Dependence of Computational Models on Input Dimension: Tractability of Approximation and Optimization Tasks

Paul C. Kainen, Věra Kůrková, and Marcello Sanguineti

***Abstract*—The role of input dimension d is studied in approximating, in various norms, target sets of d-variable functions using linear combinations of adjustable computational units. Results from the literature, which emphasize the number n of terms in the linear combination, are reformulated, and in some cases improved, with particular attention to dependence on d. For worst-case error, upper bounds are given in the factorized form $\xi(d)\kappa(n)$, where κ is nonincreasing (typically $\kappa(n) \sim n^{-1/2}$). Target sets of functions are described for which the function ξ is a polynomial. Some important cases are highlighted where ξ decreases to zero as $d \to \infty$. For target functions, extent (e.g., the size of domains in $\mathbb{R}^d$ where they are defined), scale (e.g., maximum norms of target functions), and smoothness (e.g., the order of square-integrable partial derivatives) may depend on d, and the influence of such dimension-dependent parameters on model complexity is considered. Results are applied to approximation and solution of optimization problems by neural networks with perceptron and Gaussian radial computational units.**

***Index Terms*—Dictionary-based computational models, high-dimensional approximation and optimization, model complexity, polynomial upper bounds.**

I. Introduction

MANY tasks in engineering, operations research, biology, etc. require optimization of decision functions dependent on a large number d of variables. Such functions may represent, e.g., routing strategies for telecommunication networks, stochastic decision problems, resource-allocation in computer networks, releasing policies in management of water-reservoir networks, scheduling for queueing networks in manufacturing systems, etc.

Experimental results have shown that optimization over decision functions built from relatively few computational units with a simple structure may obtain surprisingly good performance in high-dimensional optimization tasks (see, e.g., [1]–[9] and the references therein). In these models, the decision functions take on the form of linear combinations of input-output maps computed by units belonging to some *dictionary* [10]–[12]. Examples of dictionaries are various parameterized sets of functions such as those computable by perceptrons, radial or kernel units, Hermite functions, trigonometric polynomials, and splines.

Sometimes approximation from a dictionary is called *variable-basis approximation* [13], [5] in contrast to *linear approximation* [14], where the n-tuple is fixed (it is formed by the first n elements of a set with a fixed linear ordering) and only the coefficients of linear combination are adjustable. The number n of units in the linear combination plays the role of *model complexity* as it corresponds to the number of computational units in the so-called "hidden layer" of neural networks [15]. Estimates of model complexity needed to guarantee a desired accuracy in approximation of some family of functions can be derived from upper bounds on rates of approximation. Such upper bounds typically include several factors, one of which involves the number n of terms in the linear combinations, while another involves the number d of variables (i.e., the dimension of the inputs of computational units).

Emphasis on model complexity n is certainly reasonable when the number d of variables is fixed, but in modern research, where technology allows ever-increasing amounts of data to be collected, it is natural to also take into account the role of d. Further, if d has a combinatorial aspect (e.g., as the number of admissible paths in a communication network), then it may grow very dramatically and so make the computational requirements unfeasible: algorithms will require an exponential growth in time and resources because of the combinatorial explosion of possible substructures [16] as d increases.

Also, dependence on dimension d may be cryptic; i.e., estimates involve parameters that are constant with respect to n but *do* depend on d and the manner of dependence is not specified; see, e.g., [17]–[25]. Most available upper bounds take the factorized form

$$\xi(d)\kappa(n) \tag{1}$$

(see Section VII for a discussion of their tightness). In some literature (see, e.g., [18]) the terms depending on d are referred

Manuscript received October 29, 2009; revised March 17, 2011; accepted July 20, 2011. Date of current version February 08, 2012. The work of P. C. Kairen was supported in part by Georgetown University. The work of V. Kůrková was supported in part by GA ČR Grants 201/08/1744 and P202/11/1368 and the Institutional Research Plan AV0Z10300504. The work of M. Sanguineti was supported in part by a PRIN grant from the Italian Ministry for University and Research, project "Adaptive State Estimation and Optimal Control". The collaborative work of V. Kůrková and M. Sanguineti was supported in part by the CNR—AV ČR Project 2010–2012 "Complexity of Neural-Network and Kernel Computational Models." The collaborative work of V. Kůrková and P. C. Kairen was supported in part by MŠMT project KONTAKT ALNN ME10023.

P. C. Kainen is with the Department of Mathematics and Statistics, Georgetown University, Washington, DC 20057-1233 USA (e-mail: kainen@georgetown.edu).

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

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

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

Digital Object Identifier 10.1109/TIT.2011.2169531

to as "constants" since these papers focus on the number n of computational units and assume a fixed value for the dimension d of the input space. Such estimates are formulated as bounds on $O(\kappa(n))$, so the dependence on d is hidden in the so-called "big O" notation (e.g., [26]). However, it has been shown that such "constants" can actually grow at an exponential rate in d [27], [28]. In fact, the families of functions for which the estimates are valid may become negligibly small for large d [29].

As remarked in [30], in general the dependence of approximation errors on the input dimension d—i.e., the function $\xi(d)$—is much harder to estimate than dependence on the number n of computational units. Not many such estimates are available. Deriving them can help to determine when machine-learning tasks are computationally feasible as the dimension d of the input space grows. More than 15 years ago Juditsky *et al.* [2] (see especially Section 3 therein) warned that dictionary-based computational models are not a *panacea* for high-dimensional optimization and approximation tasks, but the word of caution contained in their paper seems to have been forgotten.

This paper unifies a number of recent studies of dictionary-based computational models from the standpoint of input dimension. It is shown that many "dimension-independent" computations *do*, in fact, depend on the input dimension d, though the dependence may be hidden. We investigate previous upper bounds on approximation from this perspective, pointing out how they depend not only on the number of variables d but also on *extent* (i.e., the Lebesgue measure of domains in $\mathbb{R}^d$ on which the target functions are defined), *scale* (e.g., norms of the target functions—that is, the radius of the ball to be approximated), and *smoothness* (e.g., the order of square-integrable partial derivatives). Moreover, we derive some new estimates for rates of approximation and optimization via dictionaries, with explicitly-stated dependence on input dimension and these additional parameters. In contrast to "big O" estimates, where unspecified "constants" may increase exponentially with dimension, our explicit bounds sometimes involve numerical factors that *decrease* exponentially with dimension. The volume of the unit ball in d-dimensions is one such example.

We utilize and extend the concept of "tractability" in approximation of multivariable functions by variable-basis computational models expressed as linear combinations of all n-tuples of functions computable by units belonging to a dictionary. Tractability was first introduced in studying *information-based complexity*; see, e.g., [31], [30], [32]). It was also used in *worst-case analysis* as given in [33] and [34]. We call the problem of determining the error in approximating a family of functions of d-variables using n computational units *tractable* (with respect to d) iff in the factorized estimate (1) the function κ is nonincreasing and the function ξ is bounded above by a polynomial in d. Often, the function of model complexity take the form $\kappa(n) = n^{-1/m}$, where $m > 0$ is related to the smoothness of the functions to be approximated, e.g., m is the order up to which partial derivatives are square integrable. In such cases, input-output functions with

$$n \geq \left(\frac{\xi(d)}{\varepsilon}\right)^m$$

computational units can approximate the given class of functions within ε. Hence, when $\xi(d)$ is polynomial, the model complexity needed to achieve a given accuracy grows at most polynomially with input dimension. Note that estimates of approximation errors in spaces with square-integrable partial derivatives up to some order m from linear approximating sets and ridge functions were obtained in [35, Chapter VII] and [36], resp. However, their estimates are asymptotic, not in the factorized form separating d and n, and have multiplicative factors which depend on d in an unspecified way.

Polynomial growth for $\xi(d)$ does not provide sufficient control of model complexity unless the degree of the polynomial is quite small. For large dimension d of the input space, even quadratic approximation is not going to be sufficient. But there are interesting situations, described below, where dependence on d is *linear* or better, and we highlight cases in which the function $\xi(d)$ *decreases to zero*—even exponentially fast—with dimension. In this "hyper-tractable" case, a single computational unit can approximate arbitrarily well provided the dimension is large enough.

The conditions that we derive to guarantee tractable or hyper-tractable approximation by various dictionaries define sets large enough to include many smooth functions on $\mathbb{R}^d$; for example, d-variable Gaussian functions on $\mathbb{R}^d$ can be tractably approximated by perceptron neural networks.

A preliminary version of some results contained in this work was presented in [37].

The paper is organized as follows. In Section II, the concept of tractability with respect to the dimension d of the worst-case approximation from dictionaries is introduced. In Section III, results from nonlinear approximation are used to describe sets of functions for which approximation is tractable. These results are applied in Section IV to approximation by Gaussian radial basis function networks and by perceptron networks in Section V. In Section VI, the results from Section III are applied to minimization of functionals over dictionaries. Section VII contains some concluding remarks.

II. Worst-Case Tractability in Approximation From a Dictionary

Let S, T be two subsets of a normed linear space. The worst-case approximation error in approximating elements of S by elements of T is formalized by the concept of deviation of S from T.

Definition 1: Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed linear space; $S, T \subseteq \mathcal{X}$ nonempty. The *deviation* of the target set S from the approximating set T in $\mathcal{X}$ is defined as

$$\delta(S,T) = \delta(S,T)_{\mathcal{X}} := \sup_{f\in S} \inf_{g\in T} \|f-g\|_{\mathcal{X}}.$$

Clearly, $\delta(S,T)_{\mathcal{X}} = 0$ if and only if S is contained in the $\mathcal{X}$-closure of T. Moreover, deviation is monotonic

$$\delta(S',T') \leq \delta(S,T) \text{ if } S' \subseteq S \text{ and } T' \supseteq T$$

and homogeneous

$$\delta(c\,S, cT) = c\,\delta(S,T)$$

where

$$c\,A := \{c\,a \,|\, a \in A\}.$$

For a subset G of a linear space, let

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

denote the set of all n-fold linear combinations of its elements. The sets G and $\mathrm{span}_n G$ are sometimes called a *dictionary* [12] and a *variable-basis approximation scheme* [13], [5] (or *approximation from a dictionary*), respectively. We use these sets as our approximating set; i.e.,

$$T = \mathrm{span}_n\, G.$$

The set $S = A_d$ of targets is a subset of a space of d-variable functions, usually the ball in some appropriate space or its intersection with another suitable space. By homogeneity, if the scale of the data (i.e., the radius of this ball) changes (as it may when the dimension changes), then the same scale term will appear in the deviation because the sets $\mathrm{span}_n\, G$ are not changed by scalar multiplication.

With appropriate choices of G, the sets $\mathrm{span}_n G$ provide dictionary-based models used in applications and n, the *model complexity*, is the number of computational units. For example, consider an input set $\Omega_d \subseteq \mathbb{R}^d$, a real-valued function $\phi : \mathbb{R}^q \times \Omega_d \to \mathbb{R}$ of two vector variables, and let

$$G_d^{\phi} := \{\phi(u, \cdot) \,|\, u \in \mathbb{R}^q\},$$

where $\phi(u, \cdot) : x \mapsto \phi(u, x)$ for all $x \in \Omega_d$.

For suitable choices of computational unit ϕ, the sets

$$\mathrm{span}_n\, G_d^{\phi} := \left\{ \sum_{i=1}^{n} w_i \phi(u_i, \cdot) \,|\, w_i \in \mathbb{R}, u_i \in \mathbb{R}^q \right\}$$

consist of functions computable by one-hidden layer neural networks, radial basis functions, kernel models, splines with free nodes, trigonometric polynomials with free frequencies, Hermite functions, etc. [29], [38]. For example, if $q = d + 1$ and $\phi((v, b), x) = \psi(v{\cdot}x{+}b)$, then functions in $\mathrm{span}_n\, G_d^{\psi}$ are called *perceptron networks*. If $q = d + 1$, ψ is positive and even, and $\phi((v, b), x) = \psi(b\|x - v\|)$, then functions in $\mathrm{span}_n\, G_d^{\psi}$ are called *radial basis function (RBF)* networks.

Let $\mathbb{N}'$ be any infinite subset of $\mathbb{N}_+$, the set of positive integers. We focus on upper bounds on rates of approximation from dictionaries (see, e.g., [17]–[20], [24], [39]) of functions in A_d by $\mathrm{span}_n\, G_d$ taking on the factorized form

$$\delta(A_d, \mathrm{span}_n\, G_d)_{\mathcal{X}_d} \le \xi(d)\, \kappa(n) \tag{2}$$

where $\xi : \mathbb{N}' \to \mathbb{R}_+$ is a function of the number d of variables of the functions in $\mathcal{X}_d$ and $\kappa : \mathbb{N}_+ \to \mathbb{R}$.

Definition 2: The problem of approximating A_d by elements of $\mathrm{span}_n\, G_d$ is called *tractable with respect to* d *in the worst case* or simply *tractable* iff in upper bound (2) for every $d \in \mathbb{N}'$ one has $\xi(d) \le d^{\nu}$ for some $\nu > 0$ and κ is a nonincreasing nonnegative function of the model complexity n. The problem is called *hyper-tractable* iff it is tractable and, in addition, $\lim_{d\to\infty} \xi(d) = 0$.

We also study rates of approximation of sets of d-variable functions of the form $r_d\, A_d$ for various scaling factors r_d. In particular, for the unit ball $B_1(\|\cdot\|)$ we investigate scaled sets $r_d\, B_1(\|{\cdot}\|) = B_{r_d}(\|{\cdot}\|)$. If approximation of A_d by $\mathrm{span}_n\, G_d$ is hyper-tractable, then the scaled problem of approximating $r_d A_d$ by $\mathrm{span}_n\, G_d$ is tractable, unless r_d grows faster than $\xi(d)^{-1}$. If $\xi(d)$ goes to zero at an exponential rate, then the scaled problem is hyper-tractable if r_d is at most polynomial in d.

III. Sets of Tractable Functions

In this section, we describe some sets of d-variable functions that in various function spaces can be tractably approximated by linear combinations of functions from general dictionaries. The sets can be described as balls in certain norms, tailored to such dictionaries. Their tractability depends on the speed of growth of their radii with the dimension d.

For any nonempty bounded subset G of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, the closure of its symmetric convex hull $\mathrm{cl}_{\mathcal{X}}\ \mathrm{conv}\ (G \cup -G)$ uniquely determines a norm for which it forms the unit ball. Such a norm is the Minkowski functional [40, p. 131] of the set $\mathrm{cl}_{\mathcal{X}}\ \mathrm{conv}\ (G \cup -G)$.

Definition 3: Let G be a nonempty bounded subset of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, $f \in \mathcal{X}$, and $s(G) := \sup_{g\in G} \|g\|_{\mathcal{X}}$. The *$G$-variation* of f is defined as

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

Note that G-variation can be infinite. It is a norm on the subspace of $\mathcal{X}$ formed by functions with finite G-variation. The concept of G-variation was introduced in [17] for families of characteristic functions and extended in [41] to arbitrary bounded sets of functions. Clearly, $\|f\|_X \le s(G)\, \|f\|_G$. Also for $G \subset \mathcal{Y} \subset \mathcal{X}$, where $\mathcal{X}, \mathcal{Y}$ are normed linear spaces with some constant $C > 0$ such that for all $f \in \mathcal{Y}$, $\|f\|_{\mathcal{X}} \le C\|f\|_{\mathcal{Y}}$ (i.e., with $\mathcal{Y}$ continuously embedded in $\mathcal{X}$) and with G nonempty and bounded in $\mathcal{Y}$, for every $f \in \mathcal{Y}$ one has $\|f\|_{G,\mathcal{Y}} \le \|f\|_{G,\mathcal{X}}$ [41].

Such variational norms play a crucial role in upper bounds on rates of approximation and optimization by sets of the form $\mathrm{span}_n G$ [29], [42]. Before stating these bounds, we present an estimate of variational norm for parameterized sets G of the form $G = G^{\phi} = \{\phi(\cdot, a) \,|\, a \in A\}$. The following theorem from [43] (see also [44]) gives an upper bound on G^{ϕ}-variation for functions which can be represented as

$$f(x) = \int_A w(a)\phi(x, a) d\mu(a).$$

For a measurable set $\Omega \subseteq \mathbb{R}^d$, a measure ρ on Ω, and $1 \le p < \infty$, we denote by $(\mathcal{L}^p(\Omega, \rho), \|\cdot\|_{\mathcal{L}^p(\Omega,\rho)})$ the space of (equivalence classes of) real-valued functions on Ω that have integrable pth power with respect to the measure, endowed with the standard norm. (See [39] for a sup-norm example.)

Theorem 1: Let $\Omega \subseteq \mathbb{R}^d$, $A \subseteq \mathbb{R}^s$, μ a Borel measure on A, ρ a σ-finite measure on Ω, and $\phi : \Omega \times A \to \mathbb{R}$ a mapping such that $G^{\phi} := \{\phi(\cdot, a) \,|\, a \in A\}$ is a bounded subset of $\mathcal{L}^p(\Omega,\rho)$ for some $p \in [1,\infty)$. If $f \in \mathcal{L}^p(\Omega,\rho)$ is such that for some $w \in \mathcal{L}^1(A,\mu)$ and for ρ-almost every $x \in \Omega$, $f(x) = \int_A w(a)\phi(x,a)d\mu(a)$. Then letting $s(G^\phi) := \sup_{a\in A} \|\phi(\cdot,a)\|_{\mathcal{L}^p}$,

$$\|f\|_{G^\phi,\mathcal{L}^p(\Omega,\rho)} \leq \|w\|_{\mathcal{L}^1(A,\mu)}\, s(G^\phi).$$

The next theorem [3] reformulates estimates from [45], [24], [18], and [20] and, together with its corollary below, is used repeatedly in the sequel.

Theorem 2: Let $(\mathcal{X}, \|\cdot\|_\mathcal{X})$ be a Banach space, G a bounded nonempty subset, and $s(G) := \sup_{g\in G}\|g\|_\mathcal{X}$, $f \in \mathcal{X}$. Then for every positive integer n

(i) for $(\mathcal{X}, \|\cdot\|_\mathcal{X})$ a Hilbert space,

$$\|f - \mathrm{span}_n G\|_\mathcal{X} \leq s(G)\sqrt{\|f\|_G^2 - \|f\|_\mathcal{X}^2}\; n^{-1/2};$$

(ii) for $(\mathcal{X}, \|\cdot\|_\mathcal{X}) = (\mathcal{L}^q(\Omega,\rho), \|\cdot\|_{\mathcal{L}^q(\Omega,\rho)})$ $q \in (1,\infty)$, and a measure ρ on $\Omega \subseteq \mathbb{R}^d$:

$$\|f - \mathrm{span}_n G\|_{\mathcal{L}^p(\Omega,\rho)} \leq (2^{1+1/a}\, s(G)\|f\|_G)\, n^{-1/b}$$

where $a = \min(q, \frac{q}{q-1})$, and $b = \max(q, \frac{q}{q-1})$.
Theorem 2 implies upper bounds on the worst-case errors for balls in G-variation. The bounds are of the factorized form $\xi(d)\kappa(n)$ with $\kappa(n) = n^{-1/2}$ or $\kappa(n) = n^{-1/b}$.

The next corollary follows from Theorem 2 and the definition of deviation.

Corollary 1: Let d be a positive integer, $(\mathcal{X}_d, \|\cdot\|_{\mathcal{X}_d})$ a Banach space of d-variable functions, and G_d some bounded nonempty subset with $s(G_d) := \sup_{g\in G_d}\|g\|_{\mathcal{X}_d}$. Then for every positive integer n

(i) for $(\mathcal{X}_d, \|\cdot\|_{\mathcal{X}_d})$ a Hilbert space:

$$\delta(B_{r_d}(\|\cdot\|_{G_d,\mathcal{X}_d}), \mathrm{span}_n\, G_d)_{\mathcal{X}_d} \leq s(G_d)\, r_d\, n^{-1/2};$$

(ii) for $(\mathcal{X}_d, \|\cdot\|_{\mathcal{X}_d}) = (\mathcal{L}^p(\Omega_d,\rho), \|\cdot\|_{\mathcal{L}^p(\Omega_d,\rho)})$ with $p \in (1,\infty)$ and a measure ρ on $\Omega_d \subseteq \mathbb{R}^d$:

$$\delta(B_{r_d}(\|\cdot\|_{G_d,\mathcal{L}^p(\Omega_d,\rho)}), \mathrm{span}_n\, G_d)_{\mathcal{L}^p(\Omega_d,\rho)} \\ \leq 2^{1+1/a}\, s(G_d)\, r_d\, n^{-1/b}$$

where $a = \min(q, \frac{q}{q-1})$, and $b = \max(q, \frac{q}{q-1})$.
So in the Hilbert space case, approximation of the ball of radius r_d in variational norm for the class G^ϕ is tractable provided that $s(G_d)\, r_d$ is polynomial.

It was shown in [21] that the upper bound (ii) from Theorem 2 does not hold for general bounded subsets of $\mathcal{L}^1$ or $\mathcal{L}^\infty$-spaces. However, for special cases of certain sets G with finite VC-dimension, one can derive upper bounds in this form. Recall that the *VC-dimension* of a subset of the set of characteristic functions of a set Ω is defined as follows. The *characteristic* or *indicator* function of $S \subseteq \Omega$ is defined for $x \in \Omega$ as $\chi_S(x) = 1$ if $x \in S$, otherwise $\chi_S(x) = 0$. Let $\mathcal{F}$ be any family of characteristic functions of subsets of Ω, $\mathcal{S} = \{S \subseteq \Omega \,|\, \chi_S \in \mathcal{F}\}$ the family of the corresponding subsets of Ω, and A a subset of Ω. The set A is said to be *shattered* by $\mathcal{F}$ if $\{S \cap A \,|\, S \in \mathcal{S}\}$ is the whole power set of A. The *VC-dimension* of $\mathcal{F}$ is the largest cardinality of any subset A that is shattered by $\mathcal{F}$. The *coVC-dimension* of $\mathcal{F}$ is the VC-dimension of the set $\mathcal{F}' := \{ev_x \,|\, x \in \Omega\}$, where the *evaluation* $ev_x : \mathcal{F} \to \{0,1\}$ is defined for every $\chi_S \in \mathcal{F}$ as $ev_x(\chi_S) = \chi_S(x)$.

Let $\Omega \subseteq \mathbb{R}^d$; by $(\mathcal{M}(\Omega), \|\cdot\|_{\mathcal{M}(\Omega)})$ we denote the space of all Lebesgue-measurable functions $f : \Omega \to \mathbb{R}$ with finite supremum on Ω, with

$$\|f\|_{\mathcal{M}(\Omega)} := \sup_{x\in\Omega}(|f(x)|)$$

The following theorem is a reformulation of the estimates from [23, Th. 3] in terms of G-variation.

Theorem 3: Let $\Omega \subseteq \mathbb{R}^d$ and G be a subset of the set of characteristic functions on Ω such that the co-VC-dimension h_G^* is finite. Then in $(\mathcal{M}(\Omega), \|\cdot\|_{\mathcal{M}(\Omega)})$ for every positive integer n

$$\|f - \mathrm{span}_n G\|_{\mathcal{M}(\Omega)} \\ \leq 6\sqrt{3}\,\|f\|_{G,\mathcal{M}(\Omega)}\,(h_G^*)^{1/2}\,(\log n)^{1/2}\, n^{-1/2}. \tag{3}$$

As a corollary of Theorem 3, we have the following upper bound on deviation in the supremum norm of balls in G-variation for sets G of characteristic functions with finite VC-dimension.

Corollary 2: Let $\Omega_d \subseteq \mathbb{R}^d$ and G_d be a subset of the sets of characteristic functions on Ω_d such that the co-VC-dimension $h_{G_d}^*$ of G_d is finite. Then in $(\mathcal{M}(\Omega_d), \|\cdot\|_{\mathcal{M}(\Omega_d)})$ for every $n \in \mathbb{N}_+$

$$\delta(B_{r_d}(\|\cdot\|_{G_d,\mathcal{M}(\Omega_d)}), \mathrm{span}_n G_d)_{\mathcal{M}(\Omega_d)} \\ \leq 6\sqrt{3}\,\left(h_{G_d}^*\right)^{1/2} r_d\,(\log n)^{1/2}\, n^{-1/2}.$$

In the upper bounds from Corollaries 1 and 2, we have $\xi(d) = s(G_d)\, r_d$ and $\xi(d) = 6\sqrt{3}\,\left(h_{G_d}^*\right)^{1/2} r_d$, resp. These estimates imply tractability for the scaled problem (estimating deviation from $\mathrm{span}_n G$ for the ball of radius r_d) when $s(G_d)\, r_d$ and $\left(h_{G_d}^*\right)^{1/2} r_d$, resp., grow polynomially with d increasing. While $s(G_d)$ and $h_{G_d}^*$ are determined by the choice of G_d, one may be able to specify r_d in such a way that $\xi(d)$ is a polynomial.

Table I summarizes the estimates provided by Corollaries 1 and 2.

IV. Tractability of Approximation by Gaussian RBF Networks

In this section, we apply the estimates from the previous Section III to Gaussian radial-basis-function (RBF) networks. Let $\gamma_{d,b} : \mathbb{R}^d \to \mathbb{R}$ denote the d-dimensional Gaussian function of *width* $b > 0$ and *center* $0 = (0,\ldots,0)$ in $\mathbb{R}^d$:

$$\gamma_{d,b}(x) := e^{-b\|x\|^2}.$$

When $b = 1$, we write merely γ_d instead of $\gamma_{d,1}$. Note that larger values of b correspond to sharper peaks, so "width" is parametrized oppositely to what one might expect. For $b > 0$, let

$$G_d^\gamma(b) := \left\{\tau_y(\gamma_{d,b}) \,|\, y \in \mathbb{R}^d\right\}$$

TABLE I
ESTIMATES PROVIDED BY COROLLARIES 1 AND 2 FOR VARIATIONAL NORMS

ambient space $\mathcal{X}_d$	dictionary G_d	$\xi(d)$	$\kappa(n)$
Hilbert space	bounded	$s(G_d)\, r_d$	$n^{-1/2}$
$(\mathcal{L}^p(\Omega_d), \|.\|_{\mathcal{L}^p(\Omega_d)})$ $p \in (1,\infty)$	bounded	$2^{1+1/\bar{p}}\, s(G_d)\, r_d$	$n^{-1/\bar{q}}$
$(\mathcal{M}(\Omega_d), \|.\|_{\mathcal{M}(\Omega_d)})$	subset of the set of char. functions with finite VC-dim.	$6\sqrt{3}\, r_d\, s(G_d)\, \left(h^*_{G_d}\right)^{1/2}$	$(\log n)^{1/2}\ n^{-1/2}$

denote the set of d-variable Gaussian RBFs with width $b > 0$ and all possible centers, where τ_y denotes the translation operator defined for any $g : \mathbb{R}^d \to \mathbb{R}$ as

$$\tau_y(g)(x) := g(x-y).$$

By translation invariance of Lebesgue measure, translation has no effect on norm in $\mathcal{L}^p(\mathbb{R}^d)$, $p \in [1,\infty]$. We use

$$G_d^\gamma := \bigcup_{b>0} G_d^\gamma(b).$$

to denote the set of Gaussians with varying widths. Since $\int_{-\infty}^{\infty} e^{-u^2} du = \sqrt{\pi}$ (e.g., [46, p. 174]), a simple calculation shows that for any $b > 0$ and $q \in (0,\infty)$,

$$\|\gamma_{d,b}\|_{\mathcal{L}^q(\mathbb{R}^d)} = (\pi/qb)^{d/2q}.$$

The next corollary gives an upper bound on deviation of balls in $G_d^\gamma(b)$-variation with respect to the ambient space $\mathcal{L}^2(\mathbb{R}^d)$. For $q \in (1,\infty)$, similar bounds hold in $\mathcal{L}^q(\mathbb{R}^d)$ using Corollary 1(ii).

Corollary 3: Let d be a positive integer, $b > 0$, and $r_d > 0$. Then for every positive integer n,

$$\delta(B_{r_d}(\|\cdot\|_{G_d^\gamma(b),\mathcal{L}^2(\mathbb{R}^d)}), \mathrm{span}_n\, G_d^\gamma(b))_{\mathcal{L}^2(\mathbb{R}^d)} \leq r_d \left(\frac{\pi}{2b}\right)^{d/4} n^{-1/2}.$$

Proof: Use Corollary 1(i) to approximate by Gaussians with fixed widths. ■

In the upper bound from Corollary 3, $\xi(d) = (\pi/2b)^{d/4}\, r_d$. Thus for $b = \pi/2$, the estimate implies tractability for r_d growing with d polynomially, while for $b > \pi/2$, it implies tractability even when r_d increases exponentially fast. Hence, the width b of Gaussians has a strong impact on the size of radii r_d of balls in $G_d^\gamma(b)$-variation for which $\xi(d)$ is a polynomial. The narrower the Gaussians, the larger the balls for which Corollary 3 implies tractability.

Unlike $G_d^\gamma(b)$, the set G_d^γ of *all* Gaussians (with varying widths) is not bounded in $\mathcal{L}^2(\mathbb{R}^d)$. Thus, variation with respect to this set is not defined. Nevertheless, Corollary 3 provides a description of sets that can be tractably approximated using these dictionaries. For a subset G of $\mathcal{L}^2$, let G^o denote the set of its normalized elements, i.e.,

$$G^o = \{g^o \,|\, 0 \neq g \in G\}, \text{ where } g^o := g/\|g\|_{\mathcal{L}^2}.$$

As $\mathrm{span}_n G = \mathrm{span}_n G^o$, Corollary 3 implies that balls in $G_d^{\gamma^o}$-variation with suitable radii can be tractably approximated by Gaussian radial-basis networks of varying widths and centers.

To describe subsets of balls in $G_d^{\gamma^o}$-variation, we combine Theorem 1 with a representation of functions from Sobolev spaces as integrals of Gaussians. We need some machinery. For $m > 0$, the *Bessel potential of order* m on $\mathbb{R}^d$ is the unique function $\beta_{d,m}$ with Fourier transform

$$\hat{\beta}_{d,m}(\omega) = (1 + \|\omega\|^2)^{-m/2}$$

where we parameterize the *Fourier transform* $\mathcal{F}(f) := \hat{f}$ as

$$\hat{f}(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{ix\cdot\omega} dx.$$

For $m > 0$ and $q \in [1,\infty)$, let

$$L^{q,m}(\mathbb{R}^d) := \{f \,|\, f = w * \beta_{d,m},\, w \in \mathcal{L}^q(\mathbb{R}^d)\}$$

be the *Bessel potential space* which is formed by convolutions of functions from $\mathcal{L}^q(\mathbb{R}^d)$ with $\beta_{d,m}$. For $m > 0$, it is known that $\beta_{d,m}$ is non-negative, radial, exponentially decreasing at infinity, analytic except at the origin, and belongs to $\mathcal{L}^1(\mathbb{R}^d)$ [47, p. 296]. There is a norm defined by

$$\|f\|_{L^{q,m}(\mathbb{R}^d)} := \|w_f\|_{\mathcal{L}^q(\mathbb{R}^d)} \quad \text{for} \quad f = w_f * \beta_{d,m}.$$

Since, for our parameterization, the Fourier transform of a convolution of two functions is $(2\pi)^{d/2}$ times the product of the transforms, we have $\hat{w_f} = (2\pi)^{-d/2}\hat{f}/\hat{\beta}_{d,m}$. Thus, w_f is uniquely determined by f and so the Bessel potential norm is well-defined.

For $q \in [1,\infty)$ and *integer* $m > 0$, the *Sobolev space* $\mathcal{W}^{q,m}(\mathbb{R}^d)$ is the set of functions having tth order partial derivatives in $\mathcal{L}^q(\mathbb{R}^d)$ for all $t \leq m$, with norm given by

$$\|f\|_{\mathcal{W}^{q,m}(\mathbb{R}^d)} := \left(\sum_{|\alpha|\leq m} \|D^\alpha f\|^q_{\mathcal{L}^q(\mathbb{R}^d)}\right)^{1/q}$$

where α denotes a multi-index (i.e., a vector of non-negative integers), D^α the corresponding partial derivative operator, and $|\alpha| = \alpha_1 + \cdots + \alpha_d$.

It is well-known (e.g., [48, pp. 134–136]) that for every positive integer m and all $q \in (1, \infty)$, the Sobolev space $\mathcal{W}^{q,m}(\mathbb{R}^d)$ as a linear space is identical to the Bessel potential space $L^{q,m}(\mathbb{R}^d)$, and their norms are equivalent in the sense that each is bounded by a multiple of the other; i.e., they induce the same topology. As we did not find in the literature an explicit estimate of the coefficients of equivalence of these two norms, in [39] we derived the upper bound

$$\|f\|_{L^{2,m}(\mathbb{R}^d)} \leq (2\pi)^{-d/2}(m!)^{1/2}\|f\|_{\mathcal{W}^{2,m}(\mathbb{R}^d)}.$$

For the reader's convenience, we derive a well-known integral formula (e.g., [48, p. 132]) using our parameterization of the Fourier transform [see (6)]. Let $d \in \mathbb{N}_+$, $m > 0$; for $z > 0$, let $\Gamma(z) := \int_0^\infty t^{z-1}e^{-t}\,dt$. Then we have

$$\hat{\beta}_{d,m}(\omega) = \frac{1}{\Gamma(m/2)}\int_0^\infty u^{m/2-1}\,e^{-u}\gamma_{d,u}(\omega)\,du. \quad (4)$$

Indeed, for any $\omega \in \mathbb{R}^d$, let $I_\omega = I := \int_0^\infty u^{m/2-1}\,e^{-u(1+\|\omega\|^2)}\,du$ and set $v = u(1+\|\omega\|^2)$. Then $I = (1+\|\omega\|^2)^{-m/2}\int_0^\infty v^{m/2-1}\,e^{-v}\,dv = \hat{\beta}_m(\omega)\Gamma(m/2)$. The Fourier transform of the Gaussian function is a scaled Gaussian multiplied by a scalar; for our parameterization, for every $b > 0$

$$\widehat{\gamma_{d,b}}(\omega) = (2b)^{-d/2}\gamma_{d,1/4b}(\omega). \quad (5)$$

By (4) and (5) with $b = u$, linearity and continuity of inverse Fourier transform, one obtains

$$\begin{aligned}\beta_{d,m}(x) &= \frac{2^{-d/2}}{\Gamma(m/2)}\int_0^\infty u^{\frac{m-d}{2}-1}\,e^{-u}\,\gamma_{d,1/4u}(x)\,du\\ &= \int_0^\infty v_m(u)\,\gamma^o_{d,1/4u}(x)\,du \quad (6)\end{aligned}$$

where

$$v_m(u) := c_1(m,d)\,e^{-u}\,u^{m/2-d/4-1} \quad (7)$$

and

$$c_1(m,d) := (\pi/2)^{d/4}/\Gamma(m/2).$$

The next theorem is the rigorous form of an idea in [49].

Theorem 4: Let $m > 0, d \in \mathbb{N}_+$, $q \in [1, \infty)$. Then every $f \in L^{q,m}(\mathbb{R}^d)$ can be represented as

$$f(x) = \int_{\mathbb{R}^d}\int_{\mathbb{R}_+} w_f(y)\,v_m(u)\,\gamma^o_{d,1/4u}(x-y)\,du\,dy$$

where v_m is as in (7) and w_f is the unique function in $\mathcal{L}^q(\mathbb{R}^d)$ such that $f = \beta_{d,m} * w_f$.

Proof: By definition of Bessel potential space, every $f \in L^{q,m}(\mathbb{R}^d)$ can be represented as $f(x) = \int_{\mathbb{R}^d} w_f(y)\,\beta_{d,m}(x-y)\,dy$. By (6), we are done. ■

By the equivalence mentioned above, the same representation holds for $f \in W^{q,m}$ for $m, d \in \mathbb{N}_+$, $q \in (1, \infty)$. Using this representation of sufficiently smooth functions as integrals of normalized Gaussians, the next Corollary provides a description of sets of functions which can be tractably approximated by Gaussian RBF networks.

Corollary 4: Let d and n be positive integers. Then

$$\begin{aligned}(i)\quad &\delta\left(B_{r_d}(\|\cdot\|)_{L^{1,m}(\mathbb{R}^d)} \cap L^{2,m}(\mathbb{R}^d),\ \mathrm{span}_n\, G_d^\gamma\right)_{\mathcal{L}^2(\mathbb{R}^d)}\\ &\leq \left(\frac{\pi}{2}\right)^{d/4}\frac{\Gamma(m/2-d/4)}{\Gamma(m/2)}\,r_d\,n^{-1/2}\end{aligned}$$

where $m > d/2$;

$$\begin{aligned}(ii)\quad &\delta\left(B_{r_d}(\|\cdot\|)_{L^{1,m}(\mathbb{R}^d)} \cap L^{q,m}(\mathbb{R}^d),\ \mathrm{span}_n\, G_d^\gamma\right)_{\mathcal{L}^q(\mathbb{R}^d)}\\ &\leq \left(\frac{\pi}{2}\right)^{d/2q}\frac{\Gamma(m/2-d/2q)}{\Gamma(m/2)}\,r_d\,2^{1+1/a}\,n^{-1/b}\end{aligned}$$

where $q \in (1, \infty)$, $a = \min(q, \frac{q}{q-1})$, $b = \max(q, \frac{q}{q-1})$, and $m > d/q$.

Proof: (i) Let $f \in B_{r_d}(\|\cdot\|_{L^{1,m}(\mathbb{R}^d)} \cap L^{2,m}(\mathbb{R}^d)$. By Theorems 4 and 1, we have

$$\|f\|_{G_d^{\gamma^o},\mathcal{L}^2(\mathbb{R}^d)} \leq \int_{\mathbb{R}^d}\int_{\mathbb{R}_+}|w_f(y)|\,|v_m(t)|dt\,dy.$$

For $m > d/2$ and v_m as in (7), we have the following (see [34])

$$\|v_m\|_{\mathcal{L}^1(\mathbb{R}_+)} = \int_0^\infty v_m(u)\,du = \frac{(\pi/2)^{d/4}\Gamma(m/2-d/4)}{\Gamma(m/2)}.$$

Thus

$$\begin{aligned}\|f\|_{G_d^{\gamma^o},\,\mathcal{L}^2(\mathbb{R}^d)} &\leq \left(\frac{\pi}{2}\right)^{d/4}\frac{\Gamma(m/2-d/4)}{\Gamma(m/2)}\|w_f\|_{\mathcal{L}^1(\mathbb{R}^d)}\\ &= \left(\frac{\pi}{2}\right)^{d/4}\frac{\Gamma(m/2-d/4)}{\Gamma(m/2)}\|f\|_{L^{1,m}(\mathbb{R}^d)}.\end{aligned}$$

Then the statement follows by Corollary 1(i) since $\mathrm{span}_n G_d^\gamma = \mathrm{span}_n G_d^{\gamma^o}$ and $\|f\|_{L^{1,m}} \leq r_d$.

(ii) Let $g^{o,q}$ denote $g/\|g\|_{\mathcal{L}^q}$, $1 < q < \infty$ (normalizing w.r.t. $\mathcal{L}^q$ instead of $\mathcal{L}^2$). Then we have

$$\beta_{d,m}(x) = \int_0^\infty v_{m,q}(u)\,\gamma^{o,q}_{d,1/4u}(x)\,du \quad (8)$$

where

$$v_{m,q}(u) := c_{1,q}(m,d)\,e^{-u}\,u^{m/2-d/2q-1}$$

and

$$c_{1,q}(m,d) := (\pi/2)^{d/2q}/\Gamma(m/2).$$

For $m > d/q$, $\|v_{m,q}\|_{\mathcal{L}^1(\mathbb{R}_+)} = (\pi/2)^{d/2q}\Gamma(m/2 - d/2q)/\Gamma(m/2)$ so (ii) follows by Corollary 1(ii). ■

For every $m > d/2$, the upper bound from Corollary 4(i) on the worst-case error in approximation by Gaussian-basis-function networks is of the factorized form $\xi(d)\kappa(n)$, where $\kappa(n) = n^{-1/2}$ and

$$\xi(d) = r_d\,\left(\frac{\pi}{2}\right)^{d/4}\frac{\Gamma(m/2-d/4)}{\Gamma(m/2)}.$$

Let $h > 0$ and put $m_d = d/2 + h$. Then $\xi(d)/r_d = \left(\frac{\pi}{2}\right)^{d/4}\frac{\Gamma(h/2)}{\Gamma(h/2+d/4)}$, which goes to zero exponentially fast with

TABLE II
ESTIMATES PROVIDED BY COROLLARIES 3 AND 4 FOR GAUSSIAN RBF NETWORKS

ambient space	dictionary	approximated functions	$\xi(d)$	$\kappa(n)$
$(\mathcal{L}_2(\mathbb{R}^d), \|\cdot\|_{\mathcal{L}^2(\mathbb{R}^d)})$	$G_d^\gamma(b)$	$B_{r_d}(\|\cdot\|_{G_d^\gamma(b)})$	$r_d\,\left(\frac{\pi}{2b}\right)^{d/4}$	$n^{-1/2}$
$(\mathcal{L}_2(\mathbb{R}^d), \|\cdot\|_{\mathcal{L}^2(\mathbb{R}^d)})$	G_d^γ	$B_{r_d}(\|\cdot\|_{L^{1,m}}) \cap L^{2,m}$	$\left(\frac{\pi}{2}\right)^{d/4} \frac{\Gamma(m/2-d/4)}{\Gamma(m/2)}\, r_d$	$n^{-1/2}$

increasing d. So for $h > 0$ and $m_d \geq d/2 + h$, the approximation problem (2) is hyper-tractable for $\mathcal{X}_d = \mathcal{L}^2(\mathbb{R}^d)$, $A_d = B_{r_d}(\|\cdot\|_{L^{1,m_d}(\mathbb{R}^d)}) \cap L^{2,m_d}(\mathbb{R}^d)$, and $G_d = G_d^\gamma$.

We now replace the Bessel potential by a general kernel. Let $\Omega \subseteq \mathbb{R}^d$ and $K : \Omega \times \Omega \to \mathbb{R}$ be a kernel. In the following, we state a general estimate for families of functions defined by convolution of a bounded kernel with an absolutely integrable function. Let $\|w\|_{\mathcal{L}^1(\Omega)} := \int_\Omega |f(x)|dx$. For $r > 0$,

$$\begin{aligned} A_r^K(\Omega) \quad &:= \quad \Big\{ f : \Omega \to \mathbb{R} \,|\, f(x) \\ &= \quad \int_\Omega K(x,t)\, w(t)\, dt, \;\; \|w\|_{\mathcal{L}^1(\Omega)} \leq r \Big\}. \end{aligned}$$

If K is bounded on Ω and $w \in \mathcal{L}^1(\Omega)$, the integrals $\int_\Omega K(x,t)\, w(t)\, dt$ are finite for each x. Let

$$G_d^K := \{K(\cdot, y) \,|\, y \in \Omega_d\}; \quad h_{d,K} := VC\left(G_d^K\right)$$

where $VC(\,\cdot\,)$ denotes VC-dimension. (See after Theorem 1.)

For K such a bounded kernel, in [49] tools from statistical learning theory were utilized to show that there exists $C = C(K, \Omega)$ such that

$$\begin{aligned} &\delta\left(A_{r_d}^K(\Omega), \operatorname{span}_n G_d^K\right)_{\mathcal{M}(\Omega)} \\ &\leq C\, r_d \left(h_{d,K} \ln \frac{2\,e\,n}{h_{d,K}} + \ln 4 \right)^{1/2} n^{-1/2}. \qquad (9) \end{aligned}$$

For $\Omega = \mathbb{R}^d$, the result was applied in [49] to Bessel and Gaussian kernels. However, the bound (9) and its improvements from [50], [51] are not in the factorized form $\xi(d)\,\kappa(n)$. The following theorem [52, Ths. 4.5, 5.2] extends and improves the estimate (9) to a factorized form.

Theorem 5: Let $\Omega_d \subseteq \mathbb{R}^d$, $K : \Omega_d \times \Omega_d \to \mathbb{R}$ bounded and $h_{d,K}$ the VC dimension of G_d^K. Then there exists $C = C(K, \Omega_d)$ such that for all positive integers n

$$\delta\left(A_{r_d}^K(\Omega_d), \operatorname{span}_n G_d^K\right)_{\mathcal{M}(\Omega_d)} \leq C\, r_d\, h_{d,K}^{1/2}\, n^{-1/2}. \qquad (10)$$

The bound from Theorem 5 guarantees tractability when $\xi(d) = r_d\, h_{d,K}^{1/2}$ grows at most polynomially with the number d of variables.

Table II summarizes the estimates of this section.

V. TRACTABILITY OF APPROXIMATION BY PERCEPTRON NETWORKS

In this section, we investigate tractability of worst-case errors in approximation by linear combinations of perceptrons. Perceptrons with an *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$ compute functions from $\mathbb{R}^d$ to $\mathbb{R}$ given by

$$x \mapsto \sigma(v \cdot x + b)$$

where v is a *weight vector* and b is a *bias*. Typically, σ is a *sigmoid*, i.e., a measurable function such that $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$; usually, it is also assumed that σ is nondecreasing. The *Heaviside function* $\vartheta : \mathbb{R} \to \mathbb{R}$, defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$, is a sigmoid.

Let $\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ denote the sphere in $\mathbb{R}^d$. For a sigmoid σ let

$$H_d^\sigma := \{x \mapsto \sigma(v \cdot x + b) \,|\, v \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

When $\sigma = \vartheta$, we simply write H_d. Since for $t \neq 0$, $\vartheta(t) = \vartheta(t/|t|)$, one has

$$H_d = H_d^\vartheta = \{x \mapsto \vartheta(e \cdot x + b) \,|\, e \in \mathcal{S}^{d-1}, b \in \mathbb{R}\}$$

so H_d is the set of characteristic functions of closed half-spaces of $\mathbb{R}^d$. One also calls H_d-variation *variation with respect to half-spaces* [17].

For any family $\mathcal{F}$ of functions on R^d and $\Omega \subseteq \mathbb{R}^d$, let

$$\mathcal{F}|_\Omega := \{f|_\Omega \,|\, f \in \mathcal{F}\}$$

where $f|_\Omega$ is the restriction of f to Ω. We also use the phrase "variation with respect to half-spaces" for the restrictions of H_d. For simplicity, we sometimes write H_d instead of $H_d|_\Omega$.

Remark 1: When $\Omega_d \subset \mathbb{R}^d$ has finite Lebesgue measure, for any continuous nondecreasing sigmoid σ variation with respect to half-spaces is equal to $H_d^\sigma|_{\Omega_d}$-variation in $\mathcal{L}^2(\Omega_d)$ [53], i.e.,

$$\|\cdot\|_{H_d^\sigma|_{\Omega_d}} = \|\cdot\|_{H_d|_{\Omega_d}}\; \sigma \text{ continuous nondecreasing sigmoid.}$$

Hence, investigating tractability of balls in variation with respect to half-spaces has implications for approximation by perceptron networks with arbitrary continuous nondecreasing sigmoids. For simplicity, in Corollaries 5 and 6, Theorem 7, and Table III, we state the estimates only for the dictionary H_d, but when $\Omega_d \subset \mathbb{R}^d$ has finite Lebesgue measure, the bounds hold

TABLE III
ESTIMATES PROVIDED BY COROLLARIES 5 AND 6 AND THEOREM 7 FOR PERCEPTRON NETWORKS. WHEN $\Omega_d \subset \mathbb{R}^d$ HAS FINITE LEBESGUE MEASURE, THE BOUNDS HOLD ALSO FOR THE DICTIONARY H_d^σ WITH ANY CONTINUOUS NONDECREASING SIGMOID σ

ambient space	dictionary G_d	target set $\mathcal{F}$ to be approx.	$\xi(d)$	$\kappa(n)$
$(\mathcal{M}(\Omega), \|\cdot\|_{\mathcal{M}(\Omega)})$	$H_d(\Omega)$	$B_{r_d}(\|\cdot\|_{H_d(\Omega),\mathcal{M}(\Omega)})$	$6\sqrt{3}\, r_d\, d^{1/2}$	$(\log\, n)^{1/2} n^{-1/2}$
$(\mathcal{L}^2(\Omega_d), \|\cdot\|_{\mathcal{L}^2(\Omega_d)})$	$H_d\|_{\Omega_d}$	$B_{r_d}(\|\cdot\|_{H_d\|_{\Omega_d},\mathcal{L}^2(\Omega_d)})$	$\lambda(\Omega_d)\, r_d$	$n^{-1/2}$
$(\mathcal{L}^2(\Omega_d), \|\cdot\|_{\mathcal{L}^2(\Omega_d)})$ $\Omega_d \subset \mathbb{R}^d$, d odd	$H_d\|_{\Omega_d}$	A_{r_d}	$k_d \lambda(\Omega_d)^{1/2}\, r_d$	$n^{-1/2}$
$(\mathcal{L}^2(\Omega_d), \|\cdot\|_{\mathcal{L}^2(\Omega_d)})$ $\Omega_d \subset \mathbb{R}^d$, d odd	$H_d\|_{\Omega_d}$	$G_d^{\gamma,1}\|_{\Omega_d}$	$(2\,\pi\, d)^{3/4}\, \lambda(\Omega_d)^{1/2}$	$n^{-1/2}$

also for the dictionary H_d^σ with any continuous nondecreasing sigmoid σ.

The next corollary estimates deviation of balls in variation with respect to half-spaces.

Corollary 5: Let d be a positive integer, $\Omega_d \subseteq \mathbb{R}^d$. Then for every positive integer n

$$(i)\quad \delta(B_{r_d}(\|\cdot\|_{H_d,\mathcal{M}(\Omega_d)}), \mathrm{span}_n\, H_d)_{\mathcal{M}(\Omega_d)} \leq 6\sqrt{3}\, d^{1/2}\, r_d\, (\log\, n)^{1/2}\, n^{-1/2};$$

(ii) if Ω_d has finite Lebesgue measure, then

$$\delta\left(B_{r_d}\left(\|\cdot\|_{H_d,\mathcal{L}^2(\Omega_d)}\right), \mathrm{span}_n\, H_d\right)_{\mathcal{L}^2(\Omega_d)} \leq \lambda(\Omega_d)\, r_d\, n^{-1/2}.$$

Proof: The coVC-dimension of the set of characteristic functions of half-spaces of $\mathbb{R}^d$ is equal to d [23, p. 162]. Thus, the statement follows by Corollaries 2 and 1. ■

Corollary 5 implies that approximation of functions from balls of radii r_d in variation with respect to half-spaces is tractable in $\mathcal{M}(\mathbb{R}^d)$ when r_d grows polynomially. In $(\mathcal{L}^2(\Omega_d), \|\cdot\|_{\mathcal{L}^2(\Omega_d)})$, this approximation is tractable when r_d times $\lambda(\Omega_d)$ grows polynomially with d. If for all $d \in \mathbb{N}'$, Ω_d is the unit ball in $\mathbb{R}^d$, then this approximation is hyper-tractable unless r_d is exponentially growing.

It is shown in [39] that functions with continuous dth order partials that either are compactly supported or, together with their derivatives, have sufficiently rapid decay at infinity, can be expressed as networks with infinitely many Heaviside perceptrons and so, by Theorem 1, their variation with respect to half-spaces is bounded above by the $\mathcal{L}^1$-norm of the output weight function.

A real-valued function f on $\mathbb{R}^d$, d odd, is of *weakly-controlled decay* [39] if f is d-times continuously differentiable and for all multi-indices $\alpha \in \mathbb{N}^d$ with $|\alpha| = \sum_{i=1}^d \alpha_i$ and $D^\alpha = \partial^{\alpha_1} \cdot \ldots \cdot \partial^{\alpha_d}$,

$$\left(|\alpha| < d \Rightarrow \lim_{\|x\|\to\infty} D^\alpha f(x) = 0\right)$$

and

$$\left(|\alpha| = d \Rightarrow \exists \varepsilon > 0 \setminus \lim_{\|x\|\to\infty} D^\alpha f(x) \|x\|^{d+1+\varepsilon} = 0\right).$$

We denote by $\mathcal{V}(\mathbb{R}^d)$ the set of functions of weakly controlled decay on $\mathbb{R}^d$. This set includes the Schwartz class of smooth functions rapidly decreasing at infinity as well as the class of d-times continuously differentiable functions with compact support. In particular, it includes the Gaussian function. Also, if $f \in \mathcal{V}(\mathbb{R}^d)$, then $\|D^\alpha f\|_{\mathcal{L}^1(\mathbb{R}^d)} < \infty$ if $|\alpha| = d$. The maximum over all α with $|\alpha| = d$ is called the *Sobolev seminorm* of f and is denoted $\|f\|_{d,1,\infty}$.

The following theorem from [54] gives an integral representation of smooth functions as networks with infinitely many Heaviside perceptrons. The output weight function w_f can be interpreted as a flow of the order d through the hyperplane $H_{e,b} = \{x \in \mathbb{R}^d \,|\, x \cdot e + b = 0\}$ scaled by $a(d)$, which goes to zero exponentially fast with d increasing. By $D_e^{(d)}$ is denoted the directional derivative of the order d in the direction e.

Theorem 6: Let $d \in \mathbb{N}_+$ be odd, $f \in \mathcal{V}(\mathbb{R}^d)$ of weakly-controlled decay. Then for every $x \in \mathbb{R}^d$

$$f(x) = \int_{\mathcal{S}^{d-1}\times\mathbb{R}} w_f(e,b)\, \vartheta(e\cdot x + b)\, de\, db$$

where $w_f(e,b) = a(d) \int_{H_{e,b}} D_e^{(d)}(f)(y)\, dy$ and $a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$.

The representation of Theorem 6 was first derived in [55] (see Th. 3.1, Prop. 2.2, and an equation in [55, p. 387]) using the Radon transform (see, e.g., [56, p. 251]) for all functions from the Schwartz class. In [53], the same formula was derived for all compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$ with d odd, via an integral formula for the Dirac delta function. In [54], the representation was extended to functions of weakly-controlled decay. Representation of f as a network with infinitely many perceptrons also holds for d even, but the output weight function is more complicated (see [55] for the case when f is in the Schwartz class and [54] for the case of f satisfying certain milder conditions on smoothness and behavior at infinity).

Let A_{r_d} denote the intersection of $\mathcal{V}(\mathbb{R}^d)$ with the ball $B_{r_d}(\|\cdot\|_{d,1,\infty})$ of radius r_d in the Sobolev seminorm $\|\cdot\|_{d,1,\infty}$. Then

$$A_{r_d} = \mathcal{V}(\mathbb{R}^d) \cap B_{r_d}(\|\cdot\|_{d,1,\infty}) = r_d\, A_1.$$

Theorem 7: Let $d \in \mathbb{N}_+$ be odd, $\Omega_d \subset \mathbb{R}^d$ with $\lambda(\Omega_d) < \infty$, and $k_d = 2^{1-d}\pi^{1-d/2}d^{d/2}/\Gamma(d/2)$. Then for every positive integer n

$$\delta(A_{r_d}|_{\Omega_d}, \mathrm{span}_n\, H_d|_{\Omega_d})_{\mathcal{L}^2(\Omega_d)} \le k_d \lambda(\Omega_d)^{1/2}\, r_d\, n^{-1/2}.$$

Proof: Let $f \in A_{r_d}$. If Ω_d has finite Lebesgue measure, then

$$\|f|_{\Omega_d}\|_{H_d|_{\Omega_d}, \mathcal{L}^2(\Omega_d)} \le \|f|_{\Omega_d}\|_{H_d|_{\Omega_d}, \mathcal{M}(\Omega_d)} \le \|f\|_{H_d, \mathcal{M}(\mathbb{R}^d)}.$$

Indeed, the first inequality is our remark after the definition of G-variation and the second is a similar formality. Combining the integral representation of Theorem 6 and the consequent bound on variational norm given by Theorem 1 gives the first inequality below; the second inequality is [39, Cor. 4.3]:

$$\|f\|_{H_d, \mathcal{M}(\mathbb{R}^d)} \le \|w_f\|_{\mathcal{L}^1} \le k_d \|f\|_{d,1,\infty}.$$

As $\sup_{g \in H_d|_{\Omega_d}} \|g\|_{\mathcal{L}^2(\Omega_d)} = \lambda(\Omega_d)^{1/2}$, the theorem follows by Corollary 1(i). ■

By the remark preceding Corollary 5.1, using [53], one can replace H_d by H_d^σ, for any continuous nondecreasing sigmoid σ in the conclusion of Theorem 7. The next corollary estimates the worst-case $\mathcal{L}^2$-errors in approximation by perceptron networks of the set

$$G_d^{\gamma,1} := \left\{\tau_y(\gamma_d) \,|\, y \in \mathbb{R}^d\right\}$$

of d-variable Gaussians with widths equal to 1 and varying centers.

Corollary 6: Let $d \in \mathbb{N}_+$ be odd, $\Omega_d \subset \mathbb{R}^d$ with $\lambda(\Omega_d) < \infty$. Then for every positive integer n

$$\delta\left(G_d^{\gamma,1}|_{\Omega_d}, \mathrm{span}_n\, H_d\right)_{\mathcal{L}^2(\Omega_d)} \le (2\,\pi\, d)^{3/4}\, \lambda(\Omega_d)^{1/2}\, n^{-1/2}.$$

Proof: Let $\mathcal{X}$ be any normed linear space of real-valued functions (or equivalence classes of functions) on $\mathbb{R}^d$. It is easy to see that for every bounded $G \subset \mathcal{X}$ closed under translation, every f, and every $y \in \mathbb{R}^d$, one has $\|\tau_y(f)\|_{G,\mathcal{X}} = \|f\|_{G,\mathcal{X}}$. This remark with $\mathcal{X} = \mathcal{M}(\mathbb{R}^d)$ and $G = H_d$ gives

$$\begin{aligned}
\|\tau_y(\gamma_d)|_{\Omega_d}\|_{H_d|_{\Omega_d}, \mathcal{L}^2(\Omega_d)} &\le \|\tau_y(\gamma_d)|_{\Omega_d}\|_{H_d|\Omega_d, \mathcal{M}(\Omega_d)} \\
&\le \|\tau_y(\gamma_d)\|_{H_d, \mathcal{M}(\mathbb{R}^d)} \\
&= \|\gamma_d\|_{H_d, \mathcal{M}(\mathbb{R}^d)}.
\end{aligned}$$

But $\|\gamma_d\|_{H_d, \mathcal{M}(\mathbb{R}^d)} \le (2\,\pi\, d)^{3/4}$ [39, Cor. 6.2]. As $\sup_{g \in H_d(\Omega_d)} \|g\|_{\mathcal{L}^2(\Omega_d)} = \lambda(\Omega_d)^{1/2}$, by Corollary 1(i) we are done. ■

In the upper bound from Corollary 6, we have $\xi(d) = (2\pi d^{3/4})\, \lambda(\Omega_d)^{1/2}$. This implies that approximation of d-variable Gaussians on a domain Ω_d by perceptron networks is tractable when the Lebesgue measure $\lambda(\Omega_d)$ grows polynomially with d, while if the domains Ω_d are unit balls in $\mathbb{R}^d$, then the approximation is hyper-tractable.

Table III contains the estimates derived in this section. Note that k_d which appears in rows 3 and 4 is exponentially decreasing since, by Stirling's approximation for the Gamma function:

$$k_d = 2^{1-d}\pi^{1-d/2}d^{d/2}/\Gamma(d/2) \sim (\pi d)^{1/2}(e/2\pi)^{d/2}.$$

Hence, if $r_d\lambda(\Omega_d)^{1/2}$ is at most polynomial, then the approximation problem is hyper-tractable. The estimates in rows 2–5 take a convenient form when all the domains Ω_d have unit volume (i.e., $\lambda(\Omega_d) = 1$). For d-dimensional *cubes*, to this end the sides must be 1, but for d-dimensional *balls* in the Euclidean norm, the radii have to be proportional to $\sqrt{d}$. Indeed, the volume of a radius ρ_d-ball in d dimensions is $\rho_d^d \pi^{d/2}/\Gamma(1 + d/2)$, e.g., [57, p. 304]; to get unit volume, by Stirling's formula, one needs a radius of

$$\begin{aligned}
\rho_d &= \left(\pi^{-d/2}\Gamma(1 + d/2)\right)^{1/d} \\
&\sim \pi^{-1/2}\frac{(1 + d/2)^{(1/d)+(1/2)}}{e^{(1/d)+(1/2)}} \sim c\sqrt{d}
\end{aligned}$$

where $c = 1/\sqrt{2e\pi} = 0.24197\ldots$. In this way, one can see that our methods allow tractable approximation when the Ω_d are balls of radii $r_d = c\sqrt{d}$ (and so $\lambda(\Omega_d) = 1$).

VI. Worst-Case Tractability for Optimization

The techniques developed in the previous sections can also be applied to optimization.

Let S_d be a nonempty subset of a normed linear space $(\mathcal{X}_d, \|\cdot\|_{\mathcal{X}_d})$ of d-variable functions and let $\Phi : \mathcal{X}_d \to \mathbb{R}$ be a proper functional. We consider the optimization problem of *minimizing* Φ *on* S_d:

$$\inf \Phi(f) \;\text{ s.t. }\; f \in S_d. \tag{11}$$

This entails an *infinite-programming problem* [58], [59], also called *functional optimization problem* [8], [9], as the admissible solutions are elements of an infinite-dimensional space [60].

When a solution to the problem (11) cannot be found in closed form, an approximate solution can be obtained by iterative methods, which entail the construction of a minimizing sequence converging to an element of the admissible set S_d. The *classical Ritz method* [61] constructs a minimizing sequence considering for every positive integer n the subproblems $\inf_{f \in \mathcal{X}_{d,n}} \Phi(f)$, where $\mathcal{X}_{d,n}$ is an n-dimensional subspace of $\mathcal{X}_d$ and so $\mathcal{X}_{d,n} \subseteq \mathcal{X}_{d,n+1}$.

For an input set $\Omega_d \subseteq \mathbb{R}^d$ and a computational unit $\phi : \mathbb{R}^q \times \Omega_d \to \mathbb{R}$, let

$$G_d^\phi := \{\phi(u, \cdot) \,|\, u \in \mathbb{R}^q\}$$

and suppose that $G_d^\phi \subset \mathcal{X}_d$ has $\mathrm{cl}_{\mathcal{X}_d} \mathrm{span}\, G_d^\phi \supseteq S_d$, where for any G contained in a linear space $\mathcal{X}$, $\mathrm{span}\, G$ is the intersection of all linear subspaces of $\mathcal{X}$ which contain G; i.e.,

$\mathrm{span}\, G := \bigcup_{n\geq 1} \mathrm{span}_n\, G$. The *extended Ritz method*, formalized in [9] and investigated in [1], [62], [63], [3], [64], considers approximate minimization over $\mathrm{span}_n\, G_d^{\phi}$, i.e.,

$$\inf \Phi(f) \;\; \text{s.t.} \quad f \in S_d \cap \mathrm{span}_n\, G_d^{\phi}. \tag{12}$$

With suitable choices of the computational unit ϕ, the optimization problem (12) formalizes the use of computational models such as radial basis function and perceptron networks for the solution of tasks in which a function that is optimal, in a sense specified by a cost functional, has to be found among a set of candidate admissible functions. Such functions may represent routing strategies in telecommunication networks, movement strategies for decision makers in a partially unknown environment, exploration strategies in graphs with stochastic costs, input/output mappings of a device that learns from examples; see, e.g., [1]–[3], [6]–[9] and references therein).

When investigating the tractability of optimization over $\mathrm{span}_n\, G_d^{\phi}$, to simplify the notations we may suppose that the problems (11) and (12) have solutions f^o and f_n^o, respectively. If the infima in (11) and (12) are not achieved, then the results can be restated, at the expense of more cumbersome notations, in terms of ε-near minimum points.

In order to approximate (11) by (12), one needs to estimate

$$\Phi\left(f_n^o\right) - \Phi(f^o) \;\; \text{and} \;\; \|f_n^o - f^o\|_{\mathcal{X}_d}.$$

Definition 4: The approximation of the problem (11) by the problem (12) is called *tractable with respect to* d *in the worst case* or simply *tractable* iff there exist $\nu > 0, \bar{\nu} > 0$ such that

$$\Phi\left(f_n^o\right) - \Phi(f^o) \leq \xi(d)\, \kappa(n)$$
$$\text{and}$$
$$\|f_n^o - f^o\|_{\mathcal{X}_d} \leq \bar{\xi}(d)\, \bar{\kappa}(n)$$

hold with $\xi(d) \leq d^{\nu}$, $\bar{\xi}(d) \leq d^{\bar{\nu}}$ for every $d \in \mathbb{N}'$ and every $n \in \mathbb{N}_+$, where $\kappa, \bar{\kappa}$ are nonincreasing and nonnegative.

For standard terminology (e.g., modulus of convexity), see [65] or [66]. Recall that the problem (11) is *Tikhonov well-posed* if it has a unique minimum to which every minimizing sequence converges [66, p. 1]. The *modulus of Tikhonov well-posedness* of the problem (11) at a minimum point f^o is the function $\varsigma_{f^o} : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varsigma_{f^o}(t) = \inf\{\Phi(g) - \Phi(f^o) \,|\, \|f - f^o\|_{\mathcal{X}_d} = t\}$ for all t.

The next theorem investigates tractability of the approximate solution of problem (11) with S_d equal to the ball of radius r_d in $(\mathcal{X}_d, \|\cdot\|_{\mathcal{X}_d})$, i.e.,

$$\inf \Phi(f) \;\; \text{s.t.} \quad \|f\|_{\mathcal{X}_d} \leq r_d \tag{13}$$

by the problems obtained from (12) with such a choice of S_d, i.e.,

$$\inf \Phi(f) \;\; \text{s.t.} \quad \|f\|_{\mathcal{X}_d} \leq r_d \quad \text{and} \quad f \in \mathrm{span}_n\, G_d^{\phi}. \tag{14}$$

Analogously with our suppositions for (11) and (12), without loss of generality we may assume that the infima in (13) and (14) are achieved at f^o and f_n^o, resp. (otherwise, ε-near minimum points have to be considered).

Theorem 8: Let $(\mathcal{X}_d, \|\cdot\|_{\mathcal{X}_d})$ be a normed linear space, $G_d^{\phi} \subset \mathcal{X}_d$, $s(G_d) := \sup_{f\in G_d} \|f\|_{\mathcal{X}_d}$, and let $\Phi : \mathcal{X}_d \to (-\infty, +\infty]$ be a proper functional, uniformly convex on $\mathcal{X}_d$ with modulus of convexity ϱ. Let ς_{f^o} be the modulus of Tikhonov well-posedness for the problem (13) at a minimum point f^o, Φ continuous at f^o with a modulus of continuity ω_{f^o}, and f_n^o the minimum point of the problem (14). If there exist $a, b > 0$ such that for all $t \geq 0$ $\omega_{f^o}(t) \leq t^a$ and $\min\{\varrho^{-1}(t), \varsigma_{f^o}^{-1}(t)\} \leq t^{1/b}$, then for every positive integer n the following hold:

(i) $\Phi(f_n^o) - \Phi(f^o) \leq (2\, s(G_d)\, r_d)^a\, n^{-a/2}$;

(ii) $\|f_n^o - f^o\|_{\mathcal{X}_d} \leq (s(G_d)\, r_d)^{a/b}\, n^{-a/2b}$.

Proof: (i) As the Minkowski functional of the ball $B_{r_d}(\|\cdot\|_{\mathcal{X}_d})$ is equal to $1/r_d$, by [3, Th. 4.2] (i) with $c = 1/r_d$ we get

$$\begin{aligned}\Phi(f_n^o) - \Phi(f^o) \quad &\leq \quad \omega_{f^o}\left(2\, s(G_d)\, r_d\, n^{-1/2}\right)\\ &\leq \quad (s(G_d)\, r_d)^a\, n^{-a/2}.\end{aligned}$$

(ii) By [3, Th. 4.2] (ii)–(iii) we have

$$\begin{aligned}\|f_n^o - f^o\|_{\mathcal{X}_d} \quad &\leq \quad \min\Big\{\varrho^{-1}\left((s(G_d)\, r_d)^a\, n^{-a/2}\right),\\ &\qquad\quad \varsigma_{f^o}^{-1}\left((s(G_d)\, r_d)^a\, n^{-a/2}\right)\Big\}\\ &\leq \quad (s(G_d)\, r_d)^{a/b}\, n^{-a/2b}.\end{aligned}$$

■

Thus, the problem of minimization of functionals by approximation schemes $\mathrm{span}_n\, G_d^{\phi}$ is tractable provided that the moduli of continuity, convexity, and well-posedness satisfy suitable constraints.

For example, Theorem 8 can be applied to the optimization problem associated with a sample $z := \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, \,|\, i = 1, \ldots, m\}$ of empirical data, modeled as minimization over perceptron networks or Gaussian RBF networks, of the *empirical error functional* [67]–[69]

$$\mathcal{E}_z(f) := \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

Hence, tractability can be investigated using Theorem 8 and empirical error.

VII. Remarks

Several authors [17], [13], [70], [25] derived tight improvements of the factor $\kappa(n) = n^{-1/2}$ for various dictionaries G. In the case of orthonormal dictionaries, tight bounds were established in [13], [28]. In [13] it was shown that for a general dictionary, the factor $\kappa(n) = n^{-1/2}$ cannot be substantially improved; in particular, for such dictionaries improvement is at best to $(1/2)\,(n-1)^{-1/2}$.

For perceptron networks with certain sigmoidal functions, the impossibility of improving the factor $n^{-1/2}$ in the estimate of Theorem 2 (i) over $n^{-1/2-1/d}$ was proven in [17] via a probabilistic argument and in [25] via estimates of covering numbers. The term $n^{-1/2-1/d}$ cannot be expressed in a factorized form, as the dependencies on d and n cannot be separated, but for every integer $n \geq 1$ and every positive integer d one has $n^{-1/2-1/d} < n^{-1/2}$, so to investigate tractability the extra term in the exponent can be neglected. In [70], the tightness result derived in [25] was extended to every dictionary G with (i) certain properties of its covering numbers and (ii) a sufficient "capacity" of its symmetric convex hull $\mathrm{conv}\,(G \cup -G)$.

In some cases (see Section III), the function $\xi(d)$ in the factorized estimate contains the G-variation norm. Examples of functions with variation with respect to Heaviside perceptrons growing exponentially with the number of variables d were given in [28]. However, such exponential lower bounds on variation with respect to half-spaces are only lower bounds to an upper bound on rates of approximation. Finding whether these exponentially large upper bounds are tight seems to be a difficult task related to some open problems in the theory of complexity of Boolean circuits [28].

Finally, we address the significance of our results. In several cases, given sequences of target spaces, we found that approximation is hyper-tractable. That is, even with $n = 1$, one can well approximate once d is sufficiently large. To be approximable by a single member $g \in G$ (a single hidden-unit) means that the distance from f to $\mathrm{span}_1\, g$ is small. The easiest way for this to happen is if f is near zero. But interesting functions such as the Gaussian can't be approximated with only one unit, so one sees that, in high-dimensional situations, ambient function-space norms are likely to be astronomically big. Only functions very near zero can be in the unit-ball. But what is not to be expected is that a reasonable function such as the unit-width Gaussian has $\xi(d)$ growing at less than a linear rate [39].

References

[1] S. Giulini and M. Sanguineti, "Approximation schemes for functional optimization problems," *J. Optim. Theory Applicat.*, vol. 140, pp. 33–54, 2009.

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

[3] V. Kůrková and M. Sanguineti, "Learning with generalization capability by kernel methods of bounded complexity," *J. Complexity*, vol. 21, pp. 350–367, 2005.

[4] V. Kůrková and M. Sanguineti, "Approximate minimization of the regularized expected error over kernel models," *Math. Oper. Res.*, vol. 33, pp. 747–756, 2008.

[5] V. Kůrková and M. Sanguineti, "Geometric upper bounds on rates of variable-basis approximation," *IEEE Trans. Inf. Theory*, vol. 54, pp. 5681–5688, 2008.

[6] K. S. Narendra and S. Mukhopadhyay, "Adaptive control using neural networks and approximate models," *IEEE Trans. Neural Netw.*, vol. 8, pp. 475–485, 1997.

[7] K. A. Smith, "Neural networks for combinatorial optimization: A review of more than a decade of research," *Informs J. Comput.*, vol. 11, pp. 15–34, 1999.

[8] R. Zoppoli, T. Parisini, M. Sanguineti, and M. Baglietto, *Neural Approximations for Optimal Control and Decision*. London, U.K.: Springer, in preparation.

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

[10] G. Gnecco, V. Kůrková, and M. Sanguineti, "Can dictionary-based computational models outperform the best linear ones?," *Neural Netw.*, 2011, doi:10.1016/j.neunet.2011.05.014.

[11] G. Gnecco, V. Kůrková, and M. Sanguineti, "Some comparisons of complexity in dictionary-based and linear computational models," *Neural Netw.*, vol. 24, pp. 171–182, 2011.

[12] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Trans. Inf. Theory*, vol. 52, pp. 255–261, 2006.

[13] V. Kůrková and M. Sanguineti, "Bounds on rates of variable-basis and neural-network approximation," *IEEE Trans. Inf. Theory*, vol. 47, pp. 2659–2665, 2001.

[14] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*. Berlin, Heidelberg, Germany: Springer, 1970.

[15] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2d ed. Upper Saddle River, NJ: Prentice-Hall, 1998.

[16] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[17] A. R. Barron, "Neural net approximation," in *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, K. Narendra, Ed., New Haven, CT, 1992, pp. 69–72.

[18] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, pp. 930–945, 1993.

[19] L. Breiman, "Hinging hyperplanes for regression, classification and function approximation," *IEEE Trans. Inf. Theory*, vol. 39, pp. 999–1013, 1993.

[20] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, "Rate of approximation results motivated by robust neural network learning," in *Proc. 6th Annu. ACM Conf. Computational Learning Theory*, New York, 1993, pp. 303–309.

[21] M. Donahue, L. Gurvits, C. Darken, and E. Sontag, "Rates of convex approximation in non-Hilbert spaces," *Construct. Approx.*, vol. 13, pp. 187–220, 1997.

[22] F. Girosi and G. Anzellotti, "Rates of convergence for radial basis functions and neural networks," in *Artif. Neural Netw. Speech Vis.*, R. J. Mammone, Ed. London, U.K.: Chapman & Hall, 1993, pp. 97–113.

[23] L. Gurvits and P. Koiran, "Approximation and learning of convex superpositions," *J. Comput. Syst. Sci.*, vol. 55, pp. 161–170, 1997.

[24] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, pp. 608–613, 1992.

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

[26] D. E. Knuth, "Big omicron and big omega and big theta," *SIGACT News*, vol. 8, pp. 18–24, 1976.

[27] V. Kůrková, "Minimization of error functionals over perceptron networks," *Neural Comput.*, vol. 20, pp. 252–270, 2008.

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

[29] V. Kůrková and M. Sanguineti, "Comparison of worst-case errors in linear and neural network approximation," *IEEE Trans. Inf. Theory*, vol. 28, pp. 264–275, 2002.

[30] G. W. Wasilkowski and H. Woźniakowski, "Complexity of weighted approximation over $\mathbb{R}^d$," *J. Complexity*, vol. 17, pp. 722–740, 2001.

[31] J. F. Traub and A. G. Werschulz, *Complexity and Information*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[32] H. Woźniakowski, "Tractability and strong tractability of linear multivariate problems," *J. Complexity*, vol. 10, pp. 96–128, 1994.

[33] H. N. Mhaskar, "On the tractability of multivariate integration and approximation by neural networks," *J. Complexity*, vol. 20, pp. 561–590, 2004.

[34] P. C. Kainen, V. Kůrková, and M. Sanguineti, "Complexity of Gaussian radial basis networks approximating smooth functions," *J. Complexity*, vol. 25, pp. 63–74, 2009.

[35] A. Pinkus, *n-Widths in Approximation Theory*. Berlin, Heidelberg, Germany: Springer, 1985.

[36] V. Maiorov, "On best approximation by ridge functions," *J. Approx. Theory*, vol. 99, pp. 68–94, 1999.

[37] P. C. Kainen, V. Kůrková, and M. Sanguineti, "On tractability of neural-network approximation," in *Lecture Notes in Computer Science*, M. Kolehmainen, P. Toivanen, and B. Beliczynski, Eds., Berlin, Heidelberg, Germany, 2009, vol. 5495, pp. 11–21, (Proc. ICANNGA 2009). Springer.

[38] B. Beliczyinski and B. Ribeiro, "Some enhancement to approximation of one-variable functions by orthonormal basis," *Neural Netw. World*, vol. 19, pp. 401–412, 2009.

[39] P. C. Kainen, V. Kůrková, and A. Vogt, "A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves," *J. Approx. Theory*, vol. 147, pp. 1–10, 2007.

[40] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*. New York: Dover, 1970.

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

[42] V. Kůrková and M. Sanguineti, "Error estimates for approximate optimization by the extended Ritz method," *SIAM J. Optim.*, vol. 15, pp. 461–487, 2005.

[43] V. Kůrková, "Model complexity of neural networks and integral transforms," in *Lecture Notes in Computer Science*, M. Polycarpou, C. Panayiotou, C. Alippi, and G. Ellinas, Eds. Berlin Heidelberg: Springer, 2009, vol. 5768, pp. 708–718, (Proc. ICANN 2009).

[44] P. C. Kainen and V. Kůrková, "An integral upper bound for neural network approximation," *Neural Comput.*, vol. 21, pp. 2970–2989, 2009.

[45] G. Pisier, "Remarques sur un résultat non publié de B. Maurey," in *Sém. Anal. Fonctionnelle 1980–81*, Palaiseau, France, vol. I, no. 12, École Polytechnique, Centre de Mathématiques.

[46] R. W. Hamming, *Coding and Information Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[47] C. Martínez and M. Sanz, *The Theory of Fractional Powers of Operators*. Amsterdam, The Netherlands: Elsevier, 2001.

[48] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*. Princeton, NJ: Princeton Univ. Press, 1970.

[49] F. Girosi, "Approximation error bounds that use VC- bounds," in *Proc. 5th Int. Conf. on Artificial Neural Networks*, Paris, France, 1995, pp. 295–302.

[50] M. A. Kon and L. A. Raphael, "Approximating functions in reproducing kernel Hilbert spaces via statistical learning theory," in *Wavelets Splines*, G. Chen and M. J. Lai, Eds. Nashville, TN: Nashboro Press, 2006, pp. 271–286.

[51] M. A. Kon, L. A. Raphael, and D. A. Williams, "Extending Girosi's approximation estimates for functions in Sobolev spaces via statistical learning theory," *J. Anal. Applicat.*, vol. 3, pp. 67–90, 2005.

[52] G. Gnecco and M. Sanguineti, "Approximation error bounds via Rademacher complexity," *Appl. Math. Sci.*, vol. 2, pp. 153–176, 2008.

[53] V. Kůrková, P. C. Kainen, and V. Kreinovich, "Estimates of the number of hidden units and variation with respect to half-spaces," *Neural Netw.*, vol. 10, pp. 1061–1068, 1997.

[54] P. C. Kainen, V. Kůrková, and A. Vogt, "Integral combinations of Heavisides," *Math. Nachrichten*, vol. 283, pp. 854–878, 2010.

[55] Y. Ito, "Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory," *Neural Netw.*, vol. 4, pp. 385–394, 1991.

[56] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*. Amsterdam, The Netherlands: Academic, 2003.

[57] R. Courant, *Differential and Integral Calculus*. New York: Wiley-Interscience, 1988, vol. II.

[58] E. J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*. New York: Wiley, 1987.

[59] O. Hernandez-Lerma and J. Lasserre, "Approximation schemes for infinite linear programs," *SIAM J. Optim.*, vol. 8, pp. 973–988, 1998.

[60] I. Ekeland and T. Turnbull, *Infinite-Dimensional Optimization and Convexity*. Chicago, IL: Univ. of Chicago Press, 1983.

[61] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. Englewood Cliffs, NJ: Prentice-Hall, 1963.

[62] G. Gnecco and M. Sanguineti, "Estimates of variation with respect to a set and applications to optimization problems," *J. Optim. Theory Applicat.*, vol. 145, pp. 53–75, 2010.

[63] G. Gnecco and M. Sanguineti, "Suboptimal solutions to dynamic optimization problems via approximations of the policy functions," *J. Optim. Theory Applicat.*, vol. 146, pp. 764–794, 2010.

[64] T. Zolezzi, "Condition numbers and Ritz type methods in unconstrained optimization," *Contr. Cybern.*, vol. 36, pp. 811–822, 2007.

[65] J. W. Daniel, *The Approximate Minimization of Functionals*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[66] A. L. Dontchev, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, ser. Lecture Notes in Control and Information Sciences, 52. Berlin,Heidelberg, Germany: Springer, 1983.

[67] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. AMS*, vol. 39, pp. 1–49, 2001.

[68] T. Poggio and S. Smale, "The mathematics of learning: Dealing with data," *Notices Amer. Math. Soc.*, vol. 50, pp. 537–544, 2003.

[69] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

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

Paul C. Kainen received the Ph.D. degree in algebraic topology from Cornell University, Ithaca, NY, in 1970.

In addition to neural networks, he has worked extensively in graph theory, and is co-author, with T. L. Saathy, of the *Four Color Problem* (McGraw-Hill, 1977; Dover, 1986). Prior to joining Georgetown University, Washington, DC, in 1977, he taught at the Case Western University, was a member of technical staff at Bell Telephone Laboratories, and worked as a System Engineer.

Dr. Kainen organized two conferences on Topology in Biology (2002 and 2007) as a part of the Series *Knots in Washington*. He is Director of the Laboratory for Visual Mathematics at Georgetown University.

Věra Kůrková received the Ph.D. in topology from Charles University, Prague, Czech Republic.

Since 1990, she has bee a Scientist with the Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague. From 2002 to 2008, she was the Head of the Department of Theoretical Computer Science. Her research interests include mathematical theory of neurocomputing and learning and nonlinear approximation theory.

Dr. Kůrková is a member of the Board of the European Neural Networks Society and was the Chair of conferences ICANNGA 2001 and ICANN 2008. In 2010 she was awarder by the Czech Academy of Sciences the Bolzano Medal for her contributions to mathematics. She is a member of the Editorial Boards of the journals *Neural Networks* and *Neural Processing Letters*, and in 2008–2009 was a member of the Editorial Board of the IEEE TRANSACTIONS ON NEURAL NETWORKS.

Marcello Sanguineti received the "Laurea" degree in electronic engineering in 1992 and the Ph.D. degree in electronic engineering and computer science in 1996 from the University of Genoa, Italy.

He is currently Associate Professor at the University of Genoa and a Research Associate at Institute of Intelligent Systems for Automation of the National Research Council of Italy. He coordinated several international research projects on approximate solution of optimization problems. His main research interests are: infinite programming, nonlinear programming in learning from data, network optimization, optimal control, and neural networks for optimization.

Dr. Sanguineti is a Member of the Editorial Boards of the IEEE TRANSACTIONS ON NEURAL NETWORKS, the *International Mathematical Forum*, and *Mathematics in Engineering, Science and Aerospace*. He was the Chair of the Organizing Committee of the conference ICNPAA 2008.