Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2012, Article ID 806945, 17 pages
doi:10.1155/2012/806945

Research Article

A Comparison between Fixed-Basis and Variable-Basis Schemes for Function Approximation and Functional Optimization

Giorgio Gnecco

Department of Communication, Computer and System Sciences (DIST), University of Genova, Via Opera Pia 13, 16145 Genova, Italy

Correspondence should be addressed to Giorgio Gnecco, giorgio.gnecco@dist.unige.it

Received 30 July 2011; Accepted 14 November 2011

Academic Editor: Jacek Rokicki

Fixed-basis and variable-basis approximation schemes are compared for the problems of function approximation and functional optimization (also known as infinite programming). Classes of problems are investigated for which variable-basis schemes with sigmoidal computational units perform better than fixed-basis ones, in terms of the minimum number of computational units needed to achieve a desired error in function approximation or approximate optimization. Previously known bounds on the accuracy are extended, with better rates, to families of d-variable functions whose actual dependence is on a subset of $d' \ll d$ variables, where the indices of these d' variables are not known a priori.

1. Introduction

In functional optimization problems, also known as infinite programming problems, functionals have to be minimized with respect to functions belonging to subsets of function spaces. Function-approximation problems, the classical problems of the calculus of variations [1] and, more generally, all optimization tasks in which one has to find a function that is optimal in a sense specified by a cost functional belong to this family of problems. Such functions may express, for example, the routing strategies in communication networks, the decision functions in optimal control problems and economic ones, and the input/output mappings of devices that learn from examples.

Experience has shown that optimization of functionals over admissible sets of functions made up of linear combinations of relatively few basis functions with a simple structure and depending nonlinearly on a set of "inner" parameters (e.g., feedforward neural networks with one hidden layer and linear output activation units) often provides surprisingly good

suboptimal solutions. In such approximation schemes, each function depends on both external parameters (the coefficients of the linear combination) and inner parameters (the ones inside the basis functions). These are examples of *variable-basis approximators* since the basis functions are not fixed but their choice depends on the one of the inner parameters. In contrast, classical approximation schemes (such as the *Ritz method* in the calculus of variations [1]) do not use inner parameters but employ *fixed basis functions*, and the corresponding approximators exhibit only a linear dependence on the external parameters. Then, they are called *fixed-basis* or *linear approximators*. In [2], certain variable-basis approximators were applied to obtain approximate solutions to functional optimization problems. This technique was later formalized as the *extended Ritz method* (ERIM) [3] and was motivated by the innovative and successful application of feedforward neural networks in the late 80 s. For experimental results and theoretical investigations about the ERIM, see [2–7] and the references therein.

The basic motivation to search for suboptimal solutions of these forms is quite intuitive: when the number of basis functions becomes sufficiently large, the convergence of the sequence of suboptimal solutions to an optimal one may be ensured by suitable properties of the set of basis functions, the admissible set of functions, and the functional to be optimized [1, 5, 8]. Computational feasibility requirements (i.e., memory occupancy and time needed to find sufficiently good values for the parameters) make it crucial to estimate the minimum number of computational units needed by an approximator to guarantee that suboptimal solutions are "sufficiently close" to an optimal one. Such a number plays the role of "model complexity" of the approximator and can be studied with tools from linear and nonlinear approximation theory [9, 10].

As compared to fixed-basis approximators, in variable-basis ones the nonlinearity of the parametrization of the variable basis functions may cause the loss of useful properties of best approximation operators [11], such as uniqueness, homogeneity, and continuity, but may allow improved rates of approximation or approximate optimization [9, 12–14]. Then, to justify the use of variable-basis schemes instead of fixed-basis ones, it is crucial to investigate families of function-approximation and functional optimization problems for which, for a given desired accuracy, variable-basis schemes require a smaller number of computational units than fixed-basis ones. This is the aim of this work.

In the paper, the approximate solution of certain function-approximation and functional optimization problems via fixed- and variable-basis schemes is investigated. In particular, families of problems are presented, for which variable-basis schemes of a certain kind perform better than any fixed-basis one, in terms of the minimum number of computational units needed to achieve a desired worst-case error. Propositions 2.4, 2.7, 2.8, and 3.2 are the main contributions, which are presented after the exposition of results available in the literature.

The paper is organized as follows. Section 2 compares variable- and fixed-basis approximation schemes for function-approximation problems, which are particular instances of functional optimization. Section 3 extends the estimates to some more general families of functional optimization problems through the concepts of modulus of continuity and modulus of convexity of a functional. Section 4 is a short discussion.

2. Comparison of Bounds for Fixed- and Variable-Basis Approximation

Here and in the following, the "big O," "big Ω," and "big Θ" notations [18] are used. For two functions $f, g : (0, +\infty) \to \mathbb{R}$, one writes $f = O(g)$ if and only if there exist $M > 0$ and $x_0 > 0$ such that $|f(x)| \leq M|g(x)|$ for all $x > x_0$, $f = \Omega(g)$ if and only if $g = O(f)$, and $f = \Theta(g)$

if and only if both $f = O(g)$ and $f = \Omega(g)$ hold. In order to be able to use such notations also for multivariable functions, in the following it is assumed that all their arguments are fixed with the exception of one of them (more precisely, the argument ε).

Two approaches have been adopted in the literature to compare the approximation capabilities of fixed- and variable-basis approximation schemes (see also [15] for a discussion on this topic). In the first one, one fixes the family of functions to be approximated (e.g., the unit ball in a Sobolev space [16]), then one finds bounds on the worst-case approximation error for functions belonging to such a family, for various approximation schemes (fixed- and variable-basis ones). The second approach, initiated by Barron [12, 17], fixes a variable-basis approximation scheme (e.g., the set of one-hidden-layer perceptrons with a given upper bound on the number of sigmoidal computational units) and searches for families of functions that are well approximated by such an approximation scheme. Then, for these families of functions, the approximation capability of the variable-basis approximation scheme is compared with the ones of fixed-basis approximation schemes. In this context, one is interested in finding cases for which, the number of computational units being the same, one has upper bounds on the worst-case approximation error for certain variable-basis approximation schemes that are smaller than corresponding lower bounds for any fixed-basis one, implying that such variable-basis schemes have better approximation capabilities than every fixed-basis one.

One problem of the first approach is that, for certain families of smooth functions to be approximated, the bounds on the worst-case approximation error obtained for fixed- and variable-basis approximation schemes are very similar. In particular, typically one obtains the so-called *Jackson rate* of approximation [4] $n = \Theta(\varepsilon^{-d/m})$, where n is the number of computational units, $\varepsilon > 0$ is the worst-case approximation error, m is a measure of smoothness, and d is the number of variables on which such functions depend. Following the second approach, it was shown in [12, 17] that, for certain function-approximation problems, variable-basis schemes exhibit some advantages over fixed-basis ones (see Sections 2.1 and 2.2, where extensions of some results from [12, 17] are also derived).

In Section 2.1, some bounds in the $\mathcal{L}_2$-norm are considered, whereas Section 2.2 investigates bounds in the supnorm. Estimates in the $\mathcal{L}_2$-norm can be applied, for example, to investigate the approximation of the optimal policies in static team optimization problems [19]. Estimates in the supnorm are required, for example, to investigate the approximation of the optimal policies in dynamic optimization problems with a finite number of stages [20]. Indeed, for such problems, the supnorm can be used to analyze the error propagation from one stage to the next one, while this is not the case for the $\mathcal{L}_2$-norm [20]. Moreover, it provides guarantees on the approximation errors in the design of the optimal decision laws.

2.1. Bounds in the $\mathcal{L}_2$-Norm

The following Theorem 2.1 from [12] describes a quite general set of functions of d real variables (described in terms of their Fourier distributions) whose approximation from variable-basis approximation schemes with sigmoidal computational units requires $O(\varepsilon^{-2})$ computational units, where $\varepsilon > 0$ is the desired worst-case approximation error measured in the $\mathcal{L}_2$-norm. Recall that a sigmoidal function is defined in general as a bounded measurable function $\sigma : \mathbb{R} \to \mathbb{R}$ such that $\sigma(y) \to 1$ as $y \to +\infty$ and $\sigma(y) \to 0$ as $y \to -\infty$ [21]. For $C > 0$, d a positive integer, and B a bounded subset of $\mathbb{R}^d$ containing 0, by

$\Gamma_{B,C}$ we denote the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ having a Fourier representation of the form

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega\cdot x} \widehat{F}(d\omega) \tag{2.1}$$

for some complex-valued measure $\widehat{F}(d\omega) = e^{i\theta(\omega)}F(d\omega)$ (where $F(d\omega)$ and $\theta(\omega)$ are the magnitude distribution and the phase at the pulsation ω, resp.) such that

$$\int_{\mathbb{R}^d} \sup_{x\in B} |\langle \omega, x\rangle| F(d\omega) \leq C, \tag{2.2}$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product on $\mathbb{R}^d$. Functions in $\Gamma_{B,C}$ are continuously differentiable on B [12]. When B is the hypercube $[-1,1]^d$, the inequality (2.2) reduces to

$$\int_{\mathbb{R}^d} \|\omega\|_1 F(d\omega) \leq C, \tag{2.3}$$

where $\|\cdot\|_1$ denotes the l_1-norm.

For a probability measure μ on B, we denote by $\mathcal{L}_2(B,\mu)$ the Hilbert space of functions $g : B \to \mathbb{R}$ with inner product $\langle g_1, g_2\rangle_{\mathcal{L}_2(B,\mu)} := \int_B g_1(x)g_2(x)\mu(dx)$ and induced norm $\|g\|_{\mathcal{L}_2(B,\mu)} := \sqrt{\langle g, g\rangle_{\mathcal{L}_2(B,\mu)}}$. When there is no risk of confusion, the simpler notation $\|g\|_{\mathcal{L}_2}$ is used instead of $\|g\|_{\mathcal{L}_2(B,\mu)}$.

Theorem 2.1 (see [12, Theorem 1]). *For every $f \in \Gamma_{B,C}$, every sigmoidal function $\sigma : \mathbb{R} \to \mathbb{R}$, every probability measure μ on B, and every $n \geq 1$, there exist $a_k \in \mathbb{R}^d$, $b_k, c_k \in \mathbb{R}$, and $f_n : B \to \mathbb{R}$ of the form*

$$f_n(x) = \sum_{k=1}^{n} c_k \sigma(\langle a_k, x\rangle + b_k) + c_0, \tag{2.4}$$

such that

$$\|f - f_n\|_{\mathcal{L}_2} := \sqrt{\int_B (f(x) - f_n(x))^2 \mu(dx)} \leq \frac{2C}{\sqrt{n}}. \tag{2.5}$$

Variable-basis approximators of the form (2.4) are called *one-hidden-layer perceptrons* with n computational units. Formula (2.5) shows that *at most*

$$n_1^* = \left\lceil (2C)^2 \varepsilon^{-2} \right\rceil \tag{2.6}$$

computational units are required to guarantee a desired worst-case approximation error ε in the $\mathcal{L}_2$-norm, when variable-basis approximation schemes of the form (2.4) are used to approximate functions belonging to the set $\Gamma_{B,C}$.

In contrast to this, Theorem 2.2 from [12] shows that, when B is the unit hypercube $[0,1]^d$ and $\mu = \mu_u$ is the uniform probability measure on $[0,1]^d$, for the same set of functions $\Gamma_{B,C}$ the best linear approximation scheme requires $\Omega(\varepsilon^{-d})$ computational units in order to achieve the same worst-case approximation error ε. The set of all linear combinations of n fixed basis functions $h_1, h_2, \ldots, h_n$ in a linear space is denoted by $\operatorname{span}(h_1, h_2, \ldots, h_n)$.

Theorem 2.2 (see [12, Theorem 6]). *For every $n \geq 1$ and every choice of fixed basis functions $h_1, h_2, \ldots, h_n \in \mathcal{L}_2([0,1]^d, \mu_u)$, one has*

$$\sup_{f \in \Gamma_{[0,1]^d,C}} \inf_{f_n \in \operatorname{span}(h_1,h_2,\ldots,h_n)} \sqrt{\int_{[0,1]^d} \left(f(x) - f_n(x)\right)^2 \mu_u(dx)} \geq \frac{C}{16\pi e^{\pi-1} d} \left(\frac{1}{2n}\right)^{1/d}. \tag{2.7}$$

Remark 2.3. Inspection of the proof of [12, Theorem 6] shows that the factors $1/8$ and $1/n$, which appear in the original statement of the theorem, have to be replaced by $1/16$ and $1/2n$ in (2.7), respectively.

Inspection of the proof of Theorem 2.2 in [12] shows also that the lower bound (2.7) still holds if the set $\Gamma_{[0,1]^d,C}$ is replaced by either

$$\begin{aligned} \mathcal{S}_1 := &\left\{ f : [0,1]^d \longrightarrow \mathbb{R} : f(x) = \frac{C}{2\pi \|l\|_1} \cos(\omega \cdot x) : \omega = 2\pi l \text{ for } l \in \{0,1,\ldots\}^d,\ l \neq (0,\ldots,0) \right\} \\ &\cup \left\{ f : [0,1]^d \longrightarrow \mathbb{R} : f(x) = \frac{C}{2\pi} \right\} \end{aligned} \tag{2.8}$$

or

$$\begin{aligned} \mathcal{S}_2 := &\left\{ f : [0,1]^d \longrightarrow \mathbb{R} : f(x) = \beta \cos(\omega \cdot x) : \right. \\ &\left. |\beta| \leq \frac{C}{2\pi \|l\|_1},\ \omega = 2\pi l \text{ for } l \in \{0,1,\ldots\}^d, l \neq (0,\ldots,0) \right\} \\ &\cup \left\{ f : [0,1]^d \longrightarrow \mathbb{R} : f(x) = \beta,\ |\beta| \leq \frac{C}{2\pi} \right\}, \end{aligned} \tag{2.9}$$

where l denotes any multi-index and $\|l\|_1$ its norm (i.e., the sum of the components of l, which are nonnegative). Obviously, when B is the unit hypercube $[0,1]^d$, the upper bound (2.5) still holds under one of these two replacements, since $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \Gamma_{[0,1]^d,C}$.

The inequality (2.7) implies that for a uniform probability measure on $[0,1]^d$, *at least*

$$n_2^* = \left\lfloor \frac{1}{2} \left(\frac{C}{16\pi e^{\pi-1} d} \right)^d \varepsilon^{-d} \right\rfloor \tag{2.10}$$

computational units are required to guarantee a desired worst-case approximation error ε in the $\mathcal{L}_2$-norm, when fixed-basis approximation schemes of the form $\operatorname{span}(h_1, h_2, \ldots, h_n)$ are

used to approximate functions in $\Gamma_{[0,1]^d,C}$. Then, at least for a sufficiently small value of ε, Theorems 2.1 and 2.2 show that for $d > 2$, variable-basis approximators of the form (2.4) provide a smaller approximation error than any fixed-basis one for functions in $\Gamma_{[0,1]^d,C}$, the number of computational units being the same.

It should be noted that, for fixed C and ε, the estimate (2.6) is constant with respect to d, whereas the one (2.10) goes to 0 as d goes to $+\infty$. So, a too small value of $(1/2)(C/16\pi e^{\pi-1}d)^d$ in the bound (2.10) for fixed-basis approximation may make the theoretical advantage of variable-basis approximation of impractical use, since for large d it would be guaranteed only for sufficiently small ε (depending on C, too). In the following, families of d-variable functions are considered, for which this drawback is mitigated. These are families of d-variable functions whose actual dependence is on a subset of $d' \ll d$ variables, where the indices of these d' variables are not known a priori. These families are of interest, for example, in machine learning applications, for problems with redundant or correlated features. In this context, each of the d real variables represents a feature (e.g., a measure of some physical property of an object), and one is interested in learning a function of these features on the basis of a set of supervised examples. As it often happens in applications, only a small subset of the features is useful for the specific task (typically, classification or regression), due to the presence of redundant or correlated features. Then, one may assume that the function to be learned depends only on subset of $d' \ll d$ features but one may not know a priori which particular subset is. The problem of finding such a subset (or finding a subset of features of sufficiently small cardinality d' on which the function mostly depends, when the function depends on all the d features) is called the *feature-selection problem* [22].

For d' a positive integer and d its multiple, $\Gamma_{[0,1]^d,d',C}$ denotes the subset of functions in $\Gamma_{[0,1]^d,C}$ that depend only on d' of their possible d arguments.

Proposition 2.4. *For every $n \geq 1$ and every choice of fixed basis functions $h_1, h_2, \ldots, h_n \in \mathcal{L}_2([0,1]^d, \mu_u)$, for $n \leq (d+1)/2$ one has*

$$\sup_{f\in\Gamma_{[0,1]^d,d',C}} \inf_{f_n\in \operatorname{span}(h_1,h_2,\ldots,h_n)} \sqrt{\int_{[0,1]^d} \left(f(x) - f_n(x)\right)^2 \mu_u(dx)} \geq \frac{C}{8\pi} \tag{2.11}$$

and for $n > (d+1)/2$

$$\sup_{f\in\Gamma_{[0,1]^d,d',C}} \inf_{f_n\in \operatorname{span}(h_1,h_2,\ldots,h_n)} \sqrt{\int_{[0,1]^d} \left(f(x) - f_n(x)\right)^2 \mu_u(dx)} \geq \left(\frac{d}{d'}\right)^{1/d'} \frac{C}{16\pi e^{\pi-1} d'} \left(\frac{1}{2n}\right)^{1/d'}. \tag{2.12}$$

Proof. The proof is similar to the one of [12, Theorem 6]. The following is a list of the changes to that proof, needed to derive (2.11) and (2.12). We denote by $\|l\|_0$ the number of nonzero components of the multi-index l. Proceeding likewise in the proof of [12, Theorem 6], we get

$$\sup_{f\in\Gamma_{[0,1]^d,d',C}} \inf_{f_n\in \operatorname{span}(h_1,h_2,\ldots,h_n)} \sqrt{\int_{[0,1]^d} \left(f(x) - f_n(x)\right)^2 \mu_u(dx)} \geq \frac{C}{8\pi m^*}, \tag{2.13}$$

where m^* is the smallest positive integer m such that the number $N_{m,d,d'}$ of multi-indices $l \in \{0,1,\dots\}^d$ with norm $\|l\|_1 \leq m$ and that satisfy the constraint $\|l\|_0 \leq d'$ is larger than or equal to $2n$. More precisely, (2.13) is obtained by observing that for such an integer m the set $\mathcal{S}_2 \cap \Gamma_{[0,1]^d,d',C}$ contains at least $2n$ orthogonal cosinusoidal functions with $\mathcal{L}_2([0,1]^d, \mu_u)$-norm equal to $C/4\pi m$ and applying [12, Lemma 6], which states that for any orthonormal basis of a $2n$-dimensional space, there does not exist a linear subspace of dimension n having distance smaller than $1/2$ from every basis function in such an orthonormal basis. The constraint $\|l\|_0 \leq d'$ is not present in the proof of [12, Theorem 6] and is due to the specific form of the set $\Gamma_{[0,1]^d,d',C}$. Because of such a constraint, the functions in $\mathcal{S}_2$ with $\|l\|_0 > d'$ do not belong to $\Gamma_{[0,1]^d,d',C}$.

Then we get

$$N_{m,d,d'} = \binom{m+d}{d} \quad \text{for } d = d' \text{or } 1 \leq m \leq d', \tag{2.14}$$

$$\binom{m+d}{d} \geq N_{m,d,d'} \geq \frac{d}{d'}\binom{m+d'}{d'} \quad \text{for } \frac{d}{d'} \text{ a positive integer } > 1 \text{ and } m > 1. \tag{2.15}$$

Indeed, for $d = d'$ the equality (2.14) follows recalling that the number of different ways of placing N_o identical objects in N_b distinct boxes is $\binom{N_o+N_b-1}{N_b-1}$ [23, Theorem 5.1], and for this case it is the same estimate as the one obtained in the proof of [12, Theorem 6]. Similarly, for $1 \leq m \leq d'$ the constraint $\|l\|_0 \leq d'$ is redundant and we get again (2.14). Finally, for d/d' a positive integer larger than 1 and $m > 1$, the upper bound in (2.15) is obtained ignoring the constraint $\|l\|_0 \leq d'$, whereas the lower bound is obtained as follows. First, we partition the set of d variables into d/d' subsets of cardinality d', and then we apply to each subset the estimate $N_{m,d',d'} = \binom{m+d'}{d'}$ obtained by replacing d by d' in (2.14). In this way, the multi-index $l = 0$ is counted d/d' times (one for each subset), but the final estimate $N_{m,d,d'} \geq d/d'\binom{m+d'}{d'}$ so obtained holds since for $m > 1$ there are at least other $d/d' - 1$ multi-indices that have been not counted in this process.

In the following, we apply (2.14) and (2.15) for $m = 1$ and $m > 1$, respectively. For $m = 1$, the condition $N_{m,d,d'} \geq 2n$ becomes

$$\binom{1+d}{d} = d + 1 \geq 2n, \tag{2.16}$$

so $m^* = 1$ for $n \leq (d+1)/2$. This, combined with (2.13), proves (2.11).

Now, likewise in the proof of [12, Theorem 6], for $m > 1$ we exploit a bound from Stirling's formula, according to which $\binom{m+d'}{d'} \geq (m/e^{\pi-1}d')^{d'}$, so the condition $N_{m,d,d'} \geq 2n$ holds if we impose

$$\frac{d}{d'}\left(\frac{m}{e^{\pi-1}d'}\right)^{d'} \geq 2n, \tag{2.17}$$

which is equivalent to

$$m \geq \left\lceil e^{\pi-1} d' (2n)^{1/d'} \left(\frac{d'}{d}\right)^{1/d'} \right\rceil \tag{2.18}$$

(note that, for $n > (d+1)/2$, the value of m provided by (2.18) is indeed larger than 1, as required for the application of (2.15)). Since

$$2e^{\pi-1} d' (2n)^{1/d'} \left(\frac{d'}{d}\right)^{1/d'} \geq \left\lceil e^{\pi-1} d' (2n)^{1/d'} \left(\frac{d'}{d}\right)^{1/d'} \right\rceil \tag{2.19}$$

we conclude that $m^* \leq 2e^{\pi-1} d' (2n)^{1/d'} (d'/d)^{1/d'}$ for $n > (d+1)/2$. This, together with (2.13), proves the statement (2.12). □

For the case considered by Proposition 2.4, an uniform probability measure on $[0,1]^d$, and $0 < \varepsilon < C/8\pi$, formulas (2.11) and (2.12) show that *at least*

$$n_3^* = \max\left\{ \left\lceil \frac{d+1}{2} \right\rceil, \left\lfloor \frac{1}{2} \frac{d}{d'} \left(\frac{C}{16\pi e^{\pi-1} d'} \right)^{d'} \varepsilon^{-d'} \right\rfloor \right\} \tag{2.20}$$

computational units are required to guarantee a desired worst-case approximation error ε in the $\mathcal{L}_2$-norm, when fixed-basis approximation schemes of the form $\operatorname{span}(h_1, h_2, \ldots, h_n)$ are used to approximate functions in $\Gamma_{[0,1]^d, d', C}$.

Remark 2.5. The quantity d' in Proposition 2.4 has to be interpreted as an *effective number of variables* for the family of functions $\Gamma_{[0,1]^d, d', C}$ to be approximated. Roughly speaking, the flexibility of the neural network architecture (2.4) allows one to identify, for each $f \in \Gamma_{[0,1]^d, d', C}$, the d' variables on which it actually depends, whereas fixed-basis approximation schemes have not this flexibility. Indeed, differently from the lower bound (2.10), for fixed C, ε, and d' the lower bound (2.20) goes to $+\infty$ as d goes to $+\infty$. Finally, similar remarks as in Remark 2.3 apply to Proposition 2.4.

2.2. Bounds in the Supnorm

The next result is from [17] and is analogous to Theorem 2.1, but it measures the worst-case approximation error in the supnorm.

Theorem 2.6 (see [17, Theorem 2]). *For every $f \in \Gamma_{B,C}$ and every $n \geq 1$, there exists $f_n : B \to \mathbb{R}$ of the form* (2.4) *such that*

$$\sup_{x \in B} \left| f(x) - f_n(x) \right| \leq \frac{120C}{\sqrt{n}} d. \tag{2.21}$$

Upper bounds in the supnorm similar to the one from Theorem 2.6 are given, for example, in [24, 25]. Moreover, for $f \in \Gamma_{[0,1]^d, d', C}$, the following estimate holds.

Proposition 2.7. *For every $f \in \Gamma_{[0,1]^d,d',C}$ and every $n \geq 1$, there exists $f_n : [0,1]^d \to \mathbb{R}$ of the form* (2.4) *such that*

$$\sup_{x\in[0,1]^d} \left|f(x) - f_n(x)\right| \leq \frac{120C}{\sqrt{n}} d'. \tag{2.22}$$

Proof. Each function $f \in \Gamma_{[0,1]^d,d',C}$ depends on d' arguments; let $i_1, \ldots, i_{d'}$ be their indices. Let $\widetilde{f} : [0,1]^{d'} \to \mathbb{R}$ be defined by $\widetilde{f}(y) := f(x)$, where $x_{i_1} = y_1, \ldots, x_{i_{d'}} = y_{d'}$, and all the other components of x are arbitrary in $[0,1]^{d-d'}$. Then $\widetilde{f} \in \Gamma_{[0,1]^{d'},C}$, so by Theorem 2.6 there exists an approximation $\widetilde{f}_n : [0,1]^{d'} \to \mathbb{R}$ made up of n sigmoidal computational units and a constant term such that $\sup_{x\in[0,1]^{d'}} |\widetilde{f}(x) - \widetilde{f}_n(x)| \leq (120C/\sqrt{n})d'$. Finally, we observe that $\widetilde{f}_n$ can be extended to a function $f_n : [0,1]^d \to \mathbb{R}$ of the form (2.4) such that $\sup_{x\in[0,1]^d} |f(x) - f_n(x)| = \sup_{x\in[0,1]^{d'}} |\widetilde{f}(x) - \widetilde{f}_n(x)|$, then one obtains (2.22). □

The estimates (2.21) and (2.22) show that *at most*

$$\begin{aligned} n_4^* &= \left\lceil (120C)^2 d^2 \varepsilon^{-2} \right\rceil, \\ n_5^* &= \left\lceil (120C)^2 d'^2 \varepsilon^{-2} \right\rceil \end{aligned} \tag{2.23}$$

computational units, respectively, are required to guarantee a desired worst-case approximation error ε in the supnorm, when variable-basis approximation schemes of the form (2.4) are used to approximate functions belonging to the sets $\Gamma_{B,C}$ and $\Gamma_{[0,1]^d,d',C}$, respectively.

The next proposition, combined with Theorem 2.6 and Proposition 2.7, allows one to compare the approximation capabilities of fixed- and variable-basis schemes in the supnorm, showing cases for which the upper bounds (2.21) and (2.22) are smaller than one of the corresponding lower bounds (2.24)–(2.26), at least for n sufficiently large.

Proposition 2.8. *For every $n \geq 1$ and every choice of fixed bounded and μ_u-measurable basis functions $h_1, h_2, \ldots, h_n : [0,1]^d \to \mathbb{R}$, the following hold.*

(i) *For the approximation of functions in $\Gamma_{[0,1]^d,C}$, one has*

$$\sup_{f\in\Gamma_{[0,1]^d,C}} \inf_{f_n\in \operatorname{span}(h_1,h_2,\ldots,h_n)} \sup_{x\in[0,1]^d} \left|f(x) - f_n(x)\right| \geq \frac{C}{16\pi e^{\pi-1} d} \left(\frac{1}{2n}\right)^{1/d}. \tag{2.24}$$

(ii) *For the approximation of functions in $\Gamma_{[0,1]^d,d',C}$, for $n \leq (d+1)/2$, one has*

$$\sup_{f\in\Gamma_{[0,1]^d,d',C}} \inf_{f_n\in \operatorname{span}(h_1,h_2,\ldots,h_n)} \sup_{x\in[0,1]^d} \left|f(x) - f_n(x)\right| \geq \frac{C}{8\pi} \tag{2.25}$$

whereas for $n > (d+1)/2$

$$\sup_{f\in\Gamma_{[0,1]^d,d',C}} \inf_{f_n\in \operatorname{span}(h_1,h_2,\ldots,h_n)} \sup_{x\in[0,1]^d} \left|f(x) - f_n(x)\right| \geq \left(\frac{d}{d'}\right)^{1/d'} \frac{C}{16\pi e^{\pi-1} d'} \left(\frac{1}{2n}\right)^{1/d'}. \tag{2.26}$$

Proof. For each bounded and μ_u-measurable function $g : [0,1]^d \to \mathbb{R}$, we get

$$\sqrt{\int_{[0,1]^d} g^2(x)\mu_u(dx)} \leq \sup_{x\in[0,1]^d} \left|g(x)\right| \sqrt{\int_{[0,1]^d} \mu_u(dx)} = \sup_{x\in[0,1]^d} \left|g(x)\right|, \tag{2.27}$$

so

$$\begin{aligned}
&\sup_{f\in\Gamma_{[0,1]^d,C}} \inf_{f_n\in\operatorname{span}(h_1,h_2,\ldots,h_n)} \sqrt{\int_{[0,1]^d} \left(f(x)-f_n(x)\right)^2 \mu_u(dx)} \\
&\qquad \leq \sup_{f\in\Gamma_{[0,1]^d,C}} \inf_{f_n\in\operatorname{span}(h_1,h_2,\ldots,h_n)} \sup_{x\in[0,1]^d} \left|f(x)-f_n(x)\right|, \\
&\sup_{f\in\Gamma_{[0,1]^d,d',C}} \inf_{f_n\in\operatorname{span}(h_1,h_2,\ldots,h_n)} \sqrt{\int_{[0,1]^d} \left(f(x)-f_n(x)\right)^2 \mu_u(dx)} \\
&\qquad \leq \sup_{f\in\Gamma_{[0,1]^d,d',C}} \inf_{f_n\in\operatorname{span}(h_1,h_2,\ldots,h_n)} \sup_{x\in[0,1]^d} \left|f(x)-f_n(x)\right|.
\end{aligned} \tag{2.28}$$

Then we get the lower bounds (2.24)–(2.26) by (2.7), (2.11), and (2.12), respectively. □

For the case considered by Proposition 2.8, the estimate (2.24) implies that *at least* n_2^* computational units are required to guarantee a desired worst-case approximation error ε in the supnorm, when fixed-basis approximation schemes of the form $\operatorname{span}(h_1, h_2, \ldots, h_n)$ are used to approximate functions in $\Gamma_{[0,1]^d,C}$. Similarly, for $0 < \varepsilon < C/8\pi$, the bounds (2.25) and (2.26) imply that *at least* n_3^* computational units are required when $\Gamma_{[0,1]^d,C}$ is replaced by $\Gamma_{[0,1]^d,d',C}$. One can observe that, for each d, d' and C, each of the lower bounds (2.25) and (2.26) is larger than (2.24). Moreover, all the other parameters being fixed, the lower bound (2.24) goes to 0 as d tends to $+\infty$, whereas for $d \geq 2n-1$, the lower bound (2.25) holds, and it does not depend on the specific value of d. Finally, for $d > 2$, the upper bound (2.21) is smaller than the lower bound (2.24) for n sufficiently large, and similarly, for $d' > 2$, the upper bound (2.22) is smaller than the lower bounds (2.25) and (2.26) for n sufficiently large. For instance, in the latter case and for d' sufficiently small with respect to d, this happens for $\lceil 225d'^2/\pi^2 \rceil \leq n \leq (d+1)/2$ and for

$$n \geq \min\left\{ \left\lceil \frac{d+1}{2} \right\rceil, \ \left\lceil K_1 d^{K_2} \right\rceil \right\}, \tag{2.29}$$

where $K_1 = \left(1920 \cdot 2^{1/d'} \pi e^{\pi-1}\right)^{2d'/(d'-2)} d'^{2(2d+1)/(d'-2)}$ and $K_2 = 2/(d'-2)$.

Similar remarks as in Remark 2.3 can be made about the bounds in the supnorm derived in this section.

3. Application to Functional Optimization Problems

The results of Section 2 can be extended, with the same rates of approximation or similar ones, to the approximate solution of certain functional optimization problems. This can be done by

exploiting the concepts of modulus of continuity and modulus of convexity of a functional, provided that continuity and uniform convexity assumptions are satisfied. The basic ideas are the following (see also [5] for a similar analysis).

3.1. Rates of Approximate Optimization in Terms of the Modulus of Continuity

Let $\mathcal{X}$ be a normed linear space, $X \subseteq \mathcal{X}$, and $\Phi : X \to \mathbb{R}$ a functional. Suppose that the functional optimization problem

$$\min_{f \in X} \Phi(f) \tag{3.1}$$

has a solution f°, and let $X_1 \subseteq X_2 \subseteq \cdots \subseteq X_n \subseteq \cdots \subseteq X$ be a nested sequence of subsets of X such that

$$\inf_{f_n \in X_n} \left\| f^\circ - f_n \right\|_{\mathcal{X}} \leq \varepsilon_n \tag{3.2}$$

for some $\varepsilon_n > 0$, where $\varepsilon_n \to 0$ as $n \to +\infty$. Then, if the functional Φ is continuous, too, one has

$$\inf_{f_n \in X_n} \left| \Phi(f^\circ) - \Phi(f_n) \right| \leq \alpha_{f^\circ}(\varepsilon_n) \longrightarrow 0 \quad \text{as } n \longrightarrow +\infty, \tag{3.3}$$

where $\alpha_{f^\circ} : [0, +\infty) \to [0, +\infty)$ defined by $\alpha_{f^\circ}(t) = \sup\{|\Phi(f^\circ) - \Phi(g)| : g \in X, \|f - g\|_{\mathcal{X}} \leq t\}$ is the *modulus of continuity* of Φ at f°. For instance, if Φ is Lipschitz continuous with Lipschitz constant K_Φ, one has $\alpha_{f^\circ}(t) \leq K_\Phi t$, and by (3.2)

$$\inf_{f_n \in X_n} \left| \Phi(f^\circ) - \Phi(f_n) \right| \leq K_\Phi \varepsilon_n. \tag{3.4}$$

Then, if an upper bound on ε_n in terms of n is known (e.g., $\varepsilon_n = O(n^{-1/2})$ under the assumptions of Theorem 2.1, where $X = \Gamma_{B,C} \subset \mathcal{L}_2(B, \mu) = \mathcal{X}$ and X_n is the set of functions of the form (2.4)), then the same upper bound (up to a multiplicative constant) holds on $\inf_{f_n \in X_n} |\Phi(f^\circ) - \Phi(f_n)|$. So, investigating the approximating capabilities of the sets X_n is useful for functional optimization purposes, too.

3.2. Rates of Approximate Optimization in Terms of the Modulus of Convexity

When dealing with suboptimal solutions from a set $X_n \subseteq X$, the following question arises: suppose that $\widetilde{f}_n \in X_n$ is such that

$$\left| \Phi(f^\circ) - \Phi\left(\widetilde{f}_n\right) \right| \leq \gamma_n \tag{3.5}$$

for some $\gamma_n > 0$, where $\gamma_n \to 0$ as $n \to +\infty$. This can be guaranteed, for example, if the functional is continuous, the sets X_n satisfy the property (3.2), and one chooses $\widetilde{f}_n \in \operatorname{argmin}_{f_n \in X_n} \|f^\circ - f_n\|_{\mathcal{X}}$ assuming, almost without loss of generality, that such a set is nonempty. If this is not the case, then one can proceed as follows. For $\epsilon > 0$, let $\operatorname{argmin}_{\epsilon, f_n \in X_n} \|f - f_n\|_{\mathcal{X}} := \{f_n \in X_n : \|f^\circ - f_n\|_{\mathcal{X}} \leq \inf_{f_n \in X_n} \|f^\circ - f_n\|_{\mathcal{X}} + \epsilon\}$. Then one obtains estimates similar to the ones of this section (obtained assuming that $\operatorname{argmin}_{f_n \in X_n} \|f^\circ - f_n\|_{\mathcal{X}}$ is nonempty) by choosing $\widetilde{f}_n \in \operatorname{argmin}_{\eta\epsilon_n, f_n \in X_n} \|f^\circ - f_n\|_{\mathcal{X}}$, where $\eta > 1$ is a constant. Does the estimate (3.5) imply an upper bound on the approximation error $\|f^\circ - \widetilde{f}_n\|_{\mathcal{X}}$? A positive answer can be given when the functional Φ is uniformly convex. Recall that a functional $\Phi : X \to \mathbb{R}$ is called *convex* on a convex set $X \subseteq \mathcal{X}$ if and only if for all $h, g \in X$ and all $\lambda \in [0,1]$, one has $\Phi(\lambda h + (1-\lambda)g) \leq \lambda\Phi(h) + (1-\lambda)\Phi(g)$ and it is called *uniformly convex* if and only if there exists a nonnegative function $\delta : [0, +\infty) \to [0, +\infty)$ such that $\delta(0) = 0$, $\delta(t) > 0$ for all $t > 0$, and for all $h, g \in X$ and all $\lambda \in [0,1]$, one has

$$\Phi(\lambda h + (1-\lambda)g) \leq \lambda\Phi(h) + (1-\lambda)\Phi(g) - \lambda(1-\lambda)\delta\Big(\|h - g\|_{\mathcal{X}}\Big). \tag{3.6}$$

Any such function δ is called a *modulus of convexity* of Φ [26]. The terminology is not unified: some authors use the term "strictly uniformly convex" instead of "uniformly convex" and reserve the term "uniformly convex" for the case where $\delta : [0, +\infty) \to [0, +\infty)$ merely satisfies $\delta(0) = 0$ and $\delta(t_0) > 0$ for some $t_0 > 0$ (see, e.g., [27, 28, page 10]). Note that when $\mathcal{X}$ is a Hilbert space and $\delta(t)$ has the quadratic expression

$$\delta(t) = \frac{1}{2}\, ct^2 \tag{3.7}$$

for some constant $c > 0$, the condition (3.6) is equivalent to the convexity of the functional $\Phi(\cdot) - \delta(\|\cdot\|_{\mathcal{X}}) = \Phi(\cdot) - (1/2)c\|\cdot\|_{\mathcal{X}}^2$. Indeed, the latter property means that, for all $h, g \in X$ and all $\lambda \in [0,1]$, one has

$$\Phi(\lambda h + (1-\lambda)g) - \frac{1}{2}c\|\lambda h + (1-\lambda)g\|_{\mathcal{X}}^2 \leq \lambda\Phi(h) - \frac{\lambda}{2}c\|h\|_{\mathcal{X}}^2 + (1-\lambda)\Phi(g) - \frac{1-\lambda}{2}c\|g\|_{\mathcal{X}}^2, \tag{3.8}$$

and this is equivalent to

$$\Phi(\lambda h + (1-\lambda)g) \leq \lambda\Phi(h) + (1-\lambda)\Phi(g) - \frac{\lambda(1-\lambda)}{2}c\|h - g\|_{\mathcal{X}}^2, \tag{3.9}$$

since one can show through straightforward computations that, for $\mathcal{X}$ a Hilbert space, one has

$$\frac{1}{2}c\|\lambda h + (1-\lambda)g\|_{\mathcal{X}}^2 - \frac{\lambda}{2}c\|h\|_{\mathcal{X}}^2 - \frac{1-\lambda}{2}c\|g\|_{\mathcal{X}}^2 = -\frac{\lambda(1-\lambda)}{2}c\|h - g\|_{\mathcal{X}}^2. \tag{3.10}$$

One of the most useful properties of uniform convexity is that $f^\circ \in \operatorname{argmin}_{f\in X}\Phi(f)$ implies the lower bound

$$\left|\Phi(f^\circ) - \Phi(f)\right| \geq \delta\left(\left\|f^\circ - f\right\|_{\mathcal{X}}\right) \tag{3.11}$$

for any $f \in X$ (see, e.g., [5, Proposition 2.1(iii)]). When the modulus of convexity has the form (3.7), this implies (together with (3.5))

$$\left\|f^\circ - \tilde{f}_n\right\|_{\mathcal{X}} \leq \sqrt{2\frac{\gamma_n}{c}} \longrightarrow 0 \quad \text{as } n \longrightarrow +\infty. \tag{3.12}$$

When (3.2) holds, too, and Φ has modulus of continuity α_{f° at f°, one can take

$$\gamma_n = \alpha_{f^\circ}(\varepsilon_n) \tag{3.13}$$

in (3.12), thus obtaining

$$\left\|f^\circ - \tilde{f}_n\right\|_{\mathcal{X}} \leq \sqrt{2\frac{\alpha_{f^\circ}(\varepsilon_n)}{c}} \longrightarrow 0 \quad \text{as } n \longrightarrow +\infty. \tag{3.14}$$

Again, this allows one to extend rates of function approximation to functional optimization, supposing, as in Section 3.1, that Φ is also Lipschitz continuous with Lipschitz constant K_Φ and that $\varepsilon_n = O(n^{-1/2})$. Then, one obtains (from the choice (3.13) for γ_n and formula (3.14))

$$\left|\Phi(f^\circ) - \Phi\left(\tilde{f}_n\right)\right| = O\left(n^{-1/2}\right), \tag{3.15}$$

$$\left\|f^\circ - \tilde{f}_n\right\|_{\mathcal{X}} = O\left(n^{-1/4}\right). \tag{3.16}$$

Remark 3.1. In [29], a greedy algorithm is proposed to construct a sequence of sets X_n corresponding to variable-basis schemes and functions $\tilde{f}_n \in X_n$ that achieve the rate (3.15) for certain uniformly convex functional optimization problems. Such an algorithm can be interpreted as an extension to functional optimization of the greedy algorithm proposed in [12] for function approximation by sigmoidal neural networks.

Finally, it should be noted that the rate (3.15) is achieved in general by imposing some structure on the sets X and X_n. For instance, the set X in [29] is the convex hull of some set of functions $G \subset \mathcal{X}$, that is,

$$X = \operatorname{co} G := \left\{ \sum_{j=1}^{k} \alpha_j g_j : \alpha_j \geq 0,\ \sum_{j=1}^{k} \alpha_j = 1,\ g_j \in G,\ k \in \mathbb{Z}^+ \right\}, \tag{3.17}$$

whereas, for each $n \in \mathbb{Z}^+$, the set X_n in [29] is

$$X_n = \mathrm{co}_n G := \left\{ \sum_{j=1}^{n} \alpha_j g_j : \alpha_j \geq 0,\ \sum_{j=1}^{n} \alpha_j = 1,\ g_j \in G \right\}. \tag{3.18}$$

Functional optimization problems have in general a natural domain X larger than co G (or its closure $\overline{\mathrm{co}}\, G$ in the norm of the ambient space $\mathcal{X}$). Therefore, the choice of a set X of the form (3.17) as the domain of the functional Φ might seem unmotivated. This is not the case, because there are several examples of functional optimization problems for which, for suitable sets G and a natural domain X larger than co G (resp., $\overline{\mathrm{co}}\, G$), the set

$$\mathrm{argmin}_{f \in X} \Phi(f) \tag{3.19}$$

has a nonempty intersection with co G (resp., $\overline{\mathrm{co}}\, G$), or it is contained in it. This issue is studied in [20] for dynamic optimization problems and in [19] for static team optimization ones, where structural properties (e.g., smoothness) of the minimizers are studied.

3.3. Comparison between Fixed- and Variable-Basis Schemes for Functional Optimization

The proposition follows by combining the results derived in Sections 2.1, 3.1, and 3.2.

Proposition 3.2. *Let the functional Φ be Lipschitz continuous with Lipschitz constant K_Φ and uniformly convex with modulus of convexity of the form* (3.7)*, $X = \Gamma_{B,C}$, μ any probability measure on B, $\mathcal{X} = \mathcal{L}_2(B, \mu)$, and suppose that there exists a minimizer $f^\circ \in \mathrm{argmin}_{f \in \Gamma_{B,C}} \Phi(f)$. Then the following hold.*

(i) *For every $n \geq 1$ there exists f_n of the form* (2.4) *such that*

$$\left\| f^\circ - f_n \right\|_{\mathcal{L}_2} \leq \frac{2C}{\sqrt{n}}. \tag{3.20}$$

For each such f_n one has

$$\left| \Phi(f^\circ) - \Phi(f_n) \right| \leq K_\Phi \frac{2C}{\sqrt{n}} \tag{3.21}$$

and if $\widetilde{f}_n$ of the form (2.4) *is such that*

$$\left| \Phi(f^\circ) - \Phi\left(\widetilde{f}_n\right) \right| \leq K_\Phi \frac{2C}{\sqrt{n}}, \tag{3.22}$$

then

$$\left\| f^\circ - \widetilde{f}_n \right\|_{\mathcal{L}_2} \leq 2\sqrt{\frac{K_\Phi C}{c}} \frac{1}{\sqrt[4]{n}}. \tag{3.23}$$

(ii) *For* $B = [0,1]^d$, μ_u *equal to the uniform probability measure on* $[0,1]^d$, *every* $n \geq 1$, *and every choice of fixed-basis functions* $h_1, \ldots, h_n \in \mathcal{L}_2([0,1]^d, \mu_u)$, *there exists a uniformly convex functional* $\overline{\Phi}$ *(such a functional* $\overline{\Phi}$ *can be also chosen to be Lipschitz continuous with Lipschitz constant* $K_{\overline{\Phi}}$, *but this is not needed in the inequalities* (3.24)–(3.29), *since they do not contain* $K_{\overline{\Phi}}$*) with modulus of convexity of the form* (3.7) *and minimizer* $\overline{f}^{\circ} \in \operatorname{argmin}_{f \in \Gamma_{[0,1]^d,C}} \overline{\Phi}(f)$ *such that for every* $0 < \chi < 1$ *one has*

$$\inf_{f_n \in \operatorname{span}\{h_1,h_2,\ldots,h_n\}} \left\| \overline{f}^{\circ} - f_n \right\|_{\mathcal{L}_2} \geq \chi \frac{C}{16\pi e^{\pi-1} d} \left(\frac{1}{2n} \right)^{1/d}, \tag{3.24}$$

$$\inf_{f_n \in \operatorname{span}\{h_1,h_2,\ldots,h_n\}} \left| \overline{\Phi}\left(\overline{f}^{\circ} \right) - \overline{\Phi}(f_n) \right| \geq \frac{1}{2} c \left(\chi \frac{C}{16\pi e^{\pi-1} d} \right)^2 \left(\frac{1}{2n} \right)^{2/d}. \tag{3.25}$$

(iii) *The statements* (i) *and* (ii) *still hold by replacing the set* $\Gamma_{B,C}$ *by* $\Gamma_{[0,1]^d,d',C}$, *for* d *a multiple of* d'. *The only difference is that the estimates* (3.24) *and* (3.25) *are replaced, respectively, by*

$$\inf_{f_n \in \operatorname{span}\{h_1,h_2,\ldots,h_n\}} \left\| \overline{f}^{\circ} - f_n \right\|_{\mathcal{L}_2} \geq \chi \frac{C}{8\pi}, \tag{3.26}$$

$$\inf_{f_n \in \operatorname{span}\{h_1,h_2,\ldots,h_n\}} \left| \overline{\Phi}\left(\overline{f}^{\circ} \right) - \overline{\Phi}(f_n) \right| \geq \frac{1}{2} c \left(\chi \frac{C}{8\pi} \right)^2 \tag{3.27}$$

for $n \leq (d+1)/2$ *and by*

$$\inf_{f_n \in \operatorname{span}\{h_1,h_2,\ldots,h_n\}} \left\| \overline{f}^{\circ} - f_n \right\|_{\mathcal{L}_2} \geq \chi \left(\frac{d}{d'} \right)^{1/d'} \frac{C}{16\pi e^{\pi-1} d'} \left(\frac{1}{2n} \right)^{1/d'}, \tag{3.28}$$

$$\inf_{f_n \in \operatorname{span}\{h_1,h_2,\ldots,h_n\}} \left| \overline{\Phi}\left(\overline{f}^{\circ} \right) - \overline{\Phi}(f_n) \right| \geq \frac{1}{2} c \left(\frac{d}{d'} \right)^{2/d'} \left(\chi \frac{C}{16\pi e^{\pi-1} d'} \right)^2 \left(\frac{1}{2n} \right)^{2/d'} \tag{3.29}$$

for $n > (d+1)/2$.

Proof. (i) The estimate (3.20) follows by Theorem 2.1. The bound (3.21) follows by (3.20), the definition of modulus of continuity, and the assumption of Lipschitz continuity of Φ. Finally, (3.23) is obtained by property (3.11) of the modulus of convexity and its expression (3.7).

(ii) (3.24) comes from Theorem 2.2: the constant χ is introduced in order to remove the supremum with respect to $f \in \Gamma_{[0,1]^d,C}$ in formula (2.7) and replace it with the choice $f = \overline{f}^{\circ}$, where $\overline{f}^{\circ}$ is any function that achieves the bound (2.7) up to the constant factor χ; (3.25) follows from (3.24), (3.11), and (3.7), choosing as $\overline{\Phi}$ any functional that is uniformly convex with modulus of convexity of the form (3.7), and such that $\overline{f}^{\circ} \in \operatorname{argmin}_{f \in \Gamma_{[0,1]^d,C}} \overline{\Phi}(f)$.

(iii) The estimates (3.20), (3.21), (3.23) still hold when the set $\Gamma_{B,C}$ is replaced by $\Gamma_{[0,1]^d,d',C}$ since $\Gamma_{[0,1]^d,d',C} \subset \Gamma_{B,C}$ for $B = [0,1]^d$, whereas formulas (3.26)–(3.29) are obtained likewise formulas (3.24) and (3.25), by applying Proposition 2.4 instead of Theorem 2.2. □

4. Discussion

Classes of function-approximation and functional optimization problems have been investigated for which, for a given desired error, certain variable-basis approximation schemes with sigmoidal computational units require less parameters than fixed-basis ones. Previously known bounds on the accuracy have been extended, with better rates, to families of functions whose effective number of variables d' is much smaller than the number of their arguments d.

Proposition 3.2 shows that there is a strict connection between certain problems of function approximation and functional optimization. For such two classes of problems, indeed, the approximation error rates for the first class can be converted into rates of approximate optimization for the second one and vice versa. In particular, for $d > 2$, $X = \Gamma_{[0,1]^d,C}$, and any linear approximation scheme $\mathrm{span}\{h_1, h_2, \ldots, h_n\}$, the estimates (3.21) and (3.25) show families of functional optimization problems for which the error in approximate optimization with variable-basis schemes of sigmoidal type is smaller than the one associated with the linear scheme. For $d' > 2$ and $X = \Gamma_{[0,1]^d,d',C}$, a similar remark can be made for the estimates (3.21) and (3.27) and for the bounds (3.21) and (3.29). Finally, the bound (3.23) shows that for large n any approximate minimizer $\tilde{f}_n$ of the form (2.4) differs slightly from the true minimizer f°, even though the error in approximate optimization (3.22) and the associated approximation error (3.23) have different rates. In contrast, the estimates (3.24), (3.26), and (3.28) show that, for any linear approximation scheme $\mathrm{span}\{h_1, h_2, \ldots, h_n\}$, there exists a functional optimization problem whose minimizer $\overline{f}^\circ$ cannot be approximated with the same accuracy by the linear scheme.

The results presented in the paper provide some theoretical justification for the use of variable-basis approximation schemes (instead of fixed-basis ones) in function approximation and functional optimization.

Acknowledgment

The author was partially supported by a PRIN grant from the Italian Ministry for University and Research, project "Adaptive State Estimation and Optimal Control."

References

[1] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1963.
[2] R. Zoppoli and T. Parisini, "Learning techniques and neural networks for the solution of N-stage nonlinear nonquadratic optimal control problems," in *Systems, Models and Feedback: Theory and Applications*, A. Isidori and T. J. Tarn, Eds., pp. 193–210, Birkhäuser, Boston, Mass, USA, 1992.
[3] R. Zoppoli, M. Sanguineti, and T. Parisini, "Approximating networks and extended Ritz method for the solution of functional optimization problems," *Journal of Optimization Theory and Applications*, vol. 112, no. 2, pp. 403–440, 2002.
[4] S. Giulini and M. Sanguineti, "Approximation schemes for functional optimization problems," *Journal of Optimization Theory and Applications*, vol. 140, no. 1, pp. 33–54, 2009.
[5] V. Kůrková and M. Sanguineti, "Error estimates for approximate optimization by the extended Ritz method," *SIAM Journal on Optimization*, vol. 15, no. 2, pp. 461–487, 2005.
[6] T. Zolezzi, "Condition numbers and Ritz type methods in unconstrained optimization," *Control and Cybernetics*, vol. 36, no. 3, pp. 811–822, 2007.
[7] R. Zoppoli, T. Parisini, M. Sanguineti, and M. Baglietto, *Neural Approximations for Optimal Control and Decision*, Springer, London, UK.
[8] J. W. Daniel, *The Approximate Minimization of Functionals*, Prentice-Hall Inc., Englewood Cliffs, NJ, USA, 1971.

[9] V. Kůrková and M. Sanguineti, "Comparison of worst case errors in linear and neural network approximation," *IEEE Transactions on Information Theory*, vol. 48, no. 1, pp. 264–275, 2002.
[10] A. Pinkus, *n-Widths in Approximation Theory*, vol. 7, Springer, Berlin, Germany, 1985.
[11] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer, Berlin, Germany, 1970.
[12] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.
[13] G. Gnecco, V. Kůrková, and M. Sanguineti, "Can dictionary-based computational models outperform the best linear ones?" *Neural Networks*, vol. 24, no. 8, pp. 881–887, 2011.
[14] G. Gnecco, V. Kůrková, and M. Sanguineti, "Some comparisons of complexity in dictionary-based and linear computational models," *Neural Networks*, vol. 24, pp. 172–182, 2011.
[15] A. Pinkus, "Approximation theory of the MLP model in neural networks," in *Acta Numerica, 1999*, vol. 8, pp. 143–195, Cambridge University Press, Cambridge, UK, 1999.
[16] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, vol. 140 of *Pure and Applied Mathematics (Amsterdam)*, Elsevier/Academic Press, Amsterdam, The Netherlands, 2nd edition, 2003.
[17] A. R. Barron, "Neural net approximation," in *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems*, K. Narendra, Ed., pp. 69–72, Yale University Press, 1992.
[18] D. E. Knuth, "Big omicron and big omega and big theta," *SIGACT News*, vol. 8, pp. 18–24, 1976.
[19] G. Gnecco and M. Sanguineti, "Suboptimal solutions to network team optimization problems," in *Proceedings of the International Network Optimization Conference (INOC '09)*, April 2009.
[20] G. Gnecco and M. Sanguineti, "Suboptimal solutions to dynamic optimization problems via approximations of the policy functions," *Journal of Optimization Theory and Applications*, vol. 146, no. 3, pp. 764–794, 2010.
[21] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, 1989.
[22] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[23] M. Bóna, *A Walk through Combinatorics: An Introduction to Enumeration and Graph Theory*, World Scientific, River Edge, NJ, USA, 2002.
[24] Y. Makovoz, "Uniform approximation by neural networks," *Journal of Approximation Theory*, vol. 95, no. 2, pp. 215–228, 1998.
[25] Joseph E. Yukich, Maxwell B. Stinchcombe, and White, "Sup-norm approximation bounds for networks through probabilistic methods," *IEEE Transactions on Information Theory*, vol. 41, no. 4, pp. 1021–1027, 1995.
[26] E. S. Levitin and B. T. Polyak, "Convergence of minimizing sequences in conditional extremum problems," *Doklady Akademii Nauk SSSR*, vol. 168, pp. 764–767, 1966.
[27] A. A. Vladimirov, Y. E. Nesterov, and Y. N. Chekanov, "On uniformly convex functionals," *Vestnik Moskovskogo Universiteta. Seriya 15—Vychislitel'naya Matematika i Kibernetika*, vol. 3, pp. 12–23, 1978, English translation: Moscow University Computational Mathematics and Cybernetics, pp. 10–21, 1979.
[28] A. L. Dontchev, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, vol. 52 of *Lecture Notes in Control and Information Sciences*, Springer, Berlin, Germany, 1983.
[29] T. Zhang, "Sequential greedy approximation for certain convex optimization problems," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 682–691, 2003.