses: giorgio.gnecco@dist.unige.it (G. Gnecco), vera@cs.cas.cz (V. Kůrková), marcello@dist.unige.it (M. Sanguineti).

0893-6080/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2011.05.014

the dictionary-based approximation in the worst case for balls in certain norms tailored to the dictionary (Kůrková, 2003). Instead, the $\Omega\left(n^{-1/d}\right)$ bound for linear approximation applies to the best possible linear approximator used for the worst case in Sobolev balls. Although some embeddings of Sobolev balls into balls defined via norms tailored to certain dictionaries were derived (Barron, 1993; Kainen, Kůrková, & Sanguineti, 2009a; Kainen, Kůrková, & Vogt, 2007), the radii of these balls may depend on d exponentially.

In Barron (1993), Barron investigated worst-case errors in approximation of certain sets of functions by perceptron neural networks and compared these errors with Kolmogorov widths of related sets of functions (differing by the domains). Kůrková and Sanguineti (2002) extended his results to sets of functions on the same domains and more general dictionaries. For kernel units, Gnecco, Kůrková, and Sanguineti (2010) compared upper bounds on dictionary-based approximation due to Barron (1993), Jones (1992) and Maurey (see Pisier (1981)) with lower bounds on the Kolmogorov width. They derived their comparisons using properties of integral operators with kernels corresponding to computational units and relationships among these operators and norms induced by computational units.

In this paper, we further develop the comparison between linear and dictionary-based models. First, by analyzing constructive proofs of estimates of approximation rates by variable-basis models derived in the form $cn^{-1/2}$ by Barron (1993) and in the form c^{n-1} by Kůrková and Sanguineti (2008b), we show that these proofs do not provide arguments for faster rates of approximation by dictionary-based models than by linear ones for individual functions. Indeed, such proofs are based on incremental procedures constructing for each function to be approximated a linear ordering of the dictionary, specially tailored to this function. So, an upper bound on the distance from the linear subspace generated by the first n elements in this ordering also implies an upper bound on the distance from the set of functions computable by the dictionary-based model. To emphasize implications of the constructive proofs of upper bounds on dictionary-based approximation from Barron (1993) and Kůrková and Sanguineti (2008b), we reformulate these upper bounds as estimates of rates of linear approximators specially tailored to each function to be approximated. Thus, for individual functions the upper bounds from Barron (1993) and Kůrková and Sanguineti (2008b) do not give better estimates of rates of dictionary-based approximators than estimates of rates of linear approximators.

However, the upper bounds on dictionary-based approximation hold for all functions in rather large sets defined by suitable constraints (such as bounds on certain variational norms), so they can be exploited to compare worst-case errors in these sets for both types of approximators: dictionary-based and linear ones. Exploiting certain invariance properties of worst-case errors in linear approximation, we derive lower bounds on the n-widths of sets of functions for which upper bounds from Barron (1993) and Kůrková and Sanguineti (2008b) hold. We compare worst-case errors in linear and dictionary-based approximation, for dictionaries that are "large enough" to contain orthogonal sets, the elements of which have norms decreasing to zero "rather slowly". Taking advantage of orthonormal sets related to dictionaries formed by perceptrons with periodic or sigmoidal activation functions, we obtain proofs of better rates of approximation by such perceptron networks than by linear models.

The paper is organized as follows. In Section 2, we introduce notations and basic properties of approximation by linear and dictionary-based models. In Section 3, we analyze incremental constructions which were used in Barron (1993), Kůrková and Sanguineti (2008b) and Lavretsky (2002) to derive upper bounds on dictionary-based approximation. In Section 4, we investigate worst-case errors in approximation by linear and dictionary-based computational models for sets of functions associated with orthonormal dictionaries. We apply the results to perceptrons with certain periodic activations. Section 5 offers a comparison of linear and dictionary-based approximation for orthogonal dictionaries, with application to sigmoidal perceptron networks.

2. Approximation by linear and dictionary-based models

In traditional *linear computational models*, nested sets of the form

$$\operatorname{span}\{g_1, \ldots, g_n\} := \left\{ \sum_{i=1}^{n} w_i g_i \mid w_i \in \mathbb{R} \right\}$$

are used as approximating families ($\mathbb{R}$ denotes the set of real numbers). They are formed by linear combinations of the *first n* elements from a set $G = \{g_i \mid i \in \mathbb{N}_+\}$ ($\mathbb{N}_+$ denotes the set of positive integers) with a *fixed linear ordering* (typically, an ordered sets of polynomials). In linear regression, merely coefficients of a linear combination of these n a priori chosen functions are adjustable.

In contrast to linear approximation schemes with a fixed linear ordering of G, many computational models used in neurocomputing can be formally described as *variable-basis* schemes. They compute functions from sets

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

where G is a set of functions, sometimes called *dictionary* (Gribonval & Vandergheynst, 2006), and n is the number of *computational units*. Note that for G linearly independent with card $G > n$, the set $\operatorname{span}_n G$ is not convex. Approximation by the family $\{\operatorname{span}_n G,\ n \in \mathbb{N}_+\}$ is called *variable-basis approximation* (Kůrková & Sanguineti, 2001, 2002, 2008b) or *approximation from a dictionary* (Gribonval & Vandergheynst, 2006).

Typically, dictionaries are parameterized sets of functions of the form

$$G_\phi = G_\phi(Y) := \{\phi(\cdot, y) \mid y \in Y\},$$

where $\phi : \Omega \times Y \to \mathbb{R}$ is a function of two vector variables, $\Omega \subseteq \mathbb{R}^d$ represents the set of inputs, and $Y \subseteq \mathbb{R}^q$ the set of adjustable parameters. For example, elements of the dictionary can be perceptrons, radial-basis-functions, or kernel units. Sets $\operatorname{span}_n G_\phi$ model sets of input–output functions of one-hidden-layer neural networks, radial-basis-function networks, kernel models, splines with free nodes, trigonometric polynomials with variable frequencies and phases, etc. (Kůrková & Sanguineti, 2002). The number n of hidden units can be interpreted as *model complexity*. For example, if $q = d + 1$ and

$$\phi(\cdot, (v, b)) := \psi(\langle v, \cdot\rangle + b),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^d$, then the dictionary G_ϕ is formed by functions computable by *perceptrons* with an activation unit $\psi : \mathbb{R} \to \mathbb{R}$, where $v \in \mathbb{R}^d$ is an *input weight vector* and $b \in \mathbb{R}$ is a *bias*. If $q = d + 1$, ψ is positive and even, and

$$\phi(\cdot, (v, b)) := \psi(b\| \cdot - v\|),$$

then G_ϕ is formed by functions computable by a *radial unit* $\psi : \mathbb{R} \to \mathbb{R}_+$.

By $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ we denote a normed linear space and we write merely $\mathcal{X}$ when there is no ambiguity. The error in approximation of a function $f \in \mathcal{X}$ by functions from a set A is measured by the *distance of f from A*

$$\|f - A\|_{\mathcal{X}} := \inf_{g \in A} \|f - g\|_{\mathcal{X}}.$$

Approximation capabilities of sets of functions can be studied in terms of *worst-case errors*, formalized by the concept of *deviation*. For two subsets A and M of $\mathcal{X}$, the deviation of M from A is defined as

$$\begin{aligned}\delta(M,A) &= \delta(M,A;\mathcal{X}) = \delta(M,A;(\mathcal{X},\|\cdot\|_{\mathcal{X}}))\\ &:= \sup_{f\in M}\inf_{g\in A}\|f-g\|_{\mathcal{X}}.\end{aligned} \tag{1}$$

We use the short notations when the ambient space and/or its norm are clear from the context. When the supremum in (1) is achieved, the deviation is the *worst-case error* in approximation of functions from M by functions from A.

Sometimes, the set M of functions to be approximated is described in terms of a constraint that defines a norm $\|\cdot\|$ on $\mathcal{X}$ or on its subspace. For instance, the set M may be the ball

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

of radius r in the norm $\|\cdot\|$, centered in the origin.

To describe a theoretical lower bound on worst-case errors in approximation by optimal linear subspaces, Kolmogorov (1936) introduced the concept of *n-width* (later called *Kolmogorov n-width*). Let $\mathcal{S}_n$ denote the *family of all n-dimensional linear subspaces of* $\mathcal{X}$. The Kolmogorov n-width of a subset M of a normed linear space $(\mathcal{X},\|\cdot\|_{\mathcal{X}})$ is defined as the infimum of the deviations of M from all n-dimensional linear subspaces of $\mathcal{X}$, i.e.,

$$\begin{aligned}d_n(M) &= d_n(M;\mathcal{X}) = d_n(M;(\mathcal{X},\|\cdot\|_{\mathcal{X}}))\\ &:= \inf_{\mathcal{X}_n\in\mathcal{S}_n}\delta(M,\mathcal{X}_n;(\mathcal{X},\|\cdot\|_{\mathcal{X}}))\\ &= \inf_{\mathcal{X}_n\in\mathcal{S}_n}\sup_{f\in M}\inf_{g\in\mathcal{X}_n}\|f-g\|_{\mathcal{X}}.\end{aligned} \tag{2}$$

We adopt the short notations when there is no ambiguity. If for some subspace the infimum is achieved, then the subspace is called *optimal*. Loosely speaking, if the n-width of a set is "small", then such a set can be viewed as "almost" n-dimensional, in the sense that it is contained in a small neighborhood of some n-dimensional subspace. It follows from the definition that the n-width does not increase when a set is extended to its closure or its convex hull, i.e.,

$$d_n(M) = d_n(\mathrm{cl}_{\mathcal{X}}M) \quad\text{and}\quad d_n(M) = d_n(\mathrm{conv}\,M), \tag{3}$$

where conv denotes the *convex hull* and $\mathrm{cl}_{\mathcal{X}}$ the *closure* in the topology induced by the norm $\|\cdot\|_{\mathcal{X}}$.

We shall compare $d_n(M)$ with the deviation $\delta(M,\mathrm{conv}_n\,G)$, where

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

Since $\mathrm{conv}_n\,G \subseteq \mathrm{span}_n\,G$, upper bounds on the deviation $\delta(M,\mathrm{conv}_n\,G)$ are also upper bounds on the deviation $\delta(M,\mathrm{span}_n\,G)$. Of course, the worst-case error in linear approximation by an optimal n-dimensional subspace generated by elements of G cannot be smaller than the worst-case error in dictionary-based approximation by $\mathrm{span}_n\,G$. However, this does not exclude the possibility that among other linear approximators than those generated by elements of G, there exists one that approximates the set M better than $\mathrm{span}_n\,G$, i.e., such that

$$d_n(M) < \delta(M,\mathrm{span}_n\,G) \le \delta(M,\mathrm{conv}_n\,G).$$

Obviously, the description of cases when the inequality

$$\delta(M,\mathrm{conv}_n\,G) < d_n(M) \tag{4}$$

holds is of a great interest. For such sets M, worst-case errors in approximation by $\mathrm{conv}_n\,G$ are smaller than those in approximation by *any* linear n-dimensional subspace.

3. Inspection of upper bounds on approximation rates by dictionaries

In this section, we reformulate upper bounds on dictionary-based approximation as upper bounds on special linear approximators.

The following upper bound is a version of Jones' result (Jones, 1992) as improved by Barron (1993) (see also in Pisier (1981) an earlier estimate derived by Maurey).

Theorem 1 (*Maurey–Jones–Barron, Barron, 1993, Jones, 1992 and Pisier, 1981*)**.** *Let $(\mathcal{X},\|\cdot\|_{\mathcal{X}})$ be a Hilbert space, G its bounded nonempty subset, $s_G = \sup_{g\in G}\|g\|_{\mathcal{X}}$, and $f\in\mathrm{cl}_{\mathcal{X}}\,\mathrm{conv}\,G$. Then, for every positive integer n*

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

In Kůrková (1997) (see also Kůrková, 2003), Theorem 1 was extended using the concept of *G-variation*, defined for all functions $f\in\mathcal{X}$ as

$$\|f\|_G := \inf\{c>0 \mid f/c\in\mathrm{cl}_{\mathcal{X}}\,\mathrm{conv}(G\cup -G)\},$$

where

$$-G := \{-g \mid g\in G\}.$$

Note that $\|\cdot\|_G$ is the Minkowski functional (Kolmogorov & Fomin, 1970, p. 131) of the set $\mathrm{cl}_{\mathcal{X}}\,\mathrm{conv}(G\cup -G)$ and so it is a norm on the subspace of $\mathcal{X}$ containing the elements $f\in\mathcal{X}$ for which $\|f\|_G<\infty$.

Lavretsky (2002) noticed that the argument used by Barron (1993) and Jones (1992) can yield better rates when applied to functions satisfying a certain angular relationship with respect to G. In Kůrková and Sanguineti (2008b), Kůrková and Sanguineti showed that the estimate derived in Lavretsky (2002) holds for all functions in the convex hull of any bounded subset of any Hilbert space.

Theorem 2 (*Kůrková & Sanguineti, 2008b*)**.** *Let $(\mathcal{X},\|\cdot\|_{\mathcal{X}})$ be a Hilbert space, G its bounded nonempty subset, and $s_G = \sup_{g\in G}\|g\|$. Then, for every $f\in\mathrm{conv}\,G$ there exists $\tau_f\in[0,1)$ such that for every positive integer n*

$$\|f-\mathrm{conv}_n\,G\|_{\mathcal{X}}^2 \le \tau_f^{n-1}\left(s_G^2-\|f\|_{\mathcal{X}}^2\right).$$

Let

$$\tau(f) := \inf\{\tau>0 \mid \|f-\mathrm{conv}_nG\|_{\mathcal{X}}^2 \le \tau^{n-1}(s_G^2-\|f\|^2)\}. \tag{5}$$

For every $f\in\mathrm{conv}\,G$, the set over which this minimum is taken is nonempty, closed, and bounded, so the infimum is attained—i.e.,

$$\|f-\mathrm{conv}_nG\|_{\mathcal{X}}^2 \le \tau(f)^{n-1}(s_G^2-\|f\|_{\mathcal{X}}^2).$$

The proof of Theorem 2 is based on a constructive incremental procedure, which improves the one used to prove Theorem 1. In both these proofs, for every function $f\in\mathrm{conv}\,G$ and its every representation $f=\sum_{j=1}^m a_jg_j$ as a convex combination of elements of G, a linear ordering

$$\{g_{j_1},\dots,g_{j_m}\}$$

of the subset

$$G' := \{g_1,\dots,g_m\}$$

is constructed. Then, it is shown that for every positive integer $n\le m$ in the case of Theorem 1 one has

$$\|f-\mathrm{span}\{g_{j_1},\dots,g_{j_n}\}\|_{\mathcal{X}}^2 \le \frac{s_G^2-\|f\|_{\mathcal{X}}^2}{n},$$

whereas in Theorem 2 for some $\tau_f\in[0,1)$ one has

$$\|f - \mathrm{span}\{g_{j_1}, \ldots, g_{j_n}\}\|_{\mathcal{X}}^2 \le \tau_f^{n-1}\left(s_G^2 - \|f\|_{\mathcal{X}}^2\right).$$

Formally, the proof of Theorem 2 from Kůrková and Sanguineti (2008b) can be described as the following incremental construction. The structure of the proof of Theorem 1 from Barron (1993) is similar to the structure of the proof of Theorem 2, merely the choice of g_{j_n} is simpler.

Incremental construction

```
Choose $g_{j_1} \in \{g_j \,|\, j = 1, \ldots, m\}$ such that $\|f - g_{j_1}\|_{\mathcal{X}} = \min_{j=1,\ldots,m} \|f - g_j\|_{\mathcal{X}}$;
$f_1 := g_{j_1}$;
  For $n = 2, \ldots, m - 1$:
    begin
      for $j = 1, \ldots, m$,
      compute $\eta_j := -\frac{(f-f_{n-1})\cdot(f-g_j)}{\|f-f_{n-1}\|_{\mathcal{X}}\,\|f-g_j\|_{\mathcal{X}}}$;
      if for $j = 1, \ldots, m$ one has $\eta_j = 0$, then
        begin
          $f^* := f_{n-1}$;
          $n^* := n - 1$;
        end
    else
      begin
          $\rho_n := \max\{\eta_j > 0 \,|\, j = 1, \ldots, m\}$;
          choose $g_{j_n}$ such that $\rho_n = \eta_{j_n}$;
          compute $e_{n-1} := \|f - f_{n-1}\|_{\mathcal{X}}$;
          compute $r_n := \|f - g_{j_n}\|_{\mathcal{X}}$;
          compute $\alpha_n := \frac{\rho_n e_{n-1} r_n + r_n^2}{e_{n-1}^2 + 2\rho_n e_{n-1} r_n + r_n^2}$;
          $f_n := \alpha_n f_{n-1} + (1 - \alpha_n)\, g_{jn}$;
          $n := n + 1$.
      end
    end
Let
```

$$k := \max\{n \in \{1, \ldots, m\} \mid f_n \ne f_{n-1}\}$$
$$\tau_f := \min\{(1 - \rho_n^2) \mid n = 1, \ldots, k\}.$$

For every $\tau \in [0, 1)$, let

$$A_\tau(G) := \{f \in \mathrm{conv}\, G \mid \tau(f) = \tau\}.$$

The next proposition summarizes properties of the sets A_τ.

Proposition 1. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space and G its bounded nonempty subset. Then for every $\tau \in [0, 1)$ the following hold.*

(i) *$G \subseteq \bigcap_{\tau>0} A_\tau(G)$;*
(ii) *if $\tau_1 \le \tau_2$, then $A_{\tau_1}(G) \subseteq A_{\tau_2}(G)$;*
(iii) *$\mathrm{conv}\, G = \bigcup_{\delta\in(0,1]} A_\delta(G)$.*

Proof. (i) and (ii) follow by the definition of $A_\tau(G)$, whereas (iii) is implied by Theorem 2. □

The inspections of the above-described incremental construction from the proof of Theorem 2 and of the simpler construction from the proof of Theorem 1, allow one to reformulate such theorems in terms of linear approximators tailored to functions to be approximated. This is done in the next theorem.

Theorem 3. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space, G its bounded nonempty subset, and $s_G := \sup_{g\in G} \|g\|_{\mathcal{X}}$. Then*

(i) *for every $f \in A_\tau(G)$ there exists a positive integer m and a linear ordering $\{g_{j_1}, \ldots, g_{j_m}\}$ of a subset of m elements of G such that for all $n = 1, \ldots, m$*
$$\|f - \mathrm{span}\{g_{j_1}, \ldots, g_{j_n}\}\|_{\mathcal{X}} \le s_G \tau^{\frac{n-1}{2}};$$
(ii) *for every $f \in \mathrm{conv}\, G$ there exists a positive integer m and a linear ordering $\{g_{j_1}, \ldots, g_{j_m}\}$ of a subset of m elements of G such that for all $n = 1, \ldots, m$*
$$\|f - \mathrm{span}\{g_{j_1}, \ldots, g_{j_n}\}\|_{\mathcal{X}} \le s_G \tau^{\frac{n-1}{2}};$$
(iii) *for every positive integer n*
$$\delta(\mathrm{conv}\, G, \mathrm{conv}_n\, G) \le \frac{s_G}{\sqrt{n}}.$$
(iv) *for every $\tau \in [0, 1)$ and every positive integer n*
$$\delta(A_\tau(G), \mathrm{conv}_n\, G) \le s_G \tau^{\frac{n-1}{2}}.$$

Theorems 1–3 provide the same approximation rates for the linear and the dictionary-based approximator. However, there is a substantial difference between these two cases: in the case of dictionary-based approximation the rate holds for all functions from conv G or $A_\tau(G)$, so it also holds for the worst case. In the case of linear approximation, for each function f a specific linear ordering is constructed, so there is no guarantee of one linear ordering serving as a linear approximator with a guaranteed rate for all functions from these sets.

4. Worst-case errors for orthonormal dictionaries and perceptrons with periodic activations

In this section we compare, for sets $A_\tau(G)$ with G orthonormal, the worst-case errors by linear computational models with those by the dictionary G itself. To this end, we estimate from below the n-widths of sets $A_\tau(G)$ and compare such lower bounds with the upper bound provided by Theorem 3(iv) on deviation from $\mathrm{conv}_n\, G$. Then, we apply the results to orthonormal dictionaries corresponding to perceptrons with periodic activation functions.

We take advantage of the invariance of the Kolmogorov width under operations of symmetrization and convex closure. It follows from the definition (2) that for every Hilbert space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, every bounded nonempty set $G \subset \mathcal{X}$, and every positive integer n one has

$$d_n(G) = d_n(B_1(\|\cdot\|_G)). \quad (6)$$

Hence, lower bounds on $d_n(G)$ can be obtained by estimating $d_n(B_1(\|\cdot\|_G))$. Often this is easier, as balls $B_1(\|\cdot\|_G)$ are much larger than sets G. In some cases, such balls contain large orthogonal subsets, for which worst-case errors in linear approximation can be estimated from below.

In Kůrková and Sanguineti (2002, Corollary 2), we derived a lower bound on $d_n(G)$ via orthonormal sets contained in balls in G-variation. Let G and S be bounded nonempty subsets of a Hilbert space and S orthonormal such that $0 \ne s_{S,G} = \sup_{\alpha\in S} \|\alpha\|_G < \infty$. For S infinite, Kůrková and Sanguineti (2002, Corollary 2) states that for every positive integer n one has

$$d_n(G) \ge \frac{1}{s_{S,G}} \quad (7)$$

and that for S finite of cardinality m and every positive integer $n \le m$ one has

$$d_n(G) \ge \frac{1}{s_{S,G}}\sqrt{1 - \frac{n}{m}}. \quad (8)$$

On the other hand, Proposition 1(i) implies that for every Hilbert space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, every bounded nonempty set $G \subset \mathcal{X}$, every $\tau \in [0, 1)$, and every positive integer n one has

$$d_n(A_\tau(G)) \ge d_n(G). \quad (9)$$

Combining Eqs. (6)–(9) and taking into account Theorem 3(iv), we get the following proposition.

Proposition 2. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space, G and S its bounded nonempty subsets, $s_G := \sup_{g\in G} \|g\|_{\mathcal{X}}$, and S orthonormal with $0 \neq s_{S,G} := \sup_{\alpha\in S} \|\alpha\|_G < \infty$. Then*

(i) *for every $\tau \in [0, 1)$ and every positive integer n*

$$\delta(A_\tau(G), \mathrm{conv}_n\, G) \le s_G\, \tau^{\frac{n-1}{2}};$$

(ii) *if S is infinite, then for every positive integer n*

$$d_n(A_\tau(G)) \ge \frac{1}{s_{S,G}};$$

(iii) *if S is finite of cardinality m, then for every positive integer $n \le m$*

$$d_n(A_\tau(G)) \ge \frac{1}{s_{S,G}}\sqrt{1-\frac{n}{m}}.$$

Proposition 2(ii) implies that, whenever the unit ball in variation with respect to the dictionary G contains a ball of non-zero radius η in variation with respect to an infinite orthonormal set, $A_\tau(G)$ cannot be approximated with an error smaller than η using a linear computational model. In other words, no increase of the number n can decrease the Kolmogorov n-width of $A_\tau(G)$ below η. Instead, by Proposition 2(i) the deviation of $A_\tau(G)$ from convex combinations of n elements of the dictionary G decreases exponentially fast when n is increased.

When the dictionary G is orthonormal, we get the following corollaries.

Corollary 1. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space and G its orthonormal subset. Then*

(i) *for every $\tau \in [0, 1)$ and every positive integer n*

$$\delta(A_\tau(G), \mathrm{conv}_n\, G) \le \tau^{\frac{n-1}{2}};$$

(ii) *if G is infinite, then for every positive integer n*

$$d_n(A_\tau(G)) \ge 1;$$

(iii) *if G is finite of cardinality m, then for every positive integer $n \le m$*

$$d_n(A_\tau(G)) \ge \sqrt{1-\frac{n}{m}}.$$

Corollary 2. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space and G its infinite orthonormal subset. Then for every integer $n > 1$ and every $\tau \in [0, 1)$*

$$d_n(A_\tau(G)) > \delta(A_\tau(G), \mathrm{conv}_n G).$$

For a set $\Omega \subseteq \mathbb{R}^d$, we denote by $(\mathcal{L}_2(\Omega), \|\cdot\|_2)$ the space of Lebesgue-measurable functions that are square integrable, endowed with the $\mathcal{L}_2$-norm. Corollaries 1 and 2 can be exploited to compare the $\mathcal{L}_2$-approximation of linear models and perceptron networks with certain periodic activation functions.

Let J be a closed interval in $\mathbb{R}$. We denote by

$$P_d(\psi, J) := \left\{f : J^d \to \mathbb{R} \mid f(x) = \psi(v\cdot x + b),\, v \in \mathbb{R}^d,\, b \in \mathbb{R}\right\}$$

the set of functions on J^d computable by ψ-perceptrons. So, $\mathrm{conv}_n\, P_d(\psi, J)$ represents the set of functions on J^d computable by *ψ-perceptron networks with n hidden units*. We consider perceptrons with activations that are the sine or the *Haar function*, which is denoted by ξ and defined as $\xi(t) = 1$ for $t \in [i, i + 1/2)$ and $\xi(t) = -1$ for $t \in [i - 1/2, i)$ for all integers i. Set $J = [0, 1]$ and, to simplify the notation, for every activation function ψ let

$$P_d(\psi) := P_d(\psi, [0, 1]).$$

Proposition 3. *For all positive integers n, d and every $\tau \in [0, 1)$, the following hold in $\mathcal{L}_2([0, 1]^d)$:*

$$\delta(A_\tau(P_d(\sin)), \mathrm{conv}_n\, P_d(\sin)) \le \frac{\tau^{\frac{n-1}{2}}}{\sqrt{2}};$$

$$d_n(A_\tau(P_d(\sin))) \ge \frac{1}{\sqrt{2}},$$

$$\delta(A_\tau(P_d(\xi)), \mathrm{conv}_n\, P_d(\xi)) \le \tau^{\frac{n-1}{2}};$$

$$d_n(A_\tau(P_d(\xi))) \ge 1.$$

Proof. It is easy to check that for every positive integer d the following two families of functions are orthonormal in $\mathcal{L}_2([0, 1]^d)$:

$$S_d(\sin) := \left\{\sqrt{2}\sin(\pi v\cdot x) \mid v \in \mathbb{N}_+^d\right\};$$

$$S_d(\xi) := \left\{\xi(v\cdot x) \mid v \in \{2^j; j \in \mathbb{N}\}^d\right\}.$$

As the first family is a subset of $\sqrt{2}P_d(\sin)$ and the second one is a subset of $P_d(\xi)$, for every positive integers d, n and every $\tau \in [0, 1)$, Proposition 2(ii) gives in $\mathcal{L}_2([0, 1]^d)$ the lower bounds

$$d_n(P_d(\sin)) = d_n(B_1(\|\cdot\|_{P_d(\sin)})) \ge \frac{1}{\sqrt{2}}, \qquad (10)$$

$$d_n(P_d(\xi)) = d_n(B_1(\|\cdot\|_{P_d(\xi)})) \ge 1. \qquad (11)$$

The other estimates follow by Proposition 2(i). □

Inspection of the proof of Proposition 3 shows that one-hidden-layer perceptron networks with either the sine or the Haar function as an activation cannot be efficiently approximated by linear models. Indeed, the lower bounds (10) and (11) imply that there is no possibility of decreasing the $\mathcal{L}_2$-worst-case error in linear approximation of $P_d(\sin)$ and $P_d(\xi)$ under $1/\sqrt{2}$ and 1, resp., by increasing the number of basis functions in a linear computational model.

Corollary 3. *For every integer $n > 1$, every positive integer d, and every $\tau \in [0, 1)$, the following hold in $\mathcal{L}_2([0, 1]^d)$:*

$$d_n(A_\tau(P_d(\sin))) > \delta(A_\tau(P_d(\sin)), \mathrm{conv}_n\, P_d(\sin)),$$

$$d_n(A_\tau(P_d(\xi))) > \delta(A_\tau(P_d(\sin)), \mathrm{conv}_n\, P_d(\xi)).$$

5. Worst-case errors for orthogonal dictionaries and perceptrons with sigmoidal activations

The unit ball $B_1(\|\cdot\|_G)$ in variation with respect to the dictionary G may not be "large enough" to contain a ball of some non-zero radius in variation with respect to an infinite orthonormal set, as required by Proposition 2. However, $B_1(\|\cdot\|_G)$ may contain a ball in variation with respect to some orthogonal set, the elements of which have norms going to zero rather slowly with respect to a positive integer d. In Kůrková and Sanguineti (2002), such a slow decrease was formalized by introducing the concept of a set "not quickly vanishing with respect to d".

Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space, S its countable subset, and d a positive integer. We say that S is *not quickly vanishing with respect to d* if it can be linearly ordered as $S = \{\alpha_j \mid j \in \mathbb{N}\}$ in such a way that the norms of its elements are non-increasing and for every positive integer r one has $\|\alpha_{r^d}\| \ge 1/r$. Note that S is not quickly vanishing with respect to d if and only if it can be represented as $S = \bigcup_{r\in\mathbb{N}} S_r$, where $\mathrm{card}\, S_r \ge r^d$, $\|\alpha\| \ge 1/r$ for every $\alpha \in S_r$, and, for every positive integer $r' > r$ and $\alpha' \in S_{r'}$, one has $\|\alpha\| \ge \|\alpha'\|$.

In Kůrková and Sanguineti (2002, Corollary 3), it was proven that if an orthogonal set S is not quickly vanishing with respect to d and $0 \neq s_{S,G} := \sup_{\alpha \in S} \|\alpha\|_G < \infty$ for a dictionary G, then for $n = r^d/2$ with some integer r one has

$$d_n(G) \geq \frac{1}{s_{S,G}\sqrt{2}\sqrt[d]{2n}}. \tag{12}$$

Combining Eqs. (6), (9) and (12) and taking into account Theorem 3(ii), we get the following proposition.

Proposition 4. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space, G and S its bounded nonempty subsets, $s_G := \sup_{g \in G} \|g\|_{\mathcal{X}}$, S orthogonal not quickly vanishing with respect to a positive integer d, and $0 \neq s_{S,G} := \sup_{\alpha \in S} \|\alpha\|_G < \infty$. Then for every positive integer n and every $\tau \in [0, 1)$*

(i) $\delta(A_\tau(G), \mathrm{conv}_n\, G) \leq s_G \tau^{\frac{n-1}{2}}$;
(ii) *if* $2n = r^d$ *for some integer r, then* $d_n(A_\tau(G)) \geq \frac{1}{s_{S,G}\sqrt{2}\sqrt[d]{2n}}$.

When a dictionary G is a countable orthogonal set which is not quickly vanishing with respect to a positive integer d, we get the following estimates.

Corollary 4. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Hilbert space and G its countable orthogonal subset, not quickly vanishing with respect to a positive integer d and such that $s_G := \sup_{g \in G} \|g\|_{\mathcal{X}} \leq 1$. Then for every positive integer n and every $\tau \in [0, 1)$ the following hold:*

(i) $\delta(A_\tau(G), \mathrm{conv}_n\, G) \leq s_G \tau^{\frac{n-1}{2}}$;
(ii) *if* $2n = r^d$ *for some integer r, then* $d_n(A_\tau(G)) \geq \frac{1}{\sqrt{2}\sqrt[d]{2n}}$.

The lower bounds from Corollary 4 imply that in linear approximation of an orthogonal set of functions of d variables that is not quickly vanishing with respect to d, the number of basis functions necessary to guarantee an accuracy ε is of order $\Omega\left(1/\varepsilon^d\right)$. This lower bound exhibits the so-called "curse of dimensionality" (Bellman, 1957) (the term "dimensionality" referring to the number d of variables).

Corollary 4 can be exploited to compare linear computational models and dictionary-based models corresponding to perceptron neural networks with the most common activation functions, the so-called *sigmoidals*, i.e., bounded measurable functions $\sigma : \mathbb{R} \to \mathbb{R}$ such that $\lim_{t \to -\infty} \sigma(t) = 0$ and $\lim_{t \to +\infty} \sigma(t) = 1$. One can use both continuous sigmoidals (like the *logistic sigmoid* $1/(1 + \exp(-t))$ or the *hyperbolic tangent*) and the discontinuous *Heaviside function* ϑ, defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. Let J be a closed interval in $\mathbb{R}$ and

$$P_d(\vartheta, J) := \left\{ f : J^d \to \mathbb{R} \mid f(x) = \vartheta(v \cdot x + b),\, v \in \mathbb{R}^d,\, b \in \mathbb{R} \right\}.$$

Note that the set $P_d(\vartheta, J)$ of functions computable by Heaviside perceptrons is equal to the *set of characteristic functions of half-spaces* of $\mathbb{R}^d$ restricted to J^d. Indeed, the function $\vartheta(v \cdot . + b)$ restricted to J^d is equal to the characteristic function of the set $\{x \in J^d \mid v \cdot x + b \geq 0\}$. Analogously to Section 4, we consider $J = [0, 1]$ and to simplify the notation we let

$$P_d(\vartheta) := P_d(\vartheta, [0, 1]).$$

As variation with respect to half-spaces is bounded from below by variation with respect to perceptrons with any sigmoidal activation function σ (Kůrková & Sanguineti, 2002, Proposition 10(ii)), we get

$$d_n(B_1(\|\cdot\|_{P_d(\sigma)})) \geq d_n(B_1(\|\cdot\|_{H_d})). \tag{13}$$

Inspection of the proof of Kůrková and Sanguineti (2002, Theorem 2) shows that a lower bound on $d_n(H_d) = d_n(B_1(\|\cdot\|_{H_d}))$ was derived by using an orthogonal, not quickly vanishing set A_d obtained by a proper scaling of the elements of the set $A_d(\sin) = \left\{\sqrt{2}\sin(\pi v \cdot x) \mid v \in \mathbb{N}_+^d\right\}$.

For $d, r \in \mathbb{N}$, let

$$A_{d,r} := \left\{ \alpha_v(.) \mid v \in \{1, \ldots, r\}^d \right\} \subset \mathcal{L}_2([0, 1]^d),$$

where $\alpha_v(x) = c_v \sin(\pi v \cdot x) : [0, 1]^d \to \mathbb{R}$, $v = (v_1, \ldots, v_d) \in \mathbb{R}_+^d$ and $c_v = \frac{d\sqrt{2}}{\left\lceil \sum_{k=1}^d v_k \right\rceil}$. Set $A_d = \cup_{r \in \mathbb{N}} A_{d,r}$. First, it was proved in Kůrková and Sanguineti (2002, Theorem 2) that $A_d \subseteq B_{2d\sqrt{2}}(\|\cdot\|_{H_d})$, and that A_d is not quickly vanishing with respect to d. By embedding the sets A_d into a ball in variation with respect to sigmoidal perceptrons, in Kůrková and Sanguineti (2002, Theorem 2 and Corollary 5) the following lower bound was obtained in $\mathcal{L}_2([0, 1]^d)$ when $2n = r^d$ for some integer r:

$$d_n\left(B_1(\|\cdot\|_{P_d(\sigma)})\right) \geq \frac{1}{4d\sqrt[d]{2n}}. \tag{14}$$

By Eqs. (6), (9), (13) and (14), and taking into account Theorem 3(ii), we get the following proposition.

Proposition 5. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a sigmoidal function. Then for all positive integers d and n and every $\tau \in [0, 1)$, the following hold in $\mathcal{L}_2([0, 1]^d)$:*

(i) $\delta(A_\tau(P_d(\sigma)), \mathrm{conv}_n\, P_d(\sigma))) \leq \tau^{\frac{n-1}{2}}$;
(ii) *if* $2n = r^d$ *for some integer r, then* $d_n(A_\tau(P_d(\sigma))) \geq \frac{1}{4d\sqrt[d]{2n}}$.

Acknowledgments

G. Gnecco and M. Sanguineti were partially supported by a PRIN grant from the Italian Ministry for University and Research, project "Adaptive State Estimation and Optimal Control". V.K. was partially supported by MŠMT grant INTELLI OC10047 and the Institutional Research Plan AV0Z10300504. Collaboration of V. Kůrková with G. Gnecco and M. Sanguineti was partially supported by CNR—AV ČR 2010–2012 project "Complexity of Neural-Network and Kernel Computational Models".

References

Alessandri, A., & Sanguineti, M. (2005). Optimization of approximating networks for optimal fault diagnosis. *Optimization Methods and Software*, *20*, 235–260.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*, 930–945.

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Giulini, S., & Sanguineti, M. (2009). Approximation schemes for functional optimization problems. *Journal of Optimization Theory and Applications*, *140*, 33–54.

Gnecco, G., Kůrková, V., & Sanguineti, M. (2010). Some comparisons of model complexity in linear and neural-network approximation. In *Lecture notes in computer science*: *Vol. 6354* (pp. 358–367). Berlin, Heidelberg: Springer.

Gnecco, G., & Sanguineti, M. (2009). Accuracy of suboptimal solutions to kernel principal component analysis. *Computational Optimization and Applications*, *42*, 265–287.

Gnecco, G., & Sanguineti, M. (2010). Regularization techniques and suboptimal solutions to optimization problems in learning from data. *Neural Computation*, *22*, 793–829.

Gribonval, R., & Vandergheynst, P. (2006). On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *IEEE Transactions on Information Theory*, *52*, 255–261.

Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, *24*, 608–613.

Kainen, P. C., Kůrková, V., & Sanguineti, M. (2009a). Complexity of Gaussian radial basis networks approximating smooth functions. *Journal of Complexity*, *25*, 63–74.

Kainen, P. C., Kůrková, V., & Sanguineti, M. (2009b). On tractability of neural-network approximation. In *Lecture notes in computer science*: *Vol. 5495* (pp. 11–21).

Kainen, P. C., Kůrková, V., & Vogt, A. (2007). A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves. *Journal of Approximation Theory*, *147*, 1–10.

Knuth, D. E. (1976). Big omicron and big omega and big theta. *SIGACT News*, *8*, 18–24.

Kolmogorov, A. N. (1936). Über die beste annäherung von funktionen einer gegebenen funktionenklasse. *Annals of Mathematics*, *37*, 107–110. English translation: "On the best approximation of functions of a given class", in Selected works of A. N. Kolmogorov, vol. I; V. M. Tikhomirov Ed., pp. 202–205. Kluwer, 1991.

Kolmogorov, A. N., & Fomin, S. V. (1970). *Introductory real analysis*. New York: Dover.

Kůrková, V. (1997). Dimension-independent rates of approximation by neural networks. In K. Warwick, & M. Kárný (Eds.), *Computer-intensive methods in control and signal processing. the curse of dimensionality* (pp. 261–270). Boston: Birkhäuser.

Kůrková, V. (2003). High-dimensional approximation and optimization by neural networks. In J. Suykens, G. Horváth, S. Basu, C. Micchelli, & J. Vandewalle (Eds.), *Advances in learning theory: methods, models and applications* (pp. 69–88). Amsterdam: IOS Press, Chapter 4.

Kůrková, V., & Sanguineti, M. (2001). Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, *47*, 2659–2665.

Kůrková, V., & Sanguineti, M. (2002). Comparison of worst case errors in linear and neural network approximation. *IEEE Transactions on Information Theory*, *48*, 264–275.

Kůrková, V., & Sanguineti, M. (2008a). Approximate minimization of the regularized expected error over kernel models. *Mathematics of Operations Research*, *33*, 747–756.

Kůrková, V., & Sanguineti, M. (2008b). Geometric upper bounds on rates of variable-basis approximation. *IEEE Transactions on Information Theory*, *54*, 5681–5688.

Lavretsky, E. (2002). On the geometric convergence of neural approximations. *IEEE Transactions on Neural Networks*, *13*, 274–282.

Pinkus, A. (1985). *n-Widths in approximation theory*. Berlin, Heidelberg: Springer.

Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. In: Séminaire d'Analyse Fonctionnelle 1980–1981, vol. I, no. 12. École Polytechnique, Centre de Mathématiques, Palaiseau, France.

Zoppoli, R., Sanguineti, M., & Parisini, T. (2002). Approximating networks and extended Ritz method for the solution of functional optimization problems. *Journal of Optimization Theory and Applications*, *112*, 403–439.