ARTICLE IN PRESS

Neural Networks xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Neural Networks

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

Some comparisons of complexity in dictionary-based and linear computational models

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

Article history:
Received 23 May 2010
Received in revised form 5 October 2010
Accepted 9 October 2010

Keywords:
Linear approximation schemes
Variable-basis approximation schemes
Model complexity
Worst-case errors
Neural networks
Kernel models

ABSTRACT

Neural networks provide a more flexible approximation of functions than traditional linear regression. In the latter, one can only adjust the coefficients in linear combinations of fixed sets of functions, such as orthogonal polynomials or Hermite functions, while for neural networks, one may also adjust the parameters of the functions which are being combined. However, some useful properties of linear approximators (such as uniqueness, homogeneity, and continuity of best approximation operators) are not satisfied by neural networks. Moreover, optimization of parameters in neural networks becomes more difficult than in linear regression. Experimental results suggest that these drawbacks of neural networks are offset by substantially lower model complexity, allowing accuracy of approximation even in high-dimensional cases. We give some theoretical results comparing requirements on model complexity for two types of approximators, the traditional linear ones and so called variable-basis types, which include neural networks, radial, and kernel models. We compare upper bounds on worst-case errors in variable-basis approximation with lower bounds on such errors for any linear approximator. Using methods from nonlinear approximation and integral representations tailored to computational units, we describe some cases where neural networks outperform any linear approximator.

1. Introduction

In traditional *linear regression*, coefficients of linear combinations of given functions are searched for so that a desired functional relationship between inputs and outputs is sufficiently well approximated. Typically, a linear approximating scheme is formed by a nested family of sets, where the n-th set is generated by the first n elements of a given set of functions with a *fixed linear ordering* (e.g., a set of some orthogonal polynomials or Hermite functions of increasing degree).

In contrast, the simplest architecture of a connectionistic model is a *one-hidden-layer network with a single linear output*, in which in addition to the coefficients of linear combinations (called *output* weights), also *inner* parameters of computational units are optimized so that the entities being combined can be varied. The parameterized family of functions computable by network units is sometimes called a *dictionary* (Gribonval & Vandergheynst, 2006), it may contain finite, countably or uncountably infinite number of functions and has no fixed ordering. During learning, potentially all n-tuples of elements of the dictionary can be chosen together with the coefficients of their linear combinations. This computational model has been called a *variable-basis* approximation scheme (Kainen, Kůrková, & Sanguineti, 2009; Kůrková & Sanguineti, 2001, 2002, 2008). It includes perceptron neural networks, radial and kernel models, splines with free knots, trigonometric polynomials, etc.

In both models, linear and variable-basis, the number n of units can be interpreted as *model complexity*. Its growth with increasing accuracy of approximation can be estimated from inspection of bounds on rates of approximation.

Variable-basis models with units from various dictionaries have become a widespread tool for many classification, optimization, regression, and pattern recognition tasks (see e.g., Giulini & Sanguineti, 2000, 2009; Kecman, 2001; Kůrková & Sanguineti, 2005; Smith, 1999; Zoppoli, Sanguineti, & Parisini, 2002, and the references therein). In many high-dimensional tasks they obtained satisfactory good approximation with relatively small model complexity. The widespread utility of variable-basis models deserves theoretical treatment.

Clearly, approximation by a linear combination of n functions from a given dictionary, where both coefficients of the linear combination and the n-tuple of functions from the dictionary are optimally chosen, guarantees better accuracy than linear approximation using a fixed set of n elements from the same

* Corresponding author.
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 © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.10.002

dictionary. However, this does not exclude the possibility that a better rate might be achieved using the first n elements in another ordering of the same dictionary or the first n elements from another dictionary and ordering.

Thus one may wonder whether for some sets of functions of interest, approximation by variable-basis schemes with widely-used computational units (such as perceptrons, radial, and kernel units) provides faster rates than those achievable by *any linear approximator* (in particular, those induced by various ordered sets of polynomials). This question is especially interesting because using variable-basis approximation, one loses useful properties of linear approximators (such as uniqueness, homogeneity, and continuity of best approximation operators (Kainen, Kůrková, & Vogt, 2000a, 2000b, 2001)) and optimization of parameters becomes more difficult. So one hopes that there is a compensatory decrease in the requirements on model complexity.

Inspection of the proofs of estimates of rates of variable-basis approximation (Barron, 1993; Jones, 1992; Kůrková & Sanguineti, 2008) does not answer this question. For each function to be approximated, these proofs construct a special linear approximator. Indeed, in each step of such a construction, first a new unit is added to the previously chosen ones and then coefficients of a linear combination of all these units are recalculated. Such a proof technique can be interpreted as a construction of a linearly ordered sequence of units from the dictionary followed by an estimate of a rate of approximation by the first n elements from the dictionary with the constructed linear ordering. It should be emphasized that the linear ordering depends on the concrete function to be approximated and does not work for other functions. Moreover, algorithms based on such constructions may be inefficient; they depend on a specific representation of the approximated function as a convex combination of elements from the dictionary (Kůrková & Sanguineti, 2008).

Barron (1993) initiated a new approach to comparisons of model complexity of linear and variable-basis computational models. He proposed comparing the worst-case errors achievable by the best linear approximators (mathematically formalized by the concept of Kolmogorov's n-width (Kolmogorov, 1936)) with upper bounds on such errors in approximation by perceptron networks. Barron's estimates were extended by Kůrková and Sanguineti (2002) to sets of functions defined in terms of certain norms induced by computational units from more general dictionaries.

In this paper, we extend comparisons of worst-case errors to other dictionaries. For this goal, we develop new methods to estimate lower bounds on Kolmogorov's width based on properties of integral transforms induced by computational units. To obtain upper bounds on the worst-case errors in the variable-basis models with n computational units, we combine the upper bounds of the form $cn^{-1/2}$ by Barron (1993), Jones (1992) and Pisier (1981) with estimates that we recently obtained using integral transforms induced by computational units (Kainen & Kůrková, 2009; Kůrková, 2009, submitted for publication).

Two methods are used to derive lower bounds on theoretically optimal worst-case errors in approximation from linear subspaces. The first method provides lower bounds in terms of orthogonal subsets of large cardinality (so it utilizes geometrical properties) while the second method gives lower bounds in terms of s-numbers of integral operators induced by computational units. We describe sets of functions for which upper bounds on worst-case errors in variable-basis approximation are smaller than lower bounds on worst-case errors for any linear approximator. Such sets depend on the type of computational units and the volumes of the d-dimensional domains where the functions are defined. Some preliminary results appeared in a conference's proceedings (Gnecco, Kůrková, & Sanguineti, 2010).

Note that the requirement that worst-case errors in variable-basis approximation are smaller than lower bounds on worst-case errors by *any linear approximator* is rather strong. Even when this does not hold, approximation from some dictionary may be preferable to a theoretically better linear approximator, because finding such a linear approximator may be infeasible.

The paper is organized as follows. In Section 2, basic concepts of linear and variable-basis approximation and worst-case errors are introduced. In Section 3, upper bounds on variable-basis approximation are given in terms of certain norms induced by computational units. Section 4 is devoted to estimates of lower bounds on worst-case errors in linear approximation (n-widths). Section 5 compares lower bounds on n-widths of balls in norms induced by computational units with upper bounds on worst-case errors in variable-basis approximation. In Section 6, several examples illustrate our results. Section 7 is a brief discussion.

2. Linear and variable-basis approximation

A wide class of computational models (e.g., one-hidden-layer perceptron and radial and kernel networks) can be formally described as devices computing input–output functions from sets of the form

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

where the set of functions G is called a *dictionary* (Gribonval & Vandergheynst, 2006). The approximation by the family $\{\mathrm{span}_n G\}$ is referred to as *variable-basis approximation* (Kůrková & Sanguineti, 2001, 2002, 2008) 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 parameters.

For suitable choices of ϕ, $\mathrm{span}_n G_\phi$ models the 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., where the number n of computational units can be interpreted as the *model complexity*. For example, if $q = d + 1$ and

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

then the dictionary G_ϕ is formed by functions computable by *perceptrons* with an activation unit $\psi : \mathbb{R} \to \mathbb{R}$. 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* $\psi : \mathbb{R} \to \mathbb{R}_+$.

In contrast to variable-basis approximation, traditional *linear* models use as approximating families a nested set of the form

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

formed by linear combinations of the *first* n elements from some set $G = \{g_i \mid i \in \mathbb{N}_+\}$ with a *fixed linear ordering* (typically, some ordered set of polynomials).

By $(\mathcal{X}, \|\cdot\|_\mathcal{X})$ we denote a normed linear space and we write merely $\mathcal{X}$ when there is no ambiguity. In this paper, we shall deal with the Lebesgue spaces $\mathcal{L}^2_{\mu_\Omega}(\Omega)$ and $\mathcal{L}^1_{\mu_\Omega}(\Omega)$ endowed with the respective usual norms and $\Omega \subseteq \mathbb{R}^d$.

Please cite this article in press as: Gnecco, G., et al. Some comparisons of complexity in dictionary-based and linear computational models. Neural Networks (2010), doi:10.1016/j.neunet.2010.10.002

The error in approximation of a function $f \in \mathcal{X}$ by functions from a set A is measured by the distance

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

Approximation capabilities of whole 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 shorter 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 can be 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

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

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.,

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

We use the shorter notations when there is no ambiguity. If for some subspace the infimum is achieved, then the subspace is called *optimal*. 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.,

$$d_n(M) = d_n(\mathrm{cl}_{\mathcal{X}} M) \quad \text{and} \quad d_n(M) = d_n(\mathrm{conv} M), \tag{2}$$

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

For a dictionary $G \subset \mathcal{X}$, let $\mathcal{S}_n^G$ be the *set of all at most n-dimensional subspaces of* $\mathcal{X}$ *generated by n-tuples of elements of* G. Clearly, for every subset M of $\mathcal{X}$

$$\delta(M, \mathrm{span}_n G) \leq \inf_{\mathcal{X}_n \in \mathcal{S}_n^G} \delta(M, \mathcal{X}_n). \tag{3}$$

In other words, 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 variable-basis approximation by $\mathrm{span}_n G$. However, the inequality (3) 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

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

Description of cases when either the opposite inequality

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

holds for n greater than some n_0 or when for every $f \in M$ there exists some n_0 such that for every $n \geq n_0$ one has

$$\|f - \mathrm{span}_n G\|_{\mathcal{X}} < d_n(M) \tag{5}$$

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

The investigation of cases in which the inequality (4) holds was started by Barron (1993). He explored the dictionary

$$P^d(\sigma) := \{\sigma(\langle v, \cdot\rangle + b) \mid v \in \mathbb{R}^d,\ b \in \mathbb{R}\},$$

formed by functions computable by *perceptrons with a sigmoidal activation* σ. For $c > 0$, let

$$\Gamma_c^d := \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid \int_{\mathbb{R}^d} \|\omega\|_2 |\tilde{f}(\omega)| \mathrm{d}\omega \leq c \right\},$$

where $\tilde{f}$ denotes the Fourier transform of f and $\|\omega\|_2$ the ℓ_2-norm of $\omega \in \mathbb{R}^d$. Barron derived estimates of the n-width and the deviation from $\mathrm{span}_n P^d(\sigma)$ for the sets $\Gamma_c^d|_{[0,1]^d}$ and $\Gamma_c^d|_{B_1^d}$ made up of functions in Γ_c^d restricted to $[0, 1]^d$ and B_1^d, resp., where B_1^d is the unit ball in $\mathbb{R}^d$. In (Kůrková & Sanguineti, 2002), described properties of general dictionaries G in $\mathcal{L}^2([0, 1]^d)$, guaranteeing that certain sets of functions have n-widths larger than their deviations from $\mathrm{span}_n G$. The sets have the form of balls in norms induced by the dictionary G and are related to balls in norms defined by various smoothness conditions.

Note that the conditions (4) and (5) are rather strong as they state that worst-case errors in approximation by $\mathrm{span}_n G$ are smaller than such errors in approximation by *any* linear approximator. However, approximation by $\mathrm{span}_n G$ can be suitable even when theoretically a better linear approximator might exist as it might be difficult to find such an approximator.

3. Upper bounds for variable-basis approximation

To compare n-width with deviation from $\mathrm{span}_n G$, we take advantage of upper bounds on the latter, derived from the estimates of worst-case errors in approximation of functions from convex closures proven by Barron (1992, 1993), Jones (1992), Makovoz (1996) and Pisier (1981). We use reformulations of these estimates from Kůrková (2003), stated in terms of a norm induced by a dictionary G.

For every nonempty bounded subset G of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, its symmetric convex closure $\mathrm{cl}_{\mathcal{X}}\mathrm{conv}(G \cup -G)$ uniquely determines a norm for which it forms the unit ball. Such a norm is the Minkowski functional[1] of the set $\mathrm{cl}_{\mathcal{X}}\mathrm{conv}(G \cup -G)$. It is called *G-variation*, denoted by $\|\cdot\|_{G,\mathcal{X}}$ (shortly $\|\cdot\|_G$ when $\mathcal{X}$ is clear from the context), and defined as

$$\|f\|_{G,\mathcal{X}} = \|f\|_G := \inf\left\{c > 0 \mid c^{-1} f \in \mathrm{cl}_{\mathcal{X}}\mathrm{conv}(G \cup -G)\right\}.$$

Note that G-variation can be infinite and that it is a norm on the subspace of $\mathcal{X}$ formed by functions with finite G-variation. The general concept was introduced in Kůrková (1997), as an extension of variation with respect to sets of characteristic functions defined in Barron (1992).

The next proposition follows directly from the definitions.

Proposition 1. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space and G its bounded subset. Then for every positive integer n,*

(i) *for every $r > 0$,*
$r d_n(G) = d_n(B_r(\|\cdot\|_G))$;
(ii) *for every $M \subseteq \mathrm{cl}_{\mathcal{X}}\mathrm{conv}(G \cup -G)$,*
$d_n(G) \geq d_n(M)$.

[1] The *Minkowski functional* ν_G of a subset G of a linear space $\mathcal{X}$ is defined for every $f \in \mathcal{X}$ as $\nu_G(f) = \inf\{c \in \mathbb{R}_+ \mid f \in cG\}$ (Kolmogorov & Fomin, 1970, p. 131).

Please cite this article in press as: Gnecco, G., et al. Some comparisons of complexity in dictionary-based and linear computational models. Neural Networks (2010), doi:10.1016/j.neunet.2010.10.002

The next theorem is a reformulation, in terms of G-variation (Kůrková, 1997, 2003), of results by Barron (1993), Jones (1992) and Pisier (1981) and its extension by Makovoz (1996, Theorem 1). Recall that the *n-th entropy number* of a subset G of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is defined as

$$e_n(G) := \inf\left\{\varepsilon > 0 \mid \left(G \subseteq \bigcup_{i=1}^{n} U_i\right) \text{ and } (\forall i = 1, \ldots, n)\,(\operatorname{diam}(U_i) \leq \varepsilon)\right\},$$

where $\operatorname{diam}(U) = \sup_{x,y\in U} \|x - y\|_{\mathcal{X}}$.

Theorem 1. *Let G be a bounded subset of a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $s_G = \sup_{g\in G} \|g\|_{\mathcal{H}}$. Then for every $f \in \mathcal{H}$ and every positive integer n*

(i) $\|f - \operatorname{span}_n G\|_{\mathcal{H}} \leq \frac{s_G \|f\|_G}{\sqrt{n}}$;

(ii) $\|f - \operatorname{span}_n G\|_{\mathcal{H}} \leq \frac{\sqrt{2} s_G e_{\lfloor n/2 \rfloor}(G) \|f\|_G}{\sqrt{n}}$.

Examples for which the upper bound from Theorem 1(ii) is smaller than the one from Theorem 1(i) are given in Makovoz (1996). However, in general the bound from Theorem 1(ii) is more difficult to estimate than the one from Theorem 1(i).

As an immediate corollary of Theorem 1, we get the following upper bounds on deviations of balls in G-variation from $\operatorname{span}_n G$.

Corollary 1. *Let G be a bounded subset of a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $s_G = \sup_{g\in G} \|g\|_{\mathcal{H}}$. Then for every $r > 0$ and every positive integer n*

(i) $\delta(B_r(\|\cdot\|_G), \operatorname{span}_n G) \leq \frac{r s_G}{\sqrt{n}}$;

(ii) $\delta(B_r(\|\cdot\|_G), \operatorname{span}_n G) \leq \frac{\sqrt{2} s_G e_{\lfloor n/2 \rfloor}(G) r}{\sqrt{n}}$.

Combining upper bounds on $\delta(B_1(\|\cdot\|_G, \operatorname{span}_n G))$ from Corollary 1 with suitable lower bounds on $d_n(G)$ we can obtain some comparisons of linear and variable-basis approximation. We derive such lower bounds on $d_n(G)$ in the next section.

For a finite dictionary $G = \{g_1, \ldots, g_m\}$ and a function f representable as a linear combination of elements of G, it is easy to show that $\|f\|_G$ is the minimum of the ℓ_1-norms $\|w\|_1$ of the weight vectors $w \in \mathbb{R}^m$ for which $f = \sum_{i=1}^{m} w_i g_i$ (Kůrková, Savický, & Hlaváčková, 1998, Proposition 2.3, p. 653).

Proposition 2. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space and $G = \{g_1, \ldots, g_m\} \subset \mathcal{X}$. Then for every $f \in \operatorname{span} G$,*

$$\|f\|_G = \min\left\{\|w\|_1 \mid f = \sum_{i=1}^{m} w_i g_i,\ w \in \mathbb{R}^m\right\}.$$

Several authors investigated an analogous relationship between G-variation and $\mathcal{L}^1$-norm for infinite dictionaries (Barron, 1992; Girosi & Anzellotti, 1993; Gnecco & Sanguineti, in press; Jones, 1992; Kainen & Kůrková, 2009; Kainen, Kůrková, & Vogt, 2007; Kůrková, Kainen, & Kreinovich, 1997). Under various assumptions on the computational unit ϕ, the set of parameters Y, and the ambient function space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, they proved that for a function f represented as

$$f(x) = \int_Y w(y)\phi(x, y)\mathrm{d}\mu_Y(y), \tag{6}$$

the estimate

$$\|f\|_{G_\phi} \leq \|w\|_{\mathcal{L}^1} \tag{7}$$

holds. These results were derived using a variety of proof techniques, some of them quite sophisticated (e.g., properties of the Bochner integral (Kainen & Kůrková, 2009), characterization of G-variation in terms of linear functionals (Kůrková, 2009)). In Kůrková (2009, Theorem 3, p. 714), the estimate (7) was derived for a wide class of function spaces under minimal assumptions needed for its formulation: G_ϕ is a bounded subset of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $w \in \mathcal{L}^1_{\mu_Y}(Y)$.

In this paper, we only need the estimate (7) for functions f in $\mathcal{L}^2$-spaces having the representation (6) with finite measure μ_Y on the set of parameters Y. For this special case, we give an alternative probabilistic argument that extends an idea from Barron (1993, Theorem 2, p. 934). Recall that a measure μ_Ω on $\Omega \subseteq \mathbb{R}^d$ is *σ-finite* if and only if there exists a countable collection of measurable sets $M_i \subseteq \Omega$ such that $\Omega = \cup_{i=1}^{\infty} M_i$ and $\mu_\Omega(M_i) < \infty$ for every $i \in \mathbb{N}_+$.

Theorem 2. *Let $\Omega \subseteq \mathbb{R}^d$, μ_Ω be a σ-finite measure on Ω, $Y \subseteq \mathbb{R}^q$, μ_Y a finite measure on Y, $\phi : \Omega \times Y \to \mathbb{R}$ such that $G_\phi(Y) = \{\phi(\cdot, y) \mid y \in Y\}$ is a bounded subset of $\mathcal{L}^2_{\mu_\Omega}(\Omega)$, and $w \in \mathcal{L}^1_{\mu_Y}(Y)$. Then for every $f \in \mathcal{L}^2_{\mu_\Omega}(\Omega)$ that can be represented for every $x \in \Omega$ as $f(x) = \int_Y w(y)\phi(x, y)\mathrm{d}\mu_Y(y)$ we have*

$$\|f\|_{G_\phi(Y), \mathcal{L}^2_{\mu_\Omega}(\Omega)} \leq \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}.$$

Proof. If $\|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} = 0$, then $f = 0$ and the statement follows. If $\|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} > 0$, let $\mathrm{d}\rho_Y = \frac{|w|}{\|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}}\mathrm{d}\mu_Y$. Then

$$\begin{aligned} f(x) &= \int_Y w(y)\phi(x, y)\mathrm{d}\mu_Y(y) \\ &= \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} \int_Y \phi(x, y)\operatorname{sign} w(y) \frac{|w(y)|}{\|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}}\mathrm{d}\mu_Y(y) \\ &= \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} \int_Y \phi(x, y)\operatorname{sign} w(y)\mathrm{d}\rho_Y(y). \end{aligned}$$

Let y and $y_1, \ldots, y_n$ be i.i.d. real random variables, distributed according to the probability measure ρ_Y. For a function $h(y_1, \ldots, y_n)$, we denote by $E_{y_1,\ldots,y_n}\{h(y_1, \ldots, y_n)\}$ its expected value.

As G_ϕ is bounded, $s_{G_\phi} = \sup_{y\in Y} \|\phi(\cdot, y)\|_{\mathcal{L}^2_{\mu_\Omega}(\Omega)} < \infty$. For $i = 1, \ldots, n$, let

$$a_i(x) := \frac{f(x) - \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}\phi(x, y_i)\operatorname{sign} w(y_i)}{n}.$$

Then

$$\begin{aligned} \|a_i\|_{\mathcal{L}^2_{\mu_\Omega}(\Omega)} &\leq \frac{\|f\|_{\mathcal{L}^2_{\mu_\Omega}(\Omega)} + \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}\|\phi(\cdot, y_i)\|_{\mathcal{L}^2_{\mu_\Omega}(\Omega)}}{n} \\ &\leq \frac{\|f\|_{\mathcal{L}^2_{\mu_\Omega}(\Omega)} + \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} s_{G_\phi}}{n}. \end{aligned}$$

So we get

$$\begin{aligned} &E_{y_1,\ldots,y_n}\left\{\int_\Omega \Big(f(x) - \frac{\|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}}{n} \right. \\ &\qquad \left. \times \sum_{i=1}^{n} \phi(x, y_i)\operatorname{sign} w(y_i)\Big)^2 \mathrm{d}\mu_\Omega(x)\right\} \\ &= E_{y_1,\ldots,y_n}\left\{\int_\Omega \sum_{i,k=1}^{n} a_i(x)a_k(x)\mathrm{d}\mu_\Omega(x)\right\} \end{aligned}$$

$$= \sum_{i=1}^{n} \left(E_{y_i} \left\{ \int_{\Omega} a_i^2(x) \mathrm{d}\mu_{\Omega}(x) \right\} + \sum_{k \neq i} \int_{\Omega} E_{y_i, y_k} \{ a_i(x) a_k(x) \} \, \mathrm{d}\mu_{\Omega}(x) \right)$$

$$\leq n \frac{\left(\|f\|_{\mathcal{L}^2_{\mu_{\Omega}}(\Omega)} + \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} s_{G_\phi} \right)^2}{n^2}, \tag{8}$$

where (8) follows by Fubini's theorem (Rudin, 1987, Theorem 8.8, p. 164) (which can be applied as both μ_{Ω} and ρ_Y are σ-finite, $E_{y_i}\left\{\int_{\Omega} a_i^2(x)\mathrm{d}\mu_{\Omega}(x)\right\} < \infty$ for $i = 1, \ldots, n$, and $\int_{\Omega} E_{y_i,y_k}\{|a_i(x)a_k(x)|\}\,\mathrm{d}\mu_{\Omega}(x) < \infty$ for $k \neq i$) and (8) follows by the independence of the random variables $y_1, \ldots, y_n$ and the identity $E_{y_i}\{a_i(x)\} = 0$ for every $x \in \Omega$. So, (8) implies the existence of $\hat{y}_1, \ldots, \hat{y}_n \in Y$ such that

$$\left\| f(\cdot) - \frac{\|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}}{n} \sum_{i=1}^{n} \phi(\cdot, \hat{y}_i) \operatorname{sign} w(\hat{y}_i) \right\|^2_{\mathcal{L}^2_{\mu_{\Omega}}(\Omega)}$$

$$\leq \frac{\left(\|f\|_{\mathcal{L}^2_{\mu_{\Omega}}(\Omega)} + \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} s_{G_\phi} \right)^2}{n}.$$

Hence $\|w\|^{-1}_{\mathcal{L}^1_{\mu_Y}(Y)} f \in \operatorname{cl}\operatorname{conv}(G_\phi \cup -G_\phi)$ and so by the definition of variation $\|f\|_{G_\phi} \leq \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)}$. □

4. Lower bounds on Kolmogorov width

In this section, we illustrate two methods of derivation of lower bounds on the Kolmogorov n-width. These bounds will be used in the next section to compare linear and variable-basis approximation.

The first method can be applied when balls in the G-variation contain "sufficiently large" orthonormal subsets. It is based on the following theorem from Kůrková and Sanguineti (2002, p. 270).

Theorem 3. *Let A and G be subsets of a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, G bounded, A is finite orthonormal with $\operatorname{card} A = k$, and $c_{A,G} := \max_{f \in A} \|f\|_G$. Then for every positive integer $n \leq k$,*

$$d_n(G) = d_n(B_1(\|\cdot\|_G)) \geq \frac{1}{c_{A,G}} \sqrt{1 - \frac{n}{k}}.$$

The second method is also based on estimates of distance from orthonormal sets formed by eigenfunctions of compact self-adjoint operators. It provides a characterization of n-widths of sets of functions "large enough" to contain images of unit balls mapped by compact operators. The characterization is in terms of the singular numbers of these operators. Recall that a linear operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ between two Hilbert spaces $(\mathcal{H}_1, \|\cdot\|_{\mathcal{H}_1})$ and $(\mathcal{H}_2, \|\cdot\|_{\mathcal{H}_2})$ is *compact* if the image under T of every bounded set in $\mathcal{H}_1$ is a precompact subset of $\mathcal{H}_2$ (i.e., a set whose closure in the topology induced by $\|\cdot\|_{\mathcal{H}_2}$ is compact). The *adjoint* of T is the unique operator T^* satisfying for every $f \in \mathcal{H}_1$ and every $g \in \mathcal{H}_2$, $\langle f, T^* g \rangle_{\mathcal{H}_1} = \langle Tf, g \rangle_{\mathcal{H}_2}$. The operator T is *self-adjoint* if $T = T^*$. For a compact operator T between two Hilbert spaces, its n-th *s-number* is defined as

$$s_n(T) = \sqrt{\lambda_n(TT^*)},$$

where $\lambda_n(TT^*)$ is the n-th *eigenvalue* of the self-adjoint, non-negative, and compact operator TT^* (the eigenvalues are ordered in a non-increasing sequence counting their multiplicities). If T is self-adjoint, then its singular numbers are equal to the absolute values of its eigenvalues.

The following theorem from Pinkus (1985, p. 65) states the equality between the n-width of the image of the unit ball under a compact operator T and the $(n+1)$-th singular number of T.

Theorem 4. *Let $(\mathcal{H}_1, \|\cdot\|_{\mathcal{H}_1})$ and $(\mathcal{H}_2, \|\cdot\|_{\mathcal{H}_2})$ be Hilbert spaces and $T : \mathcal{H}_1 \to \mathcal{H}_2$ a compact linear operator. Then for every positive integer $n < \dim \mathcal{H}_2$,*

$$d_n(T(B_1(\|\cdot\|_{\mathcal{H}_1})); (\mathcal{H}_2, \|\cdot\|_{\mathcal{H}_2})) = s_{n+1}(T).$$

We apply Theorems 3 and 4 to operators from two classes. The first one contains operators induced by finite dictionaries. Let $(\mathbb{R}^m, \|\cdot\|_2)$ denote the m-dimensional Euclidean space with the ℓ_2-norm $\|\cdot\|_2$. For every fixed ordering of a finite dictionary $G = \{g_1, \ldots, g_m\} \subset \mathcal{X}$, let $T_{g_1,\ldots,g_m} : \mathbb{R}^m \to \mathcal{X}$ be the linear operator defined for every $w = (w_1, \ldots, w_m) \in \mathbb{R}^m$ as

$$T_{g_1,\ldots,g_m}(w) := \sum_{i=1}^{m} w_i g_i. \tag{9}$$

The next proposition states compactness (hence continuity) of $T_{g_1,\ldots,g_m}$.

Proposition 3. *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space and $G = \{g_1, \ldots, g_m\}$ its finite subset. Then the operator $T_{g_1,\ldots,g_m} : (\mathbb{R}^m, \|\cdot\|_2) \to (\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is compact.*

Proof. The norm of $T_{g_1,\ldots,g_m}$ satisfies

$$\|T_{g_1,\ldots,g_m}\| \leq \sqrt{m} \max_{i=1\ldots,m} \|g_i\|_{\mathcal{X}},$$

so the operator is bounded. This is equivalent to its continuity (Friedman, 1982, Theorem 4.4.2). Moreover, $T_{g_1,\ldots,g_m}$ has a finite-dimensional range and thus it is compact (Oden & Demkowicz, 1996, Section 5.15). □ Q2

For a finite subset $G = \{g_1, \ldots, g_m\}$ of a Hilbert space $\mathcal{H}$, we denote by $M(G)$ the positive-semidefinite $m \times m$ matrix formed by the inner products of elements of G, i.e.,.

$$M(G)_{ij} = \langle g_i, g_j \rangle_{\mathcal{H}}$$

and by $\lambda_n(M(G))$ its n-th eigenvalue (ordered non-increasingly and counting multiplicities).

Proposition 4. *Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a Hilbert space and $G = \{g_1, \ldots, g_m\}$ its finite subset. Then for every positive integer $n < m$,*

(i) $d_n(G) = d_n(B_1(\|\cdot\|_G)) \geq \sqrt{\dfrac{\lambda_{n+1}(M(G))}{m}}$;

(ii) *for G orthonormal* $d_n(G) \geq \dfrac{1}{\sqrt{m}}$.

Proof. (i) It is easy to check that the adjoint $T^*_{g_1,\ldots,g_m}$ satisfies $(T^*_{g_1,\ldots,g_m}(f))_i = \langle f, g_i \rangle_{\mathcal{H}}$. Thus $T^*_{g_1,\ldots,g_m} T_{g_1,\ldots,g_m} : \mathbb{R}^m \to \mathbb{R}^m$ can be represented by the matrix $M(G)$. As every operator with a finite range is compact (Friedman, 1982, p. 188), $T^*_{g_1,\ldots,g_m}$ is compact. By Proposition 3, $T_{g_1,\ldots,g_m}$ is compact, too. Moreover, the two operators have the same positive singular numbers (Weidmann, 1980, p. 170). Hence $s_{n+1}(T_{g_1,\ldots,g_m}) = \sqrt{\lambda_{n+1}(M(G))}$.

By Proposition 1, $d_n(G) = d_n(B_1(\|\cdot\|_G))$. By Proposition 2 and the Cauchy–Schwarz inequality,

$$B_1(\|\cdot\|_G) \supseteq T_{g_1,\ldots,g_m}(B_1(\|\cdot\|_1)) \supseteq T_{g_1,\ldots,g_m}(B_{1/\sqrt{m}}(\|\cdot\|_2)) = \frac{1}{\sqrt{m}} T_{g_1,\ldots,g_m}(B_1(\|\cdot\|_2)).$$

So we can apply Theorem 4 to obtain

$$d_n(B_1(\|\cdot\|_G)) \geq \frac{1}{\sqrt{m}} d_n(T_{g_1,\ldots,g_m}(B_1(\|\cdot\|_2))) = \sqrt{\frac{\lambda_{n+1}(M(G))}{m}}.$$

(ii) When G is orthonormal, $T^*_{g_1,\ldots,g_m} T_{g_1,\ldots,g_m}$ is represented by the identity matrix, so all its eigenvalues are equal to 1. □

The second class of operators to which we apply Theorems 3 and 4 is represented by integral operators with kernels ϕ corresponding to computational units. Let $\Omega \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^q$ with a measure μ_Y. For a function $\phi : \Omega \times Y \to \mathbb{R}$ and function spaces $\mathcal{F}(\Omega)$ and $\mathcal{F}(Y)$ such that the integral on the right-hand side of (10) exists for every $x \in \Omega$, an operator $T_\phi = T_{\phi,\mu_Y} : \mathcal{F}(Y) \to \mathcal{F}(\Omega)$ is defined as

$$T_\phi(w)(x) := \int_Y w(y)\phi(x,y)\mathrm{d}\mu_Y(y). \tag{10}$$

Note that $T_\phi(w)$ can be interpreted as an *input–output function of a one-hidden-layer network with infinitely many units* computing functions $\phi(\cdot, y)$ and output weights $w(y)$, for every $y \in Y$.

To describe properties of operators T_ϕ, we recall some definitions. A bounded linear operator $T : (\mathcal{H}_1, \|\cdot\|_{\mathcal{H}_1}) \to (\mathcal{H}_2, \|\cdot\|_{\mathcal{H}_2})$ between two Hilbert spaces is called a *Hilbert–Schmidt* operator if and only if for every orthonormal basis $\{\psi_\alpha \mid \alpha \in \mathcal{A}\}$ of $\mathcal{H}_1$ one has $\sum_{\alpha\in\mathcal{A}} \|T\psi_\alpha\|^2_{\mathcal{H}_2} < \infty$ (Shubin, 2001, p. 257). The value $\|T\|_{HS} = \sqrt{\sum_{\alpha\in\mathcal{A}} \|T\psi_\alpha\|^2_{\mathcal{H}_2}}$ is independent of the choice of the orthonormal basis and is called the *Hilbert–Schmidt norm* of the operator T.

The next theorem summarizes well-known properties of the operators T_ϕ on $\mathcal{L}^2$-spaces (see Shubin, 2001, Proposition A.3.1, p. 257 and Proposition A.3.2, p. 259; Weidmann, 1980, Theorem 6.11, p. 139; Akhiezer & Glazman, 1993, p. 127).

Theorem 5. *Let $\Omega \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}^q$, μ_Ω and μ_Y be σ-finite measures on Ω and Y, resp., and $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_Y}(\Omega \times Y)$. Then*

(i) *T_ϕ maps $\mathcal{L}^2_{\mu_Y}(Y)$ to $\mathcal{L}^2_{\mu_\Omega}(\Omega)$;*
(ii) *T_ϕ is a Hilbert–Schmidt operator;*
(iii) *T_ϕ is compact;*
(iv) *if $\Omega = Y$, $\mu_\Omega = \mu_Y$ and ϕ is symmetric, then T_ϕ is self-adjoint, its non zero eigenvalues form a finite or countably infinite set of reals $\{\lambda_j\}$ satisfying $\sum_{j=1}^N \lambda_j^2 < \infty$, where N is finite or $N = +\infty$, and there exists an orthonormal family $\{\psi_j\}$ in $\mathcal{L}^2_{\mu_\Omega}(\Omega)$ of eigenfunctions such that for every $x, y \in \Omega$ one has $\phi(x,y) = \sum_{j=1}^N \lambda_j\psi_j(x)\psi_j(y)$, where for $N = +\infty$ the series converges in $\mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}(\Omega \times \Omega)$ and for every $g \in \mathcal{L}^2_{\mu_\Omega}(\Omega)$,*

$$T_\phi(g) = \sum_{j=1}^N \lambda_j \langle g, \psi_j\rangle_{\mathcal{L}^2_{\mu_\Omega}} \psi_j. \tag{11}$$

Note that if $\lambda_j, j = 1, \ldots, N$ are positive and the series converges uniformly, then ϕ is positive semidefinite.

Applying Theorem 5 to T_ϕ, we get the following lower bound on the Kolmogorov width of balls in G_ϕ-variation with $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_Y}(\Omega \times Y)$, where $\mu_Y(Y)$ is finite.

Theorem 6. *Let $\Omega \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}^q$, μ_Ω be a σ-finite measure on Ω, μ_Y a finite measure on Y, $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_Y}(\Omega \times Y)$, and $G_\phi = \{\phi(\cdot, y) \mid y \in Y\}$ a bounded subset of $\mathcal{L}^2_{\mu_\Omega}(\Omega)$. Then T_ϕ is a compact operator mapping $\mathcal{L}^2_{\mu_Y}(Y)$ to $\mathcal{L}^2_{\mu_\Omega}(\Omega)$ and for every positive integer n,*

$$d_n(B_1(\|\cdot\|_{G_\phi})) = d_n(G_\phi) \geq \frac{s_{n+1}(T_\phi)}{\sqrt{\mu_Y(Y)}}.$$

Proof. By Theorem 5(i) and (iii), $T_\phi : \mathcal{L}^2_{\mu_Y}(Y) \to \mathcal{L}^2_{\mu_\Omega}(\Omega)$ is compact, so by Theorem 4 we get $d_n(T_\phi(B_1(\|\cdot\|_{\mathcal{L}^2_{\mu_Y}}))) = s_{n+1}(T_\phi)$.

As μ_Y is finite, every $w \in \mathcal{L}^2_{\mu_Y}(Y)$ is also in $\mathcal{L}^1_{\mu_Y}(Y)$ and $\|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} \leq \sqrt{\mu_Y(Y)}\|w\|_{\mathcal{L}^2_{\mu_Y}(Y)}$. By Theorem 2 with $f = T_\phi(w)$, we get

$$\|f\|_{G_\phi} \leq \|w\|_{\mathcal{L}^1_{\mu_Y}(Y)} \leq \sqrt{\mu_Y(Y)}\|w\|_{\mathcal{L}^2_{\mu_Y}(Y)}.$$

Thus $B_1(\|\cdot\|_{G_\phi}) \supseteq \frac{1}{\sqrt{\mu_Y(Y)}} T_\phi(B_1(\|\cdot\|_{\mathcal{L}^2}))$ and the statement follows. □

When ϕ is symmetric, Theorem 6 implies on the Kolmogorov width of the ball $B_1(\|\cdot\|_{G_\phi})$ the following lower bound in terms of the eigenvalues of the operator T_ϕ.

Corollary 2. *Let $\Omega \subseteq \mathbb{R}^d$, μ_Ω be a finite measure on Ω, $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}(\Omega \times \Omega)$ symmetric such that G_ϕ is bounded in $\mathcal{L}^2_{\mu_\Omega}(\Omega)$, and $\{\lambda_j\}$ a sequence of eigenvalues of T_ϕ ordered non-increasingly in absolute values. Then for every positive integer n,*

$$d_n(B_1(\|\cdot\|_{G_\phi})) = d_n(G_\phi) \geq \frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}}.$$

Proof. By Theorem 6, $d_n(B_1(\|\cdot\|_{G_\phi})) \geq \frac{s_{n+1}(T_\phi)}{\sqrt{\mu_\Omega(\Omega)}}$. As ϕ is symmetric, T_ϕ is self-adjoint. Hence $s_{n+1}(T_\phi) = |\lambda_{n+1}|$ and the statement follows. □

Using as a lower bound on Kolmogorov n-width the estimate from Theorem 3 instead of the one from Theorem 4, we get the next theorem.

Theorem 7. *Let $\Omega \subseteq \mathbb{R}^d$, μ_Ω be a finite measure on Ω, and $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}(\Omega \times \Omega)$ symmetric such that G_ϕ is bounded in $\mathcal{L}^2_{\mu_\Omega}(\Omega)$. Then there exists an orthonormal set of eigenfunctions $\{\psi_j\}$ of $T_\phi : \mathcal{L}^2_{\mu_\Omega}(\Omega) \to \mathcal{L}^2_{\mu_\Omega}(\Omega)$ and a sequence $\{\lambda_j\}$ of corresponding eigenvalues ordered non-increasingly in absolute values, such that for all positive integers n, k with $n \leq k$ and $c_{k,G_\phi} = \max_{j=1,\ldots,k}\|\psi_j\|_{G_\phi}$,*

$$d_n(B_1(\|\cdot\|_{G_\phi})) = d_n(G_\phi) \geq \frac{1}{c_{k,G_\phi}}\sqrt{1-\frac{n}{k}} \geq \frac{|\lambda_k|}{\sqrt{\mu_\Omega(\Omega)}}\sqrt{1-\frac{n}{k}}.$$

Proof. The existence of eigenfunctions and eigenvalues of T_ϕ follows by Theorem 5. Applying Theorem 3 to $A = \{\psi_1, \ldots, \psi_k\}$, we get $d_n(B_1(\|\cdot\|_{G_\phi})) \geq \frac{1}{c_{k,G_\phi}}\sqrt{1-\frac{n}{k}}$.

As $\{\psi_j\}$ are eigenfunctions of T_ϕ, we have $\lambda_j\psi_j(x) = \int_\Omega \psi_j(y)\phi(x,y)\mathrm{d}\mu_\Omega(y)$. By the Cauchy–Schwarz inequality we get $\|\psi_j\|_{\mathcal{L}^1_{\mu_\Omega}} \leq \sqrt{\mu_\Omega(\Omega)}\|\psi_j\|_{\mathcal{L}^2_{\mu_\Omega}}$, hence $\psi_j \in \mathcal{L}^1_{\mu_\Omega}(\Omega)$. As G_ϕ is bounded and μ_Ω is finite, the assumptions of Theorem 2 are satisfied and so

$$\|\psi_j\|_{G_\phi} \leq \frac{1}{|\lambda_j|}\|\psi_j\|_{\mathcal{L}^1_{\mu_\Omega}} \leq \frac{\sqrt{\mu_\Omega(\Omega)}}{|\lambda_j|}\|\psi_j\|_{\mathcal{L}^2_{\mu_\Omega}} = \frac{\sqrt{\mu_\Omega(\Omega)}}{|\lambda_j|}.$$

Thus for every positive integer k we have

$$c_{k,G_\phi} = \max_{j=1,\ldots,k}\|\psi_j\|_{G_\phi} \leq \frac{\sqrt{\mu_\Omega(\Omega)}}{|\lambda_k|}$$

and so by Theorem 3

$$d_n(B_1(\|\cdot\|_{G_\phi})) = d_n(G_\phi) \geq \frac{|\lambda_k|}{\sqrt{\mu_\Omega(\Omega)}}\sqrt{1-\frac{n}{k}}.$$ □

Note that for $k = n+1$, the lower bound

$$\frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}}\sqrt{1-\frac{n}{n+1}} = \frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}}\sqrt{\frac{1}{n+1}}$$

from Theorem 7 is smaller than the lower bound $\frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}}$ from Corollary 2. Thus the proof method based on singular numbers gives better results than the method based on combining the inclusion of orthonormal subsets of eigenfunctions with estimates of variational norms of these eigenfunctions derived from Theorem 2. However, in cases where better estimates of G_ϕ-variations of eigenfunctions can be obtained, Theorem 7 may provide larger lower bounds than the ones from Corollary 2. In Section 6.1, we shall give an example of such a case. In other cases, even larger lower bounds can be obtained when different orthonormal sets than those formed by eigenfunctions are used (Kůrková & Sanguineti, 2002).

Note that the lower bounds on the n-width depend on the number d of variables, although we have not emphasized this in the notation. When $\mu_\Omega = \mu_d$ is the d-dimensional Lebesgue measure, then the term $\mu_d(\Omega_d)$ may either grow to infinity exponentially fast (e.g., when $\Omega_d = [-1,1]^d$), or go to zero exponentially fast (e.g., when Ω_d is the unit Euclidean ball), or be constant (e.g., when $\Omega_d = [0,1]^d$). In particular, when the domain is the unit d-dimensional Euclidean ball (whose volume is $\pi^{d/2}/\Gamma((d+2)/2)$, see Courant (1988, p. 304)), our estimates of the n-width may give large values for those n for which λ_n is considerably smaller than $\sqrt{\pi^{d/2}/\Gamma((d+2)/2)}$, i.e. when the ratio

$$\frac{|\lambda_{n+1}|}{\sqrt{\pi^{d/2}/\Gamma((d+2)/2)}}$$

is large.

5. Comparisons of worst-case errors

In this section, we compare the upper bounds on deviation from $\mathrm{span}_n G$ derived in Section 3 with the lower bounds on n-width from Section 4. For a subset M of $\mathcal{L}^2_{\mu_\Omega}(\Omega)$, we denote by

$$\Delta_n(M) := d_n(M) - \delta(M, \mathrm{span}_n G)$$

the difference between its n-width and its deviation from $\mathrm{span}_n G$. When $\Delta_n(M)$ is positive, the worst-case error in approximation of M by $\mathrm{span}_n G$ is smaller than the worst-case errors in its approximation by *any* linear approximator.

As $d_n(G) = d_n(B_1(\|\cdot\|_G))$, the same worst-case error (or nearly worst-case error when the infimum is not a minimum) as the one in approximation of functions from the ball $B_1(\|\cdot\|_G)$ must be achieved by some function in G. Moreover, $\delta(G, \mathrm{span}_n G) = 0$ and so

$$\Delta_n(G) = d_n(G) = d_n(B_1(\|\cdot\|_G)) \geq 0.$$

Hence by Theorems 1, 6 and 7, and Corollary 2 we get the following estimates.

Corollary 3. *Let $\Omega \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}^q$, μ_Ω and μ_Y be σ-finite measures on Ω and Y, resp., $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_Y}(\Omega\times Y)$ such that $G_\phi = \{\phi(\cdot,y) \mid y \in Y\}$ is a bounded subset of $\mathcal{L}^2_{\mu_\Omega}(\Omega)$, and $s_{G_\phi} = \sup_{y\in Y}\|\phi(\cdot,y)\|_{\mathcal{L}^2_{\mu_\Omega}}$. Then there exists an orthonormal set of eigenfunctions $\{\psi_j\}$ and a sequence $\{\hat{\lambda}_j\}$ of corresponding eigenvalues of T_ϕ, ordered non-increasingly in absolute values, such that for every positive integer n,*

$$\text{(i) } \Delta_n(G_\phi) \geq \frac{s_{n+1}(T_\phi)}{\sqrt{\mu_Y(Y)}}.$$

For the case in which $\Omega = Y$, $\mu_\Omega = \mu_Y$, and ϕ is symmetric, for every $k \geq n$ and $c_{k,G_\phi} = \max_{j=1,\ldots,k}\|\psi_j\|_{G_\phi}$ we get

$$\text{(ii) } \Delta_n(G_\phi) = d_n(G_\phi) \geq \frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}};$$

$$\text{(iii) } \Delta_n(G_\phi) = d_n(G_\phi) \geq \frac{1}{c_{k,G_\phi}}\sqrt{1-\frac{n}{k}}.$$

Corollary 3(ii) and (iii) show that for every $\varepsilon > 0$, every linear approximator, and every computational unit defined by a symmetric function $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}(\Omega\times\Omega)$ for which the set G_ϕ is bounded, there exists a parameter y such that $\phi(\cdot,y)$ in linear approximation has an error larger than $\frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}} - \varepsilon$ or $\frac{1}{c_{k,G_\phi}}\sqrt{1-\frac{n}{k}} - \varepsilon$, resp. When for some n these values are large, such computational units cannot be efficiently approximated by n-dimensional subspaces.

As discussed at the end of Section 4, the choice of the d-dimensional domain Ω may provide substantially different behaviors of the estimates with respect to d. When the domain is the unit ball in d dimensions, for instance, the lower bound from Corollary 3(ii) grows exponentially fast with d, while for the cube $[-1,1]^d$ it converges exponentially fast to zero.

Theorems 1, 3, and Proposition 4 provide the following estimates of $\Delta_n(B_1(\|\cdot\|_G))$ for finite dictionaries G.

Corollary 4. *Let A and G be finite subsets of a Hilbert space $(\mathcal{H}, \|\cdot\|_\mathcal{H})$ such that $\mathrm{card}\ G = m$, $s_G = \sup_{g\in G}\|g\|_\mathcal{H}$, $\mathrm{card}\ A = k$, A orthonormal with $\max_{f\in A}\|f\|_G = c_{A,G}$, and let $M(G)$ be the $m\times m$ matrix defined as $M(G)_{ij} = \langle g_i, g_j\rangle_\mathcal{H}$. Then for every positive integer $n \leq k$,*

$$\text{(i) } \Delta_n(B_1(\|\cdot\|_G)) \geq \frac{1}{c_{A,G}}\sqrt{1-\frac{n}{k}} - \frac{s_G}{\sqrt{n}};$$

$$\text{(ii) } \Delta_n(B_1(\|\cdot\|_G)) \geq \frac{1}{c_{A,G}}\sqrt{1-\frac{n}{k}} - \frac{\sqrt{2}s_G e_{\lfloor n/2\rfloor}(G)}{\sqrt{n}};$$

and for every positive integer $n < m$,

$$\text{(iii) } \Delta_n(B_1(\|\cdot\|_G)) \geq \sqrt{\frac{|\lambda_{n+1}(M(G))|}{m}} - \frac{\sqrt{2}s_G e_{\lfloor n/2\rfloor}(G)}{\sqrt{n}}.$$

Corollaries 1 and 2 and Theorem 7 provide for infinite dictionaries G_ϕ the following estimates.

Corollary 5. *Let $\Omega \subseteq \mathbb{R}^d$, μ_Ω be a finite measure on Ω, $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}(\Omega\times\Omega)$ symmetric such that G_ϕ is a bounded subset of $\mathcal{L}^2_{\mu_\Omega}(\Omega)$, and $s_{G_\phi} = \sup_{y\in\Omega}\|\phi(\cdot,y)\|_{\mathcal{L}^2_{\mu_\Omega}}$. Then there exists an orthonormal set of eigenfunctions $\{\psi_j\}$ of T_ϕ and a sequence $\{\hat{\lambda}_j\}$ of corresponding eigenvalues of T_ϕ, ordered non-increasingly in absolute values, such that for all positive integers n, k with $n \leq k$ and $c_{k,G_\phi} = \max_{j=1,\ldots,k}\|\psi_j\|_{G_\phi}$,*

$$\text{(i) } \Delta_n(B_1(\|\cdot\|_{G_\phi})) \geq \frac{1}{c_{k,G_\phi}}\sqrt{1-\frac{n}{k}} - \frac{s_{G_\phi}}{\sqrt{n}};$$

$$\text{(ii) } \Delta_n(B_1(\|\cdot\|_{G_\phi})) \geq \frac{1}{c_{k,G_\phi}}\sqrt{1-\frac{n}{k}} - \frac{\sqrt{2}s_{G_\phi} e_{\lfloor n/2\rfloor}(G_\phi)}{\sqrt{n}};$$

$$\text{(iii) } \Delta_n(B_1(\|\cdot\|_{G_\phi})) \geq \frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}} - \frac{\sqrt{2}s_{G_\phi} e_{\lfloor n/2\rfloor}(G_\phi)}{\sqrt{n}}.$$

Note that to compare the n-width with the deviation from $\mathrm{span}_n G$, in Corollaries 4(iii) and 5(iii) we have used the upper bound in Corollary 1(ii) instead of the weaker one in Corollary 1(i).

The next proposition states a relationship between s_{G_ϕ} and the eigenvalues of a Hilbert–Schmidt operator which implies that the weaker upper bound from Corollary 1(i) cannot give for every Hilbert–Schmidt operator T_ϕ a positive value for $\Delta_n(B_1(\|\cdot\|_{G_\phi}))$.

Proposition 5. *Let $\Omega \subseteq \mathbb{R}^d$, μ_Ω be a finite measure on Ω, $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}(\Omega\times\Omega)$ symmetric, $G_\phi = \{\phi(\cdot, y) \mid y \in \Omega\}$ a bounded subset of $\mathcal{L}^2_{\mu_\Omega}(\Omega)$, $s_{G_\phi} = \sup_{y\in\Omega}\|\phi(\cdot,y)\|_{\mathcal{L}^2_{\mu_\Omega}}$, and $\{\lambda_j\}$ eigenvalues of T_ϕ, ordered non-increasingly in absolute values. Then for every positive integer n,*

$$\frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}} < \frac{s_{G_\phi}}{\sqrt{n}}.$$

Proof. Since μ_Ω is finite (hence σ-finite) and $\int_\Omega \mathrm{d}\mu_\Omega(y)\int_\Omega |\phi(x,y)|^2\mathrm{d}\mu_\Omega(x) < +\infty$, one can apply Fubini's theorem (Rudin, 1987, Theorem 8.8, p. 164). So, computing the integral $\int_{\Omega\times\Omega}|\phi(x,y)|^2\mathrm{d}(\mu_\Omega\times\mu_\Omega)(x,y)$ first with respect to x and then to y we get

$$\begin{aligned}\|\phi\|^2_{\mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}} &= \int_{\Omega\times\Omega}|\phi(x,y)|^2\mathrm{d}(\mu_\Omega\times\mu_\Omega)(x,y)\\ &= \int_\Omega \mathrm{d}\mu_\Omega(y)\int_\Omega|\phi(x,y)|^2\mathrm{d}\mu_\Omega(x)\\ &\le s^2_{G_\phi}\mu_\Omega(\Omega) < +\infty. \end{aligned}\tag{12}$$

By Theorem 5(iii), T_ϕ is Hilbert–Schmidt and so $\sum_{j=1}^{+\infty}\lambda_j^2 < +\infty$. Moreover, by Theorem 5(iv) the kernel ϕ can be expressed as $\phi(x,y) = \sum_{j=1}^{+\infty}\lambda_j\psi_j(x)\psi_j(y)$, where the family $\{\psi_j\}$ is orthonormal. Thus

$$\|\phi\|^2_{\mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}} = \int_{\Omega\times\Omega}|\phi(x,y)|^2\mathrm{d}(\mu_\Omega\times\mu_\Omega)(x,y) = \sum_{j=1}^{+\infty}\lambda_j^2,$$

which, combined with (12), gives $s_{G_\phi} \ge \frac{\sqrt{\sum_{j=1}^{+\infty}\lambda_j^2}}{\sqrt{\mu_\Omega(\Omega)}}$. Hence

$$\begin{aligned}\frac{s_{G_\phi}}{\sqrt{n}} &\ge \frac{\sqrt{\sum_{j=1}^{+\infty}\lambda_j^2}}{\sqrt{\mu_\Omega(\Omega)}\sqrt{n}} \ge \frac{\sqrt{\sum_{j=1}^{n+1}\lambda_{n+1}^2}}{\sqrt{\mu_\Omega(\Omega)}\sqrt{n}}\\ &= \frac{\sqrt{n+1}}{\sqrt{n}}\frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}} > \frac{|\lambda_{n+1}|}{\sqrt{\mu_\Omega(\Omega)}}. \quad\square\end{aligned}$$

A similar relationship between s_G and eigenvalues holds for finite dictionaries.

Proposition 6. *Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a Hilbert space, $G = \{g_1,\dots,g_m\}$ its finite subset, $s_G = \sup_{g\in G}\|g\|_{\mathcal{H}}$, and $M(G)$ the matrix defined as $M(G)_{ij} = \langle g_i, g_j\rangle_{\mathcal{H}}$. Then for every positive integer $n < m$,*

$$\sqrt{\frac{\lambda_{n+1}(M(G))}{m}} < \frac{s_G}{\sqrt{n}}.$$

Proof. By the definition of the trace of a matrix and the equality between the trace of a matrix and the sum of its eigenvalues (Golub & Loan, 1996, p. 310), we get

$$\mathrm{Tr}(M(G)) = \sum_{j=1}^m M(G)_{jj} = \sum_{j=1}^m\langle g_j,g_j\rangle_{\mathcal{H}} = \sum_{j=1}^m\lambda_j(M(G)).$$

This, combined with $\langle g_j,g_j\rangle_{\mathcal{H}} \le s_G^2$ and $\lambda_j(M(G)) \ge 0$, $j = 1,\dots,m$, gives $s_G \ge \frac{\sqrt{\sum_{j=1}^m\lambda_j(M(G))}}{\sqrt{m}}$. Hence

$$\begin{aligned}\frac{s_G}{\sqrt{n}} &\ge \frac{\sqrt{\sum_{j=1}^m\lambda_j(M(G))}}{\sqrt{m}\sqrt{n}}\\ &\ge \frac{\sqrt{\sum_{j=1}^{n+1}\lambda_{n+1}(M(G))}}{\sqrt{m}\sqrt{n}}\\ &= \frac{\sqrt{n+1}}{\sqrt{n}}\frac{\sqrt{\lambda_{n+1}(M(G))}}{\sqrt{m}}\\ &> \frac{\sqrt{\lambda_{n+1}(M(G))}}{\sqrt{m}}. \quad\square\end{aligned}$$

6. Examples

In this section, we illustrate our results by some examples.

6.1. Example 1

Our first example describes two cases in which better estimates of G_ϕ-variations of eigenfunctions of T_ϕ than the ones in the form $\frac{\sqrt{\mu_\Omega(\Omega)}}{|\lambda_j|}$ (which are used in the proof of Theorem 7) can be derived. Recall that the function sinc: $\mathbb{R}\to\mathbb{R}$ is defined as

$$\mathrm{sinc}(x) = \begin{cases}\dfrac{\sin\pi x}{\pi x} & \text{for } x\neq 0,\\ 1 & \text{for } x = 0.\end{cases}$$

Let $\Omega \subseteq \mathbb{R}^d$, μ_Ω be a σ-finite measure on Ω, and $\phi \in \mathcal{L}^2_{\mu_\Omega\times\mu_\Omega}(\Omega\times\Omega)$ a symmetric function that can be represented as

$$\phi(x,y) = \sum_{j=1}^N\lambda_j\psi_j(x)\psi_j(y),$$

where N is finite or $N = +\infty$, the family $\{\psi_j\}$ is orthonormal in $\mathcal{L}^2_{\mu_\Omega}(\Omega)$, the sequence $\{|\lambda_j|\}$ is ordered non-increasingly, $\sum_{j=1}^N\lambda_j^2 < +\infty$, and for every $j = 1,\dots,N$ we have $\lambda_j \neq 0$.

Assume that either (a) or (b) holds, where

(a) the functions ψ_j have mutually disjoint supports, $|\psi_j| \le 1$, and for every $j = 1,\dots,N$ there exists $y_j \in \Omega$ such that $\psi_j(y_j) = 1$;
(b) $\Omega = \mathbb{R}$, μ_Ω is the Lebesgue measure μ and for every $j = 1,\dots,N$ we have $\psi_j(x) = \mathrm{sinc}(x-j)$.

Then for every $m \in \mathbb{N}_+$ or $m \le N$ in the finite case and every $n \le m$

$$d_n(B_1(\|\cdot\|_{G_\phi})) \ge |\lambda_m|\sqrt{1-\frac{n}{m}}\tag{13}$$

and

$$\delta(B_1(\|\cdot\|_{G_\phi}), \mathrm{span}_n G_\phi) \le \frac{|\lambda_1|}{\sqrt{n}},\tag{14}$$

so

$$\Delta_n B_1(\|\cdot\|_{G_\phi}) \ge |\lambda_m|\sqrt{1-\frac{n}{m}} - \frac{|\lambda_1|}{\sqrt{n}}.\tag{15}$$

To prove (13), we first show that in both cases (a) and (b) one has $\{\lambda_j\psi_j \mid j = 1,\dots,N\} \subseteq G_\phi$.

In the case (a), we have $\phi(\cdot,y_j) = \sum_{j=1}^N\lambda_j\psi_j(\cdot)\psi_j(y_j) = \lambda_j\psi_j(\cdot)$, as the functions ψ_j have mutually disjoint supports.

In the case (b), as $\mathrm{sinc}(0) = 1$ and $\mathrm{sinc}(j) = 0$ for every $j \in \mathbb{N}_+$ we get

$$\phi(\cdot,j) = \sum_{i=1}^N\lambda_i\mathrm{sinc}(\cdot - i)\mathrm{sinc}(j-i) = \lambda_j\mathrm{sinc}(\cdot - j).$$

So, in both cases we have $c_{m,G_\phi} = \max_{j=1,\dots,m} \|\psi_j\|_{G_\phi} \leq \frac{1}{|\lambda_m|}$. Hence, by Theorem 7 we get the lower bound (13) on $d_n(B_1(\|\cdot\|_{G_\phi}))$.

The upper bound (14) on $\delta(B_1(\|\cdot\|_{G_\phi}), \mathrm{span}_n G_\phi)$ follows by Theorem 1, after showing that $s_{G_\phi} \leq |\lambda_1|$.

In the case (a), this holds as the functions ψ_j have disjoint supports and thus for every $y \in \Omega$ there exists at most one j_y such that $\psi_{j_y}(y) \neq 0$. As $|\psi_j| \leq 1$, for every $y \in \Omega$ we get $\left\|\sum_{j=1}^N \lambda_j \psi_j(\cdot)\psi_j(y)\right\|_{\mathcal{L}^2_{\mu_\Omega}} = \left\|\lambda_{j_y}\psi_{j_y}(\cdot)\psi_{j_y}(y)\right\|_{\mathcal{L}^2_{\mu_\Omega}} \leq |\lambda_1|\|\psi_{j_y}\|_{\mathcal{L}^2_{\mu_\Omega}} = |\lambda_1|$. Thus

$$s_{G_\phi} = \sup_{y\in\Omega} \left\|\sum_{j=1}^N \lambda_j \psi_j(\cdot)\psi_j(y)\right\|_{\mathcal{L}^2_{\mu_\Omega}(\Omega)} \leq |\lambda_1|.$$

In the case (b), the inequality $s_{G_\phi} \leq |\lambda_1|$ can be verified as follows. By the orthonormality of the functions $\mathrm{sinc}(\cdot - j)$ in $\mathcal{L}^2_\mu(\mathbb{R})$, we get

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \left[\sum_{j=1}^N \lambda_j \mathrm{sinc}(x-j)\mathrm{sinc}(y-j)\right] \\
&\quad \times \left[\sum_{k=1}^N \lambda_k \mathrm{sinc}(x-k)\mathrm{sinc}(y-k)\right] \mathrm{d}x \\
&= \sum_{j=1}^N \sum_{k=1}^N \lambda_j\lambda_k \mathrm{sinc}(y-j)\mathrm{sinc}(y-k) \\
&\quad \times \int_{-\infty}^{+\infty} \mathrm{sinc}(x-j)\mathrm{sinc}(x-k)\mathrm{d}x \\
&= \sum_{j=1}^N \lambda_j^2 \mathrm{sinc}^2(y-j).
\end{aligned}$$

Hence

$$\begin{aligned}
s_{G_\phi} &= \sup_{y\in\mathbb{R}} \sqrt{\sum_{j=1}^N \lambda_j^2 \mathrm{sinc}^2(y-j)} \\
&\leq |\lambda_1| \sup_{y\in\mathbb{R}} \sqrt{\sum_{j=1}^N \mathrm{sinc}^2(y-j)}.
\end{aligned}$$

Then, denoting by ι the imaginary unit and by $\Lambda(\omega)$ the Fourier transform of $\mathrm{sinc}^2(x)$, since $\Lambda(2\pi j) = 1$ for $j = 0$ and 0 otherwise Poisson's sum formula (Papoulis, 1962, p. 47) gives

$$\begin{aligned}
\sup_{y\in\mathbb{R}} \sqrt{\sum_{j=1}^N \mathrm{sinc}^2(y-j)} &\leq \sup_{y\in\mathbb{R}} \sqrt{\sum_{j=-\infty}^{+\infty} \mathrm{sinc}^2(y-j)} \\
&= \sup_{y\in\mathbb{R}} \sqrt{\sum_{j=-\infty}^{+\infty} \mathrm{e}^{-\iota 2\pi j y} \Lambda(2\pi j)} \\
&= \sqrt{\Lambda(0)} = 1.
\end{aligned}$$

This concludes the proof of (14). Finally, (15) is obtained combining (13) and (14).

Inspection of the upper bound (15) provides various conditions guaranteeing that $\Delta_n(B_1(\|\cdot\|_{G_\phi}))$ is positive. For example, suppose that $|\lambda_m| \geq \frac{2|\lambda_1|}{\sqrt{m}}$ and let

$$z_1 := \frac{m}{2}\left(1 - \frac{\sqrt{|\lambda_m|^2 - 4\frac{|\lambda_1|^2}{m}}}{|\lambda_m|}\right)$$

and

$$z_2 := \frac{m}{2}\left(1 + \frac{\sqrt{|\lambda_m|^2 - 4\frac{|\lambda_1|^2}{m}}}{|\lambda_m|}\right).$$

Then for every positive integer $n \leq m$ such that $n \in (z_1, z_2)$ one has

$$\Delta_n(B_1(\|\cdot\|_{G_\phi})) \geq |\lambda_m|\sqrt{1 - \frac{n}{m}} - \frac{|\lambda_1|}{\sqrt{n}} > 0. \tag{16}$$

The proof of (16) amounts at finding conditions on a positive integer n guaranteeing that

$$|\lambda_m|\sqrt{1 - \frac{n}{m}} - \frac{|\lambda_1|}{\sqrt{n}} > 0. \tag{17}$$

Since n and m are positive, this is equivalent to

$$|\lambda_m|^2 n^2 - |\lambda_m|^2 mn + |\lambda_1|^2 m < 0. \tag{18}$$

Both roots z_1 and z_2 of the associated equation $|\lambda_m|^2 z^2 - |\lambda_m|^2 mz + |\lambda_1|^2 m = 0$ (with z a complex number) are real, as by assumption $|\lambda_m| \geq \frac{2|\lambda_1|}{\sqrt{m}}$. So, (18) holds for every positive integer n such that $n \in (z_1, z_2)$ and the proof of (16) is concluded.

Note that the condition $z_1 + 2 \leq z_2$ implies that the interval (z_1, z_2) contains at least one positive integer. For instance, taking $m = 100$ and $|\lambda_m| = \frac{|\lambda_1|}{4}$, one has $z_1 = 20$, $z_2 = 80$, and for every positive integer $n \in [21, 79]$

$$\Delta_n(B_1(\|\cdot\|_{G_\phi})) \geq |\lambda_m|\sqrt{1 - \frac{n}{100}} - \frac{|\lambda_1|}{\sqrt{n}} > 0.$$

For finite dictionaries G, the next two examples show cases for which the lower bound $\frac{1}{c_{A,G}}\sqrt{1 - \frac{n}{k}} - \frac{s_G}{\sqrt{n}}$ provided by Corollary 4(i) on $\Delta_n(B_1(\|\cdot\|_G))$ is positive.

6.2. Example 2

For $A = G$ finite and orthonormal with $\mathrm{card}\, G = m$, we have $c_{A,G} = 1$ and $s_G = 1$, so Corollary 4(i) with $k = m$ gives

$$\Delta_n(B_1(\|\cdot\|_G)) \geq \frac{1}{c_{A,G}}\sqrt{1 - \frac{n}{m}} - \frac{s_G}{\sqrt{n}} = \sqrt{1 - \frac{n}{m}} - \frac{1}{\sqrt{n}}.$$

Let $m \geq 4$,

$$z_1 := \frac{m}{2}\left(1 - \sqrt{1 - \frac{4}{m}}\right),$$

and

$$z_2 := \frac{m}{2}\left(1 + \sqrt{1 - \frac{4}{m}}\right).$$

Calculations similar to those made in the last part of Section 6.1 show that for every positive integer $n \leq m$ such that $n \in (z_1, z_2)$

$$\Delta_n(B_1(\|\cdot\|_G)) \geq \sqrt{1 - \frac{n}{m}} - \frac{1}{\sqrt{n}} > 0.$$

For instance, with $m = 36$ we have $z_1 = 18\left(1 - \frac{2\sqrt{2}}{3}\right) \approx 1.029$, $z_2 = 18\left(1 + \frac{2\sqrt{2}}{3}\right) \approx 34.969$, and for every positive integer $n \in [2, 34]$,

$$\Delta_n(B_1(\|\cdot\|_{G_\phi})) \geq \sqrt{1 - \frac{n}{36}} - \frac{1}{\sqrt{n}} > 0.$$

6.3. Example 3

Let G be finite and linearly independent with card $G = m$ and B a nonsingular $m \times m$ matrix such that for some orthonormal basis $A = \{a_k\}_{k=1}^m$ of span$G$ and $j = 1, \ldots, m$, one has $g_j = \sum_{k=1}^m B_{j,k} a_k$.

Choose as B a unit bidiagonal matrix (Vandebril, Van Barel, & Mastronardi, 2008, p. 154) B, B^{-1} as given in Box I

Let $\gamma := \max_{j=1,\ldots,m} |\beta_j|$, and suppose $\gamma \neq 1$. Then for every positive integer $n \leq m$

$$\Delta_n(B_1(\|\cdot\|_G)) \geq \frac{1}{c_{A,G}}\sqrt{1-\frac{n}{m}} - \frac{s_G}{\sqrt{n}} \geq \frac{1-\gamma}{1-\gamma^m}\sqrt{1-\frac{n}{m}} - \frac{\sqrt{1+\gamma^2}}{\sqrt{n}}. \tag{19}$$

To prove (19), we proceed as follows. By the definition of s_G and the structure of the matrix B, we get

$$s_G = \max_{j=1,\ldots,m} \sqrt{\sum_{k=1}^m |B_{j,k}|^2} \leq \sqrt{1+\gamma^2}. \tag{20}$$

Moreover, for $j = 1, \ldots, m$ we have $a_j = \sum_{k=1}^m B_{j,k}^{-1} g_k$, where the matrix B^{-1} is given by Vandebril et al. (2008, p. 154). Hence Proposition 2 gives $\|a_j\|_G = \sum_{k=1}^m |B_{j,k}^{-1}|$ and so

$$c_{A,G} = \max_{j=1,\ldots,m} \sum_{k=1}^m |B_{j,k}^{-1}| \leq \frac{1-\gamma^m}{1-\gamma}. \tag{21}$$

Thus, by (20), (21), and Corollary 4(i) (with $k = m$), we get (19).

Now, suppose

$$\frac{1-\gamma}{1-\gamma^m} \geq \frac{2\sqrt{1+\gamma^2}}{\sqrt{m}}$$

and let

$$z_1 := \frac{m}{2}\left(1 - \frac{\sqrt{\left(\frac{1-\gamma}{1-\gamma^m}\right)^2 - 4\frac{1+\gamma^2}{m}}}{\frac{1-\gamma}{1-\gamma^m}}\right)$$

and

$$z_2 := \frac{m}{2}\left(1 + \frac{\sqrt{\left(\frac{1-\gamma}{1-\gamma^m}\right)^2 - 4\frac{1+\gamma^2}{m}}}{\frac{1-\gamma}{1-\gamma^m}}\right).$$

Computations similar to those made in the last part of Section 6.1 show that for every positive integer $n \leq m$ such that $n \in (z_1, z_2)$,

$$\Delta_n(B_1(\|\cdot\|_G)) \geq \frac{1}{c_{A,G}}\sqrt{1-\frac{n}{m}} - \frac{s_G}{\sqrt{n}} \geq \frac{1-\gamma}{1-\gamma^m}\sqrt{1-\frac{n}{m}} - \frac{\sqrt{1+\gamma^2}}{\sqrt{n}} > 0.$$

For instance, taking $m = 6$ and $\gamma = 10^{-1}$ we get

$$z_1 = \frac{6}{2}\left(1 - \frac{\sqrt{\left(\frac{1-10^{-1}}{1-10^{-6}}\right)^2 - 4\frac{1+10^{-2}}{6}}}{\frac{1-10^{-1}}{1-10^{-6}}}\right) \approx 1.768$$

and

$$z_2 = \frac{6}{2}\left(1 + \frac{\sqrt{\left(\frac{1-10^{-1}}{1-10^{-6}}\right)^2 - 4\frac{1+10^{-2}}{6}}}{\frac{1-10^{-1}}{1-10^{-6}}}\right) \approx 4.232$$

and for every positive integer $n \in [2, 4]$

$$\Delta_n(B_1(\|\cdot\|_{G_\phi})) \geq \sqrt{1-\frac{n}{6}} - \frac{1}{\sqrt{n}} > 0.$$

6.4. Example 4

Finally, we give an example in which the upper bound on variable-basis approximation from Theorem 1(ii) and the lower bound on linear approximation from Corollary 2 have the same orders and the former is worse than the latter by a constant factor. This case is still of interest, as it shows that approximation from span$_n G_\phi$ is as at least as good as the one from a so-called *asymptotically optimal linear subspace*, as defined in Pinkus (2003, p. 908).

Let $\Omega = Y = [0, 1]$, $\mu_\Omega = \mu_Y = \mu$ be the Lebesgue measure, and $\phi(x, y) = h(x - y)$, where

$$h(t) = \sum_{j=0}^N A_j \cos(\omega_j t), \tag{22}$$

N is finite or $N = +\infty$, $\omega_j = 2\pi j$, $\theta_j \in [0, 2\pi)$, and $|A_0|^2 + \frac{1}{2}\sum_{j=0}^N |A_j|^2 < +\infty$. In this case, the eigenvalues of the integral operator T_ϕ are real and their absolute values coincide with the corresponding singular numbers. It is easy to check that the functions 1, $\sqrt{2}\cos(\omega_j y)$, and $\sqrt{2}\sin(\omega_j y)$ are orthonormal eigenfunctions of T_ϕ, whose eigenvalues are $a_0 = A_0$ with multiplicity 1 and $a_j = \frac{A_j}{2}$ with multiplicity 2, for $j \geq 1$ (Pinkus, 1985, p. 95). Note that

$$s_{G_\phi} = \sup_{y\in[0,1]} \sqrt{\int_0^1 |h(x-y)|^2 dx} = \sqrt{\int_0^1 |h(t)|^2 dt} = \sqrt{|A_0|^2 + \frac{1}{2}\sum_{j=0}^N |A_j|^2}.$$

In particular, consider the case

$$h(t) = \frac{4}{\pi}\sum_{j\ \text{odd}} \frac{1}{j}\sin\left(2\pi j\left(t - \frac{1}{4}\right)\right) = \frac{4}{\pi}\sum_{j\ \text{odd}} \frac{(-1)^{\lceil j/2\rceil}}{j}\cos(2\pi j t),$$

i.e., $h(t)$ is the square wave shown in Fig. 1(a). Then for every integer $n \geq 4$ the following bounds hold:

$$d_n(B_1(\|\cdot\|_{G_\phi})) \geq \frac{2}{\pi(2\lceil (n+1)/2\rceil - 1)} = \begin{cases} \frac{2}{\pi n} & \text{for } n \text{ odd}, \\ \frac{2}{\pi(n+1)} & \text{for } n \text{ even} \end{cases} \tag{23}$$

and

$$\delta(B_1(\|\cdot\|_{G_\phi}), \text{span}_n G_\phi) \leq \sqrt{\frac{2}{3}}\frac{4}{2\lceil (n+1)/2\rceil - 2} = \begin{cases} \sqrt{\frac{2}{3}}\frac{4}{n-1} & \text{for } n \text{ odd}, \\ \sqrt{\frac{2}{3}}\frac{4}{n} & \text{for } n \text{ even}. \end{cases} \tag{24}$$

To prove (23) and (24), we proceed as follows. As $s_{n+1}(T_\phi) = |\lambda_{n+1}(T_\phi)| = \frac{2}{\pi(2\lceil (n+1)/2\rceil - 1)}$, the lower bound (23) follows by Corollary 2 with $\mu([0, 1]) = 1$. Let us prove the upper bound

$$B = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \beta_1 & 1 & \ddots & & \vdots \\ 0 & \beta_2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \beta_{m-1} & 1 \end{pmatrix},$$

$$B^{-1} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ (-\beta_1) & 1 & \ddots & & & & \vdots \\ (-\beta_1)(-\beta_2) & (-\beta_2) & 1 & \ddots & & & \vdots \\ (-\beta_1)(-\beta_2)(-\beta_3) & (-\beta_1)(-\beta_2) & (-\beta_3) & 1 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \ddots & 0 \\ (-\beta_1)\cdots(-\beta_{m-1}) & (-\beta_2)\cdots(-\beta_{m-1}) & (-\beta_3)\cdots(-\beta_{m-1}) & \cdots & \cdots & (-\beta_{m-1}) & 1 \end{pmatrix}.$$

Box I.

(24). First, suppose that $n \geq 4$ is even and consider the periodic extension $f_{n/2}$ to $\mathbb{R}$ of the function

$$\hat{f}_{n/2}(t) = \begin{cases} -1, & \text{if } 0 \leq t \leq \frac{1}{4} - \frac{1}{n}, \\ n\left(t - \frac{1}{4}\right), & \text{if } \frac{1}{4} - \frac{1}{n} \leq t \leq \frac{1}{4} + \frac{1}{n}, \\ 1, & \text{if } \frac{1}{4} + \frac{1}{n} \leq t \leq \frac{3}{4} - \frac{1}{n}, \\ -n\left(t - \frac{3}{4}\right), & \text{if } \frac{3}{4} - \frac{1}{n} \leq t \leq \frac{3}{4} + \frac{1}{n}, \\ -1, & \text{if } \frac{3}{4} + \frac{1}{n} \leq t \leq 1. \end{cases}$$

Then, for every even integer $n \geq 4$, some calculations provide an error $e_{n/2}(G_\phi)$ in approximating the elements of G_ϕ by $f_{n/2}$ and its $\frac{n}{2} - 1$ translates by multiples of $\frac{2}{n}$ the upper bound $e_{n/2}(G_\phi) \leq \frac{4}{\sqrt{3n}}$ (one can easily see that the function $h\left(t - \frac{1}{n}\right)$ is one of the translates of h for which one has the worst approximation error in $\mathcal{L}^2_\mu([0,1])$ by $f_{n/2}$ and its $\frac{n}{2} - 1$ translates; see Fig. 1(b)). So, by Theorem 1(ii) we get (24). The estimate for every odd integer $n \geq 5$ is obtained from the previous one by noting that $\delta(B_1(\|\cdot\|_{G_\phi}), \mathrm{span}_n G_\phi) \leq \delta(B_1(\|\cdot\|_{G_\phi}), \mathrm{span}_{n-1} G_\phi)$.

7. Discussion

We have compared theoretical lower bounds on the worst-case approximation error achievable via optimal linear methods (i.e., via subspaces generated by any optimal n-tuple of elements of the ambient space) with upper bounds on variable-basis approximation by $\mathrm{span}_n G$ (i.e., approximation by all n-tuples of elements of a dictionary G of computational units in the ambient space).

For dictionaries G with finite cardinality m, we have provided examples in which approximation of the unit ball in G-variation by elements of $\mathrm{span}_n G$ is better than *every* linear approximation of dimension $n \leq m$, for values of n belonging to suitable intervals. This case is of practical interest, as often in applications one needs sparse approximations, i.e. approximations with a "reasonably small" number n of computational units.

When the operator T_ϕ associated with a variable-basis model $\mathrm{span}_n G_\phi$ is symmetric and its eigenfunctions belong to the set G_ϕ, we have exhibited cases in which approximation of the unit ball in G_ϕ-variation by elements of $\mathrm{span}_n G_\phi$ is better than every linear approximation of dimension n.

These results are rather strong, as they prove that in such cases the worst-case errors achievable by the dictionaries that we have described are smaller than those provided by *any* linear approximator.

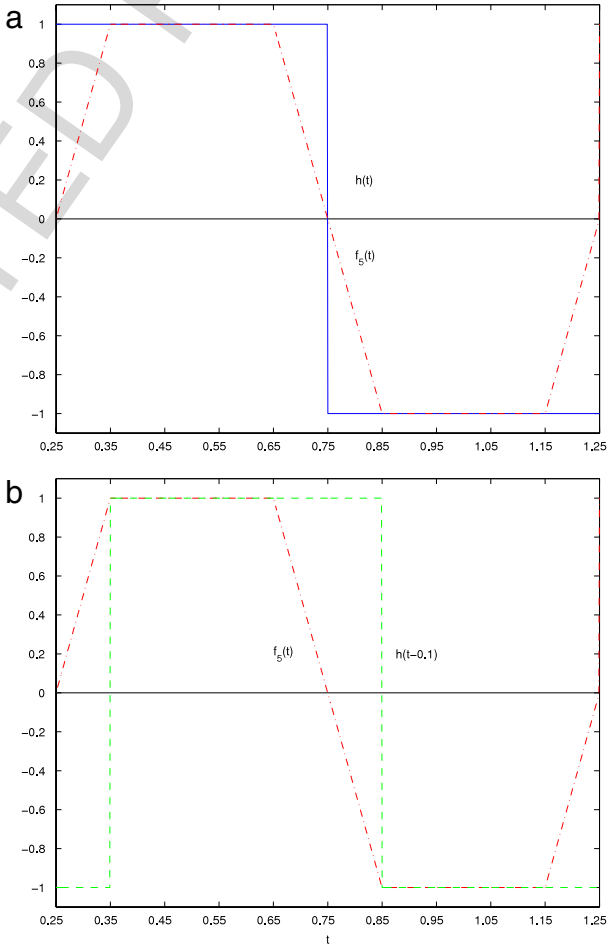

Fig. 1. Plots of one period of the functions $h(t)$, $h(t-0.1)$, and $f_5(t)$ in the example of Section 6.4 with $n = 10$.

However, variable-basis approximation may be advantageous even when theoretically a better linear approximator might exist, because finding the latter may be an unfeasible task. To illustrate such a case, we have provided an example in which an upper bound on variable-basis approximation and a lower bound on linear approximation have asymptotically the same orders.

Acknowledgements

The authors are grateful to P.C. Kainen for his careful reading of the manuscript and his suggestions to improve the style of presentation.

Giorgio Gnecco and Marcello Sanguineti were partially supported by a PRIN grant from the Italian Ministry for University and Research, project "Adaptive State Estimation and Optimal Control". Věra Kůrková was partially supported by GA ČR grant 201/08/1744 and the Institutional Research Plan AV0Z10300504. The collaboration of Věra Kůrková with Marcello Sanguineti and Giorgio Gnecco was partially supported by CNR - AV ČR project 2010-2012 "Complexity of Neural-Network and Kernel Computational Models".

References

Akhiezer, N. I., & Glazman, I. M. (1993). *Theory of linear operators in Hilbert space*. New York: Dover.

Barron, A. R. (1992). Neural net approximation. In *Proc. 7th Yale workshop on adaptive and learning systems* (pp. 69–72). Yale University Press.

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

Courant, R. (1988). *Differential and integral calculus*: *Vol. II*. Wiley.

Friedman, A. (1982). *Foundations of modern analysis*. New York: Dover.

Girosi, F., & Anzellotti, G. (1993). Rates of convergence for radial basis functions and neural networks. In R. J. Mammone (Ed.), *Artificial neural networks for speech and vision* (pp. 97–113). Chapman & Hall.

Giulini, S., & Sanguineti, M. (2000). On dimension-independent approximation by neural networks and linear approximators. In *Proc. int. joint conference on neural networks* (pp. I283–I288).

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 K. Diamantaras, W. Duch, & L. Iliadis (Eds.), *LNCS*: *Vol. 6352. ICANN 2010* (pp. 358–367). Springer-Verlag.

Gnecco, G., & Sanguineti, M. On a variational norm tailored to variable-basis approximation schemes. IEEE Transactions on Information Theory (in press).

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. London: John Hopkins University Press.

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. (2009). An integral upper bound for neural network approximation. *Neural Computation*, *21*, 2970–2989.

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

Kainen, P. C., Kůrková, V., & Vogt, A. (2000a). Best approximation by Heaviside perceptron networks. *Neural Networks*, *13*, 695–697.

Kainen, P. C., Kůrková, V., & Vogt, A. (2000b). Geometry and topology of continuous best and near best approximations. *Journal of Approximation Theory*, *105*, 252–262.

Kainen, P. C., Kůrková, V., & Vogt, A. (2001). Continuity of approximation by neural networks in L_p-spaces. *Annals of Operations Research*, *101*, 143–147.

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.

Kecman, V. (2001). *Learning and soft computing*. Cambridge: MIT Press.

Kolmogorov, A. N. (1936). Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse. *Annals of Mathematics*, *37*(1), 107–110. (English translation: "On the best approximation of functions of a given class", In: V.M. Tikhomirov (Ed.), *Selected works of A.N. Kolmogorov: Vol. I* (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. (2009). Model complexity of neural networks and integral transforms. In M. Polycarpou, C. Panayiotou, C. Alippi, & G. Ellinas (Eds.), *Lecture notes in computer science*: *Vol. 5768*. *Proc. ICANN 2009* (pp. 708–718). Berlin, Heidelberg: Springer.

Kůrková, V. (2010). Integral transforms and norms induced by computational units (submitted for publication).

Kůrková, V., Kainen, P. C., & Kreinovich, V. (1997). Estimates of the number of hidden units and variation with respect to half-spaces. *Neural Networks*, *10*, 1061–1068.

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. (2005). Error estimates for approximate optimization by the extended Ritz method. *SIAM Journal on Optimization*, *15*, 461–487.

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

Kůrková, V., Savický, P., & Hlaváčková, K. (1998). Representations and rates of approximation of real-valued Boolean functions by neural networks. *Neural Networks*, *11*, 651–659.

Makovoz, Y. (1996). Random approximants and neural networks. *Journal of Approximation Theory*, *85*, 98–109.

Oden, J. T., & Demkowicz, L. F. (1996). *Applied functional analysis*. CRC Press.

Papoulis, A. (1962). *The Fourier integral and its applications*. USA: McGraw-Hill.

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

Pinkus, A. (2003). Negative theorems in approximation theory. *American Mathematical Monthly*, *110*, 900–911.

Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. In *Séminaire d'analyse fonctionnelle 1980–81: Vol. I*, no. 12. École Polytechnique, Centre de Mathématiques, Palaiseau, France.

Rudin, W. (1987). *Real and complex analysis*. New York: McGraw-Hill.

Shubin, M. A. (2001). *Pseudodifferential operators and spectral theory*. Berlin, Heidelberg: Springer.

Smith, K. A. (1999). Neural networks for combinatorial optimization: a review of more than a decade of research. *INFORMS Journal on Computing*, *11*, 15–34.

Vandebril, R., Van Barel, M., & Mastronardi, N. (2008). *Matrix computations and semiseparable matrices: linear systems*. The Johns Hopkins University Press.

Weidmann, J. (1980). *Linear operators in Hilbert spaces*. New York: Springer.

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.